Solutions Manual
for the book
Introduction to Probability, Second Edition
Joseph K. Blitzstein and Jessica Hwang
© Chapman & Hall/CRC Press, 2019

Joseph K. Blitzstein and Jessica Hwang
Departments of Statistics, Harvard University and Stanford University

November 10, 2020

Contents

1

Probability and counting

Counting

1. How many ways are there to permute the letters in the word MISSISSIPPI?

   Solution:

   The word has 4 S's, 4 I's, 2 P's, and 1 M. Let's choose where to put the S's, then where to put the I's, then where to put the P's, and then the location of the M is determined. By the multiplication rule, there are

   $$\binom{11}{4}\binom{7}{4}\binom{3}{2} = 34650$$

   possibilities. Alternatively, we can start with 11! and adjust for overcounting:

   $$\frac{11!}{4!4!2!} = 34650.$$

2. (a) How many 7-digit phone numbers are possible, assuming that the first digit can't be a 0 or a 1?

   (b) Re-solve (a), except now assume also that the phone number is not allowed to start with 911 (since this is reserved for emergency use, and it would not be desirable for the system to wait to see whether more digits were going to be dialed after someone has dialed 911).

   Solution:

   (a) By the multiplication rule, there are $8 \cdot 10^6$ possibilities.

   (b) There are 10^4 phone numbers in (a) that start with 911 (again by the multiplication rule, since the first 3 digits are 911 and the remaining 4 digits are unconstrained). Excluding these and using the result of (a), the number of possibilities is

   $$8 \cdot 10^6 - 10^4 = 7990000.$$

3. Fred is planning to go out to dinner each night of a certain week, Monday through Friday, with each dinner being at one of his ten favorite restaurants.

   (a) How many possibilities are there for Fred's schedule of dinners for that Monday through Friday, if Fred is not willing to eat at the same restaurant more than once?

   (b) How many possibilities are there for Fred's schedule of dinners for that Monday through Friday, if Fred is willing to eat at the same restaurant more than once, but is not willing to eat at the same place twice in a row (or more)?

   Solution:

   (a) By the multiplication rule, there are $10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 = 30240$ possibilities.

   (b) By the multiplication rule, there are $10 \cdot 9^4 = 65610$ possibilities, since Monday's dinner can be at any of the 10 restaurants, and for Tuesday through Friday, each dinner can be at any of the 10 restaurants except the one where Fred ate on the previous night.

4. A *round-robin tournament* is being held with n tennis players; this means that every player will play against every other player exactly once.

(a) How many possible outcomes are there for the tournament (the outcome lists out who won and who lost for each game)?

(b) How many games are played in total?

Solution:

(a) For each of the $\binom{n}{2}$ unordered pairs of players, there is 1 game, so there are $2^{\binom{n}{2}}$ possible outcomes for the tournament.

(b) There are $\binom{n}{2}$ games played, as noted in (a).

5. A *knock-out tournament* is being held with 2^n tennis players. This means that for each round, the winners move on to the next round and the losers are eliminated, until only one person remains. For example, if initially there are $2^4 = 16$ players, then there are 8 games in the first round, then the 8 winners move on to round 2, then the 4 winners move on to round 3, then the 2 winners move on to round 4, the winner of which is declared the winner of the tournament. (There are various systems for determining who plays whom within a round, but these do not matter for this problem.)

(a) How many rounds are there?

(b) Count how many games in total are played, by adding up the numbers of games played in each round.

(c) Count how many games in total are played, this time by directly thinking about it without doing almost any calculation.

Hint: How many players need to be eliminated?

Solution:

(a) There are n rounds, since each round cuts the number of remaining players in half.

(b) There are $2^n/2 = 2^{n-1}$ games in the first round, then 2^{n-2} games in the second round, and so on, until there are only 2 players left in the final round. Using the formula for the sum of a finite geometric series (see the math appendix), the total number of games is

$$1 + 2 + 2^2 + \cdots + 2^{n-1} = \frac{1 - 2^n}{1 - 2} = 2^n - 1.$$

(c) A much easier way to see that the number of games is $2^n - 1$ is to note that each game eliminates one player, and $2^n - 1$ players need to be eliminated to leave one winner.

6. There are 20 people at a chess club on a certain day. They each find opponents and start playing. How many possibilities are there for how they are matched up, assuming that in each game it *does* matter who has the white pieces (in a chess game, one player has the white pieces and the other player has the black pieces)?

Solution: There are $\frac{20!}{2^{10} \cdot 10!}$ ways to determine who plays whom without considering color, by the multiplication rule or the result of Example 1.5.4 (Partnerships). For each game, there are 2 choices for who has the white pieces, so overall the number of possibilities is

$$\frac{2^{10} \cdot 20!}{2^{10} \cdot 10!} = \frac{20!}{10!} = 670442572800.$$

Alternatively, imagine a long table with 10 chessboards in a row, with 10 chairs on each

side of the table (for the chess players to sit in). Chess pieces are set up at each board, oriented so that the white pieces are all on one side of the table and the black pieces are all on the other side of the table. There are 20! ways for the chess players to be assigned to chairs, and once the players are seated in their chairs, a configuration for the chess games has been determined. This procedure overcounts by a factor of 10! since, after matching up the players, we can permute the players on one side of the table in any way, as long as we also permute the players on the other side of the table in the same way. So again the number of possibilities is $\frac{20!}{10!}$.

7. Two chess players, A and B, are going to play 7 games. Each game has three possible outcomes: a win for A (which is a loss for B), a draw (tie), and a loss for A (which is a win for B). A win is worth 1 point, a draw is worth 0.5 points, and a loss is worth 0 points.

(a) How many possible outcomes for the individual games are there, such that overall player A ends up with 3 wins, 2 draws, and 2 losses?

(b) How many possible outcomes for the individual games are there, such that A ends up with 4 points and B ends up with 3 points?

(c) Now assume that they are playing a best-of-7 match, where the match will end when either player has 4 points or when 7 games have been played, whichever is first. For example, if after 6 games the score is 4 to 2 in favor of A, then A wins the match and they don't play a 7th game. How many possible outcomes for the individual games are there, such that the match lasts for 7 games and A wins by a score of 4 to 3?

Solution:

(a) Writing W for win, D for draw, and L for loss (for player A), an outcome of the desired form is any permutation of WWWDDLL. So there are

$$\frac{7!}{3!2!2!} = 210$$

possible outcomes of the desired form.

(b) To end up with 4 points, A needs to have one of the following results: (i) 4 wins and 3 losses; (ii) 3 wins, 2 draws, and 2 losses; (iii) 2 wins, 4 draws, and 1 loss; or (iv) 1 win and 6 draws. Reasoning as in (a) and adding up these possibilities, there are

$$\frac{7!}{4!3!} + \frac{7!}{3!2!2!} + \frac{7!}{2!4!1!} + \frac{7!}{1!6!} = 357$$

possible outcomes of the desired form.

(c) For the desired outcomes, either (i) player A is ahead 3.5 to 2.5 after 6 games and then draws game 7, or (ii) the match is tied (3 to 3) after 6 games and then player A wins game 7. Reasoning as in (b), there are

$$\frac{6!}{3!1!2!} + \frac{6!}{2!3!1!} + \frac{6!}{1!5!} = 126$$

possibilities of type (i) and

$$\frac{6!}{3!3!} + \frac{6!}{2!2!2!} + \frac{6!}{1!4!1!} + 1 = 141$$

possibilities of type (ii), so overall there are

$$126 + 141 = 267$$

possible outcomes of the desired form.

8. Ⓢ (a) How many ways are there to split a dozen people into 3 teams, where one team has 2 people, and the other two teams have 5 people each?

(b) How many ways are there to split a dozen people into 3 teams, where each team has 4 people?

Solution:

(a) Pick any 2 of the 12 people to make the 2 person team, and then any 5 of the remaining 10 for the first team of 5, and then the remaining 5 are on the other team of 5; this overcounts by a factor of 2 though, since there is no designated "first" team of 5. So the number of possibilities is $\binom{12}{2}\binom{10}{5}/2 = 8316$. Alternatively, politely ask the 12 people to line up, and then let the first 2 be the team of 2, the next 5 be a team of 5, and then last 5 be a team of 5. There are 12! ways for them to line up, but it does not matter which order they line up in *within* each group, nor does the order of the 2 teams of 5 matter, so the number of possibilities is $\frac{12!}{2!5!5!\cdot 2} = 8316$.

(b) By either of the approaches above, there are $\frac{12!}{4!4!4!}$ ways to divide the people into a Team A, a Team B, and a Team C, if we care about which team is which (this is called a *multinomial coefficient*). Since here it doesn't matter which team is which, this over counts by a factor of 3!, so the number of possibilities is $\frac{12!}{4!4!4!3!} = 5775$.

9. Ⓢ (a) How many paths are there from the point $(0,0)$ to the point $(110, 111)$ in the plane such that each step either consists of going one unit up or one unit to the right?

(b) How many paths are there from $(0,0)$ to $(210, 211)$, where each step consists of going one unit up or one unit to the right, and the path has to go through $(110, 111)$?

Solution:

(a) Encode a path as a sequence of U's and R's, like $URURURUUUR\ldots UR$, where U and R stand for "up" and "right" respectively. The sequence must consist of 110 R's and 111 U's, and to determine the sequence we just need to specify where the R's are located. So there are $\binom{221}{110}$ possible paths.

(b) There are $\binom{221}{110}$ paths to $(110, 111)$, as above. From there, we need 100 R's and 100 U's to get to $(210, 211)$, so by the multiplication rule the number of possible paths is $\binom{221}{110} \cdot \binom{200}{100}$.

10. To fulfill the requirements for a certain degree, a student can choose to take any 7 out of a list of 20 courses, with the constraint that at least 1 of the 7 courses must be a statistics course. Suppose that 5 of the 20 courses are statistics courses.

(a) How many choices are there for which 7 courses to take?

(b) Explain intuitively why the answer to (a) is *not* $\binom{5}{1} \cdot \binom{19}{6}$.

Solution:

(a) There are $\binom{20}{7}$ ways to choose 7 courses if there are no constraints, but $\binom{15}{7}$ of these have no statistics courses. So there are

$$\binom{20}{7} - \binom{15}{7} = 71085$$

sets of 7 courses that contain at least one statistics course.

(b) An incorrect argument would be "there are $\binom{5}{1}$ to choose a statistics course (let's knock that requirement out of the way, then we can choose any other 6 courses) and then $\binom{19}{6}$ choices for the remaining 6 courses." This is incorrect since it's possible (and

often a good idea!) to take more than one statistics course. A possibility containing, for example, the statistics courses Stat 110 and Stat 111 together with 5 non-statistics courses would be counted twice in $\binom{5}{1} \cdot \binom{19}{6}$, once with Stat 110 as the choice for the $\binom{5}{1}$ term and once with Stat 111 as the choice. So it makes sense that the true answer is much less than $\binom{5}{1} \cdot \binom{19}{6}$.

11. Let A and B be sets with $|A| = n, |B| = m$.

(a) How many functions are there from A to B (i.e., functions with domain A, assigning an element of B to each element of A)?

(b) How many one-to-one functions are there from A to B (see Section A.2.1 of the math appendix for information about one-to-one functions)?

Solution:

(a) By the multiplication rule, there are m^n functions f from A to B, since for each $a \in A$ there are m possible ways to define $f(a)$.

(b) Now values can't be repeated, so there are $m \cdot (m - 1) \cdot (m - 2) \cdots (m - n + 1)$ possibilities for $n \le m$, and there are no possibilities for $n > m$.

12. Four players, named A, B, C, and D, are playing a card game. A standard, well-shuffled deck of cards is dealt to the players (so each player receives a 13-card hand).

(a) How many possibilities are there for the hand that player A will get? (Within a hand, the order in which cards were received doesn't matter.)

(b) How many possibilities are there overall for what hands everyone will get, assuming that it matters which player gets which hand, but not the order of cards within a hand?

(c) Explain intuitively why the answer to Part (b) is not the fourth power of the answer to Part (a).

Solution:

(a) There are $\binom{52}{13}$ possibilities since player A gets 13 out of 52 cards, without replacement and with order not mattering.

(b) There are $\binom{52}{13}$ possibilities for A's hand. For each of these, there are $\binom{39}{13}$ possibilities for B's hand. For each of these, there are $\binom{26}{13}$ possibilities for C's hand. After 3 hands have been determined, the 4th is also determined. So the number of possibilities is

$$\binom{52}{13}\binom{39}{13}\binom{26}{13} = \frac{52!}{(13!)^4} \approx 5.36 \times 10^{28}.$$

The expression with factorials could have been obtained directly by imagining shuffling the cards and giving the first 13 to A, the next 13 to B etc., and then adjusting for the fact that the order of cards *within* each player's hand doesn't matter.

As Wikipedia remarks (at `http://en.wikipedia.org/wiki/Bridge_probabilities` as of December 1, 2014), "The immenseness of this number can be understood by answering the question '*How large an area would you need to spread all possible bridge deals if each deal would occupy only one square millimeter?*'. The answer is: *an area more than a hundred million times the total area of Earth.*"

(c) The answer to (b), though an extremely large number, is much smaller than the fourth power of the answer to (a) since the cards are dealt *without replacement*. This makes the number of possibilities $\binom{52}{13}\binom{39}{13}\binom{26}{13}\binom{13}{13}$ rather than $\binom{52}{13}\binom{52}{13}\binom{52}{13}\binom{52}{13}$.

13. A certain casino uses 10 standard decks of cards mixed together into one big deck, which we will call a *superdeck*. Thus, the superdeck has $52 \cdot 10 = 520$ cards, with 10 copies of each card. How many different 10-card hands can be dealt from the superdeck? The order of the cards does not matter, nor does it matter which of the original 10 decks the cards came from. Express your answer as a binomial coefficient.

    Hint: Bose-Einstein.

    Solution: A hand is determined by specifying how many times each of the 52 different cards occurs. Number the 52 cards in a standard deck as $1, 2, \ldots, 52$, and let x_i be the number of times card i occurs in a hand (e.g., 3 for the Ace of Spades in the above example). Then the x_i are nonnegative integers with

    $$x_1 + x_2 + \cdots + x_{52} = 10.$$

    By Bose-Einstein, the number of solutions is

    $$\binom{52 + 10 - 1}{10} = \binom{61}{10} \approx 9.018 \times 10^{10}.$$

14. You are ordering two pizzas. A pizza can be small, medium, large, or extra large, with any combination of 8 possible toppings (getting no toppings is allowed, as is getting all 8). How many possibilities are there for your two pizzas?

    Solution: For one pizza, there are $4 \cdot 2^8 = 2^{10}$ possibilities. For two pizzas, there are $\binom{2^{10}}{2}$ ways to choose two distinct kinds of pizza, and 2^{10} possibilities with two copies of the same kind of pizza. So the number of possibilities is

    $$\binom{2^{10}}{2} + 2^{10} = 524800.$$

    Alternatively, think of choosing 2 pizza types with replacement, where order doesn't matter. Then Bose-Einstein gives the same answer: the number of possibilities is

    $$\binom{2^{10} + 2 - 1}{2} = 524800.$$

Story proofs

15. ⓢ Give a story proof that $\sum_{k=0}^{n} \binom{n}{k} = 2^n$.

    Solution: Consider picking a subset of n people. There are $\binom{n}{k}$ choices with size k, on the one hand, and on the other hand there are 2^n subsets by the multiplication rule.

16. ⓢ Show that for all positive integers n and k with $n \geq k$,

    $$\binom{n}{k} + \binom{n}{k-1} = \binom{n+1}{k},$$

    doing this in two ways: (a) algebraically and (b) with a story, giving an interpretation for why both sides count the same thing.

    Hint for the story proof: Imagine $n + 1$ people, with one of them pre-designated as "president".

    Solution:

(a) For the algebraic proof, start with the definition of the binomial coefficients in the left-hand side, and do some algebraic manipulation as follows:

$$
\begin{aligned}
\binom{n}{k} + \binom{n}{k-1} &= \frac{n!}{k!(n-k)!} + \frac{n!}{(k-1)!(n-k+1)!} \\
&= \frac{(n-k+1)n! + (k)n!}{k!(n-k+1)!} \\
&= \frac{n!(n+1)}{k!(n-k+1)!} \\
&= \binom{n+1}{k}.
\end{aligned}
$$

(b) For the story proof, consider $n+1$ people, with one of them pre-designated as "president". The right-hand side is the number of ways to choose k out of these $n+1$ people, with order not mattering. The left-hand side counts the same thing in a different way, by considering two cases: the president is or isn't in the chosen group.

The number of groups of size k which include the president is $\binom{n}{k-1}$, since once we fix the president as a member of the group, we only need to choose another $k-1$ members out of the remaining n people. Similarly, there are $\binom{n}{k}$ groups of size k that don't include the president. Thus, the two sides of the equation are equal.

17. Give a story proof that

$$
\sum_{k=0}^{n} \binom{n}{k}^2 = \binom{2n}{n},
$$

for all positive integers n.

Solution: Consider a club consisting of n sophomores and n juniors, from which a committee of size n will be selected. There are $\binom{2n}{n}$ possibilities for the composition of the committee; this is the right-hand side of the identity. If the committee has exactly k sophomores, then it must have exactly $n-k$ juniors. The number of possibilities for the committee with exactly k sophomores is

$$
\binom{n}{k}\binom{n}{n-k} = \binom{n}{k}^2,
$$

by the multiplication rule and Example 1.5.1. Summing over all possible values of k yields the left-hand side of the identity. Note that the identity can also be thought of as a special case of Vandermonde's identity, after writing

$$
\sum_{k=0}^{n} \binom{n}{k}^2 = \sum_{k=0}^{n} \binom{n}{k}\binom{n}{n-k}.
$$

18. Give a story proof that

$$
\sum_{k=1}^{n} k\binom{n}{k}^2 = n\binom{2n-1}{n-1},
$$

for all positive integers n.

Hint: Consider choosing a committee of size n from two groups of size n each, where only one of the two groups has people eligible to become the chair of the committee.

Solution:

Imagine that there are n juniors and n seniors in a certain club. A committee of size n is

chosen, and one of these people becomes the chair. Suppose though that the chair must be a senior. Letting k be the number of seniors on the committee, there are $\binom{n}{k}$ ways to choose the seniors, $\binom{n}{n-k} = \binom{n}{k}$ ways to choose the juniors, and after these choices are made there are k choices of chair. So the overall number of possibilities is the left-hand side of the identity.

Alternatively, we can choose the chair *first* (as any of the n seniors), and then choose any $n - 1$ of the remaining $2n - 1$ people to form the rest of the committee. This gives the right-hand side of the identity.

19. Give a story proof that

$$\sum_{k=2}^{n} \binom{k}{2}\binom{n-k+2}{2} = \binom{n+3}{5},$$

for all integers $n \geq 2$.

Hint: Consider the middle number in a subset of $\{1, 2, \ldots, n+3\}$ of size 5.

Solution: There are $\binom{n+3}{5}$ subsets of $\{1, 2, \ldots, n+3\}$ of size 5; this is the right-hand side of the identity. Consider such a subset, and order it from smallest to largest. Let j be the middle number (i.e., the third number on the sorted list), and note that $3 \leq j \leq n+1$. Given j, there are $\binom{j-1}{2}$ choices for the 2 numbers in the subset that are less than j, since these can be any two numbers in $1, 2, \ldots, j-1$. Similarly, there are $\binom{(n+3)-(j+1)+1}{2}$ choices for the 2 numbers in the subset that are greater than j, since these can be any two numbers in $j + 1, \ldots, n + 3$. Summing over the possible values of j gives

$$\sum_{j=3}^{n+1} \binom{j-1}{2}\binom{(n+3)-(j+1)+1}{2} = \sum_{k=2}^{n} \binom{k}{2}\binom{n-k+2}{2},$$

which is the left-hand side of the identity.

20. Ⓢ (a) Show using a story proof that

$$\binom{k}{k} + \binom{k+1}{k} + \binom{k+2}{k} + \cdots + \binom{n}{k} = \binom{n+1}{k+1},$$

where n and k are positive integers with $n \geq k$. This is called the *hockey stick identity*. Hint: Imagine arranging a group of people by age, and then think about the oldest person in a chosen subgroup.

(b) Suppose that a large pack of Haribo gummi bears can have anywhere between 30 and 50 gummi bears. There are 5 delicious flavors: pineapple (clear), raspberry (red), orange (orange), strawberry (green, mysteriously), and lemon (yellow). There are 0 non-delicious flavors. How many possibilities are there for the composition of such a pack of gummi bears? You can leave your answer in terms of a couple binomial coefficients, but not a sum of lots of binomial coefficients.

Solution:

(a) Consider choosing $k+1$ people out of a group of $n+1$ people. Call the oldest person in the subgroup "Aemon". If Aemon is also the oldest person in the full group, then there are $\binom{n}{k}$ choices for the rest of the subgroup. If Aemon is the second oldest in the full group, then there are $\binom{n-1}{k}$ choices since the oldest person in the full group can't be chosen. In general, if there are j people in the full group who are younger than Aemon, then there are $\binom{j}{k}$ possible choices for the rest of the subgroup. Thus,

$$\sum_{j=k}^{n} \binom{j}{k} = \binom{n+1}{k+1}.$$

(b) For a pack of i gummi bears, there are $\binom{5+i-1}{i} = \binom{i+4}{i} = \binom{i+4}{4}$ possibilities since the situation is equivalent to getting a sample of size i from the $n = 5$ flavors (with replacement, and with order not mattering). So the total number of possibilities is

$$\sum_{i=30}^{50} \binom{i+4}{4} = \sum_{j=34}^{54} \binom{j}{4}.$$

Applying the previous part, we can simplify this by writing

$$\sum_{j=34}^{54} \binom{j}{4} = \sum_{j=4}^{54} \binom{j}{4} - \sum_{j=4}^{33} \binom{j}{4} = \binom{55}{5} - \binom{34}{5}.$$

(This works out to 3200505 possibilities!)

21. Define $\left\{ {n \atop k} \right\}$ as the number of ways to partition $\{1, 2, \ldots, n\}$ into k nonempty subsets, or the number of ways to have n students split up into k groups such that each group has at least one student. For example, $\left\{ {4 \atop 2} \right\} = 7$ because we have the following possibilities:

- $\{1\}, \{2, 3, 4\}$
- $\{2\}, \{1, 3, 4\}$
- $\{3\}, \{1, 2, 4\}$
- $\{4\}, \{1, 2, 3\}$

- $\{1, 2\}, \{3, 4\}$
- $\{1, 3\}, \{2, 4\}$
- $\{1, 4\}, \{2, 3\}$

Prove the following identities:

(a)
$$\left\{ {n+1 \atop k} \right\} = \left\{ {n \atop k-1} \right\} + k \left\{ {n \atop k} \right\}.$$

Hint: I'm either in a group by myself or I'm not.

(b)
$$\sum_{j=k}^{n} \binom{n}{j} \left\{ {j \atop k} \right\} = \left\{ {n+1 \atop k+1} \right\}.$$

Hint: First decide how many people are not going to be in my group.

Solution:

(a) The left-hand side is the number of ways to divide $n + 1$ people into k nonempty groups. Now let's count this a different way. Say I'm the $(n+1)$st person. Either I'm in a group by myself or I'm not. If I'm in a group by myself, then there are $\left\{ {n \atop k-1} \right\}$ ways to divide the remaining n people into $k - 1$ nonempty groups. Otherwise, the n people other than me form k nonempty groups, which can be done in $\left\{ {n \atop k} \right\}$ ways, and then I can join any of those k groups. So in total, there are $\left\{ {n \atop k-1} \right\} + k \left\{ {n \atop k} \right\}$ possibilities, which is the right-hand side.

(b) The right-hand side is the number of ways to divide $n+1$ people into $k+1$ nonempty groups. Say I'm the $(n + 1)$st person. Let j be the number of people *not* in my group. Then $k \le j \le n$. The number of possible divisions with j people not in my group is $\binom{n}{j} \left\{ {j \atop k} \right\}$ since we have $\binom{n}{j}$ possibilities for which j specific people are not in my group (and then it's determined who *is* in my group) and then $\left\{ {j \atop k} \right\}$ possibilities for how to divide those j people into k groups that are not my group. Summing over all possible j gives the left-hand side.

Note: The quantities $\left\{ {n \atop k} \right\}$ are known as *Stirling numbers of the second kind*.

22. The Dutch mathematician R.J. Stroeker remarked:

    Every beginning student of number theory surely must have marveled at the miraculous fact that for each natural number n the sum of the first n positive consecutive cubes is a perfect square. [29]

    Furthermore, it is the square of the sum of the first n positive integers! That is,

    $$1^3 + 2^3 + \cdots + n^3 = (1 + 2 + \cdots + n)^2.$$

    Usually this identity is proven by induction, but that does not give much insight into why the result is true, nor does it help much if we wanted to compute the left-hand side but didn't already know this result. In this problem, you will give a story proof of the identity.

    (a) Give a story proof of the identity

    $$1 + 2 + \cdots + n = \binom{n+1}{2}.$$

    Hint: Consider a round-robin tournament (see Exercise 4).

    (b) Give a story proof of the identity

    $$1^3 + 2^3 + \cdots + n^3 = 6\binom{n+1}{4} + 6\binom{n+1}{3} + \binom{n+1}{2}.$$

    It is then just basic algebra (not required for this problem) to check that the square of the right-hand side in (a) is the right-hand side in (b).

    Hint: Imagine choosing a number between 1 and n and then choosing 3 numbers between 0 and n smaller than the original number, with replacement. Then consider cases based on how many distinct numbers were chosen.

    Solution:

    (a) Consider a chess tournament with $n + 1$ players, where everyone plays against everyone else once. A total of $\binom{n+1}{2}$ games are played. Label the players $0, 1, \ldots, n$. Player 0 plays n games, player 1 plays $n - 1$ games not already accounted for, player 2 plays $n - 2$ games not already accounted for, etc. So

    $$n + (n - 1) + (n - 2) + \cdots + 1 = \binom{n+1}{2}.$$

    (b) Following the hint, let us count the number of choices of (i, j, k, l) where i is greater than j, k, l. Given i, there are i^3 choices for (j, k, l), which gives the left-hand side. On the other hand, consider 3 cases: there could be 2, 3, or 4 distinct numbers chosen. There are $\binom{n+1}{2}$ ways to choose 2 distinct numbers from $0, 1, \ldots, n$, giving, e.g., $(3, 1, 1, 1)$. There are $\binom{n+1}{4}$ ways to choose 4 distinct numbers, giving, e.g., $(5, 2, 1, 4)$, but the $(2, 1, 4)$ could be permuted in any order so we multiply by 6. There are $\binom{n+1}{3}$ ways to choose 3 distinct numbers, giving, e.g., $(4, 2, 2, 1)$, but the $2, 2, 1$ can be in any order and could have been $1, 1, 2$ in any order also, again giving a factor of 6. Adding these cases gives the right-hand side.

Naive definition of probability

23. Three people get into an empty elevator at the first floor of a building that has 10 floors. Each presses the button for their desired floor (unless one of the others has already pressed that button). Assume that they are equally likely to want to go to

floors 2 through 10 (independently of each other). What is the probability that the buttons for 3 consecutive floors are pressed?

Solution: The number of possible outcomes for who is going to which floor is 9^3. There are 7 possibilities for which buttons are pressed such that there are 3 consecutive floors: $(2, 3, 4), (3, 4, 5), \ldots, (8, 9, 10)$. For each of these 7 possibilities, there are 3! ways to choose who is going to which floor. So by the naive definition, the probability is

$$\frac{3! \cdot 7}{9^3} = \frac{42}{729} = \frac{14}{243} \approx 0.0576.$$

24. Ⓢ A certain family has 6 children, consisting of 3 boys and 3 girls. Assuming that all birth orders are equally likely, what is the probability that the 3 eldest children are the 3 girls?

Solution: Label the girls as $1, 2, 3$ and the boys as $4, 5, 6$. Think of the birth order is a permutation of $1, 2, 3, 4, 5, 6$, e.g., we can interpret 314265 as meaning that child 3 was born first, then child 1, etc. The number of possible permutations of the birth orders is 6!. Now we need to count how many of these have all of $1, 2, 3$ appear before all of $4, 5, 6$. This means that the sequence must be a permutation of $1, 2, 3$ followed by a permutation of $4, 5, 6$. So with all birth orders equally likely, we have

$$P(\text{the 3 girls are the 3 eldest children}) = \frac{(3!)^2}{6!} = 0.05.$$

Alternatively, we can use the fact that there are $\binom{6}{3}$ ways to choose where the girls appear in the birth order (without taking into account the ordering of the girls amongst themselves). These are all equally likely. Of these possibilities, there is only 1 where the 3 girls are the 3 eldest children. So again the probability is $\frac{1}{\binom{6}{3}} = 0.05$.

25. Ⓢ A city with 6 districts has 6 robberies in a particular week. Assume the robberies are located randomly, with all possibilities for which robbery occurred where equally likely. What is the probability that some district had more than 1 robbery?

Solution: There are 6^6 possible configurations for which robbery occurred where. There are 6! configurations where each district had exactly 1 of the 6, so the probability of the complement of the desired event is $6!/6^6$. So the probability of some district having more than 1 robbery is

$$1 - 6!/6^6 \approx 0.9846.$$

Note that this also says that if a fair die is rolled 6 times, there's over a 98% chance that some value is repeated!

26. A survey is being conducted in a city with 1 million residents. It would be far too expensive to survey all of the residents, so a random sample of size 1000 is chosen (in practice, there are many challenges with sampling, such as obtaining a complete list of everyone in the city, and dealing with people who refuse to participate). The survey is conducted by choosing people one at a time, *with* replacement and with equal probabilities.

(a) Explain how sampling with vs. without replacement here relates to the birthday problem.

(b) Find the probability that at least one person will get chosen more than once.

Solution:

(a) In the survey problem, people are sampled one by one, and each person randomly is any of the 10^6 residents in the city; in the birthday problem, people show up at a party one by one, and each person randomly has any of 365 possible birthdays. The fact

that the same person can be chosen more than once when sampling with replacement is analogous to the fact that more than one person can have the same birthday.

(b) This problem is isomorphic to (has the same structure as) the birthday problem. By the naive definition of probability, the probability of no match is

$$\frac{10^6(10^6 - 1)(10^6 - 2)\cdots(10^6 - 999)}{(10^6)^{1000}} = \left(1 - \frac{1}{10^6}\right)\left(1 - \frac{2}{10^6}\right)\cdots\left(1 - \frac{999}{10^6}\right).$$

The probability of at least one person being chosen more than once is

$$1 - \left(1 - \frac{1}{10^6}\right)\left(1 - \frac{2}{10^6}\right)\cdots\left(1 - \frac{999}{10^6}\right) \approx 0.393.$$

27. A *hash table* is a commonly used data structure in computer science, allowing for fast information retrieval. For example, suppose we want to store some people's phone numbers. Assume that no two of the people have the same name. For each name x, a *hash function h* is used, letting $h(x)$ be the location that will be used to store x's phone number. After such a table has been computed, to look up x's phone number one just recomputes $h(x)$ and then looks up what is stored in that location.

The hash function h is deterministic, since we don't want to get different results every time we compute $h(x)$. But h is often chosen to be *pseudorandom*. For this problem, assume that true randomness is used. Let there be k people, with each person's phone number stored in a random location (with equal probabilities for each location, independently of where the other people's numbers are stored), represented by an integer between 1 and n. Find the probability that at least one location has more than one phone number stored there.

Solution:

This problem has the same structure as the birthday problem. For $k > n$, the probability is 1 since then there are more people than locations. For $k \leq n$, the probability is

$$1 - \frac{n(n-1)\ldots(n-k+1)}{n^k}.$$

28. Ⓢ A college has 10 (non-overlapping) time slots for its courses, and blithely assigns courses to time slots randomly and independently. A student randomly chooses 3 of the courses to enroll in. What is the probability that there is a conflict in the student's schedule?

Solution: The probability of no conflict is $\frac{10\cdot9\cdot8}{10^3} = 0.72$. So the probability of there being at least one scheduling conflict is 0.28.

29. Ⓢ For each part, decide whether the blank should be filled in with $=, <,$ or $>$, and give a clear explanation.

(a) (probability that the total after rolling 4 fair dice is 21) ____ (probability that the total after rolling 4 fair dice is 22)

(b) (probability that a random 2-letter word is a palindrome[1]) ____ (probability that a random 3-letter word is a palindrome)

Solution:

---

[1]A *palindrome* is an expression such as "A man, a plan, a canal: Panama" that reads the same backwards as forwards (ignoring spaces, capitalization, and punctuation). Assume for this problem that all words of the specified length are equally likely, that there are no spaces or punctuation, and that the alphabet consists of the lowercase letters a,b,...,z.

(a) $\boxed{>}$. All *ordered* outcomes are equally likely here. So for example with two dice, obtaining a total of 11 is more likely than obtaining a total of 12 since there are two ways to get a 5 and a 6, and only one way to get two 6's. To get a 21, the outcome must be a permutation of $(6, 6, 6, 3)$ (4 possibilities), $(6, 5, 5, 5)$ (4 possibilities), or $(6, 6, 5, 4)$ $(4!/2 = 12$ possibilities). To get a 22, the outcome must be a permutation of $(6, 6, 6, 4)$ (4 possibilities), or $(6, 6, 5, 5)$ $(4!/2^2 = 6$ possibilities). So getting a 21 is more likely; in fact, it is exactly twice as likely as getting a 22.

(b) $\boxed{=}$. The probabilities are equal, since for both 2-letter and 3-letter words, being a palindrome means that the first and last letter are the same.

30. With definitions as in the previous problem, find the probability that a random n-letter word is a palindrome for $n = 7$ and for $n = 8$.

    Solution:

    The probability of a random 7-letter word being a palindrome is

    $$\frac{26^4}{26^7} = \frac{1}{26^3} \approx 5.69 \times 10^{-5},$$

    since the first 4 letters can be chosen arbitrarily and then the last 3 letters are determined. Similarly, the probability for a random 8-letter word is

    $$\frac{26^4}{26^8} = \frac{1}{26^4} \approx 2.19 \times 10^{-6}.$$

31. Ⓢ Elk dwell in a certain forest. There are N elk, of which a simple random sample of size n are captured and tagged ("simple random sample" means that all $\binom{N}{n}$ sets of n elk are equally likely). The captured elk are returned to the population, and then a new sample is drawn, this time with size m. This is an important method that is widely used in ecology, known as *capture-recapture*. What is the probability that exactly k of the m elk in the new sample were previously tagged? (Assume that an elk that was captured before doesn't become more or less likely to be captured again.)

    Solution: We can use the naive definition here since we're assuming all samples of size m are equally likely. To have exactly k be tagged elk, we need to choose k of the n tagged elk, and then $m - k$ from the $N - n$ untagged elk. So the probability is

    $$\frac{\binom{n}{k} \cdot \binom{N-n}{m-k}}{\binom{N}{m}},$$

    for k such that $0 \le k \le n$ and $0 \le m - k \le N - n$, and the probability is 0 for all other values of k (for example, if $k > n$ the probability is 0 since then there aren't even k tagged elk in the entire population!). This is known as a *Hypergeometric* probability; we will encounter it again in Chapter 3.

32. Four cards are face down on a table. You are told that two are red and two are black, and you need to guess which two are red and which two are black. You do this by pointing to the two cards you're guessing are red (and then implicitly you're guessing that the other two are black). Assume that all configurations are equally likely, and that you do not have psychic powers. Find the probability that exactly j of your guesses are correct, for $j = 0, 1, 2, 3, 4$.

    Solution:

    There are $\binom{4}{2} = 6$ possibilities for where the two red cards are, all equally likely. So there is a 1/6 chance that you will pick both locations of red cards correctly (in which case you also get the locations of the black cards right). And there is a 1/6 chance that both

locations you choose actually contain black cards (in which case none of your guesses are correct). This leaves a 4/6 chance that the locations you picked as having red cards consist of 1 red card and 1 black card (in which case the other 2 locations also consist of 1 red card and 1 black card). Thus, for $j = 0, 1, 2, 3, 4$, the desired probabilities are $1/6, 0, 2/3, 0, 1/6$, respectively.

33. $\text{\textcircled{s}}$ A jar contains r red balls and g green balls, where r and g are fixed positive integers. A ball is drawn from the jar randomly (with all possibilities equally likely), and then a second ball is drawn randomly.

(a) Explain intuitively why the probability of the second ball being green is the same as the probability of the first ball being green.

(b) Define notation for the sample space of the problem, and use this to compute the probabilities from (a) and show that they are the same.

(c) Suppose that there are 16 balls in total, and that the probability that the two balls are the same color is the same as the probability that they are different colors. What are r and g (list all possibilities)?

Solution:

(a) This is true by *symmetry*. The first ball is equally likely to be any of the $g + r$ balls, so the probability of it being green is $g/(g+r)$. But the second ball is also equally likely to be any of the $g + r$ balls (there aren't certain balls that enjoy being chosen second and others that have an aversion to being chosen second); once we know whether the first ball is green we have information that affects our uncertainty about the second ball, but before we have this information, the second ball is equally likely to be any of the balls.

Alternatively, intuitively it shouldn't matter if we pick one ball at a time, or take one ball with the left hand and one with the right hand at the same time. By symmetry, the probabilities for the ball drawn with the left hand should be the same as those for the ball drawn with the right hand.

(b) Label the balls as $1, 2, \ldots, g + r$, such that $1, 2, \ldots, g$ are green and $g + 1, \ldots, g + r$ are red. The sample space can be taken to be the set of all pairs (a, b) with $a, b \in \{1, \ldots, g+r\}$ and $a \neq b$ (there are other possible ways to define the sample space, but it is important to specify all possible outcomes using clear notation, and it make sense to be as richly detailed as possible in the specification of possible outcomes, to avoid losing information). Each of these pairs is equally likely, so we can use the naive definition of probability. Let G_i be the event that the ith ball drawn is green.

The denominator is $(g+r)(g+r-1)$ by the multiplication rule. For G_1, the numerator is $g(g+r-1)$, again by the multiplication rule. For G_2, the numerator is also $g(g+r-1)$, since in counting favorable cases, there are g possibilities for the second ball, and for each of those there are $g + r - 1$ favorable possibilities for the first ball (note that the multiplication rule doesn't require the experiments to be listed in chronological order!); alternatively, there are $g(g-1) + rg = g(g+r-1)$ favorable possibilities for the second ball being green, as seen by considering 2 cases (first ball green and first ball red). Thus,

$$P(G_i) = \frac{g(g+r-1)}{(g+r)(g+r-1)} = \frac{g}{g+r},$$

for $i \in \{1, 2\}$, which concurs with (a).

(c) Let A be the event of getting one ball of each color. In set notation, we can write $A = (G_1 \cap G_2^c) \cup (G_1^c \cap G_2)$. We are given that $P(A) = P(A^c)$, so $P(A) = 1/2$. Then

$$P(A) = \frac{2gr}{(g+r)(g+r-1)} = \frac{1}{2},$$

giving the quadratic equation

$$g^2 + r^2 - 2gr - g - r = 0,$$

i.e.,

$$(g - r)^2 = g + r.$$

But $g + r = 16$, so $g - r$ is 4 or -4. Thus, either $g = 10, r = 6$, or $g = 6, r = 10$.

34. ⓢ A random 5-card poker hand is dealt from a standard deck of cards. Find the probability of each of the following possibilities (in terms of binomial coefficients).

(a) A flush (all 5 cards being of the same suit; do not count a royal flush, which is a flush with an ace, king, queen, jack, and 10).

(b) Two pair (e.g., two 3's, two 7's, and an ace).

Solution:

(a) A flush can occur in any of the 4 suits (imagine the tree, and for concreteness suppose the suit is Hearts); there are $\binom{13}{5}$ ways to choose the cards in that suit, except for one way to have a royal flush in that suit. So the probability is

$$\frac{4\left(\binom{13}{5} - 1\right)}{\binom{52}{5}}.$$

(b) Choose the two ranks of the pairs, which specific cards to have for those 4 cards, and then choose the extraneous card (which can be any of the $52 - 8$ cards not of the two chosen ranks). This gives that the probability of getting two pairs is

$$\frac{\binom{13}{2} \cdot \binom{4}{2}^2 \cdot 44}{\binom{52}{5}}.$$

35. A random 13-card hand is dealt from a standard deck of cards. What is the probability that the hand contains at least 3 cards of every suit?

Solution: The only to have at least 3 cards of every suit is to have exactly 4 cards from one suit and exactly 3 cards from each of the other suits. Choosing which suit the hand has 4 of, and then the specific cards from each suit, the probability is

$$\frac{4 \cdot \binom{13}{4} \cdot \binom{13}{3}^3}{\binom{52}{13}} \approx 0.1054.$$

36. A group of 30 dice are thrown. What is the probability that 5 of each of the values $1, 2, 3, 4, 5, 6$ appear?

Solution:

To get 5 dice for each of the 6 values, we can choose the 5 out of 30 dice that will be 1's, then the 5 out of the remaining 25 that will be 2's, etc. This gives

$$\binom{30}{5} \cdot \binom{25}{5} \cdot \binom{20}{5} \cdot \binom{15}{5} \cdot \binom{10}{5} \cdot \binom{5}{5} = \frac{30!}{(5!)^6}$$

possibilities; this is known as a *multinomial coefficient*, sometimes denoted as $\binom{30}{5,5,5,5,5,5}$. The right-hand side can also be obtained directly: it is the number of permutations of the sequence $1111122222\ldots66666$. By the naive definition, the probability is

$$\frac{30!}{(5!)^6 \cdot 6^{30}} \approx 0.000402.$$

37. A deck of cards is shuffled well. The cards are dealt one by one, until the first time an ace appears.

(a) Find the probability that no kings, queens, or jacks appear before the first ace.

(b) Find the probability that exactly one king, exactly one queen, and exactly one jack appear (in any order) before the first ace.

Solution:

(a) The 2's through 10's are irrelevant, so we can assume the deck consists of aces, kings, queens, and jacks. The event of interest is that the first card is an ace. This has probability 1/4 since the first card is equally likely to be any card.

(b) Continue as in (a). The probability that the deck starts as KQJA is

$$\frac{4^4 \cdot 12!}{16!} = \frac{8}{1365}.$$

The KQJ could be in any order, so the desired probability is

$$\frac{3! \cdot 8}{1365} = \frac{16}{455} \approx 0.0352.$$

Alternatively, note that there are $16 \cdot 15 \cdot 14 \cdot 13$ possibilities for the first 4 cards, of which $12 \cdot 8 \cdot 4 \cdot 4$ are favorable. So by the naive definition, the probability is

$$\frac{12 \cdot 8 \cdot 4 \cdot 4}{16 \cdot 15 \cdot 14 \cdot 13} \approx 0.0352.$$

38. Tyrion, Cersei, and ten other people are sitting at a round table, with their seating arrangement having been randomly assigned. What is the probability that Tyrion and Cersei are sitting next to each other? Find this in two ways:

(a) using a sample space of size 12!, where an outcome is fully detailed about the seating;

(b) using a much smaller sample space, which focuses on Tyrion and Cersei.

Solution:

(a) Label the seats in clockwise order as $1, 2, \ldots, 12$, starting from some fixed seat. Give the people other than Tyrion and Cersei ID numbers $1, 2, \ldots, 10$. The outcome is $(t, c, s_1, \ldots, s_{10})$, where t is Tyrion's seat assignment, c is Cersei's, and s_j is person j's.

To count the number of ways in which Tyrion and Cersei can be seated together, let Tyrion sit anywhere (12 possibilities), Cersei sit either to Tyrion's left or to his right (2 possibilities), and let everyone else fill the remaining 10 seats in any way (10! possibilities). By the multiplication rule and the naive definition, the probability is

$$\frac{12 \cdot 2 \cdot 10!}{12!} = \frac{12 \cdot 2 \cdot 10!}{12 \cdot 11 \cdot 10!} = \frac{2}{11}.$$

(b) Now let's just consider the $\binom{12}{2}$ possible seat assignments of Tyrion and Cersei, not worrying about which of these 2 seats goes to Tyrion or the details of where the other 10 people will sit. There are 12 assignments in which they sit together (without caring about order): $\{1, 2\}, \{2, 3\}, \ldots, \{11, 12\}, \{12, 1\}$. So the probability is

$$\frac{12}{\binom{12}{2}} = \frac{2}{11},$$

in agreement with (a).

39. An organization with $2n$ people consists of n married couples. A committee of size k is selected, with all possibilities equally likely. Find the probability that there are exactly j married couples within the committee.

    Solution: The probability is 0 if $j > n$ (not enough couples) or $k - 2j < 0$ (not enough space on the committee), so assume $0 \leq j \leq n$ and $k - 2j \geq 0$. There are $\binom{2n}{k}$ possible compositions of the committee. There are $\binom{n}{j}$ ways to choose which j married couples are on the committee. Once they are chosen, there are $\binom{n-j}{k-2j}$ ways to choose which of the other married couples are represented on the committee. For each of those $k - 2j$ couples, we then need to choose which person within the couple will be on the committee. Overall, the probability is
    $$\frac{\binom{n}{j}\binom{n-j}{k-2j}2^{k-2j}}{\binom{2n}{k}}.$$

40. There are n balls in a jar, labeled with the numbers $1, 2, \ldots, n$. A total of k balls are drawn, one by one *with replacement*, to obtain a sequence of numbers.

    (a) What is the probability that the sequence obtained is strictly increasing?

    (b) What is the probability that the sequence obtained is increasing? (Note: In this book, "increasing" means "nondecreasing".)

    Solution:

    (a) There is a one-to-one correspondence between strictly increasing sequences $a_1 < \cdots < a_k$ and subsets $\{a_1, \ldots, a_k\}$ of size k, so the probability is $\binom{n}{k}/n^k$.

    (b) There is a one-to-one correspondence between increasing sequences of length k and ways of choosing k balls with replacement, so the probability is $\binom{n+k-1}{k}/n^k$.

41. Each of n balls is independently placed into one of n boxes, with all boxes equally likely. What is the probability that exactly one box is empty?

    Solution: In order to have exactly one empty box, there must be one empty box, one box with two balls, and $n - 2$ boxes with one ball (if two or more boxes each had at least two balls, then there would not be enough balls left to avoid having more than one empty box). Choose which box is empty, then which has two balls, then assign balls to the boxes with one ball, and then it is determined which balls are in the box with two balls. This gives that the probability is
    $$\frac{n(n-1)n(n-1)(n-2)\ldots 3}{n^n} = \frac{n!(n-1)}{2 \cdot n^{n-1}}.$$

42. ⓢ A *norepeatword* is a sequence of at least one (and possibly all) of the usual 26 letters a,b,c,...,z, with repetitions not allowed. For example, "course" is a norepeatword, but "statistics" is not. Order matters, e.g., "course" is not the same as "source".
    A norepeatword is chosen randomly, with all norepeatwords equally likely. Show that the probability that it uses all 26 letters is very close to $1/e$.

    Solution: The number of norepeatwords having all 26 letters is the number of ordered arrangements of 26 letters: 26!. To construct a norepeatword with $k \leq 26$ letters, we first select k letters from the alphabet ($\binom{26}{k}$ selections) and then arrange them into a word ($k!$ arrangements). Hence there are $\binom{26}{k}k!$ norepeatwords with k letters, with k ranging from 1 to 26. With all norepeatwords equally likely, we have

$$
\begin{aligned}
P(\text{norepeatword having all 26 letters}) &= \frac{\#\text{ norepeatwords having all 26 letters}}{\#\text{ norepeatwords}} \\
&= \frac{26!}{\sum_{k=1}^{26}\binom{26}{k}k!} = \frac{26!}{\sum_{k=1}^{26}\frac{26!}{k!(26-k)!}k!} \\
&= \frac{1}{\frac{1}{25!} + \frac{1}{24!} + \ldots + \frac{1}{1!} + 1}.
\end{aligned}
$$

The denominator is the first 26 terms in the Taylor series $e^x = 1 + x + x^2/2! + \ldots$, evaluated at $x = 1$. Thus the probability is approximately $1/e$ (this is an *extremely* good approximation since the series for e converges very quickly; the approximation for e differs from the truth by less than 10^{-26}).

Axioms of probability

43. Show that for any events A and B,

$$P(A) + P(B) - 1 \leq P(A \cap B) \leq P(A \cup B) \leq P(A) + P(B).$$

For each of these three inequalities, give a simple criterion for when the inequality is actually an equality (e.g., give a simple condition such that $P(A \cap B) = P(A \cup B)$ if and only if the condition holds).

Solution: By Theorem 1.6.2, we have $P(A \cap B) \leq P(A \cup B)$ since $A \cap B \subseteq A \cup B$. Using the fact that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ and the fact that probability is always between 0 and 1, we have

$$P(A \cap B) = P(A) + P(B) - P(A \cup B) \geq P(A) + P(B) - 1$$

and

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \leq P(A) + P(B).$$

Now let us investigate when equality occurs for the above inequalities. Writing

$$P(A \cup B) = P(A \cap B) + P(A \cap B^c) + P(A^c \cap B),$$

we have that $P(A \cap B) = P(A \cup B)$ if and only if $P(A \cap B^c)$ and $P(A^c \cap B)$ are both 0. (This says that there is no probability mass in A but not in B or vice versa. The main example where this will hold is when A and B are the same event.)

Looking at the proof of $P(A) + P(B) - 1 \leq P(A \cap B)$, we see that equality will hold if and only if $P(A \cup B) = 1$. And looking at the proof of $P(A \cup B) \leq P(A) + P(B)$, we see that equality will hold if and only if $P(A \cap B) = 0$.

44. Let A and B be events. The *difference* $B - A$ is defined to be the set of all elements of B that are not in A. Show that if $A \subseteq B$, then

$$P(B - A) = P(B) - P(A),$$

directly using the axioms of probability.

Solution: Let $A \subseteq B$. The events A and $B - A$ are disjoint and their union is B, so

$$P(A) + P(B - A) = P(A \cup (B - A)) = P(B),$$

as desired.

45. Let A and B be events. The *symmetric difference* $A \triangle B$ is defined to be the set of all elements that are in A or B but not both. In logic and engineering, this event is also called the *XOR* (*exclusive or*) of A and B. Show that

$$P(A \triangle B) = P(A) + P(B) - 2P(A \cap B),$$

directly using the axioms of probability.

Solution: We have

$$P(A \triangle B) = P(A \cap B^c) + P(A^c \cap B),$$

since $A \triangle B$ is the union of the disjoint events $A \cap B^c$ and $A^c \cap B$. Similarly, we have

$$P(A) = P(A \cap B^c) + P(A \cap B),$$
$$P(B) = P(B \cap A^c) + P(B \cap A).$$

Adding the above two equations gives

$$P(A) + P(B) = P(A \cap B^c) + P(A^c \cap B) + 2P(A \cap B).$$

Thus,

$$P(A \triangle B) = P(A \cap B^c) + P(A^c \cap B) = P(A) + P(B) - 2P(A \cap B).$$

46. Let $A_1, A_2, \ldots, A_n$ be events. Let B_k be the event that exactly k of the A_i occur, and C_k be the event that at least k of the A_i occur, for $0 \leq k \leq n$. Find a simple expression for $P(B_k)$ in terms of $P(C_k)$ and $P(C_{k+1})$.

Solution: Saying that at least k of the A_i occur amounts to saying that either exactly k of the A_i occur or at least $k + 1$ occur. These are disjoint possibilities. So

$$P(C_k) = P(B_k) + P(C_{k+1}),$$

which gives

$$P(B_k) = P(C_k) - P(C_{k+1}).$$

47. Events A and B are *independent* if $P(A \cap B) = P(A)P(B)$ (independence is explored in detail in the next chapter).

(a) Give an example of independent events A and B in a finite sample space S (with neither equal to $\emptyset$ or S), and illustrate it with a Pebble World diagram.

(b) Consider the experiment of picking a random point in the rectangle

$$R = \{(x, y) : 0 < x < 1, 0 < y < 1\},$$

where the probability of the point being in any particular region contained within R is the area of that region. Let A_1 and B_1 be rectangles contained within R, with areas not equal to 0 or 1. Let A be the event that the random point is in A_1, and B be the event that the random point is in B_1. Give a geometric description of when it is true that A and B are independent. Also, give an example where they are independent and another example where they are not independent.

(c) Show that if A and B are independent, then

$$P(A \cup B) = P(A) + P(B) - P(A)P(B) = 1 - P(A^c)P(B^c).$$

Solution:

(a) Consider a sample space S consisting of 4 pebbles, each with probability 1/4. Let A consist of two of the pebbles and B consist of two of the pebbles, with $A \cap B$ consisting of a single pebble, as illustrated below.
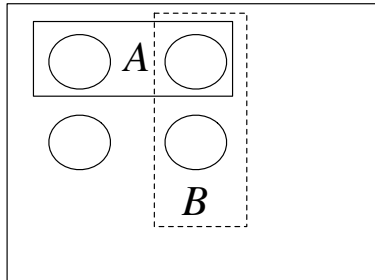
Then A and B are independent since $P(A \cap B) = 1/4 = P(A)P(B)$.

(b) Geometrically, independence of A and B says that the area of the intersection of A_1 and B_1 is the product of the areas of A_1 and B_1. An example where independence does not hold is when A_1 and B_1 are disjoint (non-overlapping). An example where independence holds is when A_1 is the left half of R and B_1 is the lower half. A more general example where independence holds is when $A_1 = \{(x, y) \in R : x \leq a\}$ and $B_1 = \{(x, y) \in R : y \leq b\}$, where a and b are constants in $(0, 1)$.

(c) Let A and B be independent. Then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = P(A) + P(B) - P(A)P(B).$$

Factoring out $P(A)$ from the terms containing it and later doing likewise with $P(B^c)$, we can also write $P(A \cup B)$ as

$$P(A)(1 - P(B)) + P(B) = (1 - P(A^c))P(B^c) + 1 - P(B^c) = 1 - P(A^c)P(B^c).$$

48. Ⓢ Arby has a belief system assigning a number $P_{\text{Arby}}(A)$ between 0 and 1 to every event A (for some sample space). This represents Arby's degree of belief about how likely A is to occur. For any event A, Arby is willing to pay a price of $1000 \cdot P_{\text{Arby}}(A)$ dollars to buy a certificate such as the one shown below:

> **Certificate**
>
> The owner of this certificate can redeem it for $1000 if A occurs. No value if A does not occur, except as required by federal, state, or local law. No expiration date.

Likewise, Arby is willing to sell such a certificate at the same price. Indeed, Arby is willing to buy or sell any number of certificates at this price, as Arby considers it the "fair" price.

Arby stubbornly refuses to accept the axioms of probability. In particular, suppose that there are two *disjoint* events A and B with

$$P_{\text{Arby}}(A \cup B) \neq P_{\text{Arby}}(A) + P_{\text{Arby}}(B).$$

Show how to make Arby go bankrupt, by giving a list of transactions Arby is willing to make that will *guarantee* that Arby will lose money (you can assume it will be known whether A occurred and whether B occurred the day after any certificates are bought/sold).

Solution: Suppose first that

$$P_{\text{Arby}}(A \cup B) < P_{\text{Arby}}(A) + P_{\text{Arby}}(B).$$

Call a certificate like the one show above, with any event C in place of A, a C-certificate. Measuring money in units of thousands of dollars, Arby is willing to pay $P_{\text{Arby}}(A) + P_{\text{Arby}}(B)$ to buy an A-certificate and a B-certificate, and is willing to sell an $(A \cup B)$-certificate for $P_{\text{Arby}}(A \cup B)$. In those transactions, Arby loses $P_{\text{Arby}}(A) + P_{\text{Arby}}(B) - P_{\text{Arby}}(A \cup B)$ and will not recoup any of that loss because if A or B occurs, Arby will have to pay out an amount equal to the amount Arby receives (since it's impossible for both A and B to occur).

Now suppose instead that

$$P_{\text{Arby}}(A \cup B) > P_{\text{Arby}}(A) + P_{\text{Arby}}(B).$$

Measuring money in units of thousands of dollars, Arby is willing to sell an A-certificate for $P_{\text{Arby}}(A)$, sell a B-certificate for $P_{\text{Arby}}(B)$, and buy a $(A \cup B)$-certificate for $P_{\text{Arby}}(A \cup B)$. In so doing, Arby loses $P_{\text{Arby}}(A \cup B) - (P_{\text{Arby}}(A) + P_{\text{Arby}}(B))$, and Arby won't recoup any of this loss, similarly to the above. (In fact, in this case, even if A and B are not disjoint, Arby will not recoup any of the loss, and will lose more money if both A and B occur.)

By buying/selling a sufficiently large number of certificates from/to Arby as described above, you can guarantee that you'll get all of Arby's money; this is called an *arbitrage opportunity*. This problem illustrates the fact that the axioms of probability are not arbitrary, but rather are *essential* for coherent thought (at least the first axiom, and the second with finite unions rather than countably infinite unions).

Arbitrary axioms allow arbitrage attacks; principled properties and perspectives on probability potentially prevent perdition.

Inclusion-exclusion

49. A fair die is rolled n times. What is the probability that at least 1 of the 6 values never appears?

Solution: Let A_j be the event that the value j never appears. Then

$$P(A_1 \cap A_2 \cap \cdots \cap A_k) = \left(\frac{6-k}{6}\right)^n$$

for $1 \le k \le 5$, since there is a $\frac{6-k}{6}$ chance that any particular roll is *not* in $\{1, 2, \ldots, k\}$. By inclusion-exclusion and symmetry

$$P(A_1 \cup \cdots \cup A_6) = 6\left(\frac{5}{6}\right)^n - \binom{6}{2}\left(\frac{4}{6}\right)^n + \binom{6}{3}\left(\frac{3}{6}\right)^n - \binom{6}{4}\left(\frac{2}{6}\right)^n + \binom{6}{5}\left(\frac{1}{6}\right)^n$$

$$= 6\left(\frac{5}{6}\right)^n - 15\left(\frac{2}{3}\right)^n + 20\left(\frac{1}{2}\right)^n - 15\left(\frac{1}{3}\right)^n + 6\left(\frac{1}{6}\right)^n.$$

Note that this reduces to 1 for $n \in \{1, 2, \ldots, 5\}$, as it must since there is no way to obtain all 6 possible values in fewer than 6 tosses.

50. Ⓢ A card player is dealt a 13-card hand from a well-shuffled, standard deck of cards. What is the probability that the hand is void in at least one suit ("void in a suit" means having no cards of that suit)?

Solution: Let S, H, D, C be the events of being void in Spades, Hearts, Diamonds, Clubs, respectively. We want to find $P(S \cup D \cup H \cup C)$. By inclusion-exclusion and symmetry,

$$P(S \cup D \cup H \cup C) = 4P(S) - 6P(S \cap H) + 4P(S \cap H \cap D) - P(S \cap H \cap D \cap C).$$

The probability of being void in a specific suit is $\frac{\binom{39}{13}}{\binom{52}{13}}$. The probability of being void in 2 specific suits is $\frac{\binom{26}{13}}{\binom{52}{13}}$. The probability of being void in 3 specific suits is $\frac{1}{\binom{52}{13}}$. And the last term is 0 since it's impossible to be void in everything. So the probability is

$$4\frac{\binom{39}{13}}{\binom{52}{13}} - 6\frac{\binom{26}{13}}{\binom{52}{13}} + \frac{4}{\binom{52}{13}} \approx 0.051.$$

51. Ⓢ For a group of 7 people, find the probability that all 4 seasons (winter, spring, summer, fall) occur at least once each among their birthdays, assuming that all seasons are equally likely.

Solution: Let A_i be the event that there are no birthdays in the ith season (with respect to some ordering of the seasons). The probability that all seasons occur at least once is $1 - P(A_1 \cup A_2 \cup A_3 \cup A_4)$. Note that $A_1 \cap A_2 \cap A_3 \cap A_4 = \emptyset$ (the most extreme case is when everyone is born in the same season). By inclusion-exclusion and symmetry,

$$P(A_1 \cup A_2 \cup A_3 \cup A_4) \quad = \quad 4P(A_1) - \binom{4}{2}P(A_1 \cap A_2) + \binom{4}{3}P(A_1 \cap A_2 \cap A_3).$$

We have $P(A_1) = (3/4)^7, P(A_1 \cap A_2) = (2/4)^7, P(A_1 \cap A_2 \cap A_3) = (1/4)^7$, so

$$1 - P(A_1 \cup A_2 \cup A_3 \cup A_4) = 1 - \left(4\left(\frac{3}{4}\right)^7 - \frac{6}{2^7} + \frac{4}{4^7}\right) \approx 0.513.$$

52. A certain class has 20 students, and meets on Mondays and Wednesdays in a classroom with exactly 20 seats. In a certain week, everyone in the class attends both days. On both days, the students choose their seats completely randomly (with one student per seat). Find the probability that no one sits in the same seat on both days of that week.

Solution: This problem is similar to the matching problem (Example 1.6.4). Label the students from 1 to 20, and then let A_i be the event that student i sits in the same seat on both days. Then

$$P(A_i) = \frac{1}{20}, P(A_i \cap A_j) = \frac{18!}{20!} = \frac{1}{19 \cdot 20}$$

for $i \neq j$, etc. By inclusion-exclusion, simplifying as in the solution to the matching problem,

$$P(A_1 \cup \cdots \cup A_{20}) = 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots - \frac{1}{20!}.$$

Thus, the probability that no one sits in the same seat on both days is

$$\frac{1}{0!} - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \cdots + \frac{1}{20!} \approx \frac{1}{e} \approx 0.37.$$

53. Fred needs to choose a password for a certain website. Assume that he will choose an 8-character password, and that the legal characters are the lowercase letters a, b, c, ..., z, the uppercase letters A, B, C, ..., Z, and the numbers 0, 1, ..., 9.

(a) How many possibilities are there if he is required to have at least one lowercase letter in his password?

(b) How many possibilities are there if he is required to have at least one lowercase letter and at least one uppercase letter in his password?

(c) How many possibilities are there if he is required to have at least one lowercase letter, at least one uppercase letter, and at least one number in his password?

Solution:

(a) There are 62^8 possible passwords if there are no restrictions, but we must exclude the 36^8 passwords that consist only of numbers and uppercase letters. So there are

$$62^8 - 36^8 \approx 2.155 \times 10^{14}$$

possibilities.

(b) Excluding the 36^8 passwords with no uppercase letters and the 36^8 passwords with

no lowercase letters, but being careful not to exclude the 10^8 numerical-only passwords twice, we have that there are

$$62^8 - 36^8 - 36^8 + 10^8 \approx 2.127 \times 10^{14}$$

possibilities.

(c) Again we use an inclusion-exclusion type argument. We exclude passwords with no uppercase letters, then with no lowercase letters, then with no numbers, but have to add back some terms to reflect the fact that some passwords are uppercase- only, lowercase-only, or numerical-only. This gives that there are

$$62^8 - 36^8 - 36^8 - 52^8 + 10^8 + 26^8 + 26^8 \approx 1.597 \times 10^{14}$$

possibilities.

54. ⓢ Alice attends a small college in which each class meets only once a week. She is deciding between 30 non-overlapping classes. There are 6 classes to choose from for each day of the week, Monday through Friday. Trusting in the benevolence of randomness, Alice decides to register for 7 randomly selected classes out of the 30, with all choices equally likely. What is the probability that she will have classes every day, Monday through Friday? (This problem can be done either directly using the naive definition of probability, or using inclusion-exclusion.)

Solution: We will solve this both by direct counting and using inclusion-exclusion.

Direct Counting Method: There are two general ways that Alice can have class every day: either she has 2 days with 2 classes and 3 days with 1 class, or she has 1 day with 3 classes, and has 1 class on each of the other 4 days. The number of possibilities for the former is $\binom{5}{2}\binom{6}{2}^2 6^3$ (choose the 2 days when she has 2 classes, and then select 2 classes on those days and 1 class for the other days). The number of possibilities for the latter is $\binom{5}{1}\binom{6}{3}6^4$. So the probability is

$$\frac{\binom{5}{2}\binom{6}{2}^2 6^3 + \binom{5}{1}\binom{6}{3}6^4}{\binom{30}{7}} = \frac{114}{377} \approx 0.302.$$

Inclusion-Exclusion Method: We will use inclusion-exclusion to find the probability of the complement, which is the event that she has at least one day with no classes. Let $B_i = A_i^c$. Then

$$P(B_1 \cup B_2 \cdots \cup B_5) = \sum_i P(B_i) - \sum_{i<j} P(B_i \cap B_j) + \sum_{i<j<k} P(B_i \cap B_j \cap B_k)$$

(terms with the intersection of 4 or more B_i's are not needed since Alice must have classes on at least 2 days). We have

$$P(B_1) = \frac{\binom{24}{7}}{\binom{30}{7}}, P(B_1 \cap B_2) = \frac{\binom{18}{7}}{\binom{30}{7}}, P(B_1 \cap B_2 \cap B_3) = \frac{\binom{12}{7}}{\binom{30}{7}}$$

and similarly for the other intersections. So

$$P(B_1 \cup \cdots \cup B_5) = 5\frac{\binom{24}{7}}{\binom{30}{7}} - \binom{5}{2}\frac{\binom{18}{7}}{\binom{30}{7}} + \binom{5}{3}\frac{\binom{12}{7}}{\binom{30}{7}} = \frac{263}{377}.$$

Therefore,

$$P(A_1 \cap A_2 \cap A_3 \cap A_4 \cap A_5) = \frac{114}{377} \approx 0.302.$$

55. A club consists of 10 seniors, 12 juniors, and 15 sophomores. An organizing committee of size 5 is chosen randomly (with all subsets of size 5 equally likely).

(a) Find the probability that there are exactly 3 sophomores in the committee.

(b) Find the probability that the committee has at least one representative from each of the senior, junior, and sophomore classes.

Solution:

(a) For a favorable outcome, we must choose 3 sophomores and 2 non-sophomores:

$$P(\text{exactly 3 sophomores}) = \frac{\binom{15}{3}\binom{22}{2}}{\binom{37}{5}} \approx 0.241.$$

(b) Let A_2, A_3, A_4 be the events that there are representatives from the sophomore, junior, and senior classes respectively. By inclusion-exclusion,

$$P(A_2^c \cup A_3^c \cup A_4^c) = P(A_2^c) + P(A_3^c) + P(A_4^c) - P(A_2^c \cap A_3^c) - P(A_2^c \cap A_4^c) - P(A_3^c \cap A_4^c)$$

$$= \frac{\binom{22}{5}}{\binom{37}{5}} + \frac{\binom{25}{5}}{\binom{37}{5}} + \frac{\binom{27}{5}}{\binom{37}{5}} - \frac{\binom{10}{5}}{\binom{37}{5}} - \frac{\binom{12}{5}}{\binom{37}{5}} - \frac{\binom{15}{5}}{\binom{37}{5}} = \frac{156147}{435897},$$

so $P(A_2 \cap A_3 \cap A_4) = 1 - \frac{156147}{435897} \approx 0.642$.

Mixed practice

56. For each part, decide whether the blank should be filled in with $=$, $<$, or $>$, and give a clear explanation. In (a) and (b), order doesn't matter.

(a) (number of ways to choose 5 people out of 10) ____ (number of ways to choose 6 people out of 10)

(b) (number of ways to break 10 people into 2 teams of 5) ____ (number of ways to break 10 people into a team of 6 and a team of 4)

(c) (probability that all 3 people in a group of 3 were born on January 1) ____ (probability that in a group of 3 people, 1 was born on each of January 1, 2, and 3)

Martin and Gale play an exciting game of "toss the coin," where they toss a fair coin until the pattern HH occurs (two consecutive Heads) or the pattern TH occurs (Tails followed immediately by Heads). Martin wins the game if and only if HH occurs before TH occurs.

(d) (probability that Martin wins) ____ 1/2

Solution:

(a) (number of ways to choose 5 people out of 10) $>$ (number of ways to choose 6 people out of 10)

Explanation: Using the fact that $n! = n \cdot (n-1)!$, we see that $\binom{10}{5} = \frac{10!}{5!5!} > \binom{10}{6} = \frac{10!}{4!6!}$ reduces to $6 > 5$. In general, $\binom{n}{k}$ is maximized at $k = n/2$ when n is even.

(b) (number of ways to break 10 people into 2 teams of 5) $<$ (number of ways to break 10 people into a team of 6 and a team of 4)

Explanation: The righthand side is $\binom{10}{6}$ since the choice of the team of 6 determines the team of 4. But the lefthand side is $\frac{1}{2}\binom{10}{5}$ since choosing a team of 5 is equivalent to choosing the complementary 5 people. The inequality then reduces to $3 < 5$.

(c) (probability that all 3 people in a group of 3 were born on January 1) < (probability that in a group of 3 people, 1 was born on each of January 1, 2, and 3)

Explanation: The righthand side is 6 times as large as the lefthand side, since there are 3! ways the righthand event can occur, but only 1 way that the people could all be born on January 1.

Martin and Gale play an exciting game of "toss the coin," where they toss a fair coin until the pattern HH occurs (two consecutive Heads) or the pattern TH occurs (Tails followed immediately by Heads). Martin wins the game if and only if HH occurs before TH occurs.

(d) (probability that Martin wins) < 1/2

Explanation: Consider the first toss. If it's Tails, we're *guaranteed* to see TH before we see HH. If it's Heads, we could still see either TH or HH first. Hence, the probability of HH occurring sooner than TH is less than 1/2.

Alternatively, consider the first two tosses. If it's HH, then Martin has already won. If it's TH, then Gale has already won. If it's HT or TT, then Gale will eventually win (she will win the next time that an H appears). So

$$P(\text{Martin wins}) = \frac{1}{4} < \frac{1}{2}.$$

57. Take a deep breath before attempting this problem. In the book *Innumeracy*, John Allen Paulos writes:

> Now for better news of a kind of immortal persistence. First, take a deep breath. Assume Shakespeare's account is accurate and Julius Caesar gasped ["Et tu, Brute!"] before breathing his last. What are the chances you just inhaled a molecule which Caesar exhaled in his dying breath?

Assume that one breath of air contains 10^{22} molecules, and that there are 10^{44} molecules in the atmosphere. (These are slightly simpler numbers than the estimates that Paulos gives; for the purposes of this problem, assume that these are exact. Of course, in reality there are many complications such as different types of molecules in the atmosphere, chemical reactions, variation in lung capacities, etc.)

Suppose that the molecules in the atmosphere now are the same as those in the atmosphere when Caesar was alive, and that in the 2000 years or so since Caesar, these molecules have been scattered completely randomly through the atmosphere. Also assume that Caesar's last breath was sampled *without* replacement but that your breathing is sampled *with* replacement (without replacement makes more sense but with replacement is easier to work with, and is a good approximation since the number of molecules in the atmosphere is so much larger than the number of molecules in one breath).

Find the probability that at least one molecule in the breath you just took was shared with Caesar's last breath, and give a simple approximation in terms of e.

Hint: As discussed in the math appendix, $(1 + \frac{x}{n})^n \approx e^x$ for n large.

Solution: Let $N = 10^{44}$ and $n = 10^{22}$. Each molecule breathed in has probability $(N - n)/N$ of not being from Caesar's last breath. The molecules breathed in are independent if we assume sampling with replacement. So the probability of at least one molecule being shared with Caesar's last breath is

$$1 - \frac{(N-n)^n}{N^n} = 1 - \left(1 - \frac{n}{N}\right)^n = 1 - \left(1 - \frac{1}{10^{22}}\right)^{10^{22}} \approx 1 - \frac{1}{e}.$$

Amazingly, this is about a 63% chance!

58. A widget inspector inspects 12 widgets and finds that exactly 3 are defective. Unfortunately, the widgets then get all mixed up and the inspector has to find the 3 defective widgets again by testing widgets one by one.

(a) Find the probability that the inspector will now have to test at least 9 widgets.

(b) Find the probability that the inspector will now have to test at least 10 widgets.

Solution:

(a) Imagine that the widgets are lined up in a row, ready to be tested (in that order). Let's find the probability of the complement. The event that the inspector has to test at most 8 widgets is the same as the event that all 3 defective widgets are among the first 8 widgets in line. This has probability $\binom{8}{3}/\binom{12}{3}$ by the naive definition: the denominator is the number of choices for where to place the defective widgets, and the numerator is the number of such choices such that the defective widgets are among the first 8 widgets. So the desired probability is

$$1 - \frac{\binom{8}{3}}{\binom{12}{3}} = \frac{41}{55} \approx 0.745.$$

Alternatively, note that the inspector has to test at most 8 widgets if and only if none of the last 4 widgets are defective. Choosing which particular widgets are among the last 4 widgets, the desired probability is

$$1 - \frac{\binom{9}{4}}{\binom{12}{4}} \approx 0.745,$$

which agrees with the result of the earlier approach.

(b) There are *two* disjoint ways that the inspector could be done after at most 9 widgets: either *all* 3 defective widgets have turned up among the first 9 widgets, or *none* of them have turned up after inspecting the first 9. So the desired probability is

$$1 - \frac{\binom{9}{3}}{\binom{12}{3}} - \frac{1}{\binom{12}{3}} = \frac{27}{44} \approx 0.614.$$

59. There are 15 chocolate bars and 10 children. In how many ways can the chocolate bars be distributed to the children, in each of the following scenarios?

(a) The chocolate bars are fungible (interchangeable).

(b) The chocolate bars are fungible, and each child must receive at least one.

Hint: First give each child a chocolate bar, and then decide what to do with the rest.

(c) The chocolate bars are not fungible (it matters which particular bar goes where).

(d) The chocolate bars are not fungible, and each child must receive at least one.

Hint: The strategy suggested in (b) does not apply. Instead, consider *randomly* giving the chocolate bars to the children, and apply inclusion-exclusion.

Solution:

(a) If we only care how many chocolate bars each child receives, not which specific bars go where, then we are in the realm of Bose-Einstein (with chocolate bars as indistinguishable particles and children as distinguishable boxes). So there are

$$\binom{10 + 15 - 1}{15} = \binom{24}{15} = 1307504$$

possibilities.

(b) As in the hint, first give each child one chocolate bar (there is only one way to do this, if the bars are treated as if they were indistinguishable). This leaves 5 bars. So by Bose-Einstein, there are

$$\binom{10 + 5 - 1}{5} = \binom{14}{5} = 2002$$

possibilities.

(c) By the multiplication rule, there are 10^{15} possibilities.

(d) Consider randomly distributing the bars to the children, with all of the 10^{15} possibilities equally likely. Let A_i be the event that child i does not receive any chocolate bars, and note that

$$P(A_1 \cap A_2 \cap \cdots \cap A_k) = \left(\frac{10 - k}{10}\right)^{15},$$

for $1 \leq k \leq 10$. By inclusion-exclusion and symmetry, the probability that at least one child does not receive a bar is

$$10\left(\frac{9}{10}\right)^{15} - \binom{10}{2}\left(\frac{8}{10}\right)^{15} + \binom{10}{3}\left(\frac{7}{10}\right)^{15} - \cdots + \binom{10}{9}\left(\frac{1}{10}\right)^{15}.$$

By the naive definition of probability, this is also $(10^{15} - a)/10^{15}$, where a is the number of possibilities such that each child receives at least one bar. So

$$a = 10^{15} - \left(10 \cdot 9^{15} - \binom{10}{2}8^{15} + \cdots + \binom{10}{9}\right) \approx 4.595 \times 10^{13}.$$

60. Given $n \geq 2$ numbers $(a_1, a_2, \ldots, a_n)$ with no repetitions, a *bootstrap sample* is a sequence $(x_1, x_2, \ldots, x_n)$ formed from the a_j's by sampling with replacement with equal probabilities. Bootstrap samples arise in a widely used statistical method known as the *bootstrap*. For example, if $n = 2$ and $(a_1, a_2) = (3, 1)$, then the possible bootstrap samples are $(3, 3), (3, 1), (1, 3),$ and $(1, 1)$.

(a) How many possible bootstrap samples are there for $(a_1, \ldots, a_n)$?

(b) How many possible bootstrap samples are there for $(a_1, \ldots, a_n)$, if order does not matter (in the sense that it only matters how many times each a_j was chosen, not the order in which they were chosen)?

(c) One random bootstrap sample is chosen (by sampling from $a_1, \ldots, a_n$ with replacement, as described above). Show that not all unordered bootstrap samples (in the sense of (b)) are equally likely. Find an unordered bootstrap sample $\mathbf{b}_1$ that is as likely as possible, and an unordered bootstrap sample $\mathbf{b}_2$ that is as unlikely as possible. Let p_1 be the probability of getting $\mathbf{b}_1$ and p_2 be the probability of getting $\mathbf{b}_2$ (so p_i is the probability of getting the *specific* unordered bootstrap sample $\mathbf{b}_i$). What is p_1/p_2? What is the ratio of the probability of getting an unordered bootstrap sample whose probability is p_1 to the probability of getting an unordered sample whose probability is p_2?

Solution:

(a) By the multiplication rule, there are n^n possibilities.

(b) By Bose-Einstein, there are $\binom{n+n-1}{n} = \binom{2n-1}{n}$ possibilities.

(c) We can take $\mathbf{b}_1 = [a_1, a_2, \ldots, a_n]$ and $\mathbf{b}_2 = [a_1, a_1, \ldots, a_1]$ (using square brackets to distinguish these *unordered* lists from the ordered bootstrap samples); these are the extreme cases since the former has $n!$ orders in which it could have been generated,

while the latter only has 1 such order. Then $p_1 = n!/n^n, p_2 = 1/n^n$, so $p_1/p_2 = n!$. There is only 1 unordered sample of the form of $\mathbf{b}_1$ but there are n of the form of $\mathbf{b}_2$, so the ratio of the probability of getting a sample whose probability is p_1 to the probability of getting a sample whose probability is p_2 is $(n!/n^n)/(n/n^n) = (n-1)!$.

61. Ⓢ There are 100 passengers lined up to board an airplane with 100 seats (with each seat assigned to one of the passengers). The first passenger in line crazily decides to sit in a randomly chosen seat (with all seats equally likely). Each subsequent passenger takes their assigned seat if available, and otherwise sits in a random available seat. What is the probability that the last passenger in line gets to sit in their assigned seat? (This is a common interview problem, and a beautiful example of the power of symmetry.)

Hint: Call the seat assigned to the jth passenger in line "seat j" (regardless of whether the airline calls it seat 23A or whatever). What are the possibilities for which seats are available to the last passenger in line, and what is the probability of each of these possibilities?

Solution: The seat for the last passenger is either seat 1 or seat 100; for example, seat 42 can't be available to the last passenger since the 42nd passenger in line would have sat there if possible. Seat 1 and seat 100 are equally likely to be available to the last passenger, since the previous 99 passengers view these two seats symmetrically. So the probability that the last passenger gets seat 100 is 1/2.

62. In the birthday problem, we assumed that all 365 days of the year are equally likely (and excluded February 29). In reality, some days are slightly more likely as birthdays than others. For example, scientists have long struggled to understand why more babies are born 9 months after a holiday. Let $\mathbf{p} = (p_1, p_2, \ldots, p_{365})$ be the vector of birthday probabilities, with p_j the probability of being born on the jth day of the year (February 29 is still excluded, with no offense intended to Leap Dayers).

The kth *elementary symmetric polynomial* in the variables $x_1, \ldots, x_n$ is defined by

$$e_k(x_1, \ldots, x_n) = \sum_{1 \le j_1 < j_2 < \cdots < j_k \le n} x_{j_1} \ldots x_{j_k}.$$

This just says to add up all of the $\binom{n}{k}$ terms we can get by choosing and multiplying k of the variables. For example, $e_1(x_1, x_2, x_3) = x_1 + x_2 + x_3$, $e_2(x_1, x_2, x_3) = x_1 x_2 + x_1 x_3 + x_2 x_3$, and $e_3(x_1, x_2, x_3) = x_1 x_2 x_3$.

Now let $k \ge 2$ be the number of people.

(a) Find a simple expression for the probability that there is at least one birthday match, in terms of $\mathbf{p}$ and an elementary symmetric polynomial.

(b) Explain intuitively why it makes sense that $P(\text{at least one birthday match})$ is minimized when $p_j = \frac{1}{365}$ for all j, by considering simple and extreme cases.

(c) The famous *arithmetic mean-geometric mean inequality* says that for $x, y \ge 0$,

$$\frac{x+y}{2} \ge \sqrt{xy}.$$

This inequality follows from adding $4xy$ to both sides of $x^2 - 2xy + y^2 = (x-y)^2 \ge 0$.

Define $\mathbf{r} = (r_1, \ldots, r_{365})$ by $r_1 = r_2 = (p_1 + p_2)/2$, $r_j = p_j$ for $3 \le j \le 365$. Using the arithmetic mean-geometric mean bound and the fact, which you should verify, that

$$e_k(x_1, \ldots, x_n) = x_1 x_2 e_{k-2}(x_3, \ldots, x_n) + (x_1 + x_2)e_{k-1}(x_3, \ldots, x_n) + e_k(x_3, \ldots, x_n),$$

show that

$$P(\text{at least one birthday match}|\mathbf{p}) \ge P(\text{at least one birthday match}|\mathbf{r}),$$

with strict inequality if $\mathbf{p} \ne \mathbf{r}$, where the "given $\mathbf{r}$" notation means that the birthday

probabilities are given by $\mathbf{r}$. Using this, show that the value of $\mathbf{p}$ that minimizes the probability of at least one birthday match is given by $p_j = \frac{1}{365}$ for all j.

Solution:

(a) One way to have no match is for the birthdays to be on days $1, 2, \ldots, k$, in any order. This has probability $k! p_1 p_2 \ldots p_k$. Similarly, we can have any other selection of k distinct days. Thus,

$$P(\text{at least one birthday match}) = 1 - k! e_k(\mathbf{p}).$$

(b) An extremely extreme case would be if $p_j = 1$ for some j, i.e., everyone is always born on the same day; then a match is guaranteed if there are at least 2 people. For another simple case, suppose that there are only 2 days in a year, with probabilities p and $q = 1 - p$. For 2 people, the probability of a match is $p^2 + q^2 = p^2 + (1-p)^2$, which is minimized at $p = 1/2$. In the general case, imagine starting with probabilities $1/365$ for all days and shifting some "probability mass" from some p_i to another p_j. This makes it less likely to have a match on day i and more likely for there to be a match on day j, but from the above it makes sense that the latter outweighs the former.

(c) The identity for $e_k(x_1, \ldots, x_n)$ is true since it is just partitioning the terms into 3 cases: terms with both x_1 and x_2, terms with one but not the other, and terms with neither x_1 nor x_2. Let $n = 365$. Note that $p_1 + p_2 = r_1 + r_2$ and by the arithmetic mean-geometric mean inequality, $p_1 p_2 \le ((p_1 + p_2)/2)^2 = r_1 r_2$. So

$$\begin{aligned}
e_k(p_1, \ldots, x_n) &= p_1 p_2 e_{k-2}(p_3, \ldots, p_n) + (p_1 + p_2) e_{k-1}(p_3, \ldots, p_n) + e_k(p_3, \ldots, p_n) \\
&\le r_1 r_2 e_{k-2}(r_3, \ldots, r_n) + (r_1 + r_2) e_{k-1}(r_3, \ldots, r_n) + e_k(r_3, \ldots, r_n) \\
&= e_k(r_1, \ldots, r_n).
\end{aligned}$$

So by (a), the probability of a birthday match when the probabilities are $\mathbf{p}$ is at least as large was when they are $\mathbf{r}$. The inequality is strict unless $p_1 = p_2$.

Now let $\mathbf{p}_0$ be a vector of birthday probabilities that minimizes the probability of at least one birthday match. If two components of $\mathbf{p}_0$ are not equal, the above shows that we could replace those two components by their average in order to obtain a smaller chance of a match, but this would contradict $\mathbf{p}_0$ minimizing the probability of a match. Thus, $\mathbf{p}_0$ has all components equal.

2

Conditional probability

Conditioning on evidence

1.   Ⓢ A spam filter is designed by looking at commonly occurring phrases in spam. Suppose that 80% of email is spam. In 10% of the spam emails, the phrase "free money" is used, whereas this phrase is only used in 1% of non-spam emails. A new email has just arrived, which does mention "free money". What is the probability that it is spam?

     Solution: Let S be the event that an email is spam and F be the event that an email has the "free money" phrase. By Bayes' rule,

     $$P(S|F) = \frac{P(F|S)P(S)}{P(F)} = \frac{0.1 \cdot 0.8}{0.1 \cdot 0.8 + 0.01 \cdot 0.2} = \frac{80/1000}{82/1000} = \frac{80}{82} \approx 0.9756.$$

2.   Ⓢ A woman is pregnant with twin boys. Twins may be either identical or fraternal. Suppose that $1/3$ of twins born are identical, that identical twins have a 50% chance of being both boys and a 50% chance of being both girls, and that for fraternal twins each twin independently has a 50% chance of being a boy and a 50% chance of being a girl. Given the above information, what is the probability that the woman's twins are identical?

     Solution: By Bayes' rule,

     $$P(\text{identical}|BB) = \frac{P(BB|\text{identical})P(\text{identical})}{P(BB)} = \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2} \cdot \frac{1}{3} + \frac{1}{4} \cdot \frac{2}{3}} = 1/2.$$

3.   According to the CDC (Centers for Disease Control and Prevention), men who smoke are 23 times more likely to develop lung cancer than men who don't smoke. Also according to the CDC, 21.6% of men in the U.S. smoke. What is the probability that a man in the U.S. is a smoker, given that he develops lung cancer?

     Solution: Let S be the event that a man in the U.S. smokes and L be the event that he gets lung cancer. We are given that $P(S) = 0.216$ and $P(L|S) = 23P(L|S^c)$. By Bayes' rule and the law of total probability, we have

     $$P(S|L) = \frac{P(L|S)P(S)}{P(L|S)P(S) + P(L|S^c)P(S^c)} = \frac{P(L|S)P(S)}{P(L|S)P(S) + \frac{1}{23}P(L|S)P(S^c)}.$$

     We don't know $P(L|S)$, but it cancels out! Thus,

     $$P(S|L) = \frac{0.216}{0.216 + (1 - 0.216)/23} \approx 0.864.$$

4.   Fred is answering a multiple-choice problem on an exam, and has to choose one of n options (exactly one of which is correct). Let K be the event that he knows the answer, and R be the event that he gets the problem right (either through knowledge or through luck). Suppose that if he knows the right answer he will definitely get the problem right, but if he does not know then he will guess completely randomly. Let $P(K) = p$.

(a) Find $P(K|R)$ (in terms of p and n).

(b) Show that $P(K|R) \geq p$, and explain why this makes sense intuitively. When (if ever) does $P(K|R)$ equal p?

Solution:

(a) By Bayes' rule and the law of total probability,

$$P(K|R) = \frac{P(R|K)P(K)}{P(R|K)P(K) + P(R|K^c)P(K^c)} = \frac{p}{p + (1-p)/n}.$$

(b) For the extreme case $p = 0$, we have $P(K|R) = 0 = p$. So assume $p > 0$. By the result of (a), $P(K|R) \geq p$ is equivalent to $p + (1-p)/n \leq 1$, which is a true statement since $p + (1-p)/n \leq p + 1 - p = 1$. This makes sense intuitively since getting the question right should increase our confidence that Fred knows the answer. Equality holds if and only if one of the extreme cases $n = 1$, $p = 0$, or $p = 1$ holds. If $n = 1$, it's not really a multiple-choice problem, and Fred getting the problem right is completely uninformative; if $p = 0$ or $p = 1$, then whether Fred knows the answer is a foregone conclusion, and no evidence will make us more (or less) sure that Fred knows the answer.

5. Three cards are dealt from a standard, well-shuffled deck. The first two cards are flipped over, revealing the Ace of Spades as the first card and the 8 of Clubs as the second card. Given this information, find the probability that the third card is an ace in two ways: using the definition of conditional probability, and by symmetry.

   Solution: Let A be the event that the first card is the Ace of Spades, B be the event that the second card is the 8 of Clubs, and C be the event that the third card is an ace. By definition of conditional probability,

   $$P(C|A, B) = \frac{P(C, A, B)}{P(A, B)} = \frac{P(A, B, C)}{P(A, B)}.$$

   By the naive definition of probability,

   $$P(A, B) = \frac{50!}{52!} = \frac{1}{51 \cdot 52}$$

   and

   $$P(A, B, C) = \frac{3 \cdot 49!}{52!} = \frac{3}{50 \cdot 51 \cdot 52}.$$

   So $P(C|A, B) = 3/50$.

   A simpler way is to see this is to use symmetry directly. Given the evidence, the third card is equally likely to be any card other than the Ace of Spades or 8 of Clubs, so it has probability $3/50$ of being an ace.

6. A hat contains 100 coins, where 99 are fair but one is double-headed (always landing Heads). A coin is chosen uniformly at random. The chosen coin is flipped 7 times, and it lands Heads all 7 times. Given this information, what is the probability that the chosen coin is double-headed? (Of course, another approach here would be to *look at both sides of the coin*—but this is a metaphorical coin.)

   Solution: Let A be the event that the chosen coin lands Heads all 7 times, and B be the event that the chosen coin is double-headed. Then

   $$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)} = \frac{0.01}{0.01 + (1/2)^7 \cdot 0.99} = \frac{128}{227} \approx 0.564.$$

7. A hat contains 100 coins, where *at least* 99 are fair, but there may be one that is double-headed (always landing Heads); if there is no such coin, then all 100 are fair. Let D be the event that there is such a coin, and suppose that $P(D) = 1/2$. A coin is chosen uniformly at random. The chosen coin is flipped 7 times, and it lands Heads all 7 times.

   (a) Given this information, what is the probability that one of the coins is double-headed?

   (b) Given this information, what is the probability that the chosen coin is double-headed?

   Solution:

   (a) Let A be the event that the chosen coin lands Heads all 7 times, and C be the event that the chosen coin is double-headed. By Bayes' rule and LOTP,

   $$P(D|A) = \frac{P(A|D)P(D)}{P(A|D)P(D) + P(A|D^c)P(D^c)}.$$

   We have $P(D) = P(D^c) = 1/2$ and $P(A|D^c) = 1/2^7$, so the only remaining ingredient that we need to find is $P(A|D)$. We can do this using LOTP with extra conditioning (it would be useful to know whether the *chosen* coin is double-headed, not just whether *somewhere* there is a double-headed coin, so we condition on whether or not C occurs):

   $$P(A|D) = P(A|D,C)P(C|D) + P(A|D,C^c)P(C^c|D) = \frac{1}{100} + \frac{1}{2^7} \cdot \frac{99}{100}.$$

   Plugging in these results, we have

   $$P(D|A) = \frac{227}{327} = 0.694.$$

   (b) By LOTP with extra conditioning (it would be useful to know whether there *is* a double-headed coin),

   $$P(C|A) = P(C|A,D)P(D|A) + P(C|A,D^c)P(D^c|A),$$

   with notation as in (a). But $P(C|A,D^c) = 0$, and we already found $P(D|A)$ in (a). Also, $P(C|A,D) = \frac{128}{227}$, as shown in Exercise 6 (conditioning on D and A puts us exactly in the setup of that exercise). Thus,

   $$P(C|A) = \frac{128}{227} \cdot \frac{227}{327} = \frac{128}{327} \approx 0.391.$$

8. The screens used for a certain type of cell phone are manufactured by 3 companies, A, B, and C. The proportions of screens supplied by A, B, and C are 0.5, 0.3, and 0.2, respectively, and their screens are defective with probabilities 0.01, 0.02, and 0.03, respectively. Given that the screen on such a phone is defective, what is the probability that Company A manufactured it?

   Solution: Let A, B, and C be the events that the screen was manufactured by Company A, B, and C, respectively, and let D be the event that the screen is defective. By Bayes' rule and LOTP,

   $$\begin{aligned} P(A|D) &= \frac{P(D|A)P(A)}{P(D|A)P(A) + P(D|B)P(B) + P(D|C)P(C)} \\ &= \frac{0.01 \cdot 0.5}{0.01 \cdot 0.5 + 0.02 \cdot 0.3 + 0.03 \cdot 0.2} \\ &\approx 0.294. \end{aligned}$$

9. (a) Show that if events A_1 and A_2 have the same *prior* probability $P(A_1) = P(A_2)$, A_1 implies B, and A_2 implies B, then A_1 and A_2 have the same *posterior* probability $P(A_1|B) = P(A_2|B)$ if it is observed that B occurred.

(b) Explain why (a) makes sense intuitively, and give a concrete example.

Solution:

(a) Suppose that $P(A_1) = P(A_2)$, A_1 implies B, and A_2 implies B. Then

$$P(A_1|B) = \frac{P(A_1, B)}{P(B)} = \frac{P(A_1)}{P(B)} = \frac{P(A_2)}{P(B)} = \frac{P(A_2, B)}{P(B)} = P(A_2|B).$$

(b) The result in (a) makes sense intuitively since, thinking in terms of Pebble World, observing that B occurred entails restricting the sample space by removing the pebbles in B^c. But none of the removed pebbles are in A_1 or in A_2, so the updated probabilities for A_1 and A_2 are just rescaled versions of the original probabilities, scaled by a constant chosen to make the total mass 1.

For a simple example, let A_1 be the event that the top card in a well-shuffled standard deck is a diamond, let A_2 be the event that it is a heart, and let B be the event that it is a red card. Then $P(A_1) = P(A_2) = 1/4$ and $P(A_1|B) = P(A_2|B) = 1/2$.

10. Fred is working on a major project. In planning the project, two milestones are set up, with dates by which they should be accomplished. This serves as a way to track Fred's progress. Let A_1 be the event that Fred completes the first milestone on time, A_2 be the event that he completes the second milestone on time, and A_3 be the event that he completes the project on time.

Suppose that $P(A_{j+1}|A_j) = 0.8$ but $P(A_{j+1}|A_j^c) = 0.3$ for $j = 1, 2$, since if Fred falls behind on his schedule it will be hard for him to get caught up. Also, assume that the second milestone supersedes the first, in the sense that once we know whether he is on time in completing the second milestone, it no longer matters what happened with the first milestone. We can express this by saying that A_1 and A_3 are conditionally independent given A_2 and they're also conditionally independent given A_2^c.

(a) Find the probability that Fred will finish the project on time, given that he completes the first milestone on time. Also find the probability that Fred will finish the project on time, given that he is late for the first milestone.

(b) Suppose that $P(A_1) = 0.75$. Find the probability that Fred will finish the project on time.

Solution:

(a) We need to find $P(A_3|A_1)$ and $P(A_3|A_1^c)$. To do so, let's use LOTP to condition on whether or not A_2 occurs:

$$P(A_3|A_1) = P(A_3|A_1, A_2)P(A_2|A_1) + P(A_3|A_1, A_2^c)P(A_2^c|A_1).$$

Using the conditional independence assumptions, this becomes

$$P(A_3|A_2)P(A_2|A_1) + P(A_3|A_2^c)P(A_2^c|A_1) = (0.8)(0.8) + (0.3)(0.2) = 0.7.$$

Similarly,

$$P(A_3|A_1^c) = P(A_3|A_2)P(A_2|A_1^c) + P(A_3|A_2^c)P(A_2^c|A_1^c) = (0.8)(0.3) + (0.3)(0.7) = 0.45.$$

(b) By LOTP and Part (a),

$$P(A_3) = P(A_3|A_1)P(A_1) + P(A_3|A_1^c)P(A_1^c) = (0.7)(0.75) + (0.45)(0.25) = 0.6375.$$

11. An *exit poll* in an election is a survey taken of voters just after they have voted. One major use of exit polls has been so that news organizations can try to figure out as soon as possible who won the election, before the votes are officially counted. This has been notoriously inaccurate in various elections, sometimes because of *selection bias*: the sample of people who are invited to and agree to participate in the survey may not be similar enough to the overall population of voters.

Consider an election with two candidates, Candidate A and Candidate B. Every voter is invited to participate in an exit poll, where they are asked whom they voted for; some accept and some refuse. For a randomly selected voter, let A be the event that they voted for A, and W be the event that they are willing to participate in the exit poll. Suppose that $P(W|A) = 0.7$ but $P(W|A^c) = 0.3$. In the exit poll, 60% of the respondents say they voted for A (assume that they are all honest), suggesting a comfortable victory for A. Find $P(A)$, the true proportion of people who voted for A.

Solution: We have $P(A|W) = 0.6$ since 60% of the respondents voted for A. Let $p = P(A)$. Then

$$0.6 = P(A|W) = \frac{P(W|A)P(A)}{P(W|A)P(A) + P(W|A^c)P(A^c)} = \frac{0.7p}{0.7p + 0.3(1-p)}.$$

Solving for p, we obtain

$$P(A) = \frac{9}{23} \approx 0.391.$$

So actually A received fewer than half of the votes!

12. Alice is trying to communicate with Bob, by sending a message (encoded in binary) across a channel.

(a) Suppose for this part that she sends only one bit (a 0 or 1), with equal probabilities. If she sends a 0, there is a 5% chance of an error occurring, resulting in Bob receiving a 1; if she sends a 1, there is a 10% chance of an error occurring, resulting in Bob receiving a 0. Given that Bob receives a 1, what is the probability that Alice actually sent a 1?

(b) To reduce the chance of miscommunication, Alice and Bob decide to use a *repetition code*. Again Alice wants to convey a 0 or a 1, but this time she repeats it two more times, so that she sends 000 to convey 0 and 111 to convey 1. Bob will decode the message by going with what the majority of the bits were. Assume that the error probabilities are as in (a), with error events for different bits independent of each other. Given that Bob receives 110, what is the probability that Alice intended to convey a 1?

Solution:

(a) Let A_1 be the event that Alice sent a 1, and B_1 be the event that Bob receives a 1. Then

$$P(A_1|B_1) = \frac{P(B_1|A_1)P(A_1)}{P(B_1|A_1)P(A_1) + P(B_1|A_1^c)P(A_1^c)} = \frac{(0.9)(0.5)}{(0.9)(0.5) + (0.05)(0.5)} \approx 0.9474.$$

(b) Now let A_1 be the event that Alice intended to convey a 1, and B_{110} be the event that Bob receives 110. Then

$$\begin{aligned}
P(A_1|B_{110}) &= \frac{P(B_{110}|A_1)P(A_1)}{P(B_{110}|A_1)P(A_1) + P(B_{110}|A_1^c)P(A_1^c)} \\
&= \frac{(0.9 \cdot 0.9 \cdot 0.1)(0.5)}{(0.9 \cdot 0.9 \cdot 0.1)(0.5) + (0.05 \cdot 0.05 \cdot 0.95)(0.5)} \\
&\approx 0.9715.
\end{aligned}$$

13. Company A has just developed a diagnostic test for a certain disease. The disease afflicts 1% of the population. As defined in Example 2.3.9, the *sensitivity* of the test is the probability of someone testing positive, given that they have the disease, and the *specificity* of the test is the probability that of someone testing negative, given that they don't have the disease. Assume that, as in Example 2.3.9, the sensitivity and specificity are both 0.95.

Company B, which is a rival of Company A, offers a competing test for the disease. Company B claims that their test is faster and less expensive to perform than Company A's test, is less painful (Company A's test requires an incision), and yet has a higher overall success rate, where overall success rate is defined as the probability that a random person gets diagnosed correctly.

(a) It turns out that Company B's test can be described and performed very simply: no matter who the patient is, diagnose that they do not have the disease. Check whether Company B's claim about overall success rates is true.

(b) Explain why Company A's test may still be useful.

(c) Company A wants to develop a new test such that the overall success rate is higher than that of Company B's test. If the sensitivity and specificity are equal, how high does the sensitivity have to be to achieve their goal? If (amazingly) they can get the sensitivity equal to 1, how high does the specificity have to be to achieve their goal? If (amazingly) they can get the specificity equal to 1, how high does the sensitivity have to be to achieve their goal?

Solution:

(a) For Company B's test, the probability that a random person in the population is diagnosed correctly is 0.99, since 99% of the people do not have the disease. For a random member of the population, let C be the event that Company A's test yields the correct result, T be the event of testing positive in Company A's test, and D be the event of having the disease. Then

$$\begin{aligned}
P(C) &= P(C|D)P(D) + P(C|D^c)P(D^c) \\
&= P(T|D)P(D) + P(T^c|D^c)P(D^c) \\
&= (0.95)(0.01) + (0.95)(0.99) \\
&= 0.95,
\end{aligned}$$

which makes sense intuitively since the sensitivity and specificity of Company A's test are both 0.95. So Company B is correct about having a higher overall success rate.

(b) Despite the result of (a), Company A's test may still provide very useful information, whereas Company B's test is uninformative. If Fred tests positive on Company A's test, Example 2.3.9 shows that his probability of having the disease increases from 0.01 to 0.16 (so it is still fairly unlikely that he has the disease, but it is much more likely than it was before the test result; further testing may well be advisable). In contrast, Fred's probability of having the disease does not change after undergoing Company's B test, since the test result is a foregone conclusion.

(c) Let s be the sensitivity and p be the specificity of A's new test. With notation as in the solution to (a), we have
$$P(C) = 0.01s + 0.99p.$$
If $s = p$, then $P(C) = s$, so Company A needs $s > 0.99$.
If $s = 1$, then $P(C) = 0.01 + 0.99p > 0.99$ if $p > 98/99 \approx 0.9899$.
If $p = 1$, then $P(C) = 0.01s + 0.99$ is automatically greater than 0.99 (unless $s = 0$, in which case both companies have tests with sensitivity 0 and specificity 1).

14. Consider the following scenario, from Tversky and Kahneman:

> Let A be the event that before the end of next year, Peter will have installed a burglar alarm system in his home. Let B denote the event that Peter's home will be burglarized before the end of next year.

(a) Intuitively, which do you think is bigger, $P(A|B)$ or $P(A|B^c)$? Explain your intuition.

(b) Intuitively, which do you think is bigger, $P(B|A)$ or $P(B|A^c)$? Explain your intuition.

(c) Show that for *any* events A and B (with probabilities not equal to 0 or 1), $P(A|B) > P(A|B^c)$ is equivalent to $P(B|A) > P(B|A^c)$.

(d) Tversky and Kahneman report that 131 out of 162 people whom they posed (a) and (b) to said that $P(A|B) > P(A|B^c)$ and $P(B|A) < P(B|A^c)$. What is a plausible explanation for why this was such a popular opinion despite (c) showing that it is impossible for these inequalities both to hold?

Solution:

(a) Intuitively, $P(A|B)$ seems larger than $P(A|B^c)$ since if Peter's home is burglarized, he is likely to take increased precautions (such as installing an alarm) against future attempted burglaries.

(b) Intuitively, $P(B|A^c)$ seems larger than $P(B|A)$, since presumably having an alarm system in place deters prospective burglars from attempting a burglary and hampers their chances of being able to burglarize the home. However, this is in conflict with (a), according to (c). Alternatively, we could argue that $P(B|A)$ should be larger than $P(B|A^c)$, since observing that an alarm system is in place could be evidence that the neighborhood has frequent burglaries.

(c) First note that $P(A|B) > P(A|B^c)$ is equivalent to $P(A|B) > P(A)$, since LOTP says that $P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$ is between $P(A|B)$ and $P(A|B^c)$ (in words, $P(A)$ is a weighted average of $P(A|B)$ and $P(A|B^c)$). But $P(A|B) > P(A)$ is equivalent to $P(A, B) > P(A)P(B)$, by definition of conditional probability. Likewise, $P(B|A) > P(B|A^c)$ is equivalent to $P(B|A) > P(B)$, which in turn is equivalent to $P(A, B) > P(A)P(B)$.

(d) It is reasonable to assume that a burglary at his home might cause Peter to install an alarm system and that having an alarm systems might reduce the chance of a future burglary. People with inconsistent beliefs about (a) and (b) may be thinking intuitively in causal terms, interpreting a probability $P(D|C)$ in terms of C causing D. But the definition of $P(D|C)$ does not invoke causality and does not require C's occurrence to precede D's occurrence or non-occurrence temporally.

15. Let A and B be events with $0 < P(A \cap B) < P(A) < P(B) < P(A \cup B) < 1$. You are hoping that *both* A and B occurred. Which of the following pieces of information would you be happiest to observe: that A occurred, that B occurred, or that $A \cup B$ occurred?

Solution: If C is one of the events $A, B, A \cup B$, then

$$P(A \cap B|C) = \frac{P(A \cap B \cap C)}{P(C)} = \frac{P(A \cap B)}{P(C)}.$$

So among the three options for C, $P(A \cap B|C)$ is maximized when C is the event A.

16. Show that $P(A|B) \leq P(A)$ implies $P(A|B^c) \geq P(A)$, and give an intuitive explanation of why this makes sense.

Solution: By LOTP,

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c).$$

So $P(A)$ is between $P(A|B)$ and $P(A|B^c)$; it is a weighted average of these two conditional probabilities. To see this in more detail, let $x = \min(P(A|B), P(A|B^c)), y = \max(P(A|B), P(A|B^c))$. Then

$$P(A) \geq xP(B) + xP(B^c) = x$$

and

$$P(A) \leq yP(B) + yP(B^c) = y,$$

so $x \leq P(A) \leq y$. Therefore, if $P(A|B) \leq P(A)$, then $P(A) \leq P(A|B^c)$.

It makes sense intuitively that B and B^c should work in opposite directions as evidence regarding A. If both B and B^c were evidence in favor of A, then $P(A)$ should have already reflected this.

17. In deterministic logic, the statement "A implies B" is equivalent to its *contrapositive*, "not B implies not A". In this problem we will consider analogous statements in probability, the logic of uncertainty. Let A and B be events with probabilities not equal to 0 or 1.

(a) Show that if $P(B|A) = 1$, then $P(A^c|B^c) = 1$.

Hint: Apply Bayes' rule and LOTP.

(b) Show however that the result in (a) does not hold in general if $=$ is replaced by $\approx$. In particular, find an example where $P(B|A)$ is very close to 1 but $P(A^c|B^c)$ is very close to 0.

Hint: What happens if A and B are independent?

Solution:

(a) Let $P(B|A) = 1$. Then $P(B^c|A) = 0$. So by Bayes' rule and LOTP,

$$P(A^c|B^c) = \frac{P(B^c|A^c)P(A^c)}{P(B^c|A^c)P(A^c) + P(B^c|A)P(A)} = \frac{P(B^c|A^c)P(A^c)}{P(B^c|A^c)P(A^c)} = 1.$$

(b) For a simple counterexample if $=$ is replaced by $\approx$ in (a), let A and B be independent events with $P(A)$ and $P(B)$ both extremely close to 1. For example, this can be done in the context of flipping a coin 1000 times, where A is an extremely likely (but not certain) event based on the first 500 tosses and B is an extremely likely (but not certain) event based on the last 500 tosses. Then $P(B|A) = P(B) \approx 1$, but $P(A^c|B^c) = P(A^c) \approx 0$.

18. Show that if $P(A) = 1$, then $P(A|B) = 1$ for any B with $P(B) > 0$. Intuitively, this says that if someone dogmatically believes something with absolute certainty, then no amount of evidence will change their mind. The principle of avoiding assigning probabilities of 0 or 1 to any event (except for mathematical certainties) was named *Cromwell's rule* by the statistician Dennis Lindley, due to Cromwell saying to the Church of Scotland, "think it possible you may be mistaken."

Hint: Write $P(B) = P(B \cap A) + P(B \cap A^c)$, and then show that $P(B \cap A^c) = 0$.

Solution: Let $P(A) = 1$. Then $P(B \cap A^c) \leq P(A^c) = 0$ since $B \cap A^c \subseteq A^c$, which shows that $P(B \cap A^c) = 0$. So

$$P(B) = P(B \cap A) + P(B \cap A^c) = P(A \cap B).$$

Thus,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A \cap B)}{P(A \cap B)} = 1.$$

19. Explain the following Sherlock Holmes saying in terms of conditional probability, carefully distinguishing between prior and posterior probabilities: "It is an old maxim of mine that when you have excluded the impossible, whatever remains, however improbable, must be the truth."

Solution: Let E be the observed evidence after a crime has taken place, and let $A_1, A_2, \ldots, A_n$ be an exhaustive list of events, any one of which (if it occurred) would serve as an explanation of how the crime occurred. Assuming that the list $A_1, \ldots, A_n$ exhausts all possible explanations for the crime, we have

$$P(A_1 \cup A_2 \cup \cdots \cup A_n | E) = 1.$$

Sherlock's maxim says that

$$P(A_n | E, A_1^c, A_1^c, \ldots, A_{n-1}^c) = 1,$$

i.e., if we have determined that all explanations other than A_n can be ruled out, then the remaining explanation, A_n, must be the truth, even if $P(A_n)$ and $P(A_n | E)$ are small. To prove Sherlock's maxim, note that

$$P(A_1^c, \ldots, A_{n-1}^c | E) = P(A_1^c, \ldots, A_{n-1}^c, A_n^c | E) + P(A_1^c, \ldots, A_{n-1}^c, A_n | E),$$

where the first term on the right-hand side is 0 by De Morgan's laws. So

$$P(A_n | E, A_1^c, A_1^c, \ldots, A_{n-1}^c) = \frac{P(A_1^c, A_1^c, \ldots, A_{n-1}^c, A_n | E)}{P(A_1^c, A_1^c, \ldots, A_{n-1}^c | E)} = 1.$$

20. The Jack of Spades (with cider), Jack of Hearts (with tarts), Queen of Spades (with a wink), and Queen of Hearts (without tarts) are taken from a deck of cards. These four cards are shuffled, and then two are dealt. Note: Literary references to cider, tarts, and winks do not need to be considered when solving this problem.

(a) Find the probability that both of these two cards are queens, given that the first card dealt is a queen.

(b) Find the probability that both are queens, given that at least one is a queen.

(c) Find the probability that both are queens, given that one is the Queen of Hearts.

Solution:

(a) Let Q_i be the event that the ith card dealt is a queen, for $i = 1, 2$. Then $P(Q_i) = 1/2$ since the ith card dealt is equally likely to be any of the cards. Also,

$$P(Q_1, Q_2) = P(Q_1)P(Q_2 | Q_1) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}.$$

As a check, note that by the naive definition of probability,

$$P(Q_1, Q_2) = \frac{1}{\binom{4}{2}} = \frac{1}{6}.$$

Thus,

$$P(Q_1 \cap Q_2 | Q_1) = \frac{P(Q_1 \cap Q_2)}{P(Q_1)} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}.$$

(b) Continuing as in (a),

$$P(Q_1 \cap Q_2 | Q_1 \cup Q_2) = \frac{P(Q_1 \cap Q_2)}{P(Q_1 \cup Q_2)} = \frac{P(Q_1 \cap Q_2)}{P(Q_1) + P(Q_2) - P(Q_1 \cap Q_2)} = \frac{\frac{1}{6}}{\frac{1}{2} + \frac{1}{2} - \frac{1}{6}} = \frac{1}{5}.$$

Another way to see this is to note that there are 6 possible 2-card hands, all equally likely, of which 1 (the "double-jack pebble") is eliminated by our conditioning; then by definition of conditional probability, we are left with 5 "pebbles" of equal mass.

(c) Let H_i be the event that the ith card dealt is a heart, for $i = 1, 2$. Then

$$P(Q_1 \cap Q_2 | (Q_1 \cap H_1) \cup (Q_2 \cap H_2)) = \frac{P(Q_1 \cap H_1 \cap Q_2) + P(Q_1 \cap Q_2 \cap H_2)}{P(Q_1 \cap H_1) + P(Q_2 \cap H_2)}$$
$$= \frac{\frac{1}{4} \cdot \frac{1}{3} + \frac{1}{4} \cdot \frac{1}{3}}{\frac{1}{4} + \frac{1}{4}}$$
$$= \frac{1}{3},$$

using the fact that $Q_1 \cap H_1$ and $Q_2 \cap H_2$ are disjoint. Alternatively, note that the conditioning reduces the sample space down to 3 possibilities, which are equally likely, and 1 of the 3 has both cards queens.

21. A fair coin is flipped 3 times. The toss results are recorded on separate slips of paper (writing "H" if Heads and "T" if Tails), and the 3 slips of paper are thrown into a hat.

(a) Find the probability that all 3 tosses landed Heads, given that at least 2 were Heads.

(b) Two of the slips of paper are randomly drawn from the hat, and both show the letter H. Given this information, what is the probability that all 3 tosses landed Heads?

Solution:

(a) Let A be the event that all 3 tosses landed Heads, and B be the event that at least 2 landed Heads. Then

$$P(A|B) = \frac{P(A, B)}{P(B)} = \frac{P(A)}{P(2 \text{ or } 3 \text{ Heads})} = \frac{1/8}{4/8} = \frac{1}{4}.$$

(b) Let C be the event that the two randomly chosen slips of paper show Heads. Then

$$P(A|C) = \frac{P(C|A)P(A)}{P(C)}$$
$$= \frac{P(C|A)P(A)}{P(C|A)P(A) + P(C|2 \text{ Heads})P(2 \text{ Heads}) + P(C|1 \text{ or } 0 \text{ Heads})P(1 \text{ or } 0 \text{ Heads})}$$
$$= \frac{\frac{1}{8}}{\frac{1}{8} + \frac{1}{3} \cdot \frac{3}{8} + 0 \cdot \frac{1}{2}}$$
$$= \frac{1}{2}.$$

Alternatively, let A_i be the event that the ith toss was Heads. Note that

$$P(A|A_i, A_j) = \frac{P(A)}{P(A_i, A_j)} = \frac{1/8}{1/4} = \frac{1}{2}$$

for any $i \neq j$. Since this probability is $1/2$ regardless of which 2 slips of paper were drawn, conditioning on which 2 slips were drawn gives

$$P(A|C) = \frac{1}{2}.$$

22. ⑤ A bag contains one marble which is either green or blue, with equal probabilities. A green marble is put in the bag (so there are 2 marbles now), and then a random marble is taken out. The marble taken out is green. What is the probability that the remaining marble is also green?

Solution: Let A be the event that the initial marble is green, B be the event that the removed marble is green, and C be the event that the remaining marble is green. We need to find $P(C|B)$. There are several ways to find this; one natural way is to condition on whether the initial marble is green:

$$P(C|B) = P(C|B, A)P(A|B) + P(C|B, A^c)P(A^c|B) = 1P(A|B) + 0P(A^c|B).$$

To find $P(A|B)$, use Bayes' rule:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{1/2}{P(B|A)P(A) + P(B|A^c)P(A^c)} = \frac{1/2}{1/2 + 1/4} = \frac{2}{3}.$$

So $P(C|B) = 2/3$.

Historical note: This problem was first posed by Lewis Carroll in 1893.

23. ⓢ Let G be the event that a certain individual is guilty of a certain robbery. In gathering evidence, it is learned that an event E_1 occurred, and a little later it is also learned that another event E_2 also occurred. Is it possible that individually, these pieces of evidence increase the chance of guilt (so $P(G|E_1) > P(G)$ and $P(G|E_2) > P(G)$), but together they decrease the chance of guilt (so $P(G|E_1, E_2) < P(G)$)?

Solution: Yes, this is possible. In fact, it is possible to have two events which separately provide evidence in favor of G, yet which together preclude G! For example, suppose that the crime was committed between 1 pm and 3 pm on a certain day. Let E_1 be the event that the suspect was at a specific nearby coffeeshop from 1 pm to 2 pm that day, and let E_2 be the event that the suspect was at the nearby coffeeshop from 2 pm to 3 pm that day. Then $P(G|E_1) > P(G), P(G|E_2) > P(G)$ (assuming that being in the vicinity helps show that the suspect had the opportunity to commit the crime), yet $P(G|E_1 \cap E_2) < P(G)$ (as being in the coffeehouse from 1 pm to 3 pm gives the suspect an alibi for the full time).

24. Is it possible to have events A_1, A_2, B, C with $P(A_1|B) > P(A_1|C)$ and $P(A_2|B) > P(A_2|C)$, yet $P(A_1 \cup A_2|B) < P(A_1 \cup A_2|C)$? If so, find an example (with a "story" interpreting the events, as well as giving specific numbers); otherwise, show that it is impossible for this phenomenon to happen.

Solution: Yes, this is possible. First note that $P(A_1 \cup A_2|B) = P(A_1|B) + P(A_2|B) - P(A_1 \cap A_2|B)$, so it is *not* possible if A_1 and A_2 are disjoint, and that it is crucial to consider the intersection. So let's choose examples where $P(A_1 \cap A_2|B)$ is much larger than $P(A_1 \cap A_2|C)$, to offset the other inequalities.

Story 1: Consider two basketball players, one of whom is randomly chosen to shoot two free throws. The first player is very streaky, and always either makes both or misses both free throws, with probability 0.8 of making both (this is an extreme example chosen for simplicity, but we could also make it so the player has good days (on which there is a high chance of making both shots) and bad days (on which there is a high chance of missing both shots) without requiring *always* making both or missing both). The second player's free throws go in with probability 0.7, independently. Define the events as A_j: the jth free throw goes in; B: the free throw shooter is the first player; $C = B^c$. Then

$$P(A_1|B) = P(A_2|B) = P(A_1 \cap A_2|B) = P(A_1 \cup A_2|B) = 0.8,$$

$$P(A_1|C) = P(A_2|C) = 0.7, P(A_1 \cap A_2|C) = 0.49, P(A_1 \cup A_2|C) = 2 \cdot 0.7 - 0.49 = 0.91.$$

Story 2: Suppose that you can either take Good Class or Other Class, but not both. If you take Good Class, you'll attend lecture 70% of the time, and you will understand the material if and only if you attend lecture. If you take Other Class, you'll attend lecture 40% of the time and understand the material 40% of the time, but because the class is so poorly taught, the only way you understand the material is by studying on your own

and not attending lecture. Defining the events as A_1: attend lecture; A_2: understand material; B: take Good Class; C: take Other Class,

$$P(A_1|B) = P(A_2|B) = P(A_1 \cap A_2|B) = P(A_1 \cup A_2|B) = 0.7,$$

$$P(A_1|C) = P(A_2|C) = 0.4, P(A_1 \cap A_2|C) = 0, P(A_1 \cup A_2|C) = 2 \cdot 0.4 = 0.8.$$

25. Ⓢ A crime is committed by one of two suspects, A and B. Initially, there is equal evidence against both of them. In further investigation at the crime scene, it is found that the guilty party had a blood type found in 10% of the population. Suspect A does match this blood type, whereas the blood type of Suspect B is unknown.

(a) Given this new information, what is the probability that A is the guilty party?

(b) Given this new information, what is the probability that B's blood type matches that found at the crime scene?

Solution:

(a) Let M be the event that A's blood type matches the guilty party's and for brevity, write A for "A is guilty" and B for "B is guilty". By Bayes' rule,

$$P(A|M) = \frac{P(M|A)P(A)}{P(M|A)P(A) + P(M|B)P(B)} = \frac{1/2}{1/2 + (1/10)(1/2)} = \frac{10}{11}.$$

(We have $P(M|B) = 1/10$ since, given that B is guilty, the probability that A's blood type matches the guilty party's is the same probability as for the general population.)

(b) Let C be the event that B's blood type matches, and condition on whether B is guilty. This gives

$$P(C|M) = P(C|M, A)P(A|M) + P(C|M, B)P(B|M) = \frac{1}{10} \cdot \frac{10}{11} + \frac{1}{11} = \frac{2}{11}.$$

26. Ⓢ To battle against spam, Bob installs two anti-spam programs. An email arrives, which is either legitimate (event L) or spam (event L^c), and which program j marks as legitimate (event M_j) or marks as spam (event M_j^c) for $j \in \{1, 2\}$. Assume that 10% of Bob's email is legitimate and that the two programs are each "90% accurate" in the sense that $P(M_j|L) = P(M_j^c|L^c) = 9/10$. Also assume that given whether an email is spam, the two programs' outputs are conditionally independent.

(a) Find the probability that the email is legitimate, given that the 1st program marks it as legitimate (simplify).

(b) Find the probability that the email is legitimate, given that both programs mark it as legitimate (simplify).

(c) Bob runs the 1st program and M_1 occurs. He updates his probabilities and then runs the 2nd program. Let $\tilde{P}(A) = P(A|M_1)$ be the updated probability function after running the 1st program. Explain briefly in words whether or not $\tilde{P}(L|M_2) = P(L|M_1 \cap M_2)$: is conditioning on $M_1 \cap M_2$ in one step equivalent to first conditioning on M_1, then updating probabilities, and then conditioning on M_2?

Solution:

(a) By Bayes' rule,

$$P(L|M_1) = \frac{P(M_1|L)P(L)}{P(M_1)} = \frac{\frac{9}{10} \cdot \frac{1}{10}}{\frac{9}{10} \cdot \frac{1}{10} + \frac{1}{10} \cdot \frac{9}{10}} = \frac{1}{2}.$$

(b) By Bayes' rule,

$$P(L|M_1, M_2) = \frac{P(M_1, M_2|L)P(L)}{P(M_1, M_2)} = \frac{(\frac{9}{10})^2 \cdot \frac{1}{10}}{(\frac{9}{10})^2 \cdot \frac{1}{10} + (\frac{1}{10})^2 \cdot \frac{9}{10}} = \frac{9}{10}.$$

(c) Yes, they are the same, since Bayes' rule is coherent. The probability of an event given various pieces of evidence does not depend on the order in which the pieces of evidence are incorporated into the updated probabilities.

27. Suppose that there are 5 blood types in the population, named type 1 through type 5, with probabilities $p_1, p_2, \ldots, p_5$. A crime was committed by two individuals. A suspect, who has blood type 1, has prior probability p of being guilty. At the crime scene blood evidence is collected, which shows that one of the criminals has type 1 and the other has type 2.

    Find the posterior probability that the suspect is guilty, given the evidence. Does the evidence make it more likely or less likely that the suspect is guilty, or does this depend on the values of the parameters $p, p_1, \ldots, p_5$? If it depends, give a simple criterion for when the evidence makes it more likely that the suspect is guilty.

    Solution: Let B be the event that the criminals have blood types 1 and 2 and G be the event that the suspect is guilty, so $P(G) = p$. Then

    $$P(G|B) = \frac{P(B|G)P(G)}{P(B|G)P(G) + P(B|G^c)P(G^c)} = \frac{p_2 p}{p_2 p + 2p_1 p_2(1-p)} = \frac{p}{p + 2p_1(1-p)},$$

    since given G, event B occurs if and only if the other criminal has blood type 2, while given G^c, the probability is $p_1 p_2$ that the elder criminal and the younger criminal have blood types 1 and 2 respectively, and also is $p_1 p_2$ for the other way around.

    Note that p_2 canceled out and p_3, p_4, p_5 are irrelevant. If $p_1 = 1/2$, then $P(G|B) = P(G)$. If $p_1 < 1/2$, then $P(G|B) > P(G)$, which means that the evidence increases the probability of guilt. But if $p_1 > 1/2$, then $P(G|B) < P(G)$, so the evidence decreases the probability of guilt, even though the evidence includes finding blood at the scene of the crime that matches the suspect's blood type!

28. Fred has just tested positive for a certain disease.

    (a) Given this information, find the posterior odds that he has the disease, in terms of the prior odds, the sensitivity of the test, and the specificity of the test.

    (b) Not surprisingly, Fred is much more interested in $P(\text{have disease}|\text{test positive})$, known as the *positive predictive value*, than in the sensitivity $P(\text{test positive}|\text{have disease})$. A handy rule of thumb in biostatistics and epidemiology is as follows:

    For a rare disease and a reasonably good test, specificity matters much more than sensitivity in determining the positive predictive value.

    Explain intuitively why this rule of thumb works. For this part you can make up some specific numbers and interpret probabilities in a frequentist way as proportions in a large population, e.g., assume the disease afflicts 1% of a population of 10000 people and then consider various possibilities for the sensitivity and specificity.

    Solution:

    (a) Let D be the event that Fred has the disease, and T be the event that he tests positive. Let sens $= P(T|D)$, spec $= P(T^c|D^c)$ be the sensitivity and specificity (respectively). By the odds form of Bayes' rule (or using Bayes' rule in the numerator and the denominator), the posterior odds of having the disease are

    $$\frac{P(D|T)}{P(D^c|T)} = \frac{P(D)}{P(D^c)} \frac{P(T|D)}{P(T|D^c)} = (\text{prior odds of } D) \left(\frac{\text{sens}}{1 - \text{spec}} \right).$$

(b) Let p be the prior probability of having the disease and $q = 1 - p$. Let PPV be the positive predictive value. By (a) or directly using Bayes' rule, we have

$$\text{PPV} = \frac{\text{sens}}{\text{sens} + \frac{q}{p}(1 - \text{spec})}.$$

For a concrete example to build intuition, let $p = 0.01$ and take sens = spec = 0.9 as a baseline. Then PPV ≈ 0.083. In the calculations below, we describe what happens if sensitivity is changed while specificity is held constant at 0.9 or vice versa. If we can improve the sensitivity to 0.95, the PPV improves slightly, to 0.088. But if we can improve the specificity to 0.95, the PPV improves to 0.15, a much bigger improvement. If we can improve the sensitivity to 0.99, the PPV improves to 0.091, but the other way around the PPV improves drastically more, to 0.48. Even in the extreme case that we can make the sensitivity 1, the PPV only improves to 0.092. But in the extreme case that we can make the specificity 1, the PPV becomes 1, the best value possible!

To further the intuitive picture, imagine a population of 10000 people, in which 1% (i.e., 100 people) have the disease. Again take sens = spec = 0.9 as a baseline. On average, there will be 90 true positives (correctly diagnosed diseased people), 10 false negatives (misdiagnosed diseased people), 8910 true negatives (correctly diagnosed healthy people), and 990 false positives (misdiagnosed healthy people). This is illustrated in the figure below (not to scale).
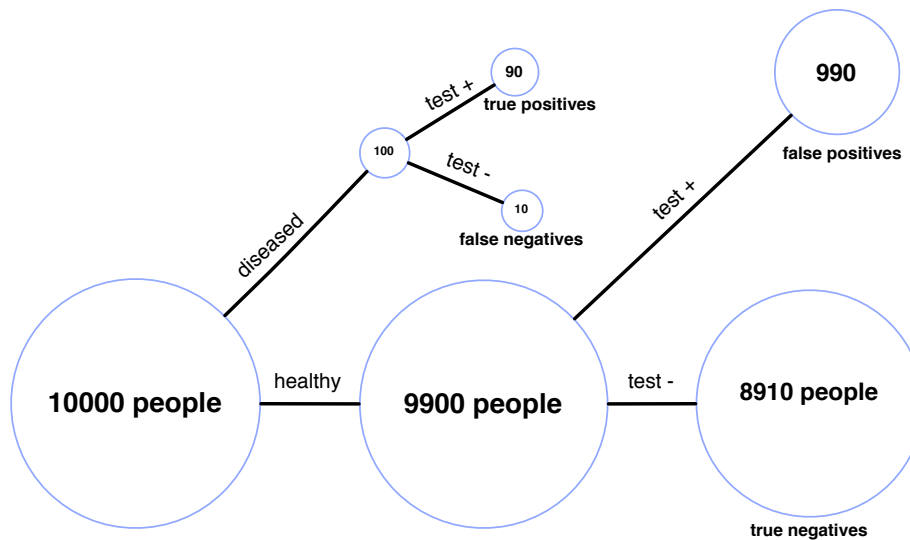
The PPV corresponds to the number of true positives over the number of positives, which is $90/(90+990) \approx 0.083$ in this example. Increasing specificity could dramatically decrease the number of false positives, replacing 990 by a much lower number; on the other hand, increasing sensitivity could at best increase the number of true positives from 90 to 100 here.

29. A family has two children. Let C be a characteristic that a child can have, and assume that each child has characteristic C with probability p, independently of each other and of gender. For example, C could be the characteristic "born in winter" as in Example 2.2.7. Under the assumptions of Example 2.2.5, show that the probability that both children are girls given that at least one is a girl with characteristic C is $\frac{2-p}{4-p}$. Note that this is $1/3$ if $p = 1$ (agreeing with the first part of Example 2.2.5) and approaches $1/2$ from below as $p \to 0$ (agreeing with Example 2.2.7).

Solution: Let G be the event that both children are girls, A be the event that at least

one child is a girl with characteristic C, and B be the event that at least one child has characteristic C. Note that $G \cap A = G \cap B$, and G is independent of B. Then

$$
\begin{aligned}
P(G|A) &= \frac{P(G, A)}{P(A)} \\
&= \frac{P(G, B)}{P(A)} \\
&= \frac{P(G)P(B)}{P(A)} \\
&= \frac{\frac{1}{4}(1 - (1 - p)^2)}{1 - (1 - \frac{p}{2})^2} \\
&= \frac{1 - (1 - 2p + p^2)}{4 - (4 - 4p + p^2)} \\
&= \frac{2 - p}{4 - p}.
\end{aligned}
$$

This is $1/3$ if $p = 1$ and approaches $1/2$ as $p \to 0$, but is less than $1/2$ for all $p > 0$ since $\frac{2-p}{4-p} < \frac{1}{2}$ is equivalent to $4 - 2p < 4 - p$, which in turn is equivalent to $p > 0$.

Independence and conditional independence

30. ⓢ A family has 3 children, creatively named A, B, and C.

    (a) Discuss intuitively (but clearly) whether the event "A is older than B" is independent of the event "A is older than C".

    (b) Find the probability that A is older than B, given that A is older than C.

    Solution:

    (a) They are not independent: knowing that A is older than B makes it more likely that A is older than C, as if A is older than B, then the only way that A can be younger than C is if the birth order is CAB, whereas the birth orders ABC and ACB are both compatible with A being older than B. To make this more intuitive, think of an extreme case where there are 100 children instead of 3, call them $A_1, \ldots, A_{100}$. Given that A_1 is older than all of $A_2, A_3, \ldots, A_{99}$, it's clear that A_1 is very old (relatively), whereas there isn't evidence about where A_{100} fits into the birth order.

    (b) Writing $x > y$ to mean that x is older than y,

    $$
    P(A > B | A > C) = \frac{P(A > B, A > C)}{P(A > C)} = \frac{1/3}{1/2} = \frac{2}{3}
    $$

    since $P(A > B, A > C) = P(A \text{ is the eldest child}) = 1/3$ (unconditionally, any of the 3 children is equally likely to be the eldest).

31. ⓢ Is it possible that an event is independent of itself? If so, when is this the case?

    Solution: Let A be an event. If A is independent of itself, then $P(A) = P(A \cap A) = P(A)^2$, so $P(A)$ is 0 or 1. So this is only possible in the extreme cases that the event has probability 0 or 1.

32. ⓢ Consider four nonstandard dice (the *Efron dice*), whose sides are labeled as follows (the 6 sides on each die are equally likely).

    A: $4, 4, 4, 4, 0, 0$
    B: $3, 3, 3, 3, 3, 3$

C: $6, 6, 2, 2, 2, 2$

D: $5, 5, 5, 1, 1, 1$

These four dice are each rolled once. Let A be the result for die A, B be the result for die B, etc.

(a) Find $P(A > B), P(B > C), P(C > D)$, and $P(D > A)$.

(b) Is the event $A > B$ independent of the event $B > C$? Is the event $B > C$ independent of the event $C > D$? Explain.

Solution:

(a) We have

$$P(A > B) = P(A = 4) = 2/3.$$
$$P(B > C) = P(C = 2) = 2/3.$$
$$P(C > D) = P(C = 6) + P(C = 2, D = 1) = 2/3.$$
$$P(D > A) = P(D = 5) + P(D = 1, A = 0) = 2/3.$$

(b) The event $A > B$ is independent of the event $B > C$ since $A > B$ is the same thing as $A = 4$, knowledge of which gives no information about $B > C$ (which is the same thing as $C = 2$). On the other hand, $B > C$ is *not* independent of $C > D$ since $P(C > D|C = 2) = 1/2 \neq 1 = P(C > D|C \neq 2)$.

33. Alice, Bob, and 100 other people live in a small town. Let C be the set consisting of the 100 other people, let A be the set of people in C who are friends with Alice, and let B be the set of people in C who are friends with Bob. Suppose that for each person in C, Alice is friends with that person with probability $1/2$, and likewise for Bob, with all of these friendship statuses independent.

    (a) Let $D \subseteq C$. Find $P(A = D)$.

    (b) Find $P(A \subseteq B)$.

    (c) Find $P(A \cup B = C)$.

    Solution:

    (a) More generally, let p be the probability of Alice being friends with any specific person in C (without assuming $p = 1/2$), and let $k = |D|$ (the size of D). Then

    $$P(A = D) = p^k (1 - p)^{100-k},$$

    by independence. For the case $p = 1/2$, this reduces to

    $$P(A = D) = 1/2^{100} \approx 7.89 \times 10^{-31}.$$

    That is, A is equally likely to be any subset of C.

    (b) The event $A \subseteq B$ says that for each person in C, it is not the case that they are friends with Alice but not with Bob. The event "friends with Alice but not Bob" has a probability of $1/4$ for each person in C, so by independence the overall probability is $(3/4)^{100} \approx 3.21 \times 10^{-13}$.

    (c) The event $A \cup B = C$ says that everyone in C is friends with Alice or Bob (inclusive of the possibility of being friends with both). For each person in C, there is a $3/4$ chance that they are friends with Alice or Bob, so by independence there is a $(3/4)^{100} \approx 3.21 \times 10^{-13}$ chance that everyone in C is friends with Alice or Bob.

34. Suppose that there are two types of drivers: good drivers and bad drivers. Let G be the event that a certain man is a good driver, A be the event that he gets into a car accident next year, and B be the event that he gets into a car accident the following year. Let $P(G) = g$ and $P(A|G) = P(B|G) = p_1, P(A|G^c) = P(B|G^c) = p_2$, with $p_1 < p_2$. Suppose that given the information of whether or not the man is a good driver, A and B are independent (for simplicity and to avoid being morbid, assume that the accidents being considered are minor and wouldn't make the man unable to drive).

    (a) Explain intuitively whether or not A and B are independent.

    (b) Find $P(G|A^c)$.

    (c) Find $P(B|A^c)$.

    Solution:

    (a) Intuitively, A and B are *not* independent, since learning that the man has a car accident next year makes it more likely that he is a bad driver, which in turn makes it more likely that he will have another accident the following year. We have that A and B are conditionally independent given G (and conditionally independent given G^c), but they are not independent since A gives information about whether the man is a good driver, and this information is very relevant for assessing how likely B is.

    (b) By Bayes's rule and LOTP,

    $$P(G|A^c) = \frac{P(A^c|G)P(G)}{P(A^c)} = \frac{(1-p_1)g}{(1-p_1)g + (1-p_2)(1-g)}.$$

    (c) Condition on whether or not the man is a good driver, using LOTP with extra conditioning:

    $$\begin{aligned} P(B|A^c) &= P(B|A^c, G)P(G|A^c) + P(B|A^c, G^c)P(G^c|A^c) \\ &= P(B|G)P(G|A^c) + P(B|G^c)P(G^c|A^c) \\ &= p_1 P(G|A^c) + p_2(1 - P(G|A^c)) \\ &= \frac{p_1(1-p_1)g + p_2(1-p_2)(1-g)}{(1-p_1)g + (1-p_2)(1-g)}. \end{aligned}$$

35. ⓢ You are going to play 2 games of chess with an opponent whom you have never played against before (for the sake of this problem). Your opponent is equally likely to be a beginner, intermediate, or a master. Depending on which, your chances of winning an individual game are 90%, 50%, or 30%, respectively.

    (a) What is your probability of winning the first game?

    (b) Congratulations: you won the first game! Given this information, what is the probability that you will also win the second game (assume that, given the skill level of your opponent, the outcomes of the games are independent)?

    (c) Explain the distinction between assuming that the outcomes of the games are independent and assuming that they are conditionally independent given the opponent's skill level. Which of these assumptions seems more reasonable, and why?

    Solution:

    (a) Let W_i be the event of winning the ith game. By the law of total probability,

    $$P(W_1) = (0.9 + 0.5 + 0.3)/3 = 17/30.$$

(b) We have $P(W_2|W_1) = P(W_2, W_1)/P(W_1)$. The denominator is known from (a), while the numerator can be found by conditioning on the skill level of the opponent:

$$P(W_1, W_2) = \frac{1}{3}P(W_1, W_2|\text{beginner}) + \frac{1}{3}P(W_1, W_2|\text{intermediate}) + \frac{1}{3}P(W_1, W_2|\text{expert}).$$

Since W_1 and W_2 are conditionally independent given the skill level of the opponent, this becomes

$$P(W_1, W_2) = (0.9^2 + 0.5^2 + 0.3^2)/3 = 23/60.$$

So

$$P(W_2|W_1) = \frac{23/60}{17/30} = 23/34.$$

(c) Independence here means that knowing one game's outcome gives no information about the other game's outcome, while conditional independence is the same statement where all probabilities are conditional on the opponent's skill level. Conditional independence given the opponent's skill level is a more reasonable assumption here. This is because winning the first game gives information about the opponent's skill level, which in turn gives information about the result of the second game.

That is, if the opponent's skill level is treated as fixed and known, then it may be reasonable to assume independence of games given this information; with the opponent's skill level random, earlier games can be used to help infer the opponent's skill level, which affects the probabilities for future games.

36. (a) Suppose that in the population of college applicants, being good at baseball is independent of having a good math score on a certain standardized test (with respect to some measure of "good"). A certain college has a simple admissions procedure: admit an applicant if and only if the applicant is good at baseball or has a good math score on the test.

Give an intuitive explanation of why it makes sense that among students that the college admits, having a good math score is *negatively associated* with being good at baseball, i.e., conditioning on having a good math score decreases the chance of being good at baseball.

(b) Show that if A and B are independent and $C = A \cup B$, then A and B are conditionally dependent given C (as long as $P(A \cap B) > 0$ and $P(A \cup B) < 1$), with

$$P(A|B, C) < P(A|C).$$

This phenomenon is known as *Berkson's paradox*, especially in the context of admissions to a school, hospital, etc.

Solution:

(a) Even though baseball skill and the math score are independent in the general population of applicants, it makes sense that they will become dependent (with a negative association) when restricting only to the students who are admitted. This is because within this sub-population, having a bad math score implies being good at baseball (else the student would not have been admitted). So having a good math score decreases the chance of being good in baseball (as shown in Exercise 16, if an event B is evidence in favor of an event A, then B^c is evidence against A).

As another explanation, note that 3 types of students are admitted: (i) good math score, good at baseball; (ii) good math score, bad at baseball; (iii) bad math score, good at baseball. Conditioning on having good math score removes students of type (iii) from consideration, which decreases the proportion of students who are good at baseball.

(b) Assume that A, B, C are as described. Then

$$P(A|B \cap C) = P(A|B) = P(A),$$

since A and B are independent and $B \cap C = B$. In contrast,

$$P(A|C) = \frac{P(A \cap C)}{P(C)} = \frac{P(A)}{P(C)} > P(A),$$

since $0 < P(C) < 1$. Therefore, $P(A|B,C) = P(A) < P(A|C)$.

37. Two different diseases cause a certain weird symptom; anyone who has either or both of these diseases will experience the symptom. Let D_1 be the event of having the first disease, D_2 be the event of having the second disease, and W be the event of having the weird symptom. Suppose that D_1 and D_2 are independent with $P(D_j) = p_j$, and that a person with neither of these diseases will have the weird symptom with probability w_0. Let $q_j = 1 - p_j$, and assume that $0 < p_j < 1$.

(a) Find $P(W)$.

(b) Find $P(D_1|W), P(D_2|W)$, and $P(D_1, D_2|W)$.

(c) Determine algebraically whether or not D_1 and D_2 are conditionally independent given W.

(d) Suppose for this part only that $w_0 = 0$. Give a clear, convincing intuitive explanation in words of whether D_1 and D_2 are conditionally independent given W.

Solution:

(a) Using LOTP to condition on whether the person has at least one of the diseases, we have

$$\begin{aligned}
P(W) &= P(W|D_1 \cup D_2)P(D_1 \cup D_2) + P(W|D_1^c \cap D_2^c)P(D_1^c \cap D_2^c) \\
&= 1 - q_1 q_2 + w_0 q_1 q_2 \\
&= 1 - (1 - w_0) q_1 q_2.
\end{aligned}$$

Alternatively, we can use LOTP to condition on the disease statuses:

$$\begin{aligned}
P(W) &= P(W|D_1, D_2)p_1 p_2 + P(W|D_1, D_2^c)p_1 q_2 + P(W|D_1^c, D_2)q_1 p_2 + P(W|D_1^c, D_2^c)q_1 q_2 \\
&= p_1 p_2 + p_1 q_2 + q_1 p_2 + w_0 q_1 q_2 \\
&= p_1 + p_2 - p_1 p_2 + w_0 q_1 q_2,
\end{aligned}$$

which (after a bit of algebra) agrees with the previous result.

(b) By Bayes' rule, we have

$$P(D_1|W) = \frac{P(W|D_1)P(D_1)}{P(W)} = \frac{p_1}{1 - (1 - w_0) q_1 q_2}.$$

Similarly,

$$P(D_2|W) = \frac{P(W|D_2)P(D_2)}{P(W)} = \frac{p_2}{1 - (1 - w_0) q_1 q_2}$$

and

$$P(D_1, D_2|W) = \frac{P(W|D_1, D_2)P(D_1, D_2)}{P(W)} = \frac{p_1 p_2}{1 - (1 - w_0) q_1 q_2}.$$

(c) We will check that D_1 and D_2 are *not* conditionally independent given W, i.e.,

$$P(D_1, D_2|W) \neq P(D_1|W)P(D_2|W),$$

Let

$$w = P(W).$$

We have

$$P(D_1, D_2|W) = \frac{p_1 p_2}{w},$$

whereas

$$P(D_1|W)P(D_2|W) = \frac{p_1}{w} \cdot \frac{p_2}{w} = \frac{p_1 p_2}{w^2},$$

so

$$P(D_1, D_2|W) < P(D_1|W)P(D_2|W),$$

since $0 < w < 1$.

(d) In the previous part we showed algebraically that D_1 and D_2 are not conditionally independent given W. For $w_0 = 0$, a simple intuitive explanation of this fact is as follows. The only possible explanations for the weird symptom are the two diseases. If someone has the symptom and doesn't have the first disease, then they *must* have the second disease.

38. We want to design a spam filter for email. As described in Exercise 1, a major strategy is to find phrases that are much more likely to appear in a spam email than in a non-spam email. In that exercise, we only consider one such phrase: "free money". More realistically, suppose that we have created a list of 100 words or phrases that are much more likely to be used in spam than in non-spam.

Let W_j be the event that an email contains the jth word or phrase on the list. Let

$$p = P(\text{spam}), p_j = P(W_j|\text{spam}), r_j = P(W_j|\text{not spam}),$$

where "spam" is shorthand for the event that the email is spam.

Assume that $W_1, \ldots, W_{100}$ are conditionally independent given that the email is spam, and conditionally independent given that it is not spam. A method for classifying emails (or other objects) based on this kind of assumption is called a *naive Bayes classifier*. (Here "naive" refers to the fact that the conditional independence is a strong assumption, not to Bayes being naive. The assumption may or may not be realistic, but naive Bayes classifiers sometimes work well in practice even if the assumption is not realistic.)

Under this assumption we know, for example, that

$$P(W_1, W_2, W_3^c, W_4^c, \ldots, W_{100}^c|\text{spam}) = p_1 p_2 (1-p_3)(1-p_4)\ldots(1-p_{100}).$$

Without the naive Bayes assumption, there would be vastly more statistical and computational difficulties since we would need to consider $2^{100} \approx 1.3 \times 10^{30}$ events of the form $A_1 \cap A_2 \cdots \cap A_{100}$ with each A_j equal to either W_j or W_j^c. A new email has just arrived, and it includes the 23rd, 64th, and 65th words or phrases on the list (but not the other 97). So we want to compute

$$P(\text{spam}|W_1^c, \ldots, W_{22}^c, W_{23}, W_{24}^c, \ldots, W_{63}^c, W_{64}, W_{65}, W_{66}^c, \ldots, W_{100}^c).$$

Note that we need to condition on *all* the evidence, not just the fact that $W_{23} \cap W_{64} \cap W_{65}$ occurred. Find the conditional probability that the new email is spam (in terms of p and the p_j and r_j).

Solution:

Let

$$E = W_1^c \cap \cdots \cap W_{22}^c \cap W_{23} \cap W_{24}^c \cap \cdots \cap W_{63}^c \cap W_{64} \cap W_{65} \cap W_{66}^c \cap \cdots \cap W_{100}^c$$

be the observed evidence. By Bayes' rule, LOTP, and conditional independence,

$$\begin{aligned}
P(\text{spam}|E) &= \frac{P(E|\text{spam})P(\text{spam})}{P(E|\text{spam})P(\text{spam}) + P(E|\text{non-spam})P(\text{non-spam})} \\
&= \frac{ap}{ap + b(1-p)},
\end{aligned}$$

where

$$a = (1-p_1)\ldots(1-p_{22})p_{23}(1-p_{24})\ldots(1-p_{63})p_{64}p_{65}(1-p_{66})\ldots(1-p_{100}),$$
$$b = (1-r_1)\ldots(1-r_{22})r_{23}(1-r_{24})\ldots(1-r_{63})r_{64}r_{65}(1-r_{66})\ldots(1-r_{100}).$$

Monty Hall

39. Ⓢ (a) Consider the following 7-door version of the Monty Hall problem. There are 7 doors, behind one of which there is a car (which you want), and behind the rest of which there are goats (which you don't want). Initially, all possibilities are equally likely for where the car is. You choose a door. Monty Hall then opens 3 goat doors, and offers you the option of switching to any of the remaining 3 doors.

    Assume that Monty Hall knows which door has the car, will always open 3 goat doors and offer the option of switching, and that Monty chooses with equal probabilities from all his choices of which goat doors to open. Should you switch? What is your probability of success if you switch to one of the remaining 3 doors?

    (b) Generalize the above to a Monty Hall problem where there are $n \geq 3$ doors, of which Monty opens m goat doors, with $1 \leq m \leq n - 2$.

    Solution:

    (a) Assume the doors are labeled such that you choose door 1 (to simplify notation), and suppose first that you follow the "stick to your original choice" strategy. Let S be the event of success in getting the car, and let C_j be the event that the car is behind door j. Conditioning on which door has the car, we have

    $$P(S) = P(S|C_1)P(C_1) + \cdots + P(S|C_7)P(C_7) = P(C_1) = \frac{1}{7}.$$

    Let M_{ijk} be the event that Monty opens doors i, j, k. Then

    $$P(S) = \sum_{i,j,k} P(S|M_{ijk})P(M_{ijk})$$

    (summed over all i, j, k with $2 \leq i < j < k \leq 7$.) By symmetry, this gives

    $$P(S|M_{ijk}) = P(S) = \frac{1}{7}$$

    for all i, j, k with $2 \leq i < j < k \leq 7$. Thus, the conditional probability that the car is behind 1 of the remaining 3 doors is 6/7, which gives 2/7 for each. So you should switch, thus making your probability of success 2/7 rather than 1/7.

    (b) By the same reasoning, the probability of success for "stick to your original choice" is $\frac{1}{n}$, both unconditionally and conditionally. Each of the $n - m - 1$ remaining doors has conditional probability $\frac{n-1}{(n-m-1)n}$ of having the car. This value is greater than $\frac{1}{n}$, so you should switch, thus obtaining probability $\frac{n-1}{(n-m-1)n}$ of success (both conditionally and unconditionally).

40. Ⓢ Consider the Monty Hall problem, except that Monty enjoys opening door 2 more than he enjoys opening door 3, and if he has a choice between opening these two doors, he opens door 2 with probability p, where $\frac{1}{2} \leq p \leq 1$.

    To recap: there are three doors, behind one of which there is a car (which you want), and behind the other two of which there are goats (which you don't want). Initially, all possibilities are equally likely for where the car is. You choose a door, which for concreteness we assume is door 1. Monty Hall then opens a door to reveal a goat, and offers you the option of switching. Assume that Monty Hall knows which door has the car, will always open a goat door and offer the option of switching, and as above assume that if Monty Hall has a choice between opening door 2 and door 3, he chooses door 2 with probability p (with $\frac{1}{2} \leq p \leq 1$).

    (a) Find the unconditional probability that the strategy of always switching succeeds (unconditional in the sense that we do not condition on which of doors 2 or 3 Monty opens).

(b) Find the probability that the strategy of always switching succeeds, given that Monty opens door 2.

(c) Find the probability that the strategy of always switching succeeds, given that Monty opens door 3.

Solution:

(a) Let C_j be the event that the car is hidden behind door j and let W be the event that we win using the switching strategy. Using the law of total probability, we can find the unconditional probability of winning:

$$\begin{aligned} P(W) &= P(W|C_1)P(C_1) + P(W|C_2)P(C_2) + P(W|C_3)P(C_3) \\ &= 0 \cdot 1/3 + 1 \cdot 1/3 + 1 \cdot 1/3 = 2/3. \end{aligned}$$

(b) A tree method works well here (delete the paths which are no longer relevant after the conditioning, and reweight the remaining values by dividing by their sum), or we can use Bayes' rule and the law of total probability (as below).

Let D_i be the event that Monty opens door i. Note that we are looking for $P(W|D_2)$, which is the same as $P(C_3|D_2)$ as we first choose door 1 and then switch to door 3. By Bayes' rule and the law of total probability,

$$\begin{aligned} P(C_3|D_2) &= \frac{P(D_2|C_3)P(C_3)}{P(D_2)} \\ &= \frac{P(D_2|C_3)P(C_3)}{P(D_2|C_1)P(C_1) + P(D_2|C_2)P(C_2) + P(D_2|C_3)P(C_3)} \\ &= \frac{1 \cdot 1/3}{p \cdot 1/3 + 0 \cdot 1/3 + 1 \cdot 1/3} \\ &= \frac{1}{1+p}. \end{aligned}$$

(c) The structure of the problem is the same as Part (b) (except for the condition that $p \geq 1/2$, which was not needed above). Imagine repainting doors 2 and 3, reversing which is called which. By Part (b) with $1 - p$ in place of p,

$$P(C_2|D_3) = \frac{1}{1+(1-p)} = \frac{1}{2-p}.$$

41. The ratings of Monty Hall's show have dropped slightly, and a panicking executive producer complains to Monty that the part of the show where he opens a door lacks suspense: Monty always opens a door with a goat. Monty replies that the reason is so that the game is never spoiled by him revealing the car, but he agrees to update the game as follows.

Before each show, Monty secretly flips a coin with probability p of Heads. If the coin lands Heads, Monty resolves to open a goat door (with equal probabilities if there is a choice). Otherwise, Monty resolves to open a random unopened door, with equal probabilities. The contestant knows p but does not know the outcome of the coin flip. When the show starts, the contestant chooses a door. Monty (who knows where the car is) then opens a door. If the car is revealed, the game is over; if a goat is revealed, the contestant is offered the option of switching. Now suppose it turns out that the contestant opens door 1 and then Monty opens door 2, revealing a goat. What is the contestant's probability of success if they switch to door 3?

Solution: For $j = 1, 2, 3$, let C_j be the event that the car is behind door j, $G_j = C_j^c$,

and M_j be the event that Monty opens door j. Let R be the event that Monty is in "random mode" (i.e., the coin lands Tails). By the law of total probability,

$$P(C_3|M_2, G_2) = P(R|M_2, G_2)P(C_3|M_2, G_2, R) + P(R^c|M_2, G_2)P(C_3|M_2, G_2, R^c),$$

where $P(C_3|M_2, G_2, R^c) = 2/3$ (since given R^c, Monty is operating as in the usual Monty Hall problem) and

$$P(C_3|M_2, G_2, R) = \frac{P(M_2, G_2|C_3, R)P(C_3|R)}{P(M_2, G_2|R)} = \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2} \cdot \frac{2}{3}} = \frac{1}{2}.$$

For the denominator above, note that M_2 and G_2 are conditionally independent given R; for the numerator, note also that $P(M_2, G_2|C_3, R) = P(M_2|C_3, R) = P(M_2|R)$. The posterior probability that Monty is in random mode is

$$P(R|M_2, G_2) = \frac{P(M_2, G_2|R)P(R)}{P(M_2, G_2|R)P(R) + P(M_2, G_2|R^c)P(R^c)} = \frac{\frac{1}{2} \cdot \frac{2}{3}(1-p)}{\frac{1}{2} \cdot \frac{2}{3}(1-p) + \frac{1}{2}p} = \frac{2(1-p)}{2+p},$$

again since M_2 and G_2 are conditionally independent given R, and since given R^c, M_2 implies G_2. Putting these results together, we have

$$P(C_3|M_2, G_2) = \frac{1}{2}\left(\frac{2(1-p)}{2+p}\right) + \frac{2}{3}\left(1 - \frac{2(1-p)}{2+p}\right) = \frac{1+p}{2+p}.$$

42. Consider the following variation of the Monty Hall problem, where in some situations Monty may *not* open a door and give the contestant the choice of whether to switch doors. Specifically, there are 3 doors, with 2 containing goats and 1 containing a car. The car is equally likely to be anywhere, and Monty knows where the car is. Let $0 \leq p \leq 1$.

The contestant chooses a door. If this initial choice has the car, Monty *will* open another door, revealing a goat (choosing with equal probabilities among his two choices of door), and then offer the contestant the choice of whether to switch to the other unopened door.

If the contestant's initial choice has a goat, then with probability p Monty *will* open another door, revealing a goat, and then offer the contestant the choice of whether to switch to the other unopened door; but with probability $1 - p$, Monty will *not* open a door, and the contestant must stick with their initial choice.

The contestant decides in advance to use the following strategy: initially choose door 1. Then, if Monty opens a door and offers the choice of whether to switch, do switch.

(a) Find the unconditional probability that the contestant will get the car. Also, check what your answer reduces to in the extreme cases $p = 0$ and $p = 1$, and briefly explain why your answer makes sense in these two cases.

(b) Monty now opens door 2, revealing a goat. So the contestant switches to door 3. Given this information, find the conditional probability that the contestant will get the car.

Solution:

(a) Let W be the event that the contestant gets the car and C_j be the event that the car is behind door j. By LOTP,

$$P(W) = P(W|C_1)P(C_1) + P(W|C_2)P(C_2) + P(W|C_3)P(C_3) = 0 \cdot \frac{1}{3} + p \cdot \frac{1}{3} + p \cdot \frac{1}{3} = \frac{2p}{3}.$$

For $p = 0$ this is 0, which makes sense: if C_1 occurs, the contestant will switch (and not get the car); if $C_2 \cup C_3$ occurs, Monty won't let them switch (so again they won't get the car).

For $p = 1$ this is $2/3$, which also makes sense: the problem then reduces to the original Monty Hall problem, for which switching gives probability $2/3$ of getting the car.

(b) Let M_j be the event that Monty opens door j. By Bayes' rule,

$$P(C_3|M_2) = \frac{P(M_2|C_3)P(C_3)}{P(M_2|C_3)P(C_3) + P(M_2|C_1)P(C_1)} = \frac{p \cdot \frac{1}{3}}{p \cdot \frac{1}{3} + \frac{1}{2} \cdot \frac{1}{3}} = \frac{2p}{2p+1}.$$

Again it makes sense that this reduces to 0 when $p = 0$, and $2/3$ when $p = 1$.

43. You are the contestant on the Monty Hall show. Monty is trying out a new version of his game, with rules as follows. You get to choose one of three doors. One door has a car behind it, another has a computer, and the other door has a goat (with all permutations equally likely). Monty, who knows which prize is behind each door, will open a door (but not the one you chose) and then let you choose whether to switch from your current choice to the other unopened door.

Assume that you prefer the car to the computer, the computer to the goat, and (by transitivity) the car to the goat.

(a) Suppose for this part only that Monty always opens the door that reveals your less preferred prize out of the two alternatives, e.g., if he is faced with the choice between revealing the goat or the computer, he will reveal the goat. Monty opens a door, revealing a goat (this is again for this part only). Given this information, should you switch? If you do switch, what is your probability of success in getting the car?

(b) Now suppose that Monty reveals your less preferred prize with probability p, and your more preferred prize with probability $q = 1 - p$. Monty opens a door, revealing a computer. Given this information, should you switch (your answer can depend on p)? If you do switch, what is your probability of success in getting the car (in terms of p)?

Solution:

(a) Let C be the event that the car is behind the door you originally chose, and let M_{goat} be the event that Monty reveals a goat when he opens a door. By Bayes' rule,

$$P(C|M_{\text{goat}}) = \frac{P(M_{\text{goat}}|C)P(C)}{P(M_{\text{goat}})} = \frac{(1)(1/3)}{2/3} = 1/2,$$

where the denominator comes from the fact that the two doors other than your initial choice are equally likely to have $\{\text{car}, \text{computer}\}$, $\{\text{computer}, \text{goat}\}$, or $\{\text{car}, \text{goat}\}$, and only in the first of these cases will Monty not reveal a goat.

So you should be indifferent between switching and not switching; either way, your conditional probability of getting the car is $1/2$. (Note though that the *unconditional* probability that switching would get you the car, before Monty revealed the goat, is $2/3$ since you will succeed by switching if and only if your initial door does not have the car.)

(b) Let C, R, G be the events that behind the door you originally chose is a car, computer, goat, respectively, and let M_{com} be the event that Monty reveals a computer when he opens a door. By Bayes' rule and LOTP,

$$P(C|M_{\text{com}}) = \frac{P(M_{\text{com}}|C)P(C)}{P(M_{\text{com}})} = \frac{P(M_{\text{com}}|C)P(C)}{P(M_{\text{com}}|C)P(C) + P(M_{\text{com}}|R)P(R) + P(M_{\text{com}}|G)P(G)}.$$

We have $P(M_{\text{com}}|C) = q, P(M_{\text{com}}|R) = 0, P(M_{\text{com}}|G) = p$, so

$$P(C|M_{\text{com}}) = \frac{q/3}{q/3 + 0 + p/3} = q.$$

Thus, your conditional probability of success if you follow the switching strategy is p. For $p < 1/2$, you should not switch, for $p = 1/2$, you should be indifferent about switching, and for $p > 1/2$, you should switch.

44. Monty Hall has introduced a new twist in his game, by generalizing the assumption that the initial probabilities for where the car is are $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Specifically, there are three doors, behind one of which there is a car (which the contestant wants), and behind the other two of which there are goats (which the contestant doesn't want). Initially, door i has probability p_i of having the car, where p_1, p_2, p_3 are known constants such that $0 < p_1 \leq p_2 \leq p_3 < 1$ and $p_1 + p_2 + p_3 = 1$. The contestant chooses a door. Then Monty opens a door (other than the one the contestant chose) and offers the contestant the option of switching to the other unopened door.

(a) Assume for this part that Monty knows in advance which door has the car. He always opens a door to reveal a goat, and if he has a choice of which door to open he chooses with equal probabilities. Suppose for this part that the contestant initially chooses door 3, and then Monty opens door 2, revealing a goat. Given the above information, find the conditional probability that door 3 has the car. Should the contestant switch doors? (If whether to switch depends on the p_i's, give a fully simplified criterion in terms of the p_i's.)

(b) Now assume instead that Monty does *not* know in advance where the car is. He randomly chooses which door to open (other than the one the contestant chose), with equal probabilities. (The game is spoiled if he reveals the car.) Suppose again that the contestant initially chooses door 3, and then Monty opens door 2, revealing a goat. Given the above information, find the conditional probability that door 3 has the car. Should the contestant switch doors? (If whether to switch depends on the p_i's, give a fully simplified criterion in terms of the p_i's.)

(c) Repeat (a), except with the contestant initially choosing door 1 rather than door 3.

(d) Repeat (b), except with the contestant initially choosing door 1 rather than door 3.

Solution:

(a) Let C_i be the event that door i has the car, $G_i = C_i^c$, and M_i be the event that Monty opens door i. By Bayes' rule,

$$P(C_3|M_2, G_2) = P(C_3|M_2) = \frac{P(M_2|C_3)P(C_3)}{P(M_2)} = \frac{\frac{1}{2}p_3}{P(M_2)}.$$

By LOTP, conditioning on where the car is, we have

$$P(M_2) = P(M_2|C_1)P(C_1) + P(M_2|C_2)P(C_2) + P(M_2|C_3)P(C_3) = p_1 + \frac{p_3}{2}.$$

Thus,

$$P(C_3|M_2, G_2) = \frac{p_3}{2p_1 + p_3}.$$

Also,

$$P(C_1|M_2, G_2) = 1 - P(C_3|M_2, G_2) = \frac{2p_1}{2p_1 + p_3}.$$

So the contestant should switch to door 1 if $2p_1 > p_3$, should not switch if $2p_1 < p_3$, and should be indifferent about switching if $2p_1 = p_3$.

(b) We will use notation as in (a). Note that now M_2 is independent of G_2, whereas in Part (a), M_2 implied G_2. We have

$$P(C_3|M_2, G_2) = \frac{P(M_2, G_2|C_3)P(C_3)}{P(M_2, G_2)} = \frac{P(M_2|C_3)P(C_3)}{P(M_2)P(G_2)} = \frac{\frac{1}{2}p_3}{\frac{1}{2}(1 - p_2)} = \frac{p_3}{p_1 + p_3}.$$

Then

$$P(C_1|M_2, G_2) = 1 - P(C_3|M_2, G_2) = \frac{p_1}{p_1 + p_3} \leq P(C_3|M_2, G_2),$$

since $p_1 \leq p_3$. So the contestant should not switch (unless $p_1 = p_3$, in which case the contestant should be indifferent about switching).

(c) By the same method as in (a),

$$P(C_1|M_2, G_2) = \frac{p_1}{2p_3 + p_1},$$

$$P(C_3|M_2, G_2) = \frac{2p_3}{2p_3 + p_1},$$

Since $p_1 \leq p_3$, we have $p_1 < 2p_3$, so the contestant should switch to door 3.

(d) By the same method as in (b),

$$P(C_1|M_2, G_2) = \frac{p_1}{p_3 + p_1},$$

$$P(C_3|M_2, G_2) = \frac{p_3}{p_3 + p_1}.$$

So the contestant should switch (unless $p_1 = p_3$, in which case the contestant should be indifferent about switching).

45. Monty Hall is trying out a new version of his game. In this version, instead of there always being 1 car and 2 goats, the prizes behind the doors are generated *independently*, with each door having probability p of having a car and $q = 1 - p$ of having a goat. In detail: There are three doors, behind each of which there is one prize: either a car or a goat. For each door, there is probability p that there is a car behind it and $q = 1 - p$ that there is a goat, independent of the other doors.

The contestant chooses a door. Monty, who knows the contents of each door, then opens one of the two remaining doors. In choosing which door to open, Monty will always reveal a goat if possible. If both of the remaining doors have the same kind of prize, Monty chooses randomly (with equal probabilities). After opening a door, Monty offers the contestant the option of switching to the other unopened door.

The contestant decides in advance to use the following strategy: first choose door 1. Then, after Monty opens a door, switch to the other unopened door.

(a) Find the unconditional probability that the contestant will get a car.

(b) Monty now opens door 2, revealing a goat. Given this information, find the conditional probability that the contestant will get a car.

Solution:

(a) If doors 2 and 3 both have goats, the contestant will get a goat; if both have cars, the contestant will get a car; and if one has a car and the other has a goat, Monty will reveal the goat and the contestant will get the car. So

$$P(\text{get car}) = P(\text{Door 2 has a car or Door 3 has a car}) = 1 - (1 - p)^2 = 2p - p^2.$$

(b) Let C_i, G_i, M_i be the events that door i has a car, has a goat, and is opened by Monty, respectively. By Bayes' rule,

$$P(C_3|G_2, M_2) = \frac{P(G_2, M_2|C_3)P(C_3)}{P(G_2, M_2)}.$$

The numerator is

$$P(G_2|C_3)P(M_2|G_2, C_3)P(C_3) = q \cdot 1 \cdot p = pq,$$

and by LOTP, conditioning on what's behind Door 3, the denominator is

$$P(G_2|C_3)P(M_2|G_2, C_3)P(C_3) + P(G_2|G_3)P(M_2|G_2, G_3)P(G_3) = q \cdot 1 \cdot p + q \cdot \frac{1}{2} \cdot q = pq + \frac{1}{2}q^2.$$

So

$$P(C_3|G_2, M_2) = \frac{pq}{pq + \frac{1}{2}q^2} = \frac{2p}{2p + q} = \frac{2p}{p+1}.$$

Alternatively, let W be the event that the contestant wins and R be the event that Monty reveals a goat. By symmetry and the definition of conditional probability,

$$P(W|M_2, R) = P(W|R) = \frac{P(W, R)}{P(R)} = \frac{P(\text{exactly one goat in doors 2, 3})}{P(\text{at least one goat in doors 2, 3})} = \frac{2p(1-p)}{1 - p^2} = \frac{2p}{p+1}.$$

46. Monty Hall is trying out a new version of his game, with rules as follows. The contestant gets to choose one of *four* doors. One door has a car behind it, another has an apple, another has a book, and another has a goat. All 24 permutations for which door has which prize are equally likely. In order from least preferred to most preferred, the contestant's preferences are: goat, apple, book, car.

Monty, who knows which prize is behind each door, will open a door (other than the contestant's initial choice) and then let the contestant choose whether to switch to another unopened door. Monty will reveal the least preferred prize (among the 3 doors other than the contestant's initial choice) with probability p, the intermediately preferred prize with probability $1 - p$, and the most preferred prize never.

The contestant decides in advance to use the following strategy: Initially choose door 1. After Monty opens a door, switch to one of the other two unopened doors, randomly choosing between them (with probability $1/2$ each).

(a) Find the unconditional probability that the contestant will get the car.

Hint: Condition on where the car is.

(b) Find the unconditional probability that Monty will reveal the apple.

Hint: Condition on what is behind door 1.

(c) Monty now opens a door, revealing the apple. Given this information, find the conditional probability that the contestant will get the car.

Solution:

(a) Let W be the event that the contestant wins the car, and G_j, A_j, B_j, and C_j be the events that Door j has a goat, apple, book, and car, respectively. By LOTP,

$$P(W) = \sum_{j=1}^{4} P(W|C_j)P(C_j) = \frac{1}{4}\left(0 + \frac{1}{2} + \frac{1}{2} + \frac{1}{2}\right) = \frac{3}{8}.$$

(b) Let A be the event that Monty reveals the apple. By LOTP,

$$P(A) = P(A|G_1)P(G_1) + P(A|A_1)P(A_1) + P(A|B_1)P(B_1) + P(A|C_1)P(C_1)$$
$$= \frac{p + 0 + (1 - p) + (1 - p)}{4}$$
$$= \frac{2 - p}{4}.$$

(c) By Bayes' rule,

$$P(C_1|A) = \frac{P(A|C_1)P(C_1)}{P(A)} = \frac{(1 - p)/4}{(2 - p)/4} = \frac{1 - p}{2 - p}.$$

For the contestant to win the car, C_1 must *not* occur, and then the contestant's random choice of which door to switch to must be the door with the car. So

$$P(W|A) = \left(1 - \frac{1 - p}{2 - p}\right) \cdot \frac{1}{2} = \frac{1}{2(2 - p)}.$$

47. You are the contestant on Monty Hall's game show. Hoping to double the excitement of the game, Monty will offer you *two* opportunities to switch to another door. Specifically, the new rules are as follows. There are *four* doors. Behind one door there is a car (which you want); behind the other three doors there are goats (which you don't want). Initially, all possibilities are equally likely for where the car is. Monty knows where the car is, and when he has a choice of which door to open, he chooses with equal probabilities.

You choose a door, which for concreteness we assume is door 1. Monty opens a door (other than door 1), revealing a goat, and then offers you the option to switch to another door. Monty then opens *another* door (other than your currently selected door), revealing another goat. So now there are two open doors (with goats) and two unopened doors. Again Monty offers you the option to switch.

You decide in advance to use one of the following four strategies: stay-stay, stay-switch, switch-stay, switch-switch, where, for example, "stay-switch" means that the first time Monty offers you the choice of switching, you stay with your current selection, but then the second time Monty offers you the choice, you do switch doors. In each part below the goal is to find or compare *unconditional* probabilities, i.e., from a vantage point of before the game has started.

(a) Find the probability of winning the car if you follow the stay-stay strategy.

(b) Find the probability of winning the car if you follow the stay-switch strategy.

(c) Find the probability of winning the car if you follow the switch-stay strategy.

(d) Find the probability of winning the car if you follow the switch-switch strategy.

(e) Which of these four strategies is the best?

Solution: Let C_j be the event that door j has a car, G_j be the event that door j has a goat, and W be the event that you win the car.

(a) With the stay-stay strategy, you win if and only if door 1 has the car, so

$$P(W) = P(C_1) = \frac{1}{4}.$$

(b) With the stay-switch strategy, you win if and only if door 1 does not have the car (if door 1 has the car, the switch part of stay-switch means you will not stay with door 1; if door 1 does not have the car, Monty will have to open the two other doors that have goats, and then you will switch to the door with the car). So

$$P(W) = P(G_1) = \frac{3}{4}.$$

Note that this is 1 minus the answer to (a), which makes sense since both of these strategies first stay with door 1 and then at the time of your second opportunity to switch, one strategy will yield the car and the other strategy will yield a goat. This shows that the *conditional* probability of success with stay-stay, given Monty's actions is 1 minus the conditional probability of success with stay-switch, given Monty's actions. But it also follows that the same relationship holds between the unconditional probabilities, by LOTP.

(c) For the switch-stay strategy, expand using LOTP:

$$P(W) = P(W|C_1)P(C_1) + P(W|C_2)P(C_2) + P(W|C_3)P(C_3) + P(W|C_4)P(C_4).$$

We have $P(W|C_1) = 0$ since switch-stay always ends up on a door other than door 1. If the car is behind door 2, then Monty will open door 3 or door 4 first, and then you will win if and only if you switch to door 2 rather than whichever of doors 3,4 Monty

did not open. So $P(W|C_2) = 1/2$. By symmetry or using the same argument, we also have $P(W|C_3) = P(W|C_4) = 1/2$. Thus,

$$P(W) = 3 \cdot \frac{1}{2} \cdot \frac{1}{4} = \frac{3}{8}.$$

(d) For the switch-switch strategy, $P(W|C_1) = 1$ since if door 1 has the car, Monty will reveal a goat, you will switch doors, Monty will reveal another goat, and then you will switch back to door 1. And $P(W|C_2) = 1/2$ since if door 2 has the car, then Monty will open door 3 or door 4 first, and then you will win if and only if you do *not* switch to door 2 on your first switch. Similarly, $P(W|C_3) = P(W|C_4) = 1/2$. So

$$P(W) = 1 \cdot \frac{1}{4} + 3 \cdot \frac{1}{2} \cdot \frac{1}{4} = \frac{5}{8}.$$

Note that this is 1 minus the answer to (c), which again makes sense.

(e) The best of these strategies is stay-switch, since the answer to (b) is larger than each of the answers to (a), (c), and (d).

First-step analysis and gambler's ruin

48. Ⓢ A fair die is rolled repeatedly, and a running total is kept (which is, at each time, the total of all the rolls up until that time). Let p_n be the probability that the running total is ever *exactly* n (assume the die will always be rolled enough times so that the running total will eventually exceed n, but it may or may not ever equal n).

(a) Write down a recursive equation for p_n (relating p_n to earlier terms p_k in a simple way). Your equation should be true for all positive integers n, so give a definition of p_0 and p_k for $k < 0$ so that the recursive equation is true for small values of n.

(b) Find p_7.

(c) Give an intuitive explanation for the fact that $p_n \to 1/3.5 = 2/7$ as $n \to \infty$.

Solution:

(a) We will find something to condition on to reduce the case of interest to earlier, simpler cases. This is achieved by the useful strategy of *first step anaysis*. Let p_n be the probability that the running total is ever *exactly* n. Note that if, for example, the first throw is a 3, then the probability of reaching n exactly is p_{n-3} since starting from that point, we need to get a total of $n - 3$ exactly. So

$$p_n = \frac{1}{6}(p_{n-1} + p_{n-2} + p_{n-3} + p_{n-4} + p_{n-5} + p_{n-6}),$$

where we define $p_0 = 1$ (which makes sense anyway since the running total is 0 before the first toss) and $p_k = 0$ for $k < 0$.

(b) Using the recursive equation in (a), we have

$$p_1 = \frac{1}{6}, \quad p_2 = \frac{1}{6}(1 + \frac{1}{6}), \quad p_3 = \frac{1}{6}(1 + \frac{1}{6})^2,$$
$$p_4 = \frac{1}{6}(1 + \frac{1}{6})^3, \quad p_5 = \frac{1}{6}(1 + \frac{1}{6})^4, \quad p_6 = \frac{1}{6}(1 + \frac{1}{6})^5.$$

Hence,

$$p_7 = \frac{1}{6}(p_1 + p_2 + p_3 + p_4 + p_5 + p_6) = \frac{1}{6}\left((1 + \frac{1}{6})^6 - 1\right) \approx 0.2536.$$

(c) An intuitive explanation is as follows. The average number thrown by the die is (total of dots)/6, which is $21/6 = 7/2$, so that every throw adds on an average of $7/2$. We can therefore expect to land on 2 out of every 7 numbers, and the probability of landing on any particular number is $2/7$. A mathematical derivation (which was not requested in the problem) can be given as follows:

$$p_{n+1} + 2p_{n+2} + 3p_{n+3} + 4p_{n+4} + 5p_{n+5} + 6p_{n+6}$$
$$= p_{n+1} + 2p_{n+2} + 3p_{n+3} + 4p_{n+4} + 5p_{n+5}$$
$$\quad + p_n + p_{n+1} + p_{n+2} + p_{n+3} + p_{n+4} + p_{n+5}$$
$$= p_n + 2p_{n+1} + 3p_{n+2} + 4p_{n+3} + 5p_{n+4} + 6p_{n+5}$$
$$= \cdots$$
$$= p_{-5} + 2p_{-4} + 3p_{-3} + 4p_{-2} + 5p_{-1} + 6p_0 = 6.$$

Taking the limit of the lefthand side as n goes to ∞, we have

$$(1 + 2 + 3 + 4 + 5 + 6) \lim_{n \to \infty} p_n = 6,$$

so $\lim_{n \to \infty} p_n = 2/7$.

49. A sequence of $n \geq 1$ independent trials is performed, where each trial ends in "success" or "failure" (but not both). Let p_i be the probability of success in the ith trial, $q_i = 1 - p_i$, and $b_i = q_i - 1/2$, for $i = 1, 2, \ldots, n$. Let A_n be the event that the number of successful trials is even.

(a) Show that for $n = 2$, $P(A_2) = 1/2 + 2b_1 b_2$.

(b) Show by induction that

$$P(A_n) = 1/2 + 2^{n-1} b_1 b_2 \ldots b_n.$$

(This result is very useful in cryptography. Also, note that it implies that if n coins are flipped, then the probability of an even number of Heads is $1/2$ if and only if at least one of the coins is fair.) Hint: Group some trials into a supertrial.

(c) Check directly that the result of (b) is true in the following simple cases: $p_i = 1/2$ for some i; $p_i = 0$ for all i; $p_i = 1$ for all i.

Solution:

(a) We have

$$P(A_2) = p_1 p_2 + q_1 q_2 = (\frac{1}{2} - b_1)(\frac{1}{2} - b_2) + (\frac{1}{2} + b_1)(\frac{1}{2} + b_2) = \frac{1}{2} + 2b_1 b_2.$$

(b) For $n = 1$, $P(A_1) = P(\text{1st trial is a failure}) = q_1 = 1/2 + b_1$. For $n = 2$, the result was shown in (a). Now assume the result is true for n, and prove it for $n + 1$, where $n \geq 2$ is fixed. Think of the first n trials as 1 supertrial, where "success" is defined as the number of successes being odd. By assumption, the probability of failure for the supertrial is $\tilde{q}_1 = \frac{1}{2} + 2^{n-1} b_1 \ldots b_n$. Let $\tilde{b}_1 = \tilde{q}_1 - \frac{1}{2}$. Then by (a),

$$P(A_{n+1}) = \frac{1}{2} + 2\tilde{b}_1 b_{n+1} = \frac{1}{2} + 2^n b_1 b_2 \ldots b_n b_{n+1},$$

which completes the induction. This result, called the *piling-up lemma*, is used in cryptography to compute or bound probabilities needed for some widely-used ciphers.

(c) Let $p_i = 1/2$ for some fixed i. Then $b_i = 0$ and we need to check directly that $P(A_n) = 1/2$. Let p be the probability that the number of successes in the $n - 1$ trials other than the ith is odd. Conditioning on the result of the ith trial, we have

$$P(A_n) = p \cdot \frac{1}{2} + (1 - p) \cdot \frac{1}{2} = \frac{1}{2}.$$

For the case that $p_i = 0$ for all i, the event A_n is guaranteed to occur (since there will be 0 successes), which agrees with $P(A_n) = 1/2 + 2^{n-1} \cdot 2^{-n} = 1$. For the case that $p_i = 1$ for all i, there will be n successes, so A_n occurs if and only if n is even. This agrees with (b), which simplifies in this case to $P(A_n) = (1 + (-1)^n)/2$.

50. Ⓢ Calvin and Hobbes play a match consisting of a series of games, where Calvin has probability p of winning each game (independently). They play with a "win by two" rule: the first player to win two games more than his opponent wins the match. Find the probability that Calvin wins the match (in terms of p), in two different ways:

(a) by conditioning, using the law of total probability.

(b) by interpreting the problem as a gambler's ruin problem.

Solution:

(a) Let C be the event that Calvin wins the match, $X \sim \text{Bin}(2, p)$ be how many of the first 2 games he wins, and $q = 1 - p$. Then

$$P(C) = P(C|X = 0)q^2 + P(C|X = 1)(2pq) + P(C|X = 2)p^2 = 2pqP(C) + p^2,$$

so $P(C) = \frac{p^2}{1-2pq}$. This can also be written as $\frac{p^2}{p^2+q^2}$, since $p + q = 1$.

Sanity check: Note that this should (and does) reduce to 1 for $p = 1$, 0 for $p = 0$, and $\frac{1}{2}$ for $p = \frac{1}{2}$. Also, it makes sense that the probability of Hobbes winning, which is $1 - P(C) = \frac{q^2}{p^2+q^2}$, can also be obtained by swapping p and q.

(b) The problem can be thought of as a gambler's ruin where each player starts out with \$2. So the probability that Calvin wins the match is

$$\frac{1 - (q/p)^2}{1 - (q/p)^4} = \frac{(p^2 - q^2)/p^2}{(p^4 - q^4)/p^4} = \frac{(p^2 - q^2)/p^2}{(p^2 - q^2)(p^2 + q^2)/p^4} = \frac{p^2}{p^2 + q^2},$$

which agrees with the above.

51. Ⓢ A gambler repeatedly plays a game where in each round, he wins a dollar with probability 1/3 and loses a dollar with probability 2/3. His strategy is "quit when he is ahead by \$2". Suppose that he starts with a million dollars. Show that the probability that he'll ever be ahead by \$2 is less than 1/4.

Solution: This is a special case of the gambler's ruin problem. Let A_1 be the event that he is successful on the first play and let W be the event that he is ever ahead by \$2 before being ruined. Then by the law of total probability, we have

$$P(W) = P(W|A_1)P(A_1) + P(W|A_1^c)P(A_1^c).$$

Let a_i be the probability that the gambler achieves a profit of \$2 before being ruined, starting with a fortune of \$$i$. For our setup, $P(W) = a_i$, $P(W|A_1) = a_{i+1}$ and $P(W|A_1^c) = a_{i-1}$. Therefore,

$$a_i = a_{i+1}/3 + 2a_{i-1}/3,$$

with boundary conditions $a_0 = 0$ and $a_{i+2} = 1$. We can then solve this difference equation for a_i (directly or using the result of the gambler's ruin problem):

$$a_i = \frac{2^i - 1}{2^{2+i} - 1}.$$

This is always less than 1/4 since $\frac{2^i - 1}{2^{2+i} - 1} < \frac{1}{4}$ is equivalent to $4(2^i - 1) < 2^{2+i} - 1$, which is equivalent to the true statement $2^{2+i} - 4 < 2^{2+i} - 1$.

52. As in the gambler's ruin problem, two gamblers, A and B, make a series of bets, until one of the gamblers goes bankrupt. Let A start out with i dollars and B start out with $N - i$ dollars, and let p be the probability of A winning a bet, with $0 < p < \frac{1}{2}$. Each bet is for $\frac{1}{k}$ dollars, with k a positive integer, e.g., $k = 1$ is the original gambler's ruin problem and $k = 20$ means they're betting nickels. Find the probability that A wins the game, and determine what happens to this as $k \to \infty$.

Solution: Define 1 kidollar to be $\frac{1}{k}$ dollars. This problem is exactly the gambler's ruin problem if everything is written in terms of kidollars as the unit of currency, e.g., for $k = 20$ it's just the gambler's ruin problem, with nickels instead of dollars. Note that A starts out with ki kidollars and B starts out with $k(N - i)$ kidollars. Let $r = (1 - p)/p$. By the gambler's ruin problem, the probability that A wins the game is

$$P(\text{A wins}) = \frac{1 - r^{ki}}{1 - r^{kN}}.$$

This is 0 if $i = 0$ and 1 if $i = N$. For $1 \le i \le N - 1$, since $r > 1$ we have

$$\lim_{k \to \infty} \frac{1 - r^{ki}}{1 - r^{kN}} = \lim_{k \to \infty} \frac{r^{ki}}{r^{kN}} = \lim_{k \to \infty} r^{k(i - N)} = 0.$$

53. There are 100 equally spaced points around a circle. At 99 of the points, there are sheep, and at 1 point, there is a wolf. At each time step, the wolf randomly moves either clockwise or counterclockwise by 1 point. If there is a sheep at that point, he eats it. The sheep don't move. What is the probability that the sheep who is initially opposite the wolf is the last one remaining?

Solution: Call the sheep initially opposite the wolf Dolly. If Dolly is the last sheep surviving, then both of Dolly's neighbors get eaten before Dolly. But the converse is also true: if both of Dolly's neighbors get eaten before Dolly, then the wolf must have gone all the way around the long way after eating the first neighbor to get to the other neighbor.

Now consider the moment just after the wolf has eaten the first of Dolly's neighbors. The question then becomes whether the wolf will reach Dolly first or the other neighbor first. This is the same as the gambler's ruin problem, viewed as a random walk started at 1 and ending when it reaches either 0 or 99 (reaching 0 corresponds to eating Dolly). Thus, the probability that Dolly is the last sheep surviving is 1/99. (Similarly, it can be shown that the sheep are all equally likely to be the last sheep surviving!)

54. An immortal drunk man wanders around randomly on the integers. He starts at the origin, and at each step he moves 1 unit to the right or 1 unit to the left, with probabilities p and $q = 1 - p$ respectively, independently of all his previous steps. Let S_n be his position after n steps.

(a) Let p_k be the probability that the drunk ever reaches the value k, for all $k \ge 0$. Write down a difference equation for p_k (you do not need to solve it for this part).

(b) Find p_k, fully simplified; be sure to consider all 3 cases: $p < 1/2, p = 1/2$, and $p > 1/2$. Feel free to assume that if $A_1, A_2, \ldots$ are events with $A_j \subseteq A_{j+1}$ for all j, then $P(A_n) \to P(\cup_{j=1}^{\infty} A_j)$ as $n \to \infty$ (because it is true; this is known as *continuity of probability*).

Solution:

(a) Conditioning on the first step,

$$p_k = pp_{k-1} + qp_{k+1}$$

for all $k \ge 1$, with $p_0 = 1$.

(b) For fixed k and any positive integer j, let A_j be the event that the drunk reaches k before ever reaching $-j$. Then $A_j \subseteq A_{j+1}$ for all m since the drunk would have to walk past $-j$ to reach $-j-1$. Also, $\cup_{j=1}^{\infty} A_j$ is the event that the drunk ever reaches k, since if he reaches $-j$ before k for *all* j, then he will never have time to get to k. Now we just need to find $\lim_{j\to\infty} P(A_j)$, where we already know $P(A_j)$ from the result of the gambler's ruin problem!

For $p = 1/2$,

$$P(A_j) = \frac{j}{j+k} \to 1 \text{ as } j \to \infty.$$

For $p > 1/2$,

$$P(A_j) = \frac{1 - (\frac{q}{p})^j}{1 - (\frac{q}{p})^{j+k}} \to 1 \text{ as } j \to \infty,$$

since $(q/p)^j \to 0$. For $p < 1/2$,

$$P(A_j) = \frac{1 - (\frac{q}{p})^j}{1 - (\frac{q}{p})^{j+k}} \to \left(\frac{p}{q}\right)^k \text{ as } j \to \infty,$$

since $(q/p)^j \to \infty$ so the 1's in the numerator and denominator become negligible as $j \to \infty$.

Simpson's paradox

55. Ⓢ (a) Is it possible to have events A, B, C such that $P(A|C) < P(B|C)$ and $P(A|C^c) < P(B|C^c)$, yet $P(A) > P(B)$? That is, A is less likely than B given that C is true, and also less likely than B given that C is false, yet A is more likely than B if we're given no information about C. Show this is impossible (with a short proof) or find a counterexample (with a story interpreting A, B, C).

(b) If the scenario in (a) is possible, is it a special case of Simpson's paradox, equivalent to Simpson's paradox, or neither? If it is impossible, explain intuitively why it is impossible even though Simpson's paradox is possible.

Solution:

(a) It is *not* possible, as seen using the law of total probability:

$$P(A) = P(A|C)P(C) + P(A|C^c)P(C^c) < P(B|C)P(C) + P(B|C^c)P(C^c) = P(B).$$

(b) In Simpson's paradox, using the notation from the chapter, we can expand out $P(A|B)$ and $P(A|B^c)$ using LOTP to condition on C, but the inequality can flip because of the weights such as $P(C|B)$ on the terms (e.g., Dr. Nick performs a lot more Band-Aid removals than Dr. Hibbert). In this problem, the weights $P(C)$ and $P(C^c)$ are the same in both expansions, so the inequality is preserved.

56. Ⓢ Consider the following conversation from an episode of *The Simpsons*:

> Lisa: *Dad, I think he's an ivory dealer! His boots are ivory, his hat is ivory, and I'm pretty sure that check is ivory.*
> Homer: *Lisa, a guy who has lots of ivory is* less *likely to hurt Stampy than a guy whose ivory supplies are low.*

Here Homer and Lisa are debating the question of whether or not the man (named Blackheart) is likely to hurt Stampy the Elephant if they sell Stampy to him. They clearly disagree about how to use their observations about Blackheart to learn about the probability (conditional on the evidence) that Blackheart will hurt Stampy.

(a) Define clear notation for the various events of interest here.

(b) Express Lisa's and Homer's arguments (Lisa's is partly implicit) as conditional probability statements in terms of your notation from (a).

(c) Assume it is true that someone who has a lot of a commodity will have less desire to acquire more of the commodity. Explain what is wrong with Homer's reasoning that the evidence about Blackheart makes it less likely that he will harm Stampy.

Solution:

(a) Let H be the event that the man will hurt Stampy, let L be the event that a man has lots of ivory, and let D be the event that the man is an ivory dealer.

(b) Lisa observes that L is true. She suggests (reasonably) that this evidence makes D more likely, i.e., $P(D|L) > P(D)$. Implicitly, she suggests that this makes it likely that the man will hurt Stampy, i.e.,

$$P(H|L) > P(H|L^c).$$

Homer argues that

$$P(H|L) < P(H|L^c).$$

(c) Homer does not realize that observing that Blackheart has so much ivory makes it much more likely that Blackheart is an ivory dealer, which in turn makes it more likely that the man will hurt Stampy. This is an example of Simpson's paradox. It may be true that, *controlling for whether or not Blackheart is a dealer*, having high ivory supplies makes it less likely that he will harm Stampy: $P(H|L, D) < P(H|L^c, D)$ and $P(H|L, D^c) < P(H|L^c, D^c)$. However, this does not imply that $P(H|L) < P(H|L^c)$.

57. (a) There are two crimson jars (labeled C_1 and C_2) and two mauve jars (labeled M_1 and M_2). Each jar contains a mixture of green gummi bears and red gummi bears. Show by example that it is possible that C_1 has a much higher percentage of green gummi bears than M_1, and C_2 has a much higher percentage of green gummi bears than M_2, yet if the contents of C_1 and C_2 are merged into a new jar and likewise for M_1 and M_2, then the combination of C_1 and C_2 has a lower percentage of green gummi bears than the combination of M_1 and M_2.

(b) Explain how (a) relates to Simpson's paradox, both intuitively and by explicitly defining events A, B, C as in the statement of Simpson's paradox.

Solution:

(a) As an example, let C_1 have 9 green, 1 red; M_1 have 50 green, 50 red; C_2 have 30 green, 70 red; M_2 have 1 green, 9 red.

(b) This is a form of Simpson's paradox since which jar color is more likely to provide a green gummy bear flips depending on whether the jars get aggregated. To match this example up to the notation used in the statement of Simpson's paradox, let A be the event that a red gummi bear is chosen in the random draw, B be the event that it is drawn from a crimson jar, and C be the event that it is drawn from a jar with index 1. With the numbers from the solution to (a), $P(C|B) = 1/11$ is much less than $P(C|B^c) = 10/11$, which enables us to have $P(A|B) < P(A|B^c)$ even though the inequalities go the other way when we also condition on C or on C^c.

58. As explained in this chapter, Simpson's paradox says that it is possible to have events A, B, C such that $P(A|B, C) < P(A|B^c, C)$ and $P(A|B, C^c) < P(A|B^c, C^c)$, yet $P(A|B) > P(A|B^c)$.

(a) Can Simpson's paradox occur if A and B are independent? If so, give a concrete example (with both numbers and an interpretation); if not, prove that it is impossible.

(b) Can Simpson's paradox occur if A and C are independent? If so, give a concrete example (with both numbers and an interpretation); if not, prove that it is impossible.

(c) Can Simpson's paradox occur if B and C are independent? If so, give a concrete example (with both numbers and an interpretation); if not, prove that it is impossible.

Solution:

(a) No, since if A and B are independent, then $P(A|B) = P(A) = P(A|B^c)$, using the fact that A is also independent of B^c.

(b) No, as shown by the following. Suppose that A and C are independent and that the first two inequalities in Simpson's paradox hold. Then by LOTP,

$$\begin{aligned} P(A) &= P(A|C) \\ &= P(A|C,B)P(B|C) + P(A|C,B^c)P(B^c|C) \\ &< P(A|C,B^c)P(B|C) + P(A|C,B^c)P(B^c|C) \\ &= P(A|C,B^c), \end{aligned}$$

so

$$P(A) < P(A|C,B^c).$$

Similarly, $P(A) = P(A|C^c)$ gives

$$P(A) < P(A|C^c,B^c).$$

Thus,

$$P(A) < P(A|B^c),$$

since $P(A|B^c)$ is a weighted average of $P(A|C,B^c)$ and $P(A|C^c,B^c)$ (so is in between them). But then

$$P(A|B) < P(A|B^c),$$

since $P(A)$ is a weighted average of $P(A|B)$ and $P(A|B^c)$ (so is in between them).

(c) No, as shown by the following. Suppose that B and C are independent and that the first two inequalities in Simpson's paradox hold. Then LOTP (as on p. 60) yields

$$\begin{aligned} P(A|B) &= P(A|C,B)P(C|B) + P(A|C^c,B)P(C^c|B) \\ &< P(A|C,B^c)P(C|B^c) + P(A|C^c,B^c)P(C^c|B^c) \\ &= P(A|B^c). \end{aligned}$$

59. Ⓢ The book *Red State, Blue State, Rich State, Poor State* by Andrew Gelman [13] discusses the following election phenomenon: within any U.S. state, a wealthy voter is more likely to vote for a Republican than a poor voter, yet the wealthier states tend to favor Democratic candidates! In short: rich individuals (in any state) tend to vote for Republicans, while states with a higher percentage of rich people tend to favor Democrats.

(a) Assume for simplicity that there are only 2 states (called Red and Blue), each of which has 100 people, and that each person is either rich or poor, and either a Democrat or a Republican. Make up numbers consistent with the above, showing how this phenomenon is possible, by giving a 2×2 table for each state (listing how many people in each state are rich Democrats, etc.).

(b) In the setup of (a) (not necessarily with the numbers you made up there), let D be the event that a randomly chosen person is a Democrat (with all 200 people equally likely), and B be the event that the person lives in the Blue State. Suppose that 10 people move from the Blue State to the Red State. Write P_{old} and P_{new} for probabilities before and after they move. Assume that people do not change parties,

so we have $P_{\text{new}}(D) = P_{\text{old}}(D)$. Is it possible that *both* $P_{\text{new}}(D|B) > P_{\text{old}}(D|B)$ and $P_{\text{new}}(D|B^c) > P_{\text{old}}(D|B^c)$ are true? If so, explain how it is possible and why it does not contradict the law of total probability $P(D) = P(D|B)P(B) + P(D|B^c)P(B^c)$; if not, show that it is impossible.

Solution:

(a) Here are two tables that are as desired:

Red	Dem	Rep	**Total**
Rich	5	25	30
Poor	20	50	70
Total	25	75	100

Blue	Dem	Rep	**Total**
Rich	45	15	60
Poor	35	5	40
Total	80	20	100

In these tables, within each state a rich person is more likely to be a Republican than a poor person; but the richer state has a higher percentage of Democrats than the poorer state. Of course, there are many possible tables that work.

The above example is a form of Simpson's paradox: aggregating the two tables seems to give different conclusions than conditioning on which state a person is in. Letting D, W, B be the events that a randomly chosen person is a Democrat, wealthy, and from the Blue State (respectively), for the above numbers we have $P(D|W, B) < P(D|W^c, B)$ and $P(D|W, B^c) < P(D|W^c, B^c)$ (controlling for whether the person is in the Red State or the Blue State, a poor person is more likely to be a Democrat than a rich person), but $P(D|W) > P(D|W^c)$ (stemming from the fact that the Blue State is richer and more Democratic).

(b) Yes, it is possible. Suppose with the numbers from (a) that 10 people move from the Blue State to the Red State, of whom 5 are Democrats and 5 are Republicans. Then $P_{\text{new}}(D|B) = 75/90 > 80/100 = P_{\text{old}}(D|B)$ and $P_{\text{new}}(D|B^c) = 30/110 > 25/100 = P_{\text{old}}(D|B^c)$. Intuitively, this makes sense since the Blue State has a higher percentage of Democrats initially than the Red State, and the people who move have a percentage of Democrats which is between these two values.

This result does not contradict the law of total probability since the weights $P(B), P(B^c)$ also change: $P_{\text{new}}(B) = 90/200$, while $P_{\text{old}}(B) = 1/2$. The phenomenon could not occur if an equal number of people also move from the Red State to the Blue State (so that $P(B)$ is kept constant).

Mixed practice

60. A patient is being given a blood test for the disease conditionitis. Let p be the prior probability that the patient has conditionitis. The blood sample is sent to one of two labs for analysis, lab A or lab B. The choice of which lab to use is made randomly, independent of the patient's disease status, with probability $1/2$ for each lab.

For lab A, the probability of someone testing *positive* given that they *do* have the disease is a_1, and the probability of someone testing *negative* given that they do *not* have the disease is a_2. The corresponding probabilities for lab B are b_1 and b_2.

(a) Find the probability that the patient has the disease, given that they tested positive.

(b) Find the probability that the patient's blood sample was analyzed by lab A, given that the patient tested positive.

Solution:

(a) Let D, T, A, and B be the events that the patient has the disease, tests positive, has

their blood sample analyzed by lab A, and has their blood sample analyzed by lab B, respectively. By Bayes' rule,

$$P(D|T) = \frac{P(T|D)P(D)}{P(T)}.$$

By LOTP with extra conditioning,

$$P(T|D) = P(T|D, A)P(A|D) + P(T|D, B)P(B|D) = a_1/2 + b_1/2,$$

$$P(T|D^c) = P(T|D^c, A)P(A|D^c) + P(T|D^c, B)P(B|D^c) = (1 - a_2)/2 + (1 - b_2)/2.$$

Thus,

$$P(T) = P(T|D)P(D) + P(T|D^c)P(D^c) = (a_1 + b_1)p/2 + (2 - a_2 - b_2)(1 - p)/2,$$

and

$$P(D|T) = \frac{(a_1 + b_1)p}{(a_1 + b_1)p + (2 - a_2 - b_2)(1 - p)}.$$

(b) Again by Bayes' rule,

$$P(A|T) = \frac{P(T|A)P(A)}{P(T)} = \frac{P(T|A)}{2P(T)}.$$

We found $P(T)$ already in (a). To find $P(T|A)$, expand

$$P(T|A) = P(T|A, D)P(D|A) + P(T|A, D^c)P(D^c|A) = a_1p + (1 - a_2)(1 - p).$$

Therefore,

$$P(A|T) = \frac{a_1p + (1 - a_2)(1 - p)}{(a_1 + b_1)p + (2 - a_2 - b_2)(1 - p)}.$$

61. Fred decides to take a series of n tests, to diagnose whether he has a certain disease (any individual test is not perfectly reliable, so he hopes to reduce his uncertainty by taking multiple tests). Let D be the event that he has the disease, $p = P(D)$ be the prior probability that he has the disease, and $q = 1 - p$. Let T_j be the event that he tests positive on the jth test.

(a) Assume for this part that the test results are conditionally independent given Fred's disease status. Let $a = P(T_j|D)$ and $b = P(T_j|D^c)$, where a and b don't depend on j. Find the posterior probability that Fred has the disease, given that he tests positive on all n of the n tests.

(b) Suppose that Fred tests positive on all n tests. However, some people have a certain gene that makes them *always* test positive. Let G be the event that Fred has the gene. Assume that $P(G) = 1/2$ and that D and G are independent. If Fred does *not* have the gene, then the test results are conditionally independent given his disease status. Let $a_0 = P(T_j|D, G^c)$ and $b_0 = P(T_j|D^c, G^c)$, where a_0 and b_0 don't depend on j. Find the posterior probability that Fred has the disease, given that he tests positive on all n of the tests.

Solution:

(a) Let $T = T_1 \cap \cdots \cap T_n$ be the event that Fred tests positive on all the tests. By Bayes' rule and LOTP,

$$P(D|T) = \frac{P(T|D)P(D)}{P(T)} = \frac{pa^n}{pa^n + qb^n}.$$

(b) Let T be the event that Fred tests positive on all n tests. Then

$$P(D|T) = \frac{P(T|D)P(D)}{P(T)} = \frac{pP(T|D)}{pP(T|D) + qP(T|D^c)}.$$

Conditioning on whether or not he has the gene, we have

$$P(T|D) = P(T|D,G)P(G|D) + P(T|D,G^c)P(G^c|D) = \frac{1}{2} + \frac{a_0^n}{2},$$

$$P(T|D^c) = P(T|D^c,G)P(G|D^c) + P(T|D^c,G^c)P(G^c|D^c) = \frac{1}{2} + \frac{b_0^n}{2}.$$

Thus,

$$P(D|T) = \frac{p(1 + a_0^n)}{p(1 + a_0^n) + q(1 + b_0^n)}.$$

62. A certain hereditary disease can be passed from a mother to her children. Given that the mother has the disease, her children independently will have it with probability $1/2$. Given that she doesn't have the disease, her children won't have it either. A certain mother, who has probability $1/3$ of having the disease, has two children.

(a) Find the probability that neither child has the disease.

(b) Is whether the elder child has the disease independent of whether the younger child has the disease? Explain.

(c) The elder child is found not to have the disease. A week later, the younger child is also found not to have the disease. Given this information, find the probability that the mother has the disease.

Solution:

(a) Let M, A, B be the events that the mother, elder child, and younger child have the disease (respectively). Then

$$P(A^c, B^c) = P(A^c, B^c|M)P(M) + P(A^c, B^c|M^c)P(M^c) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{3} + \frac{2}{3} = \frac{3}{4}.$$

(b) These events are conditionally independent given the disease status of the mother, but they are not independent. Knowing whether the elder child has the disease gives information about whether the mother has the disease, which in turn gives information about whether the younger child has the disease.

(c) By Bayes' rule,

$$P(M|A^c, B^c) = \frac{P(A^c, B^c|M)P(M)}{P(A^c, B^c)} = \frac{\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{3}}{\frac{3}{4}} = \frac{1}{9}.$$

Alternatively, we can do the conditioning in two steps: first condition on A^c, giving

$$P(M|A^c) = \frac{P(A^c|M)P(M)}{P(A^c)} = \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2} \cdot \frac{1}{3} + \frac{2}{3}} = \frac{1}{5}.$$

Then do further conditioning on B^c, giving

$$P(M|A^c, B^c) = \frac{P(B^c|M, A^c)P(M|A^c)}{P(B^c|A^c)} = \frac{\frac{1}{2} \cdot \frac{1}{5}}{\frac{1}{2} \cdot \frac{1}{5} + \frac{4}{5}} = \frac{1}{9},$$

which agrees with the result of the one-step method.

63. Three fair coins are tossed at the same time. Explain what is wrong with the following argument: "there is a 50% chance that the three coins all landed the same way, since obviously it is possible to find two coins that match, and then the other coin has a 50% chance of matching those two".

Solution: The probability of all three coins landing the same way is $1/8 + 1/8 = 1/4$,

so clearly there is something wrong with the argument. What is wrong is that *which* two out of the three coins match is not specified, so "the other coin" is not well-defined. Given that the first two flips match, it is true that there is a 50% chance that the third flip matches those, but knowing that there is *at least one* pair of tosses that match does not provide a *specific* pair that matches. (The argument under discussion is known as *Galton's paradox*.)

64. An urn contains red, green, and blue balls. Let r, g, b be the proportions of red, green, blue balls, respectively ($r + g + b = 1$).

(a) Balls are drawn randomly *with replacement*. Find the probability that the first time a green ball is drawn is before the first time a blue ball is drawn.

Hint: Explain how this relates to finding the probability that a draw is green, given that it is either green or blue.

(b) Balls are drawn randomly *without replacement*. Find the probability that the first time a green ball is drawn is before the first time a blue ball is drawn. Is the answer the same or different than the answer in (a)?

Hint: Imagine the balls all lined up, in the order in which they will be drawn. Note that where the red balls are standing in this line is irrelevant.

(c) Generalize the result from (a) to the following setting. Independent trials are performed, and the outcome of each trial is classified as being exactly one of type 1, type 2, ..., or type n, with probabilities $p_1, p_2, \ldots, p_n$, respectively. Find the probability that the first trial to result in type i comes before the first trial to result in type j, for $i \neq j$.

Solution:

(a) Red balls are irrelevant here. What matters is whether the first non-red ball drawn is green or blue. The probability that this ball is green is

$$P(\text{green}|\text{green or blue}) = \frac{P(\text{green})}{P(\text{green or blue})} = \frac{g}{g + b}.$$

(b) Let N be the total number of balls, so there are rN, gN, bN red, green, and blue balls, respectively. Line the balls up in the order in which they will be drawn. Now look at the first non-red ball. By symmetry, it is equally like to be any of the $gN + bN$ non-red balls. The probability that it is green is

$$\frac{gN}{gN + bN} = \frac{g}{g + b},$$

which is the same as the answer in (a).

(c) Arguing as in (a), the probability of getting type i before type j (for $i \neq j$ is

$$P(\text{type } i|\text{type } i \text{ or type } j) = \frac{P(\text{type } i)}{P(\text{type } i \text{ or type } j)} = \frac{p_i}{p_i + p_j}.$$

65. Marilyn vos Savant was asked the following question for her column in *Parade*:

You're at a party with 199 other guests when robbers break in and announce that they are going to rob one of you. They put 199 blank pieces of paper in a hat, plus one marked "you lose." Each guest must draw, and the person who draws "you lose" will get robbed. The robbers offer you the option of drawing first, last, or at any time in between. When would you take your turn?

The draws are made *without replacement*, and for (a) are uniformly random.

(a) Determine whether it is optimal to draw first, last, or somewhere in between (or whether it does not matter), to maximize the probability of not being robbed. Give a clear, concise, and compelling explanation.

(b) More generally, suppose that there is one "you lose" piece of paper, with "weight" v, and there are n blank pieces of paper, each with "weight" w. At each stage, draws are made with probability proportional to weight, i.e., the probability of drawing a particular piece of paper is its weight divided by the sum of the weights of all the remaining pieces of paper. Determine whether it is better to draw first or second (or whether it does not matter); here $v > 0, w > 0$, and $n \geq 1$ are known constants.

Solution:

(a) By symmetry, it does not matter: unconditionally, the jth draw is equally likely to be any of the 200 pieces of paper.

(b) Drawing first, the probability of being robbed is $v/(v + nw)$. Drawing second, by LOTP the probability of being robbed is

$$0 \cdot \frac{v}{v + nw} + \frac{v}{v + (n-1)w} \cdot \frac{nw}{v + nw} = \frac{vnw}{(v + (n-1)w)(v + nw)}.$$

This is greater than $v/(v + nw)$ if and only if $nw > v + (n-1)w$, which is equivalent to $v < w$. Similarly, it is less than $v/(v + nw)$ if $v > w$, and equal if $v = w$. So it is better to draw first if $v < w$, and draw second if $v > w$ (and it does not matter if $v = w$). Interestingly, the optimal choice does not depend on n. Note that this result is correct in the simple case $n = 1$, in the case $v = w$ (which reduces to the previous part), and in the extreme case where v is much, much larger than w.

66. A fair die is rolled repeatedly, until the running total is at least 100 (at which point the rolling stops). Find the most likely value of the final running total (i.e., the value of the running total at the first time when it is at least 100).

Hint: Consider the possibilities for what the running total is just before the last roll.

Solution: Just before the last roll, the running total must be one of the six values $94, 95, \ldots, 99$. So the final running total must be one of the six $100, 101, \ldots, 105$. Let T_j be the event that the final running total is $100 + j$, for $j = 0, 1, \ldots, 5$. Let B_i be the event that the running total just before the last roll is $100 - i$, for $i = 1, \ldots, 6$.

By LOTP,

$$P(T_j) = \sum_{i=1}^{6} P(T_j|B_i)P(B_i).$$

Note that, for example, $P(T_1|B_6) = 0$ since we can't go in one step from a running total of 94 to a running total of 101. Then

$$P(T_0) = \sum_{i=1}^{6} \frac{1}{6} P(B_i), \ P(T_1) = \sum_{i=1}^{5} \frac{1}{6} P(B_i),$$

and, in general,

$$P(T_j) = \sum_{i=1}^{6-j} \frac{1}{6} P(B_i).$$

It follows, without even having to compute the $P(B_i)$, that $P(T_0)$ is the largest of the $P(T_j)$'s, so the most likely final running total is 100.

67. Homer has a box of donuts, which currently contains exactly c chocolate, g glazed, and j jelly donuts. Homer eats donuts one after another, each time choosing uniformly at random from the remaining donuts.

(a) Find the probability that the last donut remaining in the box is a chocolate donut.

(b) Find the probability of the following event: glazed is the first type of donut that Homer runs out of, and then jelly is the second type of donut that he runs out of.

Hint: Consider the last donut remaining, and the last donut that is either glazed or jelly.

Solution:

(a) Let C be the event that the last donut remaining is a chocolate donut. By symmetry, the last donut remaining is equally likely to be any of the donuts. So

$$P(C) = \frac{c}{g+j+c}.$$

(b) The event of interest occurs if and only if both of the following two events occur: (i) the last donut is chocolate and (ii) the last donut that is either glazed or jelly is jelly. Let J be the event that the last donut that is either glazed or jelly is jelly. We have

$$P(J|C) = \frac{j}{g+j}$$

since the last donut that is either glazed or jelly is equally likely to be any of the non-chocolate donuts, and the positions of the chocolate donuts are irrelevant when considering the question of whether glazed donuts run out before jelly donuts. So the desired probability is

$$P(C \cap J) = P(C)P(J|C) = \frac{c}{g+j+c}\frac{j}{g+j}.$$

68. Let D be the event that a person develops a certain disease, and C be the event that the person was exposed to a certain substance (e.g., D may correspond to lung cancer and C may correspond to smoking cigarettes). We are interested in whether exposure to the substance is related to developing the disease (and if so, how they are related).

The *odds ratio* is a very widely used measure in epidemiology of the association between disease and exposure, defined as

$$\mathrm{OR} = \frac{\mathrm{odds}(D|C)}{\mathrm{odds}(D|C^c)},$$

where conditional odds are defined analogously to unconditional odds: $\mathrm{odds}(A|B) = \frac{P(A|B)}{P(A^c|B)}$. The *relative risk* of the disease for someone exposed to the substance, another widely used measure, is

$$\mathrm{RR} = \frac{P(D|C)}{P(D|C^c)}.$$

The relative risk is especially easy to interpret, e.g., $\mathrm{RR} = 2$ says that someone exposed to the substance is twice as likely to develop the disease as someone who isn't exposed (though this does not necessarily mean that the substance *causes* the increased chance of getting the disease, nor is there necessarily a causal interpretation for the odds ratio).

(a) Show that if the disease is rare, both for exposed people and for unexposed people, then the relative risk is approximately equal to the odds ratio.

(b) Let p_{ij} for $i = 0, 1$ and $j = 0, 1$ be the probabilities in the following 2×2 table.

	D	D^c
C	p_{11}	p_{10}
C^c	p_{01}	p_{00}

For example, $p_{10} = P(C, D^c)$. Show that the odds ratio can be expressed as a *cross-product ratio*, in the sense that

$$\text{OR} = \frac{p_{11}p_{00}}{p_{10}p_{01}}.$$

(c) Show that the odds ratio has the neat symmetry property that the roles of C and D can be swapped without changing the value:

$$\text{OR} = \frac{\text{odds}(C|D)}{\text{odds}(C|D^c)}.$$

This property is one of the main reasons why the odds ratio is so widely used, since it turns out that it allows the odds ratio to be estimated in a wide variety of problems where relative risk would be hard to estimate well.

Solution:

(a) The odds ratio is related to the relative risk by

$$\text{OR} = \frac{P(D|C)/P(D^c|C)}{P(D|C^c)/P(D^c|C^c)} = \text{RR} \cdot \frac{P(D^c|C^c)}{P(D^c|C)}.$$

So $\text{OR} \approx \text{RR}$ if both $P(D^c|C^c)$ and $P(D^c|C)$ are close to 1.

(b) By definition of conditional probability,

$$\text{OR} = \frac{P(D|C)P(D^c|C^c)}{P(D|C^c)P(D^c|C)} = \frac{P(D,C)P(D^c,C^c)}{P(D,C^c)P(D^c,C)} = \frac{p_{11}p_{00}}{p_{10}p_{01}}.$$

(c) We have

$$\frac{\text{odds}(C|D)}{\text{odds}(C|D^c)} = \frac{P(C|D)P(C^c|D^c)}{P(C|D^c)P(C^c|D)} = \frac{P(C,D)P(C^c,D^c)}{P(C,D^c)P(C^c,D)} = \text{OR}.$$

69. A researcher wants to estimate the percentage of people in some population who have used illegal drugs, by conducting a survey. Concerned that a lot of people would lie when asked a sensitive question like "Have you ever used illegal drugs?", the researcher uses a method known as *randomized response*. A hat is filled with slips of paper, each of which says either "I have used illegal drugs" or "I have not used illegal drugs". Let p be the proportion of slips of paper that say "I have used illegal drugs" (p is chosen by the researcher in advance).

Each participant chooses a random slip of paper from the hat and answers (truthfully) "yes" or "no" to whether the statement on that slip is true. The slip is then returned to the hat. The researcher does not know which type of slip the participant had. Let y be the probability that a participant will say "yes", and d be the probability that a participant has used illegal drugs.

(a) Find y, in terms of d and p.

(b) What would be the worst possible choice of p that the researcher could make in designing the survey? Explain.

(c) Now consider the following alternative system. Suppose that proportion p of the slips of paper say "I have used illegal drugs", but that now the remaining $1 - p$ say "I was

born in winter" rather than "I have not used illegal drugs". Assume that $1/4$ of people are born in winter, and that a person's season of birth is independent of whether they have used illegal drugs. Find d, in terms of y and p.

Solution:

(a) Let A be the event that the participant will draw a "I have used illegal drugs" slip, and Y be the event that there is a "yes" response. By the law of total probability,

$$P(Y) = P(Y|A)P(A) + P(Y|A^c)P(A^c),$$

so

$$y = dp + (1-d)(1-p).$$

(b) The worst possible choice is $p = 1/2$, since then the survey gives no information about d, which is the main quantity of interest. Mathematically, this can be seen by trying to solve the equation from (a) for d in terms of y and p; this gives $d = (y + p - 1)/(2p - 1)$ for $p \neq 1/2$, but would involve dividing by 0 for $p = 1/2$. Intuitively, this makes sense since the $p = 1/2$ case is like dealing with someone who tells the truth half the time and lies half the time (someone who always tells the truth or always lies is much more informative!).

(c) LOTP gives

$$y = dp + \frac{1-p}{4},$$

so $d = (y - 1/4)/p + 1/4$ (for $p \neq 0$; if $p = 0$, then the survey only asks about winter births and so gives no information about d).

70. At the beginning of the play *Rosencrantz and Guildenstern are Dead* by Tom Stoppard, Guildenstern is spinning coins and Rosencrantz is betting on the outcome for each. The coins have been landing Heads over and over again, prompting the following remark:

> *Guildenstern*: A weaker man might be moved to re-examine his faith, if in nothing else at least in the law of probability.

The coin spins have resulted in Heads 92 times in a row.

(a) Fred and his friend are watching the play. Upon seeing the events described above, they have the following conversation:

> *Fred*: That outcome would be incredibly unlikely with fair coins. They must be using trick coins (maybe with double-headed coins), or the experiment must have been rigged somehow (maybe with magnets).
>
> *Fred's friend:* It's true that the string HH...H of length 92 is very unlikely; the chance is $1/2^{92} \approx 2 \times 10^{-28}$ with fair coins. But *any* other specific string of H's and T's with length 92 has *exactly* the same probability! The reason the outcome seems extremely unlikely is that the number of possible outcomes grows exponentially as the number of spins grows, so *any* outcome would seem extremely unlikely. You could just as well have made the same argument even without looking at the results of their experiment, which means you really don't have evidence against the coins being fair.

Discuss these comments, to help Fred and his friend resolve their debate.

(b) Suppose there are only two possibilities: either the coins are all fair (and spun fairly), or double-headed coins are being used (in which case the probability of Heads is 1). Let p be the prior probability that the coins are fair. Find the posterior probability that the coins are fair, given that they landed Heads in 92 out of 92 trials.

(c) Continuing from (b), for which values of p is the posterior probability that the coins are fair greater than 0.5? For which values of p is it less than 0.05?

Solution:

(a) Fred is correct that the outcome HH...H is extremely suspicious, but Fred's friend is correct that with a fair coin, any specific string of H's and T's of length 92 has the same probability. To reconcile these statements, note that there is only 1 string of length 92 with 92 H's, whereas there are a vast number of strings with about half H's and about half T's. For example, there are $\binom{92}{46} \approx 4.1 \times 10^{26}$ strings of length 92 with 46 H's and 46 T's. It is enormously more likely to have 46 H's and 46 T's (with a fair coin) than 92 H's, even though any specific string with 46 H's and 46 T's has the same probability as HH...H. Furthermore, there are an even vaster number of possibilities where the number of Heads is roughly equal to the number of Tails.

We should check, however, that Fred is not just doing *data snooping* (also known as *data fishing*) after observing the data. By trying a lot of different calculations with the data that are not prescribed in advance, it is easy to find unusual patterns in a data set, but it is highly dangerous and misleading to present such results without reporting how many calculations were tried, and when and how it was decided which calculations to do. In this case though, looking at the number of Heads in a sequence of coin tosses is a very simple, common, natural summary of the data, and there is no reason to think that Fred was fishing for something suspicious.

There could also be a *selection bias* at play: maybe millions of such experiments are performed, by people all over the world and at different points in time, and Tom Stoppard wrote about this particular incident *because* the outcome was so interesting. Likewise, it's not surprising to read in the news that someone won the lottery, when millions of people are entering the lottery; the news article reports on the person who won the lottery *because* they won. But having no Heads in 92 trials is so staggeringly unlikely (for a fair coin) that there would be good reason to be suspicious even if several billion people were performing similar experiments and only the most extreme outcome was selected to include in the play.

(b) Let F be the event that the coins are fair and A be the event that they land Heads in 92 out of 92 trials. By Bayes' rule and LOTP,

$$P(F|A) = \frac{P(A|F)P(F)}{P(A)} = \frac{2^{-92} \cdot p}{2^{-92} \cdot p + 1 - p} = \frac{p}{p + 2^{92}(1 - p)}.$$

(c) We have that

$$\frac{p}{p + 2^{92}(1 - p)} > \frac{1}{2}$$

is equivalent to

$$p > \frac{2^{92}}{2^{92} + 1}.$$

This says that p must be incredibly close to 1, since

$$\frac{2^{92}}{2^{92} + 1} = 1 - \frac{1}{2^{92} + 1} \approx 1 - 2 \times 10^{-28}.$$

On the other hand,

$$\frac{p}{p + 2^{92}(1 - p)} < 0.05$$

is equivalent to

$$p < \frac{2^{92}}{2^{92} + 19}.$$

Unless p is incredibly close to 1, the above inequality will hold.

71. There are n types of toys, which you are collecting one by one. Each time you buy a toy, it is randomly determined which type it has, with equal probabilities. Let p_{ij} be the probability that just after you have bought your ith toy, you have exactly j toy types in your collection, for $i \geq 1$ and $0 \leq j \leq n$. (This problem is in the setting of the *coupon collector* problem, a famous problem which we study in Example 4.3.11.)

(a) Find a recursive equation expressing p_{ij} in terms of $p_{i-1,j}$ and $p_{i-1,j-1}$, for $i \geq 2$ and $1 \leq j \leq n$.

(b) Describe how the recursion from (a) can be used to calculate p_{ij}.

Solution:

(a) There are two ways to have exactly j toy types just after buying your ith toy: either you have exactly $j-1$ toy types just after buying your $i-1$st toy and then the ith toy you buy is of a type you don't already have, or you already have exactly j toy types just after buying your $i-1$st toy and then the ith toy you buy is of a type you do already have. Conditioning on how many toy types you have just after buying your $i-1$st toy,

$$p_{ij} = p_{i-1,j-1}\frac{n-j+1}{n} + p_{i-1,j}\frac{j}{n}.$$

(b) First note that $p_{11} = 1$, and $p_{ij} = 0$ for $j = 0$ or $j > i$. Now suppose that we have computed $p_{i-1,1}, p_{i-1,2}, \ldots, p_{i-1,i-1}$ for some $i \geq 2$. Then we can compute $p_{i,1}, p_{i,2}, \ldots, p_{i,i}$ using the recursion from (a). We can then compute $p_{i+1,1}, p_{i+1,2}, \ldots, p_{i+1,i+1}$ using the recursion from (a), and so on. By induction, it follows that we can obtain any desired p_{ij} recursively by this method.

72. *A/B testing* is a form of randomized experiment that is used by many companies to learn about how customers will react to different treatments. For example, a company may want to see how users will respond to a new feature on their website (compared with how users respond to the current version of the website) or compare two different advertisements.

As the name suggests, two different treatments, Treatment A and Treatment B, are being studied. Users arrive one by one, and upon arrival are randomly assigned to one of the two treatments. The trial for each user is classified as "success" (e.g., the user made a purchase) or "failure". The probability that the nth user receives Treatment A is allowed to depend on the outcomes for the previous users. This set-up is known as a *two-armed bandit*.

Many algorithms for how to randomize the treatment assignments have been studied. Here is an especially simple (but fickle) algorithm, called a *stay-with-a-winner* procedure:

 (i) Randomly assign the first user to Treatment A or Treatment B, with equal probabilities.

(ii) If the trial for the nth user is a success, stay with the same treatment for the $(n+1)$st user; otherwise, switch to the other treatment for the $(n+1)$st user.

Let a be the probability of success for Treatment A, and b be the probability of success for Treatment B. Assume that $a \neq b$, but that a and b are unknown (which is why the test is needed). Let p_n be the probability of success on the nth trial and a_n be the probability that Treatment A is assigned on the nth trial (using the above algorithm).

(a) Show that

$$p_n = (a-b)a_n + b,$$
$$a_{n+1} = (a+b-1)a_n + 1 - b.$$

(b) Use the results from (a) to show that p_{n+1} satisfies the following recursive equation:

$$p_{n+1} = (a+b-1)p_n + a + b - 2ab.$$

(c) Use the result from (b) to find the long-run probability of success for this algorithm, $\lim_{n\to\infty} p_n$, assuming that this limit exists.

Solution:

(a) Conditioning on which treatment is applied to the nth user, we have

$$p_n = a_n a + (1 - a_n)b = (a - b)a_n + b.$$

Treatment A will be applied to the $(n + 1)$st user if and only if (i) Treatment A is successfully applied to the nth user or (ii) Treatment B is unsuccessfully applied to the nth user. Again conditioning on which treatment is applied to the nth user, we have

$$a_{n+1} = aa_n + (1 - b)(1 - a_n) = (a + b - 1)a_n + 1 - b.$$

(b) Plugging the recursion for a_{n+1} into the recursion

$$p_{n+1} = (a - b)a_{n+1} + b,$$

we have

$$p_{n+1} = (a - b)((a + b - 1)a_n + 1 - b) + b.$$

Replacing a_n by $(p_n - b)/(a - b)$, we have

$$\begin{aligned} p_{n+1} &= (a - b)\left((a + b - 1)\frac{p_n - b}{a - b} + 1 - b\right) + b \\ &= (a + b - 1)(p_n - b) + (a - b)(1 - b) + b \\ &= (a + b - 1)p_n + a + b - 2ab. \end{aligned}$$

(c) Let $p = \lim_{n\to\infty} p_n$ (assuming that this exists). Taking the limit as $n \to \infty$ on both sides of

$$p_{n+1} = (a + b - 1)p_n + a + b - 2ab,$$

we have

$$p = (a + b - 1)p + a + b - 2ab.$$

Therefore,

$$p = \frac{a + b - 2ab}{2 - a - b}.$$

73. In humans (and many other organisms), genes come in pairs. A certain gene comes in two types (*alleles*): type a and type A. The *genotype* of a person for that gene is the types of the two genes in the pair: AA, Aa, or aa (aA is equivalent to Aa). Assume that the *Hardy-Weinberg law* applies here, which means that the frequencies of AA, Aa, aa in the population are $p^2, 2p(1 - p), (1 - p)^2$ respectively, for some p with $0 < p < 1$.

When a woman and a man have a child, the child's gene pair consists of one gene contributed by each parent. Suppose that the mother is equally likely to contribute either of the two genes in her gene pair, and likewise for the father, independently. Also suppose that the genotypes of the parents are independent of each other (with probabilities given by the Hardy-Weinberg law).

(a) Find the probabilities of each possible genotype (AA, Aa, aa) for a child of two random parents. Explain what this says about stability of the Hardy-Weinberg law from one generation to the next.

Hint: Condition on the genotypes of the parents.

(b) A person of type AA or aa is called *homozygous* (for the gene under consideration), and a person of type Aa is called *heterozygous* (for that gene). Find the probability

that a child is homozygous, given that both parents are homozygous. Also, find the probability that a child is heterozygous, given that both parents are heterozygous.

(c) Suppose that having genotype aa results in a distinctive physical characteristic, so it is easy to tell by looking at someone whether or not they have that genotype. A mother and father, neither of whom are of type aa, have a child. The child is also not of type aa. Given this information, find the probability that the child is heterozygous.

Hint: Use the definition of conditional probability. Then expand both the numerator and the denominator using LOTP, conditioning on the genotypes of the parents.

Solution:

(a) Let M_{AA}, F_{AA}, C_{AA} be the events that the mother, father, and child (respectively) have genotype AA, and likewise define M_{Aa} etc. Then

$$
\begin{aligned}
P(C_{AA}) &= P(C_{AA}|M_{AA}, F_{AA})P(M_{AA}, F_{AA}) + P(C_{AA}|M_{AA}, F_{Aa})P(M_{AA}, F_{Aa}) \\
&\quad + P(C_{AA}|M_{Aa}, F_{AA})P(M_{Aa}, F_{AA}) + P(C_{AA}|M_{Aa}, F_{Aa})P(M_{Aa}, F_{Aa}) \\
&= p^4 + \frac{1}{2}p^2(2p(1-p)) + \frac{1}{2}p^2(2p(1-p)) + \frac{1}{4}(2p(1-p))^2 \\
&= p^4 + 2p^3 - 2p^4 + p^2(1 - 2p + p^2) \\
&= p^2.
\end{aligned}
$$

Let $q = 1 - p$. Swapping A's and a's and swapping p's and q's in the above calculation,

$$
P(C_{aa}) = q^2 = (1-p)^2.
$$

It follows that

$$
P(C_{Aa}) = 1 - P(C_{AA}) - P(C_{aa}) = 1 - p^2 - (1-p)^2 = 2p - 2p^2 = 2p(1-p).
$$

So the Hardy-Weinberg law is stable, in the sense that the probabilities for the various genotypes are preserved from one generation to the next.

(b) Let H be the event that both parents are homozygous. Then

$$
P(\text{child homozygous}|H) = P(C_{AA}|H) + P(C_{aa}|H).
$$

To find $P(C_{AA}|H)$, we can do further conditioning on the exact genotypes of the parents:

$$
\begin{aligned}
P(C_{AA}|H) &= P(C_{AA}|M_{AA}, F_{AA})P(M_{AA}, F_{AA}|H) + P(C_{AA}|M_{AA}, F_{aa})P(M_{AA}, F_{aa}|H) \\
&\quad + P(C_{AA}|M_{aa}, F_{AA})P(M_{aa}, F_{AA}|H) + P(C_{AA}|M_{aa}, F_{aa})P(M_{aa}, F_{aa}|H) \\
&= P(M_{AA}, F_{AA}|H),
\end{aligned}
$$

since all the terms except the first are zero (because, for example, an AA mother and aa father can't produce an AA child). This result can also be seen directly by noting that, given H, the child being AA is equivalent to both parents being AA. Next, we have

$$
P(M_{AA}, F_{AA}|H) = \frac{P(M_{AA}, F_{AA}, H)}{P(H)} = \frac{P(M_{AA}, F_{AA})}{P(H)} = \frac{p^4}{(p^2 + (1-p)^2)^2}.
$$

By symmetry, we have

$$
P(C_{aa}|H) = P(M_{aa}, F_{aa}|H) = \frac{(1-p)^4}{((1-p)^2 + p^2)^2}.
$$

Hence,

$$
P(\text{child homozygous}|H) = \frac{p^4 + (1-p)^4}{(p^2 + (1-p)^2)^2}.
$$

Lastly,

$$P(\text{child heterozygous}|\text{both parents heterozygous}) = \frac{1}{2}$$

since if both parents are Aa, then a child will be heterozygous if and only if they receive the mother's A and the father's a or vice versa.

(c) We wish to find $P(C_{Aa}|C_{aa}^c \cap M_{aa}^c \cap F_{aa}^c)$. Let

$$G = C_{Aa} \cap M_{aa}^c \cap F_{aa}^c \text{ and } H = C_{aa}^c \cap M_{aa}^c \cap F_{aa}^c.$$

Then

$$P(C_{Aa}|C_{aa}^c \cap M_{aa}^c \cap F_{aa}^c) = \frac{P(G)}{P(H)},$$

and

$$\begin{aligned}
P(G) &= P(G|M_{AA}, F_{AA})P(M_{AA}, F_{AA}) + P(G|M_{AA}, F_{Aa})P(M_{AA}, F_{Aa}) \\
&\quad + P(G|M_{Aa}, F_{AA})P(M_{Aa}, F_{AA}) + P(G|M_{Aa}, F_{Aa})P(M_{Aa}, F_{Aa}) \\
&= 0 + \frac{1}{2}p^2(2p(1-p)) + \frac{1}{2}p^2(2p(1-p)) + \frac{1}{2}(2p(1-p))^2 \\
&= 2p^2(1-p),
\end{aligned}$$

while

$$\begin{aligned}
P(H) &= P(H|M_{AA}, F_{AA})P(M_{AA}, F_{AA}) + P(H|M_{AA}, F_{Aa})P(M_{AA}, F_{Aa}) \\
&\quad + P(H|M_{Aa}, F_{AA})P(M_{Aa}, F_{AA}) + P(H|M_{Aa}, F_{Aa})P(M_{Aa}, F_{Aa}) \\
&= p^4 + p^2(2p(1-p)) + p^2(2p(1-p)) + \frac{3}{4}(2p(1-p))^2 \\
&= p^2(3-2p).
\end{aligned}$$

Thus,

$$P(C_{Aa}|C_{aa}^c \cap M_{aa}^c \cap F_{aa}^c) = \frac{2p^2(1-p)}{p^2(3-2p)} = \frac{2-2p}{3-2p}.$$

74. A standard deck of cards will be shuffled and then the cards will be turned over one at a time until the first ace is revealed. Let B be the event that the *next* card in the deck will also be an ace.

(a) Intuitively, how do you think $P(B)$ compares in size with $1/13$ (the overall proportion of aces in a deck of cards)? Explain your intuition. (Give an intuitive discussion rather than a mathematical calculation; the goal here is to describe your intuition explicitly.)

(b) Let C_j be the event that the first ace is at position j in the deck. Find $P(B|C_j)$ in terms of j, fully simplified.

(c) Using the law of total probability, find an expression for $P(B)$ as a sum. (The sum can be left unsimplified, but it should be something that could easily be computed in software such as R that can calculate sums.)

(d) Find a fully simplified expression for $P(B)$ using a symmetry argument.

Hint: If you were deciding whether to bet on the next card after the first ace being an ace or to bet on the last card in the deck being an ace, would you have a preference?

Solution:

(a) Intuitively, it may seem that $P(B) < 1/13$ since we know an ace has been depleted from the deck and don't know anything about how many (if any) non-aces were depleted. On the other hand, it clearly matters how many cards are needed to reach the first ace (in the extreme case where the first ace is at the 49th card, we *know* the next 3 cards

are aces). But we are not given information about the number of cards needed to reach the first ace. Of course, we can condition on it, which brings us to the next two parts.

(b) There are $52 - j$ remaining cards in the deck, of which 3 are aces. By symmetry, the next card is equally likely to be any of the remaining cards. So

$$P(B|C_j) = \frac{3}{52 - j}.$$

(c) LOTP yields

$$P(B) = \sum_{j=1}^{49} P(B|C_j)P(C_j) = \sum_{j=1}^{49} \frac{\binom{48}{j-1}}{\binom{52}{j-1}} \cdot \frac{4}{52 - j + 1} \cdot \frac{3}{52 - j},$$

where the ratio of binomial coefficients is used to get the probability of the deck starting out with $j - 1$ non-aces. Computing the sum was not required for this part, but to do so in R we can type

```
j <- 1:49
sum(choose(48,j-1)/choose(52,j-1)*(4/(52-j+1))*(3/(52-j)))
```

(d) At any stage, by symmetry all unrevealed cards are completely interchangeable (in statistics, they are said to be *exchangeable*). Let L be the event that the last card in the deck is an ace. By symmetry, $P(B|C_j) = P(L|C_j)$ for any j. So $P(B) = P(L)$ unconditionally too (by LOTP). Thus,

$$P(B) = P(L) = \frac{1}{13}.$$

3

Random variables and their distributions

PMFs and CDFs

1. People are arriving at a party one at a time. While waiting for more people to arrive they entertain themselves by comparing their birthdays. Let X be the number of people needed to obtain a birthday match, i.e., before person X arrives no two people have the same birthday, but when person X arrives there is a match. Find the PMF of X.

    Solution: We will make the usual assumptions as in the birthday problem (e.g., exclude February 29). The support of X is $\{2, 3, \ldots, 366\}$ since if there are 365 people there and no match, then every day of the year is accounted for and the 366th person will create a match. Let's start with a couple simple cases and then generalize:
    $$P(X = 2) = \frac{1}{365},$$
    since the second person has a 1/365 chance of having the same birthday as the first,
    $$P(X = 3) = \frac{364}{365} \cdot \frac{2}{365},$$
    since $X = 3$ means that the second person didn't match the first but the third person matched one of the first two. In general, for $2 \le k \le 366$ we have
    $$\begin{aligned} P(X = k) &= P(X > k - 1 \text{ and } X = k) \\ &= \frac{365 \cdot 364 \cdots (365 - k + 2)}{365^{k-1}} \cdot \frac{k - 1}{365} \\ &= \frac{(k - 1) \cdot 364 \cdot 363 \cdots (365 - k + 2)}{365^{k-1}}. \end{aligned}$$

2. (a) Independent Bernoulli trials are performed, with probability 1/2 of success, until there has been at least one success. Find the PMF of the number of trials performed.

    (b) Independent Bernoulli trials are performed, with probability 1/2 of success, until there has been at least one success and at least one failure. Find the PMF of the number of trials performed.

    Solution:

    (a) Let X be the number of trials (including the success). Then $X = n$ says that there were $n - 1$ failures followed by a success, so
    $$P(X = n) = \left(\frac{1}{2}\right)^{n-1} \frac{1}{2} = \left(\frac{1}{2}\right)^n,$$
    for $n = 1, 2, \ldots$ (This is a *First Success* distribution and is explored in Chapter 4.)

    (b) Let Y be the number of trials performed. The support of Y is $\{2, 3, \ldots\}$. Let A be the event that the first trial is a success. The PMF of Y is
    $$P(Y = n) = P(Y = n|A)P(A) + P(Y = n|A^c)P(A^c) = 2 \left(\frac{1}{2}\right)^{n-1} \frac{1}{2} = \left(\frac{1}{2}\right)^{n-1},$$
    for $n = 2, 3, \ldots.$

3. Let X be an r.v. with CDF F, and $Y = \mu + \sigma X$, where μ and σ are real numbers with $\sigma > 0$. (Then Y is called a *location-scale transformation* of X; we will encounter this concept many times in Chapter 5 and beyond.) Find the CDF of Y, in terms of F.

Solution: The CDF of Y is

$$P(Y \leq y) = P(\mu + \sigma X \leq y) = P\left(X \leq \frac{y - \mu}{\sigma}\right) = F\left(\frac{y - \mu}{\sigma}\right).$$

4. Let n be a positive integer and

$$F(x) = \frac{\lfloor x \rfloor}{n}$$

for $0 \leq x \leq n$, $F(x) = 0$ for $x < 0$, and $F(x) = 1$ for $x > n$, where $\lfloor x \rfloor$ is the greatest integer less than or equal to x. Show that F is a CDF, and find the PMF that it corresponds to.

Solution: The function F is increasing since the floor function $\lfloor x \rfloor$ is increasing in x (if $x_1 \leq x_2$, then any integer less than or equal to x_1 is also less than or equal to x_2, so $\lfloor x_1 \rfloor \leq \lfloor x_2 \rfloor$) and is 0 to the left of 0 and 1 to the right of n. And F is right-continuous since the floor function is right-continuous. Also, $F(x) \to 0$ as $x \to -\infty$ and $F(x) \to 1$ as $x \to \infty$. So F is a valid CDF.

The support of this discrete distribution is $\{1, 2, \ldots, n\}$ since those values are where the jumps in the CDF are. The jumps are all of the same height, $1/n$, so the corresponding PMF is that of an r.v. X which takes values $1, 2, \ldots, n$ with equal probabilities:

$$P(X = j) = \frac{1}{n},$$

for $j = 1, 2, \ldots, n$.

5. (a) Show that $p(n) = \left(\frac{1}{2}\right)^{n+1}$ for $n = 0, 1, 2, \ldots$ is a valid PMF for a discrete r.v.

(b) Find the CDF of a random variable with the PMF from (a).

Solution:

(a) Clearly $p(n) \geq 0$, so we just need to check that the values sum to 1. By the formula for the sum of a geometric series,

$$\sum_{n=0}^{\infty} \left(\frac{1}{2}\right)^{n+1} = \frac{\frac{1}{2}}{1 - \frac{1}{2}} = 1.$$

(b) Let X have the distribution from (a). Then $P(X \leq x) = 0$ for $x < 0$. For $x \geq 0$,

$$P(X \leq x) = P(X \in \{0, 1, \ldots, \lfloor x \rfloor\}) = \sum_{n=0}^{\lfloor x \rfloor} \left(\frac{1}{2}\right)^{n+1} = 1 - \left(\frac{1}{2}\right)^{\lfloor x \rfloor + 1},$$

by the formula for the sum of a finite geometric series. (See the math appendix for information about the floor function $\lfloor x \rfloor$.)

6. ⓢ *Benford's law* states that in a very large variety of real-life data sets, the first digit approximately follows a particular distribution with about a 30% chance of a 1, an 18% chance of a 2, and in general

$$P(D = j) = \log_{10}\left(\frac{j + 1}{j}\right), \text{ for } j \in \{1, 2, 3, \ldots, 9\},$$

where D is the first digit of a randomly chosen element. Check that this is a valid PMF (using properties of logs, not with a calculator).

Solution: The function $P(D = j)$ is nonnegative and the sum over all values is

$$\sum_{j=1}^{9} \log_{10} \frac{j+1}{j} = \sum_{j=1}^{9} (\log_{10}(j+1) - \log_{10}(j)).$$

All terms cancel except $\log_{10} 10 - \log_{10} 1 = 1$ (this is a *telescoping series*). Since the values add to 1 and are nonnegative, $P(D = j)$ is a PMF.

7. Bob is playing a video game that has 7 levels. He starts at level 1, and has probability p_1 of reaching level 2. In general, given that he reaches level j, he has probability p_j of reaching level $j + 1$, for $1 \leq j \leq 6$. Let X be the highest level that he reaches. Find the PMF of X (in terms of $p_1, \ldots, p_6$).

   Solution: We have $P(X = 1) = 1 - p_1$, $P(X = j) = p_1 p_2 \ldots p_{j-1}(1 - p_j)$ for $2 \leq j \leq 6$, and $P(X = 7) = p_1 p_2 \ldots p_6$.

8. There are 100 prizes, with one worth \$1, one worth \$2, ..., and one worth \$100. There are 100 boxes, each of which contains one of the prizes. You get 5 prizes by picking random boxes one at a time, *without replacement*. Find the PMF of how much your most valuable prize is worth (as a simple expression in terms of binomial coefficients).

   Solution: Let X be the value of your most valuable prize. The support is $5, 6, \ldots, 100$ since the worst case is getting the \$1, ..., \$5 prizes. The event $X = k$ says that you got the \$$k$ prize and 4 less valuable prizes. So the PMF of X is

   $$P(X = k) = \frac{\binom{k-1}{4}}{\binom{100}{5}}, \text{ for } k = 5, 6, \ldots, 100.$$

9. Let F_1 and F_2 be CDFs, $0 < p < 1$, and $F(x) = pF_1(x) + (1 - p)F_2(x)$ for all x.

   (a) Show directly that F has the properties of a valid CDF (see Theorem 3.6.3). The distribution defined by F is called a *mixture* of the distributions defined by F_1 and F_2.

   (b) Consider creating an r.v. in the following way. Flip a coin with probability p of Heads. If the coin lands Heads, generate an r.v. according to F_1; if the coin lands Tails, generate an r.v. according to F_2. Show that the r.v. obtained in this way has CDF F.

   Solution:

   (a) The function F is increasing since F_1 and F_2 are increasing and both p and $1 - p$ are positive. It is right-continuous since F_1 and F_2 are right-continuous: if x converges to x_0, approaching from the right, then

   $$F(x) = pF_1(x) + (1 - p)F_2(x) \to pF_1(x_0) + (1 - p)F_2(x_0) = F(x_0).$$

   The limits are as desired since $F(x) \to p \cdot 0 + (1 - p) \cdot 0 = 0$ as $x \to -\infty$ and $F(x) \to p + 1 - p = 1$ as $x \to \infty$. So F is a valid CDF.

   (b) Let X be an r.v. obtained in this way, and let H be the event that the coin lands Heads. Then

   $$P(X \leq x) = P(X \leq x|H)P(H) + P(X \leq x|H^c)P(H^c) = pF_1(x) + (1 - p)F_2(x).$$

10. (a) Is there a discrete distribution with support $1, 2, 3, \ldots$, such that the value of the PMF at n is proportional to $1/n$?

    Hint: See the math appendix for a review of some facts about series.

    (b) Is there a discrete distribution with support $1, 2, 3, \ldots$, such that the value of the PMF at n is proportional to $1/n^2$?

Solution:

(a) No, since the harmonic series $\sum_{n=1}^{\infty} \frac{1}{n}$ diverges, so it is not possible to find a constant c such that $\sum_{n=1}^{\infty} \frac{c}{n}$ converges to 1.

(b) Yes, since the series $\sum_{n=1}^{\infty} \frac{1}{n^2}$ converges. Let a be the sum of the series (it turns out that $a = \pi^2/6$) and $b = 1/a$. Then $p(n) = b/n^2$ for $n = 1, 2, \ldots$ is a valid PMF.

11. Ⓢ Let X be an r.v. whose possible values are $0, 1, 2, \ldots$, with CDF F. In some countries, rather than using a CDF, the convention is to use the function G defined by $G(x) = P(X < x)$ to specify a distribution. Find a way to convert from F to G, i.e., if F is a known function, show how to obtain $G(x)$ for all real x.

Solution: Write

$$G(x) = P(X \le x) - P(X = x) = F(x) - P(X = x).$$

If x is not a nonnegative integer, then $P(X = x) = 0$ so $G(x) = F(x)$. For x a nonnegative integer,
$$P(X = x) = F(x) - F(x - 1/2)$$

since the PMF corresponds to the lengths of the jumps in the CDF. (The $1/2$ was chosen for concreteness; we also have $F(x - 1/2) = F(x - a)$ for any $a \in (0, 1]$.) Thus,

$$G(x) = \begin{cases} F(x) \text{ if } x \notin \{0, 1, 2, \ldots\} \\ F(x - 1/2) \text{ if } x \in \{0, 1, 2, \ldots\}. \end{cases}$$

More compactly, we can also write $G(x) = \lim_{t \to x^-} F(t)$, where the $-$ denotes taking a limit from the left (recall that F is right-continuous), and $G(x) = F(\lceil x \rceil - 1)$, where $\lceil x \rceil$ is the ceiling of x (the smallest integer greater than or equal to x).

12. (a) Give an example of r.v.s X and Y such that $F_X(x) \le F_Y(x)$ for all x, where the inequality is strict for some x. Here F_X is the CDF of X and F_Y is the CDF of Y. For the example you gave, sketch the CDFs of both X and Y on the same axes. Then sketch their PMFs on a second set of axes.

(b) In Part (a), you found an example of two different CDFs where the first is less than or equal to the second everywhere. Is it possible to find two different PMFs where the first is less than or equal to the second everywhere? In other words, find discrete r.v.s X and Y such that $P(X = x) \le P(Y = x)$ for all x, where the inequality is strict for some x, or show that it is impossible to find such r.v.s.

Solution:

(a) Consider flipping a coin $n \ge 2$ times, with $p \in (0, 1)$ the probability of Heads. Let $X \sim \text{Bin}(n, p)$ be the number of Heads in the n tosses and $Y \sim \text{Bin}(m, p)$ be the number of Heads in the first m tosses, for some fixed $m < n$. Then $P(X \le x) \le P(Y \le x)$ for all x, since $X \le x$ implies $Y \le x$. The inequality is strict for some x since, for example,

$$P(X \le 0) = P(X = 0) = p^n < p^m = P(Y = 0) = P(Y \le 0)$$

and

$$P(X \le m) < 1 = P(Y \le m).$$

For the case $m = 1, n = 2, p = 0.5$, the PMFs and CDFs are plotted below. For the PMF plot, $\text{Bin}(10, 0.5)$ has solid dots and $\text{Bin}(5, 0.5)$ has open dots. For the CDF plot, $\text{Bin}(10, 0.5)$ has solid lines and $\text{Bin}(5, 0.5)$ has dashed lines.
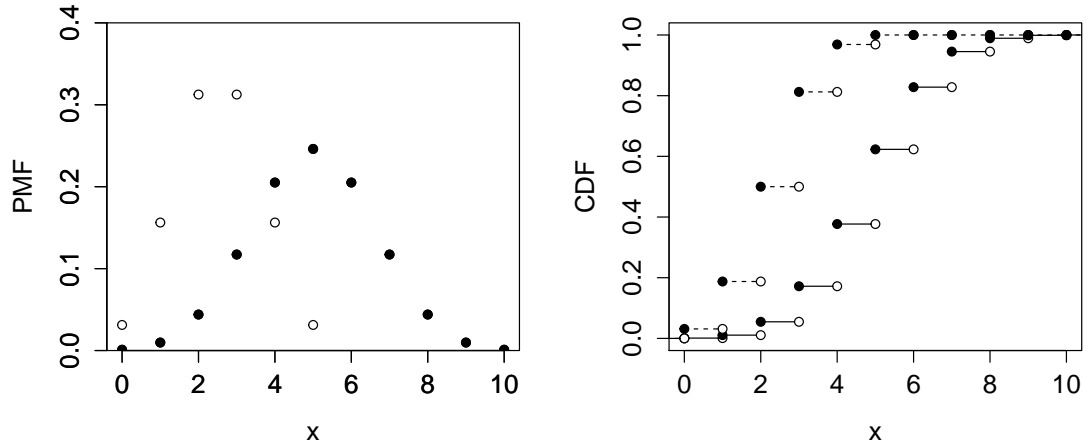
(b) It is not possible to find such X and Y. This is because a PMF must sum to 1. Suppose that $P(X = x) \leq P(Y = x)$ for all x. Then

$$1 = \sum_x P(X = x) \leq \sum_x P(Y = x) = 1,$$

where the sums are over the union of the supports of X and Y. If even one value c exists such that $P(X = c) < P(Y = c)$, then the inequality $\sum_x P(X = x) \leq \sum_x P(Y = x)$ is strict, resulting in the contradiction $1 < 1$. So $P(X = x) = P(Y = x)$ for all x.

13. Let X, Y, Z be discrete r.v.s such that X and Y have the same conditional distribution given Z, i.e., for all a and z we have

$$P(X = a | Z = z) = P(Y = a | Z = z).$$

Show that X and Y have the same distribution (unconditionally, not just when given Z).

Solution: By the law of total probability, conditioning on Z, we have

$$P(X = a) = \sum_z P(X = a | Z = z) P(Z = z).$$

Since X and Y have the same conditional distribution given Z, this becomes

$$\sum_z P(Y = a | Z = z) P(Z = z) = P(Y = a).$$

14. Let X be the number of purchases that Fred will make on the online site for a certain company (in some specified time period). Suppose that the PMF of X is $P(X = k) = e^{-\lambda} \lambda^k / k!$ for $k = 0, 1, 2, \ldots$. This distribution is called the *Poisson distribution* with parameter λ, and it will be studied extensively in later chapters.

(a) Find $P(X \geq 1)$ and $P(X \geq 2)$ without summing infinite series.

(b) Suppose that the company only knows about people who have made at least one purchase on their site (a user sets up an account to make a purchase, but someone who has never made a purchase there doesn't appear in the customer database). If the company computes the number of purchases for everyone in their database, then these data are draws from the *conditional* distribution of the number of purchases, given that

at least one purchase is made. Find the conditional PMF of X given $X \geq 1$. (This conditional distribution is called a *truncated Poisson distribution*.)

Solution:

(a) Taking complements,

$$P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-\lambda},$$
$$P(X \geq 2) = 1 - P(X \leq 1) = 1 - e^{-\lambda} - \lambda e^{-\lambda}.$$

(b) The conditional PMF of X given $X \geq 1$ is

$$P(X = k | X \geq 1) = \frac{P(X = k)}{P(X \geq 1)} = \frac{e^{-\lambda}\lambda^k}{k!(1 - e^{-\lambda})},$$

for $k = 1, 2, \ldots$.

Named distributions

15. Find the CDF of an r.v. $X \sim \text{DUnif}(1, 2, \ldots, n)$.

    Solution: As shown in Exercise 4, the CDF is F defined by

    $$F(x) = \frac{\lfloor x \rfloor}{n}$$

    for $0 \leq x \leq n$, $F(x) = 0$ for $x < 0$, and $F(x) = 1$ for $x > n$, where $\lfloor x \rfloor$ is the greatest integer less than or equal to x.

16. Let $X \sim \text{DUnif}(C)$, and B be a nonempty subset of C. Find the conditional distribution of X, given that X is in B.

    Solution: The conditional PMF of X is

    $$P(X = x | X \in B) = \frac{P(X \in B | X = x)P(X = x)}{P(X \in B)} = \begin{cases} \frac{1}{|B|} & \text{if } x \in B \\ 0 & \text{if } x \notin B, \end{cases}$$

    since $P(X \in B) = |B|/|C|$ and for $x \in B$, $P(X \in B | X = x) = 1, P(X = x) = 1/|C|$. So the conditional distribution of X, given that X is in B, is Discrete Uniform over B. This makes sense intuitively since the values in B were equally likely before we knew that X was in B; just learning that X is in B rules out values not in B but should not result in some values in B being more likely than others.

17. An airline overbooks a flight, selling more tickets for the flight than there are seats on the plane (figuring that it's likely that some people won't show up). The plane has 100 seats, and 110 people have booked the flight. Each person will show up for the flight with probability 0.9, independently. Find the probability that there will be enough seats for everyone who shows up for the flight.

    Solution: Let $X \sim \text{Bin}(110, 0.9)$ be the number of people who show up. Then

    $$P(X \leq 100) = \sum_{k=0}^{100} \binom{110}{k} 0.9^k 0.1^{110-k} \approx 0.671.$$

    (We used `pbinom(100,110,0.9)` in R to compute the sum.)

18. Ⓢ (a) In the World Series of baseball, two teams (call them A and B) play a sequence of games against each other, and the first team to win four games wins the series. Let p be the probability that A wins an individual game, and assume that the games are independent. What is the probability that team A wins the series?

    (b) Give a clear intuitive explanation of whether the answer to (a) depends on whether the teams always play 7 games (and whoever wins the majority wins the series), or the teams stop playing more games as soon as one team has won 4 games (as is actually the case in practice: once the match is decided, the two teams do not keep playing more games).

    Solution:

    (a) Let $q = 1 - p$. First let us do a direct calculation:

    $$
    \begin{aligned}
    P(\text{A wins}) \quad = \quad & P(\text{A wins in 4 games}) + P(\text{A wins in 5 games}) \\
    & + P(\text{A wins in 6 games}) + P(\text{A wins in 7 games}) \\
    = \quad & p^4 + \binom{4}{3}p^4 q + \binom{5}{3}p^4 q^2 + \binom{6}{3}p^4 q^3.
    \end{aligned}
    $$

    To understand how these probabilities are calculated, note for example that

    $$
    \begin{aligned}
    P(\text{A wins in 5}) \quad = \quad & P(\text{A wins 3 out of first 4}) \cdot P(\text{A wins 5th game} | \text{A wins 3 out of first 4}) \\
    = \quad & \binom{4}{3}p^3 q p.
    \end{aligned}
    $$

    (Each of the 4 terms in the expression for $P(\text{A wins})$ can also be found using the PMF of a distribution known as the *Negative Binomial*, which is introduced in Chapter 4.)

    An neater solution is to use the fact (explained in the solution to Part (b)) that we can assume that the teams play all 7 games no matter what. Let X be the number of wins for team A, so that $X \sim \text{Bin}(7, p)$. Then

    $$
    \begin{aligned}
    P(X \geq 4) \quad = \quad & P(X = 4) + P(X = 5) + P(X = 6) + P(X = 7) \\
    = \quad & \binom{7}{4}p^4 q^3 + \binom{7}{5}p^5 q^2 + \binom{7}{6}p^6 q + p^7,
    \end{aligned}
    $$

    which looks different from the above but is actually identical as a function of p (as can be verified by simplifying both expressions as polynomials in p).

    (b) The answer to (a) does not depend on whether the teams play all seven games no matter what. Imagine telling the players to continue playing the games even after the match has been decided, just for fun: the outcome of the match won't be affected by this, and this also means that the probability that A wins the match won't be affected by assuming that the teams always play 7 games!

19. In a chess tournament, n games are being played, independently. Each game ends in a win for one player with probability 0.4 and ends in a draw (tie) with probability 0.6. Find the PMFs of the number of games ending in a draw, and of the number of players whose games end in draws.

    Solution: Let G be the number of games ending in a draw and X be the number of players whose games end in draws, so $X = 2G$. Then $G \sim \text{Bin}(n, 0.6)$, so the PMFs are

    $$
    P(G = g) = \binom{n}{g}0.6^g \cdot 0.4^{n-g} \text{ for } g = 0, 1, \ldots, n,
    $$

    $$
    P(X = k) = P(G = k/2) = \binom{n}{k/2}0.6^{k/2} \cdot 0.4^{n-k/2} \text{ for } k = 0, 2, 4, \ldots, 2n.
    $$

20. Suppose that a lottery ticket has probability p of being a winning ticket, independently of other tickets. A gambler buys 3 tickets, hoping this will triple the chance of having at least one winning ticket.

(a) What is the distribution of how many of the 3 tickets are winning tickets?

(b) Show that the probability that at least 1 of the 3 tickets is winning is $3p - 3p^2 + p^3$, in two different ways: by using inclusion-exclusion, and by taking the complement of the desired event and then using the PMF of a certain named distribution.

(c) Show that the gambler's chances of having at least one winning ticket do not quite triple (compared with buying only one ticket), but that they do *approximately* triple if p is small.

Solution:

(a) By the story of the Binomial, the distribution is $\text{Bin}(3, p)$.

(b) Let A_i be the event that the ith ticket wins, for $i = 1, 2, 3$. By inclusion-exclusion and symmetry, we have

$$P(A_1 \cup A_2 \cup A_3) = 3P(A_1) - \binom{3}{2} P(A_1 \cap A_2) + P(A_1 \cap A_2 \cap A_3) = 3p - 3p^2 + p^3.$$

The Binomial PMF yields the same result: for $X \sim \text{Bin}(3, p)$,

$$P(X \geq 1) = 1 - P(X = 0) = 1 - (1 - p)^3 = 1 - (1 - 3p + 3p^2 - p^3) = 3p - 3p^2 + p^3.$$

(c) For any $p \in (0, 1)$, we have $p^3 < 3p^2$ (since $p < 3$), so $3p - 3p^2 + p^3 < 3p$. So the probability does not triple. But if p is small, then $3p^2$ and p^3 are *very* small, and then $3p - 3p^2 + p^3 \approx 3p$.

21. Ⓢ Let $X \sim \text{Bin}(n, p)$ and $Y \sim \text{Bin}(m, p)$, independent of X. Show that $X - Y$ is *not* Binomial.

Solution: A Binomial can't be negative, but $X - Y$ is negative with positive probability.

22. There are two coins, one with probability p_1 of Heads and the other with probability p_2 of Heads. One of the coins is randomly chosen (with equal probabilities for the two coins). It is then flipped $n \geq 2$ times. Let X be the number of times it lands Heads.

(a) Find the PMF of X.

(b) What is the distribution of X if $p_1 = p_2$?

(c) Give an intuitive explanation of why X is *not* Binomial for $p_1 \neq p_2$ (its distribution is called a *mixture* of two Binomials). You can assume that n is large for your explanation, so that the frequentist interpretation of probability can be applied.

Solution:

(a) By LOTP, conditioning on which coin is chosen, we have

$$P(X = k) = \frac{1}{2}\binom{n}{k} p_1^k (1 - p_1)^{n-k} + \frac{1}{2}\binom{n}{k} p_2^k (1 - p_2)^{n-k},$$

for $k = 0, 1, \ldots, n$.

(b) For $p_1 = p_2$, the above expression reduces to the $\text{Bin}(n, p_1)$ PMF.

(c) A mixture of two Binomials is *not* Binomial (except in the degenerate case $p_1 = p_2$).

Marginally, each toss has probability $(p_1 + p_2)/2$ of landing Heads, but the tosses are *not* independent since earlier tosses give information about which coin was chosen, which in turn gives information about later tosses.

Let n be large, and imagine repeating the entire experiment many times (each repetition consists of choosing a random coin and flipping it n times). We would expect to see *either* approximately np_1 Heads about half the time, and approximately np_2 Heads about half the time. In contrast, with a $\text{Bin}(n, p)$ distribution we would expect to see approximately np Heads; no fixed choice of p can create the behavior described above.

23. There are n people eligible to vote in a certain election. Voting requires registration. Decisions are made independently. Each of the n people will register with probability p_1. Given that a person registers, they will vote with probability p_2. Given that a person votes, they will vote for Kodos (who is one of the candidates) with probability p_3. What is the distribution of the number of votes for Kodos (give the PMF, fully simplified, or the name of the distribution, including its parameters)?

    Solution: Let X be the number of votes for Kodos. By the story of the Binomial, $X \sim \text{Bin}(n, p_1 p_2 p_3)$. The PMF is $P(X = k) = \binom{n}{k} p^k q^{n-k}$ for $k \in \{0, 1, \ldots, n\}$, with $p = p_1 p_2 p_3$ and $q = 1 - p$.

24. Let X be the number of Heads in 10 fair coin tosses.

    (a) Find the conditional PMF of X, given that the first two tosses both land Heads.

    (b) Find the conditional PMF of X, given that at least two tosses land Heads.

    Solution:

    (a) Let X_2 and X_8 be the number of Heads in the first 2 and last 8 tosses, respectively. Then the conditional PMF of X given $X_2 = 2$ is

    $$
    \begin{aligned}
    P(X = k | X_2 = 2) &= P(X_2 + X_8 = k | X_2 = 2) \\
    &= P(X_8 = k - 2 | X_2 = 2) \\
    &= P(X_8 = k - 2) \\
    &= \binom{8}{k-2} \left(\frac{1}{2}\right)^{k-2} \left(\frac{1}{2}\right)^{8-(k-2)} \\
    &= \frac{1}{256} \binom{8}{k-2},
    \end{aligned}
    $$

    for $k = 2, 3, \ldots, 10$.

    (b) The conditional PMF of X given $X \geq 2$ is

    $$
    \begin{aligned}
    P(X = k | X \geq 2) &= \frac{P(X = k, X \geq 2)}{P(X \geq 2)} \\
    &= \frac{P(X = k)}{1 - P(X = 0) - P(X = 1)} \\
    &= \frac{\binom{10}{k} \left(\frac{1}{2}\right)^{10}}{1 - \left(\frac{1}{2}\right)^{10} - 10 \left(\frac{1}{2}\right)^{10}} \\
    &= \frac{1}{1013} \binom{10}{k},
    \end{aligned}
    $$

    for $k = 2, 3, \ldots, 10$.

25. Ⓢ Alice flips a fair coin n times and Bob flips another fair coin $n+1$ times, resulting in independent $X \sim \text{Bin}(n, \frac{1}{2})$ and $Y \sim \text{Bin}(n+1, \frac{1}{2})$.

    (a) Show that $P(X < Y) = P(n - X < n + 1 - Y)$.

    (b) Compute $P(X < Y)$.

    Hint: Use (a) and the fact that X and Y are integer-valued.

    Solution:

    (a) Note that $n - X \sim \text{Bin}(n, 1/2)$ and $n + 1 - Y \sim \text{Bin}(n+1, 1/2)$ (we can interpret this by thinking of counting Tails rather than counting Heads), with $n - X$ and $n + 1 - Y$ independent. So $P(X < Y) = P(n - X < n + 1 - Y)$, since both sides have exactly the same structure.

    (b) We have

    $$P(X < Y) = P(n - X < n + 1 - Y) = P(Y < X + 1) = P(Y \leq X)$$

    since X and Y are integer-valued (e.g., $Y < 5$ is equivalent to $Y \leq 4$). But $Y \leq X$ is the complement of $X < Y$, so $P(X < Y) = 1 - P(X < Y)$. Thus, $P(X < Y) = 1/2$.

26. If $X \sim \text{HGeom}(w, b, n)$, what is the distribution of $n - X$? Give a short proof.

    Solution: By the story of the Hypergeometric, $n - X \sim \text{HGeom}(b, w, n)$. Thinking of X as the number of white balls in n draws from an urn with w white balls and b black balls, $n - X$ is the number of black balls, and has the same story except with the roles of white and black swapped.

27. Recall de Montmort's matching problem from Chapter 1: in a deck of n cards labeled 1 through n, a match occurs when the number on the card matches the card's position in the deck. Let X be the number of matching cards. Is X Binomial? Is X Hypergeometric?

    Solution: No, X is neither Binomial nor Hypergeometric (for $n \geq 2$). To see this, note that the event $X = n - 1$ is impossible: if $n - 1$ cards match, then the remaining card must match too. This constraint on the possible values of X rules out the Binomial, Hypergeometric, and any other distribution whose support contains $n - 1$.

28. Ⓢ There are n eggs, each of which hatches a chick with probability p (independently). Each of these chicks survives with probability r, independently. What is the distribution of the number of chicks that hatch? What is the distribution of the number of chicks that survive? (Give the PMFs; also give the names of the distributions and their parameters, if applicable.)

    Solution:

    Let H be the number of eggs that hatch and X be the number of hatchlings that survive. Think of each egg as a Bernoulli trial, where for H we define "success" to mean hatching, while for X we define "success" to mean surviving. For example, in the picture above, where ☺ denotes an egg that hatches with the chick surviving, ⓧ denotes an egg that hatched but whose chick died, and ▭ denotes an egg that didn't hatch, the events $H = 7$, $X = 5$ occurred. By the story of the Binomial, $H \sim \text{Bin}(n, p)$, with PMF

    $$P(H = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

    for $k = 0, 1, \ldots, n$.

The eggs independently have probability pr each of hatching a chick that survives. By the story of the Binomial, we have $X \sim \text{Bin}(n, pr)$, with PMF

$$P(X = k) = \binom{n}{k}(pr)^k(1 - pr)^{n-k}$$

for $k = 0, 1, \dots, n$.

29. Ⓢ A sequence of n independent experiments is performed. Each experiment is a success with probability p and a failure with probability $q = 1 - p$. Show that conditional on the number of successes, all valid possibilities for the list of outcomes of the experiment are equally likely.

Solution:

Let X_j be 1 if the jth experiment is a success and 0 otherwise, and let $X = X_1 + \cdots + X_n$ be the total number of successes. Then for any k and any $a_1, \dots, a_n \in \{0, 1\}$ with $a_1 + \cdots + a_n = k$,

$$\begin{aligned}
P(X_1 = a_1, \dots, X_n = a_n | X = k) &= \frac{P(X_1 = a_1, \dots, X_n = a_n, X = k)}{P(X = k)} \\
&= \frac{P(X_1 = a_1, \dots, X_n = a_n)}{P(X = k)} \\
&= \frac{p^k q^{n-k}}{\binom{n}{k} p^k q^{n-k}} \\
&= \frac{1}{\binom{n}{k}}.
\end{aligned}$$

This does not depend on $a_1, \dots, a_n$. Thus, for n independent Bernoulli trials, given that there are exactly k successes, the $\binom{n}{k}$ possible sequences consisting of k successes and $n - k$ failures are equally likely. Interestingly, the conditional probability above also does not depend on p (this is closely related to the notion of a *sufficient statistic*, which is an important concept in statistical inference).

30. A certain company has $n + m$ employees, consisting of n women and m men. The company is deciding which employees to promote.

(a) Suppose for this part that the company decides to promote t employees, where $1 \le t \le n + m$, by choosing t random employees (with equal probabilities for each set of t employees). What is the distribution of the number of women who get promoted?

(b) Now suppose that instead of having a predetermined number of promotions to give, the company decides independently for each employee, promoting the employee with probability p. Find the distributions of the number of women who are promoted, the number of women who are not promoted, and the number of employees who are promoted.

(c) In the set-up from (b), find the conditional distribution of the number of women who are promoted, given that exactly t employees are promoted.

Solution:

(a) By the story of the Hypergeometric, the distribution is $\text{HGeom}(n, m, t)$.

(b) Let W be the number of women who are promoted and T be the number of employees who are promoted. By the story of the Binomial, $W \sim \text{Bin}(n, p)$ women are promoted, $n - W \sim \text{Bin}(n, 1 - p)$ women are not promoted, and $T \sim \text{Bin}(n + m, p)$ employees are promoted.

(c) With notation as in (b), it follows from Theorem 3.9.2 (or by the same reasoning as in the Fisher exact test example) that the conditional distribution of W given $T = t$ is $\mathrm{HGeom}(n, m, t)$.

31. Once upon a time, a famous statistician offered tea to a lady. The lady claimed that she could tell whether milk had been added to the cup before or after the tea. The statistician decided to run some experiments to test her claim.

    (a) The lady is given 6 cups of tea, where it is known in advance that 3 will be milk-first and 3 will be tea-first, in a completely random order. The lady gets to taste each and then guess which 3 were milk-first. Assume for this part that she has no ability whatsoever to distinguish milk-first from tea-first cups of tea. Find the probability that at least 2 of her 3 guesses are correct.

    (b) Now the lady is given one cup of tea, with probability $1/2$ of it being milk-first. She needs to say whether she thinks it was milk-first. Let p_1 be the lady's probability of being correct given that it was milk-first, and p_2 be her probability of being correct given that it was tea-first. She claims that the cup was milk-first. Find the *posterior odds* that the cup is milk-first, given this information.

    Solution:

    (a) Let X be how many she gets right. Then $X \sim \mathrm{HGeom}(3, 3, 3)$ ("tag" the cups she chooses, and then randomly choose which 3 of the 6 cups are actually milk-first). So

    $$P(X \geq 2) = P(X = 2) + P(X = 3) = \frac{\binom{3}{2}\binom{3}{1} + \binom{3}{3}\binom{3}{0}}{\binom{6}{3}} = \frac{9 + 1}{20} = \frac{1}{2}.$$

    Alternatively, note that by symmetry, X and $3 - X$ have the same distribution, so $P(X = 2) + P(X = 3) = P(3 - X = 2) + P(3 - X = 3) = P(X = 0) + P(X = 1)$, which again gives $P(X \geq 2) = 1/2$.

    (b) Let M be the event that the cup is milk-first and C be the event that she claims it was milk-first. By Bayes' rule and LOTP,

    $$P(M|C) = \frac{P(C|M)P(M)}{P(C|M)P(M) + P(C|M^c)P(M^c)} = \frac{p_1/2}{p_1/2 + (1 - p_2)/2} = \frac{p_1}{p_1 + (1 - p_2)},$$

    so the posterior odds are $p_1/(1 - p_2)$. Alternatively, we can use the odds form of Bayes' rule. Note that the result is sensitivity/$(1 - \mathrm{specificity})$ for her test for "milk-first disease".

32. In Evan's history class, 10 out of 100 key terms will be randomly selected to appear on the final exam; Evan must then choose 7 of those 10 to define. Since he knows the format of the exam in advance, Evan is trying to decide how many key terms he should study.

    (a) Suppose that Evan decides to study s key terms, where s is an integer between 0 and 100. Let X be the number of key terms appearing on the exam that he has studied. What is the distribution of X? Give the name and parameters, in terms of s.

    (b) Using R or other software, calculate the probability that Evan knows at least 7 of the 10 key terms that appear on the exam, assuming that he studies $s = 75$ key terms.

    Solution:

    (a) Thinking of the terms Evan has studied as tagged and the terms he hasn't studied as untagged, we have $X \sim \mathrm{HGeom}(s, 100 - s, 10)$.

    (b) Using the command `1-phyper(6,75,25,10)` in R gives

    $$P(X \geq 7) = 1 - P(X \leq 6) \approx 0.785.$$

33. A book has n typos. Two proofreaders, Prue and Frida, independently read the book. Prue catches each typo with probability p_1 and misses it with probability $q_1 = 1 - p_1$, independently, and likewise for Frida, who has probabilities p_2 of catching and $q_2 = 1 - p_2$ of missing each typo. Let X_1 be the number of typos caught by Prue, X_2 be the number caught by Frida, and X be the number caught by at least one of the two proofreaders.

    (a) Find the distribution of X.

    (b) For this part only, assume that $p_1 = p_2$. Find the conditional distribution of X_1 given that $X_1 + X_2 = t$.

    Solution:

    (a) By the story of the Binomial, $X \sim \text{Bin}(n, 1 - q_1 q_2)$.

    (b) By Theorem 3.9.2 (or by noting that the structure is the same as in the Fisher exact test), $X_1 | (X_1 + X_2 = t) \sim \text{HGeom}(n, n, t)$. Alternatively, we can use Bayes' rule directly. Let $p = p_1 = p_2$ and $T = X_1 + X_2 \sim \text{Bin}(2n, p)$. Then

    $$P(X_1 = k | T = t) = \frac{P(T = t | X_1 = k) P(X_1 = k)}{P(T = t)} = \frac{\binom{n}{t-k} p^{t-k} q^{n-t+k} \binom{n}{k} p^k q^{n-k}}{\binom{2n}{t} p^t q^{2n-t}} = \frac{\binom{n}{t-k} \binom{n}{k}}{\binom{2n}{t}}$$

    for $k \in \{0, 1, \ldots, t\}$, so again the conditional distribution is $\text{HGeom}(n, n, t)$.

34. There are n students at a certain school, of whom $X \sim \text{Bin}(n, p)$ are Statistics majors. A simple random sample of size m is drawn ("simple random sample" means sampling without replacement, with all subsets of the given size equally likely).

    (a) Find the PMF of the number of Statistics majors in the sample, using the law of total probability (don't forget to say what the support is). You can leave your answer as a sum (though with some algebra it can be simplified, by writing the binomial coefficients in terms of factorials and using the binomial theorem).

    (b) Give a story proof derivation of the distribution of the number of Statistics majors in the sample; simplify fully.

    Hint: Does it matter whether the students declare their majors before or after the random sample is drawn?

    Solution:

    (a) Let Y be the number of Statistics majors in the sample. The support is $0, 1 \ldots, m$. Let $q = 1 - p$. By LOTP,

    $$P(Y = y) = \sum_{k=0}^{n} P(Y = y | X = k) P(X = k)$$
    $$= \frac{1}{\binom{n}{m}} \sum_{k=0}^{n} \binom{k}{y} \binom{n-k}{m-y} \binom{n}{k} p^k q^{n-k},$$

    for $0 \le y \le m$ (note that any term with $y > k$ or $m - y > n - k$ is 0).

    We can simplify the sum algebraically as follows (not required for this part). Despite all the exclamation points, we are much more enthusiastic about the method from (b)!

Writing the binomial coefficients in terms of factorials, we have

$$P(Y = y) = \frac{m!(n-m)!}{n!} \sum_{k=y}^{n-m+y} \frac{k!(n-k)!n!}{y!(k-y)!(m-y)!(n-m-k+y)!k!(n-k)!} p^k q^{n-k}$$

$$= \frac{m!}{(m-y)!y!} p^y q^{m-y} \sum_{k=y}^{n-m+y} \frac{(n-m)!}{(k-y)!(n-m-(k-y))!} p^{k-y} q^{n-m-(k-y)}$$

$$= \binom{m}{y} p^y q^{m-y} \sum_{j=0}^{n-m} \binom{n-m}{j} p^j q^{n-m-j}$$

$$= \binom{m}{y} p^y q^{m-y}.$$

(b) The distribution of Y has nothing to do with when the majors were declared, so we can let them be made after drawing the sample. Then by the story of the Binomial, we have $Y \sim \text{Bin}(m, p)$.

35. Ⓢ Players A and B take turns in answering trivia questions, starting with player A answering the first question. Each time A answers a question, she has probability p_1 of getting it right. Each time B plays, he has probability p_2 of getting it right.

(a) If A answers m questions, what is the PMF of the number of questions she gets right?

(b) If A answers m times and B answers n times, what is the PMF of the total number of questions they get right (you can leave your answer as a sum)? Describe exactly when/whether this is a Binomial distribution.

(c) Suppose that the first player to answer correctly wins the game (with no predetermined maximum number of questions that can be asked). Find the probability that A wins the game.

Solution:

(a) The r.v. is $\text{Bin}(m, p_1)$, so the PMF is $\binom{m}{k} p_1^k (1-p_1)^{m-k}$ for $k \in \{0, 1, \ldots, m\}$.

(b) Let T be the total number of questions they get right. To get a total of k questions right, it must be that A got 0 and B got k, or A got 1 and B got $k-1$, etc. These are disjoint events so the PMF is

$$P(T = k) = \sum_{j=0}^{k} \binom{m}{j} p_1^j (1-p_1)^{m-j} \binom{n}{k-j} p_2^{k-j} (1-p_2)^{n-(k-j)}$$

for $k \in \{0, 1, \ldots, m+n\}$, with the usual convention that $\binom{n}{k}$ is 0 for $k > n$.

This is the $\text{Bin}(m+n, p)$ distribution if $p_1 = p_2 = p$ (using the story for the Binomial, or using Vandermonde's identity). For $p_1 \neq p_2$, it's not a Binomial distribution, since the trials have different probabilities of success; having some trials with one probability of success and other trials with another probability of success isn't equivalent to having trials with some "effective" probability of success.

(c) Let $r = P(\text{A wins})$. Conditioning on the results of the first question for each player, we have

$$r = p_1 + (1-p_1)p_2 \cdot 0 + (1-p_1)(1-p_2)r,$$

which gives $r = \frac{p_1}{1-(1-p_1)(1-p_2)} = \frac{p_1}{p_1+p_2-p_1p_2}$.

36. There are n voters in an upcoming election in a certain country, where n is a large, even number. There are two candidates: Candidate A (from the Unite Party) and Candidate B (from the Untie Party). Let X be the number of people who vote for Candidate A. Suppose that each voter chooses randomly whom to vote for, independently and with equal probabilities.

(a) Find an exact expression for the probability of a tie in the election (so the candidates end up with the same number of votes).

(b) Use Stirling's approximation, which approximates the factorial function as

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n,$$

to find a simple approximation to the probability of a tie. Your answer should be of the form $1/\sqrt{cn}$, with c a constant (which you should specify).

Solution:

(a) We have $X \sim \text{Bin}(n, 1/2)$, so

$$P(\text{election is tied}) = P(X = n/2) = \binom{n}{n/2}\frac{1}{2^n}.$$

(b) By Stirling's approximation,

$$P(X = n/2) \approx \frac{\sqrt{2\pi n}(\frac{n}{e})^n}{2\pi(n/2)(\frac{n}{2e})^n} \cdot \frac{1}{2^n} = \frac{1}{\sqrt{\pi n/2}},$$

which is of the desired form with $c = \pi/2$.

37. Ⓢ A message is sent over a noisy channel. The message is a sequence $x_1, x_2, \ldots, x_n$ of n bits ($x_i \in \{0, 1\}$). Since the channel is noisy, there is a chance that any bit might be corrupted, resulting in an error (a 0 becomes a 1 or vice versa). Assume that the error events are independent. Let p be the probability that an individual bit has an error ($0 < p < 1/2$). Let $y_1, y_2, \ldots, y_n$ be the received message (so $y_i = x_i$ if there is no error in that bit, but $y_i = 1 - x_i$ if there is an error there).

To help detect errors, the nth bit is reserved for a parity check: x_n is defined to be 0 if $x_1 + x_2 + \cdots + x_{n-1}$ is even, and 1 if $x_1 + x_2 + \cdots + x_{n-1}$ is odd. When the message is received, the recipient checks whether y_n has the same parity as $y_1 + y_2 + \cdots + y_{n-1}$. If the parity is wrong, the recipient knows that at least one error occurred; otherwise, the recipient assumes that there were no errors.

(a) For $n = 5, p = 0.1$, what is the probability that the received message has errors which go undetected?

(b) For general n and p, write down an expression (as a sum) for the probability that the received message has errors which go undetected.

(c) Give a simplified expression, not involving a sum of a large number of terms, for the probability that the received message has errors which go undetected.

Hint for (c): Letting

$$a = \sum_{k \text{ even, } k \geq 0} \binom{n}{k} p^k (1-p)^{n-k} \text{ and } b = \sum_{k \text{ odd, } k \geq 1} \binom{n}{k} p^k (1-p)^{n-k},$$

the binomial theorem makes it possible to find simple expressions for $a + b$ and $a - b$, which then makes it possible to obtain a and b.

Solution:

(a) Note that $\sum_{i=1}^{n} x_i$ is even. If the number of errors is even (and nonzero), the errors will go undetected; otherwise, $\sum_{i=1}^{n} y_i$ will be odd, so the errors will be detected.

The number of errors is $\text{Bin}(n, p)$, so the probability of undetected errors when $n = 5, p = 0.1$ is

$$\binom{5}{2} p^2 (1-p)^3 + \binom{5}{4} p^4 (1-p) \approx 0.073.$$

(b) By the same reasoning as in (a), the probability of undetected errors is

$$\sum_{k \text{ even, } k \geq 2} \binom{n}{k} p^k (1-p)^{n-k}.$$

(c) Let a, b be as in the hint. Then

$$a + b = \sum_{k \geq 0} \binom{n}{k} p^k (1-p)^{n-k} = 1,$$

$$a - b = \sum_{k \geq 0} \binom{n}{k} (-p)^k (1-p)^{n-k} = (1-2p)^n.$$

Solving for a and b gives

$$a = \frac{1 + (1-2p)^n}{2} \text{ and } b = \frac{1 - (1-2p)^n}{2}.$$

$$\sum_{k \text{ even, } k \geq 0} \binom{n}{k} p^k (1-p)^{n-k} = \frac{1 + (1-2p)^n}{2}.$$

Subtracting off the possibility of no errors, we have

$$\sum_{k \text{ even, } k \geq 2} \binom{n}{k} p^k (1-p)^{n-k} = \frac{1 + (1-2p)^n}{2} - (1-p)^n.$$

Sanity check: Note that letting $n = 5, p = 0.1$ here gives 0.073, which agrees with (a); letting $p = 0$ gives 0, as it should; and letting $p = 1$ gives 0 for n odd and 1 for n even, which again makes sense.

Independence of r.v.s

38. (a) Give an example of dependent r.v.s X and Y such that $P(X < Y) = 1$.

(b) Give an example of independent r.v.s X and Y such that $P(X < Y) = 1$.

Solution:

(a) Let X be any random variable, and $Y = X + 1$. Then $P(X < Y) = P(0 < 1) = 1$.

(b) A simple way to make this happen is to choose X and Y such that Y's support is strictly to the right of X's support. For example, let $X \sim \text{Bin}(10, 1/2)$ and $Y = Z + 11$ with $Z \sim \text{Bin}(10, 1/2)$, with X and Z independent. Then X and Y are independent, but X is always less than Y.

39. Give an example of two discrete random variables X and Y on the same sample space such that X and Y have the same distribution, with support $\{1, 2, \ldots, 10\}$, but the event $X = Y$ *never* occurs. If X and Y are independent, is it still possible to construct such an example?

    Solution: Let X be Discrete Uniform on $1, 2, \ldots, 10$. Let's define Y in terms of X, making Y have the same distribution as X but making it so that the event $X = Y$ can never occur. For example, we can let

    $$Y = \begin{cases} X + 1, \text{ if } X \neq 10 \\ 1, \text{ if } X = 10. \end{cases}$$

    Such examples are impossible if X and Y are independent, since in that case we have

    $$P(X = Y) = \sum_{j=1}^{10} P(X = Y = j) = \sum_{j=1}^{10} P(X = j)P(Y = j) > 0.$$

40. Suppose X and Y are discrete r.v.s such that $P(X = Y) = 1$. This means that X and Y always take on the same value.

    (a) Do X and Y have the same PMF?

    (b) Is it possible for X and Y to be independent?

    Solution:

    (a) Yes, since X and Y always take on the same value, so the probabilities with which X takes on each possible value must agree with the corresponding probabilities for Y.

    (b) No, they are extremely dependent (except in the degenerate case where X and Y are constants): if we learn what X is, then we know (with probability 1) what Y is.

41. If X, Y, Z are r.v.s such that X and Y are independent and Y and Z are independent, does it follow that X and Z are independent?

    Hint: Think about simple and extreme examples.

    Solution: No, as seen by considering the extreme case where X and Z are the same r.v.

42. Ⓢ Let X be a random day of the week, coded so that Monday is 1, Tuesday is 2, etc. (so X takes values $1, 2, \ldots, 7$, with equal probabilities). Let Y be the next day after X (again represented as an integer between 1 and 7). Do X and Y have the same distribution? What is $P(X < Y)$?

    Solution: Yes, X and Y have the same distribution, since Y is also equally likely to represent any day of the week. However, X is likely to be less than Y. Specifically,

    $$P(X < Y) = P(X \neq 7) = \frac{6}{7}.$$

    In general, if Z and W are *independent* r.v.s with the same distribution, then $P(Z < W) = P(W < Z)$ by symmetry. Here though, X and Y are *dependent*, and we have $P(X < Y) = 6/7, P(X = Y) = 0, P(Y < X) = 1/7$.

43. (a) Is it possible to have two r.v.s X and Y such that X and Y have the same distribution but $P(X < Y) \geq p$, where:
    - $p = 0.9$?
    - $p = 0.99$?
    - $p = 0.9999999999999$?

- $p = 1$?

For each, give an example showing it is possible, or prove it is impossible.

Hint: Do the previous question first.

(b) Consider the same question as in Part (a), but now assume that X and Y are independent. Do your answers change?

Solution:

(a) The first three are possible; the fourth is impossible. Let n be a positive integer. Generalizing the idea from the previous problem, let X be Discrete Uniform on $0, 1, \ldots, 10^n - 1$ and

$$Y = \begin{cases} X + 1, \text{ if } X \neq 10^n - 1 \\ 0, \text{ if } X = 10^n - 1. \end{cases}$$

Then Y is also Discrete Uniform on $0, 1, \ldots, 10^n - 1$, and

$$P(X < Y) = P(X \neq 10^n - 1) = 1 - 10^{-n}.$$

By choosing n sufficiently large, we can make $P(X < Y) \geq p$, for any fixed $p < 1$.

On the other hand, it is impossible to make $P(X < Y) = 1$ if X and Y have the same distribution. To see this, assume (for contradiction) that X and Y do have those properties. Let's consider the event $X < a$, for any constant a. There are two disjoint ways to have $X < a$ and $X < Y$ hold: either $X < Y < a$, or $X < a < Y$. Since X and Y have the same distribution, we have

$$\begin{aligned} P(Y < a) &= P(X < a) \\ &= P(X < Y \text{ and } X < a) \\ &= P(X < Y < a) + P(X < a < Y) \\ &= P(Y < a) + P(X < a < Y), \end{aligned}$$

which shows that $P(X < a < Y) = 0$ for all a. Since there are countably many rational numbers and the union of countably many events of probability 0 has probability 0, we then have

$$P(X < a < Y \text{ for some rational } a) = 0.$$

On the other hand, $x < y$ if and only if $x < a < y$ for some rational a, so

$$P(X < a < Y \text{ for some rational } a) = P(X < Y) = 1,$$

which gives a contradiction.

(b) If X and Y are independent and have the same distribution, then

$$P(X < Y) = P(Y < X)$$

by symmetry. In the continuous case we then have, $P(X < Y) = P(Y < X) = 1/2$. In general, we have $P(X < Y) = P(Y < X) \leq 1/2$ since

$$P(X < Y) + P(X = Y) + P(X > Y) = 1.$$

So there is no $p > 1/2$ for which $P(X < Y) \geq p$.

44. For x and y binary digits (0 or 1), let $x \oplus y$ be 0 if $x = y$ and 1 if $x \neq y$ (this operation is called *exclusive or* (often abbreviated to XOR), or *addition mod 2*).

(a) Let $X \sim \text{Bern}(p)$ and $Y \sim \text{Bern}(1/2)$, independently. What is the distribution of $X \oplus Y$?

(b) With notation as in (a), is $X \oplus Y$ independent of X? Is $X \oplus Y$ independent of Y? Be sure to consider both the case $p = 1/2$ and the case $p \neq 1/2$.

(c) Let $X_1, \ldots, X_n$ be i.i.d. Bern(1/2) r.v.s. For each nonempty subset J of $\{1, 2, \ldots, n\}$, let

$$Y_J = \bigoplus_{j \in J} X_j,$$

where the notation means to "add" in the $\oplus$ sense all the elements of J; the order in which this is done doesn't matter since $x \oplus y = y \oplus x$ and $(x \oplus y) \oplus z = x \oplus (y \oplus z)$. Show that $Y_J \sim \text{Bern}(1/2)$ and that these $2^n - 1$ r.v.s are pairwise independent, but not independent. For example, we can use this to simulate 1023 pairwise independent fair coin tosses using only 10 independent fair coin tosses.

Hint: Apply the previous parts with $p = 1/2$. Show that if J and K are two different nonempty subsets of $\{1, 2, \ldots, n\}$, then we can write $Y_J = A \oplus B, Y_K = A \oplus C$, where A consists of the X_i with $i \in J \cap K$, B consists of the X_i with $i \in J \cap K^c$, and C consists of the X_i with $i \in J^c \cap K$. Then A, B, C are independent since they are based on disjoint sets of X_i. Also, at most one of these sets of X_i can be empty. If $J \cap K = \emptyset$, then $Y_J = B, Y_K = C$. Otherwise, compute $P(Y_J = y, Y_K = z)$ by conditioning on whether $A = 1$.

Solution:

(a) The distribution of $X \oplus Y$ is Bern(1/2), no matter what p is:

$$\begin{aligned}
P(X \oplus Y = 1) &= P(X \oplus Y = 1 | X = 1)P(X = 1) + P(X \oplus Y = 1 | X = 0)P(X = 0) \\
&= P(Y = 0)P(X = 1) + P(Y = 1)P(X = 0) \\
&= p/2 + (1 - p)/2 \\
&= 1/2.
\end{aligned}$$

(b) The conditional distribution of $X \oplus Y | (X = x)$ is Bern(1/2), as shown within the above calculation. This conditional distribution does not depend on x, so $X \oplus Y$ is independent of X. This result and the result from (a) make sense intuitively: adding Y destroys all information about X, resulting in a fair coin flip independent of X. Note that given $X = x$, $X \oplus Y$ is x with probability 1/2 and $1 - x$ with probability 1/2, which is another way to see that $X \oplus Y | (X = x) \sim \text{Bern}(1/2)$.

If $p = 1/2$, then the above reasoning shows that $X \oplus Y$ is independent of Y. So $X \oplus Y$ is independent of X and independent of Y, even though it is clearly not independent of the pair (X, Y).

But if $p \neq 1/2$, then $X \oplus Y$ is not independent of Y. The conditional distribution of $X \oplus Y | (Y = y)$ is Bern$(p(1 - y) + (1 - p)y)$, since

$$P(X \oplus Y = 1 | Y = y) = P(X \oplus y = 1) = P(X \neq y) = p(1 - y) + (1 - p)y.$$

(c) These r.v.s are not independent since, for example, if we know $Y_{\{1\}}$ and $Y_{\{2\}}$, then we know $Y_{\{1,2\}}$ via $Y_{\{1,2\}} = Y_1 + Y_2$. But they are *pairwise* independent. To show this, let's use the notation and approach from the hint. We can write Y_J and Y_K in their stated forms by partitioning $J \cup K$ (the set of indices that appear in Y_J or Y_K) into the sets $J \cap K$, $J \cap K^c$, and $J^c \cap K$.

Assume that $J \cap K$ is nonempty (the case where it is empty was handled in the hint).

By (a), $A \sim \text{Bern}(1/2)$. Then for $y \in \{0, 1\}, z \in \{0, 1\}$,

$$
\begin{aligned}
P(Y_J = y, Y_K = z) &= \frac{1}{2}P(A \oplus B = y, A \oplus C = z|A = 1) + \frac{1}{2}P(A \oplus B = y, A \oplus C = z|A = 0) \\
&= \frac{1}{2}P(1 \oplus B = y)P(1 \oplus C = z) + \frac{1}{2}P(0 \oplus B = y)P(0 \oplus C = z) \\
&= \frac{1}{8} + \frac{1}{8} \\
&= \frac{1}{4} \\
&= P(Y_J = y)P(Y_K = z),
\end{aligned}
$$

using the fact that A, B, C are independent and the fact that A, B, C, Y_J, Y_K are Bern(1/2). Thus, Y_J and Y_K are independent.

Mixed practice

45. Ⓢ A new treatment for a disease is being tested, to see whether it is better than the standard treatment. The existing treatment is effective on 50% of patients. It is believed initially that there is a 2/3 chance that the new treatment is effective on 60% of patients, and a 1/3 chance that the new treatment is effective on 50% of patients. In a pilot study, the new treatment is given to 20 random patients, and is effective for 15 of them.

(a) Given this information, what is the probability that the new treatment is better than the standard treatment?

(b) A second study is done later, giving the new treatment to 20 new random patients. Given the results of the first study, what is the PMF for how many of the new patients the new treatment is effective on? (Letting p be the answer to (a), your answer can be left in terms of p.)

Solution:

(a) Let B be the event that the new treatment is better than the standard treatment and let X be the number of people in the study for whom the new treatment is effective. By Bayes' rule and LOTP,

$$
\begin{aligned}
P(B|X = 15) &= \frac{P(X = 15|B)P(B)}{P(X = 15|B)P(B) + P(X = 15|B^c)P(B^c)} \\
&= \frac{\binom{20}{15}(0.6)^{15}(0.4)^5(\frac{2}{3})}{\binom{20}{15}(0.6)^{15}(0.4)^5(\frac{2}{3}) + \binom{20}{15}(0.5)^{20}(\frac{1}{3})}.
\end{aligned}
$$

(b) Let Y be how many of the new patients the new treatment is effective for and $p = P(B|X = 15)$ be the answer from (a). Then for $k \in \{0, 1, \ldots, 20\}$,

$$
\begin{aligned}
P(Y = k|X = 15) &= P(Y = k|X = 15, B)P(B|X = 15) + P(Y = k|X = 15, B^c)P(B^c|X = 15) \\
&= P(Y = k|B)P(B|X = 15) + P(Y = k|B^c)P(B^c|X = 15) \\
&= \binom{20}{k}(0.6)^k(0.4)^{20-k}p + \binom{20}{k}(0.5)^{20}(1 - p).
\end{aligned}
$$

(This distribution is *not* Binomial. As in the coin with a random bias problem, the individual outcomes are conditionally independent but not independent. Given the true probability of effectiveness of the new treatment, the pilot study is irrelevant and the distribution is Binomial, but without knowing that, we have a mixture of two different Binomial distributions.)

46. Independent Bernoulli trials are performed, with success probability $1/2$ for each trial. An important question that often comes up in such settings is how many trials to perform. Many controversies have arisen in statistics over the issue of how to analyze data coming from an experiment where the number of trials can depend on the data collected so far.

For example, if we can follow the rule "keep performing trials until there are more than twice as many failures as successes, and then stop", then naively looking at the ratio of failures to successes (if and when the process stops) will give more than 2:1 rather than the true theoretical 1:1 ratio; this could be a very misleading result! However, it might *never* happen that there are more than twice as many failures as successes; in this problem, you will find the probability of that happening.

(a) Two gamblers, A and B, make a series of bets, where each has probability $1/2$ of winning a bet, but A gets \$2 for each win and loses \$1 for each loss (a very favorable game for A!). Assume that the gamblers are allowed to borrow money, so they can and do gamble forever. Let p_k be the probability that A, starting with \$$k$, will ever reach \$0, for each $k \geq 0$. Explain how this story relates to the original problem, and how the original problem can be solved if we can find p_k.

(b) Find p_k.

Hint: As in the gambler's ruin, set up and solve a difference equation for p_k. We have $p_k \to 0$ as $k \to \infty$ (you don't need to prove this, but it should make sense since the game is so favorable to A, which will result in A's fortune going to ∞; a formal proof, not required here, could be done using the *law of large numbers*, an important theorem from Chapter 10). The solution can be written neatly in terms of the golden ratio.

(c) Find the probability of ever having more than twice as many failures as successes with independent Bern$(1/2)$ trials, as originally desired.

Solution:

(a) Think of a win for A as a success and a win for B as a failure. Let X_n be the number of successes after n bets, and $Y_n = n - X_n$ be the number of failures. If A starts with \$1, then after n rounds, A has $1 + 2X_n - Y_n$ dollars. We can visualize this as a random walk where after each bet, A moves 2 steps to the right or 1 step to the left. So A reaches 0 eventually if and only if $1 + 2X_n - Y_n \leq 0$ for some n (when moving to the left, A takes steps of size 1 and so can't jump over 0). On the other hand, $Y_n > 2X_n$ is equivalent to $Y_n \geq 2X_n + 1$. So p_1 is the probability of ever reaching a 2:1 ratio, and $1 - p_1$ is the probability of never reaching a 2:1 ratio.

(b) Conditioning on the first step, for any $k \geq 1$ we have

$$p_k = \frac{1}{2}p_{k+2} + \frac{1}{2}p_{k-1}.$$

We also have the initial condition $p_0 = 1$. The characteristic equation is

$$x = \frac{1}{2}x^3 + \frac{1}{2},$$

which factors as $(x - 1)(x^2 + x - 1) = 0$. So the solution is of the form

$$p_k = a + b\varphi^{-k} + c(-\varphi)^k,$$

where $\varphi = (1 + \sqrt{5})/2$ is the golden ratio. But $c = 0$ since otherwise letting k be a large enough even number would make p_k greater than 1, and $a = 0$ since $p_k \to 0$ as $k \to \infty$. Then $b = 1$ since $p_0 = 1$. Thus, $p_k = \varphi^{-k}$.

(c) By the previous parts, $p_1 = (-1 + \sqrt{5})/2 \approx 0.618$.

47. A copy machine is used to make n pages of copies per day. The machine has two trays in which paper gets loaded, and each page used is taken randomly and independently from one of the trays. At the beginning of the day, the trays are refilled so that they each have m pages.

(a) Let pbinom(x, n, p) be the CDF of the Bin(n, p) distribution, evaluated at x. In terms of pbinom, find a simple expression for the probability that both trays have enough paper on any particular day, when this probability is strictly between 0 and 1 (also specify the values of m for which the probability is 0 and the values for which it is 1).

Hint: Be careful about whether inequalities are strict, since the Binomial is discrete.

(b) Using a computer, find the smallest value of m for which there is at least a 95% chance that both trays have enough paper on a particular day, for $n = 10, n = 100, n = 1000$, and $n = 10000$.

Hint: If you use R, you may find the following commands useful:
`g <- function(m,n)` *[your answer from (a)]* defines a function g such that $g(m, n)$ is your answer from (a), `g(1:100,100)` gives the vector $(g(1, 100), \dots, g(100, 100))$, `which(v>0.95)` gives the indices of the components of vector **v** that exceed 0.95, and `min(w)` gives the minimum of a vector **w**.

Solution:

(a) Label the trays as first tray and second tray. Let X be the number of pages requested from the first tray, and $Y = n - X$ be the number requested from the second tray. The desired probability is 0 if $2m < n$ since clearly there won't be enough paper in that case, and it is 1 if $m \geq n$ since then each tray on its own could handle all the copying. For $n/2 \leq m < n$, the desired probability is

$$
\begin{aligned}
P(X \leq m, Y \leq m) &= P(n - m \leq X \leq m) \\
&= P(n - m - 1 < X \leq m) \\
&= \text{pbinom}(m, n, 1/2) - \text{pbinom}(n - m - 1, n, 1/2).
\end{aligned}
$$

(b) In R, entering the commands

```
g <- function(m,n) pbinom(m,n,1/2)-pbinom(n-m-1,n,1/2)
min(which(g(1:10,10)>0.95))
```

yields $m = 8$ for the case $n = 10$. Similarly, we obtain $m = 60$ for $n = 100$, $m = 531$ for $n = 1000$, and $m = 5098$ for $n = 10000$. Note that the values of m approach $n/2$ as n grows, which makes sense intuitively since for n large, the fraction of requests going to the first tray is likely to be close to $1/2$ (this can be proven using the law of large numbers, an important theorem from Chapter 10).

4

Expectation

Expectations and variances

1. Bobo, the amoeba from Chapter 2, currently lives alone in a pond. After one minute Bobo will either die, split into two amoebas, or stay the same, with equal probability. Find the expectation and variance for the number of amoebas in the pond after one minute.

   Solution: Let X be the number of amoebas in the pond after 1 minute, so $P(X = 0) = P(X = 1) = P(X = 2) = 1/3$. Then

   $$E(X) = \frac{1}{3} \cdot 0 + \frac{1}{3} \cdot 1 + \frac{1}{3} \cdot 2 = 1,$$
   $$E(X^2) = \frac{1}{3} \cdot 0^2 + \frac{1}{3} \cdot 1^2 + \frac{1}{3} \cdot 2^2 = \frac{5}{3},$$
   $$\text{Var}(X) = E(X^2) - (E(X))^2 = \frac{5}{3} - 1^2 = \frac{2}{3}.$$

2. In the Gregorian calendar, each year has either 365 days (a normal year) or 366 days (a leap year). A year is randomly chosen, with probability 3/4 of being a normal year and 1/4 of being a leap year. Find the mean and variance of the number of days in the chosen year.

   Solution: Let X be the number of days in the chosen year. Then

   $$E(X) = \frac{3}{4} \cdot 365 + \frac{1}{4} \cdot 366 = 365.25.$$
   $$E(X^2) = \frac{3}{4} \cdot 365^2 + \frac{1}{4} \cdot 366^2 = 133407.8$$
   $$\text{Var}(X) = E(X^2) - (E(X))^2 = 133407.8 - 365.25^2 = 0.2375.$$

3. (a) A fair die is rolled. Find the expected value of the roll.

   (b) Four fair dice are rolled. Find the expected total of the rolls.

   Solution:

   (a) The expected value is

   $$\frac{1}{6} \cdot (1 + 2 + \cdots + 6) = \frac{6 \cdot 7}{6 \cdot 2} = 3.5.$$

   (b) By linearity, the expected total is $4 \cdot 3.5 = 14$.

4. A fair die is rolled some number of times. You can choose whether to stop after $1, 2$, or 3 rolls, and your decision can be based on the values that have appeared so far. You receive the value shown on the last roll of the die, in dollars. What is your optimal strategy (to maximize your expected winnings)? Find the expected winnings for this strategy.

Hint: Start by considering a simpler version of this problem, where there are at most 2 rolls. For what values of the first roll should you continue for a second roll?

Solution: The expected value for one roll is 3.5 (as calculated in the previous problem). So if given the choice between 1 rolls and 2 rolls, you should continue for a second roll if the first roll is $1, 2$, or 3, and stop rolling if the first roll is $4, 5$, or 6. It follows that the expected value of the roll-at-most-twice game is

$$\frac{1}{2} \cdot (\frac{4+5+6}{3}) + \frac{1}{2} \cdot 3.5 = 4.25.$$

So for the roll-at-most-thrice game, the optimal strategy is to stop after the first roll if it's a 5 or a 6, and otherwise reroll and follow the strategy described above for the roll-at-most-twice game.

The expected winnings are then

$$\frac{1}{3} \cdot (\frac{5+6}{2}) + \frac{2}{3} \cdot 4.25 = \frac{14}{3} \approx 4.67.$$

5. Find the mean and variance of a Discrete Uniform r.v. on $1, 2, \ldots, n$.

Hint: See the math appendix for some useful facts about sums.

Solution: Let X be Discrete Uniform on $1, 2, \ldots, n$. Using facts about sums from the math appendix, we have

$$E(X) = \frac{1}{n}(1 + 2 + \cdots + n) = \frac{n+1}{2},$$
$$E(X^2) = \frac{1}{n}(1^2 + 2^2 + \cdots + n^2) = \frac{(n+1)(2n+1)}{6},$$
$$\text{Var}(X) = \frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2 = \frac{n^2 - 1}{12}.$$

6. Two teams are going to play a best-of-7 match (the match will end as soon as either team has won 4 games). Each game ends in a win for one team and a loss for the other team. Assume that each team is equally likely to win each game, and that the games played are independent. Find the mean and variance of the number of games played.

Solution: Let X be the number of games played. The PMF of X is

$$P(X = 4) = \frac{2}{2^4} = \frac{1}{8},$$
$$P(X = 5) = 2 \cdot \binom{4}{1} \cdot \frac{1}{2^4} \cdot \frac{1}{2} = \frac{1}{4},$$
$$P(X = 6) = 2 \cdot \binom{5}{3} \cdot \frac{1}{2^5} \cdot \frac{1}{2} = \frac{5}{16}$$
$$P(X = 7) = 2 \cdot \binom{6}{3} \cdot \frac{1}{2^6} \cdot \frac{1}{2} = \frac{5}{16},$$

where, for example, we found $P(X = 6)$ by noting that for a specific team to win in 6 games, it must win exactly 3 of the first 5 games and then win the 6th game. As a check, note that it makes sense that $P(X = 6) = P(X = 7)$, since if the match has gone 5 games, then one team must be ahead 3 to 2, and then the match will go 6 games if that team wins and 7 games if that team loses.

Thus,

$$E(X) = 4 \cdot \frac{1}{8} + 5 \cdot \frac{1}{4} + 6 \cdot \frac{5}{16} + 7 \cdot \frac{5}{16} = 5.8125,$$
$$E(X^2) = 4^2 \cdot \frac{1}{8} + 5^2 \cdot \frac{1}{4} + 6^2 \cdot \frac{5}{16} + 7^2 \cdot \frac{5}{16} = 34.8125,$$
$$\text{Var}(X) = 34.8125 - 5.8125^2 \approx 1.027.$$

7. A certain small town, whose population consists of 100 families, has 30 families with 1 child, 50 families with 2 children, and 20 families with 3 children. The *birth rank* of one of these children is 1 if the child is the firstborn, 2 if the child is the secondborn, and 3 if the child is the thirdborn.

(a) A random family is chosen (with equal probabilities), and then a random child within that family is chosen (with equal probabilities). Find the PMF, mean, and variance of the child's birth rank.

(b) A random child is chosen in the town (with equal probabilities). Find the PMF, mean, and variance of the child's birth rank.

Solution:

(a) Let X be the child's birth rank. Using LOTP to condition on how many children are in the random family, the PMF of X is

$$P(X = 1) = 0.3 \cdot 1 + 0.5 \cdot 0.5 + 0.2 \cdot (1/3) \approx 0.617,$$
$$P(X = 2) = 0.3 \cdot 0 + 0.5 \cdot 0.5 + 0.2 \cdot (1/3) \approx 0.317,$$
$$P(X = 3) = 0.3 \cdot 0 + 0.5 \cdot 0 + 0.2 \cdot (1/3) \approx 0.0667.$$

So

$$E(X) = P(X = 1) + 2P(X = 2) + 3P(X = 3) = 1.45,$$
$$E(X^2) = 1^2 P(X = 1) + 2^2 P(X = 2) + 3^2 P(X = 3) \approx 2.483,$$
$$\text{Var}(X) = E(X^2) - (EX)^2 \approx 0.381.$$

(b) Let Y be the child's birth rank. There are $30 + 50 \cdot 2 + 20 \cdot 3 = 190$ children, of which 100 are firstborn, 70 are second born, and 20 are thirdborn. So the PMF of Y is

$$P(Y = 1) = 100/190 \approx 0.526,$$
$$P(Y = 2) = 70/190 \approx 0.368,$$
$$P(Y = 3) = 20/190 \approx 0.105.$$

It makes sense intuitively that $P(X = 1)$ from (a) is less than $P(Y = 1)$, since the sampling method from (a) gives all families equal probabilities, whereas the sampling method from (b) gives the children equal probabilities, which effectively gives higher probabilities of being represented to larger families. Then

$$E(Y) = P(Y = 1) + 2P(Y = 2) + 3P(Y = 3) = 1.579,$$
$$E(Y^2) = 1^2 P(Y = 1) + 2^2 P(Y = 2) + 3^2 P(Y = 3) \approx 2.947,$$
$$\text{Var}(Y) = E(Y^2) - (EY)^2 \approx 0.454.$$

8. A certain country has four regions: North, East, South, and West. The populations of these regions are 3 million, 4 million, 5 million, and 8 million, respectively. There are 4 cities in the North, 3 in the East, 2 in the South, and there is only 1 city in the West. Each person in the country lives in exactly one of these cities.

(a) What is the average size of a city in the country? (This is the arithmetic mean of the populations of the cities, and is also the expected value of the population of a city chosen uniformly at random.)

Hint: Give the cities *names* (labels).

(b) Show that without further information it is impossible to find the variance of the population of a city chosen uniformly at random. That is, the variance depends on how the people within each region are allocated between the cities in that region.

(c) A region of the country is chosen uniformly at random, and then a city within that region is chosen uniformly at random. What is the expected population size of this randomly chosen city?

Hint: First find the selection probability for each city.

(d) Explain intuitively why the answer to (c) is larger than the answer to (a).

Solution:

(a) Let x_i be the population of the ith city, for $1 \leq i \leq 10$, with respect to some labeling of the cities. The sum of the city populations equals the sum of the region populations (either way, it's just the total number of people), so the average size of a city is

$$\frac{1}{10} \sum_{i=1}^{10} x_i = \frac{1}{10}(3 + 4 + 5 + 8) \text{ million} = 2 \text{ million.}$$

(b) The variance is

$$\frac{1}{10} \sum_{i=1}^{10} x_i^2 - \left(\frac{1}{10} \sum_{i=1}^{10} x_i \right)^2.$$

The second term is $(2 \times 10^6)^2$ by (a), but the first term depends on how people are allocated to cities. For example, if the two cities in the South each have 2.5 million people, then the South contributes $2(2.5 \cdot 10^6)^2 = 1.25 \cdot 10^{13}$ to the sum of x_i^2, but if one city in the South has 5 million people and the other is completely deserted, then the South contributes $(5 \cdot 10^6)^2 = 25 \cdot 10^{12}$ to the sum of x_i^2.

(c) Let x_i be the population of the ith city, with respect to an ordering where cities $1, \ldots, 4$ are in the North, $5, \ldots, 7$ are in the East, $8, 9$ are in the South, and 10 is in the West. The probability that the ith city is chosen is $1/16$ for $1 \leq i \leq 4$, $1/12$ for $5 \leq i \leq 7$, $1/8$ for $8 \leq i \leq 9$, and $1/4$ for $i = 10$. So the expected population size is

$$\frac{x_1 + \cdots + x_4}{16} + \frac{x_5 + x_6 + x_7}{12} + \frac{x_8 + x_9}{8} + \frac{x_{10}}{4} = \left(\frac{3}{16} + \frac{4}{12} + \frac{5}{8} + \frac{8}{4} \right) \cdot 10^6 \approx 3.146 \times 10^6.$$

(d) It makes sense intuitively that the answer to (c) is greater than the answer to (a), since in (a) all cities are equally likely, whereas in (c) the city in the West, which is very populous, is more likely to be chosen than any other particular city.

9. Consider the following simplified scenario based on *Who Wants to Be a Millionaire?*, a game show in which the contestant answers multiple-choice questions that have 4 choices per question. The contestant (Fred) has answered 9 questions correctly already, and is now being shown the 10th question. He has no idea what the right answers are to the 10th or 11th questions are. He has one "lifeline" available, which he can apply on any question, and which narrows the number of choices from 4 down to 2. Fred has the following options available.

(a) Walk away with $16,000.

(b) Apply his lifeline to the 10th question, and then answer it. If he gets it wrong, he will leave with $1,000. If he gets it right, he moves on to the 11th question. He then leaves with $32,000 if he gets the 11th question wrong, and $64,000 if he gets the 11th question right.

(c) Same as the previous option, except not using his lifeline on the 10th question, and instead applying it to the 11th question (if he gets the 10th question right).

Find the expected value of each of these options. Which option has the highest expected value? Which option has the lowest variance?

Solution:

For option (a), Fred always gets $16,000, so the expected value is $16,000 and the variance is 0.

For option (b), the expected value is

$$\frac{1}{2} \cdot 1 + \frac{1}{2} \left(\frac{3}{4} \cdot 32 + \frac{1}{4} \cdot 64 \right) = 20.5,$$

measured in thousands of dollars.

For option (c), the expected value is

$$\frac{3}{4} \cdot 1 + \frac{1}{4} \left(\frac{1}{2} \cdot 32 + \frac{1}{2} \cdot 64 \right) = 12.75,$$

measured in thousands of dollars.

So option (b) has the highest expected value, and option (a) has the lowest variance.

10. Consider the St. Petersburg paradox (Example 4.3.14), except that you receive n rather than 2^n if the game lasts for n rounds. What is the fair value of this game? What if the payoff is n^2?

Solution: If the payoff is n, then the fair value of the game is $\sum_{n=1}^{\infty} \frac{n}{2^n}$. This sum can be interpreted as the expected value of a First Success distribution with parameter $1/2$. So the sum is 2. If the payoff is n^2, then the fair value of the game is $\sum_{n=1}^{\infty} \frac{n^2}{2^n}$. This sum can be interpreted as $E(N^2)$ for $N \sim \text{FS}(1/2)$. By Example 4.6.4,

$$E(N^2) = \text{Var}(N) + (EN)^2 = \frac{1/2}{1/4} + 2^2 = 6.$$

11. Martin has just heard about the following exciting gambling strategy: bet $1 that a fair coin will land Heads. If it does, stop. If it lands Tails, double the bet for the next toss, now betting $2 on Heads. If it does, stop. Otherwise, double the bet for the next toss to $4. Continue in this way, doubling the bet each time and then stopping right after winning a bet. Assume that each individual bet is fair, i.e., has an expected net winnings of 0. The idea is that

$$1 + 2 + 2^2 + 2^3 + \cdots + 2^n = 2^{n+1} - 1,$$

so the gambler will be $1 ahead after winning a bet, and then can walk away with a profit.

Martin decides to try out this strategy. However, he only has $31, so he may end up walking away bankrupt rather than continuing to double his bet. On average, how much money will Martin win?

Solution:

Let X be Martin's net winnings. Let $N \sim \text{FS}(1/2)$ be the trial on which the coin first lands Heads (we can assume that the coin continues to get flipped until landing Heads, even though Martin may already have left bankrupt before this happens). If $N \leq 5$,

then Martin will keep betting until winning, and walk away with a profit of $1. For example, if $N = 5$, then Martin bets $1, $2, $4, $8 unsuccessfully, but then bets his remaining $16 on the next toss, which does land Heads, so he receives $32, giving him a net payoff of $32 - 1 - 2 - 4 - 8 - 16 = 1$ dollar. If $N > 5$, Martin will leave bankrupt. So the PMF of X is

$$P(X = -31) = P(N > 5) = P(\text{first 5 tosses land Tails}) = \frac{1}{32},$$

$$P(X = 1) = P(N \leq 5) = 1 - P(N > 5) = \frac{31}{32}.$$

The expected value is

$$E(X) = -31 \cdot \frac{1}{32} + \frac{31}{32} = 0.$$

Martin's strategy gives a high chance of a small gain but this is offset by the small (but nonzero) chance of a big loss, such that he breaks even on average.

12. Let X be a discrete r.v. with support $-n, -n + 1, \ldots, 0, \ldots, n - 1, n$ for some positive integer n. Suppose that the PMF of X satisfies the symmetry property $P(X = -k) = P(X = k)$ for all integers k. Find $E(X)$.

Solution: The symmetry property makes the mean 0:

$$E(X) = \sum_{k=1}^{n} kP(X = k) + \sum_{k=1}^{n} -kP(X = -k) = \sum_{k=1}^{n} (kP(X = k) - kP(X = k)) = 0.$$

13. Ⓢ Are there discrete random variables X and Y such that $E(X) > 100E(Y)$ but Y is greater than X with probability at least 0.99?

Solution: Yes. Consider what happens if we make X usually 0 but on rare occasions, X is extremely large (like the outcome of a lottery); Y, on the other hand, can be more moderate. For a simple example, let X be 10^6 with probability $1/100$ and 0 with probability $99/100$, and let Y be the constant 1 (which is a degenerate r.v.).

14. Let X have PMF
$$P(X = k) = cp^k/k \text{ for } k = 1, 2, \ldots,$$
where p is a parameter with $0 < p < 1$ and c is a normalizing constant. We have $c = -1/\log(1 - p)$, as seen from the Taylor series

$$-\log(1 - p) = p + \frac{p^2}{2} + \frac{p^3}{3} + \ldots.$$

This distribution is called the *Logarithmic* distribution (because of the log in the above Taylor series), and has often been used in ecology. Find the mean and variance of X.

Solution: The mean is

$$E(X) = \sum_{k=1}^{\infty} kP(X = k) = c \sum_{k=1}^{\infty} p^k = \frac{cp}{1 - p}.$$

To get the variance, let's first find $E(X^2)$ (using LOTUS):

$$E(X^2) = \sum_{k=1}^{\infty} k^2 P(X = k) = c \sum_{k=1}^{\infty} kp^k.$$

We computed a similar sum when deriving the mean of a Geometric; specifically, Example 4.3.6 shows that

$$\sum_{k=1}^{\infty} kp^k = p \sum_{k=1}^{\infty} kp^{k-1} = \frac{p}{(1 - p)^2}.$$

Therefore,

$$\text{Var}(X) = \frac{cp}{(1-p)^2} - \frac{c^2 p^2}{(1-p)^2} = \frac{cp - c^2 p^2}{(1-p)^2}.$$

15. Player A chooses a random integer between 1 and 100, with probability p_j of choosing j (for $j = 1, 2, \ldots, 100$). Player B guesses the number that player A picked, and receives from player A that amount in dollars if the guess is correct (and 0 otherwise).

    (a) Suppose for this part that player B knows the values of p_j. What is player B's optimal strategy (to maximize expected earnings)?

    (b) Show that if both players choose their numbers so that the probability of picking j is proportional to $1/j$, then neither player has an incentive to change strategies, assuming the opponent's strategy is fixed. (In game theory terminology, this says that we have found a *Nash equilibrium.*)

    (c) Find the expected earnings of player B when following the strategy from (b). Express your answer both as a sum of simple terms and as a numerical approximation. Does the value depend on what strategy player A uses?

    Solution:

    (a) If player B guesses j, then their expected earnings are jp_j. So player B should choose j to maximize jp_j (if this j is not unique then player B can choose between all such j, deterministically or randomly).

    (b) Let q_j be the probability of player B choosing j. First assume that $q_j = c/j$, where

    $$\frac{1}{c} = \sum_{j=1}^{100} \frac{1}{j},$$

    and let the p_j's be general. Then the expected payoff to B is

    $$\sum_j jp_j q_j = c \sum_j p_j = c.$$

    The constant c does not depend on the p_j, so the expected payoffs will not change if player A changes strategy. Similarly, if $p_j = c/j$ and the q_j's are general, then the expected payoffs will not change if player B changes strategy.

    (c) As shown in (b), the expected earnings of B are c, regardless of what strategy A uses. Numerically,

    $$c = \left(\sum_{j=1}^{100} \frac{1}{j} \right)^{-1} \approx 0.193,$$

    using the R command `n <- 1:100; 1/sum(1/n)`.

16. The dean of Blotchville University boasts that the average class size there is 20. But the reality experienced by the majority of students there is quite different: they find themselves in huge courses, held in huge lecture halls, with hardly enough seats or Haribo gummi bears for everyone. The purpose of this problem is to shed light on the situation. For simplicity, suppose that every student at Blotchville University takes only one course per semester.

    (a) Suppose that there are 16 seminar courses, which have 10 students each, and 2 large lecture courses, which have 100 students each. Find the dean's-eye-view average class size (the simple average of the class sizes) and the student's-eye-view average class size (the average class size experienced by students, as it would be reflected by surveying students and asking them how big their classes are). Explain the discrepancy intuitively.

(b) Give a short proof that for *any* set of class sizes (not just those given above), the dean's-eye-view average class size will be strictly less than the student's-eye-view average class size, unless all classes have exactly the same size.

Hint: Relate this to the fact that variances are nonnegative.

Solution:

(a) The dean's-eye-view average is $(16 \cdot 10 + 2 \cdot 100)/18 = 20$. There are 160 students experiencing 10-student courses and 200 students experiencing 100-student courses, so the student's-eye-view average is $(160 \cdot 10 + 200 \cdot 100)/360 = 60$, which is 3 times higher than the dean's-eye-view average! The reason is that a large class gets weighted much more heavily than a small class in the student's-eye-view, but gets weighted the same as a small class in the dean's-eye-view. This phenomenon is known as the *class size paradox*.

(b) Let $c_1, \ldots, c_n$ be the class sizes, and $c = \sum_{j=1}^n c_j$. The dean's-eye-view average is $\frac{c}{n}$, and the student's-eye-view average is $\frac{1}{c} \sum_{j=1}^n c_j^2$. Let X be the size of a randomly selected class (with equal probabilities for each class). The fact that $\mathrm{Var}(X) > 0$ can be paraphrased as $E(X^2) > (E(X))^2$, and this says that

$$\frac{1}{n} \sum_{j=1}^n c_j^2 > \left(\frac{c}{n}\right)^2.$$

Thus,

$$\frac{1}{c} \sum_{j=1}^n c_j^2 > \frac{c}{n}.$$

17. The sociologist Elizabeth Wrigley-Field posed the following puzzle:

> *American fertility fluctuated dramatically in the decades surrounding the Second World War. Parents created the smallest families during the Great Depression, and the largest families during the postwar Baby Boom. Yet children born during the Great Depression came from larger families than those born during the Baby Boom. How can this be?*

(a) For a particular era, let n_k be the number of American families with exactly k children, for each $k \geq 0$. (Assume for simplicity that American history has cleanly been separated into eras, where each era has a well-defined set of families, and each family has a well-defined set of children; we are ignoring the fact that a particular family's size may change over time, that children grow up, etc.) For each $j \geq 0$, let

$$m_j = \sum_{k=0}^{\infty} k^j n_k.$$

For a *family* selected randomly in that era (with all families equally likely), find the expected number of children in the family. Express your answer only in terms of the m_j's.

(b) For a *child* selected randomly in that era (with all children equally likely), find the expected number of children in the child's family, only in terms of the m_j's.

(c) Give an intuitive explanation in words for which of the answers to (a) and (b) is larger, or whether they are equal. Explain how this relates to the Wrigley-Field puzzle.

Solution:

(a) The number of families is $n_0 + n_1 + n_2 + n_3 + \ldots$. There are n_0 families that would

report having 0 children, n_1 that would report having 1 child, etc., so the desired average is

$$\frac{0n_0 + 1n_1 + 2n_2 + 3n_3 + \dots}{n_0 + n_1 + n_2 + n_3 + \dots} = \frac{m_1}{m_0}.$$

(b) The number of children is $0n_0 + 1n_1 + 2n_2 + 3n_3 + \dots$. There are kn_k children who would report having exactly k children in their family, so the desired average is

$$\frac{0 \cdot 0n_0 + 1 \cdot 1n_1 + 2 \cdot 2n_2 + 3 \cdot 3n_3 + \dots}{0n_0 + 1n_1 + 2n_2 + 3n_3 + \dots} = \frac{m_2}{m_1}.$$

(c) The answer to (b) is larger than the answer to (a), since in (a) all families are weighted equally, whereas in (b) larger families are weighted more. This is similar to the class size paradox encountered in the previous problem, where the student's-eye-view average class size is greater than the dean's-eye-view average class size.

Wrigley-Field says that (a)'s average, the parents'-eye-view, was smaller for the Great Depression than for the Baby Boom, but that (b)'s average, the children's-eye-view larger for the Great Depression than for the Baby Boom. Since the answers to (a) and (b) differ, the direction of inequality in comparing the two eras need not be the same for both forms of average. For example, large families during the Great Depression would contribute far more toward making (b) large than toward making (a) large.

Named distributions

18. Ⓢ A fair coin is tossed repeatedly, until it has landed Heads at least once and has landed Tails at least once. Find the expected number of tosses.

Solution: Let X be the number of tosses needed, starting with the 2nd toss, to obtain an outcome different from that of the first toss. Then $X-1$ is Geom(1/2), so $E(X) = 2$. This does not include the first toss, so the expected total number of tosses is $E(X+1) = E(X) + 1 = 3$.
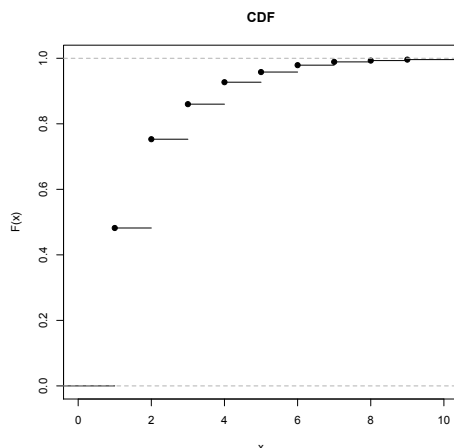
Sanity check: An answer of 2 or lower would be not make sense, since we always need at least 2 tosses, and sometimes we need more than 2 tosses. An answer of 4 or higher would also not make sense, since $4 = 2 + 2$ is the expected number of tosses if we wait for the coin to land Heads and then wait for the next time that the coin lands Tails *after* the first time the coin landed Heads.

19. Ⓢ A coin is tossed repeatedly until it lands Heads for the first time. Let X be the number of tosses that are required (including the toss that landed Heads), and let p be the probability of Heads, so that $X \sim \text{FS}(p)$. Find the CDF of X, and for $p = 1/2$ sketch its graph.

Solution: By the story of the Geometric, we have $X - 1 \sim \text{Geometric}(p)$. Using this or directly, the PMF is $P(X = k) = p(1-p)^{k-1}$ for $k \in \{1, 2, 3, \dots\}$ (and 0 otherwise). The CDF can be obtained by adding up the PMF (from $k = 1$ to $k = \lfloor x \rfloor$, where $\lfloor x \rfloor$ is the greatest integer less than or equal to x). We can also see directly that

$$P(X \le x) = 1 - P(X > x) = 1 - (1-p)^{\lfloor x \rfloor}$$

for $x \ge 1$, since $X > x$ says that the first $\lfloor x \rfloor$ flips land tails. The CDF is 0 for $x < 1$. For a fair coin, the CDF is $F(x) = 1 - \frac{1}{2^{\lfloor x \rfloor}}$ for $x \ge 1$, and $F(x) = 0$ for $x < 1$, as illustrated below.

CDF

20. Let $X \sim \text{Bin}(100, 0.9)$. For each of the following parts, construct an example showing that it is possible, or explain clearly why it is impossible. In this problem, Y is a random variable on the same probability space as X; note that X and Y are not necessarily independent.

(a) Is it possible to have $Y \sim \text{Pois}(0.01)$ with $P(X \geq Y) = 1$?

(b) Is it possible to have $Y \sim \text{Bin}(100, 0.5)$ with $P(X \geq Y) = 1$?

(c) Is it possible to have $Y \sim \text{Bin}(100, 0.5)$ with $P(X \leq Y) = 1$?

Solution:

(a) No, since there is a positive probability that Y will exceed 100, whereas X is always at most 100.

(b) Yes. To construct such an example, consider the same sequence of trials, except with two definitions of "success": a less stringent definition for X and a more stringent definition for Y (so any trial counted as a success for Y is also counted as a success for X). Specifically, we can let the experiment consist of rolling a fair 10-sided die (with sides labeled 1 through 10) 100 times, and let X and Y be the numbers of times the value of the die was at most 9 and at most 5, respectively.

(c) No, this is impossible, since if $X \leq Y$ holds with probability 1, then $E(X) \leq E(Y)$, but in fact we have $E(X) = 90 > 50 = E(Y)$.

21. ⓢ Let $X \sim \text{Bin}(n, \frac{1}{2})$ and $Y \sim \text{Bin}(n + 1, \frac{1}{2})$, independently.

(a) Let $V = \min(X, Y)$ be the smaller of X and Y, and let $W = \max(X, Y)$ be the larger of X and Y. So if X crystallizes to x and Y crystallizes to y, then V crystallizes to $\min(x, y)$ and W crystallizes to $\max(x, y)$. Find $E(V) + E(W)$.

(b) Show that $E|X - Y| = E(W) - E(V)$, with notation as in (a).

(c) Compute $\text{Var}(n - X)$ in two different ways.

Solution:

(a) Note that $V + W = X + Y$ (since adding the smaller and the larger of two numbers is the same as adding both numbers). So by linearity,

$$E(V) + E(W) = E(V + W) = E(X + Y) = E(X) + E(Y) = (2n + 1)/2 = n + \frac{1}{2}.$$

(b) Note that $|X - Y| = W - V$ (since the absolute difference between two numbers is the larger number minus the smaller number). So

$$E|X - Y| = E(W - V) = E(W) - E(V).$$

(c) We have $n - X \sim \text{Bin}(n, 1/2)$, so $\text{Var}(n - X) = n/4$. Alternatively, by properties of variance we have $\text{Var}(n - X) = \text{Var}(n) + \text{Var}(-X) = \text{Var}(X) = n/4$.

22. Ⓢ Raindrops are falling at an average rate of 20 drops per square inch per minute. What would be a reasonable distribution to use for the number of raindrops hitting a particular region measuring 5 inches^2 in t minutes? Why? Using your chosen distribution, compute the probability that the region has no rain drops in a given 3-second time interval.

Solution: A reasonable choice of distribution is $\text{Pois}(\lambda t)$, where $\lambda = 20 \cdot 5 = 100$ (the average number of raindrops per minute hitting the region). Assuming this distribution,

$$P(\text{no raindrops in 1/20 of a minute}) = e^{-100/20}(100/20)^0/0! = e^{-5} \approx 0.0067.$$

23. Ⓢ Alice and Bob have just met, and wonder whether they have a mutual friend. Each has 50 friends, out of 1000 other people who live in their town. They think that it's unlikely that they have a friend in common, saying "each of us is only friends with 5% of the people here, so it would be very unlikely that our two 5%'s overlap."

Assume that Alice's 50 friends are a random sample of the 1000 people (equally likely to be any 50 of the 1000), and similarly for Bob. Also assume that knowing who Alice's friends are gives no information about who Bob's friends are.

(a) Compute the expected number of mutual friends Alice and Bob have.

(b) Let X be the number of mutual friends they have. Find the PMF of X.

(c) Is the distribution of X one of the important distributions we have looked at? If so, which?

Solution:

(a) Let I_j be the indicator r.v. for the jth person being a mutual friend. Then

$$E\left(\sum_{j=1}^{1000} I_j\right) = 1000 E(I_1) = 1000 P(I_1 = 1) = 1000 \cdot \left(\frac{5}{100}\right)^2 = 2.5.$$

(b) Condition on who Alice's friends are, and then count the number of ways that Bob can be friends with exactly k of them. This gives

$$P(X = k) = \frac{\binom{50}{k}\binom{950}{50-k}}{\binom{1000}{50}}$$

for $0 \le k \le 50$ (and 0 otherwise).

(c) Yes, it is the Hypergeometric distribution, as shown by the PMF from (b) or by thinking of "tagging" Alice's friends (like the elk) and then seeing how many tagged people there are among Bob's friends.

24. Let $X \sim \text{Bin}(n, p)$ and $Y \sim \text{NBin}(r, p)$. Using a story about a sequence of Bernoulli trials, prove that $P(X < r) = P(Y > n - r)$.

Solution: Imagine an infinite sequence of Bernoulli trials, each with probability p of success. Let X be the number of successes among the first n trials, and Y be the number of failures obtained before the rth success. If $X < r$, then there are at most

$r-1$ successes and at least $n-(r-1)$ failures among the first n trials, so $Y \geq n-r+1$, which implies $Y > n-r$. Conversely, if $Y > n-r$, then there are at least $n-r+1$ failures before the rth success, so there are at most r successes in the first n trials (else there would be at least $(n-r+1)+r = n+1$ trials among the first n trials, which is a contradiction). Thus $X < r$ and $Y > n-r$ are the same event.

25. Ⓢ Calvin and Hobbes play a match consisting of a series of games, where Calvin has probability p of winning each game (independently). They play with a "win by two" rule: the first player to win two games more than his opponent wins the match. Find the expected number of games played.

    Hint: Consider the first two games as a pair, then the next two as a pair, etc.

    Solution: Think of the first 2 games, the 3rd and 4th, the 5th and 6th, etc. as "mini-matches." The match ends right after the first mini-match which isn't a tie. The probability of a mini-match not being a tie is $p^2 + q^2$, so the number of mini-matches needed is 1 plus a Geom$(p^2 + q^2)$ r.v. Thus, the expected number of games is $\frac{2}{p^2+q^2}$.

    Sanity check: For $p = 0$ or $p = 1$, this reduces to 2. The expected number of games is maximized when $p = \frac{1}{2}$, which makes sense intuitively. Also, it makes sense that the result is symmetric in p and q.

26. Nick and Penny are independently performing independent Bernoulli trials. For concreteness, assume that Nick is flipping a nickel with probability p_1 of Heads and Penny is flipping a penny with probability p_2 of Heads. Let $X_1, X_2, \ldots$ be Nick's results and $Y_1, Y_2, \ldots$ be Penny's results, with $X_i \sim \text{Bern}(p_1)$ and $Y_j \sim \text{Bern}(p_2)$.

    (a) Find the distribution and expected value of the first time at which they are simultaneously successful, i.e., the smallest n such that $X_n = Y_n = 1$.

    Hint: Define a new sequence of Bernoulli trials and use the story of the Geometric.

    (b) Find the expected time until at least one has a success (including the success).

    Hint: Define a new sequence of Bernoulli trials and use the story of the Geometric.

    (c) For $p_1 = p_2$, find the probability that their first successes are simultaneous, and use this to find the probability that Nick's first success precedes Penny's.

    Solution:

    (a) Let N be the time this happens. Then $N - 1 \sim \text{Geom}(p_1 p_2)$ by the story of the Geometric (in other words, N has a First Success distribution with parameter $p_1 p_2$). So the PMF is $P(N = n) = p_1 p_2 (1 - p_1 p_2)^{n-1}$ for $n = 1, 2, \ldots$, and the mean is $E(N) = 1/(p_1 p_2)$.

    (b) Let T be the time this happens, and let $q_1 = 1 - p_1, q_2 = 1 - p_2$. Define a new sequence of Bernoulli trials by saying that the jth trial is a success if at least one of the two people succeeds in the jth trial. These trials have probability $1 - q_1 q_2$ of success, which implies that $T - 1 \sim \text{Geom}(1 - q_1 q_2)$. Therefore, $E(T) = 1/(1 - q_1 q_2)$.

    (c) Let T_1 and T_2 be the first times at which Nick and Penny are successful, respectively. Let $p = p_1 = p_2$ and $q = 1 - p$. Then

    $$P(T_1 = T_2) = \sum_{n=1}^{\infty} P(T_1 = n | T_2 = n) P(T_2 = n) = \sum_{n=1}^{\infty} p^2 q^{2(n-1)} = \frac{p^2}{1 - q^2} = \frac{p}{2 - p}.$$

    Alternatively, consider the first trial where at least one of Nick and Penny succeeded. Their first successes are simultaneous if and only if they both succeed at this trial. So the probability is

    $$P(\text{both succeed}|\text{at least one succeeds}) = \frac{p^2}{1 - q^2} = \frac{p}{2 - p}.$$

Then by symmetry,

$$1 = P(T_1 < T_2) + P(T_2 < T_1) + P(T_1 = T_2) = 2P(T_1 < T_2) + P(T_1 = T_2).$$

So

$$P(T_1 < T_2) = \frac{1}{2}\left(1 - \frac{p}{2-p}\right) = \frac{1-p}{2-p}.$$

27. ⓢ Let X and Y be Pois(λ) r.v.s, and $T = X + Y$. Suppose that X and Y are *not* independent, and in fact $X = Y$. Prove or disprove the claim that $T \sim \text{Pois}(2\lambda)$ in this scenario.

Solution: The r.v. $T = 2X$ is *not* Poisson: it can only take even values $0, 2, 4, 6, \ldots$, whereas any Poisson r.v. has positive probability of being any of $0, 1, 2, 3, \ldots$.

Alternatively, we can compute the PMF of $2X$, or note that $\text{Var}(2X) = 4\lambda \neq 2\lambda = E(2X)$, whereas for any Poisson r.v. the variance equals the mean.

28. William is on a treasure hunt. There are t pieces of treasure, each of which is hidden in one of n locations. William searches these locations one by one, without replacement, until he has found all the treasure. (Assume that no location contains more than one piece of treasure, and that William *will* find the treasure piece when he searches a location that does have treasure.) Let X be the number of locations that William searches during his treasure hunt. Find the distribution of X, and find $E(X)$.

Solution: Let Y be the number of locations that William searches that do *not* have treasure. By the story of the Negative Hypergeometric, $Y \sim \text{NHGeom}(t, n-t, t)$. Including the locations that *do* have treasure gives $X = Y + t$, which has PMF

$$P(X = k) = P(Y = k - t) = \frac{\binom{t+(k-t)-1}{t-1}\binom{t+n-t-t-(k-t)}{t-t}}{\binom{t+n-t}{t}} = \frac{\binom{k-1}{t-1}}{\binom{n}{t}},$$

for $k \in \{t, t+1, \ldots, n\}$. Using the result for the mean of a Negative Hypergeometric r.v.,

$$E(X) = E(Y + t) = E(Y) + t = \frac{t(n-t)}{t+1} + t.$$

Note that in the extreme case $t = n$ (which means that *all* of the locations have treasure), this reduces to t, which makes sense since in that case William will definitely search all t rooms. Also note that

$$E(X) = \frac{t(n-t)}{t+1} + t \leq \frac{t(n-t)}{t} + t = n - t + t = n,$$

which must be true since we always have $X \leq n$. If t is very large though, $t+1$ is negligibly different from t, and then by the above (with an approximation in place of an inequality) we have $E(X) \approx n$.

29. Let $X \sim \text{Geom}(p)$, and define the function f by $f(x) = P(X = x)$, for all real x. Find $E(f(X))$. (The notation $f(X)$ means first evaluate $f(x)$ in terms of p and x, and then plug in X for x; it is *not* correct to say "$f(X) = P(X = X) = 1$".)

Solution: We have

$$f(x) = \begin{cases} pq^x, & \text{if } x \text{ is a nonnegative integer;} \\ 0, & \text{otherwise.} \end{cases}$$

Then $f(X) = pq^X$ (since X always takes on nonnegative integer values). By LOTUS,

$$E(f(X)) = \sum_{k=0}^{\infty} f(k)P(X = k) = \sum_{k=0}^{\infty} p^2 q^{2k} = p^2 \sum_{k=0}^{\infty} q^{2k}.$$

Summing the geometric series, we have

$$E(f(X)) = \frac{p^2}{1-q^2} = \frac{p^2}{(1-q)(1+q)} = \frac{p}{2-p}.$$

30. (a) Use LOTUS to show that for $X \sim \text{Pois}(\lambda)$ and any function g,

$$E(Xg(X)) = \lambda E(g(X+1)).$$

This is called the *Stein-Chen identity* for the Poisson.

(b) Find the third moment $E(X^3)$ for $X \sim \text{Pois}(\lambda)$ by using the identity from (a) and a bit of algebra to reduce the calculation to the fact that X has mean λ and variance λ.

Solution:

(a) By LOTUS and the substitution $j = k - 1$, we have

$$\begin{aligned}
E(Xg(X)) &= e^{-\lambda} \sum_{k=0}^{\infty} kg(k)\lambda^k/k! \\
&= e^{-\lambda} \sum_{k=1}^{\infty} g(k)\lambda^k/(k-1)! \\
&= \lambda e^{-\lambda} \sum_{j=0}^{\infty} g(j+1)\lambda^j/j! \\
&= \lambda E(g(X+1)).
\end{aligned}$$

(b) By the identity from (a), with $g(x) = x^2$, we have

$$E(X^3) = E(XX^2) = \lambda E(X+1)^2 = \lambda(EX^2 + 2EX + 1).$$

Plugging in $EX = \lambda$ and $EX^2 = \text{Var}(X) + (EX)^2 = \lambda + \lambda^2$, we obtain

$$E(X^3) = \lambda(\lambda + \lambda^2 + 2\lambda + 1) = \lambda^3 + 3\lambda^2 + \lambda.$$

31. In many problems about modeling count data, it is found that values of zero in the data are far more common than can be explained well using a Poisson model (we can make $P(X = 0)$ large for $X \sim \text{Pois}(\lambda)$ by making λ small, but that also constrains the mean and variance of X to be small since both are λ). The *Zero-Inflated Poisson* distribution is a modification of the Poisson to address this issue, making it easier to handle frequent zero values gracefully.

A Zero-Inflated Poisson r.v. X with parameters p and λ can be generated as follows. First flip a coin with probability of p of Heads. Given that the coin lands Heads, $X = 0$. Given that the coin lands Tails, X is distributed $\text{Pois}(\lambda)$. Note that if $X = 0$ occurs, there are two possible explanations: the coin could have landed Heads (in which case the zero is called a *structural zero*), or the coin could have landed Tails but the Poisson r.v. turned out to be zero anyway.

For example, if X is the number of chicken sandwiches consumed by a random person in a week, then $X = 0$ for vegetarians (this is a structural zero), but a chicken-eater could still have $X = 0$ occur by chance (since they might not happen to eat any chicken sandwiches that week).

(a) Find the PMF of a Zero-Inflated Poisson r.v. X.

(b) Explain why X has the same distribution as $(1 - I)Y$, where $I \sim \text{Bern}(p)$ is independent of $Y \sim \text{Pois}(\lambda)$.

(c) Find the mean of X in two different ways: directly using the PMF of X, and using

the representation from (b). For the latter, you can use the fact (which we prove in Chapter 7) that if r.v.s Z and W are independent, then $E(ZW) = E(Z)E(W)$.

(d) Find the variance X.

Solution:

(a) Let I be the indicator of the coin landing Heads. The PMF of X is given by

$$P(X = 0) = P(X = 0|I = 1)p + P(X = 0|I = 0)(1 - p) = p + (1 - p)e^{-\lambda},$$
$$P(X = k) = P(X = k|I = 1)p + P(X = k|I = 0)(1 - p) = (1 - p)e^{-\lambda}\lambda^k/k!,$$

for $k = 1, 2, \ldots$.

(b) Let $I \sim \text{Bern}(p)$ and let $Y \sim \text{Pois}(\lambda)$ be independent of I. Let $W = (1 - I)Y$. Given that $I = 1$, we have $W = 0$. Given that $I = 0$, we have that $W \sim \text{Pois}(\lambda)$ So W has the same probabilistic structure as X.

(c) Using the PMF of X,

$$E(X) = \sum_{k=0}^{\infty} kP(X = k) = (1 - p)\sum_{k=1}^{\infty} ke^{-\lambda}\lambda^k/k! = (1 - p)\lambda,$$

where we recognized the sum as the one that appeared when we showed that a $\text{Pois}(\lambda)$ r.v. has mean λ. Using the representation, we again have

$$E(X) = E((1 - I)Y) = E(1 - I)E(Y) = (1 - p)\lambda.$$

(d) Using the above representation,

$$E(X^2) = E((1 - I)^2Y^2) = E(1 - I)E(Y^2) = (1 - p)(\lambda + \lambda^2)$$
$$\text{Var}(X) = (1 - p)(\lambda + \lambda^2) - (1 - p)^2\lambda^2$$
$$= (1 - p)(1 + \lambda p)\lambda.$$

32. Ⓢ A discrete distribution has the *memoryless property* if for X a random variable with that distribution, $P(X \geq j + k|X \geq j) = P(X \geq k)$ for all nonnegative integers j, k.

(a) If X has a memoryless distribution with CDF F and PMF $p_i = P(X = i)$, find an expression for $P(X \geq j + k)$ in terms of $F(j), F(k), p_j, p_k$.

(b) Name a discrete distribution which has the memoryless property. Justify your answer with a clear interpretation in words or with a computation.

Solution:
(a) By the memoryless property,

$$P(X \geq k) = P(X \geq j + k|X \geq j) = \frac{P(X \geq j + k, X \geq j)}{P(X \geq j)} = \frac{P(X \geq j + k)}{P(X \geq j)},$$

so

$$P(X \geq j + k) = P(X \geq j)P(X \geq k) = (1 - F(j) + p_j)(1 - F(k) + p_k).$$

(b) The Geometric distribution is memoryless (in fact, it turns out to be essentially the *only* discrete memoryless distribution!). This follows from the story of the Geometric: consider Bernoulli trials, waiting for the first success (and defining waiting time to be the number of failures before the first success). Say we have already had j failures without a success. Then the additional waiting time from that point forward has the same distribution as the original waiting time (the Bernoulli trials neither are conspiring against the experimenter nor act as if the experimenter is due for a success: the trials are independent). A calculation agrees: for $X \sim \text{Geom}(p)$,

$$P(X \geq j + k|X \geq j) = \frac{P(X \geq j + k)}{P(X \geq j)} = \frac{q^{j+k}}{q^j} = q^k = P(X \geq k).$$

33. Find values of w, b, r such that the Negative Hypergeometric distribution with parameters w, b, r reduces to a Discrete Uniform on $\{0, 1, \ldots, n\}$. Justify your answer both in terms of the story of the Negative Hypergeometric and in terms of its PMF.

Solution: Let $w = 1, b = n, r = 1$. The Negative Hypergeometric distribution with these parameters is also Discrete Uniform on $\{0, 1, \ldots, n\}$. To see this via a story proof, consider a jar containing 1 white ball and n black balls, and balls are drawn randomly without replacement until all the balls have been drawn. Imagine lining up the balls in the order in which they are drawn. By symmetry, the white ball is equally likely to be anywhere in this line, so the number of black balls preceding it is equally likely to be any of the numbers $0, 1, \ldots, n$.

We can also check the fact that these distributions are the same by comparing the PMFs. The PMF of a NHGeom$(1, n, 1)$ r.v. X is
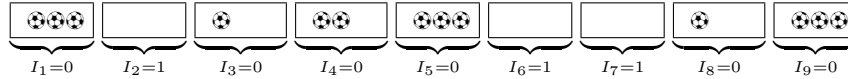
$$P(X = k) = \frac{\binom{1+k-1}{1-1}\binom{1+n-1-k}{1-1}}{\binom{1+n}{1}} = \frac{\binom{k}{0}\binom{n-k}{0}}{n+1} = \frac{1}{n+1},$$

for $k = 0, 1, \ldots, n$, which is also the DUnif$(0, 1, \ldots, n)$ PMF.

Indicator r.v.s

34. Ⓢ Randomly, k distinguishable balls are placed into n distinguishable boxes, with all possibilities equally likely. Find the expected number of empty boxes.

Solution:

Let I_j be the indicator random variable for the jth box being empty, so $I_1 + \cdots + I_n$ is the number of empty boxes (the above picture illustrates a possible outcome with 3 empty boxes, for $n = 9, k = 13$). Then $E(I_j) = P(I_j = 1) = (1 - 1/n)^k$. By linearity,

$$E\left(\sum_{j=1}^{n} I_j\right) = \sum_{j=1}^{n} E(I_j) = n(1 - 1/n)^k.$$

35. Ⓢ A group of 50 people are comparing their birthdays (as usual, assume their birthdays are independent, are not February 29, etc.). Find the expected number of pairs of people with the same birthday, and the expected number of days in the year on which at least two of these people were born.

Solution: Creating an indicator r.v. for each pair of people, we have that the expected number of pairs of people with the same birthday is $\binom{50}{2}\frac{1}{365}$ by linearity. Now create an indicator r.v. for each day of the year, taking the value 1 if at least two of the people were born that day (and 0 otherwise). Then the expected number of days on which at least two people were born is $365\left(1 - (\frac{364}{365})^{50} - 50 \cdot \frac{1}{365} \cdot (\frac{364}{365})^{49}\right)$.

36. Ⓢ A group of $n \geq 4$ people are comparing their birthdays (as usual, assume their birthdays are independent, are not February 29, etc.). Let I_{ij} be the indicator r.v. of i and j having the same birthday (for $i < j$). Is I_{12} independent of I_{34}? Is I_{12} independent of I_{13}? Are the I_{ij} independent?

Solution: The indicator I_{12} is independent of the indicator I_{34} since knowing the birthdays of persons 1 and 2 gives us no information about the birthdays of persons 3 and 4. Also, I_{12} is independent of I_{13} since even though both of these indicators involve person 1, knowing that persons 1 and 2 have the same birthday gives us no information

about whether persons 1 and 3 have the same birthday (this relies on the assumption that the 365 days are equally likely). In general, the indicator r.v.s here are pairwise independent. But they are *not* independent since, for example, if person 1 has the same birthday as person 2 and person 1 has the same birthday as person 3, then persons 2 and 3 must have the same birthday.

37. ⓢ A total of 20 bags of Haribo gummi bears are randomly distributed to 20 students. Each bag is obtained by a random student, and the outcomes of who gets which bag are independent. Find the average number of bags of gummi bears that the first three students get in total, and find the average number of students who get at least one bag.

    Solution: Let X_j be the number of bags of gummi bears that the jth student gets, and let I_j be the indicator of $X_j \geq 1$. Then $X_j \sim \text{Bin}(20, \frac{1}{20})$, so $E(X_j) = 1$. So $E(X_1 + X_2 + X_3) = 3$ by linearity.

    The average number of students who get at least one bag is

    $$E(I_1 + \cdots + I_{20}) = 20E(I_1) = 20P(I_1 = 1) = 20\left(1 - \left(\frac{19}{20}\right)^{20}\right).$$

38. Each of $n \geq 2$ people puts their name on a slip of paper (no two have the same name). The slips of paper are shuffled in a hat, and then each person draws one (uniformly at random at each stage, without replacement). Find the average number of people who draw their own names.

    Solution: Label the people as $1, 2, \ldots, n$, let I_j be the indicator of person j getting their own name, and let $X = I_1 + \cdots + I_n$. By symmetry and linearity,

    $$E(X) = nE(I_1) = n \cdot \frac{1}{n} = 1.$$

    (For large n, the Poisson paradigm says that X is approximately $\text{Pois}(1)$, so $P(X = 0) \approx 1/e$ for large n; note that this agrees with the result of the matching problem.)

39. Two researchers independently select simple random samples from a population of size N, with sample sizes m and n (for each researcher, the sampling is done without replacement, with all samples of the prescribed size equally likely). Find the expected size of the overlap of the two samples.

    Solution: Label the elements of the population $1, 2, \ldots, N$, and let I_j be the indicator of element j being in both samples. Let X be the size of the overlap. By symmetry, linearity, and the fundamental bridge,

    $$E(X) = N\left(\frac{m}{N} \cdot \frac{n}{N}\right) = \frac{mn}{N}.$$

    Alternatively, note that $X \sim \text{HGeom}(m, N - m, n)$ by the story of the Hypergeometric (imagine two sets of tags, one for each researcher). So again we have

    $$E(X) = n\frac{m}{N} = \frac{mn}{N}.$$

40. In a sequence of n independent fair coin tosses, what is the expected number of occurrences of the pattern HTH (consecutively)? Note that overlap is allowed, e.g., $HTHTH$ contains two overlapping occurrences of the pattern.

    Solution: Let I_j be the indicator of the pattern HTH occurring starting at position j, for $1 \leq j \leq n - 2$. By symmetry, linearity, and the fundamental bridge, the expected number of occurrences of HTH is $(n - 2)/8$ (assuming, of course, that $n \geq 3$; if $n < 3$, then there are no occurrences).

41. You have a well-shuffled 52-card deck. On average, how many pairs of adjacent cards are there such that both cards are red?

   Solution: Let I_j be the indicator of the jth and $(j + 1)$st cards both being red, for $1 \le j \le 51$. By symmetry, linearity, and the fundamental bridge, the desired expectation is

   $$51 \cdot \frac{26}{52} \cdot \frac{25}{51} = \frac{25}{2} = 12.5.$$

42. Suppose there are n types of toys, which you are collecting one by one. Each time you collect a toy, it is equally likely to be any of the n types. What is the expected number of distinct toy types that you have after you have collected t toys? (Assume that you will definitely collect t toys, whether or not you obtain a complete set before then.)

   Solution: Let I_j be the indicator of having the jth toy type in your collection after having collected t toys. By symmetry, linearity, and the fundamental bridge, the desired expectation is

   $$n\left(1 - \left(\frac{n-1}{n}\right)^t\right).$$

43. A building has n floors, labeled $1, 2, \ldots, n$. At the first floor, k people enter the elevator, which is going up and is empty before they enter. Independently, each decides which of floors $2, 3, \ldots, n$ to go to and presses that button (unless someone has already pressed it).

   (a) Assume for this part only that the probabilities for floors $2, 3, \ldots, n$ are equal. Find the expected number of stops the elevator makes on floors $2, 3, \ldots, n$.

   (b) Generalize (a) to the case that floors $2, 3, \ldots, n$ have probabilities $p_2, \ldots, p_n$ (respectively); you can leave your answer as a finite sum.

   Solution:

   (a) Let I_j be the indicator for a stop on the jth floor, for $j = 2, \ldots, n$. Let $X = I_2 + \cdots + I_n$ be the number of stops. By linearity and symmetry, $E(X) = (n-1)E(I_2)$. By the fundamental bridge, $E(I_2) = P(I_2 = 1) = 1 - (1 - \frac{1}{n-1})^k$. Thus,

   $$E(X) = (n-1)\left(1 - (1 - \frac{1}{n-1})^k\right).$$

   (b) With notation as above, we now have $E(I_j) = 1 - (1 - p_j)^k$, which gives

   $$E(X) = \sum_{j=2}^{n}(1 - (1 - p_j)^k) = (n-1) - \sum_{j=2}^{n}(1 - p_j)^k.$$

44. ⓢ There are 100 shoelaces in a box. At each stage, you pick two random ends and tie them together. Either this results in a longer shoelace (if the two ends came from different pieces), or it results in a loop (if the two ends came from the same piece). What are the expected number of steps until everything is in loops, and the expected number of loops after everything is in loops? (This is a famous interview problem; leave the latter answer as a sum.)

   Hint: For each step, create an indicator r.v. for whether a loop was created then, and note that the number of free ends goes down by 2 after each step.

   Solution: Initially there are 200 free ends. The number of free ends decreases by 2 each time since either two separate pieces are tied together, or a new loop is formed. So exactly 100 steps are always needed. Let I_j be the indicator r.v. for whether a new loop is formed at the jth step. At the time when there are n unlooped pieces (so $2n$ ends),

the probability of forming a new loop is $\frac{n}{\binom{2n}{2}} = \frac{1}{2n-1}$ since any 2 ends are equally likely to be chosen, and there are n ways to pick both ends of 1 of the n pieces. By linearity, the expected number of loops is

$$\sum_{n=1}^{100} \frac{1}{2n-1}.$$

45. Show that for any events $A_1, \ldots, A_n$,

$$P(A_1 \cap A_2 \cdots \cap A_n) \geq \sum_{j=1}^{n} P(A_j) - n + 1.$$

Hint: First prove a similar-looking statement about indicator r.v.s, by interpreting what the events $I(A_1 \cap A_2 \cdots \cap A_n) = 1$ and $I(A_1 \cap A_2 \cdots \cap A_n) = 0$ mean.

Solution: Note that

$$I(A_1 \cap A_2 \cdots \cap A_n) \geq \sum_{j=1}^{n} I(A_j) - n + 1,$$

since: if the left-hand side is 1 then $I(A_j) = 1$ for all j, making the right-hand side 1; if the left-hand side is 0, then $I(A_j) = 0$ for at least one value of j, so $\sum_{j=1}^{n} I(A_j) - n + 1 \leq (n-1) - n + 1 = 0$. Taking the expectation of both sides then gives the desired inequality.

46. You have a well-shuffled 52-card deck. You turn the cards face up one by one, without replacement. What is the expected number of non-aces that appear before the first ace? What is the expected number between the first ace and the second ace?

Solution: Number the non-aces from 1 to 48, and let I_j be the indicator of the jth non-ace appearing before the first ace. By symmetry, the jth non-ace and the 4 aces are equally likely to be in any order relative to each other, so $P(I_j = 1) = 1/5$. Then by linearity, the expected number of non-aces before the first aces is $48/5$. Similarly, the expected number of non-aces between the first ace and second ace is also $48/5$.

47. You are being tested for psychic powers. Suppose that you do not have psychic powers. A standard deck of cards is shuffled, and the cards are dealt face down one by one. Just after each card is dealt, you name any card (as your prediction). Let X be the number of cards you predict correctly. (See Diaconis (1978) for much more about the statistics of testing for psychic powers.)

(a) Suppose that you get no feedback about your predictions. Show that no matter what strategy you follow, the expected value of X stays the same; find this value. (On the other hand, the *variance* may be very different for different strategies. For example, saying "Ace of Spades" every time gives variance 0.)

Hint: Indicator r.v.s.

(b) Now suppose that you get partial feedback: after each prediction, you are told immediately whether or not it is right (but without the card being revealed). Suppose you use the following strategy: keep saying a specific card's name (e.g., "Ace of Spades") until you hear that you are correct. Then keep saying a different card's name (e.g., "Two of Spades") until you hear that you are correct (if ever). Continue in this way, naming the same card over and over again until you are correct and then switching to a new card, until the deck runs out. Find the expected value of X, and show that it is very close to $e - 1$.

Hint: Indicator r.v.s.

(c) Now suppose that you get complete feedback: just after each prediction, the card is

revealed. Call a strategy "stupid" if it allows, e.g., saying "Ace of Spades" as a guess after the Ace of Spades has already been revealed. Show that any non-stupid strategy gives the same expected value for X; find this value.

Hint: Indicator r.v.s.

Solution:

(a) Think of the cards as labeled from 1 to 52, so each card in the deck has a "codename" (this is just so we can use random variables rather than random cards). Let $C_1, \ldots, C_{52}$ be the labels of the cards in the shuffled deck (so the card on top of the deck has label C_1, etc.). Let $Y_1, \ldots, Y_{52}$ be your predictions, and let I_j be the indicator of your jth prediction being right. Then $X = I_1 + \cdots + I_{52}$. Since you do not have psychic powers (by assumption) and there is no feedback, $(C_1, \ldots, C_{52})$ is independent of $(Y_1, \ldots, Y_{52})$. Thus, $P(I_j = 1) = 1/52$ and

$$E(X) = E(I_1) + \cdots + E(I_{52}) = 1.$$

(b) To simplify notation, assume that the strategy is to keep saying "Card 1" until you hear that you are correct, then start saying "Card 2", etc. Let X_j be the indicator of your guessing the card with label j correctly when it is dealt. So $X = X_1 + \cdots + X_{52}$. We have $X_1 = 1$ always. For X_2, note that $X_2 = 1$ if and only if Card 1 precedes Card 2 in the deck (if Card 2 precedes Card 1, you will miss out on Card 2 while being fixated on Card 1). Then $X_3 = 1$ if and only if Cards 1, 2, and 3 are in that order, and similarly for the other X_j's. Thus,

$$E(X) = 1 + \frac{1}{2!} + \frac{1}{3!} + \cdots + \frac{1}{52!} \approx 1.718.$$

The sum is *extremely close* to $e - 1$, by the Taylor series of e^x and since factorials grow so fast; the difference is about $2 \cdot 10^{-70}$, so $e - 1$ is an incredibly accurate approximation! Note that the answer to (a) does not depend on the number of cards in the deck, and the answer to this part depends on the number of cards only very slightly, as long as there are, say, at least 10 cards.

(c) If j cards have been revealed, then by symmetry, any non-stupid guess for the next card has probability $1/(52 - j)$ of success (for $0 \leq j \leq 51$). Using indicator r.v.s as in (a), the expected value is

$$\frac{1}{52} + \frac{1}{51} + \cdots + \frac{1}{2} + 1 \approx 4.538.$$

The expected value is the sum of the first 52 terms of the harmonic series; for an n-card deck with n large, this is approximately $\log(n) + \gamma$, where $\gamma \approx 0.577$. For $n = 52$, this approximation gives 4.528, which is quite close to the true value. Note that the answers to (a), (b), (c) are in increasing order, which is very sensible since with more information it should be possible to do better (or at least not do worse!).

48. Ⓢ Let X be Hypergeometric with parameters w, b, n.

(a) Find $E\binom{X}{2}$ *by thinking*, without any complicated calculations.

(b) Use (a) to find the variance of X. You should get

$$\text{Var}(X) = \frac{N - n}{N - 1} npq,$$

where $N = w + b, p = w/N, q = 1 - p$.

Solution:

(a) In the story of the Hypergeometric, $\binom{X}{2}$ is the number of pairs of draws such that both balls are white. Creating an indicator r.v. for each pair, we have

$$E\binom{X}{2} = \binom{n}{2}\frac{w}{w+b}\frac{w-1}{w+b-1}.$$

(b) By (a),

$$EX^2 - EX = E(X(X-1)) = n(n-1)p\frac{w-1}{N-1},$$

so

$$\begin{aligned}
\text{Var}(X) &= E(X^2) - (EX)^2 \\
&= n(n-1)p\frac{w-1}{N-1} + np - n^2p^2 \\
&= np\left(\frac{(n-1)(w-1)}{N-1} + 1 - np\right) \\
&= np\left(\frac{nw - w - n + N}{N-1} - \frac{nw}{N}\right) \\
&= np\left(\frac{Nnw - Nw - Nn + N^2 - Nnw + nw}{N(N-1)}\right) \\
&= np\left(\frac{(N-n)(N-w)}{N(N-1)}\right) \\
&= \frac{N-n}{N-1}npq.
\end{aligned}$$

49.   There are n prizes, with values \$1, \$2, …, \$n. You get to choose k random prizes, without replacement. What is the expected total value of the prizes you get?

Hint: Express the total value in the form $a_1I_1 + \cdots + a_nI_n$, where the a_j are constants and the I_j are indicator r.v.s. Or find the expected value of the jth prize received directly.

Solution: Let V be the total value, and I_j be the indicator that the prize worth \$j is obtained. Then $E(I_j) = k/n$, and the total value of the prizes received is $V = I_1 + 2I_2 + \cdots + nI_n$. By linearity,

$$E(V) = \frac{k}{n}(1 + 2 + \cdots + n) = \frac{k}{n} \cdot \frac{n(n+1)}{2} = \frac{k(n+1)}{2}.$$

Alternatively, let V_j be the value of the jth prize received. Then

$$E(V_j) = (1 + 2 + \cdots + n)/n = (n+1)/2,$$

so by linearity we again have $E(V) = k(n+1)/2$.

50.   Ten random chords of a circle are chosen, independently. To generate each of these chords, two independent uniformly random points are chosen on the circle (intuitively, "uniformly" means that the choice is completely random, with no favoritism toward certain angles; formally, it means that the probability of any arc is proportional to the length of that arc). On average, how many pairs of chords intersect?

Hint: Consider two random chords. An equivalent way to generate them is to pick four independent uniformly random points on the circle, and then pair them up randomly.

Solution: Create an indicator r.v. for each of the $\binom{10}{2}$ pairs of chords. For each of those pairs, the probability is $1/3$ that they intersect, by the hint. So by symmetry, linearity, and the fundamental bridge, the expected number of pairs of chords that intersect is

$$\binom{10}{2} \cdot \frac{1}{3} = \frac{10 \cdot 9}{2 \cdot 3} = 15.$$

51. ⓢ A hash table is being used to store the phone numbers of k people, storing each person's phone number in a uniformly random location, represented by an integer between 1 and n (see Exercise 27 from Chapter 1 for a description of hash tables). Find the expected number of locations with no phone numbers stored, the expected number with exactly one phone number, and the expected number with more than one phone number (should these quantities add up to n?).

Solution: Let I_j be an indicator random variable equal to 1 if the jth location is empty, and 0 otherwise, for $1 \leq j \leq n$. Then $P(I_j = 1) = (1 - 1/n)^k$, since the phone numbers are stored in independent random locations. Then $I_1 + \cdots + I_n$ is the number of empty locations. By linearity of expectation, we have

$$E\left(\sum_{j=1}^n I_j\right) = \sum_{j=1}^n E(I_j) = n(1 - 1/n)^k.$$

Similarly, the probability of a specific location having exactly 1 phone number stored is $\frac{k}{n}(1 - \frac{1}{n})^{k-1}$, so the expected number of such locations is $k(1 - 1/n)^{k-1}$. By linearity, the sum of the three expected values is n, so the expected number of locations with more than one phone number is $n - n(1 - 1/n)^k - k(1 - 1/n)^{k-1}$.

52. A coin with probability p of Heads is flipped n times. The sequence of outcomes can be divided into *runs* (blocks of H's or blocks of T's), e.g., $HHHTTHTTTH$ becomes $\boxed{HHH}\,\boxed{TT}\,\boxed{H}\,\boxed{TTT}\,\boxed{H}$, which has 5 runs. Find the expected number of runs.

Hint: Start by finding the expected number of tosses (other than the first) where the outcome is different from the previous one.

Solution: Let I_j be the indicator for the event that position j starts a new run, for $1 \leq j \leq n$. Then $I_1 = 1$ always holds. For $2 \leq j \leq n$, $I_j = 1$ if and only if the jth toss differs from the $(j-1)$st toss. So for $2 \leq j \leq n$,

$$E(I_j) = P((j-1)\text{st toss } H \text{ and } j\text{th toss } T, \text{ or vice versa}) = 2p(1-p).$$

Hence, the expected number of runs is $1 + 2(n-1)p(1-p)$.

53. A coin with probability p of Heads is flipped 4 times. Let X be the number of occurrences of HH (for example, $THHT$ has 1 occurrence and $HHHH$ has 3 occurrences). Find $E(X)$ and $\text{Var}(X)$.

Solution: Let I_j be the indicator of HH appearing at positions $j, j+1$, for $j \in \{1, 2, 3\}$, so $X = I_1 + I_2 + I_3$. By linearity and the fundamental bridge,

$$E(X) = E(I_1) + E(I_2) + E(I_3) = 3p^2.$$

Note that I_3 is independent of I_1, but I_2 is *not* independent of I_1 since if I_1 occurs then there is an H in position 2, which makes it more likely that I_2 occurs. To find the variance of X, let's compute $E(X^2)$. Expanding X^2 and using linearity,

$$\begin{aligned} E(X^2) &= E(I_1^2) + E(I_2^2) + E(I_3^2) + 2E(I_1 I_2) + 2E(I_1 I_3) + 2E(I_2 I_3) \\ &= 3E(I_1) + 2(E(I_1 I_2) + E(I_1 I_3) + E(I_2 I_3)) \\ &= 3p^2 + 2(p^3 + p^4 + p^3) \\ &= 3p^2 + 4p^3 + 2p^4 \end{aligned}$$

since, e.g., $I_1 I_2$ is the indicator of HH appearing in positions 1,2 and in positions 2,3, which is equivalent to H appearing in positions 1,2,3. Therefore,

$$\text{Var}(X) = 3p^2 + 4p^3 + 2p^4 - (3p^2)^2 = 3p^2 + 4p^3 - 7p^4.$$

54. A population has N people, with ID numbers from 1 to N. Let y_j be the value of some numerical variable for person j, and

$$\bar{y} = \frac{1}{N} \sum_{j=1}^{N} y_j$$

be the population average of the quantity. For example, if y_j is the height of person j then $\bar{y}$ is the average height in the population, and if y_j is 1 if person j holds a certain belief and 0 otherwise, then $\bar{y}$ is the proportion of people in the population who hold that belief. In this problem, $y_1, y_2, \ldots, y_n$ are thought of as constants rather than random variables.

A researcher is interested in learning about $\bar{y}$, but it is not feasible to measure y_j for all j. Instead, the researcher gathers a random sample of size n, by choosing people one at a time, with equal probabilities at each stage and without replacement. Let W_j be the value of the numerical variable (e.g., height) for the jth person in the sample. Even though $y_1, \ldots, y_n$ are constants, W_j is a random variable because of the random sampling. A natural way to estimate the unknown quantity $\bar{y}$ is using

$$\bar{W} = \frac{1}{n} \sum_{j=1}^{n} W_j.$$

Show that $E(\bar{W}) = \bar{y}$ in two different ways:

(a) by directly evaluating $E(W_j)$ using symmetry;

(b) by showing that $\bar{W}$ can be expressed as a sum over the population by writing

$$\bar{W} = \frac{1}{n} \sum_{j=1}^{N} I_j y_j,$$

where I_j is the indicator of person j being included in the sample, and then using linearity and the fundamental bridge.

Solution:

(a) By symmetry, $E(W_j) = \bar{y}$, since the jth person in the sample is equally likely to be anyone in the population. So by linearity,

$$E(\bar{W}) = \frac{1}{n} \cdot n\bar{y} = \bar{y}.$$

(b) We can express $\bar{W}$ in the claimed form since $\frac{1}{n} \sum_{j=1}^{N} I_j y_j$ says to add up the y_j's in the sample and then divide by n, which is the same thing that the definition of $\bar{W}$ says to do. By linearity,

$$E(\bar{W}) = \frac{1}{n} \sum_{j=1}^{N} E(I_j) y_j = \frac{1}{n} \sum_{j=1}^{N} E(I_j) y_j = \frac{1}{n} \cdot \frac{n}{N} \cdot N\bar{y} = \bar{y}.$$

55. ⓢ Consider the following algorithm, known as *bubble sort*, for sorting a list of n distinct numbers into increasing order. Initially they are in a random order, with all orders equally likely. The algorithm compares the numbers in positions 1 and 2, and swaps them if needed, then it compares the new numbers in positions 2 and 3, and swaps them if needed, etc., until it has gone through the whole list. Call this one "sweep" through the list. After the first sweep, the largest number is at the end, so the second sweep (if needed) only needs to work with the first $n - 1$ positions. Similarly, the third sweep (if needed) only needs to work with the first $n - 2$ positions, etc. Sweeps are performed until $n - 1$ sweeps have been completed or there is a swapless sweep.

For example, if the initial list is 53241 (omitting commas), then the following 4 sweeps are performed to sort the list, with a total of 10 comparisons:

$$53241 \to 35241 \to 32541 \to 32451 \to 32415.$$

$$32415 \to 23415 \to 23415 \to 23145.$$

$$23145 \to 23145 \to 21345.$$

$$21345 \to 12345.$$

(a) An *inversion* is a pair of numbers that are out of order (e.g., 12345 has no inversions, while 53241 has 8 inversions). Find the expected number of inversions in the original list.

(b) Show that the expected number of comparisons is between $\frac{1}{2}\binom{n}{2}$ and $\binom{n}{2}$.

Hint: For one bound, think about how many comparisons are made if $n-1$ sweeps are done; for the other bound, use Part (a).

Solution:

(a) There are $\binom{n}{2}$ pairs of numbers, each of which is equally likely to be in either order. So by symmetry, linearity, and indicator r.v.s, the expected number of inversions is $\frac{1}{2}\binom{n}{2}$.

(b) Let X be the number of comparisons and V be the number of inversions. On the one hand, $X \geq V$ since every inversion must be repaired. So $E(X) \geq E(V) = \frac{1}{2}\binom{n}{2}$. On the other hand, there are $n-1$ comparisons needed in the first sweep, $n-2$ in the second sweep (if needed), ..., and 1 in the $(n-1)$st sweep (if needed). So

$$X \leq (n-1) + (n-2) + \cdots + 2 + 1 = \frac{n(n-1)}{2} = \binom{n}{2}.$$

Hence, $\frac{1}{2}\binom{n}{2} \leq E(X) \leq \binom{n}{2}$.

56. A certain basketball player practices shooting free throws over and over again. The shots are independent, with probability p of success.

(a) In n shots, what is the expected number of streaks of 7 consecutive successful shots? (Note that, for example, 9 in a row counts as 3 streaks.)

(b) Now suppose that the player keeps shooting until making 7 shots in a row for the first time. Let X be the number of shots taken. Show that $E(X) \leq 7/p^7$.

Hint: Consider the first 7 trials as a block, then the next 7 as a block, etc.

Solution:

(a) If $n < 7$, no streaks of 7 are possible so the expected value is 0. Now assume $n \geq 7$, and create indicator r.v.s $I_1, I_2, \ldots, I_{n-6}$, where I_j is the indicator for there being a streak of 7 starting at the jth shot. By linearity, the expectation is

$$E(I_1 + \cdots + I_{n-6}) = E(I_1) + \cdots + E(I_{n-6}) = (n-6)p^7.$$

(b) As suggested in the hint, treat the first 7 shots as the first block, the next 7 shots as the second block, and in general, the $(7k-6)$th through $(7k)$th shots as the kth block. Now let J_k be the indicator of the shots in the kth block all going in. Let Y be the first time k at which $J_k = 1$. Note that $X \leq 7Y$ since $J_k = 1$ implies a streak of 7 at the kth block (X may not equal $7Y$ since there could be a streak of 7 that crosses block boundaries). Then $Y \sim \text{FS}(p^7)$, so $E(X) \leq 7E(Y) = 7/p^7$.

57. ⓢ An urn contains red, green, and blue balls. Balls are chosen randomly with replacement (each time, the color is noted and then the ball is put back). Let r, g, b be the probabilities of drawing a red, green, blue ball, respectively ($r + g + b = 1$).

(a) Find the expected number of balls chosen before obtaining the first red ball, not including the red ball itself.

(b) Find the expected number of different *colors* of balls obtained before getting the first red ball.

(c) Find the probability that at least 2 of n balls drawn are red, given that at least 1 is red.

Solution:

(a) The distribution is Geom(r), so the expected value is $\frac{1-r}{r}$.

(b) Use indicator random variables: let I_1 be 1 if green is obtained before red, and 0 otherwise, and define I_2 similarly for blue. Then

$$E(I_1) = P(\text{green before red}) = \frac{g}{g + r}$$

since "green before red" means that the first nonblue ball is green. Similarly, $E(I_2) = b/(b + r)$, so the expected number of colors obtained before getting red is

$$E(I_1 + I_2) = \frac{g}{g + r} + \frac{b}{b + r}.$$

(c) By definition of conditional probability,

$$P(\text{at least 2 red}|\text{ at least 1 red}) = \frac{P(\text{at least 2 red})}{P(\text{at least 1 red})} = \frac{1 - (1 - r)^n - nr(1 - r)^{n-1}}{1 - (1 - r)^n}.$$

58. ⓢ Job candidates $C_1, C_2, \ldots$ are interviewed one by one, and the interviewer compares them and keeps an updated list of rankings (if n candidates have been interviewed so far, this is a list of the n candidates, from best to worst). Assume that there is no limit on the number of candidates available, that for any n the candidates $C_1, C_2, \ldots, C_n$ are equally likely to arrive in any order, and that there are no ties in the rankings given by the interview.

Let X be the index of the first candidate to come along who ranks as better than the very first candidate C_1 (so C_X is better than C_1, but the candidates after 1 but prior to X (if any) are worse than C_1. For example, if C_2 and C_3 are worse than C_1 but C_4 is better than C_1, then $X = 4$. All 4! orderings of the first 4 candidates are equally likely, so it could have happened that the first candidate was the best out of the first 4 candidates, in which case $X > 4$.

What is $E(X)$ (which is a measure of how long, on average, the interviewer needs to wait to find someone better than the very first candidate)?

Hint: Find $P(X > n)$ by interpreting what $X > n$ says about how C_1 compares with other candidates, and then apply the result of Theorem 4.4.8.

Solution: For $n \geq 2$, $P(X > n)$ is the probability that none of $C_2, C_3, \ldots, C_n$ are better candidates than C_1, i.e., the probability that the first candidate is the highest ranked out of the first n. Since any ordering of the first n candidates is equally likely, each of the first n is equally likely to be the highest ranked of the first n, so $P(X > n) = 1/n$. For $n = 0$ or $n = 1$, $P(X > n) = 1$ (note that it does not make sense to say the probability is $1/n$ when $n = 0$). By Theorem 4.4.8,

$$E(X) = \sum_{n=0}^{\infty} P(X > n) = P(X > 0) + \sum_{n=1}^{\infty} P(X > n) = 1 + \sum_{n=1}^{\infty} \frac{1}{n} = \infty$$

since the series is the *harmonic series*, which diverges.

How can the average waiting time to find someone better than the first candidate be infinite? In the real world, there are always only finitely many candidates so the expected waiting time is finite, just as in the St. Petersburg paradox there must in reality be an upper bound on the number of rounds. The harmonic series diverges very slowly, so even with millions of job candidates the average waiting time would not be very large.

59. People are arriving at a party one at a time. While waiting for more people to arrive they entertain themselves by comparing their birthdays. Let X be the number of people needed to obtain a birthday match, i.e., before person X arrives there are no two people with the same birthday, but when person X arrives there is a match.

Assume for this problem that there are 365 days in a year, all equally likely. By the result of the birthday problem from Chapter 1, for 23 people there is a 50.7% chance of a birthday match (and for 22 people there is a less than 50% chance). But this has to do with the *median* of X (defined below); we also want to know the *mean* of X, and in this problem we will find it, and see how it compares with 23.

(a) A *median* of an r.v. Y is a value m for which $P(Y \leq m) \geq 1/2$ and $P(Y \geq m) \geq 1/2$ (this is also called a median of the *distribution* of Y; note that the notion is completely determined by the CDF of Y). Every distribution has a median, but for some distributions it is not unique. Show that 23 is the *unique* median of X.

(b) Show that $X = I_1 + I_2 + \cdots + I_{366}$, where I_j is the indicator r.v. for the event $X \geq j$. Then find $E(X)$ in terms of p_j's defined by $p_1 = p_2 = 1$ and for $3 \leq j \leq 366$,

$$p_j = \left(1 - \frac{1}{365}\right)\left(1 - \frac{2}{365}\right)\cdots\left(1 - \frac{j-2}{365}\right).$$

(c) Compute $E(X)$ numerically. In R, the pithy command `cumprod(1-(0:364)/365)` produces the vector $(p_2, \ldots, p_{366})$.

(d) Find the variance of X, both in terms of the p_j's and numerically.

Hint: What is I_i^2, and what is $I_i I_j$ for $i < j$? Use this to simplify the expansion

$$X^2 = I_1^2 + \cdots + I_{366}^2 + 2\sum_{j=2}^{366}\sum_{i=1}^{j-1} I_i I_j.$$

Note: In addition to being an entertaining game for parties, the birthday problem has many applications in computer science, such as in a method called the *birthday attack* in cryptography. It can be shown that if there are n days in a year and n is large, then $E(X) \approx \sqrt{\pi n/2}$. In Volume 1 of his masterpiece *The Art of Computer Programming*, Don Knuth shows that an even better approximation is

$$E(X) \approx \sqrt{\frac{\pi n}{2}} + \frac{2}{3} + \sqrt{\frac{\pi}{288n}}.$$

Solution:

(a) The event $X > 23$ occurring means exactly that among the first 23 people, there is no birthday match. So $P(X > 23) < 1/2$ and, similarly, $P(X > 22) > 1/2$. But $P(X \leq 23) = 1 - P(X > 23) > 1/2$ and $P(X \geq 23) = P(X > 22) > 1/2$, so 23 is a median of X. It is unique since for $m < 23$, $P(X \leq m) \leq P(X < 23) = 1 - P(X \geq 23) < 1/2$ and for $m > 23$, $P(X \geq m) \leq P(X > 23) < 1/2$.

(b) The possible values of X are $2, 3, \ldots, 366$ (for 366 people there must be a match since then there are more people than days of the year, excluding February 29; this is

called the *pigeonhole principle*). Then $X = I_1 + I_2 + \cdots + I_{366}$ since if $X = k$, then the first k indicators will be 1's and the rest will be 0's.

The event $X \geq j$ is the same as the event $X > j - 1$. So

$$E(I_j) = P(X > j - 1) = P(\text{no match in a group of } j - 1 \text{ people}) = p_j.$$

Thus,

$$E(X) = p_1 + p_2 + \cdots + p_{366}.$$

(Note: this method of finding $E(X)$ is a version of the result from SP 4 #1.5.)

(c) Entering `1 + sum(cumprod(1-(0:364)/365))` in R yields $E(X) \approx 24.617$.

(d) First note that $I_i^2 = I_i$ (this always true for indicator r.v.s) and that $I_i I_j = I_j$ for $i < j$ (since it is the indicator of $\{X \geq i\} \cap \{X \geq j\}$). Therefore,

$$X^2 = I_1 + \cdots + I_{366} + 2\sum_{j=2}^{366}(j-1)I_j,$$

$$E(X^2) = p_1 + \cdots + p_{366} + 2\sum_{j=2}^{366}(j-1)p_j$$

$$= \sum_{j=1}^{366}(2j-1)p_j,$$

and the variance is

$$\mathrm{Var}(X) = E(X^2) - (EX)^2 = \sum_{j=1}^{366}(2j-1)p_j - \left(\sum_{j=1}^{366}p_j\right)^2.$$

Entering `p <- c(1,cumprod(1-(0:364)/365)); sum((2*(1:366)-1)*p)-(sum(p))^2` in R yields $\mathrm{Var}(X) \approx 148.640$.

60. Elk dwell in a certain forest. There are N elk, of which a simple random sample of size n is captured and tagged (so all $\binom{N}{n}$ sets of n elk are equally likely). The captured elk are returned to the population, and then a new sample is drawn. This is an important method that is widely used in ecology, known as *capture-recapture*. If the new sample is also a simple random sample, with some fixed size, then the number of tagged elk in the new sample is Hypergeometric.

For this problem, assume that instead of having a fixed sample size, elk are sampled one by one without replacement until m tagged elk have been recaptured, where m is specified in advance (of course, assume that $1 \leq m \leq n \leq N$). An advantage of this sampling method is that it can be used to avoid ending up with a very small number of tagged elk (maybe even zero), which would be problematic in many applications of capture-recapture. A disadvantage is not knowing how large the sample will be.

(a) Find the PMFs of the number of untagged elk in the new sample (call this X) and of the total number of elk in the new sample (call this Y).

(b) Find the expected sample size EY using symmetry, linearity, and indicator r.v.s.

Hint: We can assume that even after getting m tagged elk, they continue to be captured until all N of them have been obtained; briefly explain why this can be assumed. Express $X = X_1 + \cdots + X_m$, where X_1 is the number of untagged elk before the first tagged elk, X_2 is the number between the first and second tagged elk, etc. Then find EX_j by creating the relevant indicator r.v. for each untagged elk in the population.

(c) Suppose that m, n, N are such that EY is an integer. If the sampling is done with

a fixed sample size equal to EY rather than sampling until exactly m tagged elk are obtained, find the expected number of tagged elk in the sample. Is it less than m, equal to m, or greater than m (for $n < N$)?

Solution:

(a) The event $X = k$ says that there are $m - 1$ tagged elk and k untagged elk in the first $m + k - 1$ elk sampled, and that the $(m + k)$th elk sampled is tagged. So

$$P(X = k) = \frac{\binom{n}{m-1}\binom{N-n}{k}}{\binom{N}{m+k-1}} \cdot \frac{n - m + 1}{N - m - k + 1},$$

for $k = 0, 1, \ldots, N - n$ (note that $k = 0$ is the case where the first m elk sampled are all tagged, and $k = N - n$ is the case where we have to collect *all* the untagged elk before recapturing a tagged elk). This is known as the *Negative Hypergeometric* distribution. The PMF of Y can then be found by noting that $Y = X + m$: for $y = m, m + 1, \ldots, N - n + m$,

$$P(Y = y) = P(X = y - m) = \frac{\binom{n}{m-1}\binom{N-n}{y-m}}{\binom{N}{y-1}} \cdot \frac{n - m + 1}{N - y + 1}.$$

(b) As suggested in the hint, assume that the elk get captured until all N of them have been obtained. This is convenient since then we are just looking at a random permutation of the N elk, and it is valid since what transpires after m tagged elk have been recaptured does not affect the value of X. Define $X_1, \ldots, X_m$ as in the hint. Label the untagged elk as $1, 2, \ldots, N - n$ and write $X_1 = I_1 + \cdots + I_{N-n}$, where I_j is the indicator of Untagged Elk j being captured before any tagged elk.

By symmetry, $E(I_j) = 1/(n + 1)$ since Untagged Elk j and the n tagged elk are equally likely to be in any order. So $E(X_1) = (N - n)/(n + 1)$. For example, for $N = 10$ elk with $n = 4$ tagged, labeled $7, 8, 9, 10$ and $N - n = 6$ untagged, labeled $1, 2, 3, 4, 5, 6$, and with $m = 3$, the observed evidence (if all elk are collected) could be

$$\underbrace{❺❷❸}_{X_1} ⑨ \; \underbrace{❻}_{X_2} \; ⑦ \underbrace{❹❶}_{X_3} ⑩⑧.$$

The observed values of I_5, I_2, I_3 are 1 and of I_1, I_4, I_6 are 0. Before the data are collected, $E(I_5) = 1/5$ since Untagged Elk 5 and Tagged Elk $7, 8, 9, 10$ are equally likely to be in any order.

Similarly, $E(X_j) = (N - n)/(n + 1)$ for all $j = 1, \ldots, m$ since each untagged elk is equally likely to be positioned anywhere among the n tagged elk. Thus,

$$E(X) = \frac{m(N - n)}{n + 1}, E(Y) = m + \frac{m(N - n)}{n + 1} = \frac{m(N + 1)}{n + 1}.$$

(c) With a fixed sample size equal to EY, the number of tagged elk in the sample is Hypergeometric with mean

$$\frac{m(N + 1)}{n + 1} \cdot \frac{n}{N} = m \cdot \frac{1 + \frac{1}{N}}{1 + \frac{1}{n}} < m.$$

If n is small and N is large, then this is a major difference between the two sampling methods; if n is large, then the above expectation is approximately m.

LOTUS

61. ⓢ For $X \sim \text{Pois}(\lambda)$, find $E(X!)$ (the average factorial of X), if it is finite.

    Solution: By LOTUS,
    $$E(X!) = e^{-\lambda} \sum_{k=0}^{\infty} k! \frac{\lambda^k}{k!} = \frac{e^{-\lambda}}{1-\lambda},$$
    for $0 < \lambda < 1$ since this is a geometric series (and $E(X!)$ is infinite if $\lambda \geq 1$).

62. For $X \sim \text{Pois}(\lambda)$, find $E(2^X)$, if it is finite.

    Solution: By LOTUS,
    $$E(2^X) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(2\lambda)^k}{k!} = e^{-\lambda} e^{2\lambda} = e^{\lambda},$$
    for all $\lambda > 0$.

63. For $X \sim \text{Geom}(p)$, find $E(2^X)$ (if it is finite) and $E(2^{-X})$ (if it is finite). For each, make sure to clearly state what the values of p are for which it is finite.

    Solution: Let $q = 1 - p$. By LOTUS and the formula for the sum of a geometric series,
    $$E(2^X) = p \sum_{k=0}^{\infty} (2q)^k = \frac{p}{1-2q},$$
    for $2q < 1$ (which is equivalent to $p > 1/2$). Similarly,
    $$E(2^{-X}) = p \sum_{k=0}^{\infty} (q/2)^k = \frac{p}{1-q/2},$$
    for all $p \in (0,1)$ (since we automatically have $q/2 \leq 1/2 < 1$).

64. ⓢ Let $X \sim \text{Geom}(p)$ and let t be a constant. Find $E(e^{tX})$, as a function of t (this is known as the *moment generating function*; we will see in Chapter 6 how this function is useful).

    Solution: Letting $q = 1 - p$, we have
    $$E(e^{tX}) = p \sum_{k=0}^{\infty} e^{tk} q^k = p \sum_{k=0}^{\infty} (qe^t)^k = \frac{p}{1-qe^t},$$
    for $qe^t < 1$ (while for $qe^t \geq 1$, the series diverges).

65. ⓢ The number of fish in a certain lake is a $\text{Pois}(\lambda)$ random variable. Worried that there might be no fish at all, a statistician adds one fish to the lake. Let Y be the resulting number of fish (so Y is 1 plus a $\text{Pois}(\lambda)$ random variable).

    (a) Find $E(Y^2)$.

    (b) Find $E(1/Y)$.

    Solution:

    (a) We have $Y = X + 1$ with $X \sim \text{Pois}(\lambda)$, so $Y^2 = X^2 + 2X + 1$. So
    $$E(Y^2) = E(X^2 + 2X + 1) = E(X^2) + 2E(X) + 1 = (\lambda + \lambda^2) + 2\lambda + 1 = \lambda^2 + 3\lambda + 1,$$
    since $E(X^2) = \text{Var}(X) + (EX)^2 = \lambda + \lambda^2$.

(b) By LOTUS,

$$E\left(\frac{1}{Y}\right) = E\left(\frac{1}{X+1}\right) = \sum_{k=0}^{\infty} \frac{1}{k+1} e^{-\lambda} \frac{\lambda^k}{k!}.$$

Using the identity $k!(k+1) = (k+1)!$ and the Taylor series for e^λ, this becomes

$$e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{(k+1)!} = \frac{e^{-\lambda}}{\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{k+1}}{(k+1)!} = \frac{e^{-\lambda}}{\lambda}(e^\lambda - 1) = \frac{1}{\lambda}(1 - e^{-\lambda}).$$

66. Ⓢ Let X be a Pois(λ) random variable, where λ is fixed but unknown. Let $\theta = e^{-3\lambda}$, and suppose that we are interested in estimating θ based on the data. Since X is what we observe, our estimator is a function of X, call it $g(X)$. The *bias* of the estimator $g(X)$ is defined to be $E(g(X)) - \theta$, i.e., how far off the estimate is on average; the estimator is *unbiased* if its bias is 0.

(a) For estimating λ, the r.v. X itself is an unbiased estimator. Compute the bias of the estimator $T = e^{-3X}$. Is it unbiased for estimating θ?

(b) Show that $g(X) = (-2)^X$ is an unbiased estimator for θ. (In fact, it turns out to be the only unbiased estimator for θ.)

(c) Explain intuitively why $g(X)$ is a silly choice for estimating θ, despite (b), and show how to improve it by finding an estimator $h(X)$ for θ that is always at least as good as $g(X)$ and sometimes strictly better than $g(X)$. That is,

$$|h(X) - \theta| \le |g(X) - \theta|,$$

with the inequality sometimes strict.

Solution:

(a) The estimator is biased, with bias given by

$$E(e^{-3X}) - \theta = \sum_{k=0}^{\infty} e^{-3k} \frac{\lambda^k}{k!} e^{-\lambda} - e^{-3\lambda}$$

$$= e^{-\lambda} \sum_{k=0}^{\infty} \frac{(e^{-3}\lambda)^k}{k!} - e^{-3\lambda}$$

$$= e^{-\lambda} e^{e^{-3}\lambda} - e^{-3\lambda}$$

$$= e^{-3\lambda}(e^{(2+e^{-3})\lambda} - 1) \ne 0.$$

(b) The estimator $g(X) = (-2)^X$ is unbiased since

$$E(-2)^X - \theta = \sum_{k=0}^{\infty} (-2)^k \frac{\lambda^k}{k!} e^{-\lambda} - e^{-3\lambda}$$

$$= e^{-\lambda} e^{-2\lambda} - e^{-3\lambda} = 0.$$

(c) The estimator $g(X)$ is silly in the sense that it is sometimes negative, whereas $e^{-3\lambda}$ is positive. One simple way to get a better estimator is to modify $g(X)$ to make it nonnegative, by letting $h(X) = 0$ if $g(X) < 0$ and $h(X) = g(X)$ otherwise.

Better yet, note that $e^{-3\lambda}$ is between 0 and 1 since $\lambda > 0$, so letting $h(X) = 0$ if $g(X) < 0$ and $h(X) = 1$ if $g(X) > 0$ is clearly more sensible than using $g(X)$.

Poisson approximation

67. ⓢ Law school courses often have assigned seating to facilitate the Socratic method. Suppose that there are 100 first-year law students, and each takes the same two courses: Torts and Contracts. Both are held in the same lecture hall (which has 100 seats), and the seating is uniformly random and independent for the two courses.

(a) Find the probability that no one has the same seat for both courses (exactly; you should leave your answer as a sum).

(b) Find a simple but accurate approximation to the probability that no one has the same seat for both courses.

(c) Find a simple but accurate approximation to the probability that at least two students have the same seat for both courses.

Solution:

(a) Let N be the number of students in the same seat for both classes. The problem has essentially the same structure as the matching problem. Let E_j be the event that the jth student sits in the same seat in both classes. Then

$$P(N = 0) \;=\; 1 - P\left(\bigcup_{j=1}^{100} E_j\right).$$

By symmetry, inclusion-exclusion gives

$$P\left(\bigcup_{j=1}^{100} E_j\right) \;=\; \sum_{j=1}^{100}(-1)^{j-1}\binom{100}{j}P\left(\bigcap_{k=1}^{j} E_k\right).$$

The j-fold intersection represents j particular students sitting pat throughout the two lectures, which occurs with probability $(100-j)!/100!$. So

$$P\left(\bigcup_{j=1}^{100} E_j\right) \;=\; \sum_{j=1}^{100}(-1)^{j-1}\binom{100}{j}\frac{(100-j)!}{100!} = \sum_{j=1}^{100}(-1)^{j-1}/i!$$

$$P(N = 0) \;=\; 1 - \sum_{j=1}^{100}\frac{(-1)^{j-1}}{j!} = \sum_{j=0}^{100}\frac{(-1)^{j}}{j!}.$$

(b) Define I_i to be the indicator for student i having the same seat in both courses, so that $N = \sum_{i=1}^{100} I_i$. Then $P(I_i = 1) = 1/100$, and the I_i are weakly dependent because

$$P((I_i = 1) \cap (I_j = 1)) = \left(\frac{1}{100}\right)\left(\frac{1}{99}\right) \;\approx\; \left(\frac{1}{100}\right)^2 = P(I_i = 1)P(I_j = 1).$$

So N is close to $\text{Pois}(\lambda)$ in distribution, where $\lambda = E(N) = 100E(I_1) = 1$. Thus,

$$P(N = 0) \;\approx\; e^{-1}1^0/0! = e^{-1} \approx 0.37.$$

This agrees with the result of (a), which we recognize as the Taylor series for e^x, evaluated at $x = -1$.

(c) Using a Poisson approximation, we have

$$P(N \geq 2) \;=\; 1 - P(N = 0) - P(N = 1) \approx 1 - e^{-1} - e^{-1} = 1 - 2e^{-1} \approx 0.26.$$

68. ⓢ A group of n people play "Secret Santa" as follows: each puts their name on a slip of paper in a hat, picks a name randomly from the hat (without replacement), and then buys a gift for that person. Unfortunately, they overlook the possibility of drawing one's

own name, so some may have to buy gifts for themselves (on the bright side, some may like self-selected gifts better). Assume $n \geq 2$.

(a) Find the expected value of the number X of people who pick their own names.

(b) Find the expected number of pairs of people, A and B, such that A picks B's name and B picks A's name (where $A \neq B$ and order doesn't matter).

(c) Let X be the number of people who pick their own names. What is the *approximate* distribution of X if n is large (specify the parameter value or values)? What does $P(X = 0)$ converge to as $n \to \infty$?

Solution:

(a) Let I_j be the indicator r.v. for the jth person picking their own name. Then $E(I_j) = P(I_j = 1) = \frac{1}{n}$. By linearity, the expected number is $n \cdot E(I_j) = 1$.

(b) Let I_{ij} the the indicator r.v. for the ith and jth persons having such a swap (for $i < j$). Then $E(I_{ij}) = P(i \text{ picks } j)P(j \text{ picks } i|i \text{ picks } j) = \frac{1}{n(n-1)}$.

Alternatively, we can get this by counting: there are $n!$ permutations for who picks whom, of which $(n-2)!$ have i pick j and j pick i, giving $\frac{(n-2)!}{n!} = \frac{1}{n(n-1)}$. So by linearity, the expected number is $\binom{n}{2} \cdot \frac{1}{n(n-1)} = \frac{1}{2}$.

(c) By the Poisson paradigm, X is approximately Pois(1) for large n. As $n \to \infty$, $P(X = 0) \to 1/e$, which is the probability of a Pois(1) r.v. being 0.

69. A survey is being conducted in a city with a million (10^6) people. A sample of size 1000 is collected by choosing people in the city at random, *with* replacement and with equal probabilities for everyone in the city. Find a simple, accurate approximation to the probability that at least one person will get chosen more than once (in contrast, Exercise 26 from Chapter 1 asks for an exact answer).

Hint: Indicator r.v.s are useful here, but creating 1 indicator for each of the million people is *not* recommended since it leads to a messy calculation. Feel free to use the fact that $999 \approx 1000$.

Solution: Let I_{ij} be the indicator of the ith and jth sampled people being the same person, for $1 \leq i < j \leq 10^3$, and let X be the sum of the I_{ij}. We want to approximate $P(X \geq 1)$. By symmetry and linearity,

$$E(X) = \binom{1000}{2} \frac{1}{10^6} = \frac{1000 \cdot 999}{2 \cdot 10^6} \approx \frac{1}{2}.$$

By the Poisson paradigm, X is approximately Pois(λ) with $\lambda = E(X)$. So

$$P(X \geq 1) = 1 - P(X = 0) \approx 1 - e^{-\lambda} = 1 - e^{-1/2} \approx 0.393.$$

The true value is also 0.393 (to three decimal places), so the approximation is very accurate. Remarkably, there is almost a 40% chance of sampling someone twice, even though the sample size is only 0.1% of the population size!

70. Ⓢ Ten million people enter a certain lottery. For each person, the chance of winning is one in ten million, independently.

(a) Find a simple, good approximation for the PMF of the number of people who win the lottery.

(b) Congratulations! You won the lottery. However, there may be other winners. Assume now that the number of winners other than you is $W \sim$ Pois(1), and that if there is

more than one winner, then the prize is awarded to one randomly chosen winner. Given this information, find the probability that you win the prize (simplify).

Solution:

(a) Let X be the number of people who win. Then

$$E(X) = \frac{10^7}{10^7} = 1.$$

A Poisson approximation is very good here since X is the number of "successes" for a very large number of independent trials where the probability of success on each trial is very low. So X is approximately Pois(1), and for k a nonnegative integer,

$$P(X = k) \approx \frac{1}{e \cdot k!}.$$

(b) Let A be the event that you win the prize, and condition on W:

$$P(A) = \sum_{k=0}^{\infty} P(A|W = k)P(W = k) = \frac{1}{e} \sum_{k=0}^{\infty} \frac{1}{k+1} \frac{1}{k!} = \frac{1}{e} \sum_{k=0}^{\infty} \frac{1}{(k+1)!} = \frac{e-1}{e} = 1 - \frac{1}{e}.$$

71. In a group of 90 people, find a simple, good approximation for the probability that there is at least one pair of people such that they share a birthday *and* their biological mothers share a birthday. Assume that no one among the 90 people is the biological mother of another one of the 90 people, nor do two of the 90 people have the same biological mother. Express your answer as a fully simplified fraction in the form a/b, where a and b are positive integers and $b \leq 100$.

Make the usual assumptions as in the birthday problem. To simplify the calculation, you can use the approximations $365 \approx 360$ and $89 \approx 90$, and the fact that $e^x \approx 1 + x$ for $x \approx 0$.

Solution: Let X be the number of pairs of people for which this coincidence occurs. Create an indicator r.v. for each pair, indicating whether this coincidence occurs. By symmetry, linearity, and the fundamental bridge,

$$E(X) = \binom{90}{2} \cdot \frac{1}{365^2} = \frac{90 \cdot 89}{2 \cdot 365^2} \approx \frac{90^2}{2 \cdot 360^2} = \frac{1}{32}.$$

A Poisson approximation for X makes sense since for each pair this coincidence is very unlikely, but there are a lot of pairs, and the indicator r.v.s are approximately independent. So X is approximately Pois(λ), where $\lambda = E(X)$. Then

$$P(X \geq 1) = 1 - P(X = 0) \approx 1 - e^{-\lambda} \approx 1 - (1 - \lambda) = \lambda \approx \frac{1}{32}.$$

(The above approximation turns out to be quite accurate: as a decimal to 2 significant digits it is 0.031, while the R command `pbirthday(90,classes=365^2)` gives that the true probability, to 2 significant digits, is 0.030.)

72. Use Poisson approximations to investigate the following types of coincidences. The usual assumptions of the birthday problem apply.

(a) How many people are needed to have a 50% chance that at least one of them has the same birthday as *you*?

(b) How many people are needed to have a 50% chance that there is at least one pair of people who not only were born on the same day of the year, but also were born at the

same *hour* (e.g., two people born between 2 pm and 3 pm are considered to have been born at the same hour)?

(c) Considering that only $1/24$ of pairs of people born on the same day were born at the same hour, why isn't the answer to (b) approximately $24 \cdot 23$?

(d) With 100 people, there is a 64% chance that there is at least one set of 3 people with the same birthday (according to R, using `pbirthday(100,classes=365,coincident=3)` to compute it). Provide two different Poisson approximations for this value, one based on creating an indicator r.v. for each triplet of people, and the other based on creating an indicator r.v. for each day of the year. Which is more accurate?

Solution:

(a) Let k be the number of people, and for each create an indicator for that person having the same birthday as you. These indicator r.v.s are independent, each with probability $1/365$ of being 1. By Poisson approximation, the probability of at least one match is approximately $1 - e^{-k/365}$. The smallest k for which this is at least 0.5 is

$$k = 253.$$

This turns out to be *exactly* the right number, as seen using the exact probability of at least one match, which is $1 - (1 - 1/365)^k$. Another way to get the same approximation is to use the fact that $\log(1 + x) \approx x$ for small x to write

$$(1 - 1/365)^k = e^{k \log(1 - 1/365)} \approx e^{-k/365}.$$

(b) This is the birthday problem, with $365 \cdot 24$ categories of people rather than 365. A Poisson approximation gives $1 - e^{-\binom{k}{2}/(365 \cdot 24)}$ as the approximate probability of at least one match. The smallest k for which this is at least 0.5 is

$$k = 111.$$

This again turns out to be exactly right, as seen using `qbirthday(0.5,classes=365*24)` to compute it in R.

(c) What matters is the number of *pairs* of people, which grows quadratically as a function of the number of people. Let k be the number of people in the birthday-birthhour problem and m be the number of people in the birthday problem. To obtain about the same probability p for both problems, we need

$$e^{-\binom{k}{2}/(365 \cdot 24)} \approx e^{-\binom{m}{2}/365},$$

which reduces to

$$k \approx \sqrt{24}m,$$

using the approximation $\binom{k}{2} = k(k-1)/2 \approx k^2/2$ and similarly for $\binom{m}{2}$. For $p = 0.5$, this correctly suggests computing $\sqrt{24} \cdot 23 \approx 113$ rather than $24 \cdot 23 = 552$ as an approximate way to convert from the birthday problem to the birthday-birthhour problem.

(d) Creating an indicator for each triplet of people gives

$$1 - e^{-\binom{100}{3}/365^2} \approx 0.70$$

as the Poisson approximation for the probability of at least one triple match. An alternative method is to create an indicator for each day of the year: let I_j be the indicator for at least 3 people having been born on the jth day of the year, and $X = I_1 + \cdots + I_{365}$. Then $X = 0$ is the event that there is no triple birthday match. We have

$$E(I_j) = P(I_j = 1) = 1 - \left(\frac{364}{365}\right)^{100} - 100 \cdot \left(\frac{364}{365}\right)^{99} \cdot \left(\frac{1}{365}\right) - \binom{100}{2} \cdot \left(\frac{364}{365}\right)^{98} \cdot \left(\frac{1}{365}\right)^2,$$

and by linearity $E(X) = 365E(I_1)$. We then have the Poisson approximation

$$P(X > 0) = 1 - P(X = 0) \approx 1 - e^{-E(X)} \approx 0.63.$$

The latter is closer to 0.64 (the correct value, as stated in the problem and found using `pbirthday(100,coincident=3)` in R). An intuitive explanation for why the former approximation is less accurate is that there is a more substantial dependence in the indicators in that method: note that if persons 1, 2, and 3 all have the same birthday, say October 31, then if person 4 is born on October 31 that will automatically result in the triplets $\{1, 2, 4\}$, $\{1, 3, 4\}$, and $\{2, 3, 4\}$ also being triple birthday matches.

73. A chess tournament has 100 players. In the first round, they are randomly paired to determine who plays whom (so 50 games are played). In the second round, they are again randomly paired, independently of the first round. In both rounds, all possible pairings are equally likely. Let X be the number of people who play against the same opponent twice.

(a) Find the expected value of X.

(b) Explain why X is *not* approximately Poisson.

(c) Find good approximations to $P(X = 0)$ and $P(X = 2)$, by thinking about games in the second round such that the same pair played each other in the first round.

Solution:

(a) Label the players as $1, 2, \ldots, 100$, and let I_j be the indicator of player j having the same opponent twice. We have $P(I_1 = 1) = 99/99^2 = 1/99$. Then by symmetry, linearity, and the fundamental bridge,

$$E(X) = 100E(I_1) = 100P(I_1 = 1) = 100/99.$$

(b) The possible values of X are $0, 2, 4, \ldots, 100$; we have that X must be even since if Alice plays the same opponent twice, say Bob, then Bob also plays the same opponent twice. A Poisson distribution has possible values $0, 1, 2, 3, \ldots$, which does not make sense as an approximation for an r.v. that must be even.

(c) Let G be the number of games in the second round such that the same pair of opponents played each other in the first round; note that $G = X/2$. Let $J_1, \ldots, J_{50}$ be indicator r.v.s where J_i is the indicator of the ith game in round 2 having the same pair as a round 1 game, with respect to a pre-determined way to order games (e.g., in increasing order of the smaller of the two player IDs). A Poisson approximation for G *does* make sense since the J_i are weakly dependent and $P(J_i = 1) = 1/99$ is small. So G is approximately Pois(50/99). This gives

$$P(X = 0) = P(G = 0) \approx e^{-50/99} \approx 0.6035,$$
$$P(X = 2) = P(G = 1) \approx e^{-50/99}(50/99) \approx 0.3048.$$

Simulating this in R by generating a million random round 2 pairings (by symmetry, round 1 can be taken to be 1 plays 2, 3 plays 4, etc.) gave $P(X = 0) \approx 0.6034, P(X = 2) \approx 0.3050$, agreeing very well with the Poisson approximation.

*Existence

74. Ⓢ Each of 111 people names their 5 favorite movies out of a list of 11 movies.

(a) Alice and Bob are 2 of the 111 people. Assume *for this part only* that Alice's 5 favorite movies out of the 11 are random, with all sets of 5 equally likely, and likewise

138

for Bob, independently. Find the expected number of movies in common to Alice's and Bob's lists of favorite movies.

(b) Show that there are 2 movies such that at least 21 of the people name both of these movies as favorites.

Solution:

(a) Let I_j be the indicator for the jth movie being on both lists, for $1 \leq j \leq 11$. By symmetry and linearity, the desired expected value is

$$11 \left(\frac{5}{11} \right)^2 = \frac{25}{11}.$$

(b) Choose 2 *random* movies (one at a time, without replacement). Let X be the number of people who name both of these movies. Creating an indicator r.v. for each person,

$$E(X) = 111 P(\text{Alice names both random movies}) = 111 \left(\frac{5}{11} \cdot \frac{4}{10} \right) = \left(\frac{111}{110} \right) 20 > 20,$$

since the first chosen movie has a 5/11 chance of being on Alice's list and given that it is, the second chosen movie has a 4/10 chance of being on the list (or we can use the Hypergeometric PMF after "tagging" Alice's favorite movies).

Thus, there must exist 2 movies such that at least 21 of the people name both of them as favorites.

75. Ⓢ The circumference of a circle is colored with red and blue ink such that 2/3 of the circumference is red and 1/3 is blue. Prove that no matter how complicated the coloring scheme is, there is a way to inscribe a square in the circle such that at least three of the four corners of the square touch red ink.

Solution: Consider a random square, obtained by picking a uniformly random point on the circumference and inscribing a square with that point a corner; say that the corners are $U_1, \ldots, U_4$, in clockwise order starting with the initial point chosen. Let I_j be the indicator r.v. of U_j touching red ink. By symmetry, $E(I_j) = 2/3$ so by linearity, the expected number of corners touching red ink is 8/3. Thus, there must exist an inscribed square with at least 8/3 of its corners touching red ink. Such a square must have at least 3 of its corners touching red ink.

76. Ⓢ A hundred students have taken an exam consisting of 8 problems, and for each problem at least 65 of the students got the right answer. Show that there exist two students who collectively got everything right, in the sense that for each problem, at least one of the two got it right.

Solution: Say that the "score" of a pair of students is how many problems at least one of them got right. The expected score of a random pair of students (with all pairs equally likely) is at least $8(1 - 0.35^2) = 7.02$, as seen by creating an indicator r.v. for each problem for the event that at least one student in the pair got it right. (We can also improve the 0.35^2 to $\frac{35}{100} \cdot \frac{34}{99}$ since the students are sampled without replacement.) So some pair of students must have gotten a score of at least 7.02, which means that they got a score of at least 8.

77. Ⓢ Ten points in the plane are designated. You have ten circular coins (of the same radius). Show that you can position the coins in the plane (without stacking them) so that all ten points are covered.

Hint: Consider a *honeycomb tiling* of the plane (this is a way to divide the plane into hexagons). You can use the fact from geometry that if a circle is inscribed in a hexagon then the ratio of the area of the circle to the area of the hexagon is $\frac{\pi}{2\sqrt{3}} > 0.9$.

Solution: Take a uniformly random honeycomb tiling (to do this, start with any honeycomb tiling and then shift it horizontally and vertically by uniformly random amounts; by periodicity there is an upper bound on how large the shifts need to be). Choose the tiling so that a circle the same size as one of the coins can be inscribed in each hexagon. Then inscribe a circle in each hexagon, and let I_j be the indicator r.v. for the jth point being contained inside one the circles. We have $E(I_j) > 0.9$ by the geometric fact mentioned above, so by linearity $E(I_1 + \cdots + I_{10}) > 9$. Thus, there is a positioning of the honeycomb tiling such that all 10 points are contained inside the circles. Putting coins on top of the circles containing the points, we can cover all ten points.

78. Ⓢ Let S be a set of binary strings $a_1 \ldots a_n$ of length n (where juxtaposition means concatenation). We call S *k-complete* if for any indices $1 \leq i_1 < \cdots < i_k \leq n$ and any binary string $b_1 \ldots b_k$ of length k, there is a string $s_1 \ldots s_n$ in S such that $s_{i_1} s_{i_2} \ldots s_{i_k} = b_1 b_2 \ldots b_k$. For example, for $n = 3$, the set $S = \{001, 010, 011, 100, 101, 110\}$ is 2-complete since all 4 patterns of 0's and 1's of length 2 can be found in any 2 positions. Show that if $\binom{n}{k} 2^k (1 - 2^{-k})^m < 1$, then there exists a k-complete set of size at most m.

Solution: Generate m random strings of length n independently, using fair coin flips to determine each bit. Let S be the resulting random set of strings. If we can show that the probability that S is k-complete is *positive*, then we know that a k-complete set of size at most m must *exist*. Let A be the event that S is k-complete. Let $N = \binom{n}{k} 2^k$ and let $A_1, \ldots, A_N$ be the events of the form "S contains a string which is $b_1 \ldots b_k$ at coordinates $i_1 < \cdots < i_k$," in any fixed order. For example, if $k = 3$ then A_1 could be the event "S has an element which is 110 at positions $1, 2, 3$." Then $P(A) > 0$ since

$$P(A^c) = P(\cup_{j=1}^N A_j^c) \leq \sum_{j=1}^N P(A_j^c) = N(1 - 2^{-k})^m < 1.$$

Mixed practice

79. A hacker is trying to break into a password-protected website by randomly trying to guess the password. Let m be the number of possible passwords.

(a) Suppose for this part that the hacker makes random guesses (with equal probability), *with replacement*. Find the average number of guesses it will take until the hacker guesses the correct password (including the successful guess).

(b) Now suppose that the hacker guesses randomly, *without replacement*. Find the average number of guesses it will take until the hacker guesses the correct password (including the successful guess).
Hint: Use symmetry.

(c) Show that the answer to (a) is greater than the answer to (b) (except in the degenerate case $m = 1$), and explain why this makes sense intuitively.

(d) Now suppose that the website locks out any user after n incorrect password attempts, so the hacker can guess at most n times. Find the PMF of the number of guesses that the hacker makes, both for the case of sampling with replacement and for the case of sampling without replacement.

Solution:

(a) The number of guesses is distributed as FS$(1/m)$, so the expected value is m.

(b) Let X be the number of guesses. Think of the m possible passwords as lined up in a random order, and getting chosen as guesses one by one in that order. By symmetry, the correct password is equally likely to be anywhere in that line. So

$$P(X = k) = 1/m$$

for $k = 1, 2, \ldots, m$, i.e., X is Discrete Uniform on $1, 2, \ldots, m$. Therefore,

$$E(X) = \frac{1 + 2 + \cdots + m}{m} = \frac{m+1}{2}.$$

Alternatively, write $X = I_1 + I_2 + \cdots + I_{m-1} + 1$, where I_j is the indicator of the jth possible-but-incorrect password being chosen before the correct password. By symmetry, $E(I_j) = P(I_j = 1) = 1/2$. So by linearity,

$$E(X) = \frac{m-1}{2} + 1 = \frac{m+1}{2}.$$

As another alternative, we can apply Example 4.47, noting that the number of wrong guesses is Negative Hypergeometric.

(c) For any $m > 1$, we have $2m > m + 1$, so $m > (m+1)/2$. In makes sense intuitively that on average the hacker will figure out the password faster by sampling without replacement than by sampling with replacement, since then no time is wasted repeated guesses already known to be wrong.

(d) Let Y be the number of guesses made. For sampling with replacement, the support of Y is $\{1, 2, \ldots, n\}$ and the PMF of Y is

$$P(Y = k) = \left(\frac{m-1}{m}\right)^{k-1} \cdot \frac{1}{m} = \quad \text{for } k = 1, 2, \ldots, n-1;$$
$$P(Y = n) = \left(\frac{m-1}{m}\right)^{n-1} \cdot \frac{1}{m} + \left(\frac{m-1}{m}\right)^n.$$

(To obtain $P(Y = n)$, we considered two cases: either the nth try is correct, or after n tries the correct password has still not been found.) We can check that this PMF is valid using the result for the sum of a finite geometric series.

For sampling without replacement, the support of Y is again $\{1, 2, \ldots, n\}$. Using the first symmetry argument from the solution to (b),

$$P(Y = k) = \frac{1}{m} = \quad \text{for } k = 1, 2, \ldots, n-1;$$
$$P(Y = n) = \frac{1}{m} + \frac{m-n}{m} = \frac{m-n+1}{m},$$

where again to find $Y = n$ we considered two cases: either the nth try is correct (which says that the correct password is at position n in the line described in the solution to (b)), or the correct password has not yet been found after n guesses (which says that the correct password is *not* in any of positions $1, 2, \ldots, n$ in the line).

80. A fair 20-sided die is rolled repeatedly, until a gambler decides to stop. The gambler receives the amount shown on the die when the gambler stops. The gambler decides in advance to roll the die until a value of m or greater is obtained, and then stop (where m is a fixed integer with $1 \leq m \leq 20$).

(a) What is the expected number of rolls (simplify)?

(b) What is the expected square root of the number of rolls (as a sum)?

Solution:

(a) Let X be the number of rolls. Then $X \sim \text{FS}(p)$ with $p = (21 - m)/20$, so

$$E(X) = 20/(21 - m).$$

As a check, note that this reduces to 1 when $m = 1$ (which makes sense since then the first roll is always accepted) and reduces to 20 when $m = 20$ (which makes sense since then only a value of 20 is accepted, resulting in a FS(1/20) distribution).

(b) By LOTUS,

$$E(\sqrt{X}) = \sum_{k=1}^{\infty} \sqrt{k} \cdot p(1-p)^{k-1},$$

where $p = (21 - m)/20$.

81. Ⓢ A group of 360 people is going to be split into 120 teams of 3 (where the order of teams and the order within a team don't matter).

(a) How many ways are there to do this?

(b) The group consists of 180 married couples. A random split into teams of 3 is chosen, with all possible splits equally likely. Find the expected number of teams containing married couples.

Solution:

(a) Imagine lining the people up and saying the first 3 are a team, the next 3 are a team, etc. This overcounts by a factor of $(3!)^{120} \cdot 120!$ since the order within teams and the order of teams don't matter. So the number of ways is

$$\frac{360!}{6^{120} \cdot 120!}.$$

(b) Let I_j be the indicator for the jth team having a married couple (taking the teams to be chosen one at a time, or with respect to a random ordering). By symmetry and linearity, the desired quantity is $120E(I_1)$. We have

$$E(I_1) = P(\text{first team has a married couple}) = \frac{180 \cdot 358}{\binom{360}{3}},$$

since the first team is equally likely to be any 3 of the people, and to have a married couple on the team we need to choose a couple and then any third person. So the expected value is

$$\frac{120 \cdot 180 \cdot 358}{\binom{360}{3}}.$$

(This simplifies to $\frac{120 \cdot 180 \cdot 358}{360 \cdot 359 \cdot 358/6} = \frac{360}{359}$. Another way to find the probability that the first team has a married couple is to note that any particular pair in the team has probability $\frac{1}{359}$ of being married to each other, so since there are 3 disjoint possibilities the probability is $\frac{3}{359}$.)

82. Ⓢ The gambler de Méré asked Pascal whether it is more likely to get at least one six in 4 rolls of a die, or to get at least one double-six in 24 rolls of a pair of dice. Continuing this pattern, suppose that a group of n fair dice is rolled $4 \cdot 6^{n-1}$ times.

(a) Find the expected number of times that "all sixes" is achieved (i.e., how often among the $4 \cdot 6^{n-1}$ rolls it happens that all n dice land 6 simultaneously).

(b) Give a simple but accurate approximation of the probability of having at least one occurrence of "all sixes", for n large (in terms of e but not n).

(c) de Méré finds it tedious to re-roll so many dice. So after one normal roll of the n dice, in going from one roll to the next, with probability 6/7 he leaves the dice in the same configuration and with probability 1/7 he re-rolls. For example, if $n = 3$ and the 7th roll is $(3, 1, 4)$, then 6/7 of the time the 8th roll remains $(3, 1, 4)$ and 1/7 of the time the 8th roll is a new random outcome. Does the expected number of times that

"all sixes" is achieved stay the same, increase, or decrease (compared with (a))? Give a short but clear explanation.

Solution:

(a) Let I_j be the indicator r.v. for the event "all sixes" on the jth roll. Then $E(I_j) = 1/6^n$, so the expected value is $4 \cdot 6^{n-1}/6^n = 2/3$.

(b) By a Poisson approximation with $\lambda = 2/3$ (the expected value from (a)), the probability is approximately $1 - e^{-2/3}$.

(c) The answer stays the same, by the same reasoning as in (a), since linearity of expectation holds even for dependent r.v.s.

83. Ⓢ Five people have just won a $100 prize, and are deciding how to divide the $100 up between them. Assume that whole dollars are used, not cents. Also, for example, giving $50 to the first person and $10 to the second is different from vice versa.

(a) How many ways are there to divide up the $100, such that each gets at least $10?

(b) Assume that the $100 is randomly divided up, with all of the possible allocations counted in (a) equally likely. Find the expected amount of money that the first person receives.

(c) Let A_j be the event that the jth person receives more than the first person (for $2 \leq j \leq 5$), when the $100 is randomly allocated as in (b). Are A_2 and A_3 independent?

Solution:

(a) Give each person $10, and then distribute the remaining $50 arbitrarily. By Bose-Einstein (thinking of people as boxes and dollars as balls!), the number of ways is

$$\binom{5 + 50 - 1}{50} = \binom{54}{50} = \binom{54}{4}.$$

(b) Let X_j be the amount that j gets. By symmetry, $E(X_j)$ is the same for all j. But $X_1 + \cdots + X_5 = 100$, so by linearity $100 = 5EX_1$. Thus, EX_1 is $20.

(c) The events A_2 and A_3 are not independent since knowing that A_2 occurred makes it more likely that person 1 received a low percentage of the money, which in turn makes it more likely that A_3 occurred.

84. Ⓢ Joe's iPod has 500 different songs, consisting of 50 albums of 10 songs each. He listens to 11 random songs on his iPod, with all songs equally likely and chosen independently (so repetitions may occur).

(a) What is the PMF of how many of the 11 songs are from his favorite album?

(b) What is the probability that there are 2 (or more) songs from the same album among the 11 songs he listens to?

(c) A pair of songs is a *match* if they are from the same album. If, say, the 1st, 3rd, and 7th songs are all from the same album, this counts as 3 matches. Among the 11 songs he listens to, how many matches are there on average?

Solution:

(a) The distribution is $\text{Bin}(n, p)$ with $n = 11, p = \frac{1}{50}$ (thinking of getting a song from the favorite album as a "success"). So the PMF is

$$\binom{11}{k}\left(\frac{1}{50}\right)^k \left(\frac{49}{50}\right)^{11-k}, \text{ for } 0 \leq k \leq 11.$$

(b) This is a version of the birthday problem. We have

$$P(\text{at least 1 match}) = 1 - P(\text{no matches}) = 1 - \frac{50 \cdot 49 \cdot \cdots \cdot 40}{50^{11}} = 1 - \frac{49!}{39! \cdot 50^{10}}.$$

(c) Defining an indicator r.v. I_{jk} for the event that the jth and kth songs match, we have $E(I_{jk}) = P(I_{jk} = 1) = 1/50$, so the expected number of matches is

$$\binom{11}{2}\frac{1}{50} = \frac{11 \cdot 10}{2 \cdot 50} = \frac{110}{100} = 1.1.$$

85. Ⓢ In each day that the Mass Cash lottery is run in Massachusetts, 5 of the integers from 1 to 35 are chosen (randomly and without replacement).

(a) When playing this lottery, find the probability of guessing exactly 3 numbers right, given that you guess at least 1 of the numbers right.

(b) Find an exact expression for the expected number of days needed so that all of the $\binom{35}{5}$ possible lottery outcomes will have occurred.

(c) Approximate the probability that after 50 days of the lottery, every number from 1 to 35 has been picked at least once.

Solution:

(a) The distribution is Hypergeometric (think of capture-recapture, "tagging" the numbers you choose). So

$$P(\text{exactly 3 right}|\text{at least 1 right}) = \frac{P(\text{exactly 3 right})}{1 - P(\text{none right})}$$
$$= \frac{\binom{5}{3}\binom{30}{2}/\binom{35}{5}}{1 - \binom{5}{0}\binom{30}{5}/\binom{35}{5}}.$$

(b) Let $n = \binom{35}{5}$. By the coupon collector problem (or directly by linearity, writing the number of days as a sum of T_j's with $T_j - 1$ a Geometric), the expected value is

$$n\left(\frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{2} + 1\right).$$

(c) Let A_j be the event that j doesn't get picked, so

$$P(A_j) = (30/35)^{50} = (6/7)^{50}.$$

Let X be the number of A_j that occur. A Poisson approximation for X is reasonable since these events are rare and weakly dependent. This gives

$$P(X = 0) \approx e^{-35 \cdot (6/7)^{50}} \approx 0.98.$$

86. A certain country has three political parties, denoted by A, B, and C. Each adult in the country is a member of exactly one of the three parties. There are n adults in the country, consisting of n_A members of party A, n_B members of party B, and n_C members of party C, where n_A, n_B, n_C are positive integers with $n_A + n_B + n_C = n$.

A simple random sample of size m is chosen from the adults in the country (the sampling is done *without* replacement, and all possible samples of size m are equally likely). Let X be the number of members of party A in the sample, Y be the number of members of party B in the sample, and Z be the number of members of party C in the sample.

(a) Find $P(X = x, Y = y, Z = z)$, for x, y, z nonnegative integers with $x + y + z = m$.

(b) Find $E(X)$.

(c) Find $\text{Var}(X)$, and briefly explain why your answer makes sense in the extreme cases $m = 1$ and $m = n$.

Solution:

(a) By the naive definition, the probability is

$$\frac{\binom{n_A}{x}\binom{n_B}{y}\binom{n_C}{z}}{\binom{n}{m}}.$$

(b) Let I_j be the indicator r.v. for the jth person sampled being in party A. By symmetry, linearity, and the fundamental bridge,

$$E(X) = mE(I_1) = \frac{mn_A}{n}.$$

Alternatively, we can use the fact that X is Hypergeometric, as in the next part.

(c) By the story of the Hypergeometric, $X \sim \text{HGeom}(n_A, n - n_A, m)$. So

$$\text{Var}(X) = \left(\frac{n - m}{n - 1}\right)\frac{mn_A}{n}\left(1 - \frac{n_A}{n}\right).$$

Let $p = n_A/n$. In the extreme case $m = 1$, $\text{Var}(X)$ reduces to $p(1 - p)$, which makes sense since $m = 1$ implies $X \sim \text{Bern}(p)$, which does have variance $p(1 - p)$.

In the extreme case $m = n$, $\text{Var}(X)$ reduces to 0, which makes sense since $m = n$ implies that the sample consists of the entire population of adults in the country, and then X deterministically equals n_A.

87. The U.S. Senate consists of 100 senators, with 2 from each of the 50 states. There are d Democrats in the Senate. A committee of size c is formed, by picking a random set of senators such that all sets of size c are equally likely.

(a) Find the expected number of Democrats on the committee.

(b) Find the expected number of states represented on the committee (by at least one senator).

(c) Find the expected number of states such that both of the state's senators are on the committee.

Solution:

(a) The number of Democrats on the committee is $\text{HGeom}(d, 100 - d, c)$. Using this or directly using indicator r.v.s, the expected value is $cd/100$.

(b) Let I_j be the indicator for the jth state (in alphabetical order) being represented on the committee. By the fundamental bridge, $E(I_j) = 1 - \binom{98}{c}/\binom{100}{c}$. So by linearity, the expected number is

$$50\left(1 - \binom{98}{c}/\binom{100}{c}\right).$$

Note that this makes sense in the extreme cases $c = 0$ and $c = 100$.

(c) Assume $c \geq 2$, since otherwise it's impossible to have both of a state's senators on the committee. The expected value of the indicator for the jth state having both senators on the committee is $\binom{2}{2}\binom{98}{c-2}/\binom{100}{c}$, so by linearity the expected number is

$$50\binom{98}{c - 2}/\binom{100}{c}.$$

88. A certain college has g good courses and b bad courses, where g and b are positive integers. Alice, who is hoping to find a good course, randomly shops courses one at a time (without replacement) until she finds a good course.

    (a) Find the expected number of bad courses that Alice shops before finding a good course (as a simple expression in terms of g and b).

    (b) Should the answer to (a) be less than, equal to, or greater than b/g? Explain this using properties of the Geometric distribution.

    Solution:

    (a) Let the courses be ordered randomly. For each bad class, create an indicator r.v. for it preceding all the good courses. By symmetry, each indicator has expected value $1/(g+1)$, so by linearity the desired expectation is $b/(g+1)$. Alternatively, we can apply the Negative Hypergeometric results from Example 4.4.7.

    (b) The answer should be less than b/g (and it *is* less: $b/(g+1) < b/g$). Under sampling *with* replacement, the number of bad courses shopped by Alice would be $\text{Geom}(g/(g+b))$, which has mean b/g. Sampling without replacement should give a smaller mean since by discarding bad courses, Alice is making progress toward finding a good course.

89. A DNA sequence can be represented as a sequence of letters, where the *alphabet* has 4 letters: A,C,T,G. Suppose such a sequence is generated randomly, where the letters are independent and the probabilities of A,C,T,G are p_A, p_C, p_T, p_G, respectively.

    (a) In a DNA sequence of length 115, what is the variance of the number of occurrences of the letter C?

    (b) In a DNA sequence of length 115, what is the expected number of occurrences of the expression CATCAT? Note that, for example, the expression CATCATCAT counts as 2 occurrences.

    (c) In a DNA sequence of length 6, what is the probability that the expression CAT occurs at least once?

    Solution:

    (a) The number of occurrences of the letter C is $\text{Bin}(115, p_C)$, so the variance is $115 p_C(1 - p_C)$.

    (b) Let I_j be the indicator r.v. of the expression CATCAT appearing starting at position j, for $1 \le j \le 110$. Then $E(I_j) = (p_C p_A p_T)^2$, so the expected number is $110(p_C p_A p_T)^2$.

    (c) Let B_i be the event that CAT appears starting at position i, for $1 \le i \le 4$. By inclusion-exclusion,

    $$P(B_1 \cup B_2 \cup B_3 \cup B_4) = P(B_1) + P(B_2) + P(B_3) + P(B_4) - P(B_1 \cap B_4)$$
    $$= 4P(B_1) - P(B_1 \cap B_4)$$
    $$= 4 p_C p_A p_T - (p_C p_A p_T)^2$$

    since, e.g., $B_1 \cap B_2 = \emptyset$, and $B_1 \cap B_4$ is the event that the sequence is CATCAT.

90. Alice is conducting a survey in a town with population size 1000. She selects a simple random sample of size 100 (i.e., sampling without replacement, such that all samples of size 100 are equally likely). Bob is also conducting a survey in this town. Bob selects a simple random sample of size 20, independent of Alice's sample. Let A be the set of people in Alice's sample and B be the set of people in Bob's sample.

    (a) Find the expected number of people in $A \cap B$.

(b) Find the expected number of people in $A \cup B$.

(c) The 1000 people consist of 500 married couples. Find the expected number of couples such that both members of the couple are in Bob's sample.

Solution:

(a) Let I_j be the indicator for person j in the town being in $A \cap B$. By linearity, symmetry, and the fundamental bridge,

$$E|A \cap B| = 1000E(I_1) = 1000 \cdot \frac{100}{1000} \cdot \frac{20}{1000} = 2.$$

Alternatively, note that by the story of the Hypergeometric,

$$|A \cap B| \sim \text{HGeom}(100, 900, 20)$$

(think of Alice's sample as tagged elk and Bob's sample as recaptured elk), so

$$E|A \cap B| = \frac{20 \cdot 100}{100 + 900} = \frac{2000}{1000} = 2.$$

(b) Writing $|A \cup B| = |A| + |B| - |A \cap B|$, by linearity we have

$$E|A \cup B| = |A| + |B| - E|A \cap B| = 100 + 20 - 2 = 118.$$

(c) For each couple, create an indicator for whether both members of the couple are in Bob's sample. Let Philip and Elizabeth be one of the couples. By linearity, symmetry, and the fundamental bridge, the desired expectation is

$$500 \cdot P\,(\text{Philip in Bob's sample})P(\text{Elizabeth in Bob's sample}|\text{Philip in Bob's sample})$$
$$= 500 \cdot \frac{20}{1000} \cdot \frac{19}{999}$$
$$= \frac{190}{999}.$$

91. The *Wilcoxon rank sum test* is a widely used procedure for assessing whether two groups of observations come from the same distribution. Let group 1 consist of i.i.d. $X_1, \ldots, X_m$ with CDF F and group 2 consist of i.i.d. $Y_1, \ldots, Y_n$ with CDF G, with all of these r.v.s independent. Assume that the probability of 2 of the observations being equal is 0 (this will be true if the distributions are continuous).

After the $m + n$ observations are obtained, they are listed in increasing order, and each is assigned a *rank* between 1 and $m + n$: the smallest has rank 1, the second smallest has rank 2, etc. Let R_j be the rank of X_j among all the observations for $1 \le j \le m$, and let $R = \sum_{j=1}^{m} R_j$ be the sum of the ranks for group 1.

Intuitively, the Wilcoxon rank sum test is based on the idea that a very large value of R is evidence that observations from group 1 are usually larger than observations from group 2 (and vice versa if R is very small). But how large is "very large" and how small is "very small"? Answering this precisely requires studying the distribution of the *test statistic R*.

(a) The *null hypothesis* in this setting is that $F = G$. Show that if the null hypothesis is true, then $E(R) = m(m + n + 1)/2$.

(b) The *power* of a test is an important measure of how good the test is about saying to reject the null hypothesis if the null hypothesis is false. To study the power of the Wilcoxon rank sum stest, we need to study the distribution of R in general. So for this part, we do *not* assume $F = G$. Let $p = P(X_1 > Y_1)$. Find $E(R)$ in terms of m, n, p.

Hint: Write R_j in terms of indicator r.v.s for X_j being greater than various other r.v.s.

Solution:

(a) Assume that $F = G$. Then by symmetry, the list of ranks of the $m + n$ observations is equally likely to be any permutation of $1, 2, \ldots, m + n$. So R_j is equally likely to be any of $1, 2, \ldots, m + n$, which means that

$$E(R_j) = \frac{1}{m + n} \sum_{i=1}^{m+n} i = \frac{m + n + 1}{2}.$$

Then by linearity, $E(R) = m(m + n + 1)/2$.

(b) We can write

$$R_1 = 1 + \sum_{i=2}^{m} I(X_1 > X_i) + \sum_{i=1}^{n} I(X_1 > Y_i),$$

where for any event A, $I(A)$ is the indicator r.v. for A. By linearity, symmetry, and the fundamental bridge,

$$E(R_1) = 1 + (m - 1)/2 + nP(X_1 > Y_1).$$

By symmetry, $E(R_j) = E(R_1)$ for $1 \leq j \leq m$. So by linearity again,

$$E(R) = m + m(m - 1)/2 + mnp = m\left((m + 1)/2 + np\right).$$

Note that this agrees with (a) when $p = 1/2$.

92. The legendary Caltech physicist Richard Feynman and two editors of *The Feynman Lectures on Physics* (Michael Gottlieb and Ralph Leighton) posed the following problem about how to decide what to order at a restaurant. You plan to eat m meals at a certain restaurant, where you have never eaten before. Each time, you will order one dish.

The restaurant has n dishes on the menu, with $n \geq m$. Assume that if you had tried all the dishes, you would have a definite ranking of them from 1 (your least favorite) to n (your favorite). If you knew which your favorite was, you would be happy to order it always (you never get tired of it).

Before you've eaten at the restaurant, this ranking is completely unknown to you. After you've tried some dishes, you can rank those dishes amongst themselves, but don't know how they compare with the dishes you haven't yet tried. There is thus an *exploration-exploitation tradeoff*: should you try new dishes, or should you order your favorite among the dishes you have tried before?

A natural strategy is to have two phases in your series of visits to the restaurant: an *exploration phase*, where you try different dishes each time, and an *exploitation phase*, where you always order the best dish you obtained in the exploration phase. Let k be the length of the exploration phase (so $m - k$ is the length of the exploitation phase).

Your goal is to maximize the expected sum of the ranks of the dishes you eat there (the rank of a dish is the "true" rank from 1 to n that you would give that dish if you could try all the dishes). Show that the optimal choice is

$$k = \sqrt{2(m + 1)} - 1,$$

or this rounded up or down to an integer if needed. Do this in the following steps:

(a) Let X be the rank of the best dish that you find in the exploration phase. Find the expected sum of the ranks of all the dishes you eat (including both phases), in terms of k, n, and $E(X)$.

(b) Find the PMF of X, as a simple expression in terms of binomial coefficients.

(c) Show that

$$E(X) = \frac{k(n + 1)}{k + 1}.$$

Hint: Use Example 1.5.2 (about the team captain) and Exercise 20 from Chapter 1 (about the hockey stick identity).

(d) Use calculus to find the optimal value of k.

Solution:

(a) Let R_j be the rank of the jth dish that you try, and R be the sum of the ranks. Then
$$R = R_1 + \cdots + R_k + (m - k)X,$$
and
$$E(R_1) = \frac{1}{n} \sum_{i=1}^{n} i = (n + 1)/2.$$

The result for $E(R_1)$ can be obtained by summing the arithmetic series (see the math appendix) or using indicator r.v.s. For the latter, note that R_1 is 1 plus the number of dishes ranked below the first dish. Create an indicator for each of the $n - 1$ other dishes of whether it is ranked below the first dish. Each of those indicators has expected value $1/2$ by symmetry, so by linearity we again have $E(R_1) = 1 + (n - 1)/2 = (n + 1)/2$. Therefore,
$$E(R) = k(n + 1)/2 + (m - k)E(X).$$

(b) The support is $\{k, k + 1, \ldots, n\}$. The PMF is given by
$$P(X = j) = \frac{\binom{j-1}{k-1}}{\binom{n}{k}}$$
for j in the support, since $X = j$ is equivalent to getting the dish with rank j and $k - 1$ worse dishes. (To see that this is a valid PMF, we can use the hockey stick identity referred to in the hint for the next part.)

(c) Using the identities in the hint and then writing the binomial coefficients in terms of factorials, we have
$$E(X) = \frac{1}{\binom{n}{k}} \sum_{j=k}^{n} j \binom{j-1}{k-1} = \frac{k}{\binom{n}{k}} \sum_{j=k}^{n} \binom{j}{k} = \frac{k \binom{n+1}{k+1}}{\binom{n}{k}} = \frac{k(n+1)}{k+1}.$$

(d) Replacing the integer k by a real variable x and using the above results, we obtain
$$g(x) = \frac{x}{2} + \frac{(m - x)x}{x + 1}$$
as the function to maximize (after dividing by the constant $n + 1$). Then
$$g'(x) = \frac{1}{2} + \frac{(x + 1)(m - 2x) - (m - x)x}{(x + 1)^2} = \frac{1}{2} + \frac{m - (x + 1)^2 + 1}{(x + 1)^2} = \frac{m + 1}{(x + 1)^2} - \frac{1}{2}.$$

Setting $g'(x) = 0$, we have
$$x_0 = \sqrt{2(m + 1)} - 1$$
as the positive solution. Since g is increasing from 0 to x_0 and decreasing from x_0 to m, the local and absolute maximum of g on $[0, m]$ is at x_0, and the optimal k is x_0, rounded up or down to an integer if needed.

5

Continuous random variables

PDFs and CDFs

1. The Rayleigh distribution from Example 5.1.7 has PDF

    $$f(x) = xe^{-x^2/2}, \quad x > 0.$$

    Let X have the Rayleigh distribution.

    (a) Find $P(1 < X < 3)$.

    (b) Find the first quartile, median, and third quartile of X; these are defined to be the values q_1, q_2, q_3 (respectively) such that $P(X \le q_j) = j/4$ for $j = 1, 2, 3$.

    Solution:

    (a) We have

    $$P(1 < X < 3) = \int_1^3 xe^{-x^2/2}dx.$$

    To compute this, we can make the substitution $u = -x^2/2$, or we can use the fact from Example 5.1.7 that the CDF of X is $F(x) = 1 - e^{-x^2/2}$ for $x > 0$. Then

    $$P(1 < X < 3) = F(3) - F(1) = e^{-1/2} - e^{-9/2} \approx 0.595.$$

    (b) Using the CDF F from above, we can find q_j by setting $F(q_j) = j/4$ and solving for q_j. This gives

    $$1 - e^{-q_j^2/2} = j/4,$$

    which becomes

    $$q_j = \sqrt{-2\log(1 - j/4)}.$$

    Numerically,

    $$q_1 \approx 0.759, q_2 \approx 1.177, q_3 \approx 1.665.$$

2. (a) Make up a PDF f, with an application for which that PDF would be plausible, where $f(x) > 1$ for all x in a certain interval.

    (b) Show that if a PDF f has $f(x) > 1$ for all x in a certain interval, then that interval must have length less than 1.

    Solution:

    (a) For example, let X be the lifetime of a certain product that is not very susceptible to wear and tear, and suppose that $X \sim \text{Expo}(2)$ (measured in years). The PDF is $f(x) = 2e^{-2x}$ for $x > 0$. Then $f(x) > 1$ for $0 < x < \frac{1}{2}\log(2)$.

    (b) If $f(x) > 1$ for all x in an interval (a, b), then

    $$1 = \int_{-\infty}^{\infty} f(x)dx \ge \int_a^b f(x)dx > \int_a^b dx = b - a,$$

    so the length $b - a$ of the interval must be less than 1.

3. Let F be the CDF of a continuous r.v., and $f = F'$ be the PDF.

(a) Show that g defined by $g(x) = 2F(x)f(x)$ is also a valid PDF.

(b) Show that h defined by $h(x) = \frac{1}{2}f(-x) + \frac{1}{2}f(x)$ is also a valid PDF.

Solution:

(a) We have $g(x) \geq 0$ since both F and f are nonnegative. Making the substitution $u = F(x)$, so $du = f(x)dx$, we have

$$\int_{-\infty}^{\infty} g(x)dx = 2\int_{-\infty}^{\infty} F(x)f(x)dx = 2\int_0^1 u\, du = 1.$$

(b) We have $h(x) \geq 0$ since f is nonnegative. Then

$$\int_{-\infty}^{\infty} h(x)dx = \frac{1}{2}\int_{-\infty}^{\infty} f(-x)dx + \frac{1}{2}\int_{-\infty}^{\infty} f(x)dx = \frac{1}{2} + \frac{1}{2} = 1,$$

since

$$\int_{-\infty}^{\infty} f(-x)dx = -\int_{\infty}^{-\infty} f(u)du = \int_{-\infty}^{\infty} f(u)du = 1.$$

4. Let X be a continuous r.v. with CDF F and PDF f.

(a) Find the conditional CDF of X given $X > a$ (where a is a constant with $P(X > a) \neq 0$). That is, find $P(X \leq x | X > a)$ for all a, in terms of F.

(b) Find the conditional PDF of X given $X > a$ (this is the derivative of the conditional CDF).

(c) Check that the conditional PDF from (b) is a valid PDF, by showing directly that it is nonnegative and integrates to 1.

Solution:

(a) We have $P(X \leq x | X > a) = 0$ for $x \leq a$. For $x > a$,

$$P(X \leq x | X > a) = \frac{P(a < X \leq x)}{P(X > a)} = \frac{F(x) - F(a)}{1 - F(a)}.$$

(b) The derivative of the conditional CDF is $f(x)/(1 - F(a))$ for $x > a$, and 0 otherwise.

(c) We have $f(x)/(1 - F(a)) \geq 0$ since $f(x) \geq 0$. And

$$\int_a^{\infty} \frac{f(x)}{1 - F(a)}dx = \frac{1}{1 - F(a)}\int_a^{\infty} f(x)dx = \frac{1 - F(a)}{1 - F(a)} = 1.$$

5. A circle with a random radius $R \sim \text{Unif}(0, 1)$ is generated. Let A be its area.

(a) Find the mean and variance of A, without first finding the CDF or PDF of A.

(b) Find the CDF and PDF of A.

Solution:

(a) We have $A = \pi R^2$. By LOTUS,

$$E(A) = \pi \int_0^1 r^2 dr = \frac{\pi}{3}$$

$$E(A^2) = \pi^2 \int_0^1 r^4 dr = \frac{\pi^2}{5}$$

$$\text{Var}(A) = \frac{\pi^2}{5} - \frac{\pi^2}{9} = \frac{4\pi^2}{45}.$$

(b) The CDF of A is

$$P(A \leq a) = P(\pi R^2 \leq a) = P(R \leq \sqrt{a/\pi}) = \sqrt{a/\pi},$$

for $0 < a < \pi$ (and the CDF is 0 for $a \leq 0$ and 1 for $a \geq \pi$). So the PDF of A is

$$f(a) = \frac{1}{2\sqrt{\pi a}},$$

for $0 < a < \pi$ (and 0 otherwise).

6. The 68-95-99.7% rule gives approximate probabilities of a Normal r.v. being within $1, 2$, and 3 standard deviations of its mean. Derive analogous rules for the following distributions.

(a) Unif$(0, 1)$.

(b) Expo(1).

(c) Expo$(1/2)$. Discuss whether there is one such rule that applies to all Exponential distributions, just as the 68-95-99.7% rule applies to all Normal distributions, not just to the standard Normal.

Solution:

(a) Let $U \sim$ Unif$(0, 1)$. The mean is $\mu = 1/2$ and the standard deviation is $\sigma = \frac{1}{\sqrt{12}}$. So

$$P(|U - \mu| \leq \sigma) = P(\mu - \sigma \leq U \leq \mu + \sigma) = 2\sigma \approx 0.5774.$$

But

$$P(|U - \mu| \leq 2\sigma) = 1,$$

since $\mu + 2\sigma > 1$. Therefore, the analogous rule for the Unif$(0, 1)$ is a 58-100-100% rule.

(b) Let $X \sim$ Expo(1). The mean is $\mu = 1$ and the standard deviation is $\sigma = 1$. So

$$
\begin{aligned}
P(|X - \mu| \leq \sigma) &= P(X \leq 2) = 1 - e^{-2} \approx 0.8647, \\
P(|X - \mu| \leq 2\sigma) &= P(X \leq 3) = 1 - e^{-3} \approx 0.9502, \\
P(|X - \mu| \leq 3\sigma) &= P(X \leq 4) = 1 - e^{-4} \approx 0.9817.
\end{aligned}
$$

So the analogous rule for the Expo(1) is an 86-95-98% rule.

(c) Let $Y \sim$ Expo(λ) and $X = \lambda Y \sim$ Expo(1). Then Y has mean $\mu = 1/\lambda$ and standard deviation $\sigma = 1/\lambda$. For any real number $c \geq 1$, the probability of Y being within c standard deviations of its mean is

$$P(|Y - \mu| \leq c\sigma) = P(|\lambda Y - \lambda\mu| \leq c\lambda\sigma) = P(|X - 1| \leq c) = P(X \leq c + 1) = 1 - e^{-c-1}.$$

This probability does not depend on λ, so for any λ the analogous rule for the Expo(λ) distribution is an 86-95-98% rule.

7. Let

$$F(x) = \frac{2}{\pi} \sin^{-1}\left(\sqrt{x}\right), \text{ for } 0 < x < 1,$$

and let $F(x) = 0$ for $x \leq 0$ and $F(x) = 1$ for $x \geq 1$.

(a) Check that F is a valid CDF, and find the corresponding PDF f. This distribution is called the *Arcsine distribution*, though it also goes by the name Beta$(1/2, 1/2)$ (we will explore the Beta distribution in depth in Chapter 8).

(b) Explain how it is possible for f to be a valid PDF even though $f(x)$ goes to ∞ as x approaches 0 and as x approaches 1.

Solution:

(a) The function F is increasing since the square root and $\sin^{-1}$ functions are increasing. It is continuous since $\frac{2}{\pi}\sin^{-1}\left(\sqrt{0}\right) = 0$, $\frac{2}{\pi}\sin^{-1}\left(\sqrt{1}\right) = 1$, the square root function is continuous on $(0, \infty)$, the $\sin^{-1}$ function is continuous on $(-1, 1)$, and a constant function is continuous everywhere. And $F(x) \to 0$ as $x \to -\infty$, $F(x) \to 1$ as $x \to \infty$ (since in fact $F(x) = 0$ for $x \leq 0$ and $F(x) = 1$ for $x \geq 1$). So F is a valid CDF.

By the chain rule, the corresponding PDF is

$$f(x) = \frac{2}{\pi} \cdot \frac{1}{2\sqrt{x}} \cdot \frac{1}{\sqrt{1-x}} = \frac{1}{\pi}x^{-1/2}(1-x)^{-1/2},$$

for $0 < x < 1$ (and 0 otherwise).

(b) By (a), $f(x) \to \infty$ as $x \to 0$, and also $f(x) \to \infty$ as $x \to 1$. But the *area* under the curve is still finite (in particular, the area is 1). There is no contradiction in this. For a simpler example, note that

$$\int_0^1 \frac{1}{\sqrt{x}}dx = \lim_{a \to 0+} \int_a^1 \frac{1}{\sqrt{x}}dx = \lim_{a \to 0+} (2\sqrt{1} - 2\sqrt{a}) = 2,$$

which is finite even though $1/\sqrt{x} \to \infty$ as x approaches 0 from the right.

8. The *Beta distribution* with parameters $a = 3, b = 2$ has PDF

$$f(x) = 12x^2(1-x), \text{ for } 0 < x < 1.$$

(We will discuss the Beta distribution in detail in Chapter 8.) Let X have this distribution.

(a) Find the CDF of X.

(b) Find $P(0 < X < 1/2)$.

(c) Find the mean and variance of X (without quoting results about the Beta distribution).

Solution:

(a) The CDF F is given by

$$F(t) = \int_0^t 12x^2(1-x)dx = 12\int_0^t (x^2 - x^3)dx = 12\left(\frac{t^3}{3} - \frac{t^4}{4}\right),$$

for $0 < t < 1$ (and the CDF is 0 for $t \leq 0$ and 1 for $t \geq 1$).

(b) By (a),
$$P(0 < X < 1/2) = F(1/2) - F(0) = 0.3125.$$

(c) We have

$$E(X) = \int_0^1 xf(x)dx = 12\int_0^1 (x^3 - x^4)dx = 12\left(\frac{1}{4} - \frac{1}{5}\right) = 0.6,$$

$$E(X^2) = \int_0^1 x^2 f(x)dx = 12\int_0^1 (x^4 - x^5)dx = 12\left(\frac{1}{5} - \frac{1}{6}\right) = 0.4,$$

$$\text{Var}(X) = 0.4 - 0.6^2 = 0.04.$$

9. The *Cauchy distribution* has PDF

$$f(x) = \frac{1}{\pi(1+x^2)},$$

for all x. (We will introduce the Cauchy from another point of view in Chapter 7.) Find the CDF of a random variable with the Cauchy PDF.

Hint: Recall that the derivative of the inverse tangent function $\tan^{-1}(x)$ is $\frac{1}{1+x^2}$.

Solution: The CDF of a Cauchy is F given by

$$F(t) = \frac{1}{\pi} \int_{-\infty}^{t} \frac{1}{1+x^2} dx = \frac{1}{\pi} \tan^{-1}(x) \Big|_{-\infty}^{t} = \frac{1}{\pi} \tan^{-1}(t) + \frac{1}{2},$$

for all real t.

Uniform and universality of the Uniform

10. Let $U \sim \text{Unif}(0, 8)$.

    (a) Find $P(U \in (0, 2) \cup (3, 7))$ without using calculus.

    (b) Find the conditional distribution of U given $U \in (3, 7)$.

    Solution:

    (a) Since for a Uniform probability is proportional to length,

    $$P(U \in (0, 2) \cup (3, 7)) = \frac{2+4}{8} = \frac{3}{4}.$$

    (b) By Proposition 5.2.3, the conditional distribution is $\text{Unif}(3, 7)$.

11. Ⓢ Let U be a Uniform r.v. on the interval $(-1, 1)$ (be careful about minus signs).

    (a) Compute $E(U), \text{Var}(U)$, and $E(U^4)$.

    (b) Find the CDF and PDF of U^2. Is the distribution of U^2 Uniform on $(0, 1)$?

    Solution:

    (a) We have $E(U) = 0$ since the distribution is symmetric about 0. By LOTUS,

    $$E(U^2) = \frac{1}{2} \int_{-1}^{1} u^2 du = \frac{1}{3}.$$

    So $\text{Var}(U) = E(U^2) - (EU)^2 = E(U^2) = \frac{1}{3}$. Again by LOTUS,

    $$E(U^4) = \frac{1}{2} \int_{-1}^{1} u^4 du = \frac{1}{5}.$$

    (b) Let $G(t)$ be the CDF of U^2. Clearly $G(t) = 0$ for $t \leq 0$ and $G(t) = 1$ for $t \geq 1$, because $0 \leq U^2 \leq 1$. For $0 < t < 1$,

    $$G(t) = P(U^2 \leq t) = P(-\sqrt{t} \leq U \leq \sqrt{t}) = \sqrt{t},$$

    since the probability of U being in an interval in $(-1, 1)$ is proportional to its length. The PDF is $G'(t) = \frac{1}{2} t^{-1/2}$ for $0 < t < 1$ (and 0 otherwise). The distribution of U^2 is *not* Uniform on $(0, 1)$ as the PDF is not a constant on this interval (it is an example of a *Beta distribution*, an important distribution that is introduced in Chapter 8).

12. ⑤ A stick is broken into two pieces, at a uniformly random breakpoint. Find the CDF and average of the length of the longer piece.

    Solution: We can assume the units are chosen so that the stick has length 1. Let L be the length of the longer piece, and let the breakpoint be $U \sim \text{Unif}(0, 1)$. For any $l \in [1/2, 1]$, observe that $L < l$ is equivalent to $\{U < l \text{ and } 1 - U < l\}$, which can be written as $1 - l < U < l$. We can thus obtain L's CDF as

    $$F_L(l) = P(L < l) = P(1 - l < U < l) = 2l - 1,$$

    so $L \sim \text{Unif}(1/2, 1)$. In particular, $E(L) = 3/4$.

13. A stick of length 1 is broken at a uniformly random point, yielding two pieces. Let X and Y be the lengths of the shorter and longer pieces, respectively, and let $R = X/Y$ be the ratio of the lengths X and Y.

    (a) Find the CDF and PDF of R.

    (b) Find the expected value of R (if it exists).

    (c) Find the expected value of $1/R$ (if it exists).

    Solution:

    (a) Let $U \sim \text{Unif}(0, 1)$ be the breakpoint, so $X = \min(U, 1 - U)$. For $r \in (0, 1)$,

    $$P(R \le r) = P(X \le r(1 - X)) = P\left(X \le \frac{r}{1 + r}\right).$$

    This is the CDF of X, evaluated at $r/(1 + r)$, so we just need to find the CDF of X:

    $$P(X \le x) = 1 - P(X > x) = 1 - P(U > x, 1 - U > x) = 1 - P(x < U < 1 - x) = 2x,$$

    for $0 \le x \le 1/2$, and the CDF is 0 for $x < 0$ and 1 for $x > 1/2$. So $X \sim \text{Unif}(0, 1/2)$, and the CDF of R is

    $$P(R \le r) = P\left(X \le \frac{r}{1 + r}\right) = \frac{2r}{1 + r}$$

    for $r \in (0, 1)$, and it is 0 for $r \le 0$ and 1 for $r \ge 1$. Alternatively, we can note that

    $$P\left(X \le \frac{r}{1 + r}\right) = P\left(U \le \frac{r}{1 + r} \text{ or } 1 - U \le \frac{r}{1 + r}\right).$$

    The events $U \le r/(1 + r)$ and $1 - U \le r/(1 + r)$ are disjoint for $r \in (0, 1)$ since the latter is equivalent to $U \ge 1/(1 + r)$. Thus, again for $r \in (0, 1)$ we have

    $$P(R \le r) = P\left(U \le \frac{r}{1 + r}\right) + P\left(1 - U \le \frac{r}{1 + r}\right) = \frac{2r}{1 + r}.$$

    The PDF of R is 0 if r is not in $(0, 1)$, and for $r \in (0, 1)$ it is

    $$f(r) = \frac{2(1 + r) - 2r}{(1 + r)^2} = \frac{2}{(1 + r)^2}.$$

    (b) We have

    $$E(R) = 2 \int_0^1 \frac{r}{(1 + r)^2} dr = 2 \int_1^2 \frac{(t - 1)}{t^2} dt = 2 \int_1^2 \frac{1}{t} dt - 2 \int_1^2 \frac{1}{t^2} dt = 2 \ln 2 - 1.$$

    (c) This expected value does not exist, since $\int_0^1 \frac{1}{r(1+r)^2} dr$ diverges. To show this, note that $\int_0^1 \frac{1}{r} dr$ diverges and $\frac{1}{r(1+r)^2} \ge \frac{1}{4r}$ for $0 < r < 1$.

14. Let $U_1, \ldots, U_n$ be i.i.d. Unif$(0, 1)$, and $X = \max(U_1, \ldots, U_n)$. What is the PDF of X? What is EX?

Hint: Find the CDF of X first, by translating the event $X \leq x$ into an event involving $U_1, \ldots, U_n$.

Solution: Note that $X \leq x$ holds if and only if all of the U_j's are at most x. So the CDF of X is

$$P(X \leq x) = P(U_1 \leq x, U_2 \leq x, \ldots, U_n \leq x) = (P(U_1 \leq x))^n = x^n,$$

for $0 < x < 1$ (and the CDF is 0 for $x \leq 0$ and 1 for $x \geq 1$). So the PDF of X is

$$f(x) = nx^{n-1},$$

for $0 < x < 1$ (and 0 otherwise). Then

$$EX = \int_0^1 x(nx^{n-1})dx = n \int_0^1 x^n dx = \frac{n}{n+1}.$$

(For generalizations of these results, see the material on *order statistics* in Chapter 8.)

15. Let $U \sim \text{Unif}(0, 1)$. Using U, construct $X \sim \text{Expo}(\lambda)$.

Solution: The CDF of X is $F(x) = 1 - e^{-\lambda x}$ for $x > 0$. The inverse function is $F^{-1}(u) = -\frac{1}{\lambda} \log(1 - u)$, for $0 < u < 1$. So by universality of the Uniform,

$$X = -\frac{1}{\lambda} \log(1 - U) \sim \text{Expo}(\lambda).$$

16. ⓢ Let $U \sim \text{Unif}(0, 1)$, and

$$X = \log\left(\frac{U}{1 - U}\right).$$

Then X has the Logistic distribution, as defined in Example 5.1.6.

(a) Write down (but do not compute) an integral giving $E(X^2)$.

(b) Find $E(X)$ without using calculus.

Hint: A useful symmetry property here is that $1 - U$ has the same distribution as U.

Solution:

(a) By LOTUS,

$$E(X^2) = \int_0^1 \left(\log\left(\frac{u}{1 - u}\right)\right)^2 du.$$

(b) By the symmetry property mentioned in the hint, $1 - U$ has the same distribution as U. So by linearity,

$$E(X) = E(\log U - \log(1 - U)) = E(\log U) - E(\log(1 - U)) = 0.$$

17. Let $U \sim \text{Unif}(0, 1)$. As a function of U, create an r.v. X with CDF $F(x) = 1 - e^{-x^3}$ for $x > 0$.

Solution: The inverse function is $F^{-1}(u) = -(\log(1 - u))^{1/3}$ for $0 < u < 1$. So by universality of the Uniform,

$$X = -(\log(1 - U))^{1/3}$$

has CDF F.

18. The *Pareto distribution* with parameter $a > 0$ has PDF $f(x) = a/x^{a+1}$ for $x \geq 1$ (and 0 otherwise). This distribution is often used in statistical modeling.

(a) Find the CDF of a Pareto r.v. with parameter a; check that it is a valid CDF.

(b) Suppose that for a simulation you want to run, you need to generate i.i.d. Pareto(a) r.v.s. You have a computer that knows how to generate i.i.d. Unif(0, 1) r.v.s but does not know how to generate Pareto r.v.s. Show how to do this.

Solution:

(a) The CDF F is given by

$$F(y) = \int_1^y \frac{a}{t^{a+1}} dt = (-t^{-a})\Big|_1^y = 1 - \frac{1}{y^a}$$

for $y > 1$, and $F(y) = 0$ for $y \leq 1$. This is a valid CDF since it is increasing in y (this can be seen directly or from the fact that $F' = f$ is nonnegative), right continuous (in fact it is continuous), $F(y) \to 0$ as $y \to -\infty$, and $F(y) \to 1$ as $y \to \infty$.

(b) Let $U \sim \text{Unif}(0, 1)$. By universality of the Uniform, $F^{-1}(U) \sim \text{Pareto}(a)$. The inverse of the CDF is

$$F^{-1}(u) = \frac{1}{(1 - u)^{1/a}}.$$

So

$$Y = \frac{1}{(1 - U)^{1/a}} \sim \text{Pareto}(a).$$

To check directly that Y defined in this way is Pareto(a), we can find its CDF:

$$P(Y \leq y) = P\left(\frac{1}{y} \leq (1 - U)^{1/a}\right) = P\left(U \leq 1 - \frac{1}{y^a}\right) = 1 - \frac{1}{y^a},$$

for any $y \geq 1$, and $P(Y \leq y) = 0$ for $y < 1$. Then if we have n i.i.d. Unif(0, 1) r.v.s, we can apply this transformation to each of them to obtain n i.i.d. Pareto(a) r.v.s.

Normal

19. Let $Z \sim \mathcal{N}(0, 1)$. Create an r.v. $Y \sim \mathcal{N}(1, 4)$, as a simple-looking function of Z. Make sure to check that your Y has the correct mean and variance.

Solution: We can create Y using a location-scale transformation: $Y = 1 + 2Z$. Then Y is Normal, with $E(Y) = 1 + 2E(Z) = 1$ and $\text{Var}(Y) = \text{Var}(2Z) = 4\text{Var}(Z) = 4$.

20. Engineers sometimes work with the "error function"

$$\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-x^2} dx,$$

instead of working with the standard Normal CDF Φ.

(a) Show that the following conversion between Φ and erf holds for all z:

$$\Phi(z) = \frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{z}{\sqrt{2}}\right).$$

(b) Show that erf is an odd function, i.e., $\text{erf}(-z) = -\text{erf}(z)$.

Solution:

(a) Let $Z \sim \mathcal{N}(0,1)$. For $z \geq 0$,

$$
\begin{aligned}
\Phi(z) &= P(Z \leq z) \\
&= P(Z < 0) + P(0 \leq Z \leq z) \\
&= \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_0^z e^{-x^2/2} dx \\
&= \frac{1}{2} + \frac{1}{\sqrt{\pi}} \int_0^{z/\sqrt{2}} e^{-u^2} du \\
&= \frac{1}{2} + \frac{1}{2} \mathrm{erf}\left(\frac{z}{\sqrt{2}}\right).
\end{aligned}
$$

And for $z < 0$,

$$
\begin{aligned}
\Phi(z) &= 1 - \Phi(-z) \\
&= 1 - \left(\frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\frac{-z}{\sqrt{2}}\right)\right) \\
&= \frac{1}{2} - \frac{1}{2}\mathrm{erf}\left(\frac{-z}{\sqrt{2}}\right) \\
&= \frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\frac{z}{\sqrt{2}}\right),
\end{aligned}
$$

using the fact shown in (b).

(b) Using the substitution $u = -x$, we have

$$
\mathrm{erf}(-z) = \frac{2}{\sqrt{\pi}} \int_0^{-z} e^{-x^2} dx = -\frac{2}{\sqrt{\pi}} \int_0^z e^{-u^2} du = -\mathrm{erf}(z).
$$

21. (a) Find the points of inflection of the $\mathcal{N}(0,1)$ PDF φ, i.e., the points where the curve switches from convex (second derivative positive) to concave (second derivative negative) or vice versa.

(b) Use the result of (a) and a location-scale transformation to find the points of inflection of the $\mathcal{N}(\mu, \sigma^2)$ PDF.

Solution:

(a) The first derivative of $e^{-x^2/2}$ is $xe^{-x^2/2}$ and the second derivative is $e^{-x^2/2}(x^2 - 1)$. So φ is convex on $(-\infty, -1)$, concave on $(-1, 1)$, and convex on $(1, \infty)$, with points of inflection at -1 and 1.

(b) In terms of φ, the $\mathcal{N}(\mu, \sigma^2)$ PDF is

$$
f(x) = \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right).
$$

Then

$$
\begin{aligned}
f'(x) &= \frac{1}{\sigma^2}\varphi'\left(\frac{x-\mu}{\sigma}\right), \\
f''(x) &= \frac{1}{\sigma^3}\varphi''\left(\frac{x-\mu}{\sigma}\right).
\end{aligned}
$$

So f has points of inflection when $\frac{x-\mu}{\sigma}$ is -1 or 1, i.e., at $x = \mu - \sigma$ and $x = \mu + \sigma$.

22. The distance between two points needs to be measured, in meters. The true distance between the points is 10 meters, but due to measurement error we can't measure the distance exactly. Instead, we will observe a value of $10+\epsilon$, where the error ϵ is distributed $\mathcal{N}(0, 0.04)$. Find the probability that the observed distance is within 0.4 meters of the

true distance (10 meters). Give both an exact answer in terms of Φ and an approximate numerical answer.

Solution: Standardizing ϵ (which has mean 0 and standard deviation 0.2), the desired probability is

$$
\begin{aligned}
P(|\epsilon| \leq 0.4) &= P(-0.4 \leq \epsilon \leq 0.4) \\
&= P(-\frac{0.4}{0.2} \leq \frac{\epsilon}{0.2} \leq \frac{0.4}{0.2}) \\
&= P(-2 \leq \frac{\epsilon}{0.2} \leq 2) \\
&= \Phi(2) - \Phi(-2) \\
&= 2\Phi(2) - 1.
\end{aligned}
$$

By the 68-95-99.7% rule, this is approximately 0.95.

23. Alice is trying to transmit to Bob the answer to a yes-no question, using a noisy channel. She encodes "yes" as 1 and "no" as 0, and sends the appropriate value. However, the channel adds noise; specifically, Bob receives what Alice sends plus a $\mathcal{N}(0, \sigma^2)$ noise term (the noise is independent of what Alice sends). If Bob receives a value greater than $1/2$ he interprets it as "yes"; otherwise, he interprets it as "no".

(a) Find the probability that Bob understands Alice correctly.

(b) What happens to the result from (a) if σ is very small? What about if σ is very large? Explain intuitively why the results in these extreme cases make sense.

Solution:

(a) Let a be the value that Alice sends and ϵ be the noise, so $B = a + \epsilon$ is what Bob receives. If $a = 1$, then Bob will understand correctly if and only if $\epsilon > -1/2$. If $a = 0$, then Bob will understand correctly if and only if $\epsilon \leq 1/2$. By symmetry of the Normal, $P(\epsilon > -1/2) = P(\epsilon \leq 1/2)$, so the probability that Bob understands does not depend on a. This probability is

$$
P\left(\epsilon \leq \frac{1}{2}\right) = P\left(\frac{\epsilon}{\sigma} \leq \frac{1}{2\sigma}\right) = \Phi\left(\frac{1}{2\sigma}\right).
$$

(b) If σ is very small, then

$$
\Phi\left(\frac{1}{2\sigma}\right) \approx 1,
$$

since $\Phi(x)$ (like any CDF) goes to 1 as $x \to \infty$. This makes sense intuitively: if there is very little noise, then it's easy for Bob to understand Alice. If σ is very large, then

$$
\Phi\left(\frac{1}{2\sigma}\right) \approx \Phi(0) = 1/2.
$$

Again this makes sense intuitively: if there is a huge amount of noise, then Alice's message will get drowned out (the noise dominates over the signal).

24. A woman is pregnant, with a due date of January 10, 2020. Of course, the actual date on which she will give birth is not necessarily the due date. On a timeline, define time 0 to be the instant when January 10, 2020 begins. Suppose that the time T when the woman gives birth has a Normal distribution, centered at 0 and with standard deviation 8 days. What is the probability that she gives birth on her due date? (Your answer should be in terms of Φ, and simplified.)

Solution: We want to find $P(0 \leq T < 1)$, with $T \sim \mathcal{N}(0, 64)$ measured in days. This is

$$
P(0 \leq T/8 < 1/8) = \Phi(1/8) - \Phi(0) = \Phi(1/8) - 1/2.
$$

25. We will show in the next chapter that if X_1 and X_2 are independent with $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, then $X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. Use this result to find $P(X < Y)$ for $X \sim \mathcal{N}(a, b), Y \sim \mathcal{N}(c, d)$ with X and Y independent.

Hint: Write $P(X < Y) = P(X - Y < 0)$ and then standardize $X - Y$. Check that your answer makes sense in the special case where X and Y are i.i.d.

Solution: Standardizing $X - Y$, we have

$$P(X < Y) = P(X - Y < 0) = P\left(\frac{X - Y - (a - c)}{\sqrt{b + d}} < \frac{-(a - c)}{\sqrt{b + d}} \right) = \Phi\left(\frac{c - a}{\sqrt{b + d}} \right).$$

26. Walter and Carl both often need to travel from Location A to Location B. Walter walks, and his travel time is Normal with mean w minutes and standard deviation σ minutes (travel time can't be negative without using a tachyon beam, but assume that w is so much larger than σ that the chance of a negative travel time is negligible).

Carl drives his car, and his travel time is Normal with mean c minutes and standard deviation 2σ minutes (the standard deviation is larger for Carl due to variability in traffic conditions). Walter's travel time is independent of Carl's. On a certain day, Walter and Carl leave from Location A to Location B at the same time.

(a) Find the probability that Carl arrives first (in terms of Φ and the parameters). For this you can use the important fact, proven in the next chapter, that if X_1 and X_2 are independent with $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, then $X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

(b) Give a fully simplified criterion (*not* in terms of Φ), such that Carl has more than a 50% chance of arriving first if and only if the criterion is satisfied.

(c) Walter and Carl want to make it to a meeting at Location B that is scheduled to begin $w + 10$ minutes after they depart from Location A. Give a fully simplified criterion (*not* in terms of Φ) such that Carl is more likely than Walter to make it on time for the meeting if and only if the criterion is satisfied.

Solution:

(a) Let W be Walter's travel time and C be Carl's. Using standardization,

$$P(C < W) = P(C - W < 0) = P\left(\frac{C - W - (c - w)}{\sigma\sqrt{5}} < \frac{-(c - w)}{\sigma\sqrt{5}} \right) = \Phi\left(\frac{w - c}{\sigma\sqrt{5}} \right).$$

(b) Since Φ is a strictly increasing function with $\Phi(0) = 1/2$ (by symmetry of the Normal), $P(C < W) > 1/2$ if and only if $w > c$.

(c) For Walter, the probability of making it on time is

$$P(W \leq w + 10) = P\left(\frac{W - w}{\sigma} \leq \frac{10}{\sigma} \right) = \Phi\left(\frac{10}{\sigma} \right).$$

For Carl, the probability is

$$P(C \leq w + 10) = P\left(\frac{C - c}{2\sigma} \leq \frac{w - c + 10}{2\sigma} \right) = \Phi\left(\frac{w - c + 10}{2\sigma} \right).$$

The latter is bigger than the former if and only if $(w - c + 10)/(2\sigma) > 10/\sigma$, which simplifies to $w > c + 10$.

27. Let $Z \sim \mathcal{N}(0, 1)$. We know from the 68-95-99.7% rule that there is a 68% chance of Z being in the interval $(-1, 1)$. Give a visual explanation of whether or not there is an interval (a, b) that is shorter than the interval $(-1, 1)$, yet which has at least as large a chance as $(-1, 1)$ of containing Z.

   Solution: The PDF of Z is maximized at 0, and gets smaller and smaller (monotonically) as one moves away from 0. So there is more area under the PDF curve from -1 to 1 then there is area under the curve for any other interval of length 2. More generally, there is more area under the PDF curve from $-c$ to c then there is area under the curve for any other interval of length $2c$, for any $c > 0$. But there is less area for $(-c, c)$ than there is for $(-1, 1)$ for $c < 1$, so no such interval (a, b) exists.

28. Let $Y \sim \mathcal{N}(\mu, \sigma^2)$. Use the fact that $P(|Y - \mu| < 1.96\sigma) \approx 0.95$ to construct a random interval $(a(Y), b(Y))$ (that is, an interval whose endpoints are r.v.s), such that the probability that μ is in the interval is approximately 0.95. This interval is called a *confidence interval* for μ; such intervals are often desired in statistics when estimating unknown parameters based on data.

   Solution: Writing

   $$P(|Y - \mu| < 1.96\sigma) = P(-1.96\sigma < Y - \mu < 1.96\sigma) = P(Y - 1.96\sigma < \mu < Y + 1.96\sigma)$$

   shows that the random interval $(Y - 1.96\sigma, Y + 1.96\sigma)$ is as desired.

29. Let $Y = |X|$, with $X \sim \mathcal{N}(\mu, \sigma^2)$. This is a well-defined continuous r.v., even though the absolute value function is not differentiable at 0 (due to the sharp corner).

   (a) Find the CDF of Y in terms of Φ. Be sure to specify the CDF everywhere.

   (b) Find the PDF of Y.

   (c) Is the PDF of Y continuous at 0? If not, is this a problem as far as using the PDF to find probabilities?

   Solution:

   (a) The CDF of Y is

   $$F(y) = P(|X| \leq y) = P(-y \leq X \leq y) = \Phi\left(\frac{y - \mu}{\sigma}\right) - \Phi\left(\frac{-y - \mu}{\sigma}\right),$$

   for $y \geq 0$ (and $F(y) = 0$ for $y < 0$).

   (b) Let $\varphi(z) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}$ be the $\mathcal{N}(0, 1)$ PDF. By the chain rule, the PDF of Y is

   $$\begin{aligned} f(y) &= \frac{1}{\sigma} \cdot \varphi\left(\frac{y - \mu}{\sigma}\right) + \frac{1}{\sigma} \cdot \varphi\left(\frac{-y - \mu}{\sigma}\right) \\ &= \frac{1}{\sigma} \cdot \varphi\left(\frac{y - \mu}{\sigma}\right) + \frac{1}{\sigma} \cdot \varphi\left(\frac{y + \mu}{\sigma}\right), \end{aligned}$$

   for $y \geq 0$ (and $f(y) = 0$ for $y < 0$).

   (c) The PDF f is not continuous at 0 since $f(y) \to 0$ as $y \to 0$ from the left, whereas

   $$f(y) \to \frac{2}{\sigma}\varphi\left(\frac{\mu}{\sigma}\right) > 0$$

   as $y \to 0$ from the right. This is not a problem for using the PDF to find probabilities, since changing a function at finitely many points does not affect its integral. We have also already dealt with various PDFs that are discontinuous at a finite number of points, such as the $\text{Unif}(0, 1)$ PDF, which is discontinuous at 0 and 1.

30. Ⓢ Let $Z \sim \mathcal{N}(0,1)$ and let S be a random sign independent of Z, i.e., S is 1 with probability $1/2$ and -1 with probability $1/2$. Show that $SZ \sim \mathcal{N}(0,1)$.

    Solution: Condition on S to find the CDF of SZ:

    $$
    \begin{aligned}
    P(SZ \leq x) &= P(SZ \leq x | S = 1)\frac{1}{2} + P(SZ \leq x | S = -1)\frac{1}{2} \\
    &= P(Z \leq x)\frac{1}{2} + P(Z \geq -x)\frac{1}{2} \\
    &= P(Z \leq x)\frac{1}{2} + P(Z \leq x)\frac{1}{2} \\
    &= \Phi(x),
    \end{aligned}
    $$

    where the penultimate equality is by symmetry of the Normal.

31. Ⓢ Let $Z \sim \mathcal{N}(0,1)$. Find $E\left(\Phi(Z)\right)$ *without* using LOTUS, where Φ is the CDF of Z.

    Solution: By universality of the Uniform, $F(X) \sim \text{Unif}(0,1)$ for any continuous random variable X with CDF F. Therefore, $E(\Phi(Z)) = 1/2$.

32. Ⓢ Let $Z \sim \mathcal{N}(0,1)$ and $X = Z^2$. Then the distribution of X is called *Chi-Square with 1 degree of freedom*. This distribution appears in many statistical methods.

    (a) Find a good numerical approximation to $P(1 \leq X \leq 4)$ using facts about the Normal distribution, without querying a calculator/computer/table about values of the Normal CDF.

    (b) Let Φ and φ be the CDF and PDF of Z, respectively. Show that for any $t > 0$, $I(Z > t) \leq (Z/t)I(Z > t)$. Using this and LOTUS, show that $\Phi(t) \geq 1 - \varphi(t)/t$.

    Solution:

    (a) By symmetry of the Normal,

    $$
    P(1 \leq Z^2 \leq 4) = P(-2 \leq Z \leq -1 \text{ or } 1 \leq Z \leq 2) = 2P(1 \leq Z \leq 2) = 2(\Phi(2) - \Phi(1)).
    $$

    By the 68-95-99.7% Rule, $P(-1 \leq Z \leq 1) \approx 0.68$. This says that 32% of the area under the Normal curve is outside of $[-1, 1]$, which by symmetry says that 16% is in $(1, \infty)$. So $\Phi(1) \approx 1 - 0.16 = 0.84$. Similarly, $P(-2 \leq Z \leq 2) \approx 0.95$ gives $\Phi(2) \approx 0.975$. In general, symmetry of the Normal implies that for any $t > 0$,

    $$
    P(-t \leq Z \leq t) = \Phi(t) - \Phi(-t) = 2\Phi(t) - 1.
    $$

    (b) The inequality $I(Z > t) \leq (Z/t)I(Z > t)$ is true since if the indicator is 0 then both sides are 0, and if it is 1 then $Z/t > 1$. So

    $$
    E(I(Z > t)) \leq \frac{1}{t}E(ZI(Z > t)) = \frac{1}{t}\int_{-\infty}^{\infty} zI(z > t)\varphi(z)dz = \frac{1}{t}\int_{t}^{\infty} z\varphi(z)dz.
    $$

    The integral can be done using a substitution: letting $u = z^2/2$, we have

    $$
    \int ze^{-z^2/2}dz = \int e^{-u}du = -e^{-u} + C = -e^{-z^2/2} + C.
    $$

    Thus,

    $$
    P(Z > t) = E(I(Z > t)) \leq \varphi(t)/t,
    $$

    which proves the desired bound on $\Phi(t)$.

33. Let $Z \sim \mathcal{N}(0, 1)$, with CDF Φ. The PDF of Z^2 is the function g given by $g(w) = \frac{1}{\sqrt{2\pi w}} e^{-w/2}$ for $w > 0$ and $g(w) = 0$ for $w \le 0$.

(a) Find expressions for $E(Z^4)$ as integrals in two different ways, one based on the PDF of Z and the other based on the PDF of Z^2.

(b) Find $E(Z^2 + Z + \Phi(Z))$.

(c) Find the CDF of Z^2 in terms of Φ; do this directly, not using the PDF g.

Solution:

(a) Let $W = Z^2$, so $W^2 = Z^4$. By LOTUS,

$$E(Z^4) = \int_{-\infty}^{\infty} z^4 \varphi(z) dz = \int_0^{\infty} w^2 g(w) dw,$$

where $\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ is the PDF of Z, and g is as above. (Using techniques from Chapter 6, it turns out that this reduces to a very simple answer: $E(Z^4) = 3$.)

(b) By linearity, this is $E(Z^2) + E(Z) + E(\Phi(Z))$. The second term is 0 and the first term is 1 since $E(Z) = 0, \operatorname{Var}(Z) = 1$. The third term is $1/2$ since by universality of the Uniform, $\Phi(Z) \sim \operatorname{Unif}(0, 1)$. Thus, the value is $3/2$.

(c) For $w \le 0$, the CDF of Z^2 is 0. For $w > 0$, the CDF of Z^2 is

$$P(Z^2 \le w) = P(-\sqrt{w} \le Z \le \sqrt{w}) = \Phi(\sqrt{w}) - \Phi(-\sqrt{w}) = 2\Phi(\sqrt{w}) - 1.$$

34. Ⓢ Let $Z \sim \mathcal{N}(0, 1)$. A measuring device is used to observe Z, but the device can only handle positive values, and gives a reading of 0 if $Z \le 0$; this is an example of *censored data*. So assume that $X = ZI_{Z>0}$ is observed rather than Z, where $I_{Z>0}$ is the indicator of $Z > 0$. Find $E(X)$ and $\operatorname{Var}(X)$.

Solution: By LOTUS,

$$E(X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} I_{z>0} z e^{-z^2/2} dz = \frac{1}{\sqrt{2\pi}} \int_0^{\infty} z e^{-z^2/2} dz.$$

Letting $u = z^2/2$, we have

$$E(X) = \frac{1}{\sqrt{2\pi}} \int_0^{\infty} e^{-u} du = \frac{1}{\sqrt{2\pi}}.$$

To obtain the variance, note that

$$E(X^2) = \frac{1}{\sqrt{2\pi}} \int_0^{\infty} z^2 e^{-z^2/2} dz = \frac{1}{2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-z^2/2} dz = \frac{1}{2},$$

since a $\mathcal{N}(0, 1)$ r.v. has variance 1. Thus,

$$\operatorname{Var}(X) = E(X^2) - (EX)^2 = \frac{1}{2} - \frac{1}{2\pi}.$$

Note that X is neither purely discrete nor purely continuous, since $X = 0$ with probability $1/2$ and $P(X = x) = 0$ for $x \neq 0$. So X has neither a PDF nor a PMF; but LOTUS still works, allowing us to work with the PDF of Z to study expected values of functions of Z.

Sanity check: The variance is positive, as it should be. It also makes sense that the variance is substantially less than 1 (which is the variance of Z), since we are reducing variability by making the r.v. 0 half the time, and making it nonnegative rather than roaming over the entire real line.

35. Let $Z \sim \mathcal{N}(0, 1)$, and c be a nonnegative constant. Find $E(\max(Z - c, 0))$, in terms of the standard Normal CDF Φ and PDF φ. (This kind of calculation often comes up in quantitative finance.)

Hint: Use LOTUS, and handle the max symbol by adjusting the limits of integration appropriately. As a check, make sure that your answer reduces to $1/\sqrt{2\pi}$ when $c = 0$; this must be the case since we show in Chapter 7 that $E|Z| = \sqrt{2/\pi}$, and we have $|Z| = \max(Z, 0) + \max(-Z, 0)$ so by symmetry

$$E|Z| = E(\max(Z, 0)) + E(\max(-Z, 0)) = 2E(\max(Z, 0)).$$

Solution: Let φ be the $\mathcal{N}(0, 1)$ PDF. By LOTUS,

$$\begin{aligned}
E(\max(Z - c, 0)) &= \int_{-\infty}^{\infty} \max(z - c, 0)\varphi(z)dz \\
&= \int_{c}^{\infty} (z - c)\varphi(z)dz \\
&= \int_{c}^{\infty} z\varphi(z)dz - c\int_{c}^{\infty} \varphi(z)dz \\
&= \frac{-1}{\sqrt{2\pi}}e^{-z^2/2}\Big|_{c}^{\infty} - c(1 - \Phi(c)) \\
&= \varphi(c) - c(1 - \Phi(c)).
\end{aligned}$$

Exponential

36. Ⓢ A post office has 2 clerks. Alice enters the post office while 2 other customers, Bob and Claire, are being served by the 2 clerks. She is next in line. Assume that the time a clerk spends serving a customer has the Exponential(λ) distribution.

(a) What is the probability that Alice is the last of the 3 customers to be done being served?

Hint: No integrals are needed.

(b) What is the expected total time that Alice needs to spend at the post office?

Solution:

(a) Alice begins to be served when either Bob or Claire leaves. By the memoryless property, the additional time needed to serve whichever of Bob or Claire is still there is Expo(λ). The time it takes to serve Alice is also Expo(λ), so by symmetry the probability is $1/2$ that Alice is the last to be done being served.

(b) The expected time spent waiting in line is $\frac{1}{2\lambda}$, since the minimum of two independent Expo(λ) r.v.s is Expo(2λ) (by Example 5.6.3). The expected time spent being served is $\frac{1}{\lambda}$. So the expected total time is

$$\frac{1}{2\lambda} + \frac{1}{\lambda} = \frac{3}{2\lambda}.$$

37. Let T be the time until a radioactive particle decays, and suppose (as is often done in physics and chemistry) that $T \sim \text{Expo}(\lambda)$.

(a) The *half-life* of the particle is the time at which there is a 50% chance that the particle has decayed (in statistical terminology, this is the *median* of the distribution of T). Find the half-life of the particle.

(b) Show that for ϵ a small, positive constant, the probability that the particle decays

in the time interval $[t, t + \epsilon]$, given that it has survived until time t, does not depend on t and is approximately proportional to ϵ.

Hint: $e^x \approx 1 + x$ if $x \approx 0$.

(c) Now consider n radioactive particles, with i.i.d. times until decay $T_1, \ldots, T_n \sim$ Expo(λ). Let L be the first time at which one of the particles decays. Find the CDF of L. Also, find $E(L)$ and Var(L).

(d) Continuing (c), find the mean and variance of $M = \max(T_1, \ldots, T_n)$, the *last* time at which one of the particles decays, *without using calculus*.

Hint: Draw a timeline, apply (c), and remember the memoryless property.

Solution:

(a) Setting $P(T > t) = e^{-\lambda t} = 1/2$ and solving for t, we get that the half-life is

$$t = \frac{\log 2}{\lambda} \approx \frac{0.693}{\lambda}.$$

(This is a very familiar formula in chemistry and physics, and is also closely related to the "rule of 72" or "rule of 70" or "rule of 69" from economics.)

(b) Using the definition of conditional probability and the Expo(λ) CDF,

$$P(T \in [t, t+\epsilon]|T \geq t) = \frac{P(T \in [t, t + \epsilon], T \geq t)}{P(T \geq t)} = \frac{P(t \leq T \leq t + \epsilon)}{P(T \geq t)} = \frac{e^{-\lambda t} - e^{-\lambda(t+\epsilon)}}{e^{-\lambda t}} = 1 - e^{-\lambda \epsilon},$$

which does not depend on t. Alternatively, we can get the same result using the memoryless property: given that $T \geq t$, we have a fresh Expo(λ) starting from time t, so the probability of the additional lifetime beyond t being at most ϵ is the Expo(λ) CDF evaluated at ϵ.

For $\epsilon > 0$ small, the above probability is approximately equal to $1 - (1 - \lambda \epsilon) = \lambda \epsilon$, which is proportional to ϵ.

(c) We have $L \sim$ Expo$(n\lambda)$, since

$$P(L > t) = P(T_1 > t, \ldots, T_n > t) = e^{-n\lambda t},$$

or by Example 5.6.3. The CDF of L is given by $F(t) = 1 - e^{-n\lambda t}$ for $t > 0$ and 0 for $t \leq 0$. So $E(L) = \frac{1}{n\lambda}$ and Var$(L) = \frac{1}{(n\lambda)^2}$.

(d) Let L_1 be the first time at which a particle decays, L_2 be the additional time until another particle decays, $\ldots$, and L_n be the additional time after $n - 1$ particles have decayed until the remaining particle decays. Then $M = L_1 + \cdots + L_n$. By the memoryless property and the previous part, $L_1, L_2, \ldots, L_n$ are independent with $L_j \sim$ Expo$((n - j + 1)\lambda)$. Thus,

$$E(M) = E(L_1) + \cdots + E(L_n) = \frac{1}{\lambda} \sum_{j=1}^n \frac{1}{j},$$

$$\text{Var}(M) = \text{Var}(L_1) + \cdots + \text{Var}(L_n) = \frac{1}{\lambda^2} \sum_{j=1}^n \frac{1}{j^2}.$$

38. Ⓢ Fred wants to sell his car, after moving back to Blissville (where he is happy with the bus system). He decides to sell it to the first person to offer at least \$18,000 for it. Assume that the offers are independent Exponential random variables with mean \$12,000, and that Fred is able to keep getting offers until he obtains one that meets his criterion.

(a) Find the expected number of offers Fred will have.

(b) Find the expected amount of money that Fred will get for the car.

Solution:

(a) The offers on the car are i.i.d. $X_i \sim \text{Expo}(1/12000)$. So the number of offers that are too low is $\text{Geom}(p)$ with $p = P(X_i \geq 12000) = e^{-1.5}$. Including the successful offer, the expected number of offers is thus $(1 - p)/p + 1 = 1/p = e^{1.5}$.

(b) Let N be the number of offers, so the sale price of the car is X_N. Note that

$$E(X_N) = E(X|X \geq 18000)$$

for $X \sim \text{Expo}(1/12000)$, since the successful offer is an Exponential for which our information is that the value is at least \$18,000. To compute this, remember the memoryless property! For any $a > 0$, if $X \sim \text{Expo}(\lambda)$ then the distribution of $X - a$ given $X > a$ is itself $\text{Expo}(\lambda)$. So

$$E(X|X \geq 18000) = 12000 + E(X) = 30000,$$

which shows that Fred's expected sale price is \$30,000.

39. As in the previous problem, Fred wants to sell his car, and the offers for his car are i.i.d. Exponential r.v.s with mean \$12,000. Assume now though that he will wait until he has 3 offers (no matter how large or small they are), and then accept the largest of the 3 offers. Find the expected amount of money that Fred will get for his car.

Solution: Let $\lambda = 1/12000$. Reasoning as in Example 5.6.5, the expected value of the maximum of 3 i.i.d. $\text{Expo}(\lambda)$ r.v.s is

$$\frac{1}{3\lambda} + \frac{1}{2\lambda} + \frac{1}{\lambda} = 4000 + 6000 + 12000 = 22000,$$

so the expected amount that Fred will get for his car is \$22,000.

40. (a) Fred visits Blotchville again. He finds that the city has installed an electronic display at the bus stop, showing the time when the previous bus arrived. The times between arrivals of buses are still independent Exponentials with mean 10 minutes. Fred waits for the next bus, and then records the time between that bus and the previous bus. On average, what length of time between buses does he see?

(b) Fred then visits Blunderville, where the times between buses are also 10 minutes on average, and independent. Yet to his dismay, he finds that on average he has to wait more than 1 hour for the next bus when he arrives at the bus stop! How is it possible that the average Fred-to-bus time is greater than the average bus-to-bus time even though Fred arrives at some time between two bus arrivals? Explain this intuitively, and construct a specific discrete distribution for the times between buses showing that this is possible.

Solution:

(a) Let T_1 be how long Fred missed the previous bus by, and T_2 be how long he has to wait for the next bus. So the time between buses that he records is the value of T_1+T_2. By the memoryless property, $E(T_2) = 10$. Reversing the arrow of time and again using the memoryless property, we also have $E(T_1) = 10$. Thus, $E(T_1+T_2) = E(T_1)+E(T_2) = 20$. This says that on average, Fred records a 20 minute gap between buses even though the average time between buses is only 10 minutes! This is an example of a phenomenon known as *length-biased sampling*. Intuitively, the reason is that it is easier to arrive during a long interval between buses than during a short interval between buses.

(b) Intuitively, this is possible because of the length-biasing phenomenon mentioned

above: we can choose a distribution such that most of the between-bus intervals are short but some are extremely long. Suppose that the distinct possible times between buses are $t_1, t_2, \ldots, t_n$, with probabilities $p_1, \ldots, p_n$, and let T be an r.v. with this distribution. We will derive a very neat expression for the average Fred-to-bus time in terms of the first two moments of T (deriving this general result was not required for the problem).

Fred's probability of arriving during a between bus interval of length t_j is proportional to $p_j t_j$. To see this, imagine a large number b of bus arrivals, with bp_j of length t_j; then the total length of the t_j intervals divided by the total length is

$$\frac{bp_j t_j}{bp_1 t_1 + \cdots + bp_n t_n} = \frac{p_j t_j}{p_1 t_1 + \cdots + p_n t_n}.$$

Given that he arrives in a t_j interval, Fred's expected waiting time is $t_j/2$. Thus, his average waiting time is

$$\frac{p_1 t_1}{p_1 t_1 + \cdots + p_n t_n} \frac{t_1}{2} + \cdots + \frac{p_n t_n}{p_1 t_1 + \cdots + p_n t_n} \frac{t_n}{2} = \frac{p_1 t_1^2 + \cdots + p_n t_n^2}{2(p_1 t_1 + \cdots + p_n t_n)} = \frac{E(T^2)}{2E(T)}.$$

For a concrete example where Fred's average waiting time is more than an hour, suppose that when a bus arrives, the next bus comes 10^4 minutes later with probability $1/10^4$, and comes $10000/1111 \approx 9.0009$ minutes later with probability $1 - 1/10^4$. Then the average time between buses is still 10 minutes. But by the above, the average waiting time that Fred experiences is approximately 504 minutes!

41. Fred and Gretchen are waiting at a bus stop in Blotchville. Two bus routes, Route 1 and Route 2, have buses that stop at this bus stop. For Route i, buses arrive according to a Poisson process with rate λ_i buses/minute. The Route 1 process is independent of the Route 2 process. Fred is waiting for a Route 1 bus, and Gretchen is waiting for a Route 2 bus.

(a) Given that Fred has already waited for 20 minutes, on average how much longer will he have to wait for his bus?

(b) Find the probability that at least n Route 1 buses will pass by before the first Route 2 bus arrives. The following result from Chapter 7 may be useful here: for independent random variables $X_1 \sim \text{Expo}(\lambda_1)$, $X_2 \sim \text{Expo}(\lambda_2)$, we have $P(X_1 < X_2) = \lambda_1/(\lambda_1 + \lambda_2)$.

(c) For this part only, assume that $\lambda_1 = \lambda_2 = \lambda$. Find the expected time it will take until both Fred and Gretchen have caught their buses.

Solution:

(a) By the memoryless property, Fred's average additional wait is $1/\lambda_1$ minutes.

(b) By the hint, there is a $\lambda_1/(\lambda_1 + \lambda_2)$ chance that a Route 1 bus will pass by first. Given that this does happen, by the memoryless property there is a $\lambda_1/(\lambda_1 + \lambda_2)$ chance that another Route 1 bus will pass by before a Route 2 bus comes. Continuing in this way, the desired probability is $(\lambda_1/(\lambda_1 + \lambda_2))^n$.

(c) Let T_1 be the time until one of them has caught their bus and T_2 be the additional time until they both have. Then $T_1 \sim \text{Expo}(2\lambda)$ since T_1 is the minimum of 2 i.i.d. $\text{Expo}(\lambda)$ r.v.s., and $T_2 \sim \text{Expo}(\lambda)$ by the memoryless property. So

$$E(T_1 + T_2) = ET_1 + ET_2 = \frac{1}{2\lambda} + \frac{1}{\lambda} = \frac{3}{2\lambda}.$$

Thus, the expected time is $\frac{3}{2\lambda}$ minutes.

42. Ⓢ Joe is waiting in continuous time for a book called *The Winds of Winter* to be released. Suppose that the waiting time T until news of the book's release is posted, measured in years relative to some starting point, has an Exponential distribution with $\lambda = 1/5$.

    Joe is not so obsessive as to check multiple times a day; instead, he checks the website *once* at the end of each day. Therefore, he observes the day on which the news was posted, rather than the exact time T. Let X be this measurement, where $X = 0$ means that the news was posted within the first day (after the starting point), $X = 1$ means it was posted on the second day, etc. (assume that there are 365 days in a year). Find the PMF of X. Is this a named distribution that we have studied?

    Solution: The event $X = k$ is the same as the event $k \leq 365T < k+1$, i.e., $X = \lfloor 365T \rfloor$, where $\lfloor t \rfloor$ is the floor function of t (the greatest integer less than or equal to t). The CDF of T is $F_T(t) = 1 - e^{-t/5}$ for $t > 0$ (and 0 for $t \leq 0$). So

    $$P(X = k) = P\left(\frac{k}{365} \leq T < \frac{k+1}{365}\right) = F_T\left(\frac{k+1}{365}\right) - F_T\left(\frac{k}{365}\right) = e^{-k/1825} - e^{-(k+1)/1825}.$$

    This factors as $\left(e^{-1/1825}\right)^k (1 - e^{-1/1825})$, which shows that $X \sim \text{Geom}(1 - e^{-1/1825})$.

    Sanity check: A Geometric distribution is plausible for a waiting time, and does take values $0, 1, 2, \ldots$. The parameter $p = 1 - e^{-1/1825} \approx 0.0005$ is very small, which reflects both the fact that there are a lot of days in a year (so each day is unlikely) and the fact that the author is not known for the celerity of his writing.

43. The Exponential is the analog of the Geometric in continuous time. This problem explores the connection between Exponential and Geometric in more detail, asking what happens to a Geometric in a limit where the Bernoulli trials are performed faster and faster but with smaller and smaller success probabilities.

    Suppose that Bernoulli trials are being performed in continuous time; rather than only thinking about first trial, second trial, etc., imagine that the trials take place at points on a timeline. Assume that the trials are at regularly spaced times $0, \Delta t, 2\Delta t, \ldots$, where Δt is a small positive number. Let the probability of success of each trial be $\lambda \Delta t$, where λ is a positive constant. Let G be the number of failures before the first success (in discrete time), and T be the time of the first success (in continuous time).

    (a) Find a simple equation relating G to T.

    Hint: Draw a timeline and try out a simple example.

    (b) Find the CDF of T.

    Hint: First find $P(T > t)$.

    (c) Show that as $\Delta t \to 0$, the CDF of T converges to the $\text{Expo}(\lambda)$ CDF, evaluating all the CDFs at a fixed $t \geq 0$.

    Hint: Use the compound interest limit (see the math appendix).

    Solution:

    (a) At time T, there will have been G failures and 1 success. The jth trial occurs at time $(j - 1)\Delta t$. So
    $$T = (G + 1 - 1)\Delta t = G\Delta t.$$

    (b) For any nonnegative integer n, $P(G > n) = (1 - \lambda \Delta t)^{n+1}$, since $G > n$ if and only if the first $n + 1$ trials are failures. More generally, for any real $x \geq 0$,
    $$P(G > x) = (1 - \lambda \Delta t)^{\lfloor x \rfloor + 1},$$

where $\lfloor x \rfloor$ is the floor function of x. Thus, for $t \geq 0$ we have

$$P(T > t) = P\left(G > \frac{t}{\Delta t}\right) = (1 - \lambda \Delta t)^{\lfloor \frac{t}{\Delta t} \rfloor + 1}.$$

The CDF of T is

$$P(T \leq t) = 1 - (1 - \lambda \Delta t)^{\lfloor \frac{t}{\Delta t} \rfloor + 1},$$

for $t \geq 0$ (and $P(T \leq t) = 0$ for $t < 0$).

(c) For $t = 0$, the CDF is

$$P(T \leq 0) = P(T = 0) = \lambda \Delta t,$$

which goes to 0 as $\Delta t \to 0$. Now fix $t > 0$. Take $\Delta t = \frac{1}{n}$, and let $n \to \infty$. Since

$$nt - 1 < \lfloor nt \rfloor \leq nt,$$

we can omit the floor of $P(T \leq t)$ when taking the limit (the 1 in $nt - 1$ will have no effect on the limit). By the compound interest limit,

$$\lim_{n \to \infty} P(T \leq t) = 1 - \lim_{n \to \infty} \left(1 - \frac{\lambda}{n}\right)^{nt+1} = 1 - e^{-\lambda t},$$

as desired.

44. The *Laplace distribution* has PDF
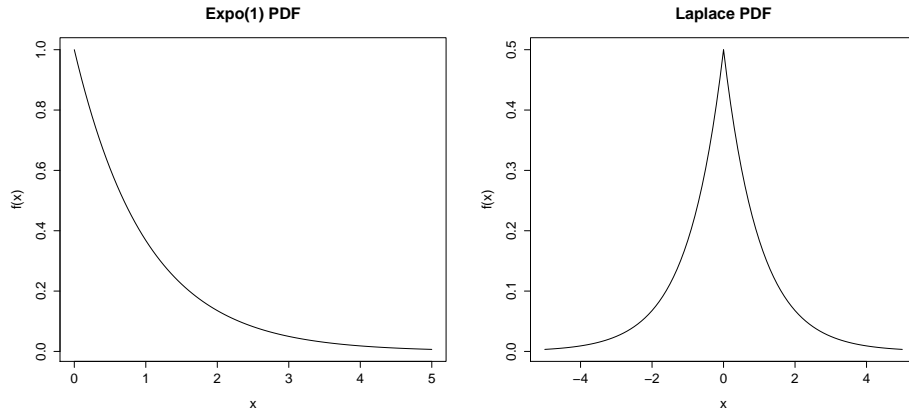
$$f(x) = \frac{1}{2} e^{-|x|}$$

for all real x. The Laplace distribution is also called a *symmetrized Exponential* distribution. Explain this in the following two ways.

(a) Plot the PDFs and explain how they relate.

(b) Let $X \sim \text{Expo}(1)$ and S be a random sign (1 or -1, with equal probabilities), with S and X independent. Find the PDF of SX (by first finding the CDF), and compare the PDF of SX and the Laplace PDF.

Solution:

(a) The Expo(1) PDF and Laplace PDF are plotted above. Note the sharp cusp of the

Laplace PDF at 0. To obtain the Laplace PDF from the Expo(1) PDF, we can reflect the Expo(1) about the vertical axis to obtain an even function. The area under the

curve is now 2, so we rescale by dividing by 2, which gives the PDF of a distribution that is symmetric about 0.

(b) The CDF of SX is given by

$$P(SX \leq a) = P(X \leq a|S = 1)P(S = 1) + P(X \geq -a|S = -1)P(S = -1)$$

$$= \frac{1}{2}P(X \leq a) + \frac{1}{2}P(X \geq -a)$$

$$= \begin{cases} \frac{1}{2}(1 - e^{-a}) + \frac{1}{2}, & \text{if } a \geq 0, \\ \frac{1}{2}e^a, & \text{if } a < 0. \end{cases}$$

So the PDF of SX is given by

$$f(a) = \begin{cases} \frac{1}{2}e^{-a}, & \text{if } a \geq 0, \\ \frac{1}{2}e^a, & \text{if } a < 0, \end{cases}$$

which simplifies to $f(a) = \frac{1}{2}e^{-|a|}$, for all real a.

45. Emails arrive in an inbox according to a Poisson process with rate 20 emails per hour. Let T be the time at which the 3rd email arrives, measured in hours after a certain fixed starting time. Find $P(T > 0.1)$ without using calculus.

Hint: Apply the count-time duality.

Solution: By the count-time duality, $P(T > 0.1) = P(N \leq 2)$, where N is the number of emails that arrive in the first 0.1 hours. We have $N \sim \text{Pois}(2)$, so

$$P(T > 0.1) = P(N \leq 2) = e^{-2} + e^{-2} \cdot 2 + e^{-2} \cdot 2^2/2! \approx 0.6767.$$

46. Let T be the lifetime of a certain person (how long that person lives), and let T have CDF F and PDF f. The *hazard function* of T is defined by

$$h(t) = \frac{f(t)}{1 - F(t)}.$$

(a) Explain why h is called the hazard function and in particular, why $h(t)$ is the probability density for death at time t, given that the person survived up until then.

(b) Show that an Exponential r.v. has constant hazard function and conversely, if the hazard function of T is a constant then T must be $\text{Expo}(\lambda)$ for some λ.

Solution:

(a) Given that $T > t_0$, the conditional CDF of T is

$$P(T \leq t|T \geq t_0) = \frac{P(t_0 \leq T \leq t)}{P(T \geq t_0)} = \frac{F(t) - F(t_0)}{1 - F(t_0)}$$

for $t \geq t_0$ (and 0 otherwise). So the conditional PDF of T given $T \geq t_0$ is

$$f(t|T \geq t_0) = \frac{f(t)}{1 - F(t_0)}, \text{ for } t \geq t_0.$$

The conditional PDF of T at t_0, given that the person survived up until t_0, is then $f(t_0)/(1 - F(t_0)) = h(t_0)$. Alternatively, we can use the hybrid form of Bayes' rule:

$$f(t|T \geq t) = \frac{P(T \geq t|T = t)f(t)}{P(T \geq t)} = \frac{f(t)}{1 - F(t)}.$$

Thus, $h(t)$ gives the probability density of death at time t given that the person has

survived up until then. This is a natural way to measure the instantaneous hazard of death at some time since it accounts for the person having survived up until that time.

(b) Let $T \sim \text{Expo}(\lambda)$. Then the hazard function is $h(t) = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda$, which is a constant. Conversely, suppose that $h(t) = \lambda$ for all t, with $\lambda > 0$ a constant. Let $s = S(t) = 1 - F(t)$ (this is called the *survival function*). We have $S'(t)/S(t) = -\lambda$, which we can write as $ds/s = -\lambda dt$. Integrating both sides and noting that $s = 1$ when $t = 0$, we have $\log s = -\lambda t$, so $S(t) = e^{-\lambda t}$. Thus, $T \sim \text{Expo}(\lambda)$.

47. Let T be the lifetime of a person (or animal or gadget), with CDF F and PDF f. Let h be the hazard function, defined as in the previous problem. If we know F then we can calculate f, and then in turn we can calculate h. In this problem, we consider the reverse problem: how to recover F and f from knowing h.

(a) Show that the CDF and hazard function are related by

$$F(t) = 1 - \exp\left(-\int_0^t h(s)ds\right),$$

for all $t > 0$.

Hint: Let $G(t) = 1 - F(t)$ be the survival function, and consider the derivative of $\log G(t)$.

(b) Show that the PDF and hazard function are related by

$$f(t) = h(t)\exp\left(-\int_0^t h(s)ds\right),$$

for all $t > 0$.

Hint: Apply the result of (a).

Solution:

(a) As suggested by the hint, let's take the derivative of $\log G(t)$, where $G(t) = 1 - F(t)$:

$$\frac{d\log G(t)}{dt} = \frac{G'(t)}{G(t)} = \frac{-f(t)}{G(t)} = -h(t).$$

Integrating both sides with respect to t gives

$$\int_0^{t_0} \frac{d\log G(t)}{dt}dt = -\int_0^{t_0} h(t)dt,$$

for any $t_0 > 0$. The left-hand side is

$$\int_0^{t_0} \frac{d\log G(t)}{dt}dt = \log G(t_0) - \log G(0) = \log G(t_0),$$

since $G(0) = 1 - F(0) = 1$. Thus,

$$\log G(t) = -\int_0^t h(s)ds,$$

which shows that

$$F(t) = 1 - \exp\left(-\int_0^t h(s)ds\right).$$

(b) Differentiating both sides of the result from (a), we have

$$f(t) = h(t)\exp\left(-\int_0^t h(s)ds\right),$$

since $\frac{d}{dt}\int_0^t h(s)ds = h(t)$ (by the fundamental theorem of calculus).

48. (S) Find $E(X^3)$ for $X \sim \text{Expo}(\lambda)$, using LOTUS and the fact that $E(X) = 1/\lambda$ and $\text{Var}(X) = 1/\lambda^2$, and integration by parts at most once. In the next chapter, we'll learn how to find $E(X^n)$ for all n.

Solution: By LOTUS,

$$E(X^3) = \int_0^\infty x^3 \lambda e^{-\lambda x} dx = -x^3 e^{-\lambda x} \Big|_0^\infty + \frac{3}{\lambda} \int_0^\infty x^2 \lambda e^{-\lambda x} dx$$

$$= \frac{3}{\lambda} E(X^2) = \frac{3}{\lambda}(\text{Var}(X) + (EX)^2) = \frac{6}{\lambda^3},$$

where the second equality uses integration by parts, letting $u = x^3$ and $dv = \lambda e^{-\lambda x} dx$ and we multiply the second term by 1 written as λ/λ.

49. (S) The *Gumbel distribution* is the distribution of $-\log X$ with $X \sim \text{Expo}(1)$.

(a) Find the CDF of the Gumbel distribution.

(b) Let $X_1, X_2, \ldots$ be i.i.d. Expo(1) and let $M_n = \max(X_1, \ldots, X_n)$. Show that $M_n - \log n$ converges in distribution to the Gumbel distribution, i.e., as $n \to \infty$ the CDF of $M_n - \log n$ converges to the Gumbel CDF.

Solution:

(a) Let G be Gumbel and $X \sim \text{Expo}(1)$. The CDF is

$$P(G \le t) = P(-\log X \le t) = P(X \ge e^{-t}) = e^{-e^{-t}}$$

for all real t.

(b) The CDF of $M_n - \log n$ is

$$P(M_n - \log n \le t) = P(X_1 \le t + \log n, \ldots, X_n \le t + \log n) = P(X_1 \le t + \log n)^n.$$

Using the Expo CDF and the fact that $(1 + \frac{x}{n})^n \to e^x$ as $n \to \infty$, this becomes

$$(1 - e^{-(t + \log n)})^n = (1 - \frac{e^{-t}}{n})^n \to e^{-e^{-t}}.$$

Mixed practice

50. Explain intuitively why $P(X < Y) = P(Y < X)$ if X and Y are i.i.d., but equality may not hold if X and Y are not independent or not identically distributed.

Solution: If X and Y are i.i.d., then $P(X < Y) = P(Y < X)$ by symmetry; the problem of finding $P(X < Y)$ has exactly the same structure as that of finding $P(Y < X)$.

But X and Y are not interchangeable if they have different distributions, since then the symmetry is broken and it may be the case, for example, that the distribution of Y tends to produce much larger values than the distribution of X. As an extreme case, we can even consider examples where the support of Y lies strictly to the right of the support of X, so $P(X < Y) = 1$ and $P(Y < X) = 0$.

If X and Y are dependent, then we also can't conclude that $P(X < Y) = P(Y < X)$, since then the structure of $P(X < Y)$ is different from the structure of $P(Y < X)$. For example, the dependence could be rigged so that X is usually less than Y, even though they have the same distribution. Exercise 3.42 gives a specific example of this.

51. Let X be an r.v. (discrete or continuous) such that $0 \leq X \leq 1$ always holds. Let $\mu = E(X)$.

(a) Show that
$$\text{Var}(X) \leq \mu - \mu^2 \leq \frac{1}{4}.$$

Hint: With probability 1, we have $X^2 \leq X$.

(b) Show that there is only one possible distribution for X for which $\text{Var}(X) = 1/4$. What is the name of this distribution?

Solution:

(a) Using the hint,
$$\text{Var}(X) = E(X^2) - (EX)^2 \leq EX - (EX)^2 = \mu - \mu^2.$$

Furthermore, $\mu - \mu^2 \leq 1/4$, since $\mu - \mu^2$ is maximized at $\mu = 1/2$, where the value is $1/2 - (1/2)^2 = 1/4$.

(b) To make equality hold above, we need $X - X^2 = 0$ with probability 1, since otherwise $E(X) - E(X^2) = E(X - X^2) > 0$, which would make $\text{Var}(X) < 1/4$. So X can take on only two values, 0 or 1. And to make $\mu - \mu^2 = 1/4$, we need $\mu = 1/2$. So $\text{Bern}(1/2)$ is the only possible distribution for X which makes $\text{Var}(X) = 1/4$.

52. The Rayleigh distribution from Example 5.1.7 has PDF
$$f(x) = xe^{-x^2/2}, \quad x > 0.$$

Let X have the Rayleigh distribution.

(a) Find $E(X)$ without using much calculus, by interpreting the integral in terms of known results about the Normal distribution.

(b) Find $E(X^2)$.

Hint: A nice approach is to use LOTUS and the substitution $u = x^2/2$, and then interpret the resulting integral in terms of known results about the Exponential distribution.

Solution:

(a) By symmetry,
$$E(X) = \int_0^\infty x^2 e^{-x^2/2} dx = \frac{1}{2} \int_{-\infty}^\infty x^2 e^{-x^2/2} dx.$$

The right-hand side is reminiscent of $E(Z^2)$ for $Z \sim \mathcal{N}(0,1)$, and we know that $E(Z^2) = \text{Var}(Z) = 1$. So
$$E(X) = \frac{\sqrt{2\pi}}{2} \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} x^2 e^{-x^2/2} dx = \frac{\sqrt{2\pi}}{2}.$$

(b) Let $Y \sim \text{Expo}(1)$. By LOTUS and the substitution $y = x^2/2$,
$$E(X^2) = \int_0^\infty x^2 e^{-x^2/2} x dx = \int_0^\infty 2y e^{-y} dy = 2E(Y) = 2.$$

53. ⓢ Consider an experiment where we observe the value of a random variable X, and estimate the value of an unknown constant θ using some random variable $T = g(X)$ that is a function of X. The r.v. T is called an *estimator*. Think of X as the data observed in the experiment, and θ as an unknown parameter related to the distribution of X.

For example, consider the experiment of flipping a coin n times, where the coin has an unknown probability θ of Heads. After the experiment is performed, we have observed the value of $X \sim \text{Bin}(n, \theta)$. The most natural estimator for θ is then X/n.

The *bias* of an estimator T for θ is defined as $b(T) = E(T) - \theta$. The *mean squared error* is the average squared error when using $T(X)$ to estimate θ:

$$\text{MSE}(T) = E(T - \theta)^2.$$

Show that

$$\text{MSE}(T) = \text{Var}(T) + (b(T))^2.$$

This implies that for fixed MSE, lower bias can only be attained at the cost of higher variance and vice versa; this is a form of the *bias-variance tradeoff*, a phenomenon which arises throughout statistics.

Solution: Using the fact that adding a constant does not affect variance, we have

$$\begin{aligned}
\text{Var}(T) &= \text{Var}(T - \theta) \\
&= E(T - \theta)^2 - (E(T - \theta))^2 \\
&= \text{MSE}(T) - (b(T))^2,
\end{aligned}$$

which proves the desired identity.

54. ⓢ (a) Suppose that we have a list of the populations of every country in the world.

Guess, without looking at data yet, what percentage of the populations have the digit 1 as their first digit (e.g., a country with a population of 1,234,567 has first digit 1 and a country with population 89,012,345 does not).

(b) After having done (a), look through a list of populations and count how many start with a 1. What percentage of countries is this? *Benford's law* states that in a very large variety of real-life data sets, the first digit approximately follows a particular distribution with about a 30% chance of a 1, an 18% chance of a 2, and in general

$$P(D = j) = \log_{10}\left(\frac{j+1}{j}\right), \text{ for } j \in \{1, 2, 3, \ldots, 9\},$$

where D is the first digit of a randomly chosen element. (Exercise 6 from Chapter 3 asks for a proof that this is a valid PMF.) How closely does the percentage found in the data agree with that predicted by Benford's law?

(c) Suppose that we write the random value in some problem (e.g., the population of a random country) in scientific notation as $X \times 10^N$, where N is a nonnegative integer and $1 \leq X < 10$. Assume that X is a continuous r.v. with PDF

$$f(x) = c/x, \text{ for } 1 \leq x \leq 10,$$

and 0 otherwise, with c a constant. What is the value of c (be careful with the bases of logs)? Intuitively, we might hope that the distribution of X does not depend on the choice of units in which X is measured. To see whether this holds, let $Y = aX$ with $a > 0$. What is the PDF of Y (specifying where it is nonzero)?

(d) Show that if we have a random number $X \times 10^N$ (written in scientific notation) and X has the PDF f from (c), then the first digit (which is also the first digit of X) has Benford's law as PMF.

Hint: What does $D = j$ correspond to in terms of the values of X?

Solution:

(a) What did you guess?

(b) According to Wikipedia (as of October 15, 2011), 63 out of 225 countries have a total population whose first digit is 1, which is 28%. (This depends slightly on whether certain territories included in the Wikipedia list should be considered as "countries" but the purpose of this problem is not to delve into the sovereignty or nation-status of territories). It is striking that 28% of the countries have first digit 1, as this is so much higher than one would expect from guessing that the first digit is equally likely to be any of $1, 2, \ldots, 9$. This is an example of Benford's law; similar phenomena have been observed in many different settings (such as with lengths of rivers, physical constants, and stock prices).

(c) The PDF ($f(x) = c/x, 1 \leq x \leq 10$) must integrate to one, by definition; therefore

$$1 = c \int_1^{10} \frac{dx}{x} = c(\ln 10 - \ln 1) = c \ln 10.$$

So the constant of proportionality $c = 1/\ln 10 = \log_{10} e$). If $Y = aX$ (a *change in scale*), then Y has pdf c/y with the same value of c as before, except now $a \leq y \leq 10a$ rather than $1 \leq x \leq 10$. So the PDF takes the same form for aX as for X, but over a different range.

(d) The first digit $D = d$ when $d \leq X < d+1$. The probability of this is

$$P(D = d) = P(d \leq X < d+1) = \int_d^{d+1} \frac{1}{x \ln 10} dx,$$

which is $\log_{10}(d+1) - \log_{10}(d)$, as desired.

55. ⓢ (a) Let $X_1, X_2, \ldots$ be independent $\mathcal{N}(0, 4)$ r.v.s., and let J be the smallest value of j such that $X_j > 4$ (i.e., the index of the first X_j exceeding 4). In terms of Φ, find $E(J)$.

(b) Let f and g be PDFs with $f(x) > 0$ and $g(x) > 0$ for all x. Let X be a random variable with PDF f. Find the expected value of the ratio

$$R = \frac{g(X)}{f(X)}.$$

Such ratios come up very often in statistics, when working with a quantity known as a *likelihood ratio* and when using a computational technique known as *importance sampling*.

(c) Define

$$F(x) = e^{-e^{-x}}.$$

This is a CDF and is a continuous, strictly increasing function. Let X have CDF F, and define $W = F(X)$. What are the mean and variance of W?

Solution:

(a) We have $J - 1 \sim \text{Geom}(p)$ with $p = P(X_1 > 4) = P(X_1/2 > 2) = 1 - \Phi(2)$, so $E(J) = 1/(1 - \Phi(2))$.

(b) By LOTUS,

$$E\frac{g(X)}{f(X)} = \int_{-\infty}^{\infty} \frac{g(x)}{f(x)} f(x) dx = \int_{-\infty}^{\infty} g(x) dx = 1.$$

(c) By universality of the Uniform, $W \sim \text{Unif}(0,1)$. Alternatively, we can compute directly that the CDF of W is

$$P(W \le w) = P(F(X) \le w) = P(X \le F^{-1}(w)) = F(F^{-1}(w)) = w$$

for $0 < w < 1$, so again we have $W \sim \text{Unif}(0,1)$. Thus, $E(W) = 1/2$ and $\text{Var}(W) = 1/12$.

56. Let $X, Y, Z \sim \mathcal{N}(0,1)$ be i.i.d.

   (a) Find an expression for $E\left(Z^2 \, \Phi(Z)\right)$ as an integral.
   (b) Find $P(\Phi(Z) < 2/3)$.

   (c) Find $P(X < Y < Z)$.

   Solution:

   (a) By LOTUS,

   $$E\left(Z^2 \, \Phi(Z)\right) = \int_{-\infty}^{\infty} z^2 \, \Phi(z) \, \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz.$$

   (b) By Universality of the Uniform, $\Phi(Z) \sim \text{Unif}(0,1)$. So

   $$P(\Phi(Z) < 2/3) = (\text{length of the interval } (0, 2/3)) = 2/3.$$

   Alternatively, we can calculate the probability directly:

   $$P(\Phi(Z) < 2/3) = P(Z < \Phi^{-1}(2/3)) = \Phi(\Phi^{-1}(2/3)) = 2/3.$$

   (c) By symmetry, all orderings of X, Y, Z are equally likely, so

   $$P(X < Y < Z) = \frac{1}{3!} = \frac{1}{6}.$$

57. Let $X, Y, Z \sim \mathcal{N}(0,1)$ be i.i.d., and $W = (\Phi(Z))^2$.

   (a) Find the CDF and PDF of W.

   (b) Let f_W be the PDF of W and φ be the PDF of Z. Find unsimplified expressions for $E(W^3)$ as integrals in two different ways, one based on f_W and one based on φ.

   (c) Find $P(X + 2Y < 2Z + 3)$, in terms of Φ.

   Hint: Move all of the r.v.s to one side of the inequality.

   Solution:

   (a) Let $U = \Phi(Z)$. By Universality of the Uniform, $U \sim \text{Unif}(0,1)$. The CDF of W is

   $$F_W(w) = P(W \le w) = P(U \le \sqrt{w}) = \sqrt{w}$$

   for $0 < w < 1$. For $w \le 0$, $F_W(w) = 0$; for $w \ge 1$, $F_W(w) = 1$. The PDF of W is

   $$f_W(w) = F_W'(w) = \frac{1}{2} w^{-1/2}$$

   for $0 < w < 1$, and $f_W(w) = 0$ otherwise. (This implies that W has a *Beta distribution*, specifically Beta(1/2,1), as defined in Chapter 8.)

   (b) By LOTUS,

   $$E(W^3) = \int_0^1 w^3 f_W(w) dw = \int_{-\infty}^{\infty} (\Phi(z))^6 \, \varphi(z) dz.$$

   (c) Note that $X + 2Y - 2Z \sim \mathcal{N}(0,9)$. Therefore,

   $$P(X + 2Y < 2Z + 3) = P\left(\frac{X + 2Y - 2Z}{3} < 1\right) = \Phi(1).$$

58. Let $Z \sim \mathcal{N}(0,1)$ and $Y = \max(Z, 0)$. So Y is Z if $Z > 0$, and Y is 0 if $Z \leq 0$.

(a) Find an expression for $E(Y)$ as an integral (which can be unsimplified).

(b) Let $Y_1, Y_2, \ldots$ be independent r.v.s, each with the same distribution as Y. Let $N = \min\{n : Y_n = 0\}$, i.e., N is the smallest value such that $Y_N = 0$. Find $E(N)$.

(c) Find the CDF of Y in terms of Φ. (Be sure to specify it for all real numbers.)

Solution:

(a) By LOTUS,

$$E(Y) = \int_{-\infty}^{\infty} \max(z, 0) \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \int_0^{\infty} \frac{z}{\sqrt{2\pi}} e^{-z^2/2} dz.$$

(Making the substitution $u = z^2/2$, this integral evaluates to $\frac{1}{\sqrt{2\pi}}$.)

(b) We have $N \sim \text{FS}(1/2)$ since $P(Y_j = 0) = P(Z \leq 0) = 1/2$. Thus, $E(N) = 2$.

(c) The CDF of Y is given by

$$P(Y \leq y) = P(\max(Z, 0) \leq y) = P(Z \leq y \text{ and } 0 \leq y) = \begin{cases} \Phi(y), & \text{if } y \geq 0; \\ 0, & \text{if } y < 0. \end{cases}$$

59. The unit circle $\{(x, y) : x^2 + y^2 = 1\}$ is divided into three arcs by choosing three random points A, B, C on the circle (independently and uniformly), forming arcs between A and B, between A and C, and between B and C. Let L be the length of the arc containing the point $(1, 0)$. What is $E(L)$? Study this by working through the following steps.

(a) Explain what is wrong with the following argument: "The total length of the arcs is 2π, the circumference of the circle. So by symmetry and linearity, each arc has length $2\pi/3$ on average. Referring to the arc containing $(1, 0)$ is just a way to specify one of the arcs (it wouldn't matter if $(1, 0)$ were replaced by $(0, -1)$ or any other specific point on the circle in the statement of the problem). So the expected value of L is $2\pi/3$."

(b) Let the arc containing $(1, 0)$ be divided into two pieces: the piece extending counterclockwise from $(1, 0)$ and the piece extending clockwise from $(1, 0)$. Write $L = L_1 + L_2$, where L_1 and L_2 are the lengths of the counterclockwise and clockwise pieces, respectively. Find the CDF, PDF, and expected value of L_1.

(c) Use (b) to find $E(L)$.

Solution:

(a) It is true that the expected length of the arc from A to B is $2\pi/3$, and likewise for the arc from A to C and the arc from B to C. But the arc containing $(1, 0)$ is larger than $2\pi/3$ on average, since a larger arc is more likely to contain $(1, 0)$ than a smaller arc. The same reasoning holds if $(1, 0)$ is replaced by $(0, -1)$ or any other pre-specified point on the circle.

(b) Let $0 < x < 2\pi$. Measuring angles in radians, counterclockwise with 0 at $(1, 0)$, we have

$$P(L_1 > x) = P(\text{none of } A, B, C \text{ are at an angle in } [0, x]) = \left(\frac{2\pi - x}{2\pi}\right)^3,$$

so the CDF of L_1 is

$$P(L_1 \leq x) = 1 - \left(1 - \frac{x}{2\pi}\right)^3$$

for $x \in (0, 2\pi)$ (and the CDF is 0 for $x \leq 0$ and 1 for $x \geq 2\pi$). The PDF of L_1 is

$$f_1(x) = \frac{3}{2\pi} \left(1 - \frac{x}{2\pi} \right)^2$$

for $x \in (0, 2\pi)$ (and 0 otherwise). Then

$$E(L_1) = \frac{3}{2\pi} \int_0^{2\pi} x \left(1 - \frac{x}{2\pi} \right)^2 dx = \frac{3}{2\pi} \int_0^{2\pi} \left(\frac{x^3}{4\pi^2} - \frac{x^2}{\pi} + x \right) dx = \frac{3}{2\pi} \left(\frac{x^4}{16\pi^2} - \frac{x^3}{3\pi} + \frac{x^2}{2} \right) \bigg|_0^{2\pi},$$

which reduces to the nice answer

$$E(L_1) = \frac{\pi}{2}.$$

(c) By symmetry, $E(L_2) = E(L_1) = \pi/2$. So by linearity,

$$E(L) = E(L_1) + E(L_2) = \pi.$$

60. Ⓢ As in Example 5.7.4, athletes compete one at a time at the high jump. Let X_j be how high the jth jumper jumped, with $X_1, X_2, \ldots$ i.i.d. with a continuous distribution. We say that the jth jumper is "best in recent memory" if they jump higher than the previous 2 jumpers (for $j \geq 3$; the first 2 jumpers don't qualify).

(a) Find the expected number of best in recent memory jumpers among the 3rd through nth jumpers.

(b) Let A_j be the event that the jth jumper is the best in recent memory. Find $P(A_3 \cap A_4), P(A_3)$, and $P(A_4)$. Are A_3 and A_4 independent?

Solution:

(a) Let I_j be the indicator of the jth jumper being best in recent memory, for each $j \geq 3$. By symmetry, $E(I_j) = 1/3$ (see Section 5.7). By linearity, the desired expected value is $(n-2)/3$.

(b) The event $A_3 \cap A_4$ occurs if and only if the ranks of the first 4 jumps are $4, 3, 2, 1$ or $3, 4, 2, 1$ (where 1 denotes the best of the first 4 jumps, etc.). Since all orderings are equally likely,

$$P(A_3 \cap A_4) = \frac{2}{4!} = \frac{1}{12}.$$

As in (a), we have $P(A_3) = P(A_4) = 1/3$. So $P(A_3 \cap A_4) \neq P(A_3)P(A_4)$, which shows that A_3 and A_4 are not independent.

61. Tyrion, Cersei, and n other guests arrive at a party at i.i.d. times drawn from a continuous distribution with support $[0, 1]$, and stay until the end (time 0 is the party's start time and time 1 is the end time). The party will be boring at times when neither Tyrion nor Cersei is there, fun when exactly one of them is there, and awkward when both Tyrion and Cersei are there.

(a) On average, how many of the n other guests will arrive at times when the party is fun?

(b) Jaime and Robert are two of the other guests. By computing both sides in the definition of independence, determine whether the event "Jaime arrives at a fun time" is independent of the event "Robert arrives at a fun time".

(c) Give a clear intuitive explanation of whether the two events from (b) are independent, and whether they are conditionally independent given the arrival times of everyone else, i.e., everyone except Jaime and Robert.

Solution:

(a) Label the "other" guests as $1, 2, \ldots, n$, and let I_j be the indicator for Guest j arriving at a fun time. Since all orderings are equally likely for the arrivals of Tyrion, Cersei, and Guest j, the probability that Guest j arrives in the middle is $E(I_j) = 1/3$. So by linearity, the expected value is $n/3$.

(b) Let these events be J and R, respectively. Then $P(J) = P(R) = 1/3$, as shown above. But $P(J \cap R) = 4/4! = 1/6$ since all 4! orderings of Tyrion, Cersei, Jaime, and Robert are equally likely, and 4 of these have Jaime and Robert in the middle. So $P(J \cap R) = 1/6 > 1/9 = P(J)P(R)$, showing that J and R are *not* independent.

(c) The events J and R from (b) are *not* independent since knowing that J occurs gives evidence that the "fun interval" is long, making it more likely that R occurs. But J and R are conditionally independent given the arrival times of everyone else, since as soon as we condition on Tyrion's and Cersei's arrival times, we know exactly when the "fun interval" is, and whether Jaime arrives in any fixed interval is independent of whether Robert arrives in that interval (since their arrival times are independent).

62. Let $X_1, X_2, \ldots$ be the annual rainfalls in Boston (measured in inches) in the years 2101, 2102, $\ldots$, respectively. Assume that annual rainfalls are i.i.d. draws from a continuous distribution. A rainfall value is a *record high* if it is greater than those in all previous years (starting with 2101), and a *record low* if it is lower than those in all previous years.

(a) In the 22nd century (the years 2101 through 2200, inclusive), find the expected number of years that have either a record low or a record high rainfall.

(b) On average, in how many years in the 22nd century is there a record low followed in the next year by a record high? (Only the record low is required to be in the 22nd century, not the record high.)

(c) By definition, the year 2101 is a record high (and record low). Let N be the number of years required to get a new record high. Find $P(N > n)$ for all positive integers n, and use this to find the PMF of N.

Hint: Note that $P(N = n) + P(N > n) = P(N > n - 1)$.

(d) With notation as above, show that $E(N)$ is infinite.

Solution:

(a) The first year is automatically a record high (and record low). For $n \geq 2$, the nth year has probability $2/n$ of being a record high or record low, since by symmetry all permutations of $X_1, \ldots, X_n$ are equally likely (so the maximum of these r.v.s is equally likely to be in any position, and likewise for the minimum). By linearity, the desired expectation is

$$1 + \sum_{n=2}^{100} \frac{2}{n} = 2 + 2 \sum_{n=3}^{100} \frac{1}{n}.$$

(b) Let H_j be the indicator for year j being a record high, and L_j be the indicator for year j being a record low (where we define 2101 as year 1, 2102 as year 2, etc.). Then $L_j H_{j+1}$ is the indicator for year j being a record low and year $j+1$ being a record high. Note that L_j is independent of H_{j+1}, since knowing whether year $j + 1$ has a rainfall greater than all of $X_1, X_2, \ldots, X_j$ gives no information about the *internal* rankings among $X_1, \ldots, X_j$ ($H_{j+1} = 1$ is the same event as $X_{j+1} > \max(X_1, \ldots, X_j)$, which says nothing about how $X_1, \ldots, X_j$ compare amongst themselves). So

$$E(L_j H_{j+1}) = P(L_j = 1, H_{j+1} = 1) = P(L_j = 1)P(H_{j+1} = 1) = \frac{1}{j(j+1)}.$$

Then by linearity, the desired expectation is

$$\sum_{j=1}^{100} E(L_j H_{j+1}) = \sum_{j=1}^{100} \frac{1}{j(j+1)}.$$

This can be further simplified by writing $\frac{1}{j(j+1)} = \frac{1}{j} - \frac{1}{j+1}$, so that we then have

$$\sum_{j=1}^{100} \frac{1}{j(j+1)} = \sum_{j=1}^{100} \left(\frac{1}{j} - \frac{1}{j+1} \right) = 1 - \frac{1}{101}.$$

(c) The event $N > n$ is equivalent to saying that the largest among $X_1, X_2, \ldots, X_{n+1}$ is X_1 (i.e., after n additional years, we still haven't gotten a new record high). By symmetry,

$$P(N > n) = \frac{1}{n+1},$$

for all nonnegative integers n. The PMF can then be obtained by writing

$$P(N = n) + P(N > n) = P(N > n - 1),$$

which gives

$$P(N = n) = \frac{1}{n} - \frac{1}{n+1} = \frac{1}{n(n+1)}, \text{ for } n = 1, 2, 3, \ldots.$$

(d) Using the PMF derived in the previous part,

$$E(N) = \sum_{n=1}^{\infty} n P(N = n) = \sum_{n=1}^{\infty} \frac{1}{n+1} = \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \ldots.$$

This diverges to infinity since the harmonic series $\sum_{n=1}^{\infty} \frac{1}{n}$ diverges to infinity. Alternatively, we can apply Theorem 4.4.8:

$$E(N) = \sum_{n=0}^{\infty} P(N > n) = \sum_{n=0}^{\infty} \frac{1}{n+1},$$

which again is the harmonic series.

6

Moments

Means, medians, modes, and moments

1. Let $U \sim \text{Unif}(a, b)$. Find the median and mode of U.

   Solution: The median is the midpoint of (a, b), which is $(a + b)/2$, since U has a 50% chance of being to the left and a 50% chance of being to the right of this point. Every point in (a, b) is a mode of U, since the PDF is constant on (a, b).

2. Let $X \sim \text{Expo}(\lambda)$. Find the median and mode of X.

   Solution: Putting $P(X \leq x) = 1 - e^{-\lambda x} = 1/2$ and solving for x, the median is $(\log 2)/\lambda$. The PDF is $\lambda e^{-\lambda x}$ for $x > 0$. This function is strictly decreasing in x, so there is no mode if the PDF is defined to be 0 at $x = 0$. The mode is 0 if the PDF is defined to be $\lambda e^{-\lambda x}$ for $x \geq 0$, not just for $x > 0$ (which of these conventions is used has no effect on probability calculations involving the Exponential, since changing a function at finitely many points does not affect the integral of that function).

3. Let X have the *Pareto distribution* with parameter $a > 0$; this means that X has PDF $f(x) = a/x^{a+1}$ for $x \geq 1$ (and 0 otherwise). Find the median and mode of X.

   Solution: The CDF is $F(x) = 1 - \frac{1}{x^a}$ for $x \geq 1$ (as shown in Exercise 18 of Chapter 5). Setting this equal to $1/2$ and solving for x, the median is $2^{1/a}$. The PDF is strictly decreasing over $x \geq 1$, so the mode is 1.

4. Let $X \sim \text{Bin}(n, p)$.

   (a) For $n = 5$, $p = 1/3$, find all medians and all modes of X. How do they compare to the mean?

   (b) For $n = 6$, $p = 1/3$, find all medians and all modes of X. How do they compare to the mean?

   Solution:

   (a) Inspection of the PMF shows that there are 2 modes: at 1 and 2. We have

   $$P(X \leq 2) \approx 0.790, P(X \geq 2) = 1 - P(X \leq 1) \approx 0.539,$$

   so 2 is a median. It is the unique median since for $x < 2$,

   $$P(X \leq x) \leq P(X \leq 1) \approx 0.461$$

   and for $x > 2$,

   $$P(X \geq x) = 1 - P(X < x) \leq 1 - P(X \leq 2) \approx 0.210.$$

   The mean is 5/3, which is less than the median and in between the two modes.

   (b) Inspection of the PMF shows that there is a unique mode at 2. We have

   $$P(X \leq 2) \approx 0.680$$

and

$$P(X \geq 2) = 1 - P(X \leq 1) \approx 0.649,$$

so 2 is a median. It is the unique median since for $x < 2$,

$$P(X \leq x) \leq P(X \leq 1) \approx 0.351$$

and for $x > 2$,

$$P(X \geq x) = 1 - P(X < x) \leq 1 - P(X \leq 2) \approx 0.320.$$

The mean is also $6/3 = 2$. So the mean, median, and mode are all equal to 2 here.

5. Let X be Discrete Uniform on $1, 2, \ldots, n$. Find all medians and all modes of X (your answer can depend on whether n is even or odd).

    Solution: All of $1, 2, \ldots, n$ are modes since the PMF is constant on $1, 2, \ldots, n$.
    For n odd, the unique median is $(n+1)/2$, since

    $$P(X \leq (n+1)/2) = \frac{(n+1)/2}{n} = \frac{n+1}{2n} > 1/2$$

    and

    $$P(X \geq (n+1)/2) = 1 - P(X \leq (n-1)/2) = 1 - \frac{n-1}{2n} = \frac{n+1}{2n} > 1/2,$$

    whereas any $x < (n+1)/2$ will have $P(X \leq x) < 1/2$, and any $x > (n+1)/2$ will have $P(X \geq x) < 1/2$.
    For n even, any x between $n/2$ and $(n+2)/2$ (inclusive) is a median. To see that $n/2$ is a median, note that

    $$P(X \leq n/2) = \frac{n/2}{n} = \frac{1}{2}$$

    and

    $$P(X \geq n/2) = 1 - P(X \leq (n-2)/2) = 1 - \frac{(n-2)/2}{n} = \frac{n+2}{2n} > \frac{1}{2}.$$

    A completely analogous calculation shows that $(n+2)/2$ is also a median. It follows that any x between $n/2$ and $(n+2)/2$ (inclusive) is also a median, since any such x satisfies $P(X \leq x) \geq P(X \leq n/2)$ and $P(X \geq x) \geq P(X \geq (n+2)/2)$. There are no other medians, since any $x < n/2$ has

    $$P(X \leq x) \leq P(X \leq (n-2)/2) = \frac{n-2}{2n} < \frac{1}{2}$$

    and, similarly, any $x > (n+2)/2$ has $P(X \geq x) < 1/2$.

6. Suppose that we have data giving the amount of rainfall in a city each day in a certain year. We want useful, informative summaries of how rainy the city was that year. On the majority of days in that year, it did not rain at all in the city. Discuss and compare the following six summaries: the mean, median, and mode of the rainfall on a randomly chosen day from that year, and the mean, median, and mode of the rainfall on a randomly chosen rainy day from that year (where by "rainy day" we mean that it did rain that day in the city).

    Solution: Let R be the rainfall on a random day. The mode is 0 since the most common outcome was no rain. The median is also 0, since $P(R \leq 0) = P(R = 0) > 1/2$ and $P(R \geq 0) = 1 > 1/2$. The mean is positive (unless it never rained that year).

    The fact that the median is 0 just restates the observation that on most days it did not rain. That is useful information but crude since it doesn't give a sense of how much rainfall there was when it did rain. The mode is much cruder still: finding out that the mode is 0 just says that any other measurement, e.g., 0.1357 inches, appears less often

in the dataset than 0's appear. It is hardly surprising for that to happen, since 0 simply means no rain, whereas 0.1357 inches is a very specific measurement. Knowing the mean is equivalent to knowing the total rainfall over the year, which is very useful information for some purposes, though quite limited since it doesn't say anything about how that total rainfall was distributed over the days.

Let Y be the rainfall on a random rainy day. The mode is probably not a very useful summary; if the measurements are reasonably precise, then there probably are very few (if any) repeated values in the dataset (restricted to rainy days). If the measurement 0.12 inches appears twice in the dataset and no other value appears more than once, then 0.12 inches is the unique mode, but that doesn't say much about what an average rainy day feels like, nor is there reason to think the mode would be reasonably stable from year to year.

The median and mean of Y are both likely to be more informative than the mode. If we also know the number of days on which it rained, the mean of Y would again let us compute the total rainfall for the year. The mean, unlike the median, could be very affected by a few days with very heavy rainfall. Whether this is good or bad depends on the purpose for which the summary is being used.

Summaries such as the mean of R and the median of Y are complementary, giving very different information. In this context, it is probably best to have several summary statistics rather than trying to boil the entire dataset down to a single number.

7. Let a and b be positive constants. The *Beta distribution* with parameters a and b, which we introduce in detail in Chapter 8, has PDF proportional to $x^{a-1}(1-x)^{b-1}$ for $0 < x < 1$ (and the PDF is 0 outside of this range). Show that for $a > 1, b > 1$, the mode of the distribution is $(a-1)/(a+b-2)$.

    Hint: Take the log of the PDF first (note that this does not affect where the maximum is achieved).

    Solution: Let $a > 1, b > 1$. The PDF goes to 0 as $x \to 0$ or $x \to 1$ so the mode will not be on or near the boundary of the support. It suffices to maximize
    $$\log\left(x^{a-1}(1-x)^{b-1}\right) = (a-1)\log x + (b-1)\log(1-x),$$
    since omitting the normalizing constant and taking the log do not affect where the maximum is achieved. Setting the derivative equal to 0, we have
    $$\frac{a-1}{x} - \frac{b-1}{1-x} = 0,$$
    which rearranges to
    $$x = \frac{a-1}{a+b-2}.$$
    The second derivative is $-(a-1)/x^2 - (b-1)/(1-x)^2 < 0$, so we have found a maximum.

8. Find the median of the Beta distribution with parameters $a = 3$ and $b = 1$ (see the previous problem for information about the Beta distribution).

    Solution: For $0 < x < 1$, the PDF of the Beta$(3, 1)$ distribution is
    $$f(x) = cx^2(1-x)^0 = cx^2,$$
    where $c = \left(\int_0^1 x^2 dx\right)^{-1} = 3$, and the CDF is
    $$F(a) = 3\int_0^a x^2 dx = a^3.$$
    Setting $F(a) = 1/2$, the median is $2^{-1/3} \approx 0.794$.

9. Let Y be Log-Normal with parameters μ and σ^2. So $Y = e^X$ with $X \sim \mathcal{N}(\mu, \sigma^2)$. Three students are discussing the median and the mode of Y. Evaluate and explain whether or not each of the following arguments is correct.

(a) Student A: "The median of Y is e^μ because the median of X is μ and the exponential function is continuous and strictly increasing, so the event $Y \le e^\mu$ is the same as the event $X \le \mu$."

(b) Student B: "The mode of Y is e^μ because the mode of X is μ, which corresponds to e^μ for Y since $Y = e^X$."

(c) Student C: "The mode of Y is μ because the mode of X is μ and the exponential function is continuous and strictly increasing, so maximizing the PDF of X is equivalent to maximizing the PDF of $Y = e^X$."

Solution:

(a) Student A is right: e^μ is the median of Y, since

$$P(Y \le e^\mu) = P(X \le \mu) = 1/2.$$

(b) Student B is wrong. Figure 6.2 and the discussion of it give an example of a Log-Normal where the mode is clearly less than the median. It turns out that the mode of Y is $e^{\mu - \sigma^2}$, which is less than e^μ for any $\sigma > 0$.

If Z is a *discrete* r.v. and $W = e^Z$, then $P(W = w) = P(Z = z)$, where $z = \log w$, so if z_0 maximizes $P(Z = z)$ then $w_0 = e^{z_0}$ maximizes $P(W = w)$. But X is a *continuous* r.v., and it's *not* true that $f_Y(y) = f_X(x)$, where $x = \log y$; see Chapter 8 for a detailed discussion of how to handle transformations correctly.

(c) Student C is also wrong. In fact, saying that the mode of Y is μ is a *category error*, since μ could be negative, whereas Y is always positive. It's true (and useful) that if a function $f(x)$ is maximized at $x = x_0$, then $g(x) = e^{f(x)}$ is also maximized at $x = x_0$. But the PDF of Y is *not* the exponential of the PDF of X; to think so would be *sympathetic magic*, confusing a random variable with its PDF.

10. A distribution is called *symmetric unimodal* if it is symmetric (about some point) and has a unique mode. For example, any Normal distribution is symmetric unimodal. Let X have a continuous symmetric unimodal distribution for which the mean exists. Show that the mean, median, and mode of X are all equal.

Solution: Let X be symmetric about μ. As shown in the discussion after Definition 6.2.3, μ is the mean of the distribution and μ is also a median. The median is unique since the distribution of X is continuous. Let $\mu + c$ be the unique mode. By symmetry, $\mu - c$ is also a mode. But the mode was assumed to be unique, so $\mu + c = \mu - c$, which shows that $c = 0$. Thus, the mean, median, and mode all equal μ.

11. Let $X_1, \ldots, X_n$ be i.i.d. r.v.s with mean μ, variance σ^2, and skewness γ.

(a) Standardize the X_j by letting

$$Z_j = \frac{X_j - \mu}{\sigma}.$$

Let $\bar{X}_n$ and $\bar{Z}_n$ be the sample means of the X_j and Z_j, respectively. Show that Z_j has the same skewness as X_j, and $\bar{Z}_n$ has the same skewness as $\bar{X}_n$.

(b) Show that the skewness of the sample mean $\bar{X}_n$ is $\gamma/\sqrt{n}$. You can use the fact, shown in Chapter 7, that if X and Y are independent then $E(XY) = E(X)E(Y)$.

Hint: By (a), we can assume $\mu = 0$ and $\sigma^2 = 1$ without loss of generality; if the X_j are

not standardized initially, then we can standardize them. If $(X_1 + X_2 + \cdots + X_n)^3$ is expanded out, there are 3 types of terms: terms such as X_1^3, terms such as $3X_1^2 X_2$, and terms such as $6X_1 X_2 X_3$.

(c) What does the result of (b) say about the distribution of $\bar{X}_n$ when n is large?

Solution:

(a) Since the Z_j's are standardized,

$$\text{Skew}(X_j) = E\left(\frac{X_j - \mu}{\sigma}\right)^3 = E(Z_j^3) = \text{Skew}(Z_j).$$

Note that $\bar{Z}_n$ has mean 0, variance $1/n$, and can be written as

$$\bar{Z}_n = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{X_j - \mu}{\sigma}\right) = \frac{\bar{X}_n - \mu}{\sigma}.$$

Therefore,

$$\text{Skew}(\bar{X}_n) = E\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}\right)^3 = E\left(\frac{\bar{Z}_n}{1/\sqrt{n}}\right)^3 = \text{Skew}(\bar{Z}_n).$$

(b) Assume without loss of generality that $\mu = 0$ and $\sigma^2 = 1$. Then

$$\text{Skew}(\bar{X}_n) = E\left(\frac{\bar{X}_n}{1/\sqrt{n}}\right)^3 = n^{-3/2}E(X_1 + \cdots + X_n)^3 = n^{-3/2}\cdot nE(X_1^3) = \frac{\gamma}{\sqrt{n}},$$

since when expanding $(X_1 + \cdots + X_n)^3$ as in the hint, there are n terms of the form X_j^3, and all other terms have mean 0 (e.g., $E(3X_1^2 X_2) = 3E(X_1^2)E(X_2) = 0$ and $E(6X_1 X_2 X_3) = 6E(X_1)E(X_2)E(X_3) = 0$).

(c) The result of (b) says that the distribution of $\bar{X}_n$ is not very skewed when n is large, even if the distribution of X_j is quite skewed. Of course, the larger γ is, the larger n will have to be to make the skewness of the distribution of $\bar{X}_n$ small. In Chapter 10, we will show that $\bar{X}_n$ is approximately Normal—a symmetric distribution—when n is large. The fact that the skewness washes away as $n \to \infty$ makes this result more plausible.

12. Let c be the speed of light in a vacuum. Suppose that c is unknown, and scientists wish to estimate it. But even more so than that, they wish to estimate c^2, for use in the famous equation $E = mc^2$.

Through careful experiments, the scientists obtain n i.i.d. measurements $X_1, X_2, \ldots, X_n \sim \mathcal{N}(c, \sigma^2)$. Using these data, there are various possible ways to estimate c^2. Two natural ways are: (1) estimate c using the average of the X_j's and then square the estimated c, and (2) average the X_j^2's. So let

$$\bar{X}_n = \frac{1}{n}\sum_{j=1}^{n}X_j,$$

and consider the two estimators

$$T_1 = \bar{X}_n^2 \text{ and } T_2 = \frac{1}{n}\sum_{j=1}^{n}X_j^2.$$

Note that T_1 is the square of the first sample moment and T_2 is the second sample moment.

(a) Find $P(T_1 < T_2)$.

Hint: Start by comparing $(\frac{1}{n}\sum_{j=1}^n x_j)^2$ and $\frac{1}{n}\sum_{j=1}^n x_j^2$ when $x_1,\ldots,x_n$ are *numbers*, by considering a discrete r.v. whose possible values are $x_1,\ldots,x_n$.

(b) When an r.v. T is used to estimate an unknown parameter θ, the *bias* of the estimator T is defined to be $E(T) - \theta$. Find the bias of T_1 and the bias of T_2.

Hint: First find the distribution of $\bar{X}_n$. In general, for finding $E(Y^2)$ for an r.v. Y, it is often useful to write it as $E(Y^2) = \text{Var}(Y) + (EY)^2$.

Solution:

(a) For any real numbers $x_1,\ldots,x_n$,

$$\left(\frac{1}{n}\sum_{j=1}^n x_j\right)^2 \le \frac{1}{n}\sum_{j=1}^n x_j^2.$$

To see this, consider an r.v. X with $P(X = x_j) = 1/n$ for $j = 1,\ldots,n$ (if the x_j's are distinct; if there are repeated values, each has probability proportional to the number of times it appears in the list). Then the lefthand side of the above is $(EX)^2$ and the righthand side is $E(X^2)$, so the inequality holds since variance is nonnegative. (An alternative is to use the Cauchy-Schwarz inequality, a famous mathematical inequality whose statistical interpretation we will see in Chapter 10.) The inequality is strict unless the x_j's are all equal. Thus, $P(T_1 < T_2) = 1$.

(b) Using the fact that $\bar{X}_n \sim \mathcal{N}(c, \sigma^2/n)$,

$$E(T_1) = E(\bar{X}_n^2) = \text{Var}(\bar{X}_n) + (E\bar{X}_n)^2 = \frac{\sigma^2}{n} + c^2.$$

So T_1 has bias σ^2/n. For T_2, linearity and symmetry give

$$E(T_2) = \frac{1}{n}\sum_{j=1}^n E(X_j^2) = E(X_1^2) = \text{Var}(X_1) + (EX_1)^2 = \sigma^2 + c^2.$$

So T_2 has bias σ^2 (which is larger than that of T_1 unless $n = 1$; furthermore, the bias of T_2 does not disappear even as $n \to \infty$).

Moment generating functions

13. ⑤ A fair die is rolled twice, with outcomes X for the first roll and Y for the second roll. Find the moment generating function $M_{X+Y}(t)$ of $X + Y$ (your answer should be a function of t and can contain unsimplified finite sums).

Solution: Since X and Y are i.i.d., LOTUS gives

$$M_{X+Y}(t) = E(e^{t(X+Y)}) = E(e^{tX})E(e^{tY}) = \left(\frac{1}{6}\sum_{k=1}^6 e^{kt}\right)^2.$$

14. ⑤ Let $U_1, U_2, \ldots, U_{60}$ be i.i.d. $\text{Unif}(0,1)$ and $X = U_1 + U_2 + \cdots + U_{60}$. Find the MGF of X.

Solution: The MGF of U_1 is $E(e^{tU_1}) = \int_0^1 e^{tu}du = \frac{1}{t}(e^t - 1)$ for $t \ne 0$ (and the value is 1 for $t = 0$). Thus, the MGF of X is

$$E(e^{tX}) = E(e^{t(U_1+\cdots+U_{60})}) = \left(E(e^{tU_1})\right)^{60} = \frac{(e^t - 1)^{60}}{t^{60}},$$

for $t \ne 0$ (and the value is 1 for $t = 0$).

15. Let $W = X^2 + Y^2$, with X, Y i.i.d. $\mathcal{N}(0,1)$. The MGF of X^2 turns out to be $(1-2t)^{-1/2}$ for $t < 1/2$ (you can assume this).

    (a) Find the MGF of W.

    (b) What famous distribution that we have studied so far does W follow (be sure to state the parameters in addition to the name)? In fact, the distribution of W is also a special case of two more famous distributions that we will study in later chapters!

    Solution:

    (a) The MGF of W is $(1-2t)^{-1}$ for $t < 1/2$, since X^2 and Y^2 are i.i.d.

    (b) The MGF of W is the Expo(1/2) MGF, so $W \sim$ Expo(1/2). (It is also Gamma$(1, 1/2)$, and Chi-Square with 2 degrees of freedom; the Gamma and Chi-Square distributions are introduced in later chapters.)

16. Let $X \sim$ Expo(λ). Find the skewness of X, and explain why it is positive and why it does not depend on λ.

    Hint: Recall that $\lambda X \sim$ Expo(1), and that the nth moment of an Expo(1) r.v. is $n!$ for all n.

    Solution: The standardized version of X is

    $$\frac{X - 1/\lambda}{1/\lambda} = \lambda X - 1,$$

    whose distribution does not depend on λ, since $\lambda X \sim$ Expo(1). So the skewness of X does not depend on λ. We can then assume without loss of generality that $\lambda = 1$. Then, using the fact that $E(X^n) = n!$,

    $$\text{Skew}(X) = E(X-1)^3 = E(X^3 - 3X^2 + 3X - 1) = 3! - 3 \cdot 2 + 3 - 1 = 2.$$

    So Skew$(X) = 2 > 0$, which agrees with the observation from Figure 6.6 that the distribution is right-skewed.

17. Let $X_1, \ldots, X_n$ be i.i.d. with mean μ, variance σ^2, and MGF M. Let

    $$Z_n = \sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right).$$

    (a) Show that Z_n is a standardized quantity, i.e., it has mean 0 and variance 1.

    (b) Find the MGF of Z_n in terms of M, the MGF of each X_j.

    Solution:

    (a) Using the fact that $\bar{X}_n$ has mean μ and variance σ^2/n,

    $$E(Z_n) = \sqrt{n} \cdot \left(\frac{E(\bar{X}_n) - \mu}{\sigma}\right) = 0,$$

    $$\text{Var}(Z_n) = \frac{n}{\sigma^2} \cdot \text{Var}(\bar{X}_n) = 1.$$

    (b) The MGF of Z_n is

    $$M_n(t) = E(e^{tZ_n}) = e^{-\sqrt{n}\mu t/\sigma} E(e^{t\sqrt{n}\bar{X}_n/\sigma}) = e^{-\sqrt{n}t\mu/\sigma}\left(M\left(\frac{t}{\sigma\sqrt{n}}\right)\right)^n.$$

18. Use the MGF of the Geom(p) distribution to give another proof that the mean of this distribution is q/p and the variance is q/p^2, with $q = 1 - p$.

   Solution: As shown in Example 6.4.3, $X \sim \text{Geom}(p)$ has MGF

   $$M(t) = \frac{p}{1 - qe^t},$$

   for $qe^t < 1$. So

   $$E(X) = M'(0) = \frac{pqe^t}{(1 - qe^t)^2}\bigg|_{t=0} = \frac{pq}{(1 - q)^2} = \frac{q}{p},$$

   $$E(X^2) = M''(0) = pq \cdot \frac{(1 - qe^t)^2 e^t + 2qe^{2t}(1 - qe^t)}{(1 - qe^t)^4}\bigg|_{t=0} = pq \cdot \frac{e^t + qe^{2t}}{(1 - qe^t)^3}\bigg|_{t=0} = \frac{q(1 + q)}{p^2},$$

   $$\text{Var}(X) = \frac{q(1 + q)}{p^2} - \frac{q^2}{p^2} = \frac{q}{p^2}.$$

19. Use MGFs to determine whether $X + 2Y$ is Poisson if X and Y are i.i.d. Pois(λ).

   Solution: If $X + 2Y$ is Poisson, it can only be Pois(3λ) since $E(X + 2Y) = 3\lambda$. The MGF of $X + 2Y$ is

   $$M(t) = E(e^{tX})E(e^{2tY}) = e^{\lambda(e^t - 1)}e^{\lambda(e^{2t} - 1)} = e^{\lambda(e^t + e^{2t} - 2)}.$$

   This function is is not the same as the Pois(3λ) MGF, which is

   $$M_{\text{Pois}(3\lambda)}(t) = e^{3\lambda(e^t - 1)}.$$

   (It is also easy to show that $X + 2Y$ is not Poisson without using MGFs: note that $X + 2Y$ has mean 3λ and variance 5λ, whereas any Poisson has mean equal to variance.)

20. Ⓢ Let $X \sim \text{Pois}(\lambda)$, and let $M(t)$ be the MGF of X. The *cumulant generating function* is defined to be $g(t) = \log M(t)$. Expanding $g(t)$ as a Taylor series

   $$g(t) = \sum_{j=1}^{\infty} \frac{c_j}{j!} t^j$$

   (the sum starts at $j = 1$ because $g(0) = 0$), the coefficient c_j is called the jth *cumulant* of X. Find the jth cumulant of X, for all $j \geq 1$.

   Solution: Using the Taylor series for e^t,

   $$g(t) = \lambda(e^t - 1) = \sum_{j=1}^{\infty} \lambda \frac{t^j}{j!},$$

   so $c_j = \lambda$ for all $j \geq 1$.

21. Ⓢ Let $X_n \sim \text{Bin}(n, p_n)$ for all $n \geq 1$, where np_n is a constant $\lambda > 0$ for all n (so $p_n = \lambda/n$). Let $X \sim \text{Pois}(\lambda)$. Show that the MGF of X_n converges to the MGF of X (this gives another way to see that the $\text{Bin}(n, p)$ distribution can be well-approximated by the Pois(λ) when n is large, p is small, and $\lambda = np$ is moderate).

   Solution: Using the fact that $(1 + x/n)^n \to e^x$ as $n \to \infty$ (the compound interest limit, which is reviewed in the math appendix), we have

   $$E(e^{tX_n}) = (1 - p_n + p_n e^t)^n = (1 + \lambda(e^t - 1)/n)^n \quad \to \quad e^{\lambda(e^t - 1)} = E(e^{tX}).$$

22. Consider a setting where a Poisson approximation should work well: let $A_1, \ldots, A_n$ be independent, rare events, with n large and $p_j = P(A_j)$ small for all j. Let $X = I(A_1) + \cdots + I(A_n)$ count how many of the rare events occur, and let $\lambda = E(X)$.

   (a) Find the MGF of X.

   (b) If the approximation $1 + x \approx e^x$ (this is a good approximation when x is very close to 0 but terrible when x is not close to 0) is used to write each factor in the MGF of X as e to a power, what happens to the MGF? Explain why the result makes sense intuitively.

   Solution:

   (a) The indicator $I(A_j)$ is $\text{Bern}(p_j)$, so its MGF is $p_j e^t + 1 - p_j$. Thus, the MGF of X is

   $$M_X(t) = \prod_{j=1}^n (p_j e^t + 1 - p_j).$$

   (b) The requested approximation is

   $$M_X(t) = \prod_{j=1}^n (1 + p_j(e^t - 1)) \approx \prod_{j=1}^n e^{p_j(e^t - 1)} = e^{\lambda(e^t - 1)}.$$

   This makes sense intuitively, in view of the Poisson paradigm, since $M_{\text{Pois}}(t) = e^{\lambda(e^t - 1)}$ is the $\text{Pois}(\lambda)$ MGF.

23. Let U_1, U_2 be i.i.d. $\text{Unif}(0, 1)$. Example 8.2.5 in Chapter 8 shows that $U_1 + U_2$ has a *Triangle distribution*, with PDF given by

   $$f(t) = \begin{cases} t & \text{for } 0 < t \leq 1, \\ 2 - t & \text{for } 1 < t < 2. \end{cases}$$

   The method in Example 8.2.5 is useful but it often leads to difficult integrals, so having alternative methods is important. Show that $U_1 + U_2$ has a Triangle distribution by showing that they have the same MGF.

   Solution: Let X have a Triangle distribution. The MGF of X is

   $$M_X(t) = E(e^{tX}) = \int_0^1 xe^{tx}dx + \int_1^2 (2-x)e^{tx}dx = \int_0^1 xe^{tx}dx + 2\int_1^2 e^{tx}dx - \int_1^2 xe^{tx}dx.$$

   Integrating by parts (letting $u = x, dv = e^{tx}dx$), we have

   $$\int xe^{tx}dx = \frac{e^{tx}(tx - 1)}{t^2} + C.$$

   So

   $$M_X(t) = \frac{e^t(t-1)}{t^2} + \frac{1}{t^2} + \frac{2e^{2t}}{t} - \frac{2e^t}{t} - \frac{e^{2t}(2t-1)}{t^2} + \frac{e^t(t-1)}{t^2} = \frac{(e^t - 1)^2}{t^2}$$

   for all $t \neq 0$ (and $M_X(0) = 1$). Therefore, X and $U_1 + U_2$ have the same MGF, which implies that they have the same distribution.

24. Let X and Y be i.i.d. $\text{Expo}(1)$, and $L = X - Y$. The *Laplace distribution* has PDF

   $$f(x) = \frac{1}{2}e^{-|x|}$$

   for all real x. Use MGFs to show that the distribution of L is Laplace.

Solution: The MGF of L is

$$M_L(t) = E(e^{t(X-Y)}) = E(e^{tX})E(e^{-tY}) = \left(\frac{1}{1-t}\right)\left(\frac{1}{1+t}\right) = \frac{1}{1-t^2}$$

for $-1 < t < 1$, using the result for the MGF of an Exponential. By LOTUS, the MGF of a Laplace is

$$M_{\text{Laplace}}(t) = \frac{1}{2}\int_{-\infty}^{\infty} e^{-|w|}e^{tw}dw = \frac{1}{2}\int_{-\infty}^{0} e^{tw+w}dw + \frac{1}{2}\int_{0}^{\infty} e^{tw-w}dw,$$

which simplifies to

$$\frac{1}{2(1+t)} + \frac{1}{2(1-t)} = \frac{1}{1-t^2},$$

for $-1 < t < 1$. Thus, the MGFs are the same, which implies that L is Laplace.

25. Let $Z \sim \mathcal{N}(0,1)$, and $Y = |Z|$. So Y has the *Folded Normal* distribution, discussed in Example 5.4.7. Find *two* expressions for the MGF of Y as unsimplified integrals: one integral based on the PDF of Y, and one based on the PDF of Z.

Solution: Let f_Y be the PDF of Y and φ be the PDF of Z. As shown in Example 5.4.7, $f_Y(y) = 2\varphi(y)$ for $y \geq 0$ (and $f_Y(y) = 0$ for $y < 0$). By LOTUS, the MGF of Y is

$$E(e^{tY}) = \int_0^{\infty} e^{ty}f_Y(y)dy = \int_{-\infty}^{\infty} e^{t|z|}\varphi(z)dz.$$

26. Let $X, Y, Z, W \sim \mathcal{N}(0,1)$ be i.i.d.

(a) Find an expression for $E\left(\Phi(Z)e^{Z}\right)$ as an unsimplified integral.

(b) Find $E(\Phi(Z))$ and $E(e^{Z})$ as fully simplified numbers.

Solution:

(a) By LOTUS,

$$E\left(\Phi(Z)e^{Z}\right) = \int_{-\infty}^{\infty} \Phi(z)e^{z}\varphi(z)dz,$$

where $\varphi(z) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}$ is the $\mathcal{N}(0,1)$ PDF.

(b) By universality of the Uniform, $\Phi(Z) \sim \text{Unif}(0,1)$, so

$$E(\Phi(Z)) = \frac{1}{2}.$$

Since e^{Z} is Log-Normal,

$$E(e^{Z}) = e^{1/2}.$$

Alternatively, we can obtain $E(e^{Z})$ by evaluating the $\mathcal{N}(0,1)$ MGF at 1.
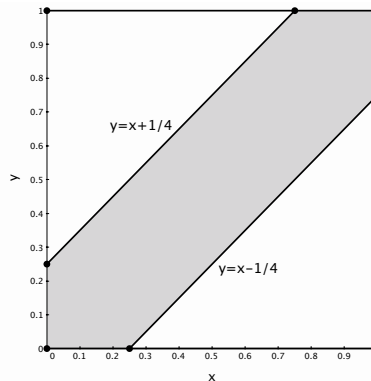
7

Joint distributions

Joint, marginal, and conditional distributions

1. Alice and Bob arrange to meet for lunch on a certain day at noon. However, neither is known for punctuality. They both arrive independently at uniformly distributed times between noon and 1 pm on that day. Each is willing to wait up to 15 minutes for the other to show up. What is the probability they will meet for lunch that day?

   Solution: Let X and Y be the arrival times of Alice and Bob respectively, measured in hours past noon. We want to find $P(|X - Y| \leq 0.25)$. This can be done by integrating the joint PDF of (X, Y) over the region $|x - y| \leq 0.25$ in the (x, y)-plane. But since (X, Y) is Uniform over the square $0 \leq x \leq 1, 0 \leq y \leq 1$, probability is proportional to area. In fact, it equals area here since the area of the square is 1. Since $|x - y| \leq 0.25$ is equivalent to $y - x \leq 0.25, y - x \geq -0.25$, we just need to find the area in the square between the lines $y = x + 0.25$ and $y = x - 0.25$, which is the shaded area below.

   The complement of the region of interest consists of two disjoint triangles, each with area $\frac{1}{2}(\frac{3}{4})^2$. Thus,

   $$P(|X - Y| \leq 0.25) = 1 - \left(\frac{3}{4}\right)^2 = \frac{7}{16}.$$

2. Alice, Bob, and Carl arrange to meet for lunch on a certain day. They arrive independently at uniformly distributed times between 1 pm and 1:30 pm on that day.

   (a) What is the probability that Carl arrives first?

   For the rest of this problem, assume that Carl arrives first at 1:10 pm, and condition on this fact.

   (b) What is the probability that Carl will be waiting alone for more than 10 minutes?

   (c) What is the probability that Carl will have to wait more than 10 minutes until his party is complete?

(d) What is the probability that the person who arrives second will have to wait more than 5 minutes for the third person to show up?

Solution:

(a) By symmetry, the probability that Carl arrives first is 1/3.

(b) There is a 50% chance that Alice will arrive within the next 10 minutes and a 50% chance that Bob will arrive within the next 10 minutes. So by independence, the probability is 1/4 that neither Alice nor Bob will arrive within the next 10 minutes.

(c) The probability is 1/4 that both Alice and Bob will arrive within the next 10 minutes, so the probability is 3/4 that Carl will have to wait more than 10 minutes in order for both Alice and Bob to have arrived.

(d) We need to find $P(|A - B| > 5)$, where A and B are i.id. Unif$(0, 20)$ r.v.s. Letting $X = A/20$ and $Y = B/20$, we need to find $P(|X - Y| > 0.25)$, where X and Y are i.i.d. Unif$(0, 1)$. This can be done geometrically, as in the solution to the previous exercise. In fact, it is the *same* calculation. By the solution to the previous exercise,

$$P(|X - Y| > 0.25) = \frac{9}{16}.$$

3. One of two doctors, Dr. Hibbert and Dr. Nick, is called upon to perform a series of n surgeries. Let H be the indicator r.v. for Dr. Hibbert performing the surgeries, and suppose that $E(H) = p$. Given that Dr. Hibbert is performing the surgeries, each surgery is successful with probability a, independently. Given that Dr. Nick is performing the surgeries, each surgery is successful with probability b, independently. Let X be the number of successful surgeries.

   (a) Find the joint PMF of H and X.

   (b) Find the marginal PMF of X.

   (c) Find the conditional PMF of H given $X = k$.

   Solution:

   (a) The joint PMF of H and X is

   $$P(H = h, X = k) = P(X = k|H = h)P(H = h) = \begin{cases} \binom{n}{k}a^k(1-a)^{n-k}p, & \text{if } h = 1; \\ \binom{n}{k}b^k(1-b)^{n-k}(1-p), & \text{if } h = 0, \end{cases}$$

   for $k \in \{0, 1, \ldots, n\}$.

   (b) The marginal PMF of X is

   $$P(X = k) = P(H = 1, X = k) + P(H = 0, X = k) = \binom{n}{k}a^k(1-a)^{n-k}p + \binom{n}{k}b^k(1-b)^{n-k}(1-p),$$

   for $k \in \{0, 1, \ldots, n\}$.

   (c) The conditional PMF of H given $X = k$ is

   $$P(H = h|X = k) = \frac{P(H = h, X = k)}{P(X = k)},$$

   where $P(H = h, X = k)$ is as in (a) and $P(X = k)$ is as in (b).

4. A fair coin is flipped twice. Let X be the number of Heads in the two tosses, and Y be the indicator r.v for the tosses landing the same way.

(a) Find the joint PMF of X and Y.

(b) Find the marginal PMFs of X and Y.

(c) Are X and Y independent?

(d) Find the conditional PMFs of Y given $X = x$ and of X given $Y = y$.

Solution:

(a) Marginally, we have $X \sim \text{Bin}(2, 1/2)$ by the story of the Binomial. And Y is a function of X: if $X = 0$ or $X = 2$, then $Y = 1$; if $X = 1$, then $Y = 0$. So the joint PMF of X and Y is

$$P(X = x, Y = y) = \begin{cases} 1/4, & \text{if } x = 0, y = 1; \\ 1/2, & \text{if } x = 1, y = 0; \\ 1/4, & \text{if } x = 2, y = 1; \\ 0, & \text{otherwise.} \end{cases}$$

(b) We have $X \sim \text{Bin}(2, 1/2)$ and $Y \sim \text{Bern}(1/2)$, so the PMF of X is

$$P(X = 0) = 1/4, P(X = 1) = 1/2, P(X = 2) = 1/4$$

for $k = 0, 1, 2$, and the PMF of Y is

$$P(Y = 0) = 1/2, P(Y = 1) = 1/2.$$

(c) No, they are extremely dependent: Y is a function of X, as shown in (a).

(d) The conditional PMF of Y given $X = x$ is degenerate:

$$P(Y = 1|X = 0) = 1, P(Y = 0|X = 1) = 1, P(Y = 1|X = 2) = 1.$$

The conditional PMF of X given $Y = 0$ is also degenerate: $P(X = 1|Y = 0) = 1$. The conditional PMF of X given $Y = 1$ is

$$P(X = 0|Y = 1) = \frac{P(X = 0, Y = 1)}{P(Y = 1)} = \frac{1}{2}, P(X = 2|Y = 1) = \frac{P(X = 2, Y = 1)}{P(Y = 1)} = \frac{1}{2}.$$

5. A fair die is rolled, and then a coin with probability p of Heads is flipped as many times as the die roll says, e.g., if the result of the die roll is a 3, then the coin is flipped 3 times. Let X be the result of the die roll and Y be the number of times the coin lands Heads.

(a) Find the joint PMF of X and Y. Are they independent?

(b) Find the marginal PMFs of X and Y.

(c) Find the conditional PMFs of Y given $X = x$ and of Y given $X = x$.

Solution:

(a) The joint PMF of X and Y is

$$P(X = x, Y = y) = P(X = x)P(Y = y|X = x) = \frac{1}{6} \binom{x}{y} p^y (1 - p)^{x-y},$$

for $x = 1, 2, \ldots, 6$ and $y = 0, 1, \ldots, x$ (and the PMF is 0 otherwise). The r.v.s X and

Y are not independent, since the joint PMF does not factor as a function of x times a function of y (we can also see that they are dependent directly, by noting that, given that $X = 1$, the support of Y is $\{0, 1\}$ but, given that $X = 6$, the support of Y is $\{0, 1, 2, \ldots, 6\}$.

(b) The marginal PMF of X is $P(X = x) = 1/6$ for $x = 1, 2, \ldots, 6$. The marginal PMF of Y is

$$P(Y = y) = \sum_{x=1}^{6} P(X = x, Y = y) = \frac{p^y}{6} \sum_{x=y}^{6} \binom{x}{y} (1-p)^{x-y},$$

for $y = 0, 1, \ldots, 6$.

(c) The conditional distribution of Y given $X = x$ is $\text{Bin}(x, p)$, so the PMF of Y given $X = x$ is

$$P(Y = y | X = x) = \binom{x}{y} p^y (1-p)^{x-y},$$

for $y = 0, 1, \ldots, x$. By Bayes' rule,

$$P(X = x | Y = y) = \frac{P(Y = y | X = x) P(X = x)}{P(Y = y)},$$

where $P(Y = y | X = x), P(X = x)$, and $P(Y = y)$ are as above.

6. A committee of size k is chosen from a group of n women and m men. All possible committees of size k are equally likely. Let X and Y be the numbers of women and men on the committee, respectively.

(a) Find the joint PMF of X and Y. Be sure to specify the support.

(b) Find the marginal PMF of X in two different ways: by doing a computation using the joint PMF, and using a story.

(c) Find the conditional PMF of Y given that $X = x$.

Solution:

(a) Since all subsets of size k are equally likely, the joint PMF of X and Y is

$$P(X = i, Y = j) = \frac{\binom{n}{i}\binom{m}{j}}{\binom{n+m}{k}},$$

for all nonnegative integers i and j with $i + j = k$ (and the joint PMF is 0 otherwise).

(b) By the story of the Hypergeometric, $X \sim \text{HGeom}(n, m, k)$. Alternatively, using (a) and the fact that $X = i$ implies $Y = k - i$,

$$P(X = i) = P(X = i, Y = k - i) = \frac{\binom{n}{i}\binom{m}{k-i}}{\binom{n+m}{k}},$$

which is indeed the $\text{HGeom}(n, m, k)$ PMF.

(c) The conditional PMF of Y given that $X = x$ is degenerate: $X = x$ implies $Y = k - x$, so $P(Y = k - x | X = x) = 1$.

7. A stick of length L (a positive constant) is broken at a uniformly random point X. Given that $X = x$, another breakpoint Y is chosen uniformly on the interval $[0, x]$.

(a) Find the joint PDF of X and Y. Be sure to specify the support.

(b) We already know that the marginal distribution of X is $\text{Unif}(0, L)$. Check that

marginalizing out Y from the joint PDF agrees that this is the marginal distribution of X.

(c) We already know that the conditional distribution of Y given $X = x$ is $\text{Unif}(0, x)$. Check that using the definition of conditional PDFs (in terms of joint and marginal PDFs) agrees that this is the conditional distribution of Y given $X = x$.

(d) Find the marginal PDF of Y.

(e) Find the conditional PDF of X given $Y = y$.

Solution:

(a) The joint PDF of X and Y is

$$f(x, y) = f_X(x)f_{Y|X}(y|x) = \frac{1}{Lx},$$

for $0 < x < L$ and $0 < y < x$ (and the joint PDF is 0 otherwise).

(b) Marginalizing out Y, while keeping in mind the constraint $y < x$, we have

$$f_X(x) = \int_0^x \frac{1}{Lx} dy = \frac{1}{L}$$

for $0 < x < L$, which agrees with the fact that $X \sim \text{Unif}(0, L)$.

(c) The conditional PDF of Y given $X = x$ is

$$f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)} = \frac{1/(Lx)}{1/L} = \frac{1}{x}$$

for $0 < y < x$, which agrees with the fact that $Y|X = x \sim \text{Unif}(0, x)$.

(d) The marginal PDF of Y is

$$f_Y(y) = \int_y^L f(x, y) dx = \frac{1}{L} \int_y^L \frac{1}{x} dx = \frac{\log L - \log y}{L},$$

for $0 < y < L$.

(e) The conditional PDF of X given $Y = y$ is

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)} = \frac{1}{x(\log L - \log y)},$$

for $y < x < L$.

8. (a) Five cards are randomly chosen from a standard deck, one at a time *with replacement*. Let X, Y, Z be the numbers of chosen queens, kings, and other cards. Find the joint PMF of X, Y, Z.

(b) Find the joint PMF of X and Y.

Hint: In summing the joint PMF of X, Y, Z over the possible values of Z, note that most terms are 0 because of the constraint that the number of chosen cards is five.

(c) Now assume instead that the sampling is without replacement (all 5-card hands are equally likely). Find the joint PMF of X, Y, Z.

Hint: Use the naive definition of probability.

Solution:

(a) By the story of the Multinomial, $(X, Y, Z) \sim \text{Mult}_3(5, \mathbf{p})$, with $\mathbf{p} = (1/13, 1/13, 11/13)$. So the joint PMF is

$$P(X = x, Y = y, Z = z) = \frac{5!}{x!y!z!} \cdot \left(\frac{1}{13}\right)^x \left(\frac{1}{13}\right)^y \left(\frac{11}{13}\right)^z = \frac{5!}{x!y!z!} \cdot \frac{11^z}{13^5},$$

for x, y, z nonnegative integers with $x + y + z = 5$.

(b) Since $X + Y + Z = 5$, the joint PMF of X and Y is

$$P(X = x, Y = y) = P(X = x, Y = y, Z = 5 - x - y) = \frac{5!}{x!y!(5 - x - y)!} \cdot \frac{11^{5-x-y}}{13^5},$$

for x, y nonnegative integers with $x + y \leq 5$.

(c) By the naive definition of probability and the multiplication rule,

$$P(X = x, Y = y, Z = z) = \frac{\binom{4}{x}\binom{4}{y}\binom{44}{z}}{\binom{52}{5}},$$

for x, y, z nonnegative integers with $x + y + z = 5$.

9. Let X and Y be i.i.d. Geom(p), and $N = X + Y$.

(a) Find the joint PMF of X, Y, N.

(b) Find the joint PMF of X and N.

(c) Find the conditional PMF of X given $N = n$, and give a simple description in words of what the result says.

Solution:

(a) Let $q = 1 - p$. Since $P(N = x + y | X = x, Y = y) = 1$, the joint PMF of X, Y, N is

$$P(X = x, Y = y, N = n) = P(X = x, Y = y) = pq^x pq^y = p^2 q^n,$$

for x, y, n nonnegative integers with $n = x + y$.

(b) If $X = x$ and $N = n$, then $Y = n - x$. So the joint PMF of X and N is

$$P(X = x, N = n) = P(X = x, Y = n - x, N = n) = p^2 q^n,$$

for x, n nonnegative integers with $x \leq n$. As a check, note that this implies

$$P(N = n) = \sum_{x=0}^{n} p^2 q^n = (n + 1)p^2 q^n,$$

which agrees with the fact that $N \sim \text{NBin}(2, p)$.

(c) The conditional PMF of X given $N = n$ is

$$P(X = x | N = n) = \frac{P(X = x, N = n)}{P(N = n)} = \frac{p^2 q^n}{(n + 1)p^2 q^n} = \frac{1}{n + 1}$$

for $x = 0, 1, \ldots, n$ since, as noted in the solution to (b), $N \sim \text{NBin}(2, p)$. This says that, given that $N = n$, X is equally likely to be any integer between 0 and n (inclusive). To describe this in terms of the story of the Negative Binomial, imagine performing independent Bernoulli trials until the second success is obtained. Let N be the number of failures before the second success. Given that $N = n$, the $(n+2)$nd trial is the second success, and the result says that the first success is equally likely to be located anywhere among the first $n + 1$ trials.

10. Let X and Y be i.i.d. Expo(λ), and $T = X + Y$.

    (a) Find the conditional CDF of T given $X = x$. Be sure to specify where it is zero.

    (b) Find the conditional PDF $f_{T|X}(t|x)$, and verify that it is a valid PDF.

    (c) Find the conditional PDF $f_{X|T}(x|t)$, and verify that it is a valid PDF.

    Hint: This can be done using Bayes' rule without having to know the marginal PDF of T, by recognizing what the conditional PDF is up to a normalizing constant—then the normalizing constant must be whatever is needed to make the conditional PDF valid.

    (d) In Example 8.2.4, we will show that the marginal PDF of T is $f_T(t) = \lambda^2 t e^{-\lambda t}$, for $t > 0$. Give a short alternative proof of this fact, based on the previous parts and Bayes' rule.

    Solution:

    (a) Let $x > 0$. The conditional CDF of T given $X = x$ is

    $$F_{T|X}(t|X = x) = P(T \le t|X = x) = P(Y \le t-x|X = x) = P(Y \le t-x) = 1-e^{-\lambda(t-x)},$$

    for $t > x$ (and the conditional CDF is 0 for $t \le x$). Independence of X and Y was used to drop the condition $X = x$ in $P(Y \le t - x|X = x)$.
    In other words, the conditional distribution of T given $X = x$ is a shifted Exponential: it is the distribution of $x + Y$ with $Y \sim$ Expo(λ).

    (b) Differentiating, the conditional PDF of T given $X = x$ is

    $$f_{T|X}(t|x) = \frac{d}{dt}F_{T|X}(t|X = x) = \lambda e^{-\lambda(t-x)},$$

    for $t > x$ (and 0 otherwise). This is a valid PDF since, letting $y = t - x$,

    $$\int_x^\infty \lambda e^{-\lambda(t-x)}dt = \int_0^\infty \lambda e^{-\lambda y}dy = 1.$$

    (c) Let $t > 0$. By Bayes' rule, for $0 < x < t$ we have

    $$f_{X|T}(x|t) \propto f_{T|X}(t|x)f_X(x) = \lambda e^{-\lambda(t-x)}\lambda e^{-\lambda x} = \lambda^2 e^{-\lambda t},$$

    which does not depend on x. Thus, $X|T \sim$ Unif($0, T$). So the conditional PDF of X given $T = t$ is $1/t$ for $0 < x < t$ (and 0 otherwise).

    (d) Rearranging Bayes' rule,

    $$f_T(t) = \frac{f_{T|X}(t|x)f_X(x)}{f_{X|T}(x|t)} = \frac{\lambda^2 e^{-\lambda t}}{1/t} = \lambda^2 t e^{-\lambda t},$$

    as desired.

11. Let X, Y, Z be r.v.s such that $X \sim \mathcal{N}(0, 1)$ and conditional on $X = x$, Y and Z are i.i.d. $\mathcal{N}(x, 1)$.

    (a) Find the joint PDF of X, Y, Z.

    (b) By definition, Y and Z are conditionally independent given X. Discuss intuitively whether or not Y and Z are also unconditionally independent.

    (c) Find the joint PDF of Y and Z. You can leave your answer as an integral, though the integral can be done with some algebra (such as completing the square) and facts about the Normal distribution.

Solution:

(a) The joint PDF of X, Y, Z is

$$f(x, y, z) = f_X(x)f_{Y|X}(y|x)f_{Z|X}(z|x) = \frac{1}{(2\pi)^{3/2}} \exp\left(-\frac{x^2}{2} - \frac{(y-x)^2}{2} - \frac{(z-x)^2}{2}\right).$$

(b) Intuitively, Y and Z are not unconditionally independent. Learning the value of Y gives information about the value of X (since, e.g., there is about a 95% chance that X is at most distance 2 away from Y), which in turn gives information about the value of Z (since, e.g., there is about a 95% chance that Z is at most distance 2 away from X).

(c) Marginalizing out X, the joint PDF of Y and Z is

$$f_{Y,Z}(y, z) = \int_{-\infty}^{\infty} f(x, y, z)dx = \frac{1}{(2\pi)^{3/2}} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2} - \frac{(y-x)^2}{2} - \frac{(z-x)^2}{2}\right) dx.$$

12. Let $X \sim \text{Expo}(\lambda)$, and let c be a positive constant.

(a) If you remember the memoryless property, you already know that the conditional distribution of X given $X > c$ is the same as the distribution of $c + X$ (think of waiting c minutes for a "success" and then having a fresh $\text{Expo}(\lambda)$ additional waiting time). Derive this in another way, by finding the conditional CDF of X given $X > c$ and the conditional PDF of X given $X > c$.

(b) Find the conditional CDF of X given $X < c$ and the conditional PDF of X given $X < c$.

Solution:

(a) Let F be the CDF of X. The conditional CDF of X given $X > c$ is

$$P(X \leq x|X > c) = \frac{P(c < X \leq x)}{P(X > c)} = \frac{F(x) - F(c)}{1 - F(c)} = \frac{e^{-\lambda c} - e^{-\lambda x}}{e^{-\lambda c}} = 1 - e^{-\lambda(x-c)},$$

for $x > c$ (and the conditional CDF is 0 for $x \leq c$). This is the CDF of $c + X$, as desired, since for $x > c$ we have

$$P(c + X \leq x) = P(X \leq x - c) = 1 - e^{-\lambda(x-c)}.$$

(b) For $x \geq c$, $P(X \leq x|X < c) = 1$. For $x < c$,

$$P(X \leq x|X < c) = \frac{P(X \leq x \text{ and } X < c)}{P(X < c)} = \frac{P(X \leq x)}{P(X < c)} = \frac{1 - e^{-\lambda x}}{1 - e^{-\lambda c}}.$$

The conditional PDF of X given $X < c$ is the derivative of the above expression:

$$f(x|X < c) = \frac{\lambda e^{-\lambda x}}{1 - e^{-\lambda c}},$$

for $x < c$ (and the conditional PDF is 0 for $x \geq c$).

13. Let X and Y be i.i.d. $\text{Expo}(\lambda)$. Find the conditional distribution of X given $X < Y$ in two different ways:

(a) by using calculus to find the conditional PDF.

(b) without using calculus, by arguing that the conditional distribution of X given $X < Y$ is the same distribution as the unconditional distribution of $\min(X, Y)$, and then applying an earlier result about the minimum of independent Exponentials.

Solution:

(a) The conditional CDF is

$$P(X \leq a | X < Y) = 1 - P(X > a | X < Y) = 1 - \frac{P(a < X < Y)}{P(X < Y)}.$$

By symmetry, $P(X < Y) = 1/2$. To find $P(a < X < Y)$, we can integrate the joint PDF of X and Y over all (x, y) with $a < x < y$:

$$\begin{aligned}
P(a < X < Y) &= \lambda^2 \int_a^\infty \int_x^\infty e^{-\lambda x} e^{-\lambda y} \, dy \, dx \\
&= \int_a^\infty \lambda e^{-\lambda x} \int_x^\infty \lambda e^{-\lambda y} \, dy \, dx \\
&= \int_a^\infty \lambda e^{-2\lambda x} \, dx \\
&= \frac{1}{2} \int_a^\infty 2\lambda e^{-2\lambda x} \, dx \\
&= \frac{1}{2} e^{-2\lambda a}.
\end{aligned}$$

(We computed $\int_x^\infty \lambda e^{-\lambda y} dy$ and $\int_a^\infty 2\lambda e^{-2\lambda x} dx$ by pattern-matching to Exponentials, e.g., the latter is $P(W > a)$ for $W \sim \text{Expo}(2\lambda)$, but these integrals can also be done using basic calculus.) Therefore,

$$P(X \leq a | X < Y) = 1 - e^{-2\lambda a}.$$

That is, the conditional distribution of X given $X < Y$ is $\text{Expo}(2\lambda)$. The conditional PDF of X given $X < Y$ is $2\lambda e^{-2\lambda x}$, for $x > 0$.

(b) Learning that $X < Y$ is the same thing as learning that X is the minimum of the two Exponentials X and Y. And as shown in Example 5.6.3, $\min(X, Y) \sim \text{Expo}(2\lambda)$.

14. (S) (a) A stick is broken into three pieces by picking two points independently and uniformly along the stick, and breaking the stick at those two points. What is the probability that the three pieces can be assembled into a triangle?
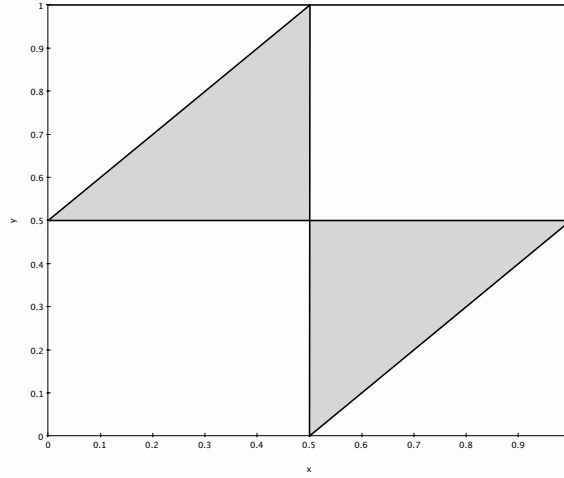
Hint: A triangle can be formed from 3 line segments of lengths a, b, c if and only if $a, b, c \in (0, 1/2)$. The probability can be interpreted geometrically as proportional to an area in the plane, avoiding all calculus, but make sure for that approach that the distribution of the random point in the plane is Uniform over some region.

(b) Three legs are positioned uniformly and independently on the perimeter of a round table. What is the probability that the table will stand?

Solution:

(a) We can assume the length is 1 (in some choice of units, the length will be 1, and the choice of units for length does not affect whether a triangle can be formed). So let X, Y be i.i.d. Unif$(0,1)$ random variables. Let x and y be the observed values of X and Y respectively. If $x < y$, then the side lengths are $x, y - x$, and $1 - y$, and a triangle can be formed if and only if $y > \frac{1}{2}, y < x + \frac{1}{2}, x < \frac{1}{2}$. Similarly, if $x > y$, then a triangle can be formed if and only if $x > \frac{1}{2}, x < y + \frac{1}{2}, y < \frac{1}{2}$.

Since (X, Y) is Uniform over the square $0 \leq x \leq 1, 0 \leq y \leq 1$, the probability of a subregion is proportional to its area. The region given by $y > 1/2, y < x + 1/2, x < 1/2$ is a triangle with area $1/8$, as is the region given by $x > 1/2, x < y + 1/2, y < 1/2$, as illustrated in the picture below. Thus, the probability that a triangle can be formed is $1/8 + 1/8 = 1/4$.

Note that the idea of interpreting probabilities as areas works here because (X, Y) is *Uniform* on the square. For other distributions, in general we would need to find the joint PDF of X, Y and integrate over the appropriate region.

(b) Think of the legs as points on a circle, chosen randomly one at a time, and choose units so that the circumference of the circle is 1. Let A, B, C be the arc lengths from one point to the next (clockwise, starting with the first point chosen). Then

$$P(\text{table falls}) = P(\text{the 3 legs are all contained in some semicircle})$$
$$= P(\text{at least one of } A, B, C \text{ is greater than } 1/2) = 3/4,$$

by Part (a). So the probability that the table will stand is $1/4$.

Alternatively, let C_j be the clockwise semicircle starting from the jth of the 3 points. Let A_j be the event that C_j contains all 3 points. Then $P(A_j) = 1/4$ and with probability 1, at most one A_j occurs. So $P(A_1 \cup A_2 \cup A_3) = 3/4$, which again shows that the probability that the table will stand is $1/4$.

15. Let X and Y be continuous r.v.s., with joint CDF $F(x, y)$. Show that the probability that (X, Y) falls into the rectangle $[a_1, a_2] \times [b_1, b_2]$ is

$$F(a_2, b_2) - F(a_1, b_2) + F(a_1, b_1) - F(a_2, b_1).$$

Solution: For any point (a, b) in the plane, let $A_{(a,b)}$ be the event that (X, Y) lies below and to the left of (a, b) (not necessarily strictly), i.e., $A_{(a,b)}$ is the event $(X, Y) \in \{(x, y) : x \leq a, y \leq b)\}$. Let R be the rectangle $[a_1, a_2] \times [b_1, b_2]$ and $B = \{(x, y) : y < b_1\}$ be the strip below the rectangle. Then

$$F(a_2, b_2) - F(a_1, b_2) = P(A_{a_2, b_2}) - P(A_{a_1, b_2}) = P(R \cup B),$$
$$F(a_2, b_1) - F(a_1, b_1) = P(A_{a_2, b_1}) - P(A_{a_1, b_1}) = P(B),$$

so

$$F(a_2, b_2) - F(a_1, b_2) + F(a_1, b_1) - F(a_2, b_1) = P(R \cup B) - P(B) = P(R).$$

16. Let X and Y have joint PDF

$$f_{X,Y}(x, y) = x + y, \text{ for } 0 < x < 1 \text{ and } 0 < y < 1.$$

(a) Check that this is a valid joint PDF.

(b) Are X and Y independent?

(c) Find the marginal PDFs of X and Y.

(d) Find the conditional PDF of Y given $X = x$.

Solution:

(a) The function $f_{X,Y}$ is nonnegative for $0 < x < 1$ and $0 < y < 1$ (and 0 otherwise), so we just have to check that it integrates to 1:

$$\int_0^1 \int_0^1 (x+y)dxdy = \int_0^1 \left(\frac{1}{2} + y\right) dy = \frac{1}{2} + \frac{1}{2} = 1.$$

(b) They are not independent, since $x + y$ does not factor as a function of x times a function of y.

(c) The marginal PDF of X is

$$f_X(x) = \int_0^1 (x+y)dy = x + \frac{1}{2},$$

for $0 < x < 1$. By an analogous calculation or by symmetry, the marginal PDF of Y is

$$f_Y(y) = y + \frac{1}{2},$$

for $0 < y < 1$.

(d) Let $0 < x < 1$. The conditional PDF of Y given $X = x$ is

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{x+y}{x + \frac{1}{2}},$$

for $0 < y < 1$.

17. Let X and Y have joint PDF

$$f_{X,Y}(x,y) = cxy, \text{ for } 0 < x < y < 1.$$

(a) Find c to make this a valid joint PDF.

(b) Are X and Y independent?

(c) Find the marginal PDFs of X and Y.

(d) Find the conditional PDF of Y given $X = x$.

Solution:

(a) We have

$$\int_0^1 \int_0^y xydxdy = \int_0^1 y \left(\int_0^y xdx\right) dy = \int_0^1 \frac{y^3}{2}dy = \frac{1}{8},$$

so $c = 8$.

(b) No, X and Y are not independent. Superficially the joint PDF might appear to factor as a function of x times a function of y, but it is not true that such a factorization holds for all real x, y. The PDF is 0 for $x \geq y$, creating the constraint that Y must be greater than X. So learning the value of X does provide information about Y, by narrowing the range of possible values of Y.

202

(c) The marginal PDF of X is

$$f_X(x) = \int_x^1 8xy dy = 8x \int_x^1 y dy = 4x - 4x^3,$$

for $0 < x < 1$. The marginal PDF of Y is

$$f_Y(y) = \int_0^y 8xy dx = 8y \int_0^y x dx = 4y^3,$$

for $0 < y < 1$.

(d) The conditional PDF of Y given $X = x$ is

$$f_{Y|X}(y|x) = \frac{8xy}{4x - 4x^3} = \frac{2y}{1 - x^2},$$

for $0 < x < y < 1$.

18. ⓢ Let (X, Y) be a uniformly random point in the triangle in the plane with vertices $(0,0), (0,1), (1,0)$. Find the joint PDF of X and Y, the marginal PDF of X, and the conditional PDF of X given Y.

Solution: The area of the triangle is $\frac{1}{2}$, so the joint PDF of (X, Y) is 2 inside the triangle and 0 outside the triangle. The triangle is given by $x \geq 0, y \geq 0, x + y \leq 1$, so the marginal PDF of X is $\int_0^{1-x} 2 dy = 2(1 - x)$, for $x \in [0, 1]$ (note that this is nonnegative and integrates to 1). The conditional PDF of X given Y is

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} = \frac{2}{2(1 - y)} = \frac{1}{1 - y},$$

for (x, y) in the triangle (and 0 otherwise). Since $\frac{1}{1-y}$ is constant with respect to x, we have $X|Y \sim \text{Unif}(0, 1 - Y)$.

19. ⓢ A random point (X, Y, Z) is chosen uniformly in the ball

$$B = \{(x, y, z) : x^2 + y^2 + z^2 \leq 1\}.$$

(a) Find the joint PDF of X, Y, Z.

(b) Find the joint PDF of X, Y.

(c) Find an expression for the marginal PDF of X, as an integral.

Solution:

(a) Just as in 2 dimensions uniform in a region means that probability is proportional to area, in 3 dimensions probability is proportional to volume. That is,

$$P((X, Y, Z) \in A) = c \cdot \text{volume}(A)$$

if A is contained in B, where c is a constant. Letting $A = B$, we have that $\frac{1}{c}$ is $\frac{4}{3}\pi$, the volume of the ball. So the joint PDF of (X, Y, Z) is

$$f(x, y, z) = \begin{cases} \frac{3}{4\pi}, & \text{if } x^2 + y^2 + z^2 \leq 1; \\ 0, & \text{otherwise.} \end{cases}$$

(b) We just need to integrate out the z from the joint PDF of X, Y, Z. The limits of

integration are found by noting that for any (x, y), we need to have z satisfy $x^2+y^2+z^2 \leq 1$.

$$f_{X,Y}(x, y) = \int_{-\infty}^{\infty} f(x, y, z)dz$$

$$= \frac{3}{4\pi} \int_{-\sqrt{1-x^2-y^2}}^{\sqrt{1-x^2-y^2}} dz$$

$$= \frac{3}{2\pi} \sqrt{1 - x^2 - y^2},$$

for $x^2 + y^2 \leq 1$ (and the PDF is 0 otherwise).

(c) We can integrate out y, z from the joint PDF of X, Y, Z, or integrate out y from the joint PDF of X, Y. Using the result of (b), we have for $-1 \leq x \leq 1$ that the marginal PDF of X is

$$f_X(x) = \frac{3}{2\pi} \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \sqrt{1 - x^2 - y^2}dy.$$

20. Ⓢ Let U_1, U_2, U_3 be i.i.d. Unif$(0, 1)$, and let $L = \min(U_1, U_2, U_3)$, $M = \max(U_1, U_2, U_3)$.

(a) Find the marginal CDF and marginal PDF of M, and the joint CDF and joint PDF of L, M.

Hint: For the latter, start by considering $P(L \geq l, M \leq m)$.

(b) Find the conditional PDF of M given L.

Solution:

(a) The event $M \leq m$ is the same as the event that all 3 of the U_j are at most m, so the CDF of M is $F_M(m) = m^3$ and the PDF is $f_M(m) = 3m^2$, for $0 \leq m \leq 1$.

The event $L \geq l, M \leq m$ is the same as the event that all 3 of the U_j are between l and m (inclusive), so

$$P(L \geq l, M \leq m) = (m - l)^3$$

for $m \geq l$ with $m, l \in [0, 1]$. By the axioms of probability, we have

$$P(M \leq m) = P(L \leq l, M \leq m) + P(L > l, M \leq m).$$

So the joint CDF is

$$P(L \leq l, M \leq m) = m^3 - (m - l)^3,$$

for $m \geq l$ with $m, l \in [0, 1]$. The joint PDF is obtained by differentiating this with respect to l and then with respect to m (or vice versa):

$$f(l, m) = 6(m - l),$$

for $m \geq l$ with $m, l \in [0, 1]$. As a check, note that getting the marginal PDF of M by finding $\int_0^m f(l, m)dl$ does recover the PDF of M (the limits of integration are from 0 to m since the min can't be more than the max).

(b) The marginal PDF of L is $f_L(l) = 3(1 - l)^2$ for $0 \leq l \leq 1$ since $P(L > l) = P(U_1 > l, U_2 > l, U_3 > l) = (1-l)^3$ (alternatively, use the PDF of M together with the symmetry that $1 - U_j$ has the same distribution as U_j, or integrate out m in the joint PDF of L, M). So the conditional PDF of M given L is

$$f_{M|L}(m|l) = \frac{f(l, m)}{f_L(l)} = \frac{2(m - l)}{(1 - l)^2},$$

for all $m, l \in [0, 1]$ with $m \geq l$.

21. Find the probability that the quadratic polynomial $Ax^2 + Bx + 1$, where the coefficients A and B are determined by drawing i.i.d. $\text{Unif}(0, 1)$ random variables, has at least one real root.

Hint: By the quadratic formula, the polynomial $ax^2 + bx + c$ has a real root if and only if $b^2 - 4ac \geq 0$.

Solution: We need to find $P(B^2 - 4A \geq 0)$. This can be done by integrating the joint PDF of A and B over the appropriate region:

$$P(B^2 - 4A \geq 0) = P(A \leq \frac{B^2}{4}) = \int_0^1 \int_0^{\min(b^2/4, 1)} da\, db = \int_0^1 \int_0^{b^2/4} da\, db = \frac{1}{4} \int_0^1 b^2 db = \frac{1}{12}.$$

22. Let X and Y each have support $(0, \infty)$ marginally, and suppose that the joint PDF $f_{X,Y}$ of X and Y is positive for $0 < x < y$ and 0 otherwise.

(a) What is the support of the conditional PDF of Y given $X = x$?

(b) Show that X and Y can't be independent.

Solution:

(a) Let $x > 0$. The support of the conditional PDF of Y given $X = x$ is the interval (x, ∞) since $f_{X,Y}(x, y)$ (which is the numerator in the definition of the conditional PDF) is positive if and only if y is in (x, ∞).

(b) The support of the conditional PDF of Y given $X = x$ depends on x, so X and Y are dependent; knowing the value of X constrains the possible values of Y.

23. The *volume* of a region in n-dimensional Euclidean space $\mathbb{R}^n$ is the integral of 1 over that region. The *unit ball* in $\mathbb{R}^n$ is $\{(x_1, \ldots, x_n) : x_1^2 + \cdots + x_n^2 \leq 1\}$, the ball of radius 1 centered at 0. As mentioned in Section A.7 of the math appendix, the volume of the unit ball in n dimensions is

$$v_n = \frac{\pi^{n/2}}{\Gamma(n/2 + 1)},$$

where Γ is the *gamma function*, a very famous function which is defined by

$$\Gamma(a) = \int_0^\infty x^a e^{-x} \frac{dx}{x}$$

for all $a > 0$, and which will play an important role in the next chapter. A few useful facts about the gamma function (which you can assume) are that $\Gamma(a + 1) = a\Gamma(a)$ for any $a > 0$, and that $\Gamma(1) = 1$ and $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. Using these facts, it follows that $\Gamma(n) = (n - 1)!$ for n a positive integer, and we can also find $\Gamma(n + \frac{1}{2})$ when n is a nonnegative integer. For practice, please verify that $v_2 = \pi$ (the area of the unit disk in 2 dimensions) and $v_3 = \frac{4}{3}\pi$ (the volume of the unit ball in 3 dimensions).

Let $U_1, U_2, \ldots, U_n \sim \text{Unif}(-1, 1)$ be i.i.d.

(a) Find the probability that $(U_1, U_2, \ldots, U_n)$ is in the unit ball in $\mathbb{R}^n$.

(b) Evaluate the result from (a) numerically for $n = 1, 2, \ldots, 10$, and plot the results (using a computer unless you are extremely good at making hand-drawn graphs). The facts above about the gamma function are sufficient so that you can do this without doing any integrals, but you can also use the command `gamma` in R to compute the gamma function.

(c) Let c be a constant with $0 < c < 1$, and let X_n count how many of the U_j satisfy $|U_j| > c$. What is the distribution of X_n?

(d) For $c = 1/\sqrt{2}$, use the result of Part (c) to give a simple, short derivation of what happens to the probability from (a) as $n \to \infty$.
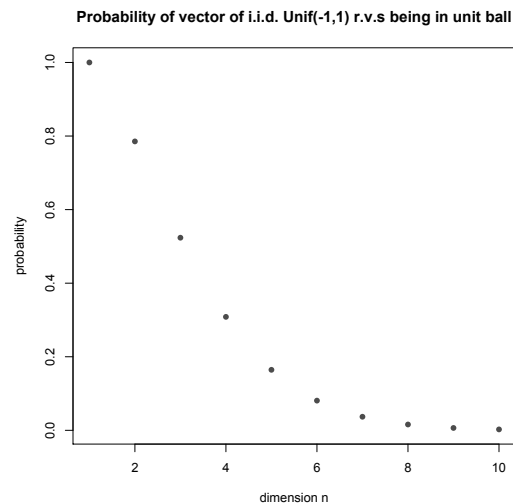
Solution:

(a) The random vector $(U_1, U_2, \ldots, U_n)$ is Uniform in $\{(x_1, \ldots, x_n) : x_i \in [-1, 1]$ for all $i\}$, a box in $\mathbb{R}^n$. So the probability of $(U_1, \ldots, U_n)$ being in a specific subset A of the box is proportional to the volume of A. Thus, the desired probability is $v_n/2^n$.

(b) Typing `n <- 1:10; pi^(n/2)/(gamma(n/2+1)*2^n)` in R gives the following results:

dimension	probability
1	1
2	0.785
3	0.524
4	0.308
5	0.164
6	0.081
7	0.037
8	0.016
9	0.006
10	0.002

Using the `plot` command, we can visualize how quickly the probability approaches 0.

Probability of vector of i.i.d. Unif(-1,1) r.v.s being in unit ball

(c) By the story of the Binomial, $X_n \sim \text{Bin}(n, p)$, with

$$p = 1 - P(-c \le U_1 \le c) = 1 - \frac{2c}{2} = 1 - c.$$

(d) If two U_j exceed $1/\sqrt{2}$ in absolute value, that already is enough to push the vector out of the unit ball. In view of this, it is not surprising that the probability from (a) goes to 0 rapidly as n increases. With notation as above,

$$P((U_1, \ldots, U_n) \text{ in unit ball}) \le P(X_n \le 1) = (1 - p)^n + np(1 - p)^{n-1} \to 0$$

rapidly as $n \to \infty$, since $(1-p)^n$ goes to 0 exponentially fast.

24. ⓢ Two students, A and B, are working independently on homework (not necessarily for the same class). Student A takes $Y_1 \sim \text{Expo}(\lambda_1)$ hours to finish their homework, while B takes $Y_2 \sim \text{Expo}(\lambda_2)$ hours.

(a) Find the CDF and PDF of Y_1/Y_2, the ratio of their problem-solving times.

(b) Find the probability that A finishes their homework before B does.

Solution:

(a) Let $t > 0$. The CDF of the ratio is

$$
\begin{aligned}
F(t) = P\left(\frac{Y_1}{Y_2} \le t\right) = P(Y_1 \le tY_2) \\
= \int_0^\infty \left(\int_0^{ty_2} \lambda_1 e^{-\lambda_1 y_1} dy_1\right) \lambda_2 e^{-\lambda_2 y_2} dy_2 \\
= \int_0^\infty (1 - e^{-\lambda_1 ty_2}) \lambda_2 e^{-\lambda_2 y_2} dy_2 \\
= 1 - \int_0^\infty \lambda_2 e^{-(\lambda_1 t + \lambda_2) y_2} dy_2 \\
= 1 - \frac{\lambda_2}{t\lambda_1 + \lambda_2} \\
= \frac{t\lambda_1}{t\lambda_1 + \lambda_2}.
\end{aligned}
$$

Of course, $F(t) = 0$ for $t \le 0$. The PDF of the ratio is

$$
f(t) = \frac{d}{dt}\left(\frac{t\lambda_1}{t\lambda_1 + \lambda_2}\right) = \frac{\lambda_1 \lambda_2}{(\lambda_1 t + \lambda_2)^2}, \text{ for } t > 0.
$$

(b) Plugging in $t = 1$ above, we have

$$
P(Y_1 < Y_2) = \frac{\lambda_1}{\lambda_1 + \lambda_2}.
$$

Alternatively, we can get the same result by applying Example 7.1.23. (The result can also be derived without using calculus by thinking about Poisson processes, as shown in Chapter 13.)

25. Two companies, Company 1 and Company 2, have just been founded. Stock market crashes occur according to a Poisson process with rate λ_0. Such a crash would put both companies out of business. For $j \in \{1, 2\}$, there may be an adverse event "of type j," which puts Company j out of business (if it is not already out of business) but does not affect the other company; such events occur according to a Poisson process with rate λ_j. If there has not been a stock market crash or an adverse event of type j, then Company j remains in business. The three Poisson processes are independent of each other. Let X_1 and X_2 be how long Company 1 and Company 2 stay in business, respectively.

(a) Find the marginal distributions of X_1 and X_2.

(b) Find $P(X_1 > x_1, X_2 > x_2)$, and use this to find the joint CDF of X_1 and X_2.

Solution:

(a) Let $T_0 \sim \text{Expo}(\lambda_0)$ be the time of the first stock market crash and $T_j \sim \text{Expo}(\lambda_j)$ the the time of the first adverse event of type j for $j = 1, 2$. Then $X_1 = \min(T_0, T_1)$

and $X_2 \sim \min(T_0, T_2)$. So (by finding the CDF or applying Example 5.6.3) $X_1 \sim$ Expo$(\lambda_0 + \lambda_1)$, $X_2 \sim$ Expo$(\lambda_0 + \lambda_2)$.

(b) For any positive x_1, x_2, we have

$$
\begin{aligned}
P(X_1 > x_1, X_2 > x_2) &= P(T_0 > x_1, T_1 > x_1, T_0 > x_2, T_2 > x_2) \\
&= P(T_0 > \max(x_1, x_2), T_1 > x_1, T_2 > x_2) \\
&= P(T_0 > \max(x_1, x_2))P(T_1 > x_1)P(T_2 > x_2) \\
&= e^{-\lambda_0 \max(x_1,x_2) - \lambda_1 x_1 - \lambda_2 x_2},
\end{aligned}
$$

so the joint CDF is

$$
\begin{aligned}
P(X_1 \le x_1, X_2 \le x_2) &= 1 - P(X_1 > x_1 \text{ or } X_2 > x_2) \\
&= 1 - (P(X_1 > x_1) + P(X_2 > x_2) - P(X_1 > x_1, X_2 > x_2)) \\
&= 1 - e^{-(\lambda_0 + \lambda_1)x_1} - e^{-(\lambda_0 + \lambda_2)x_2} + e^{-\lambda_0 \max(x_1,x_2) - \lambda_1 x_1 - \lambda_2 x_2}.
\end{aligned}
$$

26. ⓢ The bus company from Blissville decides to start service in Blotchville, sensing a promising business opportunity. Meanwhile, Fred has moved back to Blotchville. Now when Fred arrives at the bus stop, either of two independent bus lines may come by (both of which take him home). The Blissville company's bus arrival times are exactly 10 minutes apart, whereas the time from one Blotchville company bus to the next is Expo$(\frac{1}{10})$. Fred arrives at a uniformly random time on a certain day.

(a) What is the probability that the Blotchville company bus arrives first?

Hint: One good way is to use the continuous law of total probability.

(b) What is the CDF of Fred's waiting time for a bus?

Solution:

(a) Let $U \sim$ Unif$(0, 10)$ be the arrival time of the next Blissville company bus, and $X \sim$ Expo$(\frac{1}{10})$ be the arrival time of the next Blotchville company bus (the latter is $X \sim$ Expo$(\frac{1}{10})$ by the memoryless property). Then

$$
\begin{aligned}
P(X < U) &= \int_0^{10} P(X < U | U = u) \frac{1}{10} du \\
&= \frac{1}{10} \int_0^{10} P(X < u | U = u) du \\
&= \frac{1}{10} \int_0^{10} (1 - e^{-u/10}) du = \frac{1}{e}.
\end{aligned}
$$

(b) Let $T = \min(X, U)$ be the waiting time. Then

$$
P(T > t) = P(X > t, U > t) = P(X > t)P(U > t).
$$

So the CDF of T is

$$
P(T \le t) = 1 - P(X > t)P(U > t) = 1 - e^{-t/10}(1 - t/10),
$$

for $0 < t < 10$ (and 0 for $t \le 0$, and 1 for $t \ge 10$).

27. A longevity study is being conducted on n married hobbit couples. Let p be the probability that an individual hobbit lives at least until their eleventy-first birthday, and assume that the lifespans of different hobbits are independent. Let N_0, N_1, N_2 be the number of couples in which neither hobbit reaches age eleventy-one, one hobbit does but not the other, and both hobbits reach eleventy-one, respectively.

(a) Find the joint PMF of N_0, N_1, N_2.

For the rest of this problem, suppose that it is observed that exactly h of the cohort of hobbits reach their eleventy-first birthdays.

(b) Using (a) and the definition of conditional probability, find the conditional PMF of N_2 given this information, up to a normalizing constant (that is, you do not need to find the normalizing constant in this part, but just to give a simplified expression that is proportional to the conditional PMF). For simplicity, you can and should ignore multiplicative constants in this part; this includes multiplicative factors that are functions of h, since h is now being treated as a known constant.

(c) Now obtain the conditional PMF of N_2 using a direct counting argument, now including any normalizing constants needed in order to have a valid conditional PMF.

(d) Discuss intuitively whether or not p should appear in the answer to (c).

(e) What is the conditional expectation of N_2, given the above information (simplify fully)? This can be done without doing any messy sums.

Solution:

(a) Let $q = 1 - p$. By the story of the Multinomial, $(N_0, N_1, N_2) \sim \text{Mult}_3(n, (q^2, 2pq, p^2))$. The joint PMF is

$$P(N_0 = n_0, N_1 = n_1, N_2 = n_2) = \frac{n!}{n_0! n_1! n_2!} q^{2n_0} (2pq)^{n_1} p^{2n_2}$$

$$= \frac{n!}{n_0! n_1! n_2!} 2^{n_1} p^{n_1 + 2n_2} q^{2n_0 + n_1},$$

for n_0, n_1, n_2 nonnegative integers with $n_0 + n_1 + n_2 = n$.

(b) We observe that $N_1 + 2N_2 = h$. The conditional PMF of N_2 is

$$
\begin{aligned}
P(N_2 = n_2 | 2N_2 + N_1 = h) &= \frac{P(N_2 = n_2, 2N_2 + N_1 = h)}{P(2N_2 + N_1 = h)} \\
&= \frac{P(N_2 = n_2, N_1 = h - 2n_2, N_0 = n - n_2 - (h - 2n_2))}{P(2N_2 + N_1 = h)} \\
&= \frac{P(N_0 = n + n_2 - h, N_1 = h - 2n_2, N_2 = n_2)}{P(2N_2 + N_1 = h)} \\
&\propto \frac{n!}{(n + n_2 - h)!(h - 2n_2)! n_2!} 2^{h - 2n_2} p^h q^{2(n + n_2 - h) + h - 2n_2} \\
&\propto \frac{1}{2^{2n_2}(n + n_2 - h)!(h - 2n_2)! n_2!},
\end{aligned}
$$

since n and h are treated as known constants for this calculation (so, for example, the $1/P(2N_2 + N_1 = h)$ factor is just a multiplicative constant, as is the p^h). The support consists of all nonnegative integers n_2 such that $h - n \le n_2 \le h/2$, since the values of $N_0, N_1,$ and N_2 must be nonnegative integers.

(c) There are $2n$ hobbits. Let's "tag" h of them as alive at age 111. All subsets of size h are equally likely, so we can use the naive definition of probability to get

$$P(N_2 = n_2 | 2N_2 + N_1 = h) = \frac{2^{h - 2n_2} \binom{n}{n_2} \binom{n - n_2}{h - 2n_2}}{\binom{2n}{h}},$$

where the numerator was found using the multiplication rule: choose which n_2 couples have both hobbits tagged, then choose $h - 2n_2$ couples for which exactly 1 hobbit is tagged (so the remaining $n + n_2 - h$ couples have neither hobbit tagged), then for each

of those $h - 2n_2$ couples choose which partner is tagged. As in (b), the support consists of all nonnegative integers n_2 such that $h - n \le n_2 \le h/2$.

(d) The p has disappeared after conditioning in how many hobbits were alive. This is reminiscent of p disappearing after conditioning in the Fisher exact test. It makes sense intuitively that this cancellation occurs. The parameter p was important for generating h, which is a random draw from the $\text{Bin}(2n, p)$ distribution. But once h is known, p has played its part and what matters now is that we have a finite population of $2n$ hobbits in which the hobbits in a simple random sample of size h are "tagged" as alive.

(e) Create an indicator r.v. for each couple, indicating whether both hobbits in the couple reach age eleven-one. Now consider a particular couple, say Belladonna and Bungo. Then

$$P(\text{Belladonna and Bungo reach age eleven-one}|2N_2 + N_1 = h) = \frac{h(h-1)}{2n(2n-1)},$$

since Belladonna has an $h/(2n)$ chance of reaching age eleven-one, and given that she does, Bungo has an $(h-1)/(2n-1)$ chance (for $h \ge 1$). Equivalently, the probability is $\binom{h}{2}/\binom{2n}{2}$, since the conditional distribution of how many of Belladonna and Bungo reach age eleven-one is $\text{HGeom}(h, 2n - h, 2)$.

By symmetry, linearity, and the fundamental bridge,

$$E(N_2|2N_2 + N_1 = h) = \frac{h(h-1)}{2(2n-1)}.$$

28. There are n stores in a shopping center, labeled from 1 to n. Let X_i be the number of customers who visit store i in a particular month, and suppose that $X_1, X_2, \ldots, X_n$ are i.i.d. with PMF $p(x) = P(X_i = x)$. Let $I \sim \text{DUnif}(1, 2, \ldots, n)$ be the label of a randomly chosen store, so X_I is the number of customers at a randomly chosen store.

(a) For $i \ne j$, find $P(X_i = X_j)$ in terms of a sum involving the PMF $p(x)$.

(b) Find the joint PMF of I and X_I. Are they independent?

(c) Does X_I, the number of customers for a random store, have the same marginal distribution as X_1, the number of customers for store 1?

(d) Let $J \sim \text{DUnif}(1, 2, \ldots, n)$ also be the label of a randomly chosen store, with I and J independent. Find $P(X_I = X_J)$ in terms of a sum involving the PMF $p(x)$. How does $P(X_I = X_J)$ compare to $P(X_i = X_j)$ for fixed i, j with $i \ne j$?

Solution:

(a) Breaking into cases,

$$P(X_i = X_j) = \sum_x P(X_i = x, X_j = x) = \sum_x P(X_i = x)P(X_j = x) = \sum_x p(x)^2.$$

(b) Since I is independent of each X_i, the joint PMF of I and X_I is

$$P(I = i, X_I = x) = P(I = i)P(X_I = x|I = i) = P(I = i)P(X_i = x) = \frac{p(x)}{n},$$

for $i = 1, 2, \ldots, n$ and x in the support of X_1. The joint PMF factors as a function of i (actually a constant function) times a function of x, so I and X_I are independent. This independence makes sense intuitively, despite the fact that the symbol "I" appears in the notation "X_I"; knowing, for example, that store 3 was chosen does not help us predict how many customers store 3 had.

(c) By LOTP, the marginal distribution of X_I is

$$P(X_I = x) = \sum_{i=1}^{n} P(X_I = x | I = i) P(I = i) = \frac{1}{n} \sum_{i=1}^{n} p(x) = p(x).$$

So X_I has the same marginal distribution as each X_i.

(d) Let's use LOTP to condition on which stores were randomly chosen:

$$P(X_I = X_J) = \sum_{i,j} P(X_I = X_J | I = i, J = j) P(I = i, J = j) = \frac{1}{n^2} \sum_{i,j} P(X_i = X_j).$$

Let a be the answer from (a). We have $P(X_i = X_j) = 1$ if $i = j$, and $P(X_i = X_j) = a$ if $i \neq j$. Thus,

$$P(X_I = X_J) = \frac{1}{n^2} \left(n + n(n-1)a \right) = \frac{1}{n} + \left(1 - \frac{1}{n} \right) a = a + \frac{1-a}{n}.$$

This makes sense intuitively since it reflects the fact that with probability $1/n$, the same store will be chosen twice (and then $X_I = X_J$) and with probability $1 - 1/n$, distinct stores will be chosen (and then the probability is as in Part (a)). We have $P(X_I = X_J) > a$ (except in the degenerate case $a = 1$).

29. Let X and Y be i.i.d. Geom(p), $L = \min(X, Y)$, and $M = \max(X, Y)$.

(a) Find the joint PMF of L and M. Are they independent?

(b) Find the marginal distribution of L in two ways: using the joint PMF, and using a story.

(c) Find EM.

Hint: A quick way is to use (b) and the fact that $L + M = X + Y$.

(d) Find the joint PMF of L and $M - L$. Are they independent?

Solution:

(a) Let $q = 1 - p$, and let l and m be nonnegative integers. For $l < m$,

$$P(L = l, M = m) = P(X = l, Y = m) + P(X = m, Y = l) = 2p^2 q^{l+m}.$$

For $l = m$,
$$P(L = l, M = m) = P(X = l, Y = l) = p^2 q^{2l}.$$

For $l > m$, $P(L = l, M = m) = 0$. The r.v.s L and M are *not* independent since we know for sure that $L \leq M$ will occur, so learning the value of L can give us information about M. We can also write the joint PMF in one expression as

$$P(L = l, M = m) = 2^{I(l<m)} p^2 q^{l+m} I(l \leq m),$$

where $I(l < m)$ is 1 if $l < m$ and 0 otherwise, and $I(l \leq m)$ is 1 if $l \leq m$ and 0 otherwise. This way of writing it makes it easier to see why the joint PMF does not factor into the product of a function of l and a function of m.

(b) We can sum the joint PMF over all possible values of M to get the marginal distri-

bution of L:

$$P(L = l) = \sum_{m=l}^{\infty} P(L = l, M = m)$$

$$= p^2 q^{2l} + 2p^2 q^l \sum_{m=l+1}^{\infty} q^m$$

$$= p^2 q^{2l} + (2p^2 q^l q^{l+1})/(1 - q)$$

$$= p^2 q^{2l} + 2pq^{2l+1}$$

$$= q^{2l}(p^2 + 2pq)$$

$$= q^{2l}(1 - q^2),$$

where in the last step we used the fact that $p^2 + 2pq + q^2 = (p+q)^2 = 1$. An easier way to get the same result is to use the story of the Geometric: imagining two independent sequences of independent Bern(p) trials and considering whether at time n at least one of the two trials at that time was a success, we have $L \sim \text{Geom}(1 - q^2)$. This agrees with the above since the PMF of a Geom$(1 - q^2)$ r.v. is $(1 - q^2)q^{2l}$ for $l = 0, 1, \dots$.

(c) We have $EL = q^2/(1 - q^2)$ and

$$EL + EM = E(L + M) = E(X + Y) = EX + EY = 2q/p,$$

so

$$EM = \frac{2q}{p} - \frac{q^2}{1 - q^2} = \frac{(1 - p)(3 - p)}{p(2 - p)}.$$

(d) By (a),

$$P(L = l, M - L = k) = P(L = l, M = k + l) = 2^{I(k>0)} p^2 q^{2l+k},$$

where $I(k > 0)$ is 1 if $k > 0$ and 0 otherwise. This factors as

$$P(L = l, M - L = k) = f(l)g(k)$$

for all nonnegative integers l, k, where

$$f(l) = (1 - q^2)q^{2l}, g(k) = \frac{2^{I(k>0)} p^2 q^k}{1 - q^2}.$$

Thus, L and $M - L$ are independent. (Since f is the PMF of L, by summing the joint PMF over the possible values of L we also have that the PMF of $M - L$ is g. The PMF g looks complicated because of the possibility of a "tie" occurring (the event $X = Y$), but conditional on a tie not occurring we have the nice result

$$M - L - 1 | M - L > 0 \sim \text{Geom}(p),$$

which makes sense due to the memoryless property of the Geometric. To check this conditional distribution, use the definition of conditional probability and the fact that $p^2 + q^2 = 1 - 2pq$.)

30. Let X, Y have the joint CDF

$$F(x, y) = 1 - e^{-x} - e^{-y} + e^{-(x+y+\theta xy)},$$

for $x > 0, y > 0$ (and $F(x, y) = 0$ otherwise), where the parameter θ is a constant in $[0, 1]$.

(a) Find the joint PDF of X, Y. For which values of θ (if any) are they independent?

(b) Explain why we require θ to be in $[0, 1]$.

(c) Find the marginal PDFs of X and Y by working directly from the joint PDF from (a). When integrating, do *not* use integration by parts or computer assistance; rather, *pattern match* to facts we know about moments of famous distributions.

(d) Find the marginal CDFs of X and Y by working directly from the joint CDF.

Solution:

(a) Differentiating F with respect to x and then y, we have for $x > 0, y > 0$

$$\partial F(x, y)/\partial x = -(1 + \theta y)e^{-(x+y+\theta xy)},$$

$$f(x, y) = ((1 + \theta x)(1 + \theta y) - \theta)\, e^{-(x+y+\theta xy)}.$$

For $\theta = 0$, this factors as $e^{-x}e^{-y}$, showing that in this case X and Y are i.i.d. Expo(1).

(b) Note that $(1 + \theta x)(1 + \theta y) - \theta \to 1 - \theta$ as x and y go to 0. So if $\theta > 1$, we would be able to make $f(x, y) < 0$ by choosing x and y to be small enough, showing that $f(x, y)$ is not a valid joint PDF. If $\theta < 0$, then $F(x, y)$ can be made to blow up to ∞ by taking $x = y$ to be very large, since then $-(x + y + \theta xy) = -\theta x^2 - 2x \to \infty$ as $x \to \infty$, while the e^{-x} and e^{-y} terms go to 0. So F is not a valid joint CDF for $\theta < 0$.

(c) Let $\lambda = 1 + \theta x$. The marginal PDF of X is

$$\begin{aligned}
f_X(x) &= \int_0^\infty f(x, y)dy \\
&= e^{-x}\int_0^\infty (\lambda(1 + \theta y) - \theta)e^{-\lambda y}dy \\
&= e^{-x}\int_0^\infty (1 + \theta y - \theta/\lambda)\lambda e^{-\lambda y}dy \\
&= e^{-x}\left((1 - \theta/\lambda)\int_0^\infty \lambda e^{-\lambda y}dy + \theta\int_0^\infty y e^{-\lambda y}dy\right) \\
&= e^{-x}
\end{aligned}$$

for $x > 0$, since an Expo(λ) r.v. has a valid PDF and has mean $1/\lambda$. So $X \sim$ Expo(1). Similarly, $f_Y(y) = e^{-y}$ for $y > 0$, so $Y \sim$ Expo(1).

(d) The marginal CDF of X is

$$F_X(x) = P(X \le x) = \lim_{y \to \infty} P(X \le x, Y \le y) = 1 - e^{-x}$$

for $x > 0$, since letting $y \to \infty$ makes $X \le x, Y \le y$ reduce to just $X \le x$. So $X \sim$ Expo(1). Similarly, the marginal CDF of Y is

$$F_Y(y) = 1 - e^{-y}$$

for $y > 0$, so $Y \sim$ Expo(1).

2D LOTUS

31. Ⓢ Let X and Y be i.i.d. Unif$(0, 1)$. Find the standard deviation of the distance between X and Y.

Solution: Let $W = |X - Y|$. By 2-D LOTUS,

$$E(W) = \int_0^1 \int_0^1 |x - y|dxdy.$$

Split this into two parts (to get rid of the absolute values): $x < y$ and $x \geq y$ (i.e., break the square into two triangles). By symmetry the integral over $x < y$ equals the integral over $x > y$, so

$$E(W) = 2 \int_0^1 \int_0^y (y - x) dx dy = 2 \int_0^1 \frac{y^2}{2} dy = \frac{1}{3}.$$

Next, we find $E(W^2)$. This can either be done by computing the double integral

$$E(W^2) = \int_0^1 \int_0^1 (x - y)^2 dx dy,$$

or by writing

$$E(W^2) = E(X - Y)^2 = EX^2 + EY^2 - 2E(XY),$$

which is

$$2E(X^2) - 2(EX)^2 = 2\text{Var}(X) = \frac{1}{6},$$

since $E(XY) = E(X)E(Y)$ for X, Y independent, and $E(X) = E(Y)$ and $E(X^2) = E(Y^2)$ (as X and Y have the same distribution). Thus, $E(W) = 1/3$,

$$\text{Var}(W) = E(W^2) - (E(W))^2 = \frac{1}{18},$$

and the standard deviation of the distance between X and Y is $\frac{1}{\sqrt{18}} = \frac{1}{3\sqrt{2}}$.

32. Ⓢ Let X, Y be i.i.d. Expo(λ). Find $E|X - Y|$ in two different ways: (a) using 2D LOTUS and (b) using the memoryless property without any calculus.

Solution:

(a) First consider the case $\lambda = 1$. By LOTUS,

$$\begin{aligned} E|X - Y| &= \int_0^\infty \int_0^\infty |x - y| e^{-x} e^{-y} dx dy \\ &= 2 \int_0^\infty \int_y^\infty (x - y) e^{-x} e^{-y} dx dy \\ &= 2 \int_0^\infty e^{-y} \int_y^\infty (x e^{-x} - y e^{-x}) dx dy \\ &= 2 \int_0^\infty e^{-y} \left(-e^{-x}(x + 1) + y e^{-x} \right) |_y^\infty dy \\ &= 2 \int_0^\infty e^{-y} e^{-y} dy = 2 \int_0^\infty e^{-2y} dy = 1. \end{aligned}$$

For general λ, this and the fact that $\lambda X, \lambda Y$ are i.i.d. Expo(1) yield $E|X - Y| = 1/\lambda$.

(b) Write $|X - Y| = \max(X, Y) - \min(X, Y)$. By the memoryless property, this is Expo(λ), as in Example 7.3.6. So $E|X - Y| = \frac{1}{\lambda}$, which agrees with (a). (This also shows that $\text{Var}(|X - Y|) = \frac{1}{\lambda^2}$.)

33. Alice walks into a post office with 2 clerks. Both clerks are in the midst of serving customers, but Alice is next in line. The clerk on the left takes an Expo(λ_1) time to serve a customer, and the clerk on the right takes an Expo(λ_2) time to serve a customer. Let T be the amount of time Alice has to wait until it is her turn.

(a) Write down expressions for the mean and variance of T, in terms of double integrals (which you do not need to evaluate).

(b) Find the distribution, mean, and variance of T, *without using calculus*.

Solution:

(a) Let T_1 be the time it takes until the clerk on the left finishes serving their current customer, measured starting at the time of Alice's arrival, and define likewise T_2 for the clerk on the right. By the memoryless property, it does not matter how long the clerks were serving their current customers before Alice arrived, so $T_1 \sim \text{Expo}(\lambda_1)$ and $T_2 \sim \text{Expo}(\lambda_2)$. We are interested in $T = \min(T_1, T_2)$. By 2D LOTUS,

$$E(T) = \int_0^\infty \int_0^\infty \min(t_1, t_2) \lambda_1 e^{-\lambda_1 t_1} \lambda_2 e^{-\lambda_2 t_2} dt_1 dt_2,$$

$$\text{Var}(T) = \int_0^\infty \int_0^\infty \min(t_1^2, t_2^2) \lambda_1 e^{-\lambda_1 t_1} \lambda_2 e^{-\lambda_2 t_2} dt_1 dt_2 - (E(T))^2.$$

(b) By Example 5.6.3, $T \sim \text{Expo}(\lambda_1 + \lambda_2)$. We can also show this directly: for $t > 0$, $P(T \le t) = 1 - P(T > t)$, where

$$P(T > t) = P(T_1 > t, T_2 > t) = P(T_1 > t)P(T_2 > t) = e^{-(\lambda_1 + \lambda_2)t}.$$

Therefore, $E(T) = \frac{1}{\lambda_1 + \lambda_2}$ and $\text{Var}(T) = \frac{1}{(\lambda_1 + \lambda_2)^2}$.

34. Let (X, Y) be a uniformly random point in the triangle in the plane with vertices $(0,0), (0,1), (1,0)$. Find $\text{Cov}(X, Y)$. (Exercise 18 is about joint, marginal, and conditional PDFs in this setting.)

Solution: As shown in the solution to Exercise 18, the joint PDF is 2 inside the triangle (and 0 outside), and the marginal PDF of X is $2(1 - x)$ for $0 \le x \le 1$. Similarly, the marginal PDF of Y is $2(1 - y)$ for $0 \le y \le 1$. So

$$E(Y) = E(X) = \int_0^1 2x(1 - x)dx = 2(x^2/2 - x^3/3)\Big|_0^1 = \frac{1}{3}.$$

By 2D LOTUS,

$$E(XY) = \int_0^1 \int_0^{1-y} 2xy \, dx \, dy = \int_0^1 y(1-y)^2 dy = \int_0^1 (y^3 - 2y^2 + y)dy = \frac{1}{4} - \frac{2}{3} + \frac{1}{2} = \frac{1}{12}.$$

Thus,

$$\text{Cov}(XY) = E(XY) - E(X)E(Y) = \frac{1}{12} - \frac{1}{3^2} = -\frac{1}{36}.$$

35. A random point is chosen uniformly in the unit disk $\{(x, y) : x^2 + y^2 \le 1\}$. Let R be its distance from the origin.

(a) Find $E(R)$ using 2D LOTUS.

Hint: To do the integral, convert to polar coordinates (see the math appendix).

(b) Find the CDFs of R^2 and of R *without using calculus*, using the fact that for a Uniform distribution on a region, probability within that region is proportional to area. Then get the PDFs of R^2 and of R, and find $E(R)$ in two more ways: using the definition of expectation, and using a 1D LOTUS by thinking of R as a function of R^2.

Solution:

(a) By 2D LOTUS,
$$E(R) = \frac{1}{\pi} \iint_D \sqrt{x^2 + y^2} dx dy,$$

where D is the unit disk. Converting to polar coordinates, we have

$$E(R) = \frac{1}{\pi} \int_0^{2\pi} \int_0^1 r^2 dr d\theta = \frac{1}{\pi} \int_0^{2\pi} \frac{1}{3} d\theta = \frac{2}{3}.$$

(b) Let $0 < a < 1$. Since the probability of the random point being in $\{x^2 + y^2 \le a^2\}$ is proportional to the area of this disk, we have

$$P(R \le a) = P(R^2 \le a^2) = \frac{\text{area of the disk } \{x^2 + y^2 \le a\}}{\text{area of the disk } \{x^2 + y^2 \le 1\}} = a^2.$$

So the CDFs are

$$P(R^2 \le t) = t$$

for $0 < t < 1$ (and 0 for $t \le 0$, and 1 for $t \ge 1$), and

$$P(R \le r) = r^2$$

for $0 < r < 1$ (and 0 for $r \le 0$, and 1 for $r \ge 1$). Then the PDF of R^2 is

$$f_{R^2}(t) = 1$$

for $0 < t < 1$ and 0 otherwise; this just says that $R^2 \sim \text{Unif}(0, 1)$. The PDF of R is

$$f_R(r) = 2r$$

for $0 < r < 1$ (and 0 otherwise). Now we can get $E(R)$ in two more ways:

$$E(R) = \int_0^1 r f_R(r) dr = \int_0^1 2r^2 dr = (2r^3/3)\big|_0^1 = \frac{2}{3},$$

by definition. Letting $T = R^2$, we also have

$$E(R) = E(\sqrt{T}) = \int_0^1 \sqrt{t} f_T(t) dt = \int_0^1 t^{1/2} dt = (2/3)t^{3/2}\big|_0^1 = \frac{2}{3}.$$

36. Let X and Y be discrete r.v.s.

(a) Use 2D LOTUS (without assuming linearity) to show that $E(X+Y) = E(X)+E(Y)$.

(b) Now suppose that X and Y are independent. Use 2D LOTUS to show that $E(XY) = E(X)E(Y)$.

Solution:

(a) Let $p(x, y) = P(X = x, Y = y)$ be the joint PMF of X and Y. By 2D LOTUS,

$$E(X + Y) = \sum_{x,y}(x + y)p(x, y) = \sum_{x,y} xp(x, y) + \sum_{x,y} yp(x, y) = E(X) + E(Y).$$

(b) Now suppose that X and Y are independent, with joint PMF $p(x, y)$ and marginal PMFs $p_X(x)$ and $p_Y(y)$, respectively. By 2D LOTUS,

$$E(XY) = \sum_{x,y} xyp(x, y)$$

$$= \sum_x \sum_y xyp_X(x)p_Y(y)$$

$$= \sum_x \left(xp_X(x) \sum_y yp_Y(y) \right)$$

$$= \left(\sum_x xp_X(x) \right) \left(\sum_y yp_Y(y) \right)$$

$$= E(X)E(Y).$$

37. Let X and Y be i.i.d. continuous random variables with PDF f, mean μ, and variance σ^2. We know that the expected squared distance of X from its mean is σ^2, and likewise for Y; this problem is about the expected squared distance of X from Y.

(a) Use 2D LOTUS to express $E(X - Y)^2$ as a double integral.

(b) By expanding $(x - y)^2 = x^2 - 2xy + y^2$ and evaluating the double integral from (a), show that
$$E(X - Y)^2 = 2\sigma^2.$$

(c) Give an alternative proof of the result from (b), based on the trick of adding and subtracting μ:
$$(X - Y)^2 = (X - \mu + \mu - Y)^2 = (X - \mu)^2 - 2(X - \mu)(Y - \mu) + (Y - \mu)^2.$$

Solution:

(a) By 2D LOTUS,

$$E(X - Y)^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - y)^2 f(x) f(y) dx dy.$$

(b) Again by 2D LOTUS,

$$E(X - Y)^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x^2 - 2xy + y^2) f(x) f(y) dx dy$$
$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^2 f(x) f(y) dx dy - \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} 2xy f(x) f(y) dx dy + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y^2 f(x) f(y) dx dy$$
$$= E(X^2) - 2E(XY) + E(Y^2).$$

But since X and Y are i.i.d., $E(X^2) = E(Y^2)$ and $E(XY) = E(X)E(Y) = (E(X))^2$, so

$$E(X - Y)^2 = E(X^2) - 2E(XY) + E(Y^2) = 2E(X^2) - 2(E(X))^2 = 2\sigma^2.$$

(c) Using the identity from the hint and the definitions of variance and covariance,

$$E(X - Y)^2 = E(X - \mu)^2 - 2E(X - \mu)(Y - \mu) + E(Y - \mu)^2 = \sigma^2 - 2 \cdot 0 + \sigma^2 = 2\sigma^2.$$

Covariance

38. Ⓢ Let X and Y be r.v.s. Is it correct to say "$\max(X, Y) + \min(X, Y) = X + Y$"? Is it correct to say "$\text{Cov}(\max(X, Y), \min(X, Y)) = \text{Cov}(X, Y)$ since either the max is X and the min is Y or vice versa, and covariance is symmetric"? Explain.

Solution: The identity $\max(x, y) + \min(x, y) = x + y$ is true for all numbers x and y. The random variable $M = \max(X, Y)$ is *defined* by $M(s) = \max(X(s), Y(s))$; this just says to perform the random experiment, observe the numerical values of X and Y, and take their maximum. It follows that

$$\max(X, Y) + \min(X, Y) = X + Y$$

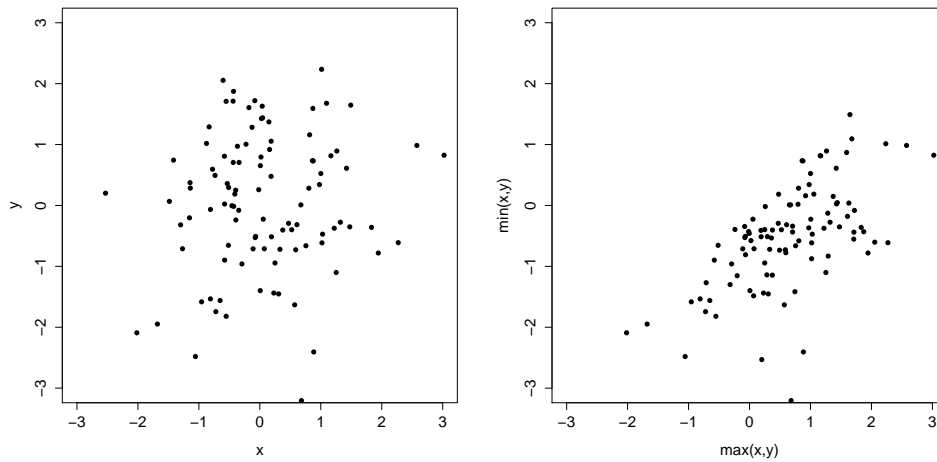for all r.v.s X and Y, since whatever the outcome s of the random experiment is, we have
$$\max(X(s), Y(s)) + \min(X(s), Y(s)) = X(s) + Y(s).$$

In contrast, the covariance of two r.v.s is a number, not an r.v.; it is *not* defined by observing the values of the two r.v.s and then taking their covariance (that would be

a useless quantity, since the covariance between two numbers is 0). It is wrong to say "$\text{Cov}(\max(X,Y), \min(X,Y)) = \text{Cov}(X,Y)$ since either the max is X and the min is Y or vice versa, and covariance is symmetric" since the r.v. X does not equal the r.v. $\max(X,Y)$, nor does it equal the r.v. $\min(X,Y)$.

To gain more intuition into this, consider a repeated sampling interpretation, where we independently repeat the same experiment many times and observe pairs $(x_1, y_1), \ldots, (x_n, y_n)$, where (x_j, y_j) is the observed value of (X,Y) for the $j$the experiment. Suppose that X and Y are independent non-constant r.v.s (and thus they are uncorrelated). Imagine a *scatter plot* of the observations (which is just a plot of the points (x_j, y_j)). Since X and Y are independent, there should be no trend in the plot.

On the other hand, imagine a scatter plot of the $(\max(x_j, y_j), \min(x_j, y_j))$ points. Here we'd expect to see a clear increasing trend (since the max is always bigger than or equal to the min, so having a large value of the min (relative to its mean) should make it more likely that we'll have a large value of the max (relative to its mean). So it makes sense that $\max(X,Y)$ and $\min(X,Y)$ should be positive correlated. This is illustrated in the plots below, in which we generated $(X_1, Y_1), \ldots, (X_{100}, Y_{100})$ with the X_i's and Y_j's i.i.d. $\mathcal{N}(0,1)$.

The simulation was done in R, using the following code:

```
x <- rnorm(100); y <- rnorm(100)
plot(x,y, xlim=c(-3,3),ylim=c(-3,3), pch=16)
plot(pmax(x,y),pmin(x,y), xlim=c(-3,3),ylim=c(-3,3), xlab="max(x,y)",
ylab = "min(x,y)", pch=16)
```

39. Ⓢ Two fair six-sided dice are rolled (one green and one orange), with outcomes X and Y respectively for the green and the orange.

(a) Compute the covariance of $X + Y$ and $X - Y$.

(b) Are $X + Y$ and $X - Y$ independent?

Solution:

(a) We have

$$\text{Cov}(X + Y, X - Y) = \text{Cov}(X, X) - \text{Cov}(X, Y) + \text{Cov}(Y, X) - \text{Cov}(Y, Y) = 0.$$

(b) They are not independent: information about $X + Y$ may give information about $X - Y$, as shown by considering an *extreme example*. Note that if $X + Y = 12$, then $X = Y = 6$, so $X - Y = 0$. Therefore, $P(X - Y = 0 | X + Y = 12) = 1 \neq P(X - Y = 0)$, which shows that $X + Y$ and $X - Y$ are not independent. Alternatively, note that $X + Y$ and $X - Y$ are both even or both odd, since the sum $(X + Y) + (X - Y) = 2X$ is even.

40. Let X and Y be i.i.d. Unif$(0, 1)$.

    (a) Compute the covariance of $X + Y$ and $X - Y$.

    (b) Are $X + Y$ and $X - Y$ independent?

    Solution:

    (a) Using the properties of covariance,

    $$\text{Cov}(X + Y, X - Y) = \text{Var}(X) - \text{Cov}(X, Y) + \text{Cov}(X, Y) - \text{Var}(Y) = 0,$$

    since $\text{Var}(X) = \text{Var}(Y)$ (as X and Y have the same distribution) and $\text{Cov}(X, Y) = 0$ (as X and Y are independent, but even if they were dependent the $-\text{Cov}(X, Y)$ term would cancel out the $\text{Cov}(X, Y)$ term).

    (b) No, $X + Y$ and $X - Y$ are dependent. To see this, note that if $X + Y$ is very close to 0, then both X and Y are very close to 0, so $X - Y$ is very close to 0. So $X + Y$ can provide information about $X - Y$.

41. Ⓢ Let X and Y be standardized r.v.s (i.e., marginally they each have mean 0 and variance 1) with correlation $\rho \in (-1, 1)$. Find a, b, c, d (in terms of ρ) such that $Z = aX + bY$ and $W = cX + dY$ are uncorrelated but still standardized.

    Solution: Let us look for a solution with $Z = X$, finding c and d to make Z and W uncorrelated:

    $$\text{Cov}(Z, W) = \text{Cov}(X, cX + dY) = \text{Cov}(X, cX) + \text{Cov}(X, dY) = c + d\rho = 0.$$

    Also, $\text{Var}(W) = c^2 + d^2 + 2cd\rho = 1$. Solving for c, d gives

    $$a = 1, b = 0, c = -\rho/\sqrt{1 - \rho^2}, d = 1/\sqrt{1 - \rho^2}.$$

42. Ⓢ Let X be the number of distinct birthdays in a group of 110 people (i.e., the number of days in a year such that at least one person in the group has that birthday). Under the usual assumptions (no February 29, all the other 365 days of the year are equally likely, and the day when one person is born is independent of the days when the other people are born), find the mean and variance of X.

    Solution: Let I_j be the indicator r.v. for the event that at least one of the people was born on the jth day of the year, so $X = \sum_{j=1}^{365} I_j$ with $I_j \sim \text{Bern}(p)$, where $p = 1 - (364/365)^{110}$. The I_j's are dependent but by linearity, we still have

    $$E(X) = 365p \approx 95.083.$$

    By symmetry, the variance is

    $$\text{Var}(X) = 365\text{Var}(I_1) + 2\binom{365}{2}\text{Cov}(I_1, I_2).$$

    To get the covariance, note that $\text{Cov}(I_1, I_2) = E(I_1 I_2) - E(I_1)E(I_2) = E(I_1 I_2) - p^2$, and $E(I_1 I_2) = P(I_1 I_2 = 1) = P(A_1 \cap A_2)$, where A_j is the event that at least one person was born on the jth day of the year. The probability of the complement is

    $$P(A_1^c \cup A_2^c) = P(A_1^c) + P(A_2^c) - P(A_1^c \cap A_2^c) = 2\left(\frac{364}{365}\right)^{110} - \left(\frac{363}{365}\right)^{110},$$

    so $\text{Var}(X) = 365p(1 - p) + 365 \cdot 364 \cdot (1 - (2\left(\frac{364}{365}\right)^{110} - \left(\frac{363}{365}\right)^{110}) - p^2) \approx 10.019$.

43. (a) Let X and Y be Bernoulli r.v.s, possibly with different parameters. Show that if X and Y are uncorrelated, then they are independent.

(b) Give an example of three Bernoulli r.v.s such that each pair of them is uncorrelated, yet the three r.v.s are dependent.

Solution:

(a) Let $X \sim \text{Bern}(p_1)$ and $Y \sim \text{Bern}(p_2)$ be uncorrelated. By the fundamental bridge,

$$P(X = 1, Y = 1) = E(XY) = E(X)E(Y) = P(X = 1)P(Y = 1).$$

Since

$$P(X = 1, Y = 1) + P(X = 1, Y = 0) = P(X = 1),$$

we have

$$P(X = 1, Y = 0) = P(X = 1) - P(X = 1, Y = 1) = p_1 - p_1 p_2 = p_1(1 - p_2) = P(X = 1)P(Y = 0).$$

By the same argument (swapping the roles of X and Y),

$$P(X = 0, Y = 1) = (1 - p_1)p_2 = P(X = 0)P(Y = 1).$$

Then

$$P(X = 0, Y = 0) = P(X = 0) - P(X = 0, Y = 1) = (1 - p_1)(1 - p_2) = P(X = 0)P(Y = 0).$$

Therefore, X and Y are independent.

(b) Pairwise independence of events doesn't imply independence of the events, as shown in Example 2.5.5. By taking the indicator r.v.s for events that are pairwise independent but not independent, we can construct r.v.s that are pairwise independent (and hence uncorrelated) but not independent.

As in Example 2.5.5, consider two fair, independent coin tosses, and let A be the event that the first toss is Heads, B be the event that the second toss is Heads, and C be the event that the tosses have the same result. Then the corresponding indicator r.v.s $I(A), I(B), I(C)$ are uncorrelated but dependent.

44. Find the variance of the number of toys needed until you have a complete set in Example 4.3.12 (the coupon collector problem), as a sum.

Solution: Using notation as in Example 4.3.11, write the number of toys needed as $N = N_1 + N_2 + \cdots + N_n$, where $N_j \sim \text{FS}((n - j + 1)/n)$ is the additional number of toys until the jth new toy type is acquired. The N_j are independent, so

$$\text{Var}(N) = \sum_{j=2}^{n} \frac{\frac{j-1}{n}}{\frac{(n-j+1)^2}{n^2}} = n \sum_{j=2}^{n} \frac{j-1}{(n-j+1)^2} = n \sum_{k=1}^{n-1} \frac{n-k}{k^2} = n^2 \sum_{k=1}^{n-1} \frac{1}{k^2} - n \sum_{k=1}^{n-1} \frac{1}{k}.$$

45. A random triangle is formed in some way, such that all pairs of angles have the same joint distribution. What is the correlation between two of the angles (assuming that the variance of the angles is nonzero)?

Solution: Let X, Y, Z be the angles (measured in radians), with $X + Y + Z = \pi$ and X, Y, Z identically distributed. Let $v = \text{Var}(X)$. Then

$$0 = \text{Var}(\pi) = \text{Var}(X + Y + Z) = 3\text{Var}(X) + 2 \cdot \binom{3}{2} \text{Cov}(X, Y),$$

so $\text{Cov}(X, Y) = -v/2$. Therefore,

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\text{SD}(X)\text{SD}(Y)} = \frac{-v/2}{v} = -\frac{1}{2}.$$

46. Each of $n \geq 2$ people puts their name on a slip of paper (no two have the same name). The slips of paper are shuffled in a hat, and then each person draws one (uniformly at random at each stage, without replacement). Find the standard deviation of the number of people who draw their own names.

   Solution: Label the people as $1, 2, \ldots, n$, let I_j be the indicator of person j getting their own name, and let $X = I_1 + \cdots + I_n$. By symmetry and linearity,

   $$E(X) = nE(I_1) = n \cdot \frac{1}{n} = 1.$$

   (This was also shown in the solution to Exercise 4.38.) To find the variance of X, we can expand the variances in terms of covariances:

   $$\begin{aligned}
   \text{Var}(X) &= n\text{Var}(I_1) + 2\binom{n}{2}\text{Cov}(I_1, I_2) \\
   &= \frac{n}{n}\left(1 - \frac{1}{n}\right) + n(n-1)(E(I_1 I_2) - E(I_1)E(I_2)) \\
   &= 1 - \frac{1}{n} + n(n-1)\left(\frac{1}{n(n-1)} - \frac{1}{n^2}\right) \\
   &= 1 - \frac{1}{n} + 1 - \frac{n-1}{n} \\
   &= 1.
   \end{aligned}$$

   Thus, the mean and standard deviation of X are both 1.

47. As in Example 4.4.7, an urn contains w white balls and b black balls. The balls are randomly drawn one by one *without replacement* until r white balls have been drawn. Let $X \sim \text{NHGeom}(w, b, r)$ be the number of black balls drawn before drawing the rth white ball. In this exercise, you will derive $\text{Var}(X)$.

   As explained in Example 4.4.7, we can assume that we continue drawing balls until the urn has been emptied out. Label the black balls as $1, 2, \ldots, b$, and write $X = \sum_{j=1}^{b} I_j$, where I_j is the indicator of black ball j being drawn before the rth white ball is drawn.

   (a) Show that
   $$E(I_j) = \frac{r}{w+1}.$$

   (b) Give an intuitive explanation of whether I_i and I_j are positively correlated, uncorrelated, or negatively correlated, for $i \neq j$.

   (c) Show that for $i \neq j$,
   $$E(I_i I_j) = \frac{\binom{r+1}{2}}{\binom{w+2}{2}} = \frac{(r+1)r}{(w+2)(w+1)}.$$

   Hint: Imagine $w + 2$ slots, into which black balls i and j and the w white balls will be placed. All orderings for these $w + 2$ balls are equally likely.

   (d) Find an expression for $\text{Var}(X)$. With some algebra (which you don't have to do), your expression should simplify to
   $$\text{Var}(X) = \frac{rb(w+b+1)(w-r+1)}{(w+1)^2(w+2)}.$$

   Solution:

   (a) For each fixed j, consider the ordering of black ball j and all the white balls. By symmetry, all $(w + 1)!$ orderings of these $w + 1$ balls are equally likely. So black ball j

is equally likely to be in any position among these $w + 1$ balls. It will be drawn before the rth white ball if and only if its position is in the set $\{1, 2, \ldots, r\}$. Therefore,

$$E(I_j) = \frac{r}{w + 1}.$$

(b) Intuitively, I_i and I_j should be positive correlated since $I_i = 1$ is evidence that there aren't a lot of white balls early on (in terms of the order in which they will be drawn), which in turn is evidence in favor of $I_j = 1$.

(c) As in the hint, consider forming an ordering of black balls i, j and the w white balls. There are $\binom{w+2}{2}$ choices for where to put black balls i, j, all of which are equally likely. Of these choices, there are $\binom{r+1}{2}$ choices for which both of black balls i and j precede the rth white ball, since such a choice is given by choosing an ordering of black balls i and j and the first $r - 1$ white balls. Thus,

$$E(I_i I_j) = \frac{\binom{r+1}{2}}{\binom{w+2}{2}} = \frac{(r + 1)r}{(w + 2)(w + 1)}.$$

(d) We have

$$
\begin{aligned}
\mathrm{Var}(X) &= \mathrm{Var}\left(\sum_{j=1}^{b} I_j\right) \\
&= b\mathrm{Var}(I_1) + b \cdot (b - 1)\mathrm{Cov}(I_1, I_2) \\
&= b \cdot \frac{r}{w + 1} \cdot \left(1 - \frac{r}{w + 1}\right) + b \cdot (b - 1)\left(\frac{(r + 1)r}{(w + 2)(w + 1)} - \left(\frac{r}{w + 1}\right)^2\right) \\
&= \frac{rb(w + b + 1)(w - r + 1)}{(w + 1)^2(w + 2)}.
\end{aligned}
$$

48. Ⓢ Athletes compete one at a time at the high jump. Let X_j be how high the jth jumper jumped, with $X_1, X_2, \ldots$ i.i.d. with a continuous distribution. We say that the jth jumper sets a *record* if X_j is greater than all of $X_{j-1}, \ldots, X_1$.

Find the variance of the number of records among the first n jumpers (as a sum). What happens to the variance as $n \to \infty$?

Solution: Let I_j be the indicator r.v. for the jth jumper setting a record. By symmetry, $E(I_j) = P(I_j = 1) = 1/j$ (as all of the first j jumps are equally likely to be the largest of those jumps). It was shown on Example 5.7.3 that I_{110} and I_{111} are independent. Similarly, I_i is independent of I_j for all i, j with $i < j$ (in fact, they are independent, not just pairwise independent). To see this, note that by symmetry, learning that the jth jumper sets a record gives no information whatsoever about how the first i jumpers rank among themselves, or compute

$$P(I_i = I_j = 1) = \frac{\binom{j-1}{j-i-1}(j - i - 1)!(i - 1)!}{j!} = \frac{(i - 1)!(j - 1)!}{i!j!} = \frac{1}{ij} = P(I_1 = 1)P(I_2 = 1),$$

where the numerator corresponds to putting the best of the first j jumps in position j, picking any $j - 1 + 1$ of the remaining jumps to fill positions $i + 1$ through $j - 1$ and putting them in any order, putting the best of the remaining i jumps in position i, and then putting the remaining $i - 1$ jumps in any order.

The variance of I_j is $\mathrm{Var}(I_j) = E(I_j^2) - (EI_j)^2 = \frac{1}{j} - \frac{1}{j^2}$. Since the I_j are pairwise independent (and thus uncorrelated), the variance of $I_1 + \cdots + I_n$ is

$$\sum_{j=1}^{n}\left(\frac{1}{j} - \frac{1}{j^2}\right),$$

which goes to ∞ as $n \to \infty$ since $\sum_{j=1}^{n} \frac{1}{j}$ diverges and $\sum_{j=1}^{n} \frac{1}{j^2}$ converges (to $\pi^2/6$, as it turns out).

49. Ⓢ A chicken lays a $\text{Pois}(\lambda)$ number N of eggs. Each egg hatches a chick with probability p, independently. Let X be the number which hatch, so $X|N = n \sim \text{Bin}(n, p)$.

Find the correlation between N (the number of eggs) and X (the number of eggs which hatch). Simplify; your final answer should work out to a simple function of p (the λ should cancel out).

Solution: By the chicken-egg story, X is independent of Y, with $X \sim \text{Pois}(\lambda p)$ and $Y \sim \text{Pois}(\lambda q)$, for $q = 1 - p$. So

$$\text{Cov}(N, X) = \text{Cov}(X + Y, X) = \text{Cov}(X, X) + \text{Cov}(Y, X) = \text{Var}(X) = \lambda p,$$

giving

$$\text{Corr}(N, X) = \frac{\lambda p}{SD(N)SD(X)} = \frac{\lambda p}{\sqrt{\lambda \lambda p}} = \sqrt{p}.$$

50. Let $X_1, \ldots, X_n$ be random variables such that $\text{Corr}(X_i, X_j) = \rho$ for all $i \neq j$. Show that $\rho \geq -\frac{1}{n-1}$. This is a bound on how negatively correlated a collection of r.v.s can all be with each other.

Hint: Assume $\text{Var}(X_i) = 1$ for all i; this can be done without loss of generality, since rescaling two r.v.s does not affect the correlation between them. Then use the fact that $\text{Var}(X_1 + \cdots + X_n) \geq 0$.

Solution: As mentioned in the hint, we can assume $\text{Var}(X_i) = 1$ without loss of generality; correlation is based on the *standardized* versions of r.v.s. Then

$$0 \leq \text{Var}(X_1 + \cdots + X_n) = n\text{Var}(X_1) + n(n-1)\text{Cov}(X_1, X_2) = n + n(n-1)\rho,$$

so

$$\rho \geq -\frac{1}{n-1}.$$

51. Let X and Y be independent r.v.s. Show that

$$\text{Var}(XY) = \text{Var}(X)\text{Var}(Y) + (EX)^2\text{Var}(Y) + (EY)^2\text{Var}(X).$$

Hint: It is often useful when working with a second moment $E(T^2)$ to write it as $\text{Var}(T) + (ET)^2$.

Solution: Note that X and Y are uncorrelated (since they are independent) and X^2 and Y^2 are uncorrelated (since they are independent). Then

$$\begin{aligned}
\text{Var}(XY) &= E(X^2Y^2) - (E(XY))^2 \\
&= E(X^2)E(Y^2) - (EX)^2(EY)^2 \\
&= (\text{Var}(X) + (EX)^2)(\text{Var}(Y) + (EY)^2) - (EX)^2(EY)^2 \\
&= \text{Var}(X)\text{Var}(Y) + (EX)^2\text{Var}(Y) + (EY)^2\text{Var}(X).
\end{aligned}$$

52. Stat 110 shirts come in 3 sizes: small, medium, and large. There are n shirts of each size (where $n \geq 2$). There are $3n$ students. For each size, n of the students have that size as the best fit. This seems ideal. But suppose that instead of giving each student the right size shirt, each student is given a shirt completely randomly (all allocations of the shirts to the students, with one shirt per student, are equally likely). Let X be the number of students who get their right size shirt.

(a) Find $E(X)$.

(b) Give each student an ID number from 1 to $3n$, such that the right size shirt is small for students 1 through n, medium for students $n + 1$ through $2n$, and large for students $2n + 1$ through $3n$. Let A_j be the event that student j gets their right size shirt. Find $P(A_1, A_2)$ and $P(A_1, A_{n+1})$.

(c) Find $\text{Var}(X)$.

Solution:

(a) Create an indicator r.v. for each student, equal to 1 if that student gets the right size shirt and 0 otherwise. By linearity, $E(X) = (3n)/3 = n$.

(b) By the multiplication rule,

$$P(A_1, A_2) = \frac{n(n-1)}{(3n)(3n-1)} = \frac{n-1}{3(3n-1)},$$

$$P(A_1, A_{n+1}) = \frac{n^2}{(3n)(3n-1)} = \frac{n}{3(3n-1)}.$$

(c) Letting I_j be the indicator for student j getting the right size shirt, we have

$$\text{Var}(X) = 3n\text{Var}(I_1) + 2\sum_{i<j} \text{Cov}(I_i, I_j).$$

By the fundamental bridge, $\text{Cov}(I_i, I_j) = P(A_i, A_j) - P(A_i)P(A_j)$. There are $3\binom{n}{2}$ pairs of students where both students have the same right shirt size, and $3n^2$ pairs of students where the two students have different right shirt sizes. So

$$\text{Var}(X) = \frac{2n}{3} + 3n(n-1)\left(\frac{n-1}{3(3n-1)} - \frac{1}{9}\right) + 6n^2\left(\frac{n}{3(3n-1)} - \frac{1}{9}\right) = \frac{2n^2}{3n-1}.$$

53. ⓢ A drunken man wanders around randomly in a large space. At each step, he moves one unit of distance North, South, East, or West, with equal probabilities. Choose coordinates such that his initial position is $(0,0)$ and if he is at (x,y) at some time, then one step later he is at $(x, y+1), (x, y-1), (x+1, y)$, or $(x-1, y)$. Let (X_n, Y_n) and R_n be his position and distance from the origin after n steps, respectively.

General hint: Note that X_n is a sum of r.v.s with possible values $-1, 0, 1$, and likewise for Y_n, but be careful throughout the problem about independence.

(a) Determine whether or not X_n is independent of Y_n.
(b) Find $\text{Cov}(X_n, Y_n)$.
(c) Find $E(R_n^2)$.

Solution:

(a) They are *not* independent, as seen by considering an *extreme case* such as the event that the drunk headed East for the entire time: note that $P(Y_n = 0|X_n = n) = 1$.

(b) Write $X_n = \sum_{i=1}^n Z_i$ and $Y_n = \sum_{j=1}^n W_j$, where Z_i is -1 if his ith step is Westward, 1 if his ith step is Eastward, and 0 otherwise, and similarly for W_j. Then Z_i is independent of W_j for $i \neq j$. But Z_i and W_i are highly dependent: exactly one of them is 0 since he moves in one direction at a time. Then $\text{Cov}(Z_i, W_i) = E(Z_i W_i) - E(Z_i)E(W_i) = 0$ since $Z_i W_i$ is always 0, and Z_i and W_i have mean 0. So

$$\text{Cov}(X_n, Y_n) = \sum_{i,j} \text{Cov}(Z_i, W_j) = 0.$$

(c) We have $R_n^2 = X_n^2 + Y_n^2$, and $E(Z_i Z_j) = 0$ for $i \neq j$. So

$$E(R_n^2) = E(X_n^2) + E(Y_n^2) = 2E(X_n^2) = 2nE(Z_1^2) = n,$$

since $Z_1^2 \sim \text{Bern}(1/2)$.

54. ⓢ A scientist makes two measurements, considered to be independent standard Normal r.v.s. Find the correlation between the larger and smaller of the values.

Hint: Note that $\max(x, y) + \min(x, y) = x + y$ and $\max(x, y) - \min(x, y) = |x - y|$.

Solution: Let X and Y be i.i.d $\mathcal{N}(0, 1)$ and $M = \max(X, Y), L = \min(X, Y)$. By the hint,
$$E(M) + E(L) = E(M + L) = E(X + Y) = E(X) + E(Y) = 0,$$
$$E(M) - E(L) = E(M - L) = E|X - Y| = \frac{2}{\sqrt{\pi}},$$
where the last equality was shown in Example 7.2.3. So $E(M) = 1/\sqrt{\pi}$, and
$$\text{Cov}(M, L) = E(ML) - E(M)E(L) = E(XY) + (EM)^2 = (EM)^2 = \frac{1}{\pi},$$
since $ML = XY$ has mean $E(XY) = E(X)E(Y) = 0$. To obtain the correlation, we also need $\text{Var}(M)$ and $\text{Var}(L)$. By symmetry of the Normal, $(-X, -Y)$ has the same distribution as (X, Y), so $\text{Var}(M) = \text{Var}(L)$; call this v. Then
$$E(X - Y)^2 = \text{Var}(X - Y) = 2, \text{ and also}$$
$$E(X - Y)^2 = E(M - L)^2 = EM^2 + EL^2 - 2E(ML) = 2v + \frac{2}{\pi}.$$
So $v = 1 - \frac{1}{\pi}$ (alternatively, we can get this by taking the variance of both sides of $\max(X, Y) + \min(X, Y) = X + Y$). Thus,
$$\text{Corr}(M, L) = \frac{\text{Cov}(M, L)}{\sqrt{\text{Var}(M)\text{Var}(L)}} = \frac{1/\pi}{1 - 1/\pi} = \frac{1}{\pi - 1}.$$

55. Let $U \sim \text{Unif}(-1, 1)$ and $V = 2|U| - 1$.

(a) Find the distribution of V (give the PDF and, if it is a named distribution we have studied, its name and parameters).

Hint: Find the support of V, and then find the CDF of V by reducing $P(V \le v)$ to probability calculations about U.

(b) Show that U and V are uncorrelated, but not independent. This is also another example illustrating the fact that knowing the marginal distributions of two r.v.s does not determine the joint distribution.

Solution:

(a) The support of V is $(-1, 1)$. The CDF of V is
$$P(V \le v) = P\left(|U| \le \frac{v + 1}{2}\right) = P\left(-\frac{v + 1}{2} \le U \le \frac{v + 1}{2}\right) = \frac{v + 1}{2}$$
for $-1 < v < 1$, since for a Uniform probability is proportional to length. The PDF of V is $f(v) = 1/2$ for $v \in (-1, 1)$ (and 0 otherwise). Therefore, $V \sim \text{Unif}(-1, 1)$.

(b) By (a), U and V are both $\text{Unif}(-1, 1)$, with mean 0. They are uncorrelated since
$$\text{Cov}(U, V) = E(UV) = 2E(U|U|) = \int_{-1}^{1} u|u|du = -\int_{-1}^{0} u^2 du + \int_{0}^{1} u^2 du = 0.$$

But they are extremely dependent, since in fact V is a deterministic function of U.

56. Ⓢ Consider the following method for creating a *bivariate Poisson* (a joint distribution for two r.v.s such that both marginals are Poissons). Let $X = V + W, Y = V + Z$ where V, W, Z are i.i.d. Pois(λ) (the idea is to have something borrowed and something new but not something old or something blue).

    (a) Find Cov(X, Y).

    (b) Are X and Y independent? Are they conditionally independent given V?

    (c) Find the joint PMF of X, Y (as a sum).

    Solution:

    (a) Using the properties of covariance, we have

    $$\text{Cov}(X, Y) = \text{Cov}(V, V) + \text{Cov}(V, Z) + \text{Cov}(W, V) + \text{Cov}(W, Z) = \text{Var}(V) = \lambda.$$

    (b) Since X and Y are correlated (with covariance $\lambda > 0$), they are not independent. Alternatively, note that $E(Y) = 2\lambda$ but $E(Y|X = 0) = \lambda$ since if $X = 0$ occurs then $V = 0$ occurs. But X and Y are conditionally independent given V, since the conditional joint PMF is

    $$\begin{aligned} P(X = x, Y = y|V = v) &= P(W = x - v, Z = y - v|V = v) \\ &= P(W = x - v, Z = y - v) \\ &= P(W = x - v)P(Z = y - v) \\ &= P(X = x|V = v)P(Y = y|V = v). \end{aligned}$$

    This makes sense intuitively since if we observe that $V = v$, then X and Y are the independent r.v.s W and Z, shifted by the constant v.

    (c) By (b), a good strategy is to condition on V:

    $$\begin{aligned} P(X = x, Y = y) &= \sum_{v=0}^{\infty} P(X = x, Y = y|V = v)P(V = v) \\ &= \sum_{v=0}^{\min(x,y)} P(X = x|V = v)P(Y = y|V = v)P(V = v) \\ &= \sum_{v=0}^{\min(x,y)} e^{-\lambda}\frac{\lambda^{x-v}}{(x-v)!}e^{-\lambda}\frac{\lambda^{y-v}}{(y-v)!}e^{-\lambda}\frac{\lambda^{v}}{v!} \\ &= e^{-3\lambda}\lambda^{x+y}\sum_{v=0}^{\min(x,y)} \frac{\lambda^{-v}}{(x-v)!(y-v)!v!}, \end{aligned}$$

    for x and y nonnegative integers. Note that we sum only up to $\min(x, y)$ since we know for sure that $V \leq X$ and $V \leq Y$.

    Sanity check: Note that $P(X = 0, Y = 0) = P(V = 0, W = 0, Z = 0) = e^{-3\lambda}$.

57. You are playing an exciting game of Battleship. Your opponent secretly positions ships on a 10 by 10 grid and you try to guess where the ships are. Each of your guesses is a *hit* if there is a ship there and a *miss* otherwise.

    The game has just started and your opponent has 3 ships: a battleship (length 4), a submarine (length 3), and a destroyer (length 2). (Usually there are 5 ships to start, but to simplify the calculations we are considering 3 here.) You are playing a variation in which you unleash a *salvo*, making 5 simultaneous guesses. Assume that your 5 guesses are a simple random sample drawn from the 100 grid positions.

    Find the mean and variance of the number of distinct ships you will hit in your salvo.

(Give exact answers in terms of binomial coefficients or factorials, and also numerical values computed using a computer.)

Hint: First work in terms of the number of ships *missed*, expressing this as a sum of indicator r.v.s. Then use the fundamental bridge and naive definition of probability, which can be applied since all sets of 5 grid positions are equally likely.

Solution: Let N be the number of ships hit and M be the number of ships missed. Then $N + M = 3$, $E(N) = 3 - E(M)$, and $\text{Var}(N) = \text{Var}(3 - M) = \text{Var}(M)$. Write

$$M = I_2 + I_3 + I_4,$$

where I_j is the indicator of missing the ship of length j for $j \in \{2, 3, 4\}$. Then for $j, k \in \{2, 3, 4\}$ with $j \neq k$, by the fundamental bridge we have

$$E(I_j) = \frac{\binom{100-j}{5}}{\binom{100}{5}}, E(I_j I_k) = \frac{\binom{100-j-k}{5}}{\binom{100}{5}}.$$

Thus,

$$E(M) = E(I_2) + E(I_3) + E(I_4) = \frac{\binom{98}{5}}{\binom{100}{5}} + \frac{\binom{97}{5}}{\binom{100}{5}} + \frac{\binom{96}{5}}{\binom{100}{5}} \approx 2.57,$$

which gives

$$E(N) = 3 - E(M) \approx 0.43.$$

For the variance, we have

$$E(M^2) = E(I_2^2) + E(I_3^2) + E(I_4^2) + 2E(I_2 I_3) + 2E(I_2 I_4) + 2E(I_3 I_4)$$

$$= \frac{\binom{98}{5}}{\binom{100}{5}} + \frac{\binom{97}{5}}{\binom{100}{5}} + \frac{\binom{96}{5}}{\binom{100}{5}} + 2\left(\frac{\binom{95}{5}}{\binom{100}{5}} + \frac{\binom{94}{5}}{\binom{100}{5}} + \frac{\binom{93}{5}}{\binom{100}{5}}\right),$$

so $\text{Var}(N) = \text{Var}(M) = E(M^2) - (E(M))^2$, with $E(M^2)$ and $E(M)$ as above. This evaluates to $\text{Var}(N) \approx 0.34$.

58. This problem explores a visual interpretation of covariance. Data are collected for $n \geq 2$ individuals, where for each individual two variables are measured (e.g., height and weight). Assume independence *across* individuals (e.g., person 1's variables gives no information about the other people), but not *within* individuals (e.g., a person's height and weight may be correlated).

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be the n data points. The data are considered here as fixed, known numbers–they are the observed values after performing an experiment. Imagine plotting all the points (x_i, y_i) in the plane, and drawing the rectangle determined by each pair of points. For example, the points $(1, 3)$ and $(4, 6)$ determine the rectangle with vertices $(1, 3), (1, 6), (4, 6), (4, 3)$.

The *signed area* contributed by (x_i, y_i) and (x_j, y_j) is the area of the rectangle they determine if the slope of the line between them is positive, and is the negative of the area of the rectangle they determine if the slope of the line between them is negative. (Define the signed area to be 0 if $x_i = x_j$ or $y_i = y_j$, since then the rectangle is degenerate.) So the signed area is positive if a higher x value goes with a higher y value for the pair of points, and negative otherwise. Assume that the x_i are all distinct and the y_i are all distinct.

(a) The *sample covariance* of the data is defined to be

$$r = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}),$$

where

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \text{ and } \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

are the sample means. (There are differing conventions about whether to divide by $n-1$ or n in the definition of sample covariance, but that need not concern us for this problem.)

Let (X, Y) be one of the (x_i, y_i) pairs, chosen uniformly at random. Determine precisely how $\text{Cov}(X, Y)$ is related to the sample covariance.

(b) Let (X, Y) be as in (a), and $(\tilde{X}, \tilde{Y})$ be an independent draw from the same distribution. That is, (X, Y) and $(\tilde{X}, \tilde{Y})$ are randomly chosen from the n points, independently (so it is possible for the same point to be chosen twice).

Express the total signed area of the rectangles as a constant times $E((X - \tilde{X})(Y - \tilde{Y}))$. Then show that the sample covariance of the data is a constant times the total signed area of the rectangles.

Hint: Consider $E((X - \tilde{X})(Y - \tilde{Y}))$ in two ways: as the average signed area of the random rectangle formed by (X, Y) and $(\tilde{X}, \tilde{Y})$, and using properties of expectation to relate it to $\text{Cov}(X, Y)$. For the former, consider the n^2 possibilities for which point (X, Y) is and which point $(\tilde{X}, \tilde{Y})$; note that n such choices result in degenerate rectangles.

(c) Based on the interpretation from (b), give intuitive explanations of why for any r.v.s W_1, W_2, W_3 and constants a_1, a_2, covariance has the following properties:

 (i) $\text{Cov}(W_1, W_2) = \text{Cov}(W_2, W_1)$;
 (ii) $\text{Cov}(a_1 W_1, a_2 W_2) = a_1 a_2 \text{Cov}(W_1, W_2)$;
(iii) $\text{Cov}(W_1 + a_1, W_2 + a_2) = \text{Cov}(W_1, W_2)$;
(iv) $\text{Cov}(W_1, W_2 + W_3) = \text{Cov}(W_1, W_2) + \text{Cov}(W_1, W_3)$.

Solution:

(a) By definition of expectation,

$$E(X) = \bar{x}, E(Y) = \bar{y}.$$

After doing the experiment of choosing randomly among the n points, (X, Y) crystallizes into some (x_i, y_i), at which time $(X - \bar{x})(Y - \bar{y})$ crystallizes to $(x_i - \bar{x})(y_i - \bar{y})$. So

$$\text{Cov}(X, Y) = E((X - \bar{x})(Y - \bar{y})) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = r.$$

(b) The total signed area is

$$A = \sum_{i<j} (x_i - x_j)(y_i - y_j).$$

Now compare this with the average signed area of the *random* rectangle formed by (X, Y) and $(\tilde{X}, \tilde{Y})$, which by linearity is

$$E((X - \tilde{X})(Y - \tilde{Y})) = E(XY) + E(\tilde{X}\tilde{Y}) - E(X\tilde{Y}) - E(\tilde{X}Y).$$

This simplifies to

$$2E(XY) - 2E(X)E(Y) = 2\text{Cov}(X, Y),$$

since $E(\tilde{X}\tilde{Y}) = E(XY)$ (because XY and $\tilde{X}\tilde{Y}$ have the same distribution), $E(X\tilde{Y}) = E(X)E(\tilde{Y}) = E(X)E(Y)$ (because X and $\tilde{Y}$ are independent, and hence uncorrelated), and $E(\tilde{X}Y) = E(X)E(Y)$.

On the other hand, $E((X - \tilde{X})(Y - \tilde{Y}))$ is the arithmetic mean of n^2 values (consider all pairs (i, j) where i is the index for which point (X, Y) is and j is the index for which point $(\tilde{X}, \tilde{Y})$ is), consisting of n 0's (for the degenerate rectangles formed when $(\tilde{X}, \tilde{Y}) = (X, Y)$) and the $\binom{n}{2}$ signed rectangle areas, listed twice each. So

$$E((X - \tilde{X})(Y - \tilde{Y})) = \frac{n \cdot 0 + 2 \sum_{i<j} (x_i - x_j)(y_i - y_j)}{n^2} = \frac{2}{n^2} A.$$

Thus,
$$\text{Cov}(X, Y) = \frac{A}{n^2}.$$

(c) Fundamental properties of covariance follow readily from the above interpretation:

(i) Reversing which axis is which has no effect on the areas of the rectangles.

(ii) Stretching (or shrinking) a rectangle along one axis changes the area by the same factor. For example, if we double all the widths and triple all the lengths of the rectangles, then the areas all increase by a factor of 6.

(iii) Shifting a rectangle horizontally or vertically has no effect on its area.

(iv) A rectangle with dimensions a by $b + c$ can be split into two rectangles, one a by b and the other a by c, and the area of the original rectangle equals the sum of the areas of the smaller rectangles. Here we are dealing with *signed* areas, but in this interpretation a positive-area rectangle splits into two positive-area rectangles and a negative-area rectangle splits into two negative-area rectangles.

59. A statistician is trying to estimate an unknown parameter θ based on some data. She has available two independent estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ (an estimator is a function of the data, used to estimate a parameter). For example, $\hat{\theta}_1$ could be the sample mean of a subset of the data and $\hat{\theta}_2$ could be the sample mean of another subset of the data, disjoint from the subset used to calculate $\hat{\theta}_1$. Assume that both of these estimators are unbiased, i.e., $E(\hat{\theta}_j) = \theta$.

Rather than having a bunch of separate estimators, the statistician wants one combined estimator. It may not make sense to give equal weights to $\hat{\theta}_1$ and $\hat{\theta}_2$ since one could be much more reliable than the other, so she decides to consider combined estimators of the form
$$\hat{\theta} = w_1 \hat{\theta}_1 + w_2 \hat{\theta}_2,$$

a weighted combination of the two estimators. The weights w_1 and w_2 are nonnegative and satisfy $w_1 + w_2 = 1$.

(a) Check that $\hat{\theta}$ is also unbiased, i.e., $E(\hat{\theta}) = \theta$.

(b) Determine the optimal weights w_1, w_2, in the sense of minimizing the mean squared error $E(\hat{\theta} - \theta)^2$. Express your answer in terms of the variances of $\hat{\theta}_1$ and $\hat{\theta}_2$. The optimal weights are known as *Fisher weights*.

Hint: As discussed in Exercise 55 from Chapter 5, mean squared error is variance plus squared bias, so in this case the mean squared error of $\hat{\theta}$ is $\text{Var}(\hat{\theta})$. Note that there is no need for multivariable calculus here, since $w_2 = 1 - w_1$.

(c) Give a simple description of what the estimator found in (b) amounts to if the data are i.i.d. random variables $X_1, \ldots, X_n, Y_1, \ldots, Y_m$, $\hat{\theta}_1$ is the sample mean of $X_1, \ldots, X_n$, and $\hat{\theta}_2$ is the sample mean of $Y_1, \ldots, Y_m$.

Solution:

(a) By linearity,
$$E(\hat{\theta}) = w_1 E(\hat{\theta}_1) + w_2 E(\hat{\theta}_2) = w_1 \theta + w_2 \theta = \theta.$$

(b) Let $v_j = \text{Var}(\hat{\theta}_j)$. Using the result mentioned in the hint,
$$E(\hat{\theta} - \theta)^2 = \text{Var}(\hat{\theta}) = w_1^2 v_1 + w_2^2 v_2 = w_1^2 v_1 + (1 - w_1)^2 v_2.$$

Setting the derivative of the mean squared error equal to 0 gives
$$2 w_1 v_1 - 2(1 - w_1) v_2 = 0,$$

resulting in the weights

$$w_1 = \frac{v_2}{v_1 + v_2}, w_1 = \frac{v_1}{v_1 + v_2}.$$

We have found a maximum since the second derivative of the mean squared error is $2v_1 + 2v_2 > 0$.

Equivalently, we can write

$$w_1 = \frac{\frac{1}{v_1}}{\frac{1}{v_1} + \frac{1}{v_2}}, w_2 = \frac{\frac{1}{v_2}}{\frac{1}{v_1} + \frac{1}{v_2}}.$$

In words, Fisher weighting says that the weight on the jth estimator should be inversely proportional to the variance of that estimator.

(c) Let σ^2 be the variance of each X_i (which is also the variance of each Y_j). Then

$$v_1 = \text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}, v_2 = \text{Var}(\bar{X}_m) = \frac{\sigma^2}{m},$$

so the Fisher weights are

$$w_1 = \frac{n}{n + m}, w_2 = \frac{m}{n + m}.$$

This seems plausible intuitively, since it says to give each sample mean weight proportional to its sample size. The estimator $\hat{\theta}$ using these weights is

$$w_1 \bar{X}_n + w_2 \bar{Y}_m = \frac{\sum_{i=1}^{n} X_i}{n + m} + \frac{\sum_{j=1}^{m} Y_j}{n + m} = \frac{X_1 + \cdots + X_n + Y_1 + \cdots + Y_m}{n + m},$$

which simply says to take the sample mean of all the observations. This estimator seems reasonable intuitively since the $n + m$ observations are i.i.d., so there is no reason (based on the information given in the problem) to weight the observations unequally.

Chicken-egg

60. Ⓢ A $\text{Pois}(\lambda)$ number of people vote in a certain election. Each voter votes for candidate A with probability p and for candidate B with probability $q = 1 - p$, independently of all the other voters. Let V be the difference in votes, defined as the number of votes for A minus the number for B.

(a) Find $E(V)$.

(b) Find $\text{Var}(V)$.

Solution:

(a) Let X and Y be the number of votes for A and B respectively, and let $N = X + Y$. Then $X|N \sim \text{Bin}(N, p)$ and $Y|N \sim \text{Bin}(N, q)$. By the chicken-egg story, $E(X) = \lambda p$ and $E(Y) = \lambda q$. So

$$E(V) = E(X - Y) = E(X) - E(Y) = \lambda(p - q).$$

(b) By the chicken-egg story, $X \sim \text{Pois}(\lambda p)$ and $Y \sim \text{Pois}(\lambda q)$ are independent. So

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) = \lambda p + \lambda q = \lambda.$$

61. A traveler gets lost $N \sim \text{Pois}(\lambda)$ times on a long journey. When lost, the traveler asks someone for directions with probability p. Let X be the number of times that the traveler is lost and asks for directions, and Y be the number of times that the traveler is lost and does not ask for directions.

(a) Find the joint PMF of N, X, Y. Are they independent?

(b) Find the joint PMF of N, X. Are they independent?

(c) Find the joint PMF of X, Y. Are they independent?

Solution:

(a) The deterministic relationship $N = X + Y$ holds, so N, X, Y are extremely dependent. Let $q = 1 - p$. By the chicken-egg story, X and Y are independent, with $X \sim \text{Pois}(\lambda p)$ and $Y \sim \text{Pois}(\lambda q)$. The joint PMF of N, X, Y is

$$P(N = n, X = x, Y = y) = P(X = x, Y = y) = \left(e^{-\lambda p}(\lambda p)^x/x!\right)\left(e^{-\lambda q}(\lambda q)^y/y!\right),$$

for n, x, y nonnegative integers satisfying $n = x + y$.

(b) We know that $N \geq X$, so X and N are dependent. The joint PMF of N, X is

$$P(N = n, X = x) = P(X = x, Y = n - x) = \left(e^{-\lambda p}(\lambda p)^x/x!\right)\left(e^{-\lambda q}(\lambda q)^{n-x}/(n-x)!\right),$$

for n, x nonnegative integers satisfying $n \geq x$.

(c) As in (a), the joint PMF of X, Y is

$$P(X = x, Y = y) = \left(e^{-\lambda p}(\lambda p)^x/x!\right)\left(e^{-\lambda q}(\lambda q)^y/y!\right),$$

for x, y nonnegative integers.

62. The number of people who visit the Leftorium store in a day is $\text{Pois}(100)$. Suppose that 10% of customers are *sinister* (left-handed), and 90% are *dexterous* (right-handed). Half of the sinister customers make purchases, but only a third of the dexterous customers make purchases. The characteristics and behavior of people are independent, with probabilities as described in the previous two sentences. On a certain day, there are 42 people who arrive at the store but leave without making a purchase. Given this information, what is the conditional PMF of the number of customers on that day who make a purchase?

Solution: Let $N \sim \text{Pois}(100)$ be the number of people who visit the store in a day, X be the number who visit and make a purchase, and Y be the number who visit but don't make a purchase. So $X + Y = N$ and $X|N \sim \text{Bin}(N, p)$, where by LOTP

$$p = 0.1 \cdot \frac{1}{2} + 0.9 \cdot \frac{1}{3} = 0.35.$$

By the chicken-egg story, X and Y are independent, with $X \sim \text{Pois}(35), Y \sim \text{Pois}(65)$. So the conditional PMF of X given $Y = 42$ is

$$P(X = i|Y = 42) = P(X = i) = e^{-35} \cdot 35^i/i!,$$

for $i = 0, 1, 2, \ldots$.

63. A chicken lays n eggs. Each egg independently does or doesn't hatch, with probability p of hatching. For each egg that hatches, the chick does or doesn't survive (independently of the other eggs), with probability s of survival. Let $N \sim \text{Bin}(n, p)$ be the number of eggs which hatch, X be the number of chicks which survive, and Y be the number of chicks which hatch but don't survive (so $X + Y = N$). Find the marginal PMF of X, and the joint PMF of X and Y. Are X and Y independent?

Solution: We will give a story proof that $X \sim \text{Bin}(n, ps)$. Consider any one of the n eggs. With probability p, it hatches. Given that it hatches, with probability s the chick

survives. So the probability is ps of the egg hatching a chick which survives. Thus, $X \sim \text{Bin}(n, ps)$, with PMF

$$P(X = k) = \binom{n}{k}(ps)^k(1 - ps)^{n-k},$$

for $k = 0, 1, 2, \ldots, n$.

The joint PMF of X and Y can be found by using LOTP to condition on N. Note that

$$P(X = i, Y = j | N = n) = 0$$

unless $n = i + j$ (all the hatched eggs must be accounted for!). For any nonnegative integers i, j with $i + j \leq n$, we then have

$$\begin{aligned}
P(X = i, Y = j) &= P(X = i, Y = j | N = i + j)P(N = i + j) \\
&= P(X = i | N = i + j)P(N = i + j) \\
&= \binom{i + j}{i}s^i(1 - s)^j \binom{n}{i + j}p^{i+j}(1 - p)^{n-i-j} \\
&= \frac{n!}{i!j!(n - i - j)!}(ps)^i(p(1 - s))^j(1 - p)^{n-i-j}.
\end{aligned}$$

As another way to see this, note that if we let Z be the number of eggs which don't hatch, then (X, Y, Z) is *Multinomial*: each egg falls into exactly 1 of 3 categories: non-hatching, hatching but not surviving, hatching and surviving.

Unlike in the chicken-egg story, X and Y are *not* independent. The constraint that $X + Y \leq n$ makes them dependent, since if we observe $X = x$ then we know $Y \leq n - x$. For example, in the extreme case where it is observed that $X = n$, we know that $Y = 0$.

64. There will be $X \sim \text{Pois}(\lambda)$ courses offered at a certain school next year.

(a) Find the expected number of choices of 4 courses (in terms of λ, fully simplified), assuming that simultaneous enrollment is allowed if there are time conflicts.

(b) Now suppose that simultaneous enrollment is not allowed. Suppose that most faculty only want to teach on Tuesdays and Thursdays, and most students only want to take courses that start at 10 am or later, and as a result there are only four possible time slots: 10 am, 11:30 am, 1 pm, 2:30 pm (each course meets Tuesday-Thursday for an hour and a half, starting at one of these times). Rather than trying to avoid major conflicts, the school schedules the courses completely randomly: after the list of courses for next year is determined, they randomly get assigned to time slots, independently and with probability 1/4 for each time slot.

Let X_{am} and X_{pm} be the number of morning and afternoon courses for next year, respectively (where "morning" means starting before noon). Find the joint PMF of X_{am} and X_{pm}, i.e., find $P(X_{\text{am}} = a, X_{\text{pm}} = b)$ for all a, b.

(c) Continuing as in (b), let X_1, X_2, X_3, X_4 be the number of 10 am, 11:30 am, 1 pm, 2:30 pm courses for next year, respectively. What is the joint distribution of X_1, X_2, X_3, X_4? (The result is completely analogous to that of $X_{\text{am}}, X_{\text{pm}}$; you can derive it by thinking conditionally, but for this part you are also allowed to just use the fact that the result is analogous to that of (b).) Use this to find the expected number of choices of 4 non-conflicting courses (in terms of λ, fully simplified). What is the ratio of the expected value from (a) to this expected value?

Solution:

(a) Let k be the number of courses chosen (the problem is asking about the case $k = 4$, but it is not harder to solve this for general k). By LOTUS and the Taylor series for e^x,

$$E\binom{X}{k} = \sum_{n=k}^{\infty} \binom{n}{k} e^{-\lambda} \frac{\lambda^n}{n!} = \frac{e^{-\lambda}}{k!} \sum_{n=k}^{\infty} \frac{\lambda^n}{(n-k)!} = \frac{e^{-\lambda}\lambda^k}{k!} \sum_{n=k}^{\infty} \frac{\lambda^{n-k}}{(n-k)!} = \frac{\lambda^k}{k!}.$$

So for $k = 4$, there are $\frac{\lambda^4}{24}$ possibilities on average. As a check, note that when $k = 1$ the above reduces to $EX = \lambda$, and when $k = 2$ the result is true since $E(X(X-1)) = E(X^2) - EX = \text{Var}(X) + (EX)^2 - EX = \lambda^2$.

(b) This problem has exactly the structure of the chicken-egg problem, so X_{am} and X_{pm} are independent $\text{Pois}(\lambda/2)$ r.v.s. (Alternatively, condition on X, and use the fact that $P(X_{\text{am}} = a, X_{\text{pm}} = b | X = n) = 0$ if $a + b \neq n$; this is essentially how we solved the chicken-egg problem.) So the joint PMF is

$$P(X_{\text{am}} = a, X_{\text{pm}} = b) = \frac{e^{-\lambda/2}(\lambda/2)^a}{a!} \cdot \frac{e^{-\lambda/2}(\lambda/2)^b}{b!},$$

for all nonnegative integers a and b.

(c) Analogously to (b), X_1, X_2, X_3, X_4 are independent Poisson r.v.s. (A proof is given below, but was not required for this problem.) By symmetry and linearity, $E(X) = E(X_1) + E(X_2) + E(X_3) + E(X_4) = 4E(X_1)$, which gives $E(X_1) = \lambda/4$. So X_1, X_2, X_3, X_4 are independent $\text{Pois}(\lambda/4)$ r.v.s.

Thus, the average number of choices of course schedules if simultaneous enrollment is allowed is a factor of $256/24 = 32/3 \approx 10.67$ times larger than the average number of choices if simultaneous enrollment is not allowed.

To show that the joint distribution of X_1, X_2, X_3, X_4 is as claimed (this was not required, but is instructive), we can proceed as in the chicken-egg problem, conditioning on X. Only one term in the law of total probability expansion is nonzero, so letting $x_1, \ldots, x_4$ be nonnegative integers and $x = x_1 + x_2 + x_3 + x_4$,

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4) = P(X_1 = x_1, \ldots, X_4 = x_4 | X = x) P(X = x)$$
$$= \frac{x!}{x_1! x_2! x_3! x_4!} \frac{1}{4^x} \frac{e^{-\lambda}\lambda^x}{x!}$$
$$= \frac{e^{-\lambda/4}(\lambda/4)^{x_1}}{x_1!} \cdots \frac{e^{-\lambda/4}(\lambda/4)^{x_4}}{x_4!},$$

which shows that $X_1, \ldots, X_4$ are i.i.d. $\text{Pois}(\lambda/4)$. (The expression $x!/(x_1! x_2! x_3! x_4!)$ is obtained by counting the number of ways to assign x_1 of x courses to be at 10 am, x_2 to be at 11:30 am, and so on: imagine listing x courses in some order and then letting the first x_1 be at 10 am, the next x_2 be at 11:30 am, etc.; we then adjust for the fact that within each time slot's block, the courses could have been listed in any order.)

Multinomial

65. Ⓢ Let $(X_1, \ldots, X_k)$ be Multinomial with parameters n and $(p_1, \ldots, p_k)$. Use indicator r.v.s to show that $\text{Cov}(X_i, X_j) = -np_i p_j$ for $i \neq j$.

Solution: First let us find $\text{Cov}(X_1, X_2)$. Consider the story of the Multinomial, where n objects are being placed into categories $1, \ldots, k$. Let I_i be the indicator r.v. for object i being in category 1, and let J_j be the indicator r.v. for object j being in category 2.

Then $X_1 = \sum_{i=1}^n I_i, X_2 = \sum_{j=1}^n J_j$. So

$$\text{Cov}(X_1, X_2) = \text{Cov}(\sum_{i=1}^n I_i, \sum_{j=1}^n J_j)$$

$$= \sum_{i,j} \text{Cov}(I_i, J_j).$$

All the terms here with $i \neq j$ are 0 since the ith object is categorized independently of the jth object. So this becomes

$$\sum_{i=1}^n \text{Cov}(I_i, J_i) = n\text{Cov}(I_1, J_1) = -np_1 p_2,$$

since

$$\text{Cov}(I_1, J_1) = E(I_1 J_1) - (EI_1)(EJ_1) = -p_1 p_2.$$

By the same method, we have $\text{Cov}(X_i, X_j) = -np_i p_j$ for all $i \neq j$.

66. Ⓢ Consider the birthdays of 100 people. Assume people's birthdays are independent, and the 365 days of the year (exclude the possibility of February 29) are equally likely. Find the covariance and correlation between how many of the people were born on January 1 and how many were born on January 2.

Solution: Let X_j be the number of people born on January j. Then

$$\text{Cov}(X_1, X_2) = -\frac{100}{365^2},$$

using the result about covariances in a Multinomial. Since $X_j \sim \text{Bin}(100, 1/365)$, we then have

$$\text{Corr}(X_1, X_2) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = -\frac{100/365^2}{100(1/365)(364/365)} = -\frac{1}{364}.$$

67. A certain course has a freshmen, b sophomores, c juniors, and d seniors. Let X be the number of freshmen and sophomores (total), Y be the number of juniors, and Z be the number of seniors in a random sample of size n, where for Part (a) the sampling is *with* replacement and for Part (b) the sampling is *without* replacement (for both parts, at each stage the allowed choices have equal probabilities).

(a) Find the joint PMF of X, Y, Z, for sampling with replacement.

(b) Find the joint PMF of X, Y, Z, for sampling without replacement.

Solution:

(a) Let $m = a + b + c + d$ be the number of students in the course. By the story of the Multinomial,

$$(X, Y, Z) \sim \text{Mult}_3(n, (\frac{a+b}{m}, \frac{c}{m}, \frac{d}{m})).$$

The joint PMF is

$$P(X = x, Y = y, Z = z) = \frac{n!}{x!y!z!} \left(\frac{a+b}{m}\right)^x \left(\frac{c}{m}\right)^y \left(\frac{d}{m}\right)^z,$$

for all nonnegative integers x, y, z with $x + y + z = n$.

(b) Analogously to the derivation of the Hypergeometric PMF, we have

$$P(X = x, Y = y, Z = z) = \frac{\binom{a+b}{x}\binom{c}{y}\binom{d}{z}}{\binom{m}{n}},$$

for all nonnegative integers x, y, z with $x + y + z = n, x \leq a + b, y \leq c, z \leq d$.
Note that the marginal distribution of X is $\text{HGeom}(a + b, c + d, n)$.

68. Ⓢ A group of $n \geq 2$ people decide to play an exciting game of Rock-Paper-Scissors. As you may recall, Rock smashes Scissors, Scissors cuts Paper, and Paper covers Rock (despite Bart Simpson saying "Good old rock, nothing beats that!").

Usually this game is played with 2 players, but it can be extended to more players as follows. If exactly 2 of the 3 choices appear when everyone reveals their choice, say $a, b \in \{\text{Rock}, \text{Paper}, \text{Scissors}\}$ where a beats b, the game is decisive: the players who chose a win, and the players who chose b lose. Otherwise, the game is indecisive and the players play again.

For example, with 5 players, if one player picks Rock, two pick Scissors, and two pick Paper, the round is indecisive and they play again. But if 3 pick Rock and 2 pick Scissors, then the Rock players win and the Scissors players lose the game.

Assume that the n players independently and randomly choose between Rock, Scissors, and Paper, with equal probabilities. Let X, Y, Z be the number of players who pick Rock, Scissors, Paper, respectively in one game.

(a) Find the joint PMF of X, Y, Z.

(b) Find the probability that the game is decisive. Simplify your answer.

(c) What is the probability that the game is decisive for $n = 5$? What is the limiting probability that a game is decisive as $n \to \infty$? Explain briefly why your answer makes sense.

Solution:

(a) The joint PMF of X, Y, Z is

$$P(X = a, Y = b, Z = c) = \frac{n!}{a!b!c!} \left(\frac{1}{3}\right)^{a+b+c}$$

where a, b, c are any nonnegative integers with $a + b + c = n$, since $(1/3)^{a+b+c}$ is the probability of any specific configuration of choices for each player with the right numbers in each category, and the coefficient in front counts the number of distinct ways to permute such a configuration.

Alternatively, we can write the joint PMF as

$$P(X = a, Y = b, Z = c) = P(X = a)P(Y = b|X = a)P(Z = c|X = a, Y = b),$$

where for $a + b + c = n$, $P(X = a)$ can be found from the $\text{Bin}(n, 1/3)$ PMF, $P(Y = b|X = a)$ can be found from the $\text{Bin}(n - a, 1/2)$ PMF, and $P(Z = c|X = a, Y = b) = 1$. This is a $\text{Mult}_3(n, (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}))$ distribution.

(b) The game is decisive if and only if exactly one of X, Y, Z is 0. These cases are disjoint so by symmetry, the probability is 3 times the probability that X is zero and Y and Z are nonzero. Note that if $X = 0$ and $Y = k$, then $Z = n - k$. This gives

$$P(\text{decisive}) = 3 \sum_{k=1}^{n-1} \frac{n!}{0!k!(n-k)!} \left(\frac{1}{3}\right)^n$$

$$= 3 \left(\frac{1}{3}\right)^n \sum_{k=1}^{n-1} \binom{n}{k}$$

$$= \frac{2^n - 2}{3^{n-1}}$$

since $\sum_{k=1}^{n-1} \binom{n}{k} = -1 - 1 + \sum_{k=0}^{n} \binom{n}{k} = 2^n - 2$ (by the binomial theorem or the fact that a set with n elements has 2^n subsets). As a check, when $n = 2$ this reduces to $2/3$, which makes sense since for 2 players, the game is decisive if and only if the two players do not pick the same choice.

(c) For $n = 5$, the probability is $(2^5 - 2)/3^4 = 30/81 \approx 0.37$. As $n \to \infty$, $(2^n - 2)/3^{n-1} \to 0$, which make sense since if the number of players is very large, it is very likely that there will be at least one of each of Rock, Paper, and Scissors.

69. ⓢ Emails arrive in an inbox according to a Poisson process with rate λ (so the number of emails in a time interval of length t is distributed as $\text{Pois}(\lambda t)$, and the numbers of emails arriving in disjoint time intervals are independent). Let X, Y, Z be the numbers of emails that arrive from 9 am to noon, noon to 6 pm, and 6 pm to midnight (respectively) on a certain day.

(a) Find the joint PMF of X, Y, Z.

(b) Find the conditional joint PMF of X, Y, Z given that $X + Y + Z = 36$.

(c) Find the conditional PMF of $X + Y$ given that $X + Y + Z = 36$, and find $E(X + Y | X + Y + Z = 36)$ and $\text{Var}(X + Y | X + Y + Z = 36)$ (conditional expectation and conditional variance given an event are defined in the same way as expectation and variance, using the conditional distribution given the event in place of the unconditional distribution).

Solution:

(a) Since $X \sim \text{Pois}(3\lambda), Y \sim \text{Pois}(6\lambda), Z \sim \text{Pois}(6\lambda)$ independently, the joint PMF is

$$P(X = x, Y = y, Z = z) = \frac{e^{-3\lambda}(3\lambda)^x}{x!} \frac{e^{-6\lambda}(6\lambda)^y}{y!} \frac{e^{-6\lambda}(6\lambda)^z}{z!},$$

for any nonnegative integers x, y, z.

(b) Let $T = X + Y + Z \sim \text{Pois}(15\lambda)$, and suppose that we observe $T = t$. The conditional PMF is 0 for $x + y + z \neq t$. For $x + y + z = t$,

$$P(X = x, Y = y, Z = z | T = t) = \frac{P(T = t | X = x, Y = y, Z = z)P(X = x, Y = y, Z = z)}{P(T = t)}$$

$$= \frac{\frac{e^{-3\lambda}(3\lambda)^x}{x!} \frac{e^{-6\lambda}(6\lambda)^y}{y} \frac{e^{-6\lambda}(6\lambda)^z}{z!}}{\frac{e^{-15\lambda}(15\lambda)^t}{t!}}$$

$$= \frac{t!}{x!y!z!}\left(\frac{3}{15}\right)^x \left(\frac{6}{15}\right)^y \left(\frac{6}{15}\right)^z.$$

Thus, (X, Y, Z) is conditionally Multinomial given $T = t$, and we have that (X, Y, Z) is conditionally $\text{Mult}_3(36, (\frac{1}{5}, \frac{2}{5}, \frac{2}{5}))$ given $T = 36$.

(c) Let $W = X + Y$ and $T = X + Y + Z$. Using the story of the Multinomial and Part (b), we can merge the categories "9 am to noon" and "noon to 6 pm" to get

$$W | T = 36 \sim \text{Bin}\left(36, \frac{9}{15}\right).$$

Therefore, $E(W | T = 36) = 36 \cdot \frac{9}{15} = 21.6$ and $\text{Var}(W | T = 36) = 36 \cdot \frac{9}{15} \cdot \frac{6}{15} = 8.64$.

70. Let X be the number of statistics majors in a certain college in the Class of 2030, viewed as an r.v. Each statistics major chooses between two tracks: a general track in statistical principles and methods, and a track in quantitative finance. Suppose that each statistics major chooses randomly which of these two tracks to follow, independently, with probability p of choosing the general track. Let Y be the number of statistics majors who choose the general track, and Z be the number of statistics majors who choose the quantitative finance track.

(a) Suppose that $X \sim \text{Pois}(\lambda)$. (This isn't the exact distribution in reality since a Poisson is unbounded, but it may be a very good approximation.) Find the correlation between X and Y.

(b) Let n be the size of the Class of 2030, where n is a known constant. For this part and the next, instead of assuming that X is Poisson, assume that each of the n students chooses to be a statistics major with probability r, independently. Find the joint distribution of Y, Z, and the number of non-statistics majors, and their marginal distributions.

(c) Continuing as in (b), find the correlation between X and Y.

Solution:

(a) By the chicken-egg story, Y and Z are independent with $Y \sim \text{Pois}(\lambda p)$, $Z \sim \text{Pois}(\lambda q)$, where $q = 1 - p$. So

$$\text{Cov}(X, Y) = \text{Cov}(Y + Z, Y) = \text{Cov}(Y, Y) + \text{Cov}(Z, Y) = \text{Var}(Y) = \lambda p.$$

Thus,

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{\lambda p}{\sqrt{\lambda \lambda p}} = \sqrt{p}.$$

(b) We now have $X \sim \text{Bin}(n, r)$. Each of the n students becomes a Statistics concentrator in the General track with probability rp, a Statistics concentrator in the Quantitative Finance track with probability rq, and a non-Statistics concentrator with probability $1 - r$. By the story of the Multinomial,

$$(Y, Z, n - X) \sim \text{Mult}_3(n, (rp, rq, 1 - r)).$$

By the story of the Binomial, the marginal distributions are

$$Y \sim \text{Bin}(n, rp), Z \sim \text{Bin}(n, rq), n - X \sim \text{Bin}(n, 1 - r).$$

(c) By Theorem 7.4.6 (the result about covariances in a Multinomial),

$$\text{Cov}(X, Y) = \text{Cov}(Y + Z, Y) = \text{Var}(Y) + \text{Cov}(Y, Z) = nrp(1 - rp) - n(rp)(rq) = npr(1 - r).$$

So

$$\text{Corr}(X, Y) = \frac{npr(1 - r)}{\sqrt{nr(1 - r)nrp(1 - rp)}} = \sqrt{\frac{p(1 - r)}{1 - pr}}.$$

Note that if $n \to \infty$ and $r \to 0$ with nr fixed at a value λ, then $\text{Cov}(X, Y) \to \lambda p$, which is the covariance from (a); this makes sense since in this limit, the $\text{Bin}(n, r)$ distribution converges to the $\text{Pois}(\lambda)$ distribution.

71. In humans (and many other organisms), genes come in pairs. Consider a gene of interest, which comes in two types (*alleles*): type a and type A. The *genotype* of a person for that gene is the types of the two genes in the pair: AA, Aa, or aa (aA is equivalent to Aa). According to the Hardy-Weinberg law, for a population in equilibrium the frequencies of AA, Aa, aa will be $p^2, 2p(1-p), (1-p)^2$ respectively, for some p with $0 < p < 1$. Suppose that the Hardy-Weinberg law holds, and that n people are drawn randomly from the population, independently. Let X_1, X_2, X_3 be the number of people in the sample with genotypes AA, Aa, aa, respectively.

(a) What is the joint PMF of X_1, X_2, X_3?

(b) What is the distribution of the number of people in the sample who have an A?

(c) What is the distribution of how many of the $2n$ genes among the people are A's?

(d) Now suppose that p is unknown, and must be estimated using the observed data X_1, X_2, X_3. The *maximum likelihood estimator* (MLE) of p is the value of p for which the observed data are as likely as possible. Find the MLE of p.

(e) Now suppose that p is unknown, and that our observations can't distinguish between AA and Aa. So for each person in the sample, we just know whether or not that person is an aa (in genetics terms, AA and Aa have the same *phenotype*, and we only get to observe the phenotypes, not the genotypes). Find the MLE of p.

Solution:

(a) By the story of the Multinomial, $(X_1, X_2, X_3) \sim \text{Mult}_3(n, (p^2, 2pq, q^2))$, where $q = 1 - p$. The PMF is

$$P(X_1 = n_1, X_2 = n_2, X_3 = n_3) = \frac{n!}{n_1!n_2!n_3!} p^{2n_1} (2pq)^{n_2} q^{2n_3},$$

for $n_1 + n_2 + n_3 = n$.

(b) By the story of the Binomial (defining "success" as having an A and "failure" as not having an A), the distribution is $\text{Bin}(n, p^2 + 2pq)$.

(c) Let Y_j be how many A's the jth person in the sample has. Then Y_j is 2 with probability p^2, 1 with probability $2pq$, and 0 with probability q^2, so $Y_j \sim \text{Bin}(2, p)$. The Y_j are also independent. Therefore, $Y_1 + \cdots + Y_n \sim \text{Bin}(2n, p)$.

(d) Let x_1, x_2, x_3 be the observed values of X_1, X_2, X_3. The MLE of p is the value of p that maximizes the function $L(p) = p^{2x_1} (pq)^{x_2} q^{2x_3} = p^{2x_1 + x_2} (1 - p)^{x_2 + 2x_3}$ (we can omit factors which are constant with respect to p, since such constants do not affect where the maximum is). Equivalently, we can maximize the log:

$$\log L(p) = (2x_1 + x_2) \log p + (x_2 + 2x_3) \log(1 - p).$$

Setting the derivative of $\log L(p)$ equal to 0, we have

$$\frac{2x_1 + x_2}{p} - \frac{x_2 + 2x_3}{1 - p} = 0,$$

which rearranges to

$$p = \frac{2x_1 + x_2}{2(x_1 + x_2 + x_3)} = \frac{2x_1 + x_2}{2n}.$$

This value of p does maximize $\log L(p)$ since the derivative of $\log L(p)$ is positive everywhere to the left of it and is negative everywhere to the right of it. Thus, the MLE of p, which we denote by $\hat{p}$, is given by $\hat{p} = (2X_1 + X_2)/(2n)$. Note that this has an intuitive interpretation: it is the fraction of A's among the $2n$ genes.

(e) Let $Y \sim \text{Bin}(n, q^2)$ be the number of aa people, and let y be the observed value of Y. We need to maximize the function $L_2(q) = q^{2y} (1 - q^2)^{n - y}$ (we will maximize over q and then find the corresponding value of p). Then

$$\log L_2(q) = 2y \log q + (n - y) \log(1 - q^2),$$

so

$$\frac{d \log L_2(q)}{dq} = \frac{2y}{q} - \frac{2q(n - y)}{1 - q^2},$$

which simplifies to $y = q^2 n$. By looking at the sign of the derivative, we see that $\log L_2(q)$ is maximized at $q = \sqrt{y/n}$. Thus, the MLE of q is $\sqrt{Y/n}$, which shows that the MLE of p is $1 - \sqrt{Y/n}$.

Multivariate Normal

72. Ⓢ Let (X, Y) be Bivariate Normal, with X and Y marginally $\mathcal{N}(0, 1)$ and with correlation ρ between X and Y.

(a) Show that $(X + Y, X - Y)$ is also Bivariate Normal.

(b) Find the joint PDF of $X + Y$ and $X - Y$ (without using calculus), assuming $-1 < \rho < 1$.

Solution:

(a) The linear combination $s(X + Y) + t(X - Y) = (s + t)X + (s - t)Y$ is also a linear combination of X and Y, so it is Normal, which shows that $(X + Y, X - Y)$ is MVN.

(b) Since $X + Y$ and $X - Y$ are uncorrelated (as $\text{Cov}(X + Y, X - Y) = \text{Var}(X) - \text{Var}(Y) = 0$) and $(X + Y, X - Y)$ is MVN, they are independent. Marginally, $X + Y \sim \mathcal{N}(0, 2 + 2\rho)$ and $X - Y \sim \mathcal{N}(0, 2 - 2\rho)$. Thus, the joint PDF is

$$f(s, t) = \frac{1}{4\pi\sqrt{1 - \rho^2}} e^{-\frac{1}{4}(s^2/(1+\rho) + t^2/(1-\rho))}.$$

73. Let the joint PDF of X and Y be

$$f_{X,Y}(x, y) = c \exp\left(-\frac{x^2}{2} - \frac{y^2}{2}\right) \text{ for all } x \text{ and } y,$$

where c is a constant.

(a) Find c to make this a valid joint PDF.

(b) What are the marginal distributions of X and Y? Are X and Y independent?

(c) Is (X, Y) Bivariate Normal?

Solution:

(a) Since for all real x, y the joint PDF factors as

$$f_{X,Y}(x, y) = c \exp\left(-\frac{x^2}{2}\right) \exp\left(-\frac{y^2}{2}\right),$$

X and Y are independent. The marginal PDF of X is proportional to $\exp(-x^2/2)$, so $X \sim \mathcal{N}(0, 1)$. Similarly, $Y \sim \mathcal{N}(0, 1)$. Using what we already know about the Normal normalizing constant, we then have

$$c = \left(\frac{1}{\sqrt{2\pi}}\right)^2 = \frac{1}{2\pi}.$$

(b) As shown in (a), X and Y are i.i.d. $\mathcal{N}(0, 1)$ r.v.s.

(c) Yes, (X, Y) is Bivariate Normal since $aX + bY$ is Normal for any constants a and b.

74. Let the joint PDF of X and Y be

$$f_{X,Y}(x, y) = c \exp\left(-\frac{x^2}{2} - \frac{y^2}{2}\right) \text{ for } xy > 0,$$

where c is a constant (the joint PDF is 0 for $xy \leq 0$).

(a) Find c to make this a valid joint PDF.

(b) What are the marginal distributions of X and Y? Are X and Y independent?

(c) Is (X, Y) Bivariate Normal?

Solution:

(a) The constraint $xy > 0$ corresponds to the two quadrants $x > 0, y > 0$ and $x < 0, y < 0$ in the plane. Let

$$g(x,y) = \exp\left(-\frac{x^2}{2}\right)\exp\left(-\frac{y^2}{2}\right).$$

We want to find

$$\int_0^\infty \int_0^\infty g(x,y)dxdy + \int_{-\infty}^0 \int_{-\infty}^0 g(x,y)dxdy.$$

We know from the Normal normalizing constant that

$$\int_0^\infty \int_0^\infty g(x,y)dxdy + \int_0^\infty \int_{-\infty}^0 g(x,y)dxdy + \int_{-\infty}^0 \int_0^\infty g(x,y)dxdy + \int_{-\infty}^0 \int_{-\infty}^0 g(x,y)dxdy = 2\pi.$$

But by symmetry (corresponding to the transformation $u = -x$ or $v = -y$), all of these four terms are equal. So each of the four terms is $\frac{\pi}{2}$. Therefore,

$$\int_0^\infty \int_0^\infty g(x,y)dxdy + \int_{-\infty}^0 \int_{-\infty}^0 g(x,y)dxdy = \pi,$$

which shows that

$$c = \frac{1}{\pi}.$$

(b) The r.v.s X and Y are *not* independent, because of the constraint $XY > 0$. If we observe that X is positive, then we know that Y is positive; if we observe that X is negative, then we know that Y is negative. For $x > 0$, the marginal PDF of X is

$$f_X(x) = \frac{1}{\pi}\int_0^\infty e^{-x^2/2}e^{-y^2/2}dy = \frac{e^{-x^2/2}}{\pi}\int_0^\infty e^{-y^2/2}dy = \frac{e^{-x^2/2}}{\pi}\cdot\frac{\sqrt{2\pi}}{2} = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}.$$

Similarly, for $x < 0$ the marginal PDF of X is

$$f_X(x) = \frac{1}{\pi}\int_{-\infty}^0 e^{-x^2/2}e^{-y^2/2}dy = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}.$$

Therefore, $X \sim \mathcal{N}(0,1)$. By symmetry, we also have $Y \sim \mathcal{N}(0,1)$.

(c) The random vector (X,Y) is not Bivariate Normal since its support consists of two of the four quadrants in the plane, whereas the support of a Bivariate Normal is the entire plane (except in degenerate cases, where the support is a line or a point).

75. Let X, Y, Z be i.i.d. $\mathcal{N}(0,1)$. Find the joint MGF of $(X + 2Y, 3X + 4Z, 5Y + 6Z)$.

Solution: Using the result for the $\mathcal{N}(0,1)$ MGF, the desired joint MGF is

$$M(a,b,c) = E\left(e^{a(X+2Y)+b(3X+4Z)+c(5Y+6Z)}\right)$$

$$= E\left(e^{(a+3b)X+(2a+5c)Y+(4b+6c)Z}\right)$$

$$= E\left(e^{(a+3b)X}\right)E\left(e^{(2a+5c)Y}\right)E\left(e^{(4b+6c)Z}\right)$$

$$= \exp\left((a+3b)^2/2 + (2a+5c)^2/2 + (4b+6c)^2/2\right).$$

76. Let X and Y be i.i.d. $\mathcal{N}(0,1)$, and let S be a random sign (1 or -1, with equal probabilities) independent of (X,Y).

(a) Determine whether or not $(X, Y, X + Y)$ is Multivariate Normal.

(b) Determine whether or not $(X, Y, SX + SY)$ is Multivariate Normal.

(c) Determine whether or not (SX, SY) is Multivariate Normal.

Solution:

(a) Yes, since $aX + bY + c(X + Y) = (a + c)X + (b + c)Y$ is Normal for any a, b, c.

(b) No, since $X + Y + (SX + SY) = (1 + S)X + (1 + S)Y$ is 0 with probability $1/2$ (since it is 0 if $S = -1$ and a non-degenerate Normal if $S = 1$).

(c) Yes. To prove this, let's show that any linear combination

$$a(SX) + b(SY) = S(aX + bY)$$

is Normal. We already know that

$$aX + bY \sim \mathcal{N}(0, a^2 + b^2).$$

By the symmetry of the Normal, as discussed in Example 7.5.2, $SZ \sim \mathcal{N}(0, 1)$ if $Z \sim \mathcal{N}(0, 1)$ and S is a random sign independent of Z. Letting

$$Z = \frac{aX + bY}{\sqrt{a^2 + b^2}},$$

we have

$$S(aX + bY) = \sqrt{a^2 + b^2} \cdot SZ \sim \mathcal{N}(0, a^2 + b^2).$$

77. Let (X, Y) be Bivariate Normal with $X \sim \mathcal{N}(0, \sigma_1^2)$ and $Y \sim \mathcal{N}(0, \sigma_2^2)$ marginally and with $\mathrm{Corr}(X, Y) = \rho$. Find a constant c such that $Y - cX$ is independent of X.

Hint: First find c (in terms of ρ, σ_1, σ_2) such that $Y - cX$ and X are uncorrelated.

Solution: Following the hint, let's find c such that $\mathrm{Cov}(Y - cX, X) = 0$. We have

$$\mathrm{Cov}(Y - cX, X) = \mathrm{Cov}(X, Y) - c\mathrm{Var}(X) = \rho\sigma_1\sigma_2 - c\sigma_1^2,$$

so

$$c = \frac{\rho\sigma_2}{\sigma_1}.$$

But $(Y - cX, X)$ is Bivariate Normal (by the argument given in Example 7.5.3), so the above choice of c makes $Y - cX$ independent of X, not just uncorrelated with X.

78. A mother and a father have 6 children. The 8 heights in the family (in inches) are $\mathcal{N}(\mu, \sigma^2)$ r.v.s (with the same distribution, but not necessarily independent).

(a) Assume for this part that the heights are all independent. On average, how many of the children are taller than *both* parents?

(b) Let X_1 be the height of the mother, X_2 be the height of the father, and $Y_1, \ldots, Y_6$ be the heights of the children. Suppose that $(X_1, X_2, Y_1, \ldots, Y_6)$ is Multivariate Normal, with $\mathcal{N}(\mu, \sigma^2)$ marginals and $\mathrm{Corr}(X_1, Y_j) = \rho$ for $1 \leq j \leq 6$, with $\rho < 1$. On average, how many of the children are more than 1 inch taller than their mother?

Solution:

(a) Let I_j be the indicator of the jth child being taller than both parents. By symmetry (all orderings of the heights of the jth child and the parents are equally likely), $E(I_j) = 1/3$. So by linearity, the desired expectation is $6/3 = 2$.

(b) Again using indicator r.v.s and linearity, the desired expectation is $6P(Y - X > 1)$, where (X, Y) is Bivariate Normal with $\mathcal{N}(\mu, \sigma^2)$ marginals and correlation ρ. Then $Y - X \sim \mathcal{N}(0, 2\sigma^2 - 2\sigma^2\rho)$, so the desired expectation is

$$6P(Y - X > 1) = 6P\left(\frac{Y - X}{\sigma\sqrt{2(1 - \rho)}} > \frac{1}{\sigma\sqrt{2(1 - \rho)}}\right) = 6\left(1 - \Phi\left(\frac{1}{\sigma\sqrt{2(1 - \rho)}}\right)\right).$$

Mixed practice

79. Cars pass by a certain point on a road according to a Poisson process with rate λ cars/minute. Let $N_t \sim \text{Pois}(\lambda t)$ be the number of cars that pass by that point in the time interval $[0, t]$, with t measured in minutes.

    (a) A certain device is able to count cars as they pass by, but it does not record the arrival times. At time 0, the counter on the device is reset to 0. At time 3 minutes, the device is observed and it is found that exactly 1 car had passed by. Given this information, find the conditional CDF of when that car arrived. Also describe in words what the result says.

    (b) In the late afternoon, you are counting blue cars. Each car that passes by is blue with probability b, independently of all other cars. Find the joint PMF and marginal PMFs of the number of blue cars and number of non-blue cars that pass by the point in 10 minutes.

    Solution:

    (a) Let T_1 be the arrival time of the first car to arrive after time 0. Unconditionally, $T_1 \sim \text{Expo}(\lambda)$. Given $N_3 = 1$, for $0 \le t \le 3$ we have

    $$P(T_1 \le t | N_3 = 1) = P(N_t \ge 1 | N_3 = 1) = \frac{P(N_t \ge 1, N_3 = 1)}{P(N_3 = 1)} = \frac{P(N_t = 1, N_3 = 1)}{P(N_3 = 1)}.$$

    By definition of Poisson process, the numerator is

    $$P(\text{exactly 1 car in } [0, t])P(\text{no cars in } (t, 3]) = e^{-\lambda t}\lambda t e^{-\lambda(3-t)} = \lambda t e^{-3\lambda},$$

    and the denominator is $e^{-3\lambda}(3\lambda)$. So the conditional CDF of T_1 is

    $$P(T_1 \le t | N_3 = 1) = t/3$$

    for $0 \le t \le 3$ (and 0 for $t < 0$ and 1 for $t > 3$). This says that the conditional distribution of the first arrival time, given that there was exactly one arrival in $[0, 3]$, is $\text{Unif}(0, 3)$.

    (b) Let X and Y be the number of blue and non-blue cars that pass by in those 10 minutes, respectively, and $N = X + Y$. Then $N \sim \text{Pois}(10\lambda)$ and $X|N \sim \text{Bin}(N, b)$. By the chicken-egg story, X and Y are *independent* with $X \sim \text{Pois}(10\lambda b)$, $Y \sim \text{Pois}(10\lambda(1-b))$. The joint PMF is the product of the marginal PMFs:

    $$P(X = i, Y = j) = \frac{e^{-10\lambda b}(10\lambda b)^i}{i!}\frac{e^{-10\lambda(1-b)}(10\lambda(1-b))^j}{j!},$$

    for all nonnegative integers i, j.

80. In a U.S. election, there will be $V \sim \text{Pois}(\lambda)$ registered voters. Suppose each registered voter is a registered Democrat with probability p and a registered Republican with probability $1 - p$, independent of other voters. Also, each registered voter shows up to the polls with probability s and stays home with probability $1 - s$, independent of other voters and independent of their own party affiliation. In this problem, we are interested in X, the number of registered Democrats who actually vote.

    (a) What is the distribution of X, before we know anything about the number of registered voters?

    (b) Suppose we learn that $V = v$; that is, v people registered to vote. What is the conditional distribution of X given this information?

    (c) Suppose we learn there were d registered Democrats and r registered Republicans (where $d + r = v$). What is the conditional distribution of X given this information?

(d) Finally, we learn in addition to all of the above information that n people showed up at the polls on election day. What is the conditional distribution of X given this information?

Solution:

(a) Each registered voter has probability ps of being a registered Democrat who will show up to vote. By the chicken-egg story, $X \sim \text{Pois}(ps\lambda)$.

(b) By the story of the Binomial, the conditional distribution of $X|V = v$ is $\text{Bin}(v, ps)$.

(c) Each of the d registered Democrats will vote with probability s, with these decisions independent of each other and of the decisions of the r registered Republicans. So the conditional distribution of X is $\text{Bin}(d, s)$.

(d) By the story of the Hypergeometric (thinking of registered Democrats as "tagged" and registered Republicans as "untagged", and choosing a random sample consisting of n of the $d + r$ registered voters to determine who actually voted, with all samples of size n equally likely), the conditional distribution of X is $\text{HGeom}(d, r, n)$.

81. A certain college has m freshmen, m sophomores, m juniors, and m seniors. A certain class there is a simple random sample of size n students, i.e., all sets of n of the $4m$ students are equally likely. Let $X_1, \ldots, X_4$ be the numbers of freshmen, $\ldots$, seniors in the class.

(a) Find the joint PMF of X_1, X_2, X_3, X_4.

(b) Give both an intuitive explanation and a mathematical justification for whether or not the distribution from (a) is Multinomial.

(c) Find $\text{Cov}(X_1, X_3)$, fully simplified.

Hint: Take the variance of both sides of $X_1 + X_2 + X_3 + X_4 = n$.

Solution:

(a) Analogously to how we obtained the Hypergeometric PMF,

$$P(X_1 = x_1, \ldots, X_4 = x_4) = \frac{\binom{m}{x_1}\binom{m}{x_2}\binom{m}{x_3}\binom{m}{x_4}}{\binom{4m}{n}},$$

for integers $x_1, \ldots, x_4$ with $0 \le x_j \le m$ and $x_1 + \cdots + x_4 = n$ (and the PMF is 0 otherwise).

(b) Let $\mathbf{X} = (X_1, \ldots, X_4)$. For the trivial case $n = 1$, $\mathbf{X}$ is Multinomial since then there is no difference between sampling with vs. without replacement. Now assume $n \ge 2$. Intuitively, the distribution is not Multinomial since the class is obtained by sampling without replacement, which makes the choices dependent; for example, drawing a senior at some step makes it less likely to draw a senior at the next step.

To prove that it's not Multinomial, note that the marginals of a Multinomial are Binomial, whereas the marginals of $\mathbf{X}$ are Hypergeometric. Alternatively, just check that the joint PMFs are not the same, though it takes some work to give a full proof that there isn't some amazing algebra that would convert one to the other.

(c) Using the hint,

$$0 = \text{Var}(X_1 + \cdots + X_4) = 4\text{Var}(X_1) + 2\binom{4}{2}\text{Cov}(X_1, X_3).$$

We have $X_1 \sim \text{HGeom}(m, 3m, n)$, so by Example 7.3.7,

$$\text{Var}(X_1) = \frac{N - n}{N - 1} np(1 - p),$$

with $N = 4m$ and $p = m/(4m) = 1/4$. Thus,

$$\text{Cov}(X_1, X_3) = -\frac{\text{Var}(X_1)}{3} = -\frac{n}{16} \cdot \frac{4m-n}{4m-1}.$$

82. Let $X \sim \text{Expo}(\lambda)$ and let Y be a nonnegative random variable, discrete or continuous, whose MGF M is finite everywhere. Assume that X and Y are independent. Show that $P(Y < X) = M(c)$ for a certain value of c (which you should specify).

    Solution: For the continuous case, let f be the PDF of Y. By LOTP and LOTUS,

    $$P(Y < X) = \int_0^\infty P(Y < X|Y = y)f(y)dy = \int_0^\infty P(X > y)f(y)dy = \int_0^\infty e^{-\lambda y}f(y)dy = M(-\lambda).$$

    For the discrete case, the derivation is analogous, and again $P(Y < X) = M(-\lambda)$.

83. A publishing company employs two proofreaders, Prue and Frida. When Prue is proofreading a book, for each typo she has probability p of catching it and $q = 1-p$ of missing it, independently. When Frida is proofreading a book, for each typo she has probability f of catching it and $g = 1 - f$ of missing it, independently.

    (a) A certain book draft has n typos. The company randomly assigns it to one of the two proofreaders, with equal probabilities. Find the distribution of the number of typos that get detected.

    (b) Another book is being written. When a draft of the book is complete, it will have a $\text{Pois}(\lambda)$ number of typos, and will be assigned to Prue to proofread. Find the probability that Prue catches exactly k typos, given that she misses exactly m typos.

    Solution:

    (a) Let T be the number of typos found and B be the event that the book is assigned to Prue. Then $T|B \sim \text{Bin}(n, p)$ and $T|B^c \sim \text{Bin}(n, f)$, so the PMF of T is

    $$P(T = k) = P(T = k|B)P(B) + P(T = k|B^c)P(B^c) = \frac{1}{2}\binom{n}{k}p^k q^{n-k} + \frac{1}{2}\binom{n}{k}f^k g^{n-k},$$

    for $k = 0, 1, \ldots, n$.

    (b) Let X be the number of typos that Prue catches, and Y be the number she misses. By the chicken-egg story, $X \sim \text{Pois}(\lambda p), Y \sim \text{Pois}(\lambda q)$, and X and Y are independent. Hence,

    $$P(X = k|Y = m) = P(X = k) = \frac{e^{-\lambda p}(\lambda p)^k}{k!}, \text{ for } k = 0, 1, 2, \ldots.$$

84. Two authors, Bob and Martha, are about to begin writing an epic co-authored book, *The Adventures of Aaron the Aardwolf*. It will take them A years to write. When they finish this book, they will immediately begin work on new, individually authored books. Bob will spend X years writing *The Bilinear Bonanza of Bonnie the Butterfly*, and Martha will spend Y years writing *Memoirs of Maude the Magnificent Mangabey*, independently. Suppose that A, X, Y are i.i.d. $\text{Expo}(\lambda)$. On a timeline where time 0 is defined as the time when they begin their collaboration, consider the following quantities.

    A: time at which *The Adventures of Aaron the Aardwolf* is completed;

    B: time at which *The Bilinear Bonanza of Bonnie the Butterfly* is completed;

    M: time at which *Memoirs of Maude the Magnificent Mangabey* is completed;

    T: time at which the last to be completed of these three books is completed.

(a) Find the distribution of B (which is also the distribution of M).

(b) Find $\text{Cov}(A, B)$.

(c) Find $E(T)$.

Solution:

(a) The CDF of $B = A + X$ is

$$F_B(b) = P(B \le b) = \int_0^\infty P(B \le b | X = x) f_X(x) dx = \int_0^\infty P(A \le b - x) f_X(x) dx.$$

Differentiating both sides with respect to b, we have that the PDF of B is

$$f_B(b) = \int_0^\infty f_A(b - x) f_X(x) dx = \int_0^b \lambda e^{-\lambda(b-x)} \lambda e^{-\lambda x} dx.$$

Simplifying,

$$f_B(b) = \lambda^2 \int_0^b e^{-\lambda b} dx = \lambda^2 b e^{-\lambda b}, \text{ for } b > 0.$$

(The distribution of B is called Gamma$(2, \lambda)$, as we will discuss in Chapter 8.)

(b) By bilinearity of covariance and the fact that independent implies uncorrelated,

$$\text{Cov}(A, B) = \text{Cov}(A, A + X) = \text{Cov}(A, A) + \text{Cov}(A, X) = \text{Var}(A) = \frac{1}{\lambda^2}.$$

(c) We have $T = A + \max(X, Y)$. Remembering the memoryless property and the fact that $\min(X, Y) \sim \text{Expo}(2\lambda)$, on average it will take $1/(2\lambda)$ years (after time A) until either author finishes their single-authored book and then an additional $1/\lambda$ years for the other author to finish. Thus,

$$E(T) = E(A) + E(\max(X, Y)) = \frac{1}{\lambda} + \frac{1}{2\lambda} + \frac{1}{\lambda} = \frac{5}{2\lambda}.$$

85. A DNA sequence can be represented as a sequence of letters, where the *alphabet* has 4 letters: A,C,G,T. Suppose that a random DNA sequence of length $n \ge 4$ is formed by independently generating letters one at a time, with p_A, p_C, p_G, p_T the probabilities of A,C,G,T, respectively, where $p_A + p_C + p_G + p_T = 1$.

(a) Find the covariance between the number of A's and the number of C's in the sequence.

(b) It is observed that the sequence contains exactly a A's, c C's, g G's, and t T's, where $a + c + g + t = n$ and $a \ge 2$. Given this information, find the probability that the first A in the sequence is followed immediately by another A.

Hint: How does this part relate to Exercise 74 in Chapter 2?

(c) Given the information from (b) about how many times each letter occurs, find the expected number of occurrences of the expression CAT in the sequence.

Solution:

(a) The vector of counts for A,C,G,T is Multinomial. So the covariance is $-n p_A p_C$.

(b) By symmetry, the probability that the next letter after the first A is an A is the same as the probability that the last letter in the sequence is an A (this is true both conditionally given where in the sequence the first A occurs, and unconditionally). Hence, the probability is a/n.

(c) Let I_j be the indicator of CAT appearing starting at position j, for $1 \leq j \leq n-2$. By symmetry and linearity, the expected number of occurrences of CAT is

$$(n-2)E(I_1) = (n-2) \cdot \frac{cat}{n(n-1)(n-2)} = \frac{cat}{n(n-1)}.$$

86. To test for a certain disease, the level of a certain substance in the blood is measured. Let T be this measurement, considered as a continuous r.v. The patient tests positive (i.e., is declared to have the disease) if $T > t_0$ and tests negative if $T \leq t_0$, where t_0 is a threshold decided upon in advance. Let D be the indicator of having the disease. As discussed in Example 2.3.9, the *sensitivity* of the test is the probability of testing positive given that the patient has the disease, and the *specificity* of the test is the probability of testing negative given that the patient does not have the disease.

    (a) The *ROC (receiver operator characteristic) curve* of the test is the plot of sensitivity vs. 1 minus specificity, where sensitivity (the vertical axis) and 1 minus specificity (the horizontal axis) are viewed as functions of the threshold t_0. ROC curves are widely used in medicine and engineering as a way to study the performance of procedures for classifying individuals into two groups (in this case, the two groups are "diseased people" and "non-diseased people").

    Given that $D = 1$, T has CDF G and PDF g; given that $D = 0$, T has CDF H and PDF h. Here g and h are positive on an interval $[a, b]$ and 0 outside this interval. Show that the area under the ROC curve is the probability that a randomly selected diseased person has a higher T value than a randomly selected non-diseased person.

    (b) Explain why the result of (a) makes sense in two extreme cases: when $g = h$, and when there is a threshold t_0 such that $P(T > t_0|D = 1)$ and $P(T \leq t_0|D = 0)$ are very close to 1.

    Solution:

    (a) Let T_1 and T_0 be the measurements for a random diseased person and a random healthy person, respectively, with T_1 and T_0 independent with marginal distributions $T_1 \sim G, T_0 \sim H$. The joint PDF of (T_1, T_0) is $f(s, t) = g(s)h(t)$, so

    $$P(T_1 > T_0) = \int_a^b \int_t^b g(s)h(t)ds dt.$$

    Now let us find the area under the ROC curve (as an integral). Let $y = $ sens and $x = 1 - $ spec. The ROC curve is the graph of y as a function of x, so the area under the curve is $\int_0^1 y dx$. For the test with threshold t, y and x as functions of t are

    $$y = \int_t^b g(s)ds, \text{ and } x = \int_t^b h(s)ds.$$

    By the fundamental theorem of calculus, $dx = -h(t)dt$. Also, note that $x = 0$ implies $t = b$ and $x = 1$ implies $t = a$. Expressing y and dx in terms of t, the area is

    $$\int_0^1 y dx = \int_a^b \left(\int_t^b g(s)ds \right) h(t)dt = \int_a^b \int_t^b g(s)h(t)ds dt = P(T_1 > T_0).$$

    (b) In the extreme case $g = h$, the test is useless since then disease status is independent of the level of the substance in the blood. Then the ROC curve is the line $y = x$ and the area is $1/2$. In this case, T_1 and T_0 as defined in the previous part are i.i.d. continuous r.v.s, so we have $P(T_1 > T_0) = 1/2$ by symmetry. Now suppose there is a threshold t_0 such that $P(T > t_0|D = 1) \approx 1$ and $P(T \leq t_0|D = 0) \approx 1$. Then $P(T_1 > T_0) \approx 1$ since with very high probability, $T_0 \leq t_0 < T_1$. The area under the ROC curve is also very close to 1 since any ROC curve is increasing and the point corresponding to threshold t_0 is approximately $(0, 1)$, so the region under the curve is approximately the region under the line $y = 1$ from $x = 0$ to $x = 1$.

87. Let J be Discrete Uniform on $\{1, 2, \ldots, n\}$.

(a) Find $E(J)$ and $\text{Var}(J)$, fully simplified, using results from Section A.8 of the math appendix.

(b) Discuss intuitively whether the results in (a) should be approximately the same as the mean and variance (respectively) of a Uniform distribution on a certain interval.

(c) Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(0,1)$ r.v.s, and let $R_1, \ldots, R_n$ be their ranks (the smallest X_i has rank 1, the next has rank 2, $\ldots$, and the largest has rank n). Explain why

$$R_n = 1 + \sum_{j=1}^{n-1} I_j,$$

where $I_j = I(X_n > X_j)$. Then use this to find $E(R_n)$ and $\text{Var}(R_n)$ directly using symmetry, linearity, the fundamental bridge, and properties of covariance.

(d) Explain how the results of (a) and (c) relate. Then prove the identities

$$\sum_{j=1}^{n} j = \frac{n(n+1)}{2} \text{ and } \sum_{j=1}^{n} j^2 = \frac{n(n+1)(2n+1)}{6},$$

using probability (rather than induction).

Solution:

(a) We have

$$E(J) = \frac{1}{n} \sum_{j=1}^{n} j = \frac{n+1}{2},$$

$$E(J^2) = \frac{1}{n} \sum_{j=1}^{n} j^2 = \frac{(n+1)(2n+1)}{6}.$$

So

$$\text{Var}(J) = \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} = \frac{(n-1)(n+1)}{12}.$$

(b) Let $U \sim \text{Unif}(1, n)$. So U takes on a *real* value between 1 and n, whereas J takes on an *integer* value between 1 and n. The distribution of J is called *Discrete Uniform* since it is the discrete analog of the continuous Uniform distribution. Now let's compare the means and variances.

$$E(U) = \frac{n+1}{2} = E(J),$$

$$\text{Var}(U) = \frac{(n-1)^2}{12} \leq \frac{(n-1)(n+1)}{12} = \text{Var}(J).$$

So the means are exactly the same, which makes sense since for both the mean is the midpoint between 1 and n.

For large n, both $(n-1)^2$ and $(n-1)(n+1)$ are approximately n^2, so the variances are approximately equal. This also makes sense intuitively since we can rescale J and U by dividing by n (so the new supports are $1/n, 2/n, \ldots, 1$ and $(0,1)$, respectively), and then think in terms of approximating a $\text{Unif}(0,1)$ r.v. by a discretized version of it (by rounding to some number of decimal places). Indeed, computers do store real numbers only to finite precision, since in general it would take infinite memory to store the infinitely many digits in a real number's decimal representation.

For small n, $\text{Var}(U)$ is substantially less than $\text{Var}(J)$. This can be explained from the substantial probability of J being at one of the extremes, 1 or n. For example, for $n = 4$ a $\text{Unif}(1, n)$ r.v. is unlikely to be very close to 1 or n, but J has a 50% chance of being at one the extreme points.

(c) By symmetry, the marginal distribution of each R_i is the same as the distribution from (a): it takes on values $1, 2, \ldots, n$, with equal probabilities. We can write $R_n = 1 + \sum_{j=1}^{n-1} I_j$ since this computes X_n's rank by seeing how many of the other X_j's it is bigger than, adding 1 since ranks start at 1. By symmetry, $E(I_j) = 1/2$ for $1 \le j \le n-1$. Again by symmetry, $E(I_j I_k) = 1/3$ for $1 \le j < k \le n-1$ since $I_j I_k = 1$ is the event that X_n is the largest of X_n, X_i, X_j (this is reminiscent of the Putnam problem). Thus,

$$\text{Cov}(I_j, I_k) = E(I_j I_k) - E(I_j)E(I_k) = 1/3 - 1/4 = 1/12$$

for $j \ne k$. So

$$E(R_n) = 1 + \frac{n-1}{2} = \frac{n+1}{2},$$

$$\text{Var}(R_n) = (n-1)\text{Var}(I_1) + 2\binom{n-1}{2}\text{Cov}(I_1, I_2)$$

$$= \frac{3(n-1)}{12} + \frac{(n-1)(n-2)}{12}$$

$$= \frac{(n-1)(n+1)}{12}.$$

(d) As noted above, R_n and J have the same distribution. So they have the same mean and variance, which means that we have computed $E(J), \text{Var}(J)$ in two different ways. One way required knowing the formulas for $\sum_{j=1}^n j$ and $\sum_{j=1}^n j^2$, and the other didn't. So as a byproduct of the above work, we have probabilistic derivations of the sum formulas:

$$\frac{1}{n}\sum_{j=1}^n j = E(J) = \frac{n+1}{2},$$

$$\frac{1}{n}\sum_{j=1}^n j^2 = E(J^2) = \text{Var}(J) + (E(J))^2 = \frac{(n-1)(n+1)}{12} + \left(\frac{n+1}{2}\right)^2,$$

which shows that
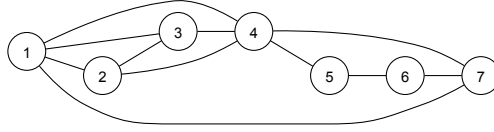
$$\sum_{j=1}^n j = \frac{n(n+1)}{2},$$

$$\sum_{j=1}^n j^2 = \frac{n(n-1)(n+1)}{12} + \frac{3n(n+1)^2}{12} = \frac{n(n+1)(2n+1)}{6}.$$

88. ⓢ A *network* consists of n *nodes*, each pair of which may or may not have an *edge* joining them. For example, a social network can be modeled as a group of n nodes (representing people), where an edge between i and j means they know each other. Assume the network is undirected and does not have edges from a node to itself (for a social network, this says that if i knows j, then j knows i and that, contrary to Socrates' advice, a person does not know himself or herself). A *clique* of size k is a set of k nodes where every node has an edge to every other node (i.e., within the clique, everyone knows everyone). An *anticlique* of size k is a set of k nodes where there are no edges between them (i.e., within the anticlique, no one knows anyone else). For example, the picture below shows a network with nodes labeled $1, 2, \ldots, 7$, where $\{1, 2, 3, 4\}$ is a clique of size 4, and $\{3, 5, 7\}$ is an anticlique of size 3.

(a) Form a random network with n nodes by independently flipping fair coins to decide for each pair $\{x, y\}$ whether there is an edge joining them. Find the expected number of cliques of size k (in terms of n and k).

(b) A *triangle* is a clique of size 3. For a random network as in (a), find the variance of the number of triangles (in terms of n).

Hint: Find the covariances of the indicator random variables for each possible clique. There are $\binom{n}{3}$ such indicator r.v.s, some pairs of which are dependent.

*(c) Suppose that $\binom{n}{k} < 2^{\binom{k}{2}-1}$. Show that there is a network with n nodes containing no cliques of size k or anticliques of size k.

Hint: Explain why it is enough to show that for a random network with n nodes, the probability of the desired property is positive; then consider the complement.

Solution:

(a) Order the $\binom{n}{k}$ subsets of people of size k in some way (i.e., give each subset of size k a code number), and let X_i be the indicator. Since $X_1 + X_2 + \cdots + X_{\binom{n}{k}}$ is the number of cliques of size k, the expected number is

$$E(X_1 + X_2 + \cdots + X_{\binom{n}{k}}) = \binom{n}{k} E(X_1) = \binom{n}{k} P(X_1 = 1) = \frac{\binom{n}{k}}{2^{\binom{k}{2}}}.$$

(b) Let $k = 3$ and the X_i be as in (a). Then

$$\text{Var}(X_1 + \cdots + X_{\binom{n}{3}}) = \text{Var}(X_1) + \cdots + \text{Var}(X_{\binom{n}{3}}) + 2 \sum_{i<j} \text{Cov}(X_i, X_j)$$

$$= \binom{n}{3} \text{Var}(X_1) + 2 \sum_{i<j} \text{Cov}(X_i, X_j),$$

with

$$\text{Var}(X_1) = 2^{-\binom{3}{2}}(1 - 2^{-\binom{3}{2}}) = \frac{7}{64}.$$

To compute $\text{Cov}(X_i, X_j)$ for $i < j$, consider how many people are in common for group i and group j. If the number of people in common is 0 or 1 (as shown in the upper and lower left cases in the figure, respectively), then the $\text{Cov}(X_i, X_j) = 0$ since the coin flips used to determine whether Group i is a clique are independent of those used for Group j. If there are 2 people in common (as shown in the lower right case of the figure), then
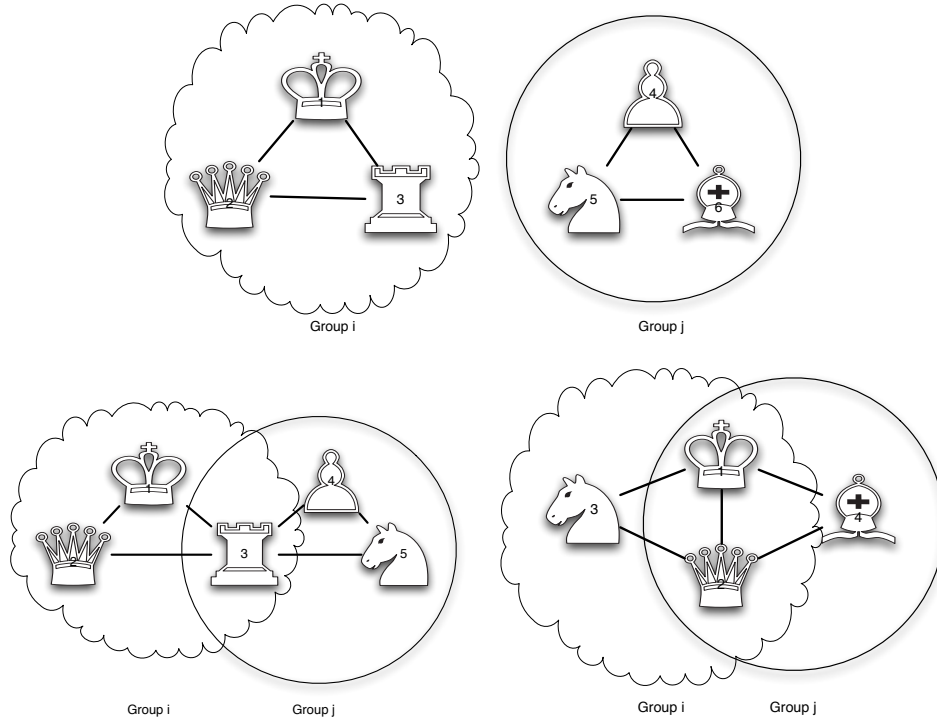
$$\text{Cov}(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j) = \frac{1}{2^5} - \left(\frac{1}{2^3}\right)^2 = \frac{1}{64},$$

since 5 distinct pairs of people must know each other to make $X_i X_j$ equal to 1.

There are $\binom{n}{4}\binom{4}{2} = 6\binom{n}{4}$ pairs of groups $\{i, j\}$ $(i \neq j)$ with 1 pair of people in common (choose 4 people out of the n, then choose which 2 of the 4 are the overlap of the groups). The remaining pairs of groups have covariance 0. Thus, the variance of the number of cliques is

$$\frac{7}{64}\binom{n}{3} + 2 \cdot 6\binom{n}{4} \cdot \frac{1}{64} = \frac{7}{64}\binom{n}{3} + \frac{3}{16}\binom{n}{4}.$$

(c) We will prove the existence of a network with the desired property by showing that

FIGURE 7.1
Two groups with 0 (upper), 1 (lower left), 2 (lower right) people in common.

the probability is positive that a random network has the property is positive (this strategy is explored in the starred Section 4.9). Form a random network as in (a), and let A_i be the event that the ith group of k people (in any fixed ordering) is neither a clique nor an anticlique. We have

$$P(\bigcup_{i=1}^{\binom{n}{k}} A_i^c) \leq \sum_{i=1}^{\binom{n}{k}} P(A_i^c) = \binom{n}{k} 2^{-\binom{k}{2}+1} < 1,$$

which shows that

$$P(\bigcap_{i=1}^{\binom{n}{k}} A_i) = 1 - P(\bigcup_{i=1}^{\binom{n}{k}} A_i^c) > 0,$$

as desired. Alternatively, let C be the number of cliques of size k and A be the number of anticliques of size k, and write $C + A = T$. Then

$$E(T) = E(C) + E(A) = \binom{n}{k} 2^{-\binom{k}{2}+1} < 1,$$

by the method of Part (a). So $P(T = 0) > 0$, since $P(T \geq 1) = 1$ would imply $E(T) \geq 1$. This again shows that there must be a network with the desired property.

89. Ⓢ Shakespeare wrote a total of 884647 words in his known works. Of course, many words are used more than once, and the number of distinct words in Shakespeare's known writings is 31534 (according to one computation). This puts a lower bound on the size of Shakespeare's vocabulary, but it is likely that Shakespeare knew words which he did not use in these known writings.

More specifically, suppose that a new poem of Shakespeare were uncovered, and consider the following (seemingly impossible) problem: give a good prediction of the number of words in the new poem that do not appear anywhere in Shakespeare's previously known works.

Ronald Thisted and Bradley Efron studied this problem in the papers [9] and [10], developing theory and methods and then applying the methods to try to determine whether Shakespeare was the author of a poem discovered by a Shakespearean scholar in 1985. A simplified version of their method is developed in the problem below. The method was originally invented by Alan Turing (the founder of computer science) and I.J. Good as part of the effort to break the German Enigma code during World War II.

Let N be the number of distinct words that Shakespeare knew, and assume these words are numbered from 1 to N. Suppose for simplicity that Shakespeare wrote only two plays, A and B. The plays are reasonably long and they are of the same length. Let X_j be the number of times that word j appears in play A, and Y_j be the number of times it appears in play B, for $1 \leq j \leq N$.

(a) Explain why it is reasonable to model X_j as being Poisson, and Y_j as being Poisson with the same parameter as X_j.

(b) Let the numbers of occurrences of the word "eyeball" (which was coined by Shakespeare) in the two plays be independent $\text{Pois}(\lambda)$ r.v.s. Show that the probability that "eyeball" is used in play B but not in play A is

$$e^{-\lambda}(\lambda - \lambda^2/2! + \lambda^3/3! - \lambda^4/4! + \dots).$$

(c) Now assume that λ from (b) is unknown and is itself taken to be a random variable to reflect this uncertainty. So let λ have a PDF f_0. Let X be the number of times the word "eyeball" appears in play A and Y be the corresponding value for play B. Assume that the conditional distribution of X, Y given λ is that they are independent $\text{Pois}(\lambda)$ r.v.s. Show that the probability that "eyeball" is used in play B but not in play A is the alternating series

$$P(X = 1) - P(X = 2) + P(X = 3) - P(X = 4) + \dots.$$

Hint: Condition on λ and use (b).

(d) Assume that every word's numbers of occurrences in A and B are distributed as in (c), where λ may be different for different words but f_0 is fixed. Let W_j be the number of words that appear exactly j times in play A. Show that the expected number of distinct words appearing in play B but not in play A is

$$E(W_1) - E(W_2) + E(W_3) - E(W_4) + \dots.$$

(This shows that $W_1 - W_2 + W_3 - W_4 + \dots$ is an *unbiased* predictor of the number of distinct words appearing in play B but not in play A: on average it is correct. Moreover, it can be computed just from having seen play A, without needing to know f_0 or any of the λ_j. This method can be extended in various ways to give predictions for unobserved plays based on observed plays.)

Solution:

(a) It is reasonable to model X_j and Y_j as Poisson, because this distribution is used to describe the number of "events" (such as emails received) happening at some average rate in a fixed interval or volume. The Poisson paradigm applies here: each individual word in a play has some very small probability of being word j, and the words are weakly dependent. Here an event means using word j, the average rate is determined by

how frequently Shakespeare uses that word overall. It is reasonable to assume that the average rate of occurrence of a particular word is the same for two plays by the same author, so we take λ to be the same for X_j and Y_j.

(b) Let X be the number of times that "eyeball" is used in play A, and Y be the number of times that it is used in play B. Since X and Y are independent $\text{Pois}(\lambda)$,

$$
\begin{aligned}
P(X = 0, Y > 0) &= P(X = 0)\,(1 - P(Y = 0)) = e^{-\lambda}\left(1 - e^{-\lambda}\right) \\
&= e^{-\lambda}\left(1 - \left(1 - \lambda + \frac{\lambda^2}{2!} - \frac{\lambda^3}{3!} + \frac{\lambda^4}{4!} - \cdots\right)\right) \\
&= e^{-\lambda}\left(\lambda - \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} - \frac{\lambda^4}{4!} + \cdots\right).
\end{aligned}
$$

(c) Now λ is a random variable. Given λ, the calculation from (b) holds. By the law of total probability,

$$
\begin{aligned}
P(X = 0, Y > 0) &= \int_0^\infty P(X = 0, Y > 0 \mid \lambda) f_0(\lambda)\,d\lambda \\
&= \int_0^\infty P(X = 1 \mid \lambda) f_0(\lambda)\,d\lambda - \int_0^\infty P(X = 2 \mid \lambda) f_0(\lambda)\,d\lambda \\
&\quad + \int_0^\infty P(X = 3 \mid \lambda) f_0(\lambda)\,d\lambda - \int_0^\infty P(X = 4 \mid \lambda) f_0(\lambda)\,d\lambda + \ldots \\
&= P(X = 1) - P(X = 2) + P(X = 3) - P(X = 4) + \ldots
\end{aligned}
$$

(d) Let X_j be the number of times word j appears in play A and let W be the number of distinct words that appear in play B but not in A. Then $W = \sum_{j=1}^N I_j$, where I_j is the indicator r.v. of the event that word j appears in play B but not in play A, and N is the total number of words. By (c), for $1 \leq j \leq N$,

$$
EI_j = \sum_{i=1}^\infty (-1)^{i+1} P(X_j = i).
$$

Also, note that the number of words that appear exactly i times in play A is

$$
W_i = I(X_1 = i) + I(X_2 = i) + I(X_3 = i) + \cdots + I(X_N = i),
$$

where $I(X_j = i)$ is the indicator of word j appearing exactly i times in play A. So

$$
EW_i = \sum_{j=1}^N EI(X_j = i) = \sum_{j=1}^N P(X_j = i).
$$

Then

$$
\begin{aligned}
EW &= \sum_{j=1}^N EI_j = \sum_{j=1}^N \sum_{i=1}^\infty (-1)^{i+1} P(X_j = i) \\
&= \sum_{i=1}^\infty (-1)^{i+1} \sum_{j=1}^N P(X_j = i) \\
&= \sum_{i=1}^\infty (-1)^{i+1} EW_i \\
&= EW_1 - EW_2 + EW_3 - EW_4 + \ldots.
\end{aligned}
$$

8

Transformations

Change of variables

1. Find the PDF of e^{-X} for $X \sim \text{Expo}(1)$.

   Solution: Let $Y = e^{-X}$ and $y = e^{-x}$, so $x = -\log y$. By the change of variables formula, the PDF of Y is
   $$f_Y(y) = f_X(-\log y) \cdot \frac{1}{y} = e^{\log y} \cdot \frac{1}{y} = 1$$
   for $0 < y < 1$. That is, $Y \sim \text{Unif}(0, 1)$. We can also see this using universality of the Uniform (which shows that $1 - e^{-X} \sim \text{Unif}(0, 1)$) and the fact that $U \sim \text{Unif}(0, 1)$ implies $1 - U \sim \text{Unif}(0, 1)$.

2. Find the PDF of X^7 for $X \sim \text{Expo}(\lambda)$.

   Solution: Let $Y = X^7$ and $y = x^7$, so $x = y^{1/7}$. By the change of variables formula, the PDF of Y is
   $$f_Y(y) = f_X(y^{1/7}) \cdot \frac{1}{7} y^{-6/7} = \frac{\lambda}{7} e^{-\lambda y^{1/7}} y^{-6/7}$$
   for $y > 0$.

3. Find the PDF of Z^3 for $Z \sim \mathcal{N}(0, 1)$.

   Solution: Let $Y = Z^3$ and $y = z^3$, so $z = y^{1/3}$. By the change of variables formula, the PDF of Y is
   $$f_Y(y) = f_Z(y^{1/3}) \cdot \frac{1}{3} y^{-2/3} = \frac{1}{3\sqrt{2\pi}} y^{-2/3} e^{-\frac{1}{2} y^{2/3}}.$$

4. Ⓢ Find the PDF of Z^4 for $Z \sim \mathcal{N}(0, 1)$.

   Solution: Let $Y = Z^4$. For $y > 0$, the CDF of Y is
   $$P(Y \le y) = P(Z^4 \le y) = P(-y^{1/4} \le Z \le y^{1/4}) = \Phi(y^{1/4}) - \Phi(-y^{1/4}) = 2\Phi(y^{1/4}) - 1.$$
   So the PDF is
   $$f_Y(y) = \frac{2}{4} y^{-3/4} \varphi(y^{1/4}) = \frac{1}{2\sqrt{2\pi}} y^{-3/4} e^{-y^{1/2}/2},$$
   for $y > 0$, where φ is the $\mathcal{N}(0, 1)$ PDF.

5. Find the PDF of $|Z|$ for $Z \sim \mathcal{N}(0, 1)$.

   Solution: Let $Y = |Z|$. The CDF of Y is
   $$P(Y \le y) = P(-y \le Z \le y) = \Phi(y) - \Phi(-y) = 2\Phi(y) - 1,$$
   for $y \ge 0$ (and the CDF is 0 for $y < 0$). So the PDF of Y is
   $$f_Y(y) = \frac{2}{\sqrt{2\pi}} e^{-y^2/2},$$
   for $y \ge 0$.

6. Ⓢ Let $U \sim \text{Unif}(0,1)$. Find the PDFs of U^2 and $\sqrt{U}$.

    Solution:

    (PDF of U^2.) Let $Y = U^2$, $0 < u < 1$, and $y = u^2$, so $u = \sqrt{y}$. The absolute Jacobian determinant is $\left| \frac{du}{dy} \right| = \left| \frac{1}{2\sqrt{y}} \right| = \frac{1}{2\sqrt{y}}$ for $0 < y < 1$. The PDF of Y for $0 < y < 1$ is

    $$f_Y(y) = f_U(u) \left| \frac{du}{dy} \right| = \frac{1}{2\sqrt{y}},$$

    with $f_Y(y) = 0$ otherwise. This is the $\text{Beta}(\frac{1}{2}, 1)$ PDF, so $Y = U^2 \sim \text{Beta}(\frac{1}{2}, 1)$.

    (PDF of $\sqrt{U}$.) Now let $Y = U^{1/2}$, $0 < u < 1$, and $y = u^{1/2}$, so $u = y^2$. The absolute Jacobian determinant is $\left| \frac{du}{dy} \right| = |2y| = 2y$ for $0 < y < 1$. The PDF of Y for $0 < y < 1$ is

    $$f_Y(y) = f_U(u) \left| \frac{du}{dy} \right| = 2y,$$

    with $f_Y(y) = 0$ otherwise. This says that Y has a Beta(2,1) distribution.

    In general, the same method shows that $U^{\frac{1}{\alpha}} \sim \text{Beta}(\alpha, 1)$ for any $\alpha > 0$.

7. Let $U \sim \text{Unif}(0, \frac{\pi}{2})$. Find the PDF of $\sin(U)$.

    Solution: Let $Y = \sin(U)$ and $y = \sin(u)$. Since sin is differentiable everywhere and strictly increasing on $(0, \pi/2)$, the change of variables formula applies, giving

    $$f_Y(y) = \frac{2}{\pi} \left(\frac{dy}{du} \right)^{-1} = \frac{2}{\pi} \left(\cos(\sin^{-1}(y)) \right)^{-1} = \frac{2}{\pi \sqrt{1 - y^2}},$$

    for $0 < y < 1$.

8. (a) Find the distribution of X^2 for $X \sim \text{DUnif}(0, 1, \dots, n)$.

    (b) Find the distribution of X^2 for $X \sim \text{DUnif}(-n, -n+1, \dots, 0, 1, \dots, n)$.

    Solution:

    (a) We have $X^2 \sim \text{DUnif}(0, 1, 4, 9, \dots, n^2)$, since $P(X^2 = j^2) = P(X = j) = 1/(n+1)$ for any $j \in \{0, 1, 2, \dots, n\}$.

    (b) Let $Y = X^2$. The support of Y is $0, 1, 4, 9, \dots, n^2$. The PMF of Y is given by $P(Y = 0) = P(X = 0) = 1/(2n+1)$ and for $j = 1, 2, \dots, n$,

    $$P(Y = j^2) = P(X = -j \text{ or } X = j) = \frac{2}{2n+1}.$$

9. Let $X \sim \text{Bern}(p)$ and let a and b be constants with $a < b$. Find a simple transformation of X that yields an r.v. that equals a with probability $1 - p$ and equals b with probability p.

    Solution: Let's look for a transformation of the form $Y = sX + t$. Then Y is t with probability $1 - p$, and $s + t$ with probability p. Putting $t = a$ and $s + t = b$, we have $s = b - a, t = a$. So $Y = (b - a)X + a$ is as desired.

10. Let $X \sim \text{Pois}(\lambda)$ and Y be the indicator of X being odd. Find the PMF of Y.

    Hint: Find $P(Y = 0) - P(Y = 1)$ by writing $P(Y = 0)$ and $P(Y = 1)$ as series and then using the fact that $(-1)^k$ is 1 if k is even and -1 if k is odd.

Solution: We have

$$P(Y = 0) = e^{-\lambda} \sum_{k \text{ even}} \lambda^k/k!$$

$$P(Y = 1) = e^{-\lambda} \sum_{k \text{ odd}} \lambda^k/k!,$$

so

$$P(Y = 0) - P(Y = 1) = e^{-\lambda} \sum_{k=0}^{\infty} (-1)^k \lambda^k/k! = e^{-\lambda} e^{-\lambda} = e^{-2\lambda}.$$

Of course, we also know that

$$P(Y = 0) + P(Y = 1) = 1.$$

Solving these two equations for $P(Y = 0)$ and $P(Y = 1)$, we have

$$P(Y = 0) = \frac{1 + e^{-2\lambda}}{2} \text{ and } P(Y = 1) = \frac{1 - e^{-2\lambda}}{2}.$$

11. Let T be a continuous r.v. and $V = 1/T$. Show that their CDFs are related as follows:

$$F_V(v) = \begin{cases} F_T(0) + 1 - F_T(\frac{1}{v}) \text{ for } v > 0, \\ F_T(0) \text{ for } v = 0, \\ F_T(0) - F_T(\frac{1}{v}) \text{ for } v < 0. \end{cases}$$

Solution: For $v > 0$, the CDF of V is

$$P(V \le v) = P\left(\frac{1}{T} \le v\right) = P\left(T < 0 \text{ or } T \ge \frac{1}{v}\right) = F_T(0) + 1 - F_T\left(\frac{1}{v}\right).$$

For $v < 0$, the CDF of V is

$$P(V \le v) = P\left(\frac{1}{T} \le v\right) = P\left(\frac{1}{v} \le T < 0\right) = F_T(0) - F_T\left(\frac{1}{v}\right).$$

For $v = 0$, the CDF of V is

$$P(V \le 0) = P(T < 0) = F_T(0).$$

12. Let T be the ratio X/Y of two i.i.d. $\mathcal{N}(0,1)$ r.v.s. X, Y. This is the *Cauchy* distribution and, as shown in Example 7.1.25, it has PDF

$$f_T(t) = \frac{1}{\pi(1 + t^2)}.$$

(a) Use the result of the previous problem to find the CDF of $1/T$. Then use calculus to find the PDF of $1/T$. (Note that the one-dimensional change of variables formula does not apply directly, since the function $g(t) = 1/t$, even though it has $g'(t) < 0$ for all $t \ne 0$, is undefined at $t = 0$ and is not a strictly decreasing function on its domain.)

(b) Show that $1/T$ has the same distribution as T without using calculus, in 140 characters or fewer.

Solution:

(a) Let F and f be the Cauchy CDF and PDF, respectively. Let $V = 1/T$. By the result of the previous problem, the CDF of V is

$$F_V(v) = \begin{cases} F(0) + 1 - F\left(\frac{1}{v}\right) \text{ for } v > 0, \\ F(0) \text{ for } v = 0, \\ F(0) - F\left(\frac{1}{v}\right) \text{ for } v < 0. \end{cases}$$

By the chain rule, the PDF of V is

$$f_V(v) = \frac{1}{v^2} f\left(\frac{1}{v}\right) = \frac{1}{v^2} \frac{1}{\pi(1 + (1/v)^2)} = \frac{1}{\pi(1 + v^2)}$$

for all $v \neq 0$. Thus, V has the same distribution as T.

(b) By symmetry, $\frac{1}{T} = \frac{Y}{X}$ has the same distribution as $\frac{X}{Y}$ since X, Y are i.i.d.

13. Let X and Y be i.i.d. Expo(λ), and $T = \log(X/Y)$. Find the CDF and PDF of T.

Solution: Without loss of generality, we can assume $\lambda = 1$, since the scale cancels out in the ratio X/Y (we can write $X = \tilde{X}/\lambda$ with $\tilde{X} \sim$ Expo(1), and $Y = \tilde{Y}/\lambda$ with $\tilde{Y} \sim$ Expo(1), and then $X/Y = \tilde{X}/\tilde{Y}$ has a distribution that does not depend on λ). The CDF of T is

$$\begin{aligned}
P(T \leq t) &= P(X \leq Ye^t) \\
&= \int_0^\infty e^{-y} \int_0^{ye^t} e^{-x} dx dy \\
&= \int_0^\infty e^{-y}(1 - e^{-ye^t}) dy \\
&= \int_0^\infty e^{-y} dy - \int_0^\infty e^{-y(1+e^t)} dy \\
&= 1 - \frac{1}{1 + e^t} \\
&= \frac{e^t}{1 + e^t},
\end{aligned}$$

for all real t. That is, T has the Logistic distribution (see Example 5.1.6). The PDF is

$$f_T(t) = \frac{e^t}{(1 + e^t)^2}.$$

14. Let X and Y have joint PDF $f_{X,Y}(x, y)$, and transform $(X, Y) \mapsto (T, W)$ linearly by letting

$$T = aX + bY \text{ and } W = cX + dY,$$

where a, b, c, d are constants such that $ad - bc \neq 0$.

(a) Find the joint PDF $f_{T,W}(t, w)$ (in terms of $f_{X,Y}$, though your answer should be written as a function of t and w).

(b) For the case where $T = X + Y, W = X - Y$, show that

$$f_{T,W}(t, w) = \frac{1}{2} f_{X,Y}\left(\frac{t + w}{2}, \frac{t - w}{2}\right).$$

Solution:

(a) Let

$$\begin{aligned}
t &= ax + by, \\
w &= cx + dy.
\end{aligned}$$

Solving this system for x and y (using matrices or by solving the first equation for one of the variables and then plugging into the second equation), we have

$$\begin{aligned}
x &= rdt - rbw, \\
y &= -rct + raw,
\end{aligned}$$

where
$$r = \frac{1}{ad - bc}.$$

The Jacobian matrix is
$$\frac{\partial(x, y)}{\partial(t, w)} = \begin{pmatrix} rd & -rb \\ -rc & ra \end{pmatrix},$$

which has determinant $r^2 ad - r^2 bc = r^2(1/r) = r$. By the change of variables formula,
$$f_{T,W}(t, w) = f_{X,Y}(rdt - rbw, -rct + raw)|r|.$$

(b) Now let $a = b = c = 1$ and $d = -1$. Then $r = 1/(ad - bc) = -1/2$, and the result from Part (a) reduces to

$$f_{T,W}(t, w) = \frac{1}{2} f_{X,Y}\left(\frac{t + w}{2}, \frac{t - w}{2}\right).$$

15. Ⓢ Let X, Y be continuous r.v.s with a *spherically symmetric* joint distribution, which means that the joint PDF is of the form $f(x, y) = g(x^2 + y^2)$ for some function g. Let (R, θ) be the polar coordinates of (X, Y), so $R^2 = X^2 + Y^2$ is the squared distance from the origin and θ is the angle (in $[0, 2\pi)$), with $X = R\cos\theta, Y = R\sin\theta$.

(a) Explain intuitively why R and θ are independent. Then prove this by finding the joint PDF of (R, θ).

(b) What is the joint PDF of (R, θ) when (X, Y) is Uniform in the unit disk $\{(x, y) : x^2 + y^2 \leq 1\}$?

(c) What is the joint PDF of (R, θ) when X and Y are i.i.d. $\mathcal{N}(0, 1)$?

Solution:

(a) Intuitively, this makes sense since the joint PDF of X, Y at a point (x, y) only depends on the distance from (x, y) to the origin, not on the angle, so knowing R gives no information about θ. The absolute Jacobian determinant is r (as shown in the math appendix), so
$$f_{R,\theta}(r, t) = f_{X,Y}(x, y)r = r \cdot g(r^2)$$

for all $r \geq 0, t \in [0, 2\pi)$. This factors as a function of r times a (constant) function of t, so R and θ are independent with $\theta \sim \text{Unif}(0, 2\pi)$.

(b) We have $f_{X,Y}(x, y) = \frac{1}{\pi}$ for $x^2 + y^2 \leq 1$, so $f_{R,\theta}(r, t) = \frac{r}{\pi}$ for $0 \leq r \leq 1, t \in [0, 2\pi)$ (and the PDF is 0 otherwise). This says that R and θ are independent with marginal PDFs $f_R(r) = 2r$ for $0 \leq r \leq 1$ and $f_\theta(t) = \frac{1}{2\pi}$ for $0 \leq t < 2\pi$.

(c) The joint PDF of X, Y is $\frac{1}{2\pi}e^{-(x^2+y^2)/2}$, so $g(r^2) = \frac{1}{2\pi}e^{-r^2/2}$ and the joint PDF of (R, θ) is $\frac{1}{2\pi}re^{-r^2/2}$. This says that R and θ are independent with marginal PDFs $f_R(r) = re^{-r^2/2}$ for $r \geq 0$ and $f_\theta(t) = \frac{1}{2\pi}$ for $0 \leq t < 2\pi$. (The distribution of R is an example of a *Weibull*; note that it is the distribution of $W^{1/2}$ for $W \sim \text{Expo}(1/2)$.)

16. Let X and Y be i.i.d. $\mathcal{N}(0, 1)$ r.v.s, $T = X + Y$, and $W = X - Y$. We know from Example 7.5.8 that T and W are independent $\mathcal{N}(0, 2)$ r.v.s (note that (T, W) is Multivariate Normal with $\text{Cov}(T, W) = 0$). Give another proof of this fact, using the change of variables theorem.

Solution: Let $t = x + y, w = x - y$, so $x = (t + w)/2, y = (t - w)/2$. The Jacobian matrix is
$$\frac{\partial(x, y)}{\partial(t, w)} = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & -1/2 \end{pmatrix},$$

which has absolute determinant $1/2$. By the change of variables formula, the joint PDF of T and W is

$$\begin{aligned} f_{T,W}(t,w) &= f_{X,Y}\left(\frac{t+w}{2}, \frac{t-w}{2}\right) \cdot \frac{1}{2} \\ &= \frac{1}{4\pi} \exp\left(-(t+w)^2/8 - (t-w)^2/8\right) \\ &= \frac{1}{4\pi} \exp\left(-t^2/4\right) \exp\left(-w^2/4\right), \end{aligned}$$

which shows that T and W are i.i.d. $\mathcal{N}(0,2)$.

17. Let X and Y be i.i.d. $\mathcal{N}(0,1)$ r.v.s, and (R,θ) be the polar coordinates for the point (X,Y), so $X = R\cos\theta$ and $Y = R\sin\theta$ with $R \geq 0$ and $\theta \in [0,2\pi)$. Find the joint PDF of R^2 and θ. Also find the marginal distributions of R^2 and θ, giving their names (and parameters) if they are distributions we have studied before.

    Solution: We have $X = R\cos(\theta), Y = R\sin(\theta)$. Let $W = R^2, T = \theta$ and mirror the relationships between the capital letters via $w = r^2 = x^2 + y^2, x = \sqrt{w}\cos(t), y = \sqrt{w}\sin(t)$. By the change of variables formula,

$$f_{W,T}(w,t) = f_{X,Y}(x,y)\left|\frac{\partial(x,y)}{\partial(w,t)}\right| = \frac{e^{-r^2/2}}{2\pi}\left|\frac{\partial(x,y)}{\partial(w,t)}\right|.$$

The Jacobian matrix is

$$\frac{\partial(x,y)}{\partial(w,t)} = \begin{pmatrix} \frac{1}{2\sqrt{w}}\cos(t) & -\sqrt{w}\sin(t) \\ \frac{1}{2\sqrt{w}}\sin(t) & \sqrt{w}\cos(t) \end{pmatrix},$$

which has absolute determinant $\frac{1}{2}\cos^2(t) + \frac{1}{2}\sin^2(t) = \frac{1}{2}$. So the joint PDF of R^2 and θ is

$$f_{R^2,\theta}(w,t) = \frac{1}{4\pi}e^{-w/2} = \frac{1}{2\pi} \cdot \frac{1}{2}e^{-w/2},$$

for $w > 0$ and $0 \leq t < 2\pi$ (and 0 otherwise). Thus, R^2 and θ are independent, with

$$R^2 \sim \text{Expo}(1/2) \text{ and } \theta \sim \text{Unif}(0,2\pi).$$

Note that this problem is Example 8.1.9 (Box-Muller) in reverse.

18. Let X and Y be independent positive r.v.s, with PDFs f_X and f_Y respectively. Let T be the ratio X/Y.

    (a) Find the joint PDF of T and X, using a Jacobian.

    (b) Find the marginal PDF of T, as a single integral.

    Solution: Let $t = x/y$ and $w = x$, so $x = w$ and $y = w/t$.

    (a) The Jacobian matrix is

$$\frac{\partial(x,y)}{\partial(t,w)} = \begin{pmatrix} 0 & 1 \\ -w/t^2 & 1/t \end{pmatrix},$$

which has absolute determinant w/t^2. By the change of variables formula,

$$f_{T,W}(t,w) = f_X(x)f_Y(y) \cdot \frac{w}{t^2} = f_X(w)f_Y\left(\frac{w}{t}\right) \cdot \frac{w}{t^2},$$

for $t > 0, w > 0$. So the joint PDF of T and X is

$$f_{T,X}(t,x) = f_X(x)f_Y\left(\frac{x}{t}\right) \cdot \frac{x}{t^2},$$

for $t > 0, x > 0$.

(b) Marginalizing out X, the marginal PDF of T is

$$f_T(t) = \int_0^\infty f_X(x) f_Y\left(\frac{x}{t}\right) \cdot \frac{x}{t^2} dx.$$

19. Let X and Y be i.i.d. Expo(λ), and transform them to $T = X + Y, W = X/Y$.

(a) Find the joint PDF of T and W. Are they independent?

(b) Find the marginal PDFs of T and W.

Solution:

(a) We can use the change of variables formula, but it is faster to relate this problem to the bank–post office story. Let $U = X/(X + Y)$. By the bank–post office story, T and U are independent, with $T \sim \text{Gamma}(2, \lambda)$ and $U \sim \text{Unif}(0, 1)$. But

$$W = \frac{X/(X + Y)}{Y/(X + Y)} = \frac{U}{1 - U}$$

is a function of U. So T and W are independent. The CDF of W is

$$P(W \le w) = P(U \le w/(w + 1)) = w/(w + 1),$$

for $w > 0$ (and 0 for $w \le 0$). So the PDF of W is

$$f_W(w) = \frac{(w + 1) - w}{(w + 1)^2} = \frac{1}{(w + 1)^2},$$

for $w > 0$. Since T and W are independent, their joint PDF is

$$f_{T,W}(t, w) = (\lambda t)^2 e^{-\lambda t} \frac{1}{t} \cdot \frac{1}{(w + 1)^2},$$

for $t > 0, w > 0$.

(b) As shown in (a), the marginal PDF of T is

$$f_T(t) = (\lambda t)^2 e^{-\lambda t} \frac{1}{t},$$

for $t > 0$, and the marginal PDF of W is

$$f_W(w) = \frac{1}{(w + 1)^2},$$

for $w > 0$.

Convolutions

20. Let $U \sim \text{Unif}(0, 1)$ and $X \sim \text{Expo}(1)$, independently. Find the PDF of $U + X$.

Solution: Let $T = U + X$. For $0 < t \le 1$, the PDF of T is

$$f_T(t) = \int_0^\infty f_X(x) f_U(t - x) dx = \int_0^t e^{-x} dx = 1 - e^{-t},$$

where the limits of integration come from the fact that we need $0 < t - x < 1$ in order for $f_U(t - x)$ to be positive. For $t > 1$, the PDF of T is

$$f_T(t) = \int_{t-1}^t e^{-x} dx = e^{-t+1} - e^{-t}.$$

21. Let X and Y be i.i.d. Expo(1). Use a convolution integral to show that the PDF of $L = X - Y$ is $f(t) = \frac{1}{2}e^{-|t|}$ for all real t; this is known as the *Laplace distribution*.

    Solution: Let $Z = -Y$. The PDF of Z is $f_Z(z) = f_Y(-z) = e^z$, for $z < 0$. So the PDF of L is

    $$f(t) = \int_0^\infty f_X(x)f_Z(t-x)dx = \int_{\max(t,0)}^\infty e^{-x}e^{t-x}dx = e^t \int_{\max(t,0)}^\infty e^{-2x}dx = \frac{e^{t-2\max(t,0)}}{2} = \frac{e^{-|t|}}{2},$$

    since if $t \geq 0$ then $t - 2\max(t,0) = t - 2t = -t = -|t|$, and if $t < 0$ then again $t - 2\max(t,0) = t - 0 = t = -|t|$. To get the limits of integration, we used the fact that we need $x > 0$ and $x > t$ in order to make the PDFs of both X and Z positive; alternatively, we could have considered the cases $t \geq 0$ and $t < 0$ separately when setting up the convolution integral.

22. Use a convolution integral to show that if $X \sim \mathcal{N}(\mu_1, \sigma^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma^2)$ are independent, then $T = X + Y \sim \mathcal{N}(\mu_1 + \mu_2, 2\sigma^2)$ (to simplify the calculation, we are assuming that the variances are equal). You can use a standardization (location-scale) idea to reduce to the standard Normal case before setting up the integral.

    Hint: Complete the square.

    Solution: Write $X = \mu_1 + \sigma Z, Y = \mu_2 + \sigma W$ where Z and W are i.i.d. $\mathcal{N}(0,1)$. Let $V = Z + W$. Then $X + Y = (\mu_1 + \mu_2) + \sigma V$, so it suffices to show $V \sim \mathcal{N}(0,2)$. The PDF of V is

    $$f_V(v) = \int_{-\infty}^\infty f_Z(z)f_W(v-z)dz = \int_{-\infty}^\infty f_Z(z)f_W(v-z)dz = \frac{1}{2\pi}\int_{-\infty}^\infty e^{-z^2/2-(v-z)^2/2}dz.$$

    Expanding the exponent and then completing the square, this becomes

    $$\frac{1}{2\pi}\int_{-\infty}^\infty e^{-z^2-v^2/2+vz}dz = \frac{e^{-v^2/2}}{2\pi}\int_{-\infty}^\infty e^{-z^2+vz}dz$$

    $$= \frac{e^{-v^2/2}}{2\pi}\int_{-\infty}^\infty e^{-(z-v/2)^2+v^2/4}dz$$

    $$= \frac{e^{-v^2/4}}{2\sqrt{\pi}}\int_{-\infty}^\infty \frac{1}{\sqrt{\pi}}e^{-(z-v/2)^2}dz$$

    $$= \frac{e^{-v^2/4}}{2\sqrt{\pi}},$$

    since the last integral above is the integral of the $\mathcal{N}(v/2, 1/2)$ PDF. So $V \sim \mathcal{N}(0,2)$.

23. Ⓢ Let X and Y be independent positive r.v.s, with PDFs f_X and f_Y respectively, and consider the product $T = XY$. When asked to find the PDF of T, Jacobno argues that "it's like a convolution, with a product instead of a sum. To have $T = t$ we need $X = x$ and $Y = t/x$ for some x; that has probability $f_X(x)f_Y(t/x)$, so summing up these possibilities we get that the PDF of T is $\int_0^\infty f_X(x)f_Y(t/x)dx$." Evaluate Jacobno's argument, while getting the PDF of T (as an integral) in 2 ways:

    (a) using the continuous version of the law of total probability to get the CDF, and then taking the derivative (you can assume that swapping the derivative and integral is valid);

    (b) by taking the log of both sides of $T = XY$ and doing a convolution (and then converting back to get the PDF of T).

    Solution:

(a) By the law of total probability (conditioning on X),

$$P(T \le t) = \int_0^\infty P(XY \le t | X = x) f_X(x) dx$$

$$= \int_0^\infty P(Y \le t/x | X = x) f_X(x) dx$$

$$= \int_0^\infty F_Y(t/x) f_X(x) dx,$$

which has derivative

$$f_T(t) = \int_0^\infty f_X(x) f_Y(t/x) \frac{dx}{x}.$$

This is *not* the same as Jacobno claimed: there is an extra x in the denominator. This stems from the fact that the transformation (X, Y) to (XY, X) is nonlinear, in contrast to the transformation (X, Y) to $(X + Y, X)$ considered in Theorem 8.2.1. Jacobno is ignoring the distinction between probabilities and probability *densities*, and is implicitly (and incorrectly) assuming that there is no Jacobian term.

(b) Let $Z = \log(T), W = \log(X), V = \log(Y)$, so $Z = W + V$. The PDF of Z is

$$f_Z(z) = \int_{-\infty}^\infty f_W(w) f_V(z - w) dw,$$

where by change of variables $f_W(w) = f_X(e^w) e^w, f_V(v) = f_Y(e^v) e^v$. So

$$f_Z(z) = \int_{-\infty}^\infty f_X(e^w) e^w f_Y(e^{z-w}) e^{z-w} dw = e^z \int_{-\infty}^\infty f_X(e^w) f_Y(e^{z-w}) dw.$$

Transforming back to T, we have

$$f_T(t) = f_Z(\log t) \frac{1}{t} = \int_{-\infty}^\infty f_X(e^w) f_Y(e^{\log(t)-w}) dw = \int_0^\infty f_X(x) f_Y(t/x) \frac{dx}{x},$$

letting $x = e^w$. This concurs with (a): Jacobno is missing the x in the denominator.

24. Let X and Y be i.i.d. Discrete Uniform r.v.s on $\{0, 1, \ldots, n\}$, where n is a positive integer. Find the PMF of $T = X + Y$.

Hint: In finding $P(T = k)$, it helps to consider the cases $0 \le k \le n$ and $n + 1 \le k \le 2n$ separately. Be careful about the range of summation in the convolution sum.

Solution: We have

$$P(T = k) = \sum_{j=0}^n P(X = j) P(Y = k - j),$$

where each term is either 0 or $1/(n + 1)^2$. A term is nonzero if and only if $0 \le j \le n$ and $0 \le k - j \le n$. Equivalently, a term is nonzero if and only if

$$\max(0, k - n) \le j \le \min(k, n).$$

For $0 \le k \le n$, this range becomes $0 \le j \le k$, so

$$P(T = k) = \frac{k + 1}{(n + 1)^2}.$$

For $n + 1 \le k \le 2n$, the range becomes $k - n \le j \le n$, so

$$P(T = k) = \frac{2n - k + 1}{(n + 1)^2}.$$

25. Let X and Y be i.i.d. Unif$(0, 1)$, and let $W = X - Y$.

(a) Find the mean and variance of W, without yet deriving the PDF.

(b) Show that the distribution of W is symmetric about 0, without yet deriving the PDF.

(c) Find the PDF of W.

(d) Use the PDF of W to verify your results from (a) and (b).

(e) How does the distribution of W relate to the distribution of $X + Y$, the Triangle distribution derived in Example 8.2.5? Give a precise description, e.g., using the concepts of location and scale.

Solution:

(a) We have

$$E(W) = E(X) - E(Y) = 0,$$
$$\text{Var}(W) = \text{Var}(X) + \text{Var}(Y) = \frac{1}{6}.$$

(b) Since X and Y are i.i.d., $-W = Y - X$ has the same distribution as $W = X - Y$ (it has the exact same structure), so the distribution of W is symmetric about 0.

(c) We can write $W = X + Z$ with $Z \sim \text{Unif}(-1, 0)$ and set up a convolution integral. Instead though, let's relate W to the result of Example 8.2.5. Using the symmetry property that $U \sim \text{Unif}(0, 1)$ implies $1 - U \sim \text{Unif}(0, 1)$, W has the same distribution as $X - (1 - Y) = X + Y - 1 = T - 1$, where $T = X + Y$ has the Triangle distribution derived in Example 8.2.5. Let $w = t - 1$, so $t = w + 1$. By the change of variables formula,

$$f_W(w) = f_T(t)\frac{dt}{dw} = f_T(w + 1) = \begin{cases} 1 + w, & \text{if } -1 < w \leq 0; \\ 1 - w, & \text{if } 0 < w < 1; \\ 0, & \text{otherwise.} \end{cases}$$

(d) Using the above PDF,

$$E(W) = \int_{-1}^{0} w(1 + w)dw + \int_{0}^{1} w(1 - w)dw = \left(\frac{w^2}{2} + \frac{w^3}{3}\right)\Big|_{-1}^{0} + \left(\frac{w^2}{2} - \frac{w^3}{3}\right)\Big|_{0}^{1} = 0,$$
$$E(W^2) = \int_{-1}^{0} w^2(1 + w)dw + \int_{0}^{1} w^2(1 - w)dw = \left(\frac{w^3}{3} + \frac{w^4}{4}\right)\Big|_{-1}^{0} + \left(\frac{w^3}{3} - \frac{w^4}{4}\right)\Big|_{0}^{1} = \frac{1}{6},$$
$$\text{Var}(W) = E(W^2) - (EW)^2 = \frac{1}{6}.$$

(e) The distribution of W is a shifted version of the distribution of $X + Y$ (a change in location). Shifting the PDF of W one unit to the right, so that it is centered at 1 rather than 0, yields the Triangle distribution from Example 8.2.5.

26. Let X and Y be i.i.d. Unif$(0, 1)$, and $T = X + Y$. We derived the distribution of T (a Triangle distribution) in Example 8.2.5, using a convolution integral. Since (X, Y) is Uniform in the unit square $\{(x, y) : 0 < x < 1, 0 < y < 1\}$, we can also interpret $P((X, Y) \in A)$ as the *area* of A, for any region A within the unit square. Use this idea to find the CDF of T, by interpreting the CDF (evaluated at some point) as an area.

Solution: Let F be the CDF of T. We can interpret $F(t) = P(T \leq t)$ as the area of the region below the line $x + y = t$ that lies in the unit square.

For $t \leq 0$, $F(t) = 0$ since then the region is empty.

Now let $0 < t < 1$. Then the region is the triangle with vertices $(0,0), (0,t), (t,0)$. The area of this triangle is $t^2/2$, so $F(t) = t^2/2$.

Next, let $1 \leq t < 2$. Then the region has area 1 minus the area of the triangle with vertices $(t-1, 1), (1, 1), (1, t-1)$, so $F(t) = 1 - (2-t)^2/2$.

For $t \geq 2$, the region is the entire unit square, so $F(t) = 1$.

Note that taking the derivative of F does yield the PDF derived in Example 8.2.5.

27. Let X, Y, Z be i.i.d. Unif$(0, 1)$, and $W = X + Y + Z$. Find the PDF of W.

Hint: We already know the PDF of $X + Y$. Be careful about limits of integration in the convolution integral; there are 3 cases that should be considered separately.

Solution: Let $T = X + Y$. As shown in Example 8.2.5, T has a triangle-shaped density:

$$f_T(t) = \begin{cases} t, & \text{if } 0 < t \leq 1; \\ 2 - t, & \text{if } 1 < t < 2; \\ 0, & \text{otherwise.} \end{cases}$$

We will find the PDF of $W = T + Z$ using a convolution:

$$f_W(w) = \int_{-\infty}^{\infty} f_T(t) f_Z(w - t) dt = \int_{w-1}^{w} f_T(t) dt,$$

since $f_Z(w - t)$ is 0 except when $0 < w - t < 1$, which is equivalent to $t > w - 1, t < w$). Consider the 3 cases $0 < w < 1, 1 < w < 2$, and $2 < w < 3$ separately.

Case 1: $0 < w < 1$.

In this case,

$$\int_{w-1}^{w} f_T(t) dt = \int_0^w t \, dt = \frac{w^2}{2}.$$

Case 2: $1 < w < 2$.

In this case,

$$\int_{w-1}^{w} f_T(t) dt = \int_{w-1}^{1} t \, dt + \int_1^w (2 - t) dt = -w^2 + 3w - \frac{3}{2}.$$

Case 3: $2 < w < 3$.

In this case,

$$\int_{w-1}^{w} f_T(t) dt = \int_{w-1}^{2} (2 - t) dt = \frac{w^2 - 6w + 9}{2}.$$

Thus, the PDF of W is the piecewise quadratic function

$$f_W(w) = \begin{cases} \frac{w^2}{2}, & \text{if } 0 < w \leq 1; \\ -w^2 + 3w - \frac{3}{2}, & \text{if } 1 < w \leq 2; \\ \frac{(w-3)^2}{2}, & \text{if } 2 < w < 3; \\ 0, & \text{otherwise.} \end{cases}$$

Beta and Gamma

28. ⓢ Let $B \sim \text{Beta}(a, b)$. Find the distribution of $1 - B$ in two ways: (a) using a change of variables and (b) using a story proof. Also explain why the result makes sense in terms of Beta being the conjugate prior for the Binomial.

    Solution:

    (a) Let $W = 1 - B$. The function $g(t) = 1 - t$ is strictly decreasing with absolute derivative $|-1| = 1$, so the PDF of W is

    $$f_W(w) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)}(1 - w)^{a-1}w^{b-1},$$

    for $0 < w < 1$, which shows that $W \sim \text{Beta}(b, a)$.

    (b) Using the bank–post office story, we can represent $B = \frac{X}{X+Y}$ with $X \sim \text{Gamma}(a, 1)$ and $Y \sim \text{Gamma}(b, 1)$ independent. Then $1 - B = \frac{Y}{X+Y} \sim \text{Beta}(b, a)$ by the same story.

    This result makes sense intuitively since if we use $\text{Beta}(a, b)$ as the prior distribution for the probability p of success in a Binomial problem, interpreting a as the number of prior successes and b as the number of prior failures, then $1 - p$ is the probability of failure and, interchanging the roles of "success" and "failure," it makes sense to have $1 - p \sim \text{Beta}(b, a)$.

29. ⓢ Let $X \sim \text{Gamma}(a, \lambda)$ and $Y \sim \text{Gamma}(b, \lambda)$ be independent, with a and b integers. Show that $X + Y \sim \text{Gamma}(a + b, \lambda)$ in three ways: (a) with a convolution integral; (b) with MGFs; (c) with a story proof.

    Solution:

    (a) The convolution integral is

    $$f_{X+Y}(t) = \int_{-\infty}^{\infty} f_X(x)f_Y(t-x)dx = \int_0^t \frac{1}{\Gamma(a)}\frac{1}{\Gamma(b)}(\lambda x)^a(\lambda(t-x))^b e^{-\lambda x}e^{-\lambda(t-x)}\frac{1}{x}\frac{1}{t-x}dx,$$

    where we integrate from 0 to t since we need $x > 0$ and $t - x > 0$. This is

    $$\lambda^{a+b}\frac{e^{-\lambda t}}{\Gamma(a)\Gamma(b)}\int_0^t x^{a-1}(t-x)^{b-1}dx = \lambda^{a+b}\frac{e^{-\lambda t}}{\Gamma(a)\Gamma(b)}\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}t^{a+b-1} = \frac{1}{\Gamma(a+b)}(\lambda t)^{a+b}e^{-\lambda t}\frac{1}{t},$$

    using a Beta integral (after letting $u = x/t$ so that we can integrate from 0 to 1 rather than 0 to t). Thus, $X + Y \sim \text{Gamma}(a + b, \lambda)$.

    (b) The MGF of $X + Y$ is $M_X(t)M_Y(t) = \frac{\lambda^a}{(\lambda-t)^a}\frac{\lambda^b}{(\lambda-t)^b} = \frac{\lambda^{a+b}}{(\lambda-t)^{a+b}} = M_{X+Y}(t)$, which again shows that $X + Y \sim \text{Gamma}(a + b, \lambda)$.

    (c) Interpret X as the time of the ath arrival in a Poisson process with rate λ, and Y as the time needed for b more arrivals to occur (which is independent of X since the times between arrivals are independent $\text{Expo}(\lambda)$ r.v.s). Then $X + Y$ is the time of the $(a + b)$th arrival, so $X + Y \sim \text{Gamma}(a + b, \lambda)$.

30. Let $B \sim \text{Beta}(a, b)$. Use integration by pattern recognition to find $E(B^k)$ for positive integers k. In particular, show that

    $$\text{Var}(B) = \frac{ab}{(a+b)^2(a+b+1)}.$$

    Solution: By LOTUS,

    $$E(B^k) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^1 x^{a+k-1}(1-x)^{b-1}dx = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(a+k)\Gamma(b)}{\Gamma(a+k+b)} = \frac{\Gamma(a+b)}{\Gamma(a)}\frac{\Gamma(a+k)}{\Gamma(a+k+b)},$$

since the integrand is an unnormalized Beta$(a + k, b)$ PDF. In particular,

$$E(B^2) = \frac{\Gamma(a + b)\Gamma(a + 2)}{\Gamma(a)\Gamma(a + b + 2)} = \frac{(a + 1)a}{(a + b + 1)(a + b)},$$

since $\Gamma(x + 1) = x\Gamma(x)$ and $\Gamma(x + 2) = (x + 1)\Gamma(x + 1) = (x + 1)x\Gamma(x)$. Hence,

$$\text{Var}(B) = \frac{(a + 1)a}{(a + b + 1)(a + b)} - \frac{a^2}{(a + b)^2} = \frac{ab}{(a + b)^2(a + b + 1)}.$$

31. (S) Fred waits $X \sim \text{Gamma}(a, \lambda)$ minutes for the bus to work, and then waits $Y \sim \text{Gamma}(b, \lambda)$ for the bus going home, with X and Y independent. Is the ratio X/Y independent of the total wait time $X + Y$?

Solution: As shown in the bank–post office story, $W = \frac{X}{X+Y}$ is independent of $X + Y$. So any function of W is independent of any function of $X + Y$. And we have that X/Y is a function of W, since

$$\frac{X}{Y} = \frac{\frac{X}{X+Y}}{\frac{Y}{X+Y}} = \frac{W}{1 - W},$$

so X/Y is independent of $X + Y$.

32. (S) The F-test is a very widely used statistical test based on the $F(m, n)$ distribution, which is the distribution of $\frac{X/m}{Y/n}$ with $X \sim \text{Gamma}(\frac{m}{2}, \frac{1}{2}), Y \sim \text{Gamma}(\frac{n}{2}, \frac{1}{2})$. Find the distribution of $mV/(n + mV)$ for $V \sim F(m, n)$.

Solution: Let $X \sim \text{Gamma}(\frac{m}{2}, \frac{1}{2}), Y \sim \text{Gamma}(\frac{n}{2}, \frac{1}{2})$, and $V = \frac{n}{m}\frac{X}{Y}$. Then

$$mV/(n + mV) = \frac{nX/Y}{n + nX/Y} = \frac{X}{X + Y} \sim \text{Beta}\left(\frac{m}{2}, \frac{n}{2}\right).$$

33. (S) Customers arrive at the Leftorium store according to a Poisson process with rate λ customers per hour. The true value of λ is unknown, so we treat it as a random variable. Suppose that our prior beliefs about λ can be expressed as $\lambda \sim \text{Expo}(3)$. Let X be the number of customers who arrive at the Leftorium between 1 pm and 3 pm tomorrow. Given that $X = 2$ is observed, find the posterior PDF of λ.

Solution: It follows from Story 8.4.5 (Gamma-Poisson conjugacy) that the posterior distribution of λ given the data is Gamma$(3, 5)$. Equivalently, we can use Bayes' rule directly. Writing f_0 for the prior PDF and f_1 for the posterior PDF, we have

$$f_1(\lambda|x) = \frac{P(X = x|\lambda)f_0(\lambda)}{P(X = x)},$$

where $f_0(\lambda) = 3e^{-3\lambda}$ for $\lambda > 0$, and $P(X = x|\lambda)$ is obtained from the Pois(2λ) PMF. For $x = 2$, the numerator is

$$\frac{e^{-2\lambda}(2\lambda)^2}{2!} \cdot 3e^{-3\lambda} = 6\lambda^2 e^{-5\lambda}.$$

The denominator does not depend on λ, so it serves as a normalizing constant for the posterior PDF. So the posterior PDF is proportional to $\lambda^2 e^{-5\lambda}$, which shows that the posterior distribution is Gamma$(3, 5)$. Including the normalizing constant for the Gamma distribution, we have

$$f_1(\lambda|2) = \frac{5^3}{\Gamma(3)}\lambda^2 e^{-5\lambda} = \frac{125}{2}\lambda^2 e^{-5\lambda},$$

for $\lambda > 0$.

34. ⓢ Let X and Y be independent, positive r.v.s. with finite expected values.

(a) Give an example where $E(\frac{X}{X+Y}) \neq \frac{E(X)}{E(X+Y)}$, computing both sides exactly.
Hint: Start by thinking about the simplest examples you can think of!

(b) If X and Y are i.i.d., then is it necessarily true that $E(\frac{X}{X+Y}) = \frac{E(X)}{E(X+Y)}$?

(c) Now let $X \sim \text{Gamma}(a, \lambda)$ and $Y \sim \text{Gamma}(b, \lambda)$. Show *without using calculus* that

$$E\left(\frac{X^c}{(X+Y)^c}\right) = \frac{E(X^c)}{E((X+Y)^c)}$$

for every real $c > 0$.

Solution:

(a) As a simple example, let X take on the values 1 and 3 with probability $1/2$ each, and let Y take on the values 3 and 5 with probability $1/2$ each. Then $E(X)/E(X+Y) = 2/(2+4) = 1/3$, but $E(X/(X+Y)) = 31/96$ (the average of the 4 possible values of $X/(X+Y)$, which are equally likely). An even simpler example is to let X be the constant 1 (a degenerate r.v.), and let Y be 1 or 3 with probability $1/2$ each. Then $E(X)/E(X+Y) = 1/(1+2) = 1/3$, but $E(X/(X+Y)) = 3/8$.

(b) Yes, since by symmetry $E(\frac{X}{X+Y}) = E(\frac{Y}{X+Y})$ and by linearity

$$E\left(\frac{X}{X+Y}\right) + E\left(\frac{Y}{X+Y}\right) = E\left(\frac{X+Y}{X+Y}\right) = 1,$$

so $E(\frac{X}{X+Y}) = 1/2$, while on the other hand

$$\frac{E(X)}{E(X+Y)} = \frac{E(X)}{E(X) + E(Y)} = \frac{E(X)}{E(X) + E(X)} = 1/2.$$

(c) The equation we need to show can be paraphrased as the statement that $X^c/(X+Y)^c$ and $(X+Y)^c$ are uncorrelated. By the bank–post office story, $X/(X+Y)$ is independent of $X + Y$. So $X^c/(X + Y)^c$ is independent of $(X + Y)^c$, which shows that they are uncorrelated.

35. Let $T = X^{1/\gamma}$, with $X \sim \text{Expo}(\lambda)$ and $\lambda, \gamma > 0$. So T has a Weibull distribution, as discussed in Example 6.5.5. Using LOTUS and the definition of the gamma function, we showed in this chapter that for $Y \sim \text{Gamma}(a, \lambda)$,

$$E(Y^c) = \frac{1}{\lambda^c} \cdot \frac{\Gamma(a + c)}{\Gamma(a)},$$

for all real $c > -a$. Use this result to show that

$$E(T) = \frac{\Gamma(1 + 1/\gamma)}{\lambda^{1/\gamma}} \text{ and } \text{Var}(T) = \frac{\Gamma(1 + 2/\gamma) - (\Gamma(1 + 1/\gamma))^2}{\lambda^{2/\gamma}}.$$

Solution: Since $X \sim \text{Gamma}(1, \lambda)$, we have

$$E(T) = E(X^{1/\gamma}) = \frac{1}{\lambda^{1/\gamma}} \cdot \frac{\Gamma(1 + 1/\gamma)}{\Gamma(1)} = \frac{\Gamma(1 + 1/\gamma)}{\lambda^{1/\gamma}},$$

$$E(T^2) = E(X^{2/\gamma}) = \frac{\Gamma(1 + 2/\gamma)}{\lambda^{2/\gamma}}.$$

Then

$$\text{Var}(T) = \frac{\Gamma(1 + 2/\gamma)}{\lambda^{2/\gamma}} - \left(\frac{\Gamma(1 + 1/\gamma)}{\lambda^{1/\gamma}}\right)^2 = \frac{\Gamma(1 + 2/\gamma) - (\Gamma(1 + 1/\gamma))^2}{\lambda^{2/\gamma}}.$$

36. Alice walks into a post office with 2 clerks. Both clerks are in the midst of serving customers, but Alice is next in line. The clerk on the left takes an Expo(λ_1) time to serve a customer, and the clerk on the right takes an Expo(λ_2) time to serve a customer. Let T_1 be the time until the clerk on the left is done serving their current customer, and define T_2 likewise for the clerk on the right.

(a) If $\lambda_1 = \lambda_2$, is T_1/T_2 independent of $T_1 + T_2$?

Hint: $T_1/T_2 = (T_1/(T_1 + T_2))/(T_2/(T_1 + T_2))$.

(b) Find $P(T_1 < T_2)$ (do not assume $\lambda_1 = \lambda_2$ here or in the next part, but do check that your answers make sense in that special case).

(c) Find the expected total amount of time that Alice spends in the post office (assuming that she leaves immediately after she is done being served).

Solution:

(a) Let $W = T_1/(T_1 + T_2)$. Then $T_1/T_2 = W/(1 - W)$. By the bank–post office story, W is independent of $T_1 + T_2$. So T_1/T_2 is also independent of $T_1 + T_2$.

(b) It was shown in Example 7.1.24 that $P(T_1 < T_2) = \lambda_1/(\lambda_1 + \lambda_2)$, using a double integral. In the special case $\lambda_1 = \lambda_2$, this gives $P(T_1 < T_2) = 1/2$, which must be true by symmetry.

Another way to derive this result is to apply the bank–post office story. This story requires two Gamma r.v.s with the same rate parameter λ, so we will first represent $T_1 = X_1/\lambda_1, T_2 = X_2/\lambda_2$ with X_1, X_2 i.i.d. Expo(1), which is Gamma(1, 1). Then

$$P(T_1 < T_2) = P\left(\frac{X_1}{X_2} < \frac{\lambda_1}{\lambda_2}\right) = P\left(\frac{X_1}{X_1 + X_2} < \frac{\lambda_1}{\lambda_1 + \lambda_2}\right) = \frac{\lambda_1}{\lambda_1 + \lambda_2},$$

since $X_1/(X_1 + X_2) \sim \text{Beta}(1, 1)$, which is Unif(0, 1).

(c) Alice's time in the post office is $T_{\text{wait}} + T_{\text{serve}}$, where $T_{\text{wait}} = \min(T_1, T_2)$ is how long she waits in line and T_{serve} is how long it takes for her to be served once it is her turn. By linearity, the expected total time is $E(T_{\text{wait}}) + E(T_{\text{serve}})$. Then

$$T_{\text{wait}} \sim \text{Expo}(\lambda_1 + \lambda_2),$$

$$E(T_{\text{wait}}) = \frac{1}{\lambda_1 + \lambda_2}.$$

Let f_{serve} be the PDF of T_{serve}. By LOTP, conditioning on which clerk serves Alice,

$$f_{\text{serve}}(t) = f_{\text{serve}}(t|T_1 < T_2)P(T_1 < T_2) + f_{\text{serve}}(t|T_1 > T_2)P(T_1 > T_2)$$

$$= \lambda_1 e^{-\lambda_1 t} \cdot \frac{\lambda_1}{\lambda_1 + \lambda_2} + \lambda_2 e^{-\lambda_2 t} \cdot \frac{\lambda_2}{\lambda_1 + \lambda_2}.$$

So

$$\begin{aligned}
E(T_{\text{serve}}) &= \int_0^\infty t f_{\text{serve}}(t) dt \\
&= \frac{\lambda_1}{\lambda_1 + \lambda_2} \int_0^\infty \lambda_1 t e^{-\lambda_1 t} dt + \frac{\lambda_2}{\lambda_1 + \lambda_2} \int_0^\infty \lambda_2 t e^{-\lambda_2 t} dt \\
&= \frac{\lambda_1}{\lambda_1 + \lambda_2} \cdot \frac{1}{\lambda_1} + \frac{\lambda_2}{\lambda_1 + \lambda_2} \cdot \frac{1}{\lambda_2} \\
&= \frac{2}{\lambda_1 + \lambda_2}.
\end{aligned}$$

Thus, the expected total time is $3/(\lambda_1 + \lambda_2)$.

In the special case $\lambda_1 = \lambda_2 = \lambda$, this simplifies to $3/(2\lambda)$, agreeing with the result of Exercise 38 from Chapter 5.

37. Let $X \sim \text{Pois}(\lambda t)$ and $Y \sim \text{Gamma}(j, \lambda)$, where j is a positive integer. Show using a story about a Poisson process that

$$P(X \geq j) = P(Y \leq t).$$

Solution: Imagine a Poisson process of rate λ. Let $X \sim \text{Pois}(\lambda)$ be the number of arrivals up until time t, and $Y \sim \text{Gamma}(j, \lambda)$ be the time of the jth arrival. Then $X \geq j$ and $Y \leq t$ are the same event: having at least j arrivals as of time t means exactly that the jth arrival has occurred, as of time t. Therefore, $P(X \geq j) = P(Y \leq t)$.

38. Visitors arrive at a certain scenic park according to a Poisson process with rate λ visitors per hour. Fred has just arrived (independent of anyone else), and will stay for an $\text{Expo}(\lambda_2)$ number of hours. Find the distribution of the number of other visitors who arrive at the park while Fred is there.

Solution: Let T be how long Fred stays at the park and Y be how many people arrive while he is there. Then

$$T \sim \text{Expo}(\lambda_2) \text{ and } Y|T \sim \text{Pois}(\lambda T).$$

As shown in Story 8.4.5 (Gamma-Poisson conjugacy), the marginal distribution of Y is

$$Y \sim \text{NBin}\left(1, \frac{\lambda_2}{\lambda_2 + \lambda}\right).$$

That is, Y is Geometric with parameter $p = \lambda_2/(\lambda_2 + \lambda)$ (and with mean λ/λ_2).

39. (a) Let $p \sim \text{Beta}(a, b)$, where a and b are positive real numbers. Find $E(p^2(1 - p)^2)$, fully simplified (Γ should not appear in your final answer).

Two teams, A and B, have an upcoming match. They will play five games and the winner will be declared to be the team that wins the majority of games. Given p, the outcomes of games are independent, with probability p of team A winning and $1 - p$ of team B winning. But you don't know p, so you decide to model it as an r.v., with $p \sim \text{Unif}(0, 1)$ a priori (before you have observed any data).

To learn more about p, you look through the historical records of previous games between these two teams, and find that the previous outcomes were, in chronological order, $AAABBAABAB$. (Assume that the true value of p has not been changing over time and will be the same for the match, though your *beliefs* about p may change over time.)

(b) Does your posterior distribution for p, given the historical record of games between A and B, depend on the specific order of outcomes or only on the fact that A won exactly 6 of the 10 games on record? Explain.

(c) Find the posterior distribution for p, given the historical data.

The posterior distribution for p from (c) becomes your new prior distribution, and the match is about to begin!

(d) Conditional on p, is the indicator of A winning the first game of the match positively correlated with, uncorrelated with, or negatively correlated with the indicator of A winning the second game? What about if we only condition on the historical data?

(e) Given the historical data, what is the expected value for the probability that the match is not yet decided when going into the fifth game (viewing this probability as an r.v. rather than a number, to reflect our uncertainty about it)?

Solution:

(a) By LOTUS,

$$E(p^2(1-p)^2) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 p^2(1-p)^2 p^{a-1}(1-p)^{b-1} dp$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 p^{a+1}(1-p)^{b+1} dp.$$

We recognize the integrand as an un-normalized Beta$(a + 2, b + 2)$ distribution. Multiplying and dividing by a constant and using the fact that PDFs integrate to 1, we obtain

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+2)\Gamma(b+2)}{\Gamma(a+b+4)} = \frac{(a+1)a(b+1)b}{(a+b+3)(a+b+2)(a+b+1)(a+b)},$$

since, e.g., $\Gamma(a+2) = (a+1)\Gamma(a+1) = (a+1)a\Gamma(a)$.

(b) Only the *numbers* of prior wins and losses for A matter, not the order in which they occurred. By the coherency of Bayes' rule (see Section 2.6), we will get the same posterior distribution from updating all at once on the observed data $AAABBAABAB$ as from updating one game at a time. So consider the one game at a time method. We start with a Beta$(1, 1)$ distribution. Each time A wins, we increment the first parameter by 1; each time B wins, we increment the second parameter by 1. Thus, the posterior distribution depends only on the number of wins for A and the number of wins for B.

(c) The prior on p is Beta$(1, 1)$. By conjugacy,

$$p|\text{historical data} \sim \text{Beta}(7, 5).$$

(d) These indicator r.v.s are conditionally independent given p, but not independent. Given p, they are therefore uncorrelated. Without being given p, the indicators are positively correlated since learning that A won the first game increases our estimate of p, which in turn increases our degree of belief that A will win the next game.

(e) Given p, the probability that the match is tied after 4 games is $\binom{4}{2}p^2(1-p)^2$. But p is unknown, so we will estimate this quantity instead. By (a) and (c), the expected value of $\binom{4}{2}p^2(1-p)^2$ given the historical data is

$$\binom{4}{2} \cdot \frac{(8)(7)(6)(5)}{(15)(14)(13)(12)} = \frac{4}{13}.$$

40. An engineer is studying the reliability of a product by performing a sequence of n trials. Reliability is defined as the probability of success. In each trial, the product succeeds with probability p and fails with probability $1 - p$. The trials are conditionally independent given p. Here p is unknown (else the study would be unnecessary!). The engineer takes a Bayesian approach, with $p \sim \text{Unif}(0, 1)$ as prior.

Let r be a desired reliability level and c be the corresponding confidence level, in the sense that, given the data, the probability is c that the true reliability p is at least r. For example, if $r = 0.9, c = 0.95$, we can be 95% sure, given the data, that the product is at least 90% reliable. Suppose that it is observed that the product succeeds all n times. Find a simple equation for c as a function of r.

Solution: The prior distribution for p is Beta$(1, 1)$, so the posterior distribution for p is Beta$(n + 1, 1)$. So

$$c = P(p \geq r|\text{data}) = (n+1) \int_r^1 p^n dp = 1 - r^{n+1}.$$

Order statistics

41. Ⓢ Let $X \sim \text{Bin}(n, p)$ and $B \sim \text{Beta}(j, n - j + 1)$, where n is a positive integer and j is a positive integer with $j \leq n$. Show using a story about order statistics that

$$P(X \geq j) = P(B \leq p).$$

This shows that the CDF of the continuous r.v. B is closely related to the CDF of the discrete r.v. X, and is another connection between the Beta and Binomial.

Solution: Let $U_1, \ldots, U_n$ be i.i.d. $\text{Unif}(0, 1)$. Think of these as Bernoulli trials, where U_j is defined to be "successful" if $U_j \leq p$ (so the probability of success is p for each trial). Let X be the number of successes. Then $X \geq j$ is the same event as $U_{(j)} \leq p$, so $P(X \geq j) = P(U_{(j)} \leq p)$.

42. Show that for i.i.d. continuous r.v.s X, Y, Z,

$$P(X < \min(Y, Z)) + P(Y < \min(X, Z)) + P(Z < \min(X, Y)) = 1.$$

Solution: Note that if x, y, z are distinct numbers (i.e., no two of them are equal), then exactly one of the statements $x < \min(y, z)$, $y < \min(x, z)$, $z < \min(x, y)$ is true (specifically, $\min(x, y, z)$ is less than the other two numbers). And since X, Y, Z are continuous r.v.s, they will be distinct with probability 1.
Let A, B, C be the events $\{X < \min(Y, Z)\}, \{Y < \min(X, Z)\}, \{Z < \min(X, Y)\}$, respectively. Then

$$1 = P(X, Y, Z \text{ are distinct}) \leq P(A \cup B \cup C) \leq 1,$$

so $P(A \cup B \cup C) = 1$. The events A, B, C are disjoint, so

$$P(A) + P(B) + P(C) = P(A \cup B \cup C) = 1.$$

43. Show that

$$\int_0^x \frac{n!}{(j - 1)!(n - j)!} t^{j-1}(1 - t)^{n-j} dt = \sum_{k=j}^n \binom{n}{k} x^k (1 - x)^{n-k},$$

without using calculus, for all $x \in [0, 1]$ and j, n positive integers with $j \leq n$.

Solution: Let $U_1, \ldots, U_n$ be i.i.d. $\text{Unif}(0, 1)$ r.v.s, and fix $x \in [0, 1]$. Define "success" to mean being at most x and "failure" to mean being greater than x. The right-hand side of the identity is the probability of having at least j successes (since it is the sum of the $\text{Bin}(n, x)$ PMF from j to n). The left-hand side is $P(U_{(j)} \leq x)$, since $U_{(j)} \sim \text{Beta}(j, n - j + 1)$. But having at least j successes is the same thing as having $U_{(j)} \leq x$, so the two sides are equal.

44. Let $X_1, \ldots, X_n$ be i.i.d. continuous r.v.s with PDF f and a strictly increasing CDF F. Suppose that we know that the jth order statistic of n i.i.d. $\text{Unif}(0, 1)$ r.v.s is a $\text{Beta}(j, n - j + 1)$, but we have forgotten the formula and derivation for the distribution of the jth order statistic of $X_1, \ldots, X_n$. Show how we can recover the PDF of $X_{(j)}$ quickly using a change of variables.

Solution: Let $U_j = F(X_j)$. By Universality of the Uniform, $U_1, \ldots, U_n$ are i.i.d. $\text{Unif}(0, 1)$. Since F is increasing, it preserves order, so $U_{(j)} = F(X_{(j)})$. To simplify notation, let $T = X_{(j)}, W = F(T) = U_{(j)}$, and $w = F(t)$. By the change of variables formula,

$$f_T(t) = f_W(w) \frac{dw}{dt}.$$

Since $W \sim \text{Beta}(j, n - j + 1)$, this is

$$\frac{\Gamma(j + n - j + 1)}{\Gamma(j)\Gamma(n - j + 1)} w^{j-1}(1 - w)^{n-j} f(t) = \frac{n!}{(n - j)!(j - 1)!} F(t)^{j-1}(1 - F(t))^{n-j} f(t).$$

45. Ⓢ Let X and Y be independent Expo(λ) r.v.s and $M = \max(X, Y)$. Show that M has the same distribution as $X + \frac{1}{2}Y$, in two ways: (a) using calculus and (b) by remembering the memoryless property and other properties of the Exponential.

Solution:

(a) The CDF of M is

$$F_M(x) = P(M \leq x) = P(X \leq x, Y \leq x) = (1 - e^{-\lambda x})^2,$$

and the CDF of $X + \frac{1}{2}Y$ is

$$F_{X + \frac{1}{2}Y}(x) = P(X + \frac{1}{2}Y \leq x) = \iint\limits_{s + \frac{1}{2}t \leq x} \lambda^2 e^{-\lambda s - \lambda t} ds dt$$

$$= \int_0^{2x} \lambda e^{-\lambda t} dt \int_0^{x - \frac{1}{2}t} \lambda e^{-\lambda s} ds$$

$$= \int_0^{2x} (1 - e^{-\lambda x - \frac{1}{2}t}) \lambda e^{-\lambda t} dt = (1 - e^{-\lambda x})^2.$$

Thus, M and $X + \frac{1}{2}Y$ have the same CDF.

(b) Reasoning as in Example 5.6.5, imagine that two students are independently trying to solve a problem. Suppose that X and Y are the times required. Let $L = \min(X, Y)$, and write $M = L + (M - L)$. $L \sim \text{Expo}(2\lambda)$ is the time it takes for the first student to solve the problem and then by the memoryless property, the additional time until the second student solves the problem is $M - L \sim \text{Expo}(\lambda)$, independent of L. Since $\frac{1}{2}Y \sim \text{Expo}(2\lambda)$ is independent of $X \sim \text{Expo}(\lambda)$, $M = L + (M - L)$ has the same distribution as $\frac{1}{2}Y + X$.

46. Ⓢ (a) If X and Y are i.i.d. continuous r.v.s with CDF $F(x)$ and PDF $f(x)$, then $M = \max(X, Y)$ has PDF $2F(x)f(x)$. Now let X and Y be discrete and i.i.d., with CDF $F(x)$ and PMF $f(x)$. Explain in words why the PMF of M is *not* $2F(x)f(x)$.

(b) Let X and Y be independent Bern(1/2) r.v.s, and let $M = \max(X, Y)$, $L = \min(X, Y)$. Find the joint PMF of M and L, i.e., $P(M = a, L = b)$, and the marginal PMFs of M and L.

Solution:

(a) The PMF is not $2F(x)f(x)$ in the discrete case due to the problem of ties: there is a nonzero chance that $X = Y$. We can write the PMF as $P(M = a) = P(X = a, Y < a) + P(Y = a, X < a) + P(X = Y = a)$ since $M = a$ means that at least one of X, Y equals a, with neither greater than a. The first two terms together become $2f(a)P(Y < a)$, but the third term may be nonzero and also $P(Y < a)$ may not equal $F(a) = P(Y \leq a)$.

(b) In order statistics notation, $L = X_{(1)}, M = X_{(2)}$. Marginally, we have $X_{(1)} \sim$ Bern(1/4), $X_{(2)} \sim$ Bern(3/4). The joint PMF is

$$
\begin{aligned}
P(X_{(1)} = 0, X_{(2)} = 0) &= 1/4 \\
P(X_{(1)} = 0, X_{(2)} = 1) &= 1/2 \\
P(X_{(1)} = 1, X_{(2)} = 0) &= 0 \\
P(X_{(1)} = 1, X_{(2)} = 1) &= 1/4.
\end{aligned}
$$

Note that these values are nonnegative and sum to 1, and that $X_{(1)}$ and $X_{(2)}$ are dependent.

47. Let $X_1, X_2, \ldots$ be i.i.d. r.v.s with CDF F, and let $M_n = \max(X_1, X_2, \ldots, X_n)$. Find the joint distribution of M_n and M_{n+1}, for each $n \geq 1$.

   Solution: We will find the joint CDF of M_n and M_{n+1}. For $a \leq b$,

   $$P(M_n \leq a, M_{n+1} \leq b) = P(X_1 \leq a, \ldots, X_n \leq a, X_{n+1} \leq b) = F(a)^n F(b).$$
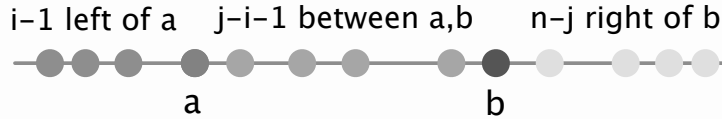
   For $a > b$,

   $$P(M_n \leq a, M_{n+1} \leq b) = P(M_{n+1} \leq b) = P(X_1 \leq b, \ldots, X_n \leq b, X_{n+1} \leq b) = F(b)^{n+1}.$$

   So the joint CDF is

   $$F_{M_n, M_{n+1}}(a, b) = \begin{cases} F(a)^n F(b), & \text{for } a \leq b, \\ F(b)^{n+1}, & \text{for } a > b. \end{cases}$$

48. Ⓢ Let $X_1, X_2, \ldots, X_n$ be i.i.d. r.v.s with CDF F and PDF f. Find the joint PDF of the order statistics $X_{(i)}$ and $X_{(j)}$ for $1 \leq i < j \leq n$, by drawing and thinking about a picture.

   Solution:

   To have $X_{(i)}$ be in a tiny interval around a and $X_{(j)}$ be in a tiny interval around b, where $a < b$, we need to have 1 of the X_k's be almost exactly at a, another be almost exactly at b, $i - 1$ of them should be to the left of a, $n - j$ should be to the right of b, and the remaining $j - i - 1$ should be between a and b, as shown in the picture. This gives that the PDF is

   $$f_{(i),(j)}(a, b) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} F(a)^{i-1} f(a) (F(b) - F(a))^{j-i-1} f(b) (1 - F(b))^{n-j},$$

   for $a < b$. The coefficient in front counts the number of ways to put the X_k's into the 5 categories "left of a," "at a," "between a and b," "at b," "right of b" with the desired number in each category (which is the same idea used to find the coefficient in front of the Multinomial PMF). Equivalently, we can write the coefficient as $n(n-1)\binom{n-2}{i-1}\binom{n-i-1}{j-i-1}$, since there are n choices for which X_k is at a, then $n - 1$ choices for which is at b, etc.

49. Ⓢ Two women are pregnant, both with the same due date. On a timeline, define time 0 to be the instant when the due date begins. Suppose that the time when the woman gives birth has a Normal distribution, centered at 0 and with standard deviation 8 days. Assume that the two birth times are i.i.d. Let T be the time of the first of the two births (in days).

   (a) Show that

   $$E(T) = \frac{-8}{\sqrt{\pi}}.$$

   Hint: For any two random variables X and Y, we have $\max(X, Y) + \min(X, Y) = X + Y$ and $\max(X, Y) - \min(X, Y) = |X - Y|$. Example 7.2.3 derives the expected distance between two i.i.d. $\mathcal{N}(0, 1)$ r.v.s.

   (b) Find $\text{Var}(T)$, in terms of integrals. You can leave your answer unsimplified for this part, but it can be shown that the answer works out to

   $$\text{Var}(T) = 64\left(1 - \frac{1}{\pi}\right).$$

Solution: Let $T = \min(T_1, T_2)$, with T_1 and T_2 the i.i.d. birth times. Standardizing, let

$$X = \frac{T_1}{8}, Y = \frac{T_2}{8}, L = \frac{T}{8},$$

so X and Y are i.i.d. $\mathcal{N}(0,1)$, and $L = \min(X, Y)$. Also, let

$$M = \max(X, Y), S = X + Y, W = |X - Y|.$$

We have $M + L = S$ and $M - L = W$, so $M = \frac{1}{2}(S + W)$ and $L = \frac{1}{2}(S - W)$. Then

$$E(S) = 0 \text{ and } E(W) = \frac{2}{\sqrt{\pi}},$$

by Example 7.2.3. Thus,

$$E(M) = \frac{1}{2}E(W) = \frac{1}{\sqrt{\pi}}, E(L) = \frac{-1}{2}E(W) = \frac{-1}{\sqrt{\pi}}.$$

It follows that

$$E(T) = 8E(L) = \frac{-8}{\sqrt{\pi}}.$$

(b) We can find the PDF of T using order statistics results, or directly using

$$P(T > t) = P(T_1 > t, T_2 > t) = (1 - \Phi(t/8))^2.$$

So the PDF of T is

$$f(t) = \frac{1}{4} \left(1 - \Phi(t/8)\right) \varphi(t/8),$$

where $\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ is the $\mathcal{N}(0,1)$ PDF. Thus,

$$\mathrm{Var}(T) = \int_{-\infty}^{\infty} t^2 f(t) dt - \left(\int_{-\infty}^{\infty} t f(t) dt\right)^2,$$

with f as above.

To get the variance in closed form (which was *not* requested in the problem), note that (with notation as above) $X + Y$ and $X - Y$ are independent since they are uncorrelated and $(X + Y, X - Y)$ is MVN. So $S = X + Y$ and $W = |X - Y|$ are independent. Thus,

$$\mathrm{Var}(L) = \frac{1}{4}(\mathrm{Var}(S) + \mathrm{Var}(W)).$$

We have $\mathrm{Var}(S) = 2$, and

$$\mathrm{Var}(W) = E(X - Y)^2 - (E(W))^2 = \mathrm{Var}(X - Y) - (E(W))^2 = 2 - \frac{4}{\pi}.$$

Therefore,

$$\mathrm{Var}(L) = 1 - \frac{1}{\pi},$$

which shows that

$$\mathrm{Var}(T) = 64\left(1 - \frac{1}{\pi}\right).$$

50. We are about to observe random variables $Y_1, Y_2, \ldots, Y_n$, i.i.d. from a continuous distribution. We will need to predict an independent future observation Y_{new}, which will also have the same distribution. The distribution is unknown, so we will construct our prediction using $Y_1, Y_2, \ldots, Y_n$ rather than the distribution of Y_{new}. In forming a prediction, we do not want to report only a single number; rather, we want to give a *predictive interval* with "high confidence" of containing Y_{new}. One approach to this is via order statistics.

(a) For fixed j and k with $1 \leq j < k \leq n$, find $P(Y_{\text{new}} \in [Y_{(j)}, Y_{(k)}])$.

Hint: By symmetry, all orderings of $Y_1, \ldots, Y_n, Y_{\text{new}}$ are equally likely.

(b) Let $n = 99$. Construct a predictive interval, as a function of $Y_1, \ldots, Y_n$, such that the probability of the interval containing Y_{new} is 0.95.

Solution:

(a) Using symmetry as in the hint, Y_{new} is equally likely to be in any of the $n+1$ possible "slots" relative to $Y_{(1)}, \ldots, Y_{(n)}$ (imagine inserting the new observation anywhere relative to the first n observations). There are $k - j$ slots that would put the new observation in between $Y_{(j)}$ and $Y_{(k)}$. So

$$P(Y_{\text{new}} \in [Y_{(j)}, Y_{(k)}]) = \frac{k - j}{n + 1}.$$

(b) Take an interval $[Y_{(j)}, Y_{(k)}]$ with $k - j = 95$. For example, the interval $[Y_{(3)}, Y_{(98)}]$ is as desired.

51. Let $X_1, \ldots, X_n$ be i.i.d. continuous r.v.s with n odd. Show that the median of the distribution of the sample median of the X_i's is the median of the distribution of the X_i's.

Hint: Start by reading the problem carefully; it is crucial to distinguish between the median of a distribution (as defined in Chapter 6) and the sample median of a collection of r.v.s (as defined in this chapter). Of course they are closely related: the sample median of i.i.d. r.v.s is a very natural way to estimate the true median of the distribution that the r.v.s are drawn from. Two approaches to evaluating a sum that might come up are (i) use the first story proof example and the first story proof exercise from Chapter 1, or (ii) use the fact that, by the story of the Binomial, $Y \sim \text{Bin}(n, 1/2)$ implies $n - Y \sim \text{Bin}(n, 1/2)$.

Solution:

Let $n = 2m + 1$, F be the CDF of X_i, and x_0 be the median of F (so $F(x_0) = 1/2$). The sample median is the order statistic $X_{(m+1)}$, which has CDF

$$P(X_{(m+1)} \leq x) = P(X_i \leq x \text{ for at least } m+1 \ X_i\text{'s}) = \sum_{k=m+1}^{n} \binom{n}{k} F(x)^k (1 - F(x))^{n-k}.$$

Evaluating this at x_0, we have

$$P(X_{(m+1)} \leq x_0) = \frac{1}{2^n} \sum_{k=m+1}^{n} \binom{n}{k} = \frac{1}{2},$$

where the last step uses the story proofs from the hint, or uses the following symmetry property of the $\text{Bin}(n, 1/2)$ distribution. Let $Y \sim \text{Bin}(n, 1/2)$. Then

$$P(Y \geq m + 1) = P(n - Y \geq m + 1) = P(Y \leq n - m - 1) = P(Y \leq m),$$

and

$$1 = P(0 \leq Y \leq n) = P(Y \leq m) + P(Y \geq m + 1),$$

so

$$\sum_{k=m+1}^{n} \binom{n}{k} \frac{1}{2^n} = P(Y \geq m + 1) = \frac{1}{2}.$$

Thus, x_0 is also the median of the distribution of $X_{(m+1)}$.

Mixed practice

52. Let $U_1, U_2, \ldots, U_n$ be i.i.d. Unif$(0,1)$, and let $X_j = -\log(U_j)$ for all j.

(a) Find the distribution of X_j. What is its name?

(b) Find the distribution of the product $U_1 U_2 \ldots U_n$.

Hint: First take the log.

Solution:

(a) Let $x = -\log(u)$, so $u = e^{-x}$. By the change of variables formula, the PDF of each X_j is

$$f_{X_j}(x) = f_{U_j}(u)\left|\frac{du}{dx}\right| = e^{-x},$$

for $x > 0$. So X_j is Exponential with rate parameter $\lambda = 1$.

(b) Let $Y = U_1 U_2 \ldots U_n$ and

$$X = -\log(Y) = -\log U_1 - \log U_2 - \cdots - \log U_n.$$

By (a) and the connection between Gamma and Exponential, $X \sim \text{Gamma}(n, 1)$. Let $y = e^{-x}$, so $x = -\log y$. By the change of variables formula, the PDF of Y is

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right| = \frac{1}{\Gamma(a)}x^a e^{-x}\frac{1}{x}\cdot\frac{1}{y} = \frac{1}{\Gamma(a)}(-\log y)^{a-1},$$

for $0 < y < 1$.

53. Ⓢ A DNA sequence can be represented as a sequence of letters, where the "alphabet" has 4 letters: A,C,T,G. Suppose such a sequence is generated randomly, where the letters are independent and the probabilities of A,C,T,G are p_1, p_2, p_3, p_4 respectively.

(a) In a DNA sequence of length 115, what is the expected number of occurrences of the expression "CATCAT" (in terms of the p_j)? (Note that, for example, the expression "CATCATCAT" counts as 2 occurrences.)

(b) What is the probability that the first A appears earlier than the first C appears, as letters are generated one by one (in terms of the p_j)?

(c) For this part, assume that the p_j are unknown. Suppose we treat p_2 as a Unif$(0,1)$ r.v. before observing any data, and that then the first 3 letters observed are "CAT". Given this information, what is the probability that the next letter is C?

Solution:

(a) Let I_j be the indicator r.v. of "CATCAT" appearing starting at position j, for $1 \le j \le 110$. Then $E(I_j) = (p_1 p_2 p_3)^2$, so the expected number is $110(p_1 p_2 p_3)^2$.

Sanity check: The number of occurrences is between 0 and 110, so the expected value must also be between 0 and 110. If any of the letters C, A, or T is very rare, then "CATCAT" will be even more rare; this is reflected in the p_j^2 factors, which will make the expected number small if any of p_1, p_2, p_3 is small.

(b) Consider the first letter which is an A or a C (call it X; alternatively, condition on the first letter of the sequence). This gives

$$P(\text{A before C}) = P(X \text{ is A}|X \text{ is A or C}) = \frac{P(X \text{ is A})}{P(X \text{ is A or C})} = \frac{p_1}{p_1 + p_2}.$$

Sanity check: The answer should be $1/2$ for $p_1 = p_2$, should go to 0 as $p_1 \to 0$, should be

increasing in p_1 and decreasing in p_2, and finding $P(\text{A before C})$ by $1 - P(\text{A before C})$ should agree with finding it by swapping p_1, p_2.

(c) Let X be the number of C's in the data (so $X = 1$ is observed here). The prior is $p_2 \sim \text{Beta}(1,1)$, so the posterior is $p_2|X = 1 \sim \text{Beta}(2,3)$ (by the connection between Beta and Binomial, or by Bayes' Rule). Given p_2, the indicator of the next letter being C is $\text{Bern}(p_2)$. So given X (but not given p_2), the probability of the next letter being C is $E(p_2|X) = \frac{2}{5}$.

Sanity check: It makes sense that the answer should be strictly in between $1/2$ (the mean of the prior distribution) and $1/3$ (the observed frequency of C's in the data).

54. Ⓢ Consider independent Bernoulli trials with probability p of success for each. Let X be the number of failures incurred before getting a total of r successes.

(a) Determine what happens to the distribution of $\frac{p}{1-p}X$ as $p \to 0$, using MGFs; what is the PDF of the limiting distribution, and its name and parameters if it is one we have studied?

Hint: Start by finding the $\text{Geom}(p)$ MGF. Then find the MGF of $\frac{p}{1-p}X$, and use the fact that if the MGFs of r.v.s Y_n converge to the MGF of an r.v. Y, then the CDFs of the Y_n converge to the CDF of Y.

(b) Explain intuitively why the result of (a) makes sense.

Solution:

(a) Let $q = 1 - p$. For $G \sim \text{Geom}(p)$, the MGF is

$$E(e^{tG}) = p\sum_{k=0}^{\infty} e^{tk}q^k = p\sum_{k=0}^{\infty}(qe^t)^k = \frac{p}{1 - qe^t},$$

for $qe^t < 1$. So the $\text{NBin}(r, p)$ MGF is $\frac{p^r}{(1 - qe^t)^r}$ for $qe^t < 1$. Then the MGF of $\frac{p}{1-p}X$ is

$$E(e^{\frac{tp}{q}X}) = \frac{p^r}{(1 - qe^{tp/q})^r}$$

for $qe^{tp/q} < 1$. Let us first consider the limit for $r = 1$. As $p \to 0$, the numerator goes to 0 and so does the denominator (since $qe^{tp/q} \to 1e^0 = 1$). By L'Hôpital's Rule,

$$\lim_{p\to 0}\frac{p}{1 - (1-p)e^{tp/(1-p)}} = \lim_{p\to 0}\frac{1}{e^{tp/(1-p)} - (1-p)t\left(\frac{1-p+p}{(1-p)^2}\right)e^{tp/(1-p)}} = \frac{1}{1-t}.$$

So for any fixed $r > 0$, as $p \to 0$ we have

$$E\left(e^{\frac{tp}{q}X}\right) = \frac{p^r}{(1 - qe^{tp/q})^r} \to \frac{1}{(1-t)^r}.$$

This is the $\text{Gamma}(r, 1)$ MGF for $t < 1$ (note also that the condition $qe^{tp/q} < 1$ is equivalent to $t < -\frac{1-p}{p}\log(1-p)$, which converges to the condition $t < 1$ since again by L'Hôpital's Rule, $\frac{-p}{\log(1-p)} \to 1$). Thus, the scaled Negative Binomial $\frac{p}{1-p}X$ converges to $\text{Gamma}(r, 1)$ in distribution as $p \to 0$.

(b) The result of (a) makes sense intuitively since the Gamma is the continuous analogue of the Negative Binomial, just as the Exponential is the continuous analog of the Geometric. To convert from discrete to continuous, imagine performing many, many trials where each is performed very, very quickly and has a very, very low chance of success. To balance the rate of trials with the chance of success, we use the scaling $\frac{p}{q}$ since this makes $E(\frac{p}{q}X) = r$, matching the $\text{Gamma}(r, 1)$ mean.

9

Conditional expectation

Conditional expectation given an event

1. Fred wants to travel from Blotchville to Blissville, and is deciding between 3 options (involving different routes or different forms of transportation). The jth option would take an average of μ_j hours, with a standard deviation of σ_j hours. Fred randomly chooses between the 3 options, with equal probabilities. Let T be how long it takes for him to get from Blotchville to Blissville.

   (a) Find $E(T)$. Is it simply $(\mu_1 + \mu_2 + \mu_3)/3$, the average of the expectations?

   (b) Find $\text{Var}(T)$. Is it simply $(\sigma_1^2 + \sigma_2^2 + \sigma_3^2)/3$, the average of the variances?

   Solution:

   (a) Let R_j be the event that Fred takes the jth route (or option). Then

   $$E(T) = E(T|R_1)P(R_1) + E(T|R_2)P(R_2) + E(T|R_3)P(R_3) = \frac{\mu_1 + \mu_2 + \mu_3}{3},$$

   which is indeed the average of the expectations for each route (or option).

   (b) We have

   $$E(T^2) = E(T^2|R_1)P(R_1) + E(T^2|R_2)P(R_2) + E(T^2|R_3)P(R_3) = \frac{(\sigma_1^2 + \mu_1^2) + (\sigma_2^2 + \mu_2^2) + (\sigma_3^2 + \mu_3^2)}{3},$$

   so

   $$\text{Var}(T) = E(T^2) - (ET)^2 = \frac{\sigma_1^2 + \sigma_2^2 + \sigma_3^2}{3} + \frac{\mu_1^2 + \mu_2^2 + \mu_3^2}{3} - \left(\frac{\mu_1 + \mu_2 + \mu_3}{3}\right)^2.$$

   This is greater than or equal to $(\sigma_1^2 + \sigma_2^2 + \sigma_3^2)/3$, with strict inequality except in the case $\mu_1 = \mu_2 = \mu_3$. To see this, note that we can also write

   $$\text{Var}(T) = \frac{\sigma_1^2 + \sigma_2^2 + \sigma_3^2}{3} + \text{Var}(M),$$

   where M is μ_1, μ_2, or μ_3, with the choice of index $(1, 2,$ or $3)$ made uniformly.

2. While Fred is sleeping one night, X legitimate emails and Y spam emails are sent to him. Suppose that X and Y are independent, with $X \sim \text{Pois}(10)$ and $Y \sim \text{Pois}(40)$. When he wakes up, he observes that he has 30 new emails in his inbox. Given this information, what is the expected value of how many new legitimate emails he has?

   Solution: By Theorem 4.8.2 (Poisson given a sum of Poissons), the conditional distribution of X given $X + Y = 30$ is $\text{Bin}(30, 10/50)$. So

   $$E(X|X + Y = 30) = 30 \cdot \frac{10}{50} = 6.$$

3. A group of 21 women and 14 men are enrolled in a medical study. Each of them has a certain disease with probability p, independently. It is then found (through extremely reliable testing) that exactly 5 of the people have the disease. Given this information, what is the expected number of women who have the disease?

Solution: Let $X \sim \text{Bin}(21, p)$ and $Y \sim \text{Bin}(14, p)$ be the number of women and men (respectively) who have the disease. By Theorem 3.9.2 or the Fisher exact test, the conditional distribution of $X|X + Y = 5 \sim \text{HGeom}(21, 14, 5)$. So

$$E(X|X + Y = 5) = \frac{5 \cdot 21}{21 + 14} = 3.$$

Note that this does not depend on p.

Alternatively, we can imagine randomly assigning the disease to 5 people, where by symmetry all sets of 5 people are equally likely, and then create an indicator r.v. for disease for each of the women. The probability that a woman has the disease is $5/35 = 1/7$, so the expected number of women with the disease is $21/7 = 3$.

4. A researcher studying crime is interested in how often people have gotten arrested. Let $X \sim \text{Pois}(\lambda)$ be the number of times that a random person got arrested in the last 10 years. However, data from police records are being used for the researcher's study, and people who were never arrested in the last 10 years do not appear in the records. In other words, the police records have a *selection bias*: they only contain information on people who *have* been arrested in the last 10 years.

So averaging the numbers of arrests for people in the police records does not directly estimate $E(X)$; it makes more sense to think of the police records as giving us information about the *conditional* distribution of how many times a person was arrest, given that the person was arrested at least once in the last 10 years. The conditional distribution of X, given that $X \geq 1$, is called a *truncated Poisson distribution* (see Exercise 14 from Chapter 3 for another example of this distribution).

(a) Find $E(X|X \geq 1)$

(b) Find $\text{Var}(X|X \geq 1)$.

Solution:

(a) The conditional PMF of X given $X \geq 1$ is

$$P(X = k|X \geq 1) = \frac{P(X = k)}{P(X \geq 1)} = \frac{e^{-\lambda}\lambda^k}{k!(1 - e^{-\lambda})},$$

for $k = 1, 2, \ldots$. Therefore,

$$E(X|X \geq 1) = \frac{e^{-\lambda}}{1 - e^{-\lambda}} \sum_{k=1}^{\infty} \frac{k\lambda^k}{k!}.$$

We have

$$e^{-\lambda} \sum_{k=1}^{\infty} \frac{k\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{k\lambda^k}{k!} = E(X) = \lambda,$$

so

$$E(X|X \geq 1) = \frac{\lambda}{1 - e^{-\lambda}}.$$

(b) We have

$$e^{-\lambda} \sum_{k=1}^{\infty} \frac{k^2\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{k^2\lambda^k}{k!} = E(X^2) = \text{Var}(X) + (EX)^2 = \lambda + \lambda^2,$$

so

$$E(X^2|X \geq 1) = \frac{e^{-\lambda}}{1 - e^{-\lambda}} \sum_{k=1}^{\infty} \frac{k^2 \lambda^k}{k!} = \frac{\lambda + \lambda^2}{1 - e^{-\lambda}},$$

$$\text{Var}(X|X \geq 1) = \frac{\lambda + \lambda^2}{1 - e^{-\lambda}} - \frac{\lambda^2}{(1 - e^{-\lambda})^2}.$$

5. A fair 20-sided die is rolled repeatedly, until a gambler decides to stop. The gambler pays \$1 per roll, and receives the amount shown on the die when the gambler stops (e.g., if the die is rolled 7 times and the gambler decides to stop then, with an 18 as the value of the last roll, then the net payoff is \$18 − \$7 = \$11). Suppose the gambler uses the following strategy: keep rolling until a value of m or greater is obtained, and then stop (where m is a fixed integer between 1 and 20).

    (a) What is the expected net payoff?

    Hint: The average of consecutive integers $a, a + 1, \ldots, a + n$ is the same as the average of the first and last of these. See the math appendix for more information about series.

    (b) Use R or other software to find the optimal value of m.

    Solution:

    (a) Let X be the value of the final roll and N be the number of rolls. The distribution of X is Discrete Uniform on $m, m + 1, \ldots, 20$, since this is the conditional distribution of a Discrete Uniform on $1, 2, \ldots, 20$ given that the value is at least m. The distribution of N is $\text{FS}((21 - m)/20)$. So the expected net payoff is

    $$E(X - N) = E(X) - E(N) = \frac{m + 20}{2} - \frac{20}{21 - m}.$$

    (b) The R code

    ```
    m <- 1:20
    v <-  (m+20)/2 - 20/(21-m)
    max(v)
    which.max(v)
    ```

    yields that the optimal expected value is $14 + \frac{1}{6}$, which is obtained when $m = 15$.

6. Let $X \sim \text{Expo}(\lambda)$. Find $E(X|X < 1)$ in two different ways:

    (a) by calculus, working with the conditional PDF of X given $X < 1$.

    (b) without calculus, by expanding $E(X)$ using the law of total expectation.

    Solution:

    (a) The conditional PDF of X given $X < 1$ is

    $$f(x|X < 1) = \frac{P(X < 1|X = x)\lambda e^{-\lambda x}}{P(X < 1)} = \frac{\lambda e^{-\lambda x}}{1 - e^{-\lambda}},$$

    for $0 < x < 1$. So

    $$E(X|X < 1) = \int_0^1 x f(x|X < 1) dx = \frac{\lambda}{1 - e^{-\lambda}} \int_0^1 x e^{-\lambda x} dx.$$

    For any constant a, integration by parts gives

    $$\int x e^{ax} dx = \frac{e^{ax}(ax - 1)}{a^2} + C,$$

where C is an arbitrary constant of integration. Therefore,

$$E(X|X < 1) = \frac{\lambda}{1 - e^{-\lambda}} \cdot \frac{e^{-\lambda}(-\lambda - 1) + 1}{\lambda^2} = \frac{1}{\lambda} - \frac{1}{e^\lambda - 1}.$$

Sanity check: Note that $E(X|X < 1) < E(X)$, which makes sense intuitively since we are conditioning on X being *below* a certain value. Also,

$$\frac{1}{\lambda} - \frac{1}{e^\lambda - 1} > 0,$$

since

$$e^\lambda = 1 + \lambda + \frac{\lambda^2}{2!} + \cdots > 1 + \lambda.$$

(b) By the memoryless property, $E(X|X > 1) = 1 + 1/\lambda$. So

$$E(X) = E(X|X > 1)P(X > 1) + E(X|X < 1)P(X < 1)$$
$$= \left(1 + \frac{1}{\lambda}\right)e^{-\lambda} + E(X|X < 1)(1 - e^{-\lambda}).$$

Plugging in $E(X) = 1/\lambda$ and solving for $E(X|X < 1)$, we have

$$E(X|X < 1) = \frac{\frac{1}{\lambda} - (1 + \frac{1}{\lambda})e^{-\lambda}}{1 - e^{-\lambda}} = \frac{1}{\lambda} - \frac{1}{e^\lambda - 1},$$

which agrees with the result from (a).

7. You are given an opportunity to bid on a mystery box containing a mystery prize! The value of the prize is completely unknown, except that it is worth at least nothing, and at most a million dollars. So the true value V of the prize is considered to be Uniform on [0,1] (measured in millions of dollars).

You can choose to bid any nonnegative amount b (in millions of dollars). If $b < \frac{1}{4}V$, then your bid is rejected and nothing is gained or lost. If $b \geq \frac{1}{4}V$, then your bid is accepted and your net payoff is $V - b$ (since you pay b to get a prize worth V).

Find your expected payoff as a function of b (be sure to specify it for all $b \geq 0$). Then find the optimal bid b, to maximize your expected payoff.

Solution: For $b \geq \frac{1}{4}$, the bid is definitely accepted, and the average payoff is $E(V - b) = \frac{1}{2} - b$. Now assume $0 \leq b < \frac{1}{4}$. By the law of total expectation, the average payoff is

$$E(V - b|b \geq \frac{1}{4}V)P(b \geq \frac{1}{4}V) = \left(E(V|V \leq 4b) - b\right)P(V \leq 4b)$$
$$= (2b - b)(4b)$$
$$= 4b^2.$$

Thus, the average payoff is

$$\begin{cases} 4b^2, & \text{for } 0 \leq b < \frac{1}{4}; \\ \frac{1}{2} - b, & \text{for } b \geq \frac{1}{4}. \end{cases}$$

This function is strictly increasing for $0 \leq b < \frac{1}{4}$ and strictly decreasing for $b > \frac{1}{4}$, so the optimal bid is $b = \frac{1}{4}$, i.e., \$250,000.

8. Ⓢ You get to choose between two envelopes, each of which contains a check for some positive amount of money. Unlike in the two-envelope paradox, it is not given that one envelope contains twice as much money as the other envelope. Instead, assume that the two values were generated independently from some distribution on the positive real numbers, with no information given about what that distribution is.

After picking an envelope, you can open it and see how much money is inside (call this value x), and then you have the option of switching. As no information has been given about the distribution, it may seem impossible to have better than a 50% chance of picking the better envelope. Intuitively, we may want to switch if x is "small" and not switch if x is "large", but how do we define "small" and "large" in the grand scheme of all possible distributions? [The last sentence was a rhetorical question.]

Consider the following strategy for deciding whether to switch. Generate a threshold $T \sim \text{Expo}(1)$, and switch envelopes if and only if the observed value x is less than the value of T. Show that this strategy succeeds in picking the envelope with more money with probability strictly greater than $1/2$.

Hint: Let t be the value of T (generated by a random draw from the Expo(1) distribution). First explain why the strategy works very well if t happens to be in between the two envelope values, and does no harm in any case (i.e., there is no case in which the strategy succeeds with probability strictly less than $1/2$).

Solution: Let a be the smaller value of the two envelopes and b be the larger value (assume $a < b$ since in the case $a = b$ it makes no difference which envelope is chosen!). Let G be the event that the strategy succeeds and A be the event that we pick the envelope with a initially. Then $P(G|A) = P(T > a) = 1 - (1 - e^{-a}) = e^{-a}$, and $P(G|A^c) = P(T \le b) = 1 - e^{-b}$. Thus, the probability that the strategy succeeds is

$$\frac{1}{2}e^{-a} + \frac{1}{2}(1 - e^{-b}) = \frac{1}{2} + \frac{1}{2}(e^{-a} - e^{-b}) > \frac{1}{2},$$

because $e^{-a} - e^{-b} > 0$.

9. There are two envelopes, each of which has a check for a Unif$(0,1)$ amount of money, measured in thousands of dollars. The amounts in the two envelopes are independent. You get to choose an envelope and open it, and then you can either keep that amount or switch to the other envelope and get whatever amount is in that envelope.

Suppose that you use the following strategy: choose an envelope and open it. If you observe U, then stick with that envelope with probability U, and switch to the other envelope with probability $1 - U$.

(a) Find the probability that you get the larger of the two amounts.

(b) Find the expected value of what you will receive.

Solution:

(a) Let $U \sim \text{Unif}(0,1)$ be the amount in the envelope you open and $V \sim \text{Unif}(0,1)$ be the amount in the other envelope. Let B be the event that you get the larger of the two amounts and I_B be the indicator r.v. for B. Given that $U = u$, the probability of B is

$$P(B|U = u) = uP(V < u) + (1 - u)P(V > u) = u^2 + (1 - u)^2.$$

Therefore,

$$P(B) = E(I_B) = E(E(I_B|U)) = E(U^2 + (1 - U)^2) = \int_0^1 u^2 du + \int_0^1 (1 - u)^2 du = \frac{2}{3}.$$

(b) Let U and V be as in (a), I be the indicator of the event that you stick with your initial choice of envelope, and X be the amount you receive. Then $I|U \sim \text{Bern}(U)$ and

$$X = UI + V(1 - I).$$

So

$$E(X|U) = U \cdot U + \frac{1}{2} \cdot (1 - U) = U^2 - \frac{U}{2} + \frac{1}{2},$$

which shows that

$$E(X) = E(E(X|U)) = \frac{1}{3} - \frac{1}{4} + \frac{1}{2} = \frac{7}{12}.$$

Thus, the expected value is $7/12$ of a thousand dollars, which is about \$583.33.

10. Suppose n people are bidding on a mystery prize that is up for auction. The bids are to be submitted in secret, and the individual who submits the highest bid wins the prize. The ith bidder receives a signal X_i, with $X_1, \ldots, X_n$ i.i.d. The value of the prize, V, is defined to be the sum of the individual bidders' signals:

$$V = X_1 + \cdots + X_n.$$

This is known in economics as the *wallet game*: we can imagine that the n people are bidding on the total amount of money in their wallets, and each person's signal is the amount of money in their own wallet. Of course, the wallet is a metaphor; the game can also be used to model company takeovers, where each of two companies bids to take over the other, and a company knows its own value but not the value of the other company.

For this problem, assume the X_i are i.i.d. Unif$(0, 1)$.

(a) Before receiving her signal, what is bidder 1's unconditional expectation for V?

(b) Conditional on receiving the signal $X_1 = x_1$, what is bidder 1's expectation for V?

(c) Suppose each bidder submits a bid equal to their conditional expectation for V, i.e., bidder i bids $E(V|X_i = x_i)$. Conditional on receiving the signal $X_1 = x_1$ and *winning the auction*, what is bidder 1's expectation for V? Explain intuitively why this quantity is always less than the quantity calculated in (b).

Solution:

(a) By linearity, $E(V) = n/2$.

(b) Given that $X_1 = x_1$, the first term in $X_1 + X_2 + \cdots + X_n$ is x_1 and the remaining terms still have expectation $1/2$ each (since the X_i are i.i.d.), so $E(V|X_1 = x_1) = x_1 + (n-1)/2$.

(c) Bidder i bids $X_i + (n-1)/2$, so whoever has the highest signal will have the highest bid. As shown in Chapter 5, a Unif$(0, 1)$ r.v. conditioned to be in a certain subinterval of $(0, 1)$ is Uniform over that subinterval. So for $2 \leq j \leq n$,

$$E(X_j|X_1 = x_1, X_2 < x_1, X_3 < x_1, \ldots, X_n < X_1) = E(X_j|X_j < x_1) = \frac{x_1}{2}.$$

Thus,

$$E(V|X_1 = x_1, X_2 < x_1, X_3 < x_1, \ldots, X_n < X_1) = x_1 + \frac{(n-1)x_1}{2}.$$

This is less than the answer from (b), which makes sense intuitively since learning that you won the bid should lower your predictions about the other signals (since it means all those signals were lower than your own).

11. $\circledS$ A coin with probability p of Heads is flipped repeatedly. For (a) and (b), suppose that p is a known constant, with $0 < p < 1$.

(a) What is the expected number of flips until the pattern HT is observed?

(b) What is the expected number of flips until the pattern HH is observed?

(c) Now suppose that p is unknown, and that we use a Beta(a, b) prior to reflect our uncertainty about p (where a and b are known constants and are greater than 2). In terms of a and b, find the corresponding answers to (a) and (b) in this setting.

Solution:

(a) This can be thought of as "Wait for Heads, then wait for the first Tails after the first Heads," so the expected value is $\frac{1}{p} + \frac{1}{q}$, with $q = 1 - p$.

(b) Let X be the waiting time for HH and condition on the first toss, writing H for the event that the first toss is Heads and T for the complement of H:

$$E(X) = E(X|H)p + E(X|T)q = E(X|H)p + (1 + EX)q.$$

To find $E(X|H)$, condition on the second toss:

$$E(X|H) = E(X|HH)p + E(X|HT)q = 2p + (2 + EX)q.$$

Solving for $E(X)$, we have

$$E(X) = \frac{1}{p} + \frac{1}{p^2}.$$

Sanity check: This gives $E(X) = 6$ when $p = 1/2$, in agreement with Example 9.1.9.

(c) Let X and Y be the number of flips until HH and until HT, respectively. By (a), $E(Y|p) = \frac{1}{p} + \frac{1}{1-p}$. So $E(Y) = E(E(Y|p)) = E(\frac{1}{p}) + E(\frac{1}{1-p})$. Likewise, by (b), $E(X) = E(E(X|p)) = E(\frac{1}{p}) + E(\frac{1}{p^2})$. By LOTUS,

$$E\left(\frac{1}{p}\right) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 p^{a-2}(1-p)^{b-1}dp = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a-1)\Gamma(b)}{\Gamma(a+b-1)} = \frac{a+b-1}{a-1},$$

$$E\left(\frac{1}{1-p}\right) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 p^{a-1}(1-p)^{b-2}dp = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a)\Gamma(b-1)}{\Gamma(a+b-1)} = \frac{a+b-1}{b-1},$$

$$E\left(\frac{1}{p^2}\right) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 p^{a-3}(1-p)^{b-1}dp = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a-2)\Gamma(b)}{\Gamma(a+b-2)} = \frac{(a+b-1)(a+b-2)}{(a-1)(a-2)}.$$

Thus,

$$E(Y) = \frac{a+b-1}{a-1} + \frac{a+b-1}{b-1},$$

$$E(X) = \frac{a+b-1}{a-1} + \frac{(a+b-1)(a+b-2)}{(a-1)(a-2)}.$$

12. A coin with probability p of Heads is flipped repeatedly, where $0 < p < 1$. The sequence of outcomes can be divided into *runs* (blocks of H's or blocks of T's), e.g., $HHHTTTTHTTTHH$ becomes $\boxed{HHH}\,\boxed{TTTT}\,\boxed{H}\,\boxed{TTT}\,\boxed{HH}$, which has 5 runs, with lengths $3, 4, 1, 3, 2$, respectively. Assume that the coin is flipped at least until the start of the third run.

(a) Find the expected length of the first run.

(b) Find the expected length of the second run.

Solution:

(a) Let X be the length of the first run, H_1 be the event that the first toss is H, and T_1 be the event that the first toss is T. By the law of total expectation,

$$E(X) = E(X|H_1)P(H_1) + E(X|T_1)P(T_1) = pE(X|H_1) + qE(X|T_1).$$

Given H_1, the first run is a run of H's, with an $FS(q)$ length (since it consists of an initial H followed by a $Geom(q)$ string of H's while waiting for a T to start the second run). Similarly, given T_1, the first run is a run of T's, with an $FS(p)$ length. Thus,

$$E(L) = \frac{p}{q} + \frac{q}{p}.$$

(b) Let L be the length of the second run. With notation as in the previous part,

$$E(L) = E(L|H_1)P(H_1) + E(L|T_1)P(T_1) = pE(L|H_1) + qE(L|T_1).$$

Given H_1, the second run is a run of T's, with an $FS(p)$ length. Given T_1, the second run is a run of H's, with an $FS(q)$ length. Thus,

$$E(L) = \frac{p}{p} + \frac{q}{q} = 2.$$

13. A fair 6-sided die is rolled once. Find the expected number of additional rolls needed to obtain a value at least as large as that of the first roll.

    Solution: Let X be the result of the first roll and Y be the number of additional rolls needed to obtain a value at least as large as X. The conditional distribution of Y given $X = k$ is FS$((7-k)/6)$. So

    $$E(Y) = \sum_{k=1}^{6} E(Y|X=k)P(X=k) = \sum_{k=1}^{6} \frac{6}{7-k} \cdot \frac{1}{6} = \sum_{j=1}^{6} \frac{1}{j} = 2.45.$$

14. A fair 6-sided die is rolled repeatedly.

    (a) Find the expected number of rolls needed to get a 1 followed right away by a 2.

    Hint: Start by conditioning on whether or not the first roll is a 1.

    (b) Find the expected number of rolls needed to get two consecutive 1's.

    (c) Let a_n be the expected number of rolls needed to get the same value n times in a row (i.e., to obtain a streak of n consecutive j's for some not-specified-in-advance value of j). Find a recursive formula for a_{n+1} in terms of a_n.

    Hint: Divide the time until there are $n + 1$ consecutive appearances of the same value into two pieces: the time until there are n consecutive appearances, and the rest.

    (d) Find a simple, explicit formula for a_n for all $n \geq 1$. What is a_7 (numerically)?

    Solution:

    (a) Let T be the number of rolls needed and X_j be the value of the jth roll. Conditioning on whether $X_1 = 1$ occurs, we have

    $$E(T) = E(T|X_1 = 1)\frac{1}{6} + E(T|X_1 \neq 1)\frac{5}{6}.$$

    Letting $a = E(T), b = E(T|X_1 = 1)$, this becomes

    $$a = \frac{b}{6} + \frac{5(a+1)}{6},$$

    since if the first roll is not a 1 then we are back in the original situation except with one "wasted" roll. Doing further conditioning to study b, we have

    $$b = E(T|X_1 = 1) = \sum_{k=1}^{6} E(T|X_1 = 1, X_2 = k)P(X_2 = k|X_1 = 1) = \frac{2}{6} + \frac{b+1}{6} + \frac{4(a+2)}{6},$$

    since given that the first roll is a 1, if the second roll is a 2 we have attained the desired pattern, if it's a 1 we are again in the situation of waiting for $1, 2$ given that we start with a 1 except with one wasted roll, and if it's anything else we are in the original situation except with two wasted rolls. We can then just solve the system of 2 equations for a and b. We have $6a = b + 5a + 5$, so $a = b + 5$. Then $6b = 2 + b + 1 + 4(b + 5 + 2)$ so $b = 31, a = 36$. Thus, $E(T) = 36$.

    (b) Let T_1 be the number of rolls needed, X_j be the value of the jth roll, $a_1 = E(T_1), b_1 = E(T_1|X_1 = 1)$. Using the method from Part (a), we have

    $$a_1 = \frac{b_1}{6} + \frac{5(a_1 + 1)}{6}, b_1 = \frac{2}{6} + \frac{5(a_1 + 2)}{6}.$$

    Then $a_1 = b_1 + 5$ and $6b_1 = 2 + 5(b_1 + 5) + 10$, so $b_1 = 37, a_1 = 42$. Thus, $E(T_1) = 42$.

    (c) Let T_n be the time until there are n consecutive appearances of the same value (we

are making new notation for this part, not continuing the notation from the previous part). Just after T_n rolls, the last n rolls have been $\underbrace{j, j, \ldots, j}_{n}$ for some j, and the roll right before this was not a j. With probability $1/6$, the next roll continues the streak, but with probability $5/6$, we need to start over as far as achieving $n + 1$ of the same value in a row is concerned. That is, T_{n+1} is $T_n + 1$ with probability $1/6$ and is $T_n + T'_{n+1}$ with probability $5/6$, where T'_{n+1} is independent of T_n and has the same distribution as T_{n+1}. For example, if we observe $(4, 1, 1, 6, 3, 2, 3, 3, 1, 5, 5, 5)$ then $T_3 = 12$ occurs; if the next roll is also a 5 then $T_4 = 13$ occurs, but otherwise we need to start anew. Therefore, $a_1 = 1$ and for $n \geq 1$,

$$a_{n+1} = \frac{a_n + 1}{6} + \frac{5(a_n + a_{n+1})}{6},$$

which simplifies to

$$a_{n+1} = 6a_n + 1.$$

(d) Using the recursion from above, $a_1 = 1, a_2 = 1 + 6, a_3 = 1 + 6(1 + 6) = 1 + 6 + 6^2, a_4 = 1 + 6 + 6^2 + 6^3$, etc. This suggests that

$$a_n = 1 + 6 + 6^2 + \cdots + 6^{n-1},$$

a fact that we can prove by induction: $a_1 = 1$ and if $a_k = 1 + 6 + 6^2 + \cdots + 6^{k-1}$, then $a_{k+1} = 6a_k + 1 = 1 + 6 + 6^2 + \cdots + 6^k$. Summing the geometric series,

$$a_n = 1 + 6 + 6^2 + \cdots + 6^{n-1} = \frac{1 - 6^n}{1 - 6} = \frac{6^n - 1}{5}.$$

In particular, $a_7 = (6^7 - 1)/5 = 55987$.

Conditional expectation given a random variable

15. Ⓢ Let X_1, X_2 be i.i.d., and let $\bar{X} = \frac{1}{2}(X_1 + X_2)$ be the sample mean. In many statistics problems, it is useful or important to obtain a conditional expectation given $\bar{X}$. As an example of this, find $E(w_1 X_1 + w_2 X_2 | \bar{X})$, where w_1, w_2 are constants with $w_1 + w_2 = 1$.

Solution: By symmetry $E(X_1 | \bar{X}) = E(X_2 | \bar{X})$ and by linearity and taking out what's known, $E(X_1 | \bar{X}) + E(X_2 | \bar{X}) = E(X_1 + X_2 | \bar{X}) = X_1 + X_2$. So $E(X_1 | \bar{X}) = E(X_2 | \bar{X}) = \bar{X}$ (see also Example 9.3.6). Thus,

$$E(w_1 X_1 + w_2 X_2 | \bar{X}) = w_1 E(X_1 | \bar{X}) + w_2 E(X_2 | \bar{X}) = w_1 \bar{X} + w_2 \bar{X} = \bar{X}.$$

16. Let $X_1, X_2, \ldots$ be i.i.d. r.v.s with mean 0, and let $S_n = X_1 + \cdots + X_n$. As shown in Example 9.3.6, the expected value of the first term given the sum of the first n terms is

$$E(X_1 | S_n) = \frac{S_n}{n}.$$

Generalize this result by finding $E(S_k | S_n)$ for all positive integers k and n.

Solution: By linearity and Example 9.3.6,

$$E(S_k | S_n) = E(X_1 | S_n) + E(X_2 | S_n) + \cdots + E(X_k | S_n) = \frac{kS_n}{n}.$$

17. Ⓢ Consider a group of n roommate pairs at a college (so there are $2n$ students). Each of these $2n$ students independently decides randomly whether to take a certain course, with probability p of success (where "success" is defined as taking the course).

Let N be the number of students among these $2n$ who take the course, and let X be

the number of roommate pairs where both roommates in the pair take the course. Find $E(X)$ and $E(X|N)$.

Solution: Create an indicator r.v. I_j for the jth roommate pair, equal to 1 if both take the course. The expected value of such an indicator r.v. is p^2, so $E(X) = np^2$ by symmetry and linearity. Similarly, $E(X|N) = nE(I_1|N)$. We have

$$E(I_1|N) = \frac{N}{2n}\frac{N-1}{2n-1}$$

since given that N of the $2n$ students take the course, the probability is $\frac{N}{2n}$ that any particular student takes Stat 110 (the p no longer matters), and given that one particular student in a roommate pair takes the course, the probability that the other roommate does is $\frac{N-1}{2n-1}$. Or write $E(I_1|N) = \frac{\binom{N}{2}}{\binom{2n}{2}}$, since given N, the number of students in the first roommate pair who are in the course is Hypergeometric! Thus,

$$E(X|N) = nE(I_1|N) = \frac{N(N-1)}{2}\frac{1}{2n-1}.$$

Historical note: An equivalent problem was first solved in the 1760s by Daniel Bernoulli, a nephew of Jacob Bernoulli. (The Bernoulli distribution is named after Jacob Bernoulli.)

18. Ⓢ Show that $E((Y - E(Y|X))^2|X) = E(Y^2|X) - (E(Y|X))^2$, so these two expressions for $\text{Var}(Y|X)$ agree.

Solution: This is the conditional version of the fact that

$$\text{Var}(Y) = E((Y - E(Y))^2) = E(Y^2) - (E(Y))^2,$$

and so must be true since conditional expectations *are* expectations, just as conditional probabilities are probabilities. Algebraically, letting $g(X) = E(Y|X)$ we have

$$E((Y-E(Y|X))^2|X) = E(Y^2-2Yg(X)+g(X)^2|X) = E(Y^2|X)-2E(Yg(X)|X)+E(g(X)^2|X),$$

and $E(Yg(X)|X) = g(X)E(Y|X) = g(X)^2, E(g(X)^2|X) = g(X)^2$ by taking out what's known, so the righthand side above simplifies to $E(Y^2|X) - g(X)^2$.

19. Let X be the height of a randomly chosen adult man, and Y be his father's height, where X and Y have been standardized to have mean 0 and standard deviation 1. Suppose that (X,Y) is Bivariate Normal, with $X, Y \sim \mathcal{N}(0,1)$ and $\text{Corr}(X,Y) = \rho$.

(a) Let $y = ax + b$ be the equation of the best line for predicting Y from X (in the sense of minimizing the mean squared error), e.g., if we were to observe $X = 1.3$ then we would predict that Y is $1.3a + b$. Now suppose that we want to use Y to predict X, rather than using X to predict Y. Give and explain an *intuitive guess* for what the slope is of the best line for predicting X from Y.

(b) Find a constant c (in terms of ρ) and an r.v. V such that $Y = cX + V$, with V independent of X.

Hint: Start by finding c such that $\text{Cov}(X, Y - cX) = 0$.

(c) Find a constant d (in terms of ρ) and an r.v. W such that $X = dY + W$, with W independent of Y.

(d) Find $E(Y|X)$ and $E(X|Y)$.

(e) Reconcile (a) and (d), giving a clear and correct intuitive explanation.

Solution:

(a) Here we are interested in the inverse problem, predicting X from Y rather than vice versa. Solving $y = ax + b$ for x gives $x = \frac{1}{a}(y - b)$, so one intuitive guess would be to use $\frac{1}{a}$, inverting the slope of the original prediction line.

(b) Setting $\mathrm{Cov}(X, Y - cX) = \mathrm{Cov}(X, Y) - c\mathrm{Cov}(X, X) = \rho - c = 0$, we have $c = \rho$. So define $V = Y - \rho X$. Then (X, V) is Bivariate Normal since any linear combination of X and V can be expressed as a linear combination of X and Y. So X and V are independent since, by construction, they are uncorrelated.

(c) By the same method as in (b), $d = \rho$ and $W = X - \rho Y$ are as desired.

(d) Continuing as in (b),

$$E(Y|X) = E(\rho X + V|X) = \rho E(X|X) + E(V|X) = \rho X + E(V) = \rho X.$$

Continuing as in (c),

$$E(X|Y) = E(\rho Y + W|Y) = \rho E(Y|Y) + E(W|Y) = \rho Y + E(W) = \rho Y.$$

(e) Part (d) shows that in this setting the slope of the line for predicting X from Y is the *same* as the slope for predicting Y from X, not the reciprocal! This must be true by symmetry since $\rho = \mathrm{Corr}(X, Y) = \mathrm{Corr}(Y, X)$. Note that this makes sense in the simple case $\rho = 0$ since then X and Y are independent, so the best guess for X given Y is $E(X) = 0$, whereas inverting ρ would be disastrous.

If it were true that X were perfectly linearly related to Y, then random (X, Y) pairs would always be on the line $y = \rho x$, and then it would make sense to invert ρ; but this only happens in the extreme cases where ρ is -1 or 1 (when we *do* have $1/\rho = \rho$).

Now assume $0 < \rho < 1$, as is true for this problem in the real world. To predict a son's height from his father's height, it would be silly to just use the father's height as the prediction, since height isn't inherited perfectly, but it would also be silly to ignore the father's height. A more reasonable approach is a compromise, giving a prediction that is in between the father's height and the population mean. If ρ is close to 1, much more weight should be put on the father's height; if ρ is close to 0, much more weight should be put on the population mean.

For example, the son of a very tall father should be predicted to be tall, but not *as* tall as his father. This phenomenon is called *regression toward the mean* (the term "regression" in statistics originates from this). It was studied by Sir Francis Galton in 1886. Note that regression toward the mean occurs in *both* directions of time, e.g., the father of a very tall son should be predicted to be tall, but not *as* tall as his son.

20. Let $\mathbf{X} \sim \mathrm{Mult}_5(n, \mathbf{p})$.

(a) Find $E(X_1|X_2)$ and $\mathrm{Var}(X_1|X_2)$.

(b) Find $E(X_1|X_2 + X_3)$.

Solution:

(a) Consider placing n objects into 5 categories. Given that $X_2 = k$, we have $n - k$ objects remaining that could be in category 1, each of which has probability

$$P(\text{category 1}|\text{not category 2}) = \frac{P(\text{category 1})}{P(\text{not category 2})} = \frac{p_1}{1 - p_2}$$

of being in category 1, independently. So $X_1|X_2 = k \sim \mathrm{Bin}(n - k, p_1/(1 - p_2))$, which shows that

$$E(X_1|X_2) = (n - X_2)\frac{p_1}{1 - p_2},$$

$$\mathrm{Var}(X_1|X_2) = (n - X_2)\frac{p_1}{1 - p_2}\left(1 - \frac{p_1}{1 - p_2}\right).$$

(b) Reasoning as in (a), $X_1|X_2 + X_3 = k \sim \text{Bin}(n - k, p_1/(1 - p_2 - p_3))$, so

$$E(X_1|X_2 + X_3) = (n - X_2 - X_3)\frac{p_1}{1 - p_2 - p_3},$$

$$\text{Var}(X_1|X_2 + X_3) = (n - X_2 - X_3)\frac{p_1}{1 - p_2 - p_3}\left(1 - \frac{p_1}{1 - p_2 - p_3}\right).$$

21. Let Y be a discrete r.v., A be an event with $0 < P(A) < 1$, and I_A be the indicator r.v. for A.

(a) Explain precisely how the r.v. $E(Y|I_A)$ relates to the numbers $E(Y|A)$ and $E(Y|A^c)$.

(b) Show that $E(Y|A) = E(YI_A)/P(A)$, directly from the definitions of expectation and conditional expectation.

Hint: First find the PMF of YI_A in terms of $P(A)$ and the conditional PMF of Y given A.

(c) Use (b) to give a short proof of the fact that $E(Y) = E(Y|A)P(A) + E(Y|A^c)P(A^c)$.

Solution:

(a) The possible values of $E(Y|I_A)$ are $E(Y|A)$ and $E(Y|A^c)$. If A occurs, then $E(Y|I_A)$ will crystallize to $E(Y|I_A = 1)$, which is $E(Y|A)$; if A does not occur, then $E(Y|I_A)$ will crystallize to $E(Y|I_A = 0)$, which is $E(Y|A^c)$.

(b) Let $X = YI_A$ and $B = \{y : P(Y = y|A) > 0\}$. The support of X is $B \cup \{0\}$, and the PMF of X is

$$P(X = x) = P(X = x|A)P(A) + P(X = x|A^c)P(A^c)$$
$$= \begin{cases} P(Y = x|A)P(A), & \text{for } x \neq 0, \\ P(Y = 0|A)P(A) + P(A^c), & \text{for } x = 0. \end{cases}$$

So

$$E(YI_A) = \sum_{x \in B} xP(Y = x|A)P(A) = P(A)\sum_{y \in B} yP(Y = y|A) = E(Y|A)P(A).$$

(c) By linearity and (b),

$$E(Y) = E(YI_A + YI_{A^c}) = E(YI_A) + E(YI_{A^c}) = E(Y|A)P(A) + E(Y|A^c)P(A^c).$$

22. Show that the following version of LOTP follows from Adam's law: for any event A and continuous r.v. X with PDF f_X,

$$P(A) = \int_{-\infty}^{\infty} P(A|X = x)f_X(x)dx.$$

Solution: Let I be the indicator r.v. for A. Let $g(x) = E(I|X = x)$, so $g(X) = E(I|X)$. By the fundamental bridge, Adam's law, and LOTUS,

$$P(A) = E(I) = E(E(I|X)) = E(g(X)) = \int_{-\infty}^{\infty} g(x)f_X(x)dx = \int_{-\infty}^{\infty} P(A|X = x)f_X(x)dx.$$

23. Ⓢ Let X and Y be random variables with finite variances, and let $W = Y - E(Y|X)$. This is a *residual*: the difference between the true value of Y and the predicted value of Y based on X.

(a) Compute $E(W)$ and $E(W|X)$.

(b) Compute $\text{Var}(W)$, for the case that $W|X \sim \mathcal{N}(0, X^2)$ with $X \sim \mathcal{N}(0, 1)$.

Solution:

(a) Adam's law, taking out what's known, and linearity give

$$
\begin{aligned}
E(W) &= EY - E(E(Y|X)) = EY - EY = 0, \\
E(W|X) &= E(Y|X) - E(E(Y|X)|X) = E(Y|X) - E(Y|X) = 0.
\end{aligned}
$$

(b) Eve's Law gives

$$
\text{Var}(W) = \text{Var}(E(W|X)) + E(\text{Var}(W|X)) = \text{Var}(0) + E(X^2) = 0 + 1 = 1.
$$

24. Ⓢ One of two identical-looking coins is picked from a hat randomly, where one coin has probability p_1 of Heads and the other has probability p_2 of Heads. Let X be the number of Heads after flipping the chosen coin n times. Find the mean and variance of X.

Solution: The distribution of X is a *mixture* of two Binomials; this is *not* Binomial unless $p_1 = p_2$. Let I be the indicator of having the p_1 coin. Then

$$
E(X) = E(X|I = 1)P(I = 1) + E(X|I = 0)P(I = 0) = \frac{1}{2}n(p_1 + p_2).
$$

Alternatively, we can represent X as $X = IX_1 + (1 - I)X_2$ with $X_j \sim \text{Bin}(n, p_j)$, and I, X_1, X_2 independent. Then

$$
E(X) = E(E(X|I)) = E(Inp_1 + (1 - I)np_2) = \frac{1}{2}n(p_1 + p_2).
$$

For the variance, note that it is *not* valid to say "$\text{Var}(X) = \text{Var}(X|I = 1)P(I = 1) + \text{Var}(X|I = 0)P(I = 0)$"; an extreme example of this mistake would be claiming that "$\text{Var}(I) = 0$ since $\text{Var}(I|I = 1)P(I = 1) + \text{Var}(I|I = 0)P(I = 0) = 0$"; of course, $\text{Var}(I) = \frac{1}{4}$). Instead, we can use Eve's Law:

$$
\text{Var}(X) = E(\text{Var}(X|I)) + \text{Var}(E(X|I)),
$$

where $\text{Var}(X|I) = Inp_1(1 - p_1) + (1 - I)np_2(1 - p_2)$ is $np_1(1 - p_1)$ with probability $1/2$ and $np_2(1 - p_2)$ with probability $1/2$, and $E(X|I) = Inp_1 + (1 - I)np_2$ is np_1 or np_2 with probability $\frac{1}{2}$ each, so

$$
\text{Var}(X) = \frac{1}{2}\left(np_1(1 - p_1) + np_2(1 - p_2)\right) + \frac{1}{4}n^2(p_1 - p_2)^2.
$$

25. Kelly makes a series of n bets, each of which she has probability p of winning, independently. Initially, she has x_0 dollars. Let X_j be the amount she has immediately after her jth bet is settled. Let f be a constant in $(0, 1)$, called the *betting fraction*. On each bet, Kelly wagers a fraction f of her wealth, and then she either wins or loses that amount. For example, if her current wealth is \$100 and $f = 0.25$, then she bets \$25 and either gains or loses that amount. (A famous choice when $p > 1/2$ is $f = 2p - 1$, which is known as the *Kelly criterion*.) Find $E(X_n)$ (in terms of n, p, f, x_0).

Hint: First find $E(X_{j+1}|X_j)$.

Solution: We have

$$
E(X_{j+1}|X_j) = (1 - f)X_j + p \cdot 2fX_j + (1 - p) \cdot 0 = (1 - f + 2pf)X_j
$$

so

$$E(X_{j+1}) = (1 - f + 2pf)E(X_j).$$

Hence,

$$E(X_n) = (1 - f + 2pf)^n x_0.$$

26. Let $N \sim \text{Pois}(\lambda_1)$ be the number of movies that will be released next year. Suppose that for each movie the number of tickets sold is $\text{Pois}(\lambda_2)$, independently. Find the mean and variance of the number of movie tickets that will be sold next year.

    Solution: Let $X_j \sim \text{Pois}(\lambda_2)$ be the number of tickets sold for the jth movie released next year, and $X = X_1 + \cdots + X_N$ be the total number of tickets sold for movies released next year. By Adam's law,

    $$E(X) = E(E(X|N)) = E(N\lambda_2) = \lambda_1\lambda_2.$$

    By Eve's law,

    $$\text{Var}(X) = E(\text{Var}(X|N)) + \text{Var}(E(X|N)) = E(N\lambda_2) + \text{Var}(N\lambda_2) = \lambda_1\lambda_2 + \lambda_1\lambda_2^2.$$

    Alternatively, we can obtain these results by applying Example 9.6.1.

27. A party is being held from 8:00 pm to midnight on a certain night, and $N \sim \text{Pois}(\lambda)$ people are going to show up. They will all arrive at uniformly random times while the party is going on, independently of each other and of N.

    (a) Find the expected time at which the first person arrives, given that at least one person shows up. Give both an exact answer in terms of λ, measured in minutes after 8:00 pm, and an answer rounded to the nearest minute for $\lambda = 20$, expressed in time notation (e.g., 8:20 pm).

    (b) Find the expected time at which the last person arrives, given that at least one person shows up. As in (a), give both an exact answer and an answer rounded to the nearest minute for $\lambda = 20$.

    Solution:

    (a) To simplify notation, write E_1 to denote an expectation that is conditioned on $N \geq 1$. In units where time 0 is the start of the party and 240 minutes is 1 unit of time, we want to find $E_1(\min(T_1, \ldots, T_N))$, with the T_j's i.i.d. $\text{Unif}(0, 1)$. This involves the minimum of a *random* number of random variables, but we can simplify the problem by conditioning on N. Let $L = \min(T_1, \ldots, T_N)$. By Adam's law,

    $$E_1(L) = E_1(E_1(L|N)) = E_1\left(\frac{1}{N+1}\right),$$

    since the minimum of n i.i.d. $\text{Unif}(0, 1)$ r.v.s is a $\text{Beta}(1, n)$ r.v., which has mean $1/(n+1)$. The conditional PMF of N given $N \geq 1$ is

    $$P(N = n|N \geq 1) = \frac{P(N = n, N \geq 1)}{P(N \geq 1)} = \frac{P(N = n)}{1 - P(N = 0)} = \frac{e^{-\lambda}\lambda^n/n!}{1 - e^{-\lambda}},$$

    for $n = 1, 2, \ldots$. By LOTUS,

    $$E_1\left(\frac{1}{N+1}\right) = \frac{e^{-\lambda}}{1 - e^{-\lambda}} \sum_{n=1}^{\infty} \frac{1}{n+1} \cdot \frac{\lambda^n}{n!} = \frac{e^{-\lambda}}{\lambda(1 - e^{-\lambda})} \sum_{n=1}^{\infty} \frac{\lambda^{n+1}}{(n+1)!} = \frac{e^{-\lambda}(e^{\lambda} - 1 - \lambda)}{\lambda(1 - e^{-\lambda})},$$

    using the fact that the sum is the Taylor series for e^{λ} except with the first two terms missing. This can be further simplified to

    $$E_1\left(\frac{1}{N+1}\right) = \frac{1 - e^{-\lambda} - \lambda e^{-\lambda}}{\lambda(1 - e^{-\lambda})} = \frac{1}{\lambda} - \frac{1}{e^{\lambda} - 1}.$$

So the expected time is $240(\frac{1}{\lambda} - \frac{1}{e^{\lambda}-1})$ minutes after 8:00 pm.

For $\lambda = 20$, this is extremely close to $240/20 = 12$, which gives an expected first arrival time of 8:12 pm.

(b) Measuring time backward from midnight rather than forward from 8:00 pm, the problem has the same structure as (a), so the expected time is $240(\frac{1}{\lambda} - \frac{1}{e^{\lambda}-1})$ minutes before midnight, which is $240 - 240(\frac{1}{\lambda} - \frac{1}{e^{\lambda}-1})$ minutes after 8:00 pm.

For $\lambda = 20$, the expected last arrival time is approximately 12 minutes before midnight, which is 11:48 pm.

Alternatively, we can apply the same method as in (a), and use the fact that the maximum of n i.i.d. Unif$(0, 1)$ r.v.s is a Beta$(n, 1)$ r.v., which has mean $n/(n + 1)$. Writing $n/(n + 1) = 1 - 1/(n + 1)$, we can then apply the LOTUS result from (a).

28. (S) We wish to estimate an unknown parameter θ, based on an r.v. X we will get to observe. As in the Bayesian perspective, assume that X and θ have a joint distribution. Let $\hat{\theta}$ be the estimator (which is a function of X). Then $\hat{\theta}$ is said to be *unbiased* if $E(\hat{\theta}|\theta) = \theta$, and $\hat{\theta}$ is said to be the *Bayes procedure* if $E(\theta|X) = \hat{\theta}$.

(a) Let $\hat{\theta}$ be unbiased. Find $E(\hat{\theta} - \theta)^2$ (the average squared difference between the estimator and the true value of θ), in terms of marginal moments of $\hat{\theta}$ and θ.

Hint: Condition on θ.

(b) Repeat (a), except in this part suppose that $\hat{\theta}$ is the *Bayes procedure* rather than assuming that it is unbiased.

Hint: Condition on X.

(c) Show that it is *impossible* for $\hat{\theta}$ to be both the Bayes procedure and unbiased, except in silly problems where we get to know θ perfectly by observing X.

Hint: If Y is a nonnegative r.v. with mean 0, then $P(Y = 0) = 1$.

Solution:

(a) Conditioning on θ, we have

$$
\begin{aligned}
E(\hat{\theta} - \theta)^2 &= E(E(\hat{\theta}^2 - 2\hat{\theta}\theta + \theta^2|\theta)) \\
&= E(E(\hat{\theta}^2|\theta)) - E(E(2\hat{\theta}\theta|\theta)) + E(E(\theta^2|\theta)) \\
&= E(\hat{\theta}^2) - 2E(\theta E(\hat{\theta}|\theta)) + E(\theta^2) \\
&= E(\hat{\theta}^2) - 2E(\theta^2) + E(\theta^2) \\
&= E(\hat{\theta}^2) - E(\theta^2).
\end{aligned}
$$

(b) By the same argument as for (a) except now conditioning on X, we have

$$
\begin{aligned}
E(\hat{\theta} - \theta)^2 &= E(E(\hat{\theta}^2 - 2\hat{\theta}\theta + \theta^2|X)) \\
&= E(E(\hat{\theta}^2|X)) - E(E(2\hat{\theta}\theta|X)) + E(E(\theta^2|X)) \\
&= E(\hat{\theta}^2) - 2E(\hat{\theta}^2) + E(\theta^2) \\
&= E(\theta^2) - E(\hat{\theta}^2).
\end{aligned}
$$

(c) Suppose that $\hat{\theta}$ is both the Bayes procedure and unbiased. By the above, we have $E(\hat{\theta} - \theta)^2 = a$ and $E(\hat{\theta} - \theta)^2 = -a$, where $a = E(\hat{\theta}^2) - E(\theta^2)$. But that implies $a = 0$, which means that $\hat{\theta} = \theta$ (with probability 1). That can only happen in the extreme situation where the observed data reveal the true θ *perfectly*; in practice, nature is much more elusive and does not reveal its deepest secrets with such alacrity.

29. Show that if $E(Y|X) = c$ is a constant, then X and Y are uncorrelated.

    Hint: Use Adam's law to find $E(Y)$ and $E(XY)$.

    Solution: Let $E(Y|X) = c$. By Adam's law and taking out what's known,

    $$E(Y) = E(E(Y|X)) = E(c) = c,$$

    $$E(XY) = E(E(XY|X)) = E(XE(Y|X)) = E(cX) = cE(X) = E(X)E(Y).$$

    So X and Y are uncorrelated.

30. Show by example that it is possible to have uncorrelated X and Y such that $E(Y|X)$ is not a constant.

    Hint: Consider a standard Normal and its square.

    Solution: Let $X \sim \mathcal{N}(0, 1)$ and $Y = X^2$. Then X and Y are uncorrelated since

    $$\text{Cov}(X, Y) = E(XY) - E(X)E(Y) = E(X^3) - E(X)E(X^2) = 0.$$

    But

    $$E(Y|X) = E(X^2|X) = X^2,$$

    which is not a constant.

31. ⓢ Emails arrive one at a time in an inbox. Let T_n be the time at which the nth email arrives (measured on a continuous scale from some starting point in time). Suppose that the waiting times between emails are i.i.d. Expo(λ), i.e., $T_1, T_2 - T_1, T_3 - T_2, \ldots$ are i.i.d. Expo(λ).

    Each email is non-spam with probability p, and spam with probability $q = 1 - p$ (independently of the other emails and of the waiting times). Let X be the time at which the first non-spam email arrives (so X is a continuous r.v., with $X = T_1$ if the 1st email is non-spam, $X = T_2$ if the 1st email is spam but the 2nd one isn't, etc.).

    (a) Find the mean and variance of X.

    (b) Find the MGF of X. What famous distribution does this imply that X has (be sure to state its parameter values)?

    Hint for both parts: Let N be the number of emails until the first non-spam (including that one), and write X as a sum of N terms; then condition on N.

    Solution:

    (a) Write $X = X_1 + X_2 + \cdots + X_N$, where X_j is the time from the $(j - 1)$th to the jth email for $j \geq 2$, and $X_1 = T_1$. Then $N - 1 \sim \text{Geom}(p)$, so

    $$E(X) = E(E(X|N)) = E(N\frac{1}{\lambda}) = \frac{1}{p\lambda}.$$

    And

    $$\text{Var}(X) = E(\text{Var}(X|N)) + \text{Var}(E(X|N)) = E(N\frac{1}{\lambda^2}) + \text{Var}(N\frac{1}{\lambda}),$$

    which is

    $$\frac{1}{p\lambda^2} + \frac{1 - p}{p^2\lambda^2} = \frac{1}{p^2\lambda^2}.$$

    (b) Again conditioning on N, the MGF is

    $$E(e^{tX}) = E(E(e^{tX_1}e^{tX_2}\cdots e^{tX_N}|N)) = E\left(E(e^{tX_1}|N)E(e^{tX_2}|N)\ldots E(e^{tX_N}|N)\right) = E(M_1(t)^N),$$

where $M_1(t)$ is the MGF of X_1 (which is $\frac{\lambda}{\lambda-t}$ for $t < \lambda$). By LOTUS, this is

$$p\sum_{n=1}^{\infty} M_1(t)^n q^{n-1} = \frac{p}{q}\sum_{n=1}^{\infty}(qM_1(t))^n = \frac{p}{q}\frac{qM_1(t)}{1-qM_1(t)} = \frac{\frac{p\lambda}{\lambda-t}}{1-\frac{q\lambda}{\lambda-t}} = \frac{p\lambda}{p\lambda-t}$$

for $t < p\lambda$ (as we need $qM_1(t) < 1$ for the series to converge). This is the $\text{Expo}(p\lambda)$ MGF, so $X \sim \text{Expo}(p\lambda)$.

32. Customers arrive at a store according to a Poisson process of rate λ customers per hour. Each makes a purchase with probability p, independently. Given that a customer makes a purchase, the amount spent has mean μ (in dollars) and variance σ^2.

(a) Find the mean and variance of how much a random customer spends (note that the customer may spend nothing).

(b) Find the mean and variance of the revenue the store obtains in an 8-hour time interval, using (a) and results from this chapter.

(c) Find the mean and variance of the revenue the store obtains in an 8-hour time interval, using the chicken-egg story and results from this chapter.

Solution:

(a) Let X be the amount that a random customer spends, and I be the indicator of the random customer making a purchase. Then

$$E(X) = E(X|I=1)P(I=1) + E(X|I=0)P(I=0) = \mu p,$$
$$E(X^2) = E(X^2|I=1)P(I=1) + E(X^2|I=0)P(I=0) = (\sigma^2+\mu^2)p,$$
$$\text{Var}(X) = E(X^2) - (EX)^2 = (\sigma^2+\mu^2)p - \mu^2 p^2 = \sigma^2 p + \mu^2 p(1-p).$$

(b) Let $N \sim \text{Pois}(8\lambda)$ be the number of customers who arrive in 8 hours. Let R be the revenue from those customers. By Example 9.6.1,

$$E(R) = 8\lambda\mu p,$$
$$\text{Var}(R) = 8\lambda(\sigma^2 p + \mu^2 p(1-p)) + 8\lambda\mu^2 p^2 = 8\lambda p(\sigma^2+\mu^2).$$

(c) By the chicken-egg story, the number of customers who arrive in 8 hours *and will make a purchase* is $\text{Pois}(8\lambda p)$. Each of those customers spends an amount with mean μ and variance σ^2. As in (b), let R be the revenue in those 8 hours (which is the total amount spent by the customers who make purchases). Then by Example 9.6.1,

$$E(R) = 8\lambda\mu p,$$
$$\text{Var}(R) = 8\lambda p\sigma^2 + 8\lambda p\mu^2,$$

in agreement with (b).

33. Fred's beloved computer will last an $\text{Expo}(\lambda)$ amount of time until it has a malfunction. When that happens, Fred will try to get it fixed. With probability p, he will be able to get it fixed. If he is able to get it fixed, the computer is good as new again and will last an additional, independent $\text{Expo}(\lambda)$ amount of time until the next malfunction (when again he is able to get it fixed with probability p, and so on). If after any malfunction Fred is unable to get it fixed, he will buy a new computer. Find the expected amount of time until Fred buys a new computer. (Assume that the time spent on computer diagnosis, repair, and shopping is negligible.)

Solution: Let $N \sim \text{FS}(1-p)$ be the number of malfunctions of the computer until Fred can no longer get it fixed (including the last malfunction). Let T_1 be the time until

the first malfunction, T_2 be the additional time until the second malfunction, etc. Let $T = T_1 + \cdots + T_N$ be the time until Fred buys a new computer. Then

$$E(T) = E(E(T_1 + \cdots + T_N | N)) = E(N/\lambda) = \frac{1}{\lambda(1-p)}.$$

Alternatively, we can condition on whether the first malfunction is fixable:

$$E(T) = (1-p)E(T_1) + pE(T_1 + T),$$

so

$$E(T) = (1-p) \cdot \frac{1}{\lambda} + \frac{p}{\lambda} + pE(T),$$

which again yields

$$E(T) = \frac{1}{\lambda(1-p)}.$$

34. A green die is rolled until it lands 1 for the first time. An orange die is rolled until it lands 6 for the first time. The dice are fair, six-sided dice. Let T_1 be the sum of the values of the rolls of the green die (including the 1 at the end) and T_6 be the sum of the values of the rolls of the orange die (including the 6 at the end). Two students are debating whether $E(T_1) = E(T_6)$ or $E(T_1) < E(T_6)$. They kindly gave permission to quote their arguments here.

Student A: We have $E(T_1) = E(T_6)$. By Adam's law, the expected sum of the rolls of a die is the expected number of rolls times the expected value of one roll, and each of these factors is the same for the two dice. In more detail, let N_1 be the number of rolls of the green die and N_6 be the number of rolls of the orange die. By Adam's law and linearity,

$$E(T_1) = E(E(T_1 | N_1)) = E(3.5N_1) = 3.5E(N_1),$$

and the same method applied to the orange die gives $3.5E(N_6)$, which equals $3.5E(N_1)$.

Student B: Actually, $E(T_1) < E(T_6)$. I agree that the expected number of rolls is the same for the two dice, but the key difference is that we *know* the last roll is a 1 for the green die and a 6 for the orange die. The expected totals are the same for the two dice *excluding* the last roll of each, and then including the last roll makes $E(T_1) < E(T_6)$.

(a) Discuss in words the extent to which Student A's argument is convincing and correct.

(b) Discuss in words the extent to which Student B's argument is convincing and correct.

(c) Give careful derivations of $E(T_1)$ and $E(T_6)$.

Solution:

(a) Student A is right that using Adam's law, conditioning on the number of rolls, is a good idea. And we will show in (c) that Student A is correct that $E(T_1) = E(T_6)$. However, Student A's argument is unconvincing since, as pointed out by Student B, the argument does not distinguish between the last roll of a die and the previous rolls of the die. Student A is not accounting for the fact that for the green die we *know* the last roll is a 1 and the other rolls are not 1's, and likewise for the orange die.

(b) Student B is right to point out that it is important to distinguish between the last roll of a die and the previous rolls of the die, but incorrectly says that the expected totals of the dice are the same if the last rolls are excluded. This is wrong since the expected value of a roll conditioned not to be a 1 is higher than the expected value of a roll conditioned not to be a 6.

(c) We will show that

$$E(T_1) = E(T_6) = 21,$$

by combining the good ideas and correcting the mistakes from the students' arguments. Let S_1 be the sum of the values of the roll of the green die except for the last roll, and S_6 be the sum of the values of the roll of the orange die except for the last roll, so $T_1 = S_1 + 1$, $T_6 = S_6 + 6$. Write

$$S_1 = X_1 + \cdots + X_N,$$

where each X_i is an individual roll and N is the random number of rolls that are added up to form S_1. (It is possible that $N = 0$, in which case we take $S_1 = 0$.) Then $N \sim \text{Geom}(1/6)$. The X_i are conditioned *not* to equal 1, and in particular they are i.i.d. Discrete Uniform on $\{2, 3, \ldots, 6\}$, independent of N. By Adam's law,

$$E(S_1) = E(E(S_1|N)) = E(N)E(X_1) = \frac{5/6}{1/6} \cdot \frac{2+3+4+5+6}{5} = 2+3+4+5+6 = 20,$$

so

$$E(T_1) = 20 + 1 = 21.$$

Similarly,

$$E(S_6) = \frac{5/6}{1/6} \cdot \frac{1+2+3+4+5}{5} = 1+2+3+4+5 = 15,$$

so

$$E(T_6) = 15 + 6 = 21.$$

35. Ⓢ Judit plays in a total of $N \sim \text{Geom}(s)$ chess tournaments in her career. Suppose that in each tournament she has probability p of winning the tournament, independently. Let T be the number of tournaments she wins in her career.

(a) Find the mean and variance of T.

(b) Find the MGF of T. What is the name of this distribution (with its parameters)?

Solution:

(a) We have $T|N \sim \text{Bin}(N, p)$. By Adam's Law,

$$E(T) = E(E(T|N)) = E(Np) = p(1-s)/s.$$

By Eve's Law,

$$\begin{aligned}
\text{Var}(T) &= E(\text{Var}(T|N)) + \text{Var}(E(T|N)) \\
&= E(Np(1-p)) + \text{Var}(Np) \\
&= p(1-p)(1-s)/s + p^2(1-s)/s^2 \\
&= \frac{p(1-s)(s + (1-s)p)}{s^2}.
\end{aligned}$$

(b) Let $I_j \sim \text{Bern}(p)$ be the indicator of Judit winning the jth tournament. Then

$$\begin{aligned}
E(e^{tT}) &= E(E(e^{tT}|N)) \\
&= E((pe^t + q)^N) \\
&= s\sum_{n=0}^{\infty}(pe^t + 1 - p)^n(1 - s)^n \\
&= \frac{s}{1 - (1-s)(pe^t + 1 - p)}.
\end{aligned}$$

This is reminiscent of the Geometric MGF, which was derived in Example 6.4.3. If

$T \sim \text{Geom}(\theta)$, we have $\theta = \frac{s}{s+p(1-s)}$, as found by setting $E(T) = \frac{1-\theta}{\theta}$ or by finding $\text{Var}(T)/E(T)$. Writing the MGF of T as

$$E(e^{tT}) = \frac{s}{s + (1-s)p - (1-s)pe^t} = \frac{\frac{s}{s+(1-s)p}}{1 - \frac{(1-s)p}{s+(1-s)p}e^t},$$

we see that $T \sim \text{Geom}(\theta)$, with $\theta = \frac{s}{s+(1-s)p}$. Note that this is consistent with (a).

The distribution of T can also be obtained by a story proof. Imagine that just before each tournament she may play in, Judit retires with probability s (if she retires, she does not play in that or future tournaments). Her tournament history can be written as a sequence of W (win), L (lose), R (retire), ending in the first R, where the probabilities of W, L, R are $(1-s)p, (1-s)(1-p), s$ respectively. For calculating T, the losses can be ignored: we want to count the number of W's before the R. The probability that a result is R given that it is W or R is $\frac{s}{s+(1-s)p}$, so we again have

$$T \sim \text{Geom}\left(\frac{s}{s + (1-s)p}\right).$$

36. In Story 8.4.5, we showed (among other things) that if $\lambda \sim \text{Gamma}(r_0, b_0)$ and $Y|\lambda \sim \text{Pois}(\lambda)$, then the marginal distribution of Y is $\text{NBin}(r_0, b_0/(b_0 + 1))$. Derive this result using Adam's law and MGFs.

Hint: Consider the conditional MGF of $Y|\lambda$.

Solution: For $Y|\lambda \sim \text{Pois}(\lambda)$, the conditional MGF of $Y|\lambda$ is

$$E(e^{tY}|\lambda) = e^{\lambda(e^t-1)}.$$

By Adam's law, the unconditional MGF of Y is

$$E(e^{tY}) = E(e^{\lambda(e^t-1)}).$$

By LOTUS, this equals

$$\int_0^\infty e^{x(e^t-1)}\Gamma(r_0)^{-1}(b_0x)^{r_0}e^{-b_0x}x^{-1}dx = \Gamma(r_0)^{-1}b_0^{r_0}\int_0^\infty e^{-(b_0+1-e^t)x}x^{r_0}x^{-1}dx.$$

Assume that $b_0 + 1 - e^t > 0$, so that the integral will converge. Making the change of variables $u = (b_0 + 1 - e^t)x$, we have

$$E(e^{tY}) = \frac{b_0^{r_0}}{\Gamma(r_0)(b_0 + 1 - e^t)^{r_0}}\int_0^\infty e^{-u}u^{r_0}u^{-1}du = \left(\frac{b_0}{(b_0 + 1 - e^t)}\right)^{r_0}.$$

In Example 6.4.3, we saw that the $\text{Geom}(p)$ MGF is

$$\frac{p}{1 - (1-p)e^t}, \text{ for } (1-p)e^t < 1.$$

So the $\text{NBin}(r_0, b_0/(b_0 + 1))$ MGF is

$$\left(\frac{b_0/(b_0 + 1)}{1 - (1/(b_0 + 1))e^t}\right)^{r_0} = \left(\frac{b_0}{b_0 + 1 - e^t}\right)^{r_0}, \text{ for } \frac{e^t}{b_0 + 1} < 1.$$

Thus, the marginal distribution of Y has the same MGF as the $\text{NBin}(r_0, b_0/(b_0 + 1))$ distribution, which shows that the marginal distribution of Y is $\text{NBin}(r_0, b_0/(b_0 + 1))$.

37. Let $X_1, \ldots, X_n$ be i.i.d. r.v.s with mean μ and variance σ^2, and $n \geq 2$. A *bootstrap sample* of $X_1, \ldots, X_n$ is a sample of n r.v.s $X_1^*, \ldots, X_n^*$ formed from the X_j by sampling with replacement with equal probabilities. Let $\bar{X}^*$ denote the sample mean of the bootstrap sample:

$$\bar{X}^* = \frac{1}{n}(X_1^* + \cdots + X_n^*).$$

(a) Calculate $E(X_j^*)$ and $\text{Var}(X_j^*)$ for each j.

(b) Calculate $E(\bar{X}^*|X_1, \ldots, X_n)$ and $\text{Var}(\bar{X}^*|X_1, \ldots, X_n)$.

Hint: Conditional on $X_1, \ldots, X_n$, the X_j^* are independent, with a PMF that puts probability $1/n$ at each of the points $X_1, \ldots, X_n$. As a check, your answers should be random variables that are functions of $X_1, \ldots, X_n$.

(c) Calculate $E(\bar{X}^*)$ and $\text{Var}(\bar{X}^*)$.

(d) Explain intuitively why $\text{Var}(\bar{X}) < \text{Var}(\bar{X}^*)$.

Solution:

(a) Note that X_j^* has the same distribution as X_1, since the way in which it is obtained from $X_1, \ldots, X_n$ is not based on the values of $X_1, \ldots, X_n$ (it is chosen completely at random from $X_1, \ldots, X_n$ rather than, for example, choosing the biggest X_j). To see this more formally, let I be the index of the X_i chosen for X_j^* and let F be the CDF of X_1. Then the CDF of X_j^* is

$$P(X_j^* \leq x) = \sum_{i=1}^{n} P(X_j^* \leq x | I = i) P(I = i) = \frac{1}{n} \sum_{i=1}^{n} F(x) = F(x).$$

In particular, $E(X_j^*) = \mu$ and $\text{Var}(X_j^*) = \sigma^2$.

(b) By linearity,

$$E(\bar{X}^*|X_1, \ldots, X_n) = \frac{1}{n}(E(X_1^*|X_1, \ldots, X_n) + \cdots + E(X_n^*|X_1, \ldots, X_n)) = E(X_1^*|X_1, \ldots, X_n).$$

By the hint,

$$E(X_1^*|X_1, \ldots, X_n) = \bar{X}.$$

For the variance, the conditional independence of the X_j^* yields

$$\text{Var}(\bar{X}^*|X_1, \ldots, X_n) = \frac{n}{n^2} \text{Var}(X_1^*|X_1, \ldots, X_n) = \frac{1}{n^2} \sum_{j=1}^{n} (X_j - \bar{X})^2.$$

(c) By Adam's law,

$$E(\bar{X}^*) = E(E(\bar{X}^*|X_1, \ldots, X_n)) = E(\bar{X}) = \mu.$$

By Eve's law,

$$\text{Var}(\bar{X}^*) = E(\text{Var}(\bar{X}^*|X_1, \ldots, X_n)) + \text{Var}(E(\bar{X}^*|X_1, \ldots, X_n))$$

$$= E\left(\frac{1}{n^2} \sum_{j=1}^{n} (X_j - \bar{X})^2\right) + \text{Var}(\bar{X})$$

$$= \frac{(n-1)\sigma^2}{n^2} + \frac{\sigma^2}{n}$$

$$= \frac{\sigma^2}{n}\left(2 - \frac{1}{n}\right).$$

Above we used the fact (shown in Theorem 6.3.4) that the sample variance $\frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})^2$ is an unbiased estimator of the population variance σ^2. Alternatively, we can compute $E(\text{Var}(\bar{X}^*|X_1, \ldots, X_n))$ directly:

$$\text{Var}(\bar{X}^*|X_1, \ldots, X_n) = \frac{n}{n^2}\text{Var}(X_1^*|X_1, \ldots, X_n)$$

$$= \frac{1}{n}E((X_1^*)^2|X_1, \ldots, X_n) - \frac{1}{n}(E(X_1^*|X_1, \ldots, X_n))^2$$

$$= \frac{1}{n^2}\sum_{j=1}^{n}X_j^2 - \frac{1}{n}\bar{X}^2,$$

so

$$E(\text{Var}(\bar{X}^*|X_1, \ldots, X_n)) = \frac{1}{n^2}\sum_{j=1}^{n}E(X_j^2) - \frac{1}{n}E(\bar{X}^2)$$

$$= \frac{1}{n}E(X_1^2) - \frac{1}{n}E(\bar{X}^2)$$

$$= \frac{1}{n}(\text{Var}(X_1) + (EX_1)^2 - \text{Var}(\bar{X}) - (E(\bar{X}))^2)$$

$$= \frac{1}{n}(\sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2)$$

$$= \frac{(n-1)\sigma^2}{n^2},$$

which again yields

$$\text{Var}(\bar{X}^*) = \frac{\sigma^2}{n}\left(2 - \frac{1}{n}\right).$$

(d) Intuitively, we should have $\text{Var}(\bar{X}^*) > \text{Var}(\bar{X})$ for any $n > 1$, since $\text{Var}(\bar{X}^*)$ has contributions from *two* sources of variability: the variability that led to the initial sample, and the variability in choosing the bootstrap sample from the initial sample. Another way to see this is to note that even though X_j^* has the same distribution as X_j, the X_j's are uncorrelated but the X_j^*'s are positively correlated (since observing a large value for X_j^* means that some X_i is that large, and that X_i could get chosen multiple times).

38. An insurance company covers disasters in two neighboring regions, R_1 and R_2. Let I_1 and I_2 be the indicator r.v.s for whether R_1 and R_2 are hit by the insured disaster, respectively. The indicators I_1 and I_2 may be dependent. Let $p_j = E(I_j)$ for $j = 1, 2$, and $p_{12} = E(I_1 I_2)$.

The company reimburses a total cost of

$$C = I_1 \cdot T_1 + I_2 \cdot T_2$$

to these regions, where T_j has mean μ_j and variance σ_j^2. Assume that T_1 and T_2 are independent of each other and that (T_1, T_2) is independent of (I_1, I_2).

(a) Find $E(C)$.

(b) Find $\text{Var}(C)$.

Solution:

(a) We have $E(C) = E(I_1)E(T_1) + E(I_2)E(T_2) = p_1\mu_1 + p_2\mu_2$.

(b) By Eve's law,

$$\text{Var}(C) = E(\text{Var}(C|I_1, I_2)) + \text{Var}(E(C|I_1, I_2))$$

$$= E(I_1\sigma_1^2 + I_2\sigma_2^2) + \text{Var}(I_1\mu_1 + I_2\mu_2)$$

$$= p_1\sigma_1^2 + p_2\sigma_2^2 + p_1(1 - p_1)\mu_1^2 + p_2(1 - p_2)\mu_2^2 + 2\mu_1\mu_2(p_{12} - p_1 p_2).$$

39. Ⓢ A certain stock has low volatility on some days and high volatility on other days. Suppose that the probability of a low volatility day is p and of a high volatility day is $q = 1 - p$, and that on low volatility days the percent change in the stock price is $\mathcal{N}(0, \sigma_1^2)$, while on high volatility days the percent change is $\mathcal{N}(0, \sigma_2^2)$, with $\sigma_1 < \sigma_2$.

Let X be the percent change of the stock on a certain day. The distribution is said to be a *mixture* of two Normal distributions, and a convenient way to represent X is as $X = I_1 X_1 + I_2 X_2$ where I_1 is the indicator r.v. of having a low volatility day, $I_2 = 1 - I_1$, $X_j \sim \mathcal{N}(0, \sigma_j^2)$, and I_1, X_1, X_2 are independent.

(a) Find $\text{Var}(X)$ in two ways: using Eve's law, and by calculating $\text{Cov}(I_1 X_1 + I_2 X_2, I_1 X_1 + I_2 X_2)$ directly.

(b) Recall from Chapter 6 that the *kurtosis* of an r.v. Y with mean μ and standard deviation σ is defined by
$$\text{Kurt}(Y) = \frac{E(Y - \mu)^4}{\sigma^4} - 3.$$
Find the kurtosis of X (in terms of $p, q, \sigma_1^2, \sigma_2^2$, fully simplified). The result will show that even though the kurtosis of any Normal distribution is 0, the kurtosis of X is positive and in fact can be very large depending on the parameter values.

Solution:

(a) By Eve's Law,

$$\text{Var}(X) = E(\text{Var}(X|I_1)) + \text{Var}(E(X|I_1)) = E(I_1^2 \sigma_1^2 + (1 - I_1)^2 \sigma_2^2) + \text{Var}(0) = p\sigma_1^2 + (1-p)\sigma_2^2,$$

since $I_1^2 = I_1, I_2^2 = I_2$. For the covariance method, expand

$$\text{Var}(X) = \text{Cov}(I_1 X_1 + I_2 X_2, I_1 X_1 + I_2 X_2) = \text{Var}(I_1 X_1) + \text{Var}(I_2 X_2) + 2\text{Cov}(I_1 X_1, I_2 X_2).$$

Then $\text{Var}(I_1 X_1) = E(I_1^2 X_1^2) - (E(I_1 X_1))^2 = E(I_1)E(X_1^2) = p\text{Var}(X_1)$ since $E(I_1 X_1) = E(I_1)E(X_1) = 0$. Similarly, $\text{Var}(I_2 X_2) = (1 - p)\text{Var}(X_2)$. And

$$\text{Cov}(I_1 X_1, I_2 X_2) = E(I_1 I_2 X_1 X_2) - E(I_1 X_1)E(I_2 X_2) = 0,$$

since $I_1 I_2$ always equals 0. So again we have $\text{Var}(X) = p\sigma_1^2 + (1 - p)\sigma_2^2$.

(b) Note that $(I_1 X_1 + I_2 X_2)^4 = I_1 X_1^4 + I_2 X_2^4$ since the cross terms disappear (because $I_1 I_2$ is always 0) and any positive power of an indicator r.v. is that indicator r.v.! So

$$E(X^4) = E(I_1 X_1^4 + I_2 X_2^4) = 3p\sigma_1^4 + 3q\sigma_2^4.$$

Alternatively, we can use $E(X^4) = E(X^4|I_1 = 1)p + E(X^4|I_1 = 0)q$ to find $E(X^4)$. The mean of X is $E(I_1 X_1) + E(I_2 X_2) = 0$, so the kurtosis of X is

$$\text{Kurt}(X) = \frac{3p\sigma_1^4 + 3q\sigma_2^4}{(p\sigma_1^2 + q\sigma_2^2)^2} - 3.$$

This becomes 0 if $\sigma_1 = \sigma_2$, since then we have a Normal distribution rather than a mixture of two different Normal distributions. For $\sigma_1 < \sigma_2$, the kurtosis is positive since

$$p\sigma_1^4 + q\sigma_2^4 > (p\sigma_1^2 + q\sigma_2^2)^2,$$

as can be seen by interpreting this as saying $E(Y^2) > (EY)^2$, where Y is σ_1^2 with probability p and σ_2^2 with probability q.

40. Let X_1, X_2, and Y be random variables, such that Y has finite variance. Let

$$A = E(Y|X_1) \text{ and } B = E(Y|X_1, X_2).$$

Show that

$$\text{Var}(A) \leq \text{Var}(B).$$

Also, check that this make sense in the extreme cases where Y is independent of X_1 and where $Y = h(X_2)$ for some function h.

Hint: Use Eve's law on B.

Solution: By Eve's law,

$$\text{Var}(B) = E(\text{Var}(B|X_1)) + \text{Var}(E(B|X_1)) \geq \text{Var}(E(B|X_1)).$$

By Adam's law,

$$E(B|X_1) = E(E(Y|X_1, X_2)|X_1) = E(Y|X_1) = A.$$

Thus,

$$\text{Var}(B) \geq \text{Var}(A).$$

In the extreme case where Y is independent of X_1, we have $A = E(Y)$, and then $\text{Var}(A) = 0$ and the inequality reduces to the true statement $\text{Var}(B) \geq 0$. In the extreme case where $Y = h(X_2)$ for some function h, we have $B = Y$, and then again it makes sense that $\text{Var}(B) \geq \text{Var}(A)$ since, by Eve's law,

$$\text{Var}(B) = \text{Var}(Y) \geq \text{Var}(E(Y|X_1)) = \text{Var}(A).$$

41. Show that for any r.v.s X and Y,

$$E(Y|E(Y|X)) = E(Y|X).$$

This has a nice intuitive interpretation if we think of $E(Y|X)$ as the prediction we would make for Y based on X: given the prediction we would use for predicting Y from X, we no longer need to know X to predict Y—we can just use the prediction we have! For example, letting $E(Y|X) = g(X)$, if we observe $g(X) = 7$, then we may or may not know what X is (since g may not be one-to-one). But even without knowing X, we know that the prediction for Y based on X is 7.

Hint: Use Adam's law with extra conditioning.

Solution: Let $g(X) = E(Y|X)$. By Adam's law with extra conditioning,

$$E(Y|g(X)) = E(E(Y|g(X), X)|g(X)) = E(E(Y|X)|g(X)) = E(g(X)|g(X)) = g(X),$$

which is what we wanted to show.

42. A researcher wishes to know whether a new treatment for the disease conditionitis is more effective than the standard treatment. It is unfortunately not feasible to do a randomized experiment, but the researcher does have the medical records of patients who received the new treatment and those who received the standard treatment. She is worried, though, that doctors tend to give the new treatment to younger, healthier patients. If this is the case, then naively comparing the outcomes of patients in the two groups would be like comparing apples and oranges.

Suppose each patient has background variables $\mathbf{X}$, which might be age, height and weight, and measurements relating to previous health status. Let Z be the indicator of receiving the new treatment. The researcher fears that Z is dependent on $\mathbf{X}$, i.e., that the distribution of $\mathbf{X}$ given $Z = 1$ is different from the distribution of $\mathbf{X}$ given $Z = 0$.

In order to compare apples to apples, the researcher wants to match every patient who received the new treatment to a patient with similar background variables who received the standard treatment. But $\mathbf{X}$ could be a high-dimensional random vector, which often makes it very difficult to find a match with a similar value of $\mathbf{X}$.

The *propensity score* reduces the possibly high-dimensional vector of background variables down to a single number (then it is much easier to match someone to a person

with a similar propensity score than to match someone to a person with a similar value of $\mathbf{X}$). The propensity score of a person with background characteristics $\mathbf{X}$ is defined as

$$S = E(Z|\mathbf{X}).$$

By the fundamental bridge, a person's propensity score is their probability of receiving the treatment, given their background characteristics. Show that conditional on S, the treatment indicator Z is independent of the background variables $\mathbf{X}$.

Hint: It helps to first solve the previous problem. Then show that $P(Z = 1|S, \mathbf{X}) = P(Z = 1|S)$. By the fundamental bridge, this is equivalent to showing $E(Z|S, \mathbf{X}) = E(Z|S)$.

Solution: Since S is a function of $\mathbf{X}$,

$$E(Z|S, \mathbf{X}) = E(Z|\mathbf{X}).$$

But as shown in the solution to the previous problem,

$$E(Z|\mathbf{X}) = E(Z|S).$$

So by the fundamental bridge,

$$P(Z = 1|S, \mathbf{X}) = E(Z|S, \mathbf{X}) = E(Z|S) = P(Z = 1|S).$$

Of course, this also implies that

$$P(Z = 0|S, \mathbf{X}) = P(Z = 0|S).$$

Thus, $\mathbf{X}$ and Z are conditionally independent given S.

43. This exercise develops a useful identity for covariance, similar in spirit to Adam's law for expectation and Eve's law for variance. First define *conditional covariance* in a manner analogous to how we defined conditional variance:

$$\text{Cov}(X, Y|Z) = E\big((X - E(X|Z))(Y - E(Y|Z))|Z\big).$$

(a) Show that

$$\text{Cov}(X, Y|Z) = E(XY|Z) - E(X|Z)E(Y|Z).$$

This should be true since it is the conditional version of the fact that

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$$

and conditional probabilities *are* probabilities, but for this problem you should prove it directly using properties of expectation and conditional expectation.

(b) *ECCE*, or the *law of total covariance*, says that

$$\text{Cov}(X, Y) = E(\text{Cov}(X, Y|Z)) + \text{Cov}(E(X|Z), E(Y|Z)).$$

That is, the covariance of X and Y is the expected value of their conditional covariance plus the covariance of their conditional expectations, where all these conditional quantities are conditional on Z. Prove this identity.

Hint: We can assume without loss of generality that $E(X) = E(Y) = 0$, since adding a constant to an r.v. has no effect on its covariance with any r.v. Then expand out the covariances on the right-hand side of the identity and apply Adam's law.

Solution:

(a) By linearity and taking out what's known, using the fact that $E(X|Z)$ and $E(Y|Z)$ are functions of Z, we have

$$
\begin{aligned}
\text{Cov}(X, Y|Z) &= E\big((X - E(X|Z))(Y - E(Y|Z))|Z\big) \\
&= E(XY|Z) - E(XE(Y|Z)|Z) - E(YE(X|Z)|Z) + E(E(X|Z)E(Y|Z)|Z) \\
&= E(XY|Z) - E(Y|Z)E(X|Z) - E(X|Z)E(Y|Z) + E(X|Z)E(Y|Z) \\
&= E(XY|Z) - E(X|Z)E(Y|Z).
\end{aligned}
$$

(b) As discussed in the hint, we can take $E(X) = E(Y) = 0$. Then the right-hand side of ECCE is

$$
\begin{aligned}
E(\text{Cov}(X, Y|Z)) + \text{Cov}(E(X|Z), E(Y|Z)) &= E(E(XY|Z) - E(X|Z)E(Y|Z)) + E(E(X|Z)E(Y|Z)) \\
&= E(XY) - E(E(X|Z)E(Y|Z)) + E(E(X|Z)E(Y|Z)) \\
&= E(XY) \\
&= \text{Cov}(X, Y),
\end{aligned}
$$

as desired.

Mixed practice

44. A group of n friends often go out for dinner together. At their dinners, they play "credit card roulette" to decide who pays the bill. This means that at each dinner, one person is chosen uniformly at random to pay the entire bill (independently of what happens at the other dinners).

(a) Find the probability that in k dinners, no one will have to pay the bill more than once (do not simplify for the case $k \leq n$, but do simplify fully for the case $k > n$).

(b) Find the expected number of dinners it takes in order for everyone to have paid at least once (you can leave your answer as a finite sum of simple-looking terms).

(c) Alice and Bob are two of the friends. Find the covariance between how many times Alice pays and how many times Bob pays in k dinners (simplify fully).

Solution:

(a) For $k > n$, the probability is 0 since then there are more dinners than diners. For $k \leq n$, by the naive definition or as in the birthday problem, the probability is

$$
\frac{n(n-1)\cdots(n-k+1)}{n^k}.
$$

(b) This is isomorphic to the coupon collector problem. The number of dinners is the sum of waiting times N_j, with $N_1 = 1$ and $N_j \sim \text{FS}(\frac{n-j+1}{n})$ for $2 \leq j \leq n$. So the expected value is

$$
1 + \frac{n}{n-1} + \frac{n}{n-2} + \cdots + \frac{n}{2} + n = n \sum_{j=1}^{n} \frac{1}{j}.
$$

(c) The counts of how many times each person pays form a $\text{Mult}_n\left(k, (\frac{1}{n}, \ldots, \frac{1}{n})\right)$ random vector, so the covariance is $-k/n^2$.

45. As in the previous problem, a group of n friends play "credit card roulette" at their dinners. In this problem, let the number of dinners be a $\text{Pois}(\lambda)$ r.v.

(a) Alice is one of the friends. Find the correlation between how many dinners Alice pays for and how many free dinners Alice gets (simplify fully).

(b) The costs of the dinners are i.i.d. $\text{Gamma}(a, b)$ r.v.s, independent of the number of dinners. Find the mean and variance of the total cost (simplify fully).

Solution:

(a) By the chicken-egg story, they are independent. So the correlation is 0.

(b) Let $X = X_1 + \cdots + X_N$ be the total cost, where X_j is the cost of the jth dinner and N is the number of dinners. Conditioning on N via Adam's law and Eve's law,

$$E(X) = E(E(X|N)) = E(Na/b) = \lambda a/b,$$

$$\text{Var}(X) = E(\text{Var}(X|N)) + \text{Var}(E(X|N)) = E(Na/b^2) + \text{Var}(Na/b) = \frac{\lambda a}{b^2} + \frac{\lambda a^2}{b^2} = \frac{\lambda a(1+a)}{b^2}.$$

46. Joe will read $N \sim \text{Pois}(\lambda)$ books next year. Each book has a $\text{Pois}(\mu)$ number of pages, with book lengths independent of each other and independent of N.

    (a) Find the expected number of book pages that Joe will read next year.

    (b) Find the variance of the number of book pages Joe will read next year.

    (c) For each of the N books, Joe likes it with probability p and dislikes it with probability $1-p$, independently. Find the conditional distribution of how many of the N books Joe likes, given that he dislikes exactly d of the books.

    Solution:

    (a) Let $X = X_1 + \cdots + X_N$ be the number of book pages Joe will read next year, with X_j the number of pages in the jth book. By Adam's law,

    $$E(X) = E(E(X|N)) = E(N\mu) = \lambda\mu.$$

    (b) By Eve's law,

    $$\text{Var}(X) = E(\text{Var}(X|N)) + \text{Var}(E(X|N)) = E(N\text{Var}(X_1)) + \text{Var}(N\mu) = \lambda\mu + \lambda\mu^2.$$

    (c) By the chicken-egg story, how many of the N books Joe likes is a $\text{Pois}(\lambda p)$ r.v. that is *independent* of how many he dislikes. So the conditional distribution is also $\text{Pois}(\lambda p)$.

47. Buses arrive at a certain bus stop according to a Poisson process of rate λ. Each bus has n seats and, at the instant when it arrives at the stop, has a $\text{Bin}(n, p)$ number of passengers. Assume that the numbers of passengers on different buses are independent of each other, and independent of the arrival times of the buses.

    Let N_t be the number of buses that arrive in the time interval $[0, t]$, and X_t be the total number of passengers on the buses that arrive in the time interval $[0, t]$.

    (a) Find the mean and variance of N_t.

    (b) Find the mean and variance of X_t.

    (c) A bus is *full* if it has exactly n passengers when it arrives at the stop. Find the probability that exactly $a + b$ buses arrive in $[0, t]$, of which a are full and b are not full.

    Solution:

    (a) By definition of Poisson process, $N_t \sim \text{Pois}(\lambda t)$, so the mean and variance are λt.

    (b) By Adam's law,

    $$E(X_t) = E(E(X_t|N_t)) = E(npN_t) = np\lambda t.$$

    By Eve's law,

    $$\text{Var}(X_t) = E(\text{Var}(X_t|N_t)) + \text{Var}(E(X_t|N_t)) = E(np(1-p)N_t) + \text{Var}(npN_t) = np(1-p)\lambda t + n^2p^2\lambda t.$$

(c) Let A_t and B_t be the number of full and not-full buses that arrive in $[0, t]$, respectively. By the chicken-egg story, $A_t \sim \text{Pois}(p^n \lambda t)$, $B_t \sim \text{Pois}((1 - p^n)\lambda t)$, and A_t and B_t are independent. Thus, for any nonnegative integers a and b,

$$P(N_t = a + b, A_t = a, B_t = b) = P(A_t = a, B_t = b)$$

$$= \frac{e^{-p^n \lambda t}(p^n \lambda t)^a}{a!} \cdot \frac{e^{-(1-p^n)\lambda t}((1 - p^n)\lambda t)^b}{b!}.$$

Alternatively, we can reason directly that

$$P(N_t = a + b, A_t = a, B_t = b) = P(N_t = a + b)P(A_t = a | N_t = a + b)P(B_t = b | N_t = a + b, A_t = a)$$

$$= \frac{e^{-\lambda t}(\lambda t)^{a+b}}{(a + b)!} \cdot \binom{a + b}{a} p^{na}(1 - p^n)^b \cdot 1$$

$$= \frac{e^{-\lambda t}(\lambda t)^{a+b}p^{na}(1 - p^n)^b}{a!b!}.$$

48. Paul and n other runners compete in a marathon. Their times are independent continuous r.v.s with CDF F.

(a) For $j = 1, 2, \ldots, n$, let A_j be the event that anonymous runner j completes the race faster than Paul. Explain whether the events A_j are independent, and whether they are conditionally independent given Paul's time to finish the race.

(b) For the rest of this problem, let N be the number of runners who finish faster than Paul. Find $E(N)$. (Your answer should depend only on n, since Paul's time is an r.v.)

(c) Find the conditional distribution of N, given that Paul's time to finish the marathon is t.

(d) Find $\text{Var}(N)$. (Your answer should depend only on n, since Paul's time is an r.v.)

Hint: Let T be Paul's time, and use Eve's law to condition on T. Alternatively, use indicator r.v.s.

Solution:

(a) The A_j are not independent. If runners 1 through $n - 1$ all complete the race faster than Paul, then this is evidence that Paul was slow, which in turn increases the chance that runner n ran faster than Paul. But the A_j *are* conditionally independent given Paul's time, since the runners' times are independent.

(b) Write $N = I_1 + \cdots + I_n$, where I_j is the indicator of runner j being faster than Paul in the race. By symmetry, $E(I_j) = 1/2$. So by linearity, $E(N) = n/2$.

(c) Given that Paul's time is t, each of the other runners has a better time than Paul with probability $F(t)$, independently. So the conditional distribution is $\text{Bin}(n, F(t))$.

(d) Let T be Paul's time. By Eve's law,

$$\begin{aligned}
\text{Var}(N) &= E(\text{Var}(N|T)) + \text{Var}(E(N|T)) \\
&= E(nF(T)(1 - F(T))) + \text{Var}(nF(T)) \\
&= \frac{n}{2} - \frac{n}{3} + \frac{n^2}{12} \\
&= \frac{n}{6} + \frac{n^2}{12},
\end{aligned}$$

since $F(T) \sim \text{Unif}(0,1)$ by universality of the Uniform.

Alternatively, write

$$N = I_1 + \cdots + I_n,$$

as in (b). Then $E(I_1) = 1/2$ (by symmetry), $\text{Var}(I_1) = 1/4$ (since $I_1 \sim \text{Bern}(1/2)$), and

$$\text{Cov}(I_1, I_2) = E(I_1 I_2) - E(I_1)E(I_2) = 1/3 - 1/4 = 1/12$$

(by symmetry). Thus,

$$\begin{aligned}
\text{Var}(N) &= \text{Var}(I_1) + \cdots + \text{Var}(I_n) + 2\sum_{i<j} \text{Cov}(I_i, I_j) \\
&= n\text{Var}(I_1) + n(n-1)\text{Cov}(I_1, I_2) \\
&= \frac{n}{4} + \frac{n(n-1)}{12} \\
&= \frac{n}{6} + \frac{n^2}{12},
\end{aligned}$$

which agrees with the result of the previous method.

49. Emails arrive in an inbox according to a Poisson process of rate λ emails per hour.

(a) Find the name and parameters of the conditional distribution of the number of emails that arrive in the first 2 hours of an 8-hour time period, given that exactly n emails arrive in that time period.

(b) Each email is legitimate with probability p and spam with probability $q = 1 - p$, independently. Find the name and parameters of the conditional distribution of the number of legitimate emails that arrive in an 8-hour time period, given that exactly s spams arrived in that time period.

(c) Reading an email takes a random amount of time, with mean μ hours and standard deviation σ hours. These reading times are i.i.d. and independent of the email arrival process. Find the (unconditional) mean and variance of the total time it takes to read all the emails that arrive in an 8-hour time period.

Solution:

(a) Let X be the number of emails that arrive in the first 2 hours and Y be the number that arrive in the next 6 hours. Then

$$P(X = k | X + Y = n) = \frac{P(X = k, Y = n - k)}{P(X + Y = n)}$$

$$= \frac{e^{-2\lambda}(2\lambda)^k}{k!} \cdot \frac{e^{-6\lambda}(6\lambda)^{n-k}}{(n-k)!} \cdot \frac{n!}{e^{-8\lambda}(8\lambda)^n}$$

$$= \binom{n}{k} \left(\frac{2}{8}\right)^k \left(\frac{6}{8}\right)^{n-k},$$

so $X | X + Y = n \sim \text{Bin}(n, 1/4)$.

(b) By the chicken–egg story, the number of legitimate emails in this time period is $\text{Pois}(8\lambda p)$, independent of the number of spams in the time period. So the conditional distribution is also $\text{Pois}(8\lambda p)$.

(c) Let T be the total time and T_j be the time it takes to read the jth email. Let $N \sim \text{Pois}(8\lambda)$ be the number of emails. By Adam's law and Eve's law,

$$E(T) = E(E(T|N)) = E(N\mu) = 8\lambda\mu,$$

$$\text{Var}(T) = E(\text{Var}(T|N)) + \text{Var}(E(T|N)) = E(N\sigma^2) + \text{Var}(N\mu) = 8\lambda(\sigma^2 + \mu^2).$$

50. An actuary wishes to estimate various quantities related to the number of insurance claims and the dollar amounts of those claims for someone named Fred. Suppose that Fred will make N claims next year, where $N|\lambda \sim \text{Pois}(\lambda)$. But λ is unknown, so the actuary, taking a Bayesian approach, gives λ a prior distribution based on past experience. Specifically, the prior is $\lambda \sim \text{Expo}(1)$. The dollar amount of a claim is Log-Normal with parameters μ and σ^2 (here μ and σ^2 are the mean and variance of the underlying Normal), with μ and σ^2 known. The dollar amounts of the claims are i.i.d. and independent of N.

(a) Find $E(N)$ and $\text{Var}(N)$ using properties of conditional expectation (your answers should not depend on λ, since λ is unknown and being treated as an r.v.!).

(b) Find the mean and variance of the total dollar amount of all the claims.

(c) Find the distribution of N. If it is a named distribution we have studied, give its name and parameters.

(d) Find the posterior distribution of λ, given that it is observed that Fred makes $N = n$ claims next year. If it is a named distribution we have studied, give its name and parameters.

Solution:

(a) Using Adam's law and Eve's law, we have

$$E(N) = E(E(N|\lambda)) = E(\lambda) = 1,$$

$$\text{Var}(N) = E(\text{Var}(N|\lambda)) + \text{Var}(E(N|\lambda)) = E(\lambda) + \text{Var}(\lambda) = 1 + 1 = 2.$$

(b) Let $X = X_1 + \cdots + X_N$, where $X_1, \ldots, X_N$ are the individual claim amounts. Let

$$m = E(X_1) = e^{\mu + \frac{1}{2}\sigma^2}.$$

By Adam's law and Eve's law,

$$E(X) = E(E(X|N)) = E(N)E(X_1) = E(X_1) = m,$$

$$\text{Var}(X) = E(\text{Var}(X|N)) + \text{Var}(E(X|N)) = E(N)\text{Var}(X_1) + m^2\text{Var}(N) = m^2(e^{\sigma^2} + 1).$$

(c) By Story 8.4.5 (Gamma-Poisson conjugacy), $N \sim \text{Geom}(1/2)$.

(d) By Story 8.4.5 (Gamma-Poisson conjugacy), $\lambda|N = n \sim \text{Gamma}(1 + n, 2)$.

51. Ⓢ Let X_1, X_2, X_3 be independent with $X_i \sim \text{Expo}(\lambda_i)$ (so with possibly different rates). Recall from Chapter 7 that

$$P(X_1 < X_2) = \frac{\lambda_1}{\lambda_1 + \lambda_2}.$$

(a) Find $E(X_1 + X_2 + X_3|X_1 > 1, X_2 > 2, X_3 > 3)$ in terms of $\lambda_1, \lambda_2, \lambda_3$.

(b) Find $P(X_1 = \min(X_1, X_2, X_3))$, the probability that the first of the three Exponentials is the smallest.

Hint: Restate this in terms of X_1 and $\min(X_2, X_3)$.

(c) For the case $\lambda_1 = \lambda_2 = \lambda_3 = 1$, find the PDF of $\max(X_1, X_2, X_3)$. Is this one of the important distributions we have studied?

Solution:

(a) By linearity, independence, and the memoryless property, we get

$$E(X_1|X_1 > 1) + E(X_2|X_2 > 2) + E(X_3|X_3 > 3) = \lambda_1^{-1} + \lambda_2^{-1} + \lambda_3^{-1} + 6.$$

(b) The desired probability is $P(X_1 \leq \min(X_2, X_3))$. Noting that $\min(X_2, X_3) \sim$ Expo$(\lambda_2 + \lambda_3)$ is independent of X_1, we have

$$P(X_1 \leq \min(X_2, X_3)) = \frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3}.$$

(c) Let $M = \max(X_1, X_2, X_3)$. Using the order statistics results from Chapter 9 or by directly computing the CDF and taking the derivative, for $x > 0$ we have

$$f_M(x) = 3(1 - e^{-x})^2 e^{-x}.$$

This is not one of the important distributions we have studied. The form is reminiscent of a Beta, but a Beta takes values between 0 and 1, while M can take any positive real value (in fact, $B \sim$ Beta$(1, 3)$ if we make the transformation $B = e^{-M}$).

52. ⓢ A task is randomly assigned to one of two people (with probability $1/2$ for each person). If assigned to the first person, the task takes an Expo(λ_1) length of time to complete (measured in hours), while if assigned to the second person it takes an Expo(λ_2) length of time to complete (independent of how long the first person would have taken). Let T be the time taken to complete the task.

(a) Find the mean and variance of T.

(b) Suppose instead that the task is assigned to *both* people, and let X be the time taken to complete it (by whoever completes it first, with the two people working independently). It is observed that after 24 hours, the task has not yet been completed. Conditional on this information, what is the expected value of X?

Solution: Write $T = IX_1 + (1 - I)X_2$, with $I \sim$ Bern$(1/2), X_1 \sim$ Expo$(\lambda_1), X_2 \sim$ Expo(λ_2) independent. Then

$$ET = \frac{1}{2}(\lambda_1^{-1} + \lambda_2^{-1}),$$

$$\begin{aligned}
\text{Var}(T) &= E(\text{Var}(T|I)) + \text{Var}(E(T|I)) \\
&= E(I^2 \frac{1}{\lambda_1^2} + (1-I)^2 \frac{1}{\lambda_2^2}) + \text{Var}\left(\frac{I}{\lambda_1} + \frac{1-I}{\lambda_2}\right) \\
&= E(I \frac{1}{\lambda_1^2} + (1-I) \frac{1}{\lambda_2^2}) + \text{Var}\left(I(\frac{1}{\lambda_1} - \frac{1}{\lambda_2})\right) \\
&= \frac{1}{2}(\frac{1}{\lambda_1^2} + \frac{1}{\lambda_2^2}) + \frac{1}{4}(\frac{1}{\lambda_1} - \frac{1}{\lambda_2})^2.
\end{aligned}$$

Sanity check: For $\lambda_1 = \lambda_2$, the two people have the same distribution so randomly assigning the task to one of the two should be equivalent to just assigning it to the first person (so the mean and variance should agree with those of an Expo(λ_1) r.v.). It makes sense that the mean is the average of the two means, as we can condition on whether $I = 1$ (though the variance is *greater* than the average of the two variances, by Eve's Law). Also, the results should be (and are) the same if we swap λ_1 and λ_2.

(b) Here $X = \min(X_1, X_2)$ with $X_1 \sim$ Expo$(\lambda_1), X_2 \sim$ Expo(λ_2) independent. Then $X \sim$ Expo$(\lambda_1 + \lambda_2)$ (since $P(X > x) = P(X_1 > x)P(X_2 > x) = e^{-(\lambda_1 + \lambda_2)x}$, or by results on order statistics). By the memoryless property,

$$E(X|X > 24) = 24 + \frac{1}{\lambda_1 + \lambda_2}.$$

308

Sanity check: The answer should be greater than 24 and should be very close to 24 if λ_1 or λ_2 is very large. Considering a Poisson process also helps make this intuitive.

53. Suppose for this problem that "true IQ" is a meaningful concept rather than a reified social construct. Suppose that in the U.S. population, the distribution of true IQs is Normal with mean 100 and SD 15. A person is chosen at random from this population to take an IQ test. The test is a noisy measure of true ability: it's correct on average but has a Normal measurement error with SD 5.

    Let μ be the person's true IQ, viewed as a random variable, and let Y be her score on the IQ test. Then we have

    $$Y|\mu \sim \mathcal{N}(\mu, 5^2)$$
    $$\mu \sim \mathcal{N}(100, 15^2).$$

    (a) Find the unconditional mean and variance of Y.

    (b) Find the marginal distribution of Y. One way is via the MGF.

    (c) Find $\text{Cov}(\mu, Y)$.
    Solution:

    (a) By Adam's law and Eve's law,

    $$E(Y) = E(E(Y|\mu)) = E(\mu) = 100,$$

    $$\text{Var}(Y) = E(\text{Var}(Y|\mu)) + \text{Var}(E(Y|\mu)) = E(5^2) + 15^2 = 250.$$

    (b) Using Adam's law and the Normal MGF, the MGF of Y is

    $$M(t) = E(e^{tY}) = E(E(e^{tY}|\mu)) = E(e^{t\mu + 25t^2/2}) = e^{25t^2/2}e^{100t + 225t^2/2} = e^{100t + 250t^2/2}.$$

    Therefore, $Y \sim \mathcal{N}(100, 250)$. Note that the mean and variance found in this part agree with those found in (a).

    (c) By Adam's law,

    $$E(\mu Y) = E(E(\mu Y|\mu)) = E(\mu E(Y|\mu)) = E(\mu^2) = \text{Var}(\mu) + (E\mu)^2 = 15^2 + 100^2.$$

    So
    $$\text{Cov}(\mu, Y) = E(\mu Y) - E(\mu)E(Y) = 15^2 + 100^2 - 100^2 = 15^2 = 225.$$

54. ⓢ A certain genetic characteristic is of interest. It can be measured numerically. Let X_1 and X_2 be the values of the genetic characteristic for two twin boys. If they are identical twins, then $X_1 = X_2$ and X_1 has mean 0 and variance σ^2; if they are fraternal twins, then X_1 and X_2 have mean 0, variance σ^2, and correlation ρ. The probability that the twins are identical is 1/2. Find $\text{Cov}(X_1, X_2)$ in terms of ρ, σ^2.

    Solution: Since the means are 0, $\text{Cov}(X_1, X_2) = E(X_1 X_2) - (EX_1)(EX_2) = E(X_1 X_2)$. Now condition on whether the twins are identical or fraternal:

    $$E(X_1 X_2) = E(X_1 X_2|\text{identical})\frac{1}{2} + E(X_1 X_2|\text{fraternal})\frac{1}{2} = E(X_1^2)\frac{1}{2} + \rho\sigma^2\frac{1}{2} = \frac{\sigma^2}{2}(1+\rho).$$

55. ⓢ The Mass Cash lottery randomly chooses 5 of the numbers from $1, 2, \ldots, 35$ each day (without repetitions within the choice of 5 numbers). Suppose that we want to know how long it will take until all numbers have been chosen. Let a_j be the average number of additional days needed if we are missing j numbers (so $a_0 = 0$ and a_{35} is the average number of days needed to collect all 35 numbers). Find a recursive formula for the a_j.

Solution: Suppose we are missing j numbers (with $0 \leq j \leq 35$), and let T_j be the additional number of days needed to complete the collection. Condition on how many "new" numbers appear the next day; call this N. This gives

$$E(T_j) = \sum_{n=0}^{5} E(T_j | N = n) P(N = n).$$

Note that N is Hypergeometric (imagine tagging the numbers that we don't already have in our collection)! Letting $a_k = 0$ for $k < 0$, we have

$$a_j = 1 + \sum_{n=0}^{5} \frac{a_{j-n} \binom{j}{n} \binom{35-j}{5-n}}{\binom{35}{5}}.$$

56. Two chess players, Vishy and Magnus, play a series of games. Given p, the game results are i.i.d. with probability p of Vishy winning, and probability $q = 1 - p$ of Magnus winning (assume that each game ends in a win for one of the two players). But p is unknown, so we will treat it as an r.v. To reflect our uncertainty about p, we use the prior $p \sim \text{Beta}(a, b)$, where a and b are known positive *integers* and $a \geq 2$.

(a) Find the expected number of games needed in order for Vishy to win a game (including the win). Simplify fully; your final answer should not use factorials or Γ.

(b) Explain in terms of independence vs. conditional independence the direction of the inequality between the answer to (a) and $1 + E(G)$ for $G \sim \text{Geom}(\frac{a}{a+b})$.

(c) Find the conditional distribution of p given that Vishy wins exactly 7 out of the first 10 games.

Solution:

(a) Let N be the number of games needed. Then since $N|p \sim \text{FS}(p)$,

$$E(N) = E(E(N|p)) = E(1/p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 p^{a-1-1}(1-p)^{b-1} dp = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a-1)\Gamma(b)}{\Gamma(a+b-1)},$$

because the $\text{Beta}(a-1, b)$ PDF integrates to 1 (or by Bayes' billiards). So

$$E(N) = \frac{(a+b-1)!(a-2)!}{(a-1)!(a+b-2)!} = \frac{a+b-1}{a-1}.$$

(b) The games are conditionally independent given p, but not independent. If p were known to be equal to its prior mean $E(p) = a/(a+b)$, the trials would be independent and then the expected time of first success would be $(a+b)/a$. But with p unknown, each time Vishy loses, the probability of him winning the next game goes down. So the expected number of games is larger than $(a+b)/a$, which agrees with (a). This inequality can also be written as $E(1/p) > 1/E(p)$, which is a special case of *Jensen's inequality*, an important inequality from Chapter 10).

(c) Since Beta is the conjugate prior for the Binomial, the posterior distribution of p given the data about the first 10 games is $\text{Beta}(a+7, b+3)$.

57. *Laplace's law of succession* says that if $X_1, X_2, \ldots, X_{n+1}$ are conditionally independent Bern(p) r.v.s given p, but p is given a Unif$(0, 1)$ prior to reflect ignorance about its value, then
$$P(X_{n+1} = 1 | X_1 + \cdots + X_n = k) = \frac{k+1}{n+2}.$$

As an example, Laplace discussed the problem of predicting whether the sun will rise tomorrow, given that the sun did rise every time for all n days of recorded history; the above formula then gives $(n+1)/(n+2)$ as the probability of the sun rising tomorrow

(of course, assuming independent trials with p unchanging over time may be a very unreasonable model for the sunrise problem).

(a) Find the posterior distribution of p given $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, and show that it only depends on the sum of the x_j (so we only need the one-dimensional quantity $x_1 + x_2 + \cdots + x_n$ to obtain the posterior distribution, rather than needing all n data points).

(b) Prove Laplace's law of succession, using a form of LOTP to find $P(X_{n+1} = 1 | X_1 + \cdots + X_n = k)$ by conditioning on p. (The next exercise, which is closely related, involves an equivalent Adam's law proof.)

Solution:

(a) By the coherency of Bayes' rule (see Section 2.6), we can update sequentially: updated based on $X_1 = x_1$, then based on $X_2 = x_2$, etc. The prior is Beta$(1, 1)$. By conjugacy of the Beta and Binomial, the posterior after observing $X_1 = x_1$ is Beta$(1 + x_1, 1 + (1 - x_1))$. This becomes the new prior, and then the new posterior after observing $X_2 = x_2$ is Beta$(1 + x_1 + x_2, 1 + (1 - x_1) + (1 - x_2))$. Continuing in this way, the posterior after all of $X_1, \ldots, X_n$ have been observed is

$$p | X_1 = x_1, \ldots, X_n = x_n \sim \text{Beta} \left(1 + \sum_{j=1}^{n} x_j, 1 + n - \sum_{j=1}^{n} x_j \right).$$

This distribution depends only on $\sum_{j=1}^{n} x_j$, and is the same posterior distribution we would have obtained if we had observed only the r.v. $X_1 + \cdots + X_n$.

(b) Let $S_n = X_1 + \cdots + X_n$. By LOTP,

$$P(X_{n+1} = 1 | S_n = k) = \int_0^1 P(X_{n+1} = 1 | p, S_n = k) f(p | S_n = k) dp,$$

where $f(p | S_n = k)$ is the posterior PDF of p given $S_n = k$. By (a) and the fact that the X_j are conditionally independent given p, we then have

$$\begin{aligned}
P(X_{n+1} = 1 | S_n = k) &= \frac{\Gamma(n+2)}{\Gamma(k+1)\Gamma(n-k+1)} \int_0^1 p p^k (1-p)^{n-k} dp \\
&= \frac{\Gamma(n+2)}{\Gamma(k+1)\Gamma(n-k+1)} \frac{\Gamma(k+2)\Gamma(n-k+1)}{\Gamma(n+3)} \\
&= \frac{k+1}{n+2},
\end{aligned}$$

where to evaluate the integral we pattern-matched to the Beta$(k + 2, n - k + 1)$ PDF.

58. Two basketball teams, A and B, play an n game match. Let X_j be the indicator of team A winning the jth game. Given p, the r.v.s $X_1, \ldots, X_n$ are i.i.d. with $X_j | p \sim \text{Bern}(p)$. But p is unknown, so we will treat it as an r.v. Let the prior distribution be $p \sim \text{Unif}(0, 1)$, and let X be the number of wins for team A.

(a) Find $E(X)$ and $\text{Var}(X)$.

(b) Use Adam's law to find the probability that team A will win game $j + 1$, given that they win exactly a of the first j games. (The previous exercise, which is closely related, involves an equivalent LOTP proof.)

Hint: Letting C be the event that team A wins exactly a of the first j games,

$$P(X_{j+1} = 1 | C) = E(X_{j+1} | C) = E(E(X_{j+1} | C, p) | C) = E(p | C).$$

(c) Find the PMF of X. (There are various ways to do this, including a very fast way to see it based on results from earlier chapters.)

(d) The Putnam exam from 2002 posed the following problem:

Shanille O'Keal shoots free throws on a basketball court. She hits the first and misses the second, and thereafter the [conditional] probability that she hits the next shot is equal to the proportion of shots she has hit so far. What is the probability she hits exactly 50 of her first 100 shots?

Solve this Putnam problem by applying the result of Part (c). Be sure to explain why it is valid to apply that result, despite the fact that the Putnam problem does not seem to be using the same model, e.g., it does not mention a prior distribution, let alone mention a Unif$(0, 1)$ prior.

Solution:

(a) By Adam's law and Eve's law,

$$E(X) = E(E(X|p)) = E(np) = n/2,$$

$$\text{Var}(X) = E(\text{Var}(X|p)) + \text{Var}(E(X|p)) = E(np(1-p)) + \text{Var}(np) = \frac{n}{6} + \frac{n^2}{12},$$

since $Ep = \frac{1}{2}, \text{Var}(p) = \frac{1}{12}$, and $E(p(1-p)) = Ep - (\text{Var}(p) + (Ep)^2) = \frac{1}{6}$.

(b) Let S_j be the number of wins for A in the first j games. Then $S_j|p \sim \text{Bin}(j, p)$ and the prior on p is Beta$(1, 1)$, so the posterior distribution of p given $S_j = a$ is Beta$(a + 1, j - a + 1)$. Thus,

$$P(X_{j+1} = 1|S_j = a) = E(p|S_j = a) = \frac{a+1}{a+1+j-a+1} = \frac{a+1}{j+2}.$$

(c) By Bayes' billiards (or pattern-matching to a Beta PDF), the PMF of X is

$$P(X = k) = \int_0^1 P(X = k|p)f(p)dp = \int_0^1 \binom{n}{k} p^k(1-p)^{n-k} dp = \frac{1}{n+1},$$

for $k = 0, 1, \ldots, n$ (and 0 otherwise), where f is the Unif$(0, 1)$ PDF.

(d) Let Y be the number of shots that Shanille makes, among her 3rd through 100th shots. Since it is given that she made exactly 1 of her first 2 shots, the question is asking for $P(Y = 49)$. Let $n = 98$. The Putnam problem is not formulated in terms of a Bayesian approach, but the probabilistic structure in (b) is the same as the probabilistic structure here: if Shanille has made exactly a out of j shots, excluding the first two shots, then her conditional probability of making her next shot is $(a+1)/(j+2)$, which exactly corresponds to the result in (b). Therefore, Y has the same distribution as X. So

$$P(Y = k) = \frac{1}{99},$$

for $k = 0, 1, \ldots, 98$. In particular,

$$P(Y = 49) = \frac{1}{99}.$$

59. Let $X|p \sim \text{Bin}(n, p)$, with $p \sim \text{Beta}(a, b)$. So X has a *Beta-Binomial distribution*, as mentioned in Story 8.3.3 and Example 8.5.3. Find $E(X)$ and $\text{Var}(X)$.

Solution: By Adams's law,

$$E(X) = E(E(X|p)) = E(np) = nE(p) = \frac{na}{a+b}.$$

By Eve's law,

$$\text{Var}(X) = E(\text{Var}(X|p)) + \text{Var}(E(X|p)) = E(np(1-p)) + \text{Var}(np) = nE(p) - nE(p^2) + n^2\text{Var}(p).$$

Let

$$c = E(p) = \frac{a}{a+b}$$

and

$$v = \text{Var}(p) = \frac{c(1-c)}{a+b+1}.$$

Then

$$\text{Var}(X) = nc - n(v+c^2) + n^2 v = nc(1-c) + n(n-1)v.$$

In terms of a, b, this simplifies to

$$\text{Var}(X) = \frac{nab(n+a+b)}{(a+b)^2(a+b+1)}.$$

Sanity check: In the extreme case $n = 1$, X is Bernoulli with parameter

$$P(X = 1) = E(X) = \frac{a}{a+b},$$

so in this case

$$\text{Var}(X) = \frac{a}{a+b}\left(1 - \frac{a}{a+b}\right) = \frac{ab}{(a+b)^2},$$

which agrees with the result

$$\text{Var}(X) = \frac{1ab(1+a+b)}{(a+b)^2(a+b+1)} = \frac{ab}{(a+b)^2}.$$

60. An election is being held. There are two candidates, A and B, and there are n voters. The probability of voting for Candidate A varies by city. There are m cities, labeled $1, 2, \ldots, m$. The jth city has n_j voters, so $n_1 + n_2 + \cdots + n_m = n$. Let X_j be the number of people in the jth city who vote for Candidate A, with $X_j|p_j \sim \text{Bin}(n_j, p_j)$. To reflect our uncertainty about the probability of voting in each city, we treat $p_1, \ldots, p_m$ as r.v.s, with prior distribution asserting that they are i.i.d. Unif$(0, 1)$. Assume that $X_1, \ldots, X_m$ are independent, both unconditionally and conditional on $p_1, \ldots, p_m$. Let X be the total number of votes for Candidate A.

    (a) Find the marginal distribution of X_1 and the posterior distribution of $p_1|X_1 = k_1$.

    (b) Find $E(X)$ and $\text{Var}(X)$ in terms of n and s, where $s = n_1^2 + n_2^2 + \cdots + n_m^2$.

    Solution:

    (a) The marginal PMF of X_1 is

    $$P(X_1 = k_1) = \int_0^1 P(X_1 = k_1|p_1)dp_1 = \int_0^1 \binom{n_1}{k_1} p_1^{k_1}(1-p_1)^{n_1-k_1} dp_1 = \frac{1}{n_1+1}$$

    for $k_1 \in \{0, 1, \ldots, n_1\}$, by Bayes' billiards or by pattern-matching to a Beta PDF. This is a *Discrete Uniform distribution* over $\{0, 1, \ldots, n_1\}$. The desired posterior distribution is $p_1|X_1 = k_1 \sim \text{Beta}(1 + k_1, 1 + n_1 - k_1)$, since the Beta is the conjugate prior for the Binomial.

    (b) By Adam's law, $E(X_j) = E(E(X_j|p_j)) = E(n_j p_j) = n_j/2$. By Eve's Law,

    $$\text{Var}(X_j) = E(\text{Var}(X_j|p_j)) + \text{Var}(E(X_j|p_j)) = E(n_j p_j(1-p_j)) + \text{Var}(n_j p_j)$$

    $$= n_j(E(p_j) - E(p_j^2)) + \frac{n_j^2}{12} = \frac{n_j}{6} + \frac{n_j^2}{12}.$$

    So by linearity (for the mean) and independence of the X_j's (for the variance),

    $$E(X) = E(X_1) + \cdots + E(X_m) = \frac{n}{2},$$

    $$\text{Var}(X) = \text{Var}(X_1) + \cdots + \text{Var}(X_m) = \frac{n}{6} + \frac{s}{12}.$$

10

Inequalities and limit theorems

Inequalities

1. Ⓢ In a national survey, a random sample of people are chosen and asked whether they support a certain policy. Assume that everyone in the population is equally likely to be surveyed at each step, and that the sampling is with replacement (sampling without replacement is typically more realistic, but with replacement will be a good approximation if the sample size is small compared to the population size). Let n be the sample size, and let $\hat{p}$ and p be the proportion of people who support the policy in the sample and in the entire population, respectively. Show that for every $c > 0$,

$$P(|\hat{p} - p| > c) \leq \frac{1}{4nc^2}.$$

Solution: We can write $\hat{p} = X/n$ with $X \sim \text{Bin}(n, p)$. So $E(\hat{p}) = p, \text{Var}(\hat{p}) = p(1-p)/n$. Then by Chebyshev's inequality,

$$P(|\hat{p} - p| > c) \leq \frac{\text{Var}(\hat{p})}{c^2} = \frac{p(1-p)}{nc^2} \leq \frac{1}{4nc^2},$$

where the last inequality is because $p(1-p)$ is maximized at $p = 1/2$.

2. Ⓢ For i.i.d. r.v.s $X_1, \ldots, X_n$ with mean μ and variance σ^2, give a value of n (as a specific number) that will ensure that there is at least a 99% chance that the sample mean will be within 2 standard deviations of the true mean μ.

Solution: We have to find n such that

$$P(|\overline{X}_n - \mu| > 2\sigma) \leq 0.01.$$

By Chebyshev's inequality (in the form $P(|Y - EY| > c) \leq \frac{\text{Var}(Y)}{c^2}$), we have

$$P(|\overline{X}_n - \mu| > 2\sigma) \leq \frac{\text{Var}\overline{X}_n}{(2\sigma)^2} = \frac{\frac{\sigma^2}{n}}{4\sigma^2} = \frac{1}{4n}.$$

So the desired inequality holds if $n \geq 25$.

3. Ⓢ Show that for any two positive r.v.s X and Y with neither a constant multiple of the other,
$$E(X/Y)E(Y/X) > 1.$$

Solution: The r.v. $W = Y/X$ is positive and non-constant, so Jensen's inequality yields

$$E(X/Y) = E(1/W) > 1/E(W) = 1/E(Y/X).$$

4. Ⓢ The famous *arithmetic mean-geometric mean* inequality says that for any positive numbers $a_1, a_2, \ldots, a_n$,

$$\frac{a_1 + a_2 + \cdots + a_n}{n} \geq (a_1 a_2 \cdots a_n)^{1/n}.$$

314

Show that this inequality follows from Jensen's inequality, by considering $E \log(X)$ for an r.v. X whose possible values are $a_1, \ldots, a_n$ (you should specify the PMF of X; if you want, you can assume that the a_j are distinct (no repetitions), but be sure to say so if you assume this).

Solution: Assume that the a_j are distinct, and let X be a random variable which takes values from $a_1, a_2, ..., a_n$ with equal probability (the case of repeated a_j's can be handled similarly, letting the probability of $X = a_j$ be m_j/n, where m_j is the number of times a_j appears in the list $a_1, \ldots, a_n$). Jensen's inequality gives $E(\log X) \leq \log(EX)$, since the log function is concave. The left-hand side is $\frac{1}{n} \sum_{i=1}^{n} \log a_i$, while the right hand-side is $\log \frac{a_1 + a_2 + \cdots + a_n}{n}$. So we have the following inequality:

$$\log \frac{a_1 + a_2 + \cdots + a_n}{n} \geq \frac{1}{n} \sum_{i=1}^{n} \log a_i$$

Thus,

$$\frac{a_1 + a_2 + \cdots + a_n}{n} \geq e^{\frac{1}{n} \sum_{i=1}^{n} \log a_i} = e^{\frac{\log(a_1 \cdots a_n)}{n}} = (a_1 \cdots a_n)^{1/n}.$$

5. Ⓢ Let X be a discrete r.v. whose distinct possible values are $x_0, x_1, \ldots$, and let $p_k = P(X = x_k)$. The entropy of X is $H(X) = \sum_{k=0}^{\infty} p_k \log_2(1/p_k)$.

(a) Find $H(X)$ for $X \sim \text{Geom}(p)$.

Hint: Use properties of logs, and interpret part of the sum as an expected value.

(b) Let X and Y be i.i.d. discrete r.v.s. Show that $P(X = Y) \geq 2^{-H(X)}$.

Hint: Consider $E(\log_2(W))$, where W is an r.v. taking value p_k with probability p_k.

Solution:

(a) We have

$$H(X) = -\sum_{k=0}^{\infty} (pq^k) \log_2(pq^k)$$
$$= -\log_2(p) \sum_{k=0}^{\infty} pq^k - \log_2(q) \sum_{k=0}^{\infty} kpq^k$$
$$= -\log_2(p) - \frac{q}{p} \log_2(q),$$

with $q = 1 - p$, since the first series is the sum of a Geom(p) PMF and the second series is the expected value of a Geom(p) r.v.

(b) Let W be as in the hint. By Jensen, $E(\log_2(W)) \leq \log_2(EW)$. But

$$E(\log_2(W)) = \sum_k p_k \log_2(p_k) = -H(X),$$
$$EW = \sum_k p_k^2 = P(X = Y),$$

so $-H(X) \leq \log_2 P(X = Y)$. Thus, $P(X = Y) \geq 2^{-H(X)}$.

6. Let X be a random variable with mean μ and variance σ^2. Show that
$$E(X - \mu)^4 \geq \sigma^4,$$
and use this to show that the kurtosis of X is at least -2.

Solution: Let $Y = (X - \mu)^2$. By Jensen's inequality,
$$E(X - \mu)^4 = E(Y^2) \geq (EY)^2 = (\text{Var}(X))^2 = \sigma^4.$$

Therefore,
$$\text{Kurt}(X) = \frac{E(X - \mu)^4}{\sigma^4} - 3 \geq 1 - 3 = -2.$$

Fill-in-the-blank inequalities

7. Ⓢ Let X and Y be i.i.d. positive r.v.s, and let $c > 0$. For each part below, fill in the appropriate equality or inequality symbol: write $=$ if the two sides are always equal, $\leq$ if the left-hand side is less than or equal to the right-hand side (but they are not necessarily equal), and similarly for $\geq$. If no relation holds in general, write ?.

(a) $E(\log(X))$ _____ $\log(E(X))$

(b) $E(X)$ _____ $\sqrt{E(X^2)}$

(c) $E(\sin^2(X)) + E(\cos^2(X))$ _____ 1

(d) $E(|X|)$ _____ $\sqrt{E(X^2)}$

(e) $P(X > c)$ _____ $\frac{E(X^3)}{c^3}$

(f) $P(X \leq Y)$ _____ $P(X \geq Y)$

(g) $E(XY)$ _____ $\sqrt{E(X^2)E(Y^2)}$

(h) $P(X + Y > 10)$ _____ $P(X > 5 \text{ or } Y > 5)$

(i) $E(\min(X, Y))$ _____ $\min(EX, EY)$

(j) $E(X/Y)$ _____ $\frac{EX}{EY}$

(k) $E(X^2(X^2 + 1))$ _____ $E(X^2(Y^2 + 1))$

(l) $E\left(\frac{X^3}{X^3+Y^3}\right)$ _____ $E(\frac{Y^3}{X^3+Y^3})$

Solution:

(a) $E(\log(X)) \leq \log(E(X))$ (by Jensen: logs are concave)

(b) $E(X) \leq \sqrt{E(X^2)}$ (since $\text{Var}(X) \geq 0$, or by Jensen)

(c) $E(\sin^2(X)) + E(\cos^2(X)) = 1$ (by linearity, trig identity)

(d) $E(|X|) \leq \sqrt{E(X^2)}$ (by (b) with $|X|$ in place of X; here $|X| = X$ anyway)

(e) $P(X > c) \leq \frac{E(X^3)}{c^3}$ (by Markov, after cubing both sides of $X > c$)

(f) $P(X \leq Y) = P(X \geq Y)$ (by symmetry, as X, Y are i.i.d.)

(g) $E(XY) \leq \sqrt{E(X^2)E(Y^2)}$ (by Cauchy-Schwarz)

(h) $P(X + Y > 10) \leq P(X > 5 \text{ or } Y > 5)$ (if $X + Y > 10$, then $X > 5$ or $Y > 5$)

(i) $E(\min(X, Y)) \leq \min(EX, EY)$ (since $\min(X, Y) \leq X$ gives $E \min(X, Y) \leq EX$, and similarly $E \min(X, Y) \leq EY$)

(j) $E(X/Y) \geq \frac{EX}{EY}$ (since $E(X/Y) = E(X)E(\frac{1}{Y})$, with $E(\frac{1}{Y}) \geq \frac{1}{EY}$ by Jensen)

(k) $E(X^2(X^2 + 1)) \geq E(X^2(Y^2 + 1))$ (since $E(X^4) \geq (EX^2)^2 = E(X^2)E(Y^2) = E(X^2Y^2)$, because X^2 and Y^2 are i.i.d. and independent implies uncorrelated)

(l) $E(\frac{X^3}{X^3+Y^3}) = E(\frac{Y^3}{X^3+Y^3})$ (by symmetry!)

8. Ⓢ Write the most appropriate of $\leq$, $\geq$, $=$, or ? in the blank for each part (where "?" means that no relation holds in general).

In (c) through (f), X and Y are i.i.d. (independent identically distributed) positive random variables. Assume that the various expected values exist.

(a) (probability that a roll of 2 fair dice totals 9) _____ (probability that a roll of 2 fair dice totals 10)

(b) (probability that at least 65% of 20 fair coin flips land Heads) _____ (probability that at least 65% of 2000 fair coin flips land Heads)

(c) $E(\sqrt{X})$ _____ $\sqrt{E(X)}$

(d) $E(\sin X)$ _____ $\sin(EX)$

(e) $P(X + Y > 4)$ _____ $P(X > 2)P(Y > 2)$

(f) $E\left((X+Y)^2\right)$ _____ $2E(X^2) + 2(EX)^2$

Solution:

(a) (probability that a roll of 2 fair dice totals 9) $\geq$ (probability that a roll of 2 fair dice totals 10)

The probability on the left is $4/36$ and that on the right is $3/36$ as there is only one way for both dice to show 5's.

(b) (probability that at least 65% of 20 fair coin flips land Heads) $\geq$ (probability that at least 65% of 2000 fair coin flips land Heads)

With a large number of coin flips, by the LLN it becomes likely that the proportion of Heads is close to $1/2$.

(c) $E(\sqrt{X}) \leq \sqrt{E(X)}$

By Jensen's inequality (or since $\text{Var}(\sqrt{X}) \geq 0$).

(d) $E(\sin X)$? $\sin(EX)$

The inequality can go in either direction. For example, let X be 0 or π with equal probabilities. Then $E(\sin X) = 0$, $\sin(EX) = 1$. But if we let X be $\pi/2$ or $5\pi/2$ with equal probabilities, then $E(\sin X) = 1$, $\sin(EX) = -1$.

(e) $P(X + Y > 4) \geq P(X > 2)P(Y > 2)$

The righthand side is $P(X > 2, Y > 2)$ by independence. The $\geq$ then holds since the event $X > 2, Y > 2$ is a subset of the event $X + Y > 4$.

(f) $E\left((X+Y)^2\right) = 2E(X^2) + 2(EX)^2$

The lefthand side is

$$E(X^2) + E(Y^2) + 2E(XY) = E(X^2) + E(Y^2) + 2E(X)E(Y) = 2E(X^2) + 2(EX)^2$$

since X and Y are i.i.d.

9. Let X and Y be i.i.d. continuous r.v.s. Assume that the various expressions below exist. Write the most appropriate of $\leq$, $\geq$, $=$, or ? in the blank for each part (where "?" means that no relation holds in general).
(a) $e^{-E(X)}$ _____ $E(e^{-X})$

(b) $P(X > Y + 3)$ _____ $P(Y > X + 3)$

(c) $P(X > Y + 3)$ _____ $P(X > Y - 3)$

(d) $E(X^4)$ ____ $(E(XY))^2$

(e) $\mathrm{Var}(Y)$ ____ $E(\mathrm{Var}(Y|X))$

(f) $P(|X + Y| > 3)$ ____ $E|X|$

Solution:

(a) $e^{-E(X)} \leq E(e^{-X})$ (by Jensen: $g(x) = e^{-x}$ is convex)

(b) $P(X > Y + 3) = P(Y > X + 3)$ (by symmetry)

(c) $P(X > Y + 3) \leq P(X > Y - 3)$ (since $X > Y + 3$ implies $X > Y - 3$)

(d) $E(X^4) \geq (E(XY))^2$ (by Jensen's inequality, after simplifying $(E(XY))^2 = (E(X)E(Y))^2 = (EX)^4$))

(e) $\mathrm{Var}(Y) = E(\mathrm{Var}(Y|X))$ (since $\mathrm{Var}(Y|X) = \mathrm{Var}(Y)$, by independence of X and Y; without independence, we would still have $\mathrm{Var}(Y) \geq E(\mathrm{Var}(Y|X))$ by Eve's law)

(f) $P(|X + Y| > 3) \leq E|X|$ (since

$$P(|X + Y| > 3) \leq \frac{1}{3}E|X + Y| \leq \frac{1}{3}(E|X| + E|Y|) = \frac{2}{3}E|X| \leq E|X|,$$

where the first inequality is by Markov's inequality)

10. ⑤ Let X and Y be positive random variables, *not necessarily independent*. Assume that the various expected values below exist. Write the most appropriate of $\leq$, $\geq$, $=$, or ? in the blank for each part (where "?" means that no relation holds in general).

(a) $(E(XY))^2$ ____ $E(X^2)E(Y^2)$

(b) $P(|X + Y| > 2)$ ____ $\frac{1}{10}E((X + Y)^4)$

(c) $E(\log(X + 3))$ ____ $\log(E(X + 3))$

(d) $E(X^2 e^X)$ ____ $E(X^2)E(e^X)$

(e) $P(X + Y = 2)$ ____ $P(X = 1)P(Y = 1)$

(f) $P(X + Y = 2)$ ____ $P(\{X \geq 1\} \cup \{Y \geq 1\})$

Solution:

(a) $(E(XY))^2 \leq E(X^2)E(Y^2)$ (by Cauchy-Schwarz)

(b) $P(|X + Y| > 2) \leq \frac{1}{10}E((X + Y)^4)$ (by Markov's inequality)

(c) $E(\log(X + 3)) \leq \log(E(X + 3))$ (by Jensen)

(d) $E(X^2 e^X) \geq E(X^2)E(e^X)$ (since X^2 and e^X are positively correlated)

(e) $P(X + Y = 2)$? $P(X = 1)P(Y = 1)$ (What if X, Y are independent? What if $X \sim \mathrm{Bern}(1/2)$ and $Y = 1 - X$?)

(f) $P(X + Y = 2) \leq P(\{X \geq 1\} \cup \{Y \geq 1\})$ (the left event is a subset of the right event)

11. Ⓢ Let X and Y be positive random variables, *not necessarily independent.* Assume that the various expected values below exist. Write the most appropriate of $\leq, \geq, =$, or ? in the blank for each part (where "?" means that no relation holds in general).

(a) $E(X^3)$ _____ $\sqrt{E(X^2)E(X^4)}$

(b) $P(|X+Y| > 2)$ _____ $\frac{1}{16}E((X+Y)^4)$

(c) $E(\sqrt{X+3})$ _____ $\sqrt{E(X+3)}$

(d) $E(\sin^2(X)) + E(\cos^2(X))$ _____ 1

(e) $E(Y|X+3)$ _____ $E(Y|X)$

(f) $E(E(Y^2|X))$ _____ $(EY)^2$

Solution:

(a) $E(X^3) \leq \sqrt{E(X^2)E(X^4)}$ (by Cauchy-Schwarz)

(b) $P(|X+Y| > 2) \leq \frac{1}{16}E((X+Y)^4)$ (by Markov, taking 4th powers first)

(c) $E(\sqrt{X+3}) \leq \sqrt{E(X+3)}$ (by Jensen with a concave function)

(d) $E(\sin^2(X)) + E(\cos^2(X)) = 1$ (by linearity)

(e) $E(Y|X+3) = E(Y|X)$ (since knowing $X+3$ is equivalent to knowing X)

(f) $E(E(Y^2|X)) \geq (EY)^2$ (by Adam's law and Jensen)

12. Ⓢ Let X and Y be positive random variables, *not necessarily independent.* Assume that the various expressions below exist. Write the most appropriate of $\leq, \geq, =$, or ? in the blank for each part (where "?" means that no relation holds in general).

(a) $P(X+Y > 2)$ _____ $\frac{EX+EY}{2}$

(b) $P(X+Y > 3)$ _____ $P(X > 3)$

(c) $E(\cos(X))$ _____ $\cos(EX)$

(d) $E(X^{1/3})$ _____ $(EX)^{1/3}$

(e) $E(X^Y)$ _____ $(EX)^{EY}$

(f) $E(E(X|Y) + E(Y|X))$ _____ $EX + EY$

Solution:

(a) $P(X+Y > 2) \leq \frac{EX+EY}{2}$ (by Markov and linearity)

(b) $P(X+Y > 3) \geq P(X > 3)$ (since $X > 3$ implies $X+Y > 3$ since $Y > 0$)

(c) $E(\cos(X))? \cos(EX)$ (e.g., let $W \sim \text{Bern}(1/2)$ and $X = aW + b$ for various a, b)

(d) $E(X^{1/3}) \leq (EX)^{1/3}$ (by Jensen)

(e) $E(X^Y)?(EX)^{EY}$ (take X constant or Y constant as examples)

(f) $E(E(X|Y) + E(Y|X)) = EX + EY$ (by linearity and Adam's law)

13. ⓢ Let X and Y be i.i.d. positive random variables. Assume that the various expressions below exist. Write the most appropriate of $\leq, \geq, =,$ or ? in the blank for each part (where "?" means that no relation holds in general).

(a) $E(e^{X+Y})$ ____ $e^{2E(X)}$

(b) $E(X^2 e^X)$ ____ $\sqrt{E(X^4)E(e^{2X})}$

(c) $E(X|3X)$ ____ $E(X|2X)$

(d) $E(X^7 Y)$ ____ $E(X^7 E(Y|X))$

(e) $E(\frac{X}{Y} + \frac{Y}{X})$ ____ 2

(f) $P(|X - Y| > 2)$ ____ $\frac{\text{Var}(X)}{2}$

Solution:

(a) $E(e^{X+Y}) \geq e^{2E(X)}$ (write $E(e^{X+Y}) = E(e^X e^Y) = E(e^X)E(e^Y) = E(e^X)E(e^X)$ using the fact that X, Y are i.i.d., and then apply Jensen)

(b) $E(X^2 e^X) \leq \sqrt{E(X^4)E(e^{2X})}$ (by Cauchy-Schwarz)

(c) $E(X|3X) = E(X|2X)$ (knowing $2X$ is equivalent to knowing $3X$)

(d) $E(X^7 Y) = E(X^7 E(Y|X))$ (by Adam's law and taking out what's known)

(e) $E(\frac{X}{Y} + \frac{Y}{X}) \geq 2$ (since $E(\frac{X}{Y}) = E(X)E(\frac{1}{Y}) \geq \frac{EX}{EY} = 1$, and similarly $E(\frac{Y}{X}) \geq 1$)

(f) $P(|X - Y| > 2) \leq \frac{\text{Var}(X)}{2}$ (by Chebyshev, applied to the r.v. $W = X - Y$, which has variance $2\text{Var}(X)$: $P(|W - E(W)| > 2) \leq \text{Var}(W)/4 = \text{Var}(X)/2$)

14. ⓢ Let X and Y be i.i.d. Gamma$(\frac{1}{2}, \frac{1}{2})$, and let $Z \sim \mathcal{N}(0,1)$ (note that X and Z may be dependent, and Y and Z may be dependent). For (a),(b),(c), write the most appropriate of $<, >, =,$ or ? in each blank; for (d),(e),(f), write the most appropriate of $\leq, \geq, =,$ or ? in each blank.

(a) $P(X < Y)$ ____ $1/2$

(b) $P(X = Z^2)$ ____ 1

(c) $P(Z \geq \frac{1}{X^4+Y^4+7})$ ____ 1

(d) $E(\frac{X}{X+Y})E((X+Y)^2)$ ____ $E(X^2) + (E(X))^2$

(e) $E(X^2 Z^2)$ ____ $\sqrt{E(X^4)E(X^2)}$

(f) $E((X + 2Y)^4)$ ____ 3^4

Solution:

(a) $P(X < Y) = 1/2$

This is since X and Y are i.i.d. continuous r.v.s.

(b) $P(X = Z^2)$? 1

This is since the probability is 0 if X and Z are independent, but it is 1 if X and Z^2 are the same r.v., which is possible since $Z^2 \sim \chi^2_1$, so $Z^2 \sim$ Gamma$(\frac{1}{2}, \frac{1}{2})$.

(c) $P(Z \geq \frac{1}{X^4+Y^4+7}) < 1$

This is since Z may be negative, and $\frac{1}{X^4+Y^4+7}$ is positive.

(d) $E(\frac{X}{X+Y})E((X+Y)^2) = E(X^2) + (E(X))^2$

By the bank–post office story, $X/(X + Y)$ and $(X + Y)^2$ are independent (and thus uncorrelated). So since X and Y are i.i.d., the lefthand side becomes

$$E(X(X + Y)) = E(X^2 + XY) = E(X^2) + E(XY) = E(X^2) + (E(X))^2.$$

(e) $E(X^2 Z^2) \le \sqrt{E(X^4)E(X^2)}$

By Cauchy-Schwarz, $E(X^2 Z^2) \le \sqrt{E(X^4)E(Z^4)}$. And $E(Z^4) = E(X^4)$ since X and Z^2 are χ_1^2, or since $E(Z^4) = 3$ (as shown in Chapter 6) and $E(X^2) = \text{Var}(X) + (E(X))^2 = 2 + 1 = 3$.

(f) $E((X + 2Y)^4) \ge 3^4$

This is true by Jensen's inequality, since $E(X + 2Y) = 1 + 2 = 3$.

15. Let X, Y, Z be i.i.d. $\mathcal{N}(0, 1)$ r.v.s. Write the most appropriate of $\le, \ge, =$, or ? in each blank (where "?" means that no relation holds in general).

(a) $P(X^2 + Y^2 + Z^2 > 6)$ ____ $1/2$

(b) $P(X^2 < 1)$ ____ $2/3$

(c) $E\left(\frac{X^2}{X^2 + Y^2 + Z^2}\right)$ ____ $1/4$

(d) $\text{Var}(\Phi(X) + \Phi(Y) + \Phi(Z))$ ____ $1/4$

(e) $E(e^{-X})$ ____ $E(e^X)$

(f) $E(|X|e^X)$ ____ $\sqrt{E(e^{2X})}$

Solution:

(a) $P(X^2 + Y^2 + Z^2 > 6) \le 1/2$.

This is true by Markov's Inequality, since $E(X^2 + Y^2 + Z^2) = 3E(X^2) = 3$.

(b) $P(X^2 < 1) \ge 2/3$.

This is since $P(X^2 < 1) = P(-1 < X < 1) \approx 0.68$ by the 68-95-99.7% Rule.

(c) $E\left(\frac{X^2}{X^2 + Y^2 + Z^2}\right) \ge 1/4$.

The left-hand side is $1/3$ since $E\left(\frac{X^2}{X^2 + Y^2 + Z^2}\right) = E\left(\frac{Y^2}{X^2 + Y^2 + Z^2}\right) = E\left(\frac{Z^2}{X^2 + Y^2 + Z^2}\right)$ by symmetry, and the sum of these 3 expectations is 1 by linearity. Alternatively, note that $X^2/(X^2 + Y^2 + Z^2) \sim \text{Beta}(1/2, 1)$ by the bank–post office story.

(d) $\text{Var}(\Phi(X) + \Phi(Y) + \Phi(Z)) = 1/4$.

By universality of the Uniform, $\Phi(X) \sim \text{Unif}(0, 1)$. Since $\Phi(X), \Phi(Y), \Phi(Z)$ are i.i.d., the variance of the sum is then $3/12 = 1/4$.

(e) $E(e^{-X}) = E(e^X)$.

This is true since $-X$ has the same distribution as X, by symmetry of the Normal.

(f) $E(|X|e^X) \le \sqrt{E(e^{2X})}$.

This is true by Cauchy-Schwarz, since $E(|X|^2) = E(X^2) = 1$.

16. Let X, Y, Z, W be i.i.d. positive r.v.s with CDF F and $E(X) = 1$. Write the most appropriate of $\leq, \geq, =,$ or ? in each blank (where "?" means that no relation holds in general).

   (a) $F(3)$ ____ $2/3$

   (b) $(F(3))^3$ ____ $P(X + Y + Z \leq 9)$

   (c) $E\left(\frac{X^2}{X^2+Y^2+Z^2+W^2}\right)$ ____ $1/4$

   (d) $E(XYZW)$ ____ $E(X^4)$

   (e) $\mathrm{Var}(E(Y|X))$ ____ $\mathrm{Var}(Y)$

   (f) $\mathrm{Cov}(X + Y, X - Y)$ ____ 0

   Solution:

   (a) $F(3) \geq 2/3$

   By Markov's inequality, $P(X > 3) \leq 1/3$, so $F(3) = 1 - P(X > 3) \geq 2/3$.

   (b) $(F(3))^3 \leq P(X + Y + Z \leq 9)$

   This is since $F(3)^3 = P(X \leq 3, Y \leq 3, Z \leq 3) \leq P(X + Y + Z \leq 9)$.

   (c) $E\left(\frac{X^2}{X^2+Y^2+Z^2+W^2}\right) = 1/4$

   By symmetry, $E\left(\frac{X^2}{X^2+Y^2+Z^2+W^2}\right) = \cdots = E\left(\frac{W^2}{X^2+Y^2+Z^2+W^2}\right)$. By linearity, these sum to 1.

   (d) $E(XYZW) \leq E(X^4)$

   By independence, $E(XYZW) = E(X)E(Y)E(Z)E(W) = (EX)^4$. By Jensen's inequality, $E(X^4) \geq (EX)^4$.

   (e) $\mathrm{Var}(E(Y|X)) \leq \mathrm{Var}(Y)$

   By Eve's law, $\mathrm{Var}(Y) = E(\mathrm{Var}(Y|X)) + \mathrm{Var}(E(Y|X)) \geq \mathrm{Var}(E(Y|X))$. (This is true even if X and Y are dependent. Here though they are independent, so in fact $E(Y|X) = E(Y)$ is a constant, which shows that $\mathrm{Var}(E(Y|X)) = 0$.)

   (f) $\mathrm{Cov}(X + Y, X - Y) = 0$

   This is since $\mathrm{Cov}(X + Y, X - Y) = \mathrm{Var}(X) - \mathrm{Cov}(X, Y) + \mathrm{Cov}(Y, X) - \mathrm{Var}(Y) = 0$.

17. Let X, Y, Z be i.i.d. (independent, identically distributed) r.v.s with a continuous distribution. Write the most appropriate of $\leq, \geq, =,$ or ? in each blank (where "?" means that no relation holds in general).

   (a) $P(X < Y < Z)$ ____ $1/6$

   (b) $P(X > 1)$ ____ $E(X)$

   (c) $P\left(\sum_{k=0}^{2015}\left(\frac{X^2+1}{X^2+2}\right)^k > 3\right)$ ____ $P(X^2 > 1)$

   (d) $E\left(\sqrt{X^2 + Y^2}\right)$ ____ $\sqrt{E(X^2) + E(Y^2)}$

   (e) $\mathrm{Var}(Y^2|Z)$ ____ $\mathrm{Var}(X^2|X)$

   (f) $\mathrm{Var}(X - 2Y + 3Z)$ ____ $14\mathrm{Var}(X)$

   Solution:

(a) $P(X < Y < Z) = 1/6$ since all 3! orderings are equally likely, by symmetry.

(b) $P(X > 1)$? $E(X)$ since, for example, if $E(X) < 0$ then $P(X > 1) > E(X)$, while if X is nonnegative then $P(X > 1) \le E(X)$ by Markov's inequality.

(c) $P\left(\sum_{k=0}^{2015} \left(\frac{X^2 + 1}{X^2 + 2}\right)^k > 3\right) \le P(X^2 > 1)$ since, letting $R = \frac{X^2+1}{X^2+2}$, we have

$$\sum_{k=0}^{2015} R^k < \sum_{k=0}^{\infty} R^k = \frac{1}{1 - R} = X^2 + 2,$$

so $\sum_{k=0}^{2015} R^k > 3$ implies $X^2 + 2 > 3$, which in turn implies $X^2 > 1$.

(d) $E\left(\sqrt{X^2 + Y^2}\right) \le \sqrt{E(X^2) + E(Y^2)}$ by Jensen's inequality and linearity.

(e) $\text{Var}(Y^2|Z) \ge \text{Var}(X^2|X)$ since $\text{Var}(X^2|X) = 0$, while $\text{Var}(Y^2|Z) = \text{Var}(Y^2) \ge 0$.

(f) $\text{Var}(X - 2Y + 3Z) = 14\text{Var}(X)$ since

$$\text{Var}(X - 2Y + 3Z) = \text{Var}(X) + 2^2\text{Var}(Y) + 3^2\text{Var}(Z) = 14\text{Var}(X).$$

18. Let $X, Y, Z \sim \text{Bin}(n, p)$ be i.i.d. (independent and identically distributed). Write the most appropriate of $\le, \ge, =,$ or ? in each blank (where "?" means that no relation holds in general).

(a) $P(X < Y < Z)$ ____ $1/6$

(b) $P(X + Y + Z > n)$ ____ $3p$

(c) $P\left(\sum_{n=0}^{2016} \frac{(X^2 + 1)^n}{n!} > e^5\right)$ ____ $P(X > 2)$

(d) $E\left(e^X\right)$ ____ $(pe + 1 - p)^n$

(e) $\text{Var}(X + Y)$ ____ $n/2$

(f) $E(X \mid X + Y = n)$ ____ $n/2$

Solution:

(a) $P(X < Y < Z) \le 1/6$

All $3! = 6$ orderings of X, Y, Z are equally likely by symmetry, but events with ties, such as $X = Y = Z$, also have positive probability since X, Y, Z are discrete.

(b) $P(X + Y + Z > n) \le 3p$

By Markov's inequality,

$$P(X + Y + Z > n) \le \frac{E(X + Y + Z)}{n} = 3p.$$

(c) $P\left(\sum_{n=0}^{2016} \frac{(X^2 + 1)^n}{n!} > e^5\right) = P(X > 2)$

Using the fact that if event A implies event B, then $P(A) \le P(B)$,

$$P\left(\sum_{n=0}^{2016} \frac{(X^2 + 1)^n}{n!} > e^5\right) \le P\left(\sum_{n=0}^{\infty} \frac{(X^2 + 1)^n}{n!} > e^5\right) = P(e^{X^2+1} > e^5) = P(X > 2).$$

Conversely, if $X > 2$ then $X \geq 3$ (since X is integer-valued), and then

$$\sum_{n=0}^{2016} \frac{(X^2+1)^n}{n!} \geq \sum_{n=0}^{2016} \frac{(3^2+1)^n}{n!} = \sum_{n=0}^{2016} \frac{10^n}{n!} \approx e^{10}.$$

(d) $E\left(e^X\right) = (pe + 1 - p)^n$

If $I \sim \text{Bern}(p)$, then $E(e^I) = pe + 1 - p$. Writing X as a sum of n i.i.d. $\text{Bern}(p)$ r.v.s, we have $E(e^X) = (pe + 1 - p)^n$. Alternatively, we can use the MGF of a Binomial.

(e) $\text{Var}(X + Y) \leq n/2$

We have $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) = 2np(1-p) \leq n/2$, since $p(1-p)$ is maximized at $p = 1/2$.

(f) $E(X \mid X + Y = n) = n/2$

By symmetry, $E(X|X + Y = n) = E(Y|X + Y = n)$. And by linearity,

$$E(X \mid X + Y = n) + E(Y \mid X + Y = n) = E(X + Y \mid X + Y = n) = n.$$

19. Let $X, Y, Z, W \sim \mathcal{N}(c, c^2)$ be i.i.d., where $c > 0$. Write the most appropriate of $\leq$, $\geq$, $=$, or ? in each blank (where "?" means that no relation holds in general).
     (a) $P(X + Y \leq Z - W)$ ____ $\Phi(-1)$

     (b) $P(X^4 - Y^8 \leq Z^4 - W^8)$ ____ $\Phi(-1)$

     (c) $E(X - c)^2$ ____ c

     (d) $E(X - c)^3$ ____ c

     (e) $E\left(\frac{1}{4}X^2 + X\right)$ ____ $\frac{1}{2}c^2 + c$

     (f) $\log\left(E(e^X)\right)$ ____ $\frac{1}{2}c^2 + c$

     Solution:

     (a) $P(X + Y \leq Z - W) = \Phi(-1)$

     since, letting $T = X + Y - Z + W \sim \mathcal{N}(2c, 4c^2)$,

     $$P(X + Y \leq Z - W) = P(T \leq 0) = P\left(\frac{T - 2c}{2c} \leq \frac{-2c}{2c}\right) = \Phi(-1).$$

     (b) $P(X^4 - Y^8 \leq Z^4 - W^8) \geq \Phi(-1)$

     since $X^4 - Y^8$ and $Z^4 - W^8$ are i.i.d., so the left-hand side is $\frac{1}{2}$, whereas

     $$\Phi(-1) < \Phi(0) = \frac{1}{2}.$$

     (c) $E(X - c)^2$? c

     since $E(X - c)^2 = \text{Var}(X) = c^2$, which is less than c if $c < 1$, equal to c if $c = 1$, and greater than c if $c > 1$.

     (d) $E(X - c)^3 \leq c$

     since $X - c \sim \mathcal{N}(0, c^2)$, which is symmetric about 0, so the odd moments of $X - c$ are 0.

(e) $E\left(\frac{1}{4}X^2 + X\right) = \frac{1}{2}c^2 + c$

since, by linearity,

$$E\left(\frac{1}{4}X^2 + X\right) = \frac{1}{4}\left(\text{Var}(X) + (EX)^2\right) + EX = \frac{1}{4}(c^2 + c^2) + c = \frac{1}{2}c^2 + c.$$

(f) $\log\left(E(e^X)\right) = \frac{1}{2}c^2 + c$

since e^X is Log-Normal, with expected value $e^{c+c^2/2}$.

20. Let $X, Y \sim \text{Pois}(\lambda)$ be i.i.d., where $\lambda > 1$. Write the most appropriate of $\leq, \geq, =$, or ? in each blank (where "?" means that no relation holds in general).

(a) $P(X \leq Y)$ _____ $1/2$

(b) $P(X + Y \leq 1)$ _____ $3e^{-2\lambda}$

(c) $E\left(e^{X+Y}\right)$ _____ $e^{2\lambda}$

(d) $E(X \mid X + Y = 4)$ _____ 2

(e) $\text{Var}(X \mid X + Y = 4)$ _____ 1

(f) $E(X^2 - Y)$ _____ λ^2

Solution:

(a) $P(X \leq Y) \geq 1/2$

By symmetry, $P(X < Y) = P(Y < X)$. So

$$1 = P(X < Y) + P(X = Y) + P(X > Y) = 2P(X > Y) + P(X = Y) > 2P(X > Y),$$

which shows that $P(X > Y) < 1/2$. Therefore, $P(X \leq Y) = 1 - P(X > Y) > 1/2$.

(b) $P(X + Y \leq 1) \geq 3e^{-2\lambda}$

We have $X + Y \sim \text{Pois}(2\lambda)$ and $\lambda > 1$, so

$$P(X + Y \leq 1) = P(X + Y = 0) + P(X + Y = 1) = (1 + 2\lambda)e^{-2\lambda} > 3e^{-2\lambda}.$$

(c) $E\left(e^{X+Y}\right) \geq e^{2\lambda}$

By Jensen's inequality, $E\left(e^{X+Y}\right) > e^{E(X+Y)} = e^{2\lambda}$.

(d) $E(X \mid X + Y = 4) = 2$

Let $c = E(X|X + Y = 4)$. By symmetry, $c = E(X|X + Y = 4) = E(Y|X + Y = 4)$, so

$$2c = E(X|X + Y = 4) + E(Y|X + Y = 4) = E(X + Y|X + Y = 4) = 4.$$

Alternatively, we can use the conditional distribution given in (e).

(e) $\text{Var}(X \mid X + Y = 4) = 1$

The conditional distribution of X given $X + Y = 4$ is $\text{Bin}(4, 1/2)$, by a result from Chapter 4 or by using Bayes' rule to compute the conditional PMF $P(X = k|X+Y = 4)$. The variance of the $\text{Bin}(4, 1/2)$ distribution is $4 \cdot \frac{1}{2} \cdot \frac{1}{2} = 1$.

(f) $E(X^2 - Y) = \lambda^2$

Note that $E(X^2) = \text{Var}(X) + (EX)^2 = \lambda + \lambda^2$. So by linearity,

$$E(X^2 - Y) = E(X^2) - EY = \lambda + \lambda^2 - \lambda = \lambda^2.$$

LLN and CLT

21. Ⓢ Let $X_1, X_2, \ldots$ be i.i.d. positive random variables with mean 2. Let $Y_1, Y_2, \ldots$ be i.i.d. positive random variables with mean 3. Show that

$$\frac{X_1 + X_2 + \cdots + X_n}{Y_1 + Y_2 + \cdots + Y_n} \to \frac{2}{3}$$

with probability 1. Does it matter whether the X_i are independent of the Y_j?

Solution: By the law of large numbers,

$$\frac{X_1 + X_2 + \cdots + X_n}{n} \to 2$$

with probability 1 and

$$\frac{Y_1 + Y_2 + \cdots + Y_n}{n} \to 3$$

with probability 1, as $n \to \infty$. Note that if two events A and B both have probability 1, then the event $A \cap B$ also has probability 1. So with probability 1, *both* the convergence involving the X_i and the convergence involving the Y_j occur. Therefore,

$$\frac{X_1 + X_2 + \cdots + X_n}{Y_1 + Y_2 + \cdots + Y_n} = \frac{(X_1 + X_2 + \cdots + X_n)/n}{(Y_1 + Y_2 + \cdots + Y_n)/n} \to \frac{2}{3} \text{ with probability 1}$$

as $n \to \infty$. It was not necessary to assume that the X_i are independent of the Y_j because of the pointwise with probability 1 convergence.

22. Ⓢ Let $U_1, U_2, \ldots, U_{60}$ be i.i.d. Unif(0,1) and $X = U_1 + U_2 + \cdots + U_{60}$.

(a) Which important distribution is the distribution of X very close to? Specify what the parameters are, and state which theorem justifies your choice.

(b) Give a simple but accurate approximation for $P(X > 17)$. Justify briefly.

Solution:

(a) By the central limit theorem, the distribution is approximately $\mathcal{N}(30, 5)$ since $E(X) = 30, \operatorname{Var}(X) = 60/12 = 5$.

(b) We have

$$P(X > 17) = 1 - P(X \le 17) = 1 - P\left(\frac{X - 30}{\sqrt{5}} \le \frac{-13}{\sqrt{5}}\right) \approx 1 - \Phi\left(\frac{-13}{\sqrt{5}}\right) = \Phi\left(\frac{13}{\sqrt{5}}\right).$$

Since $13/\sqrt{5} > 5$, and we already have $\Phi(3) \approx 0.9985$ by the 68-95-99.7% rule, the value is extremely close to 1.

23. Ⓢ Let $V_n \sim \chi_n^2$ and $T_n \sim t_n$ for all positive integers n.

(a) Find numbers a_n and b_n such that $a_n(V_n - b_n)$ converges in distribution to $\mathcal{N}(0, 1)$.

(b) Show that $T_n^2/(n + T_n^2)$ has a Beta distribution (without using calculus).

Solution:

(a) By definition of χ_n^2, we can take $V_n = Z_1^2 + \cdots + Z_n^2$, where $Z_j \sim \mathcal{N}(0, 1)$ independently. We have $E(Z_1^2) = 1$ and $E(Z_1^4) = 3$, so $\operatorname{Var}(Z_1^2) = 2$. By the CLT, if we standardize V_n it will go to $\mathcal{N}(0, 1)$:

$$\frac{Z_1^2 + \cdots + Z_n^2 - n}{\sqrt{2n}} \to \mathcal{N}(0, 1) \text{ in distribution.}$$

So we can take $a_n = \frac{1}{\sqrt{2n}}$, $b_n = n$.

(b) We can take $T_n = Z_0/\sqrt{V_n/n}$, with $Z_0 \sim \mathcal{N}(0, 1)$ independent of V_n. Then we have $T_n^2/(n + T_n^2) = Z_0^2/(Z_0^2 + V_n)$, with $Z_0^2 \sim \operatorname{Gamma}(1/2, 1/2), V_n \sim \operatorname{Gamma}(n/2, 1/2)$. By the bank–post office story, $Z_0^2/(Z_0^2 + V_n) \sim \operatorname{Beta}(1/2, n/2)$.

24. ⓢ Let $T_1, T_2, \ldots$ be i.i.d. Student-t r.v.s with $m \geq 3$ degrees of freedom. Find constants a_n and b_n (in terms of m and n) such that $a_n(T_1 + T_2 + \cdots + T_n - b_n)$ converges to $\mathcal{N}(0, 1)$ in distribution as $n \to \infty$.

Solution: First let us find the mean and variance of each T_j. Let $T = \frac{Z}{\sqrt{V/m}}$ with $Z \sim \mathcal{N}(0, 1)$ independent of $V \sim \chi_m^2$. By LOTUS, for $G \sim \text{Gamma}(a, \lambda)$, $E(G^r)$ is $\lambda^{-r}\Gamma(a+r)/\Gamma(a)$ for $r > -a$, and does not exist for $r \leq -a$. So

$$
\begin{aligned}
E(T) &= E(Z)E\left(\frac{1}{\sqrt{V/m}}\right) = 0, \\
\text{Var}(T) = E(T^2) - (ET)^2 &= mE(Z^2)E\left(\frac{1}{V}\right) \\
&= m\frac{(1/2)\Gamma(m/2 - 1)}{\Gamma(m/2)} \\
&= \frac{m\Gamma(m/2 - 1)}{2\Gamma(m/2)} \\
&= \frac{m/2}{m/2 - 1} = \frac{m}{m - 2}.
\end{aligned}
$$

By the CLT, this is true for

$$
\begin{aligned}
b_n &= E(T_1) + \ldots + E(T_n) = 0, \\
a_n &= \frac{1}{\sqrt{\text{Var}(T_1) + \cdots + \text{Var}(T_n)}} = \sqrt{\frac{m - 2}{mn}}.
\end{aligned}
$$

25. ⓢ (a) Let $Y = e^X$, with $X \sim \text{Expo}(3)$. Find the mean and variance of Y.

(b) For $Y_1, \ldots, Y_n$ i.i.d. with the same distribution as Y from (a), what is the approximate distribution of the sample mean $\bar{Y}_n = \frac{1}{n}\sum_{j=1}^n Y_j$ when n is large?

Solution:

(a) By LOTUS,

$$
E(Y) = \int_0^\infty e^x(3e^{-3x})dx = \frac{3}{2},
$$
$$
E(Y^2) = \int_0^\infty e^{2x}(3e^{-3x})dx = 3.
$$

So $E(Y) = 3/2, \text{Var}(Y) = 3 - 9/4 = 3/4$.

(b) By the CLT, $\bar{Y}_n$ is approximately $\mathcal{N}(\frac{3}{2}, \frac{3}{4n})$ for large n.

26. ⓢ (a) Explain why the $\text{Pois}(n)$ distribution is approximately Normal if n is a large positive integer (specifying what the parameters of the Normal are).

(b) Stirling's formula is an amazingly accurate approximation for factorials:

$$
n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n,
$$

where in fact the ratio of the two sides goes to 1 as $n \to \infty$. Use (a) to give a quick heuristic derivation of Stirling's formula by using a Normal approximation to the probability that a $\text{Pois}(n)$ r.v. is n, with the continuity correction: first write $P(N = n) = P(n - \frac{1}{2} < N < n + \frac{1}{2})$, where $N \sim \text{Pois}(n)$.

Solution:

(a) Let $S_n = X_1 + \cdots + X_n$, with $X_1, X_2, \ldots$ i.i.d. $\sim \text{Pois}(1)$. Then $S_n \sim \text{Pois}(n)$ and for n large, S_n is approximately $\mathcal{N}(n, n)$ by the CLT.

(b) Let $N \sim \text{Pois}(n)$ and $X \sim \mathcal{N}(n, n)$. Then

$$P(N = n) \approx P\left(n - \frac{1}{2} < X < n + \frac{1}{2}\right) = \frac{1}{\sqrt{2\pi n}} \int_{n-1/2}^{n+1/2} e^{\frac{-(x-n)^2}{2n}} dx.$$

The integral is approximately 1 since the interval of integration has length 1 and for large n the integrand is very close to 1 throughout the interval. So

$$e^{-n} n^n / n! \approx (2\pi n)^{-1/2}.$$

Rearranging this gives exactly Stirling's formula.

27. Ⓢ (a) Consider i.i.d. $\text{Pois}(\lambda)$ r.v.s $X_1, X_2, \ldots$. The MGF of X_j is $M(t) = e^{\lambda(e^t - 1)}$. Find the MGF $M_n(t)$ of the sample mean $\bar{X}_n = \frac{1}{n} \sum_{j=1}^{n} X_j$.

(b) Find the limit of $M_n(t)$ as $n \to \infty$. (You can do this with almost no calculation using a relevant theorem; or you can use (a) and the fact that $e^x \approx 1 + x$ if x is very small.)

Solution:

(a) The MGF is

$$E(e^{\frac{t}{n}(X_1 + \cdots + X_n)}) = \left(E(e^{\frac{t}{n} X_1})\right)^n = e^{n\lambda(e^{t/n} - 1)},$$

since the X_j are i.i.d. and $E(e^{\frac{t}{n} X_1})$ is the MGF of X_1 evaluated at t/n.

(b) By the law of large numbers, $\bar{X}_n \to \lambda$ with probability 1. The MGF of the constant λ (viewed as an r.v. that always equals λ) is $e^{t\lambda}$. Thus, $M_n(t) \to e^{t\lambda}$ as $n \to \infty$.

28. Let $X_n \sim \text{Pois}(n)$ for all positive integers n. Use MGFs to show that the distribution of the standardized version of X_n converges to a Normal distribution as $n \to \infty$, without invoking the CLT.

Solution: Let $Z_n = (X_n - n)/\sqrt{n}$ be the standardized version of X_n. The MGF of Z_n is

$$M(t) = E(e^{tZ_n}) = e^{-t\sqrt{n}} E(e^{tX_n/\sqrt{n}}) = \exp(ne^{t/\sqrt{n}} - t\sqrt{n} - n),$$

using the result for the Poisson MGF. Expanding $e^{t/\sqrt{n}}$ with a Taylor series, we have

$$e^{t/\sqrt{n}} = 1 + \frac{t}{\sqrt{n}} + \frac{t^2}{2n} + r_n(t),$$

where $r_n(t)$ is the remainder term, which satisfies $nr_n(t) \to 0$ as $n \to \infty$. Then

$$ne^{t/\sqrt{n}} - t\sqrt{n} - n = \frac{t^2}{2} + nr_n(t) \to \frac{t^2}{2}$$

as $n \to \infty$. Thus, the MGF of Z_n converges to $e^{t^2/2}$, which is the $\mathcal{N}(0, 1)$ MGF. This shows that Z_n converges in distribution to $\mathcal{N}(0, 1)$.

29. An important concept in frequentist statistics is that of a *confidence interval* (CI). Suppose we observe data X from a distribution with parameter θ. Unlike in Bayesian statistics, θ is treated as a fixed but unknown constant; it is not given a prior distribution. A 95% confidence interval consists of a lower bound $L(X)$ and upper bound $U(X)$ such that

$$P(L(X) < \theta < U(X)) = 0.95$$

for all possible values of θ. Note that in the above statement, $L(X)$ and $U(X)$ are random variables, as they are functions of the r.v. X, whereas θ is a constant. The

definition says that the random interval $(L(X), U(X))$ has a 95% chance of containing the true value of θ.

Imagine an army of frequentists all over the world, independently generating 95% CIs. The jth frequentist observes data X_j and makes a confidence interval for the parameter θ_j. Show that if there are n of these frequentists, then the fraction of their intervals which contain the corresponding parameter approaches 0.95 as $n \to \infty$.

Hint: Consider the indicator r.v. $I_j = I(L(X_j) < \theta_j < U(X_j))$.

Solution: Let $I_j = I(L(X_j) < \theta_j < U(X_j))$. By the assumptions of the problem, the I_j's are i.i.d Bern(0.95). So by the law of large numbers, the fraction of the frequentists' intervals that contain the corresponding θ_j's is

$$\frac{I_1 + I_2 + \cdots + I_n}{n} \to E(I_1) = 0.95$$

as $n \to \infty$.

30. This problem extends Example 10.3.7 to a more general setting. Again, suppose a very volatile stock rises 70% or drops 50% in price, with equal probabilities and with different days independent.

(a) Suppose a hedge fund manager always invests half of her current fortune into the stock each day. Let Y_n be her fortune after n days, starting from an initial fortune of $Y_0 = 100$. What happens to Y_n as $n \to \infty$?

(b) More generally, suppose the hedge fund manager always invests a fraction α of her current fortune into the stock each day (in Part (a), we took $\alpha = 1/2$). With Y_0 and Y_n defined as in Part (a), find the function $g(\alpha)$ such that

$$\frac{\log Y_n}{n} \to g(\alpha)$$

with probability 1 as $n \to \infty$, and prove that $g(\alpha)$ is maximized when $\alpha = 2/7$.

Solution:

(a) Let U_n be the number of days in which the stock rises, among the first n days. Then

$$Y_n = 100(1 + 0.7/2)^{U_n}(1 - 0.5/2)^{n - U_n},$$

so

$$\log(Y_n) = U_n \log(1 + 0.35) + (n - U_n)\log(1 - 0.25) + \log(100).$$

By LLN, with probability 1 we have

$$\frac{\log(Y_n)}{n} = \frac{U_n}{n}\log(1.35) + \left(1 - \frac{U_n}{n}\right)\log(0.75) + \frac{1}{n}\log(100) \to 0.5\log(1.35 \cdot 0.75) = \frac{1}{2}\log(1.0125),$$

as $n \to \infty$. This limiting value for $\frac{1}{n}\log(Y_n)$ is positive since $1.0125 > 1$ (specifically, the limiting value is approximately 0.00621), so $\log(Y_n) \to \infty$ with probability 1, which shows that $Y_n \to \infty$ with probability 1.

(b) Now

$$\log(Y_n) = U_n \log(1 + 0.7\alpha) + (n - U_n)\log(1 - 0.5\alpha) + \log(100),$$

so by LLN, with probability 1

$$\frac{\log(Y_n)}{n} \to 0.5\log((1 + 0.7\alpha)(1 - 0.5\alpha)),$$

as $n \to \infty$. Therefore,

$$g(\alpha) = 0.5\log((1 + 0.7\alpha)(1 - 0.5\alpha)).$$

The value of α that maximizes g is the same as the value of α that maximizes the quadratic function

$$(1 + 0.7\alpha)(1 - 0.5\alpha) = -0.35\alpha^2 + 0.2\alpha + 1.$$

So the optimal value of α is

$$\alpha = \frac{0.2}{2 \cdot 0.35} = \frac{2}{7}.$$

Mixed practice

31. As in Exercise 36 from Chapter 3, there are n voters in an upcoming election in a certain country, where n is a large, even number. There are two candidates, A and B. Each voter chooses randomly whom to vote for, independently and with equal probabilities.

    (a) Use a Normal approximation (with continuity correction) to get an approximation for the probability of a tie, in terms of Φ.

    (b) Use a first-order Taylor expansion (linear approximation) to the approximation from Part (a) to show that the probability of a tie is approximately $1/\sqrt{cn}$, where c is a constant (which you should specify).

    Solution:

    (a) Approximating X with a $\mathcal{N}(n/2, n/4)$ r.v., which is valid by the CLT since X can be thought of as a sum of a lot of i.i.d. Bern$(1/2)$ r.v.s,

    $$P(X = n/2) = P\left(\frac{n}{2} - \frac{1}{2} \leq X \leq \frac{n}{2} + \frac{1}{2}\right) = P\left(\frac{-1}{\sqrt{n}} \leq \frac{X - n/2}{\sqrt{n/4}} \leq \frac{1}{\sqrt{n}}\right)$$

    $$\approx \Phi(1/\sqrt{n}) - \Phi(-1/\sqrt{n}) = 2\Phi(1/\sqrt{n}) - 1.$$

    (b) Since $\Phi'(x) = \varphi(x)$, the first order Taylor expansion of Φ about 0 is

    $$\Phi(x) \approx \Phi(0) + \Phi'(0)x = \frac{1}{2} + \frac{x}{\sqrt{2\pi}}.$$

    Thus,

    $$2\Phi\left(\frac{1}{\sqrt{n}}\right) - 1 \approx \frac{1}{\sqrt{\pi n/2}}.$$

    This approximation is quite accurate even for moderate values of n. For example, for $n = 52$, we have $1/\sqrt{\pi n/2} \approx 0.1106$ and $P(\text{election is tied}) \approx 0.1101$, using the R command `dbinom(26,52,1/2)`.

32. Cassie enters a casino with $X_0 = 1$ dollar and repeatedly plays the following game: with probability $1/3$, the amount of money she has increases by a factor of 3; with probability $2/3$, the amount of money she has decreases by a factor of 3. Let X_n be the amount of money she has after playing this game n times. For example, X_1 is 3 with probability $1/3$ and is 3^{-1} with probability $2/3$.

    (a) Compute $E(X_1)$, $E(X_2)$ and, in general, $E(X_n)$.

    (b) What happens to $E(X_n)$ as $n \to \infty$?

    (c) Let Y_n be the number of times out of the first n games that Cassie triples her money. What happens to Y_n/n as $n \to \infty$?

    (d) What happens to X_n as $n \to \infty$?

Solution:

(a) We have
$$E(X_1) = \frac{1}{3} \cdot 3 + \frac{2}{3} \cdot \frac{1}{3} = \frac{11}{9},$$

$$E(X_2) = E(E(X_2|X_1)) = E\left(\frac{1}{3} \cdot 3X_1 + \frac{2}{3} \cdot \frac{1}{3}X_1\right) = \left(\frac{11}{9}\right)^2.$$

In general,

$$E(X_n) = E(E(X_n|X_{n-1})) = E\left(\frac{1}{3} \cdot 3X_{n-1} + \frac{2}{3} \cdot \frac{1}{3}X_{n-1}\right) = \left(\frac{11}{9}\right)^n.$$

(b) It goes to ∞ since $\frac{11}{9} > 1$.

(c) It goes to $1/3$ (with probability 1), by the strong law of large numbers.

(d) It goes to 0 (with probability 1), since
$$X_n = 3^{Y_n} \cdot 3^{-(n-Y_n)} = 3^{2Y_n - n} = 3^{n(2Y_n/n - 1)},$$

where, by (c), the $2Y_n/n - 1$ in the exponent converges to $-1/3$, a negative number, which implies that the exponent $n(2Y_n/n - 1)$ goes to $-\infty$.

33. A handy rule of thumb in statistics and life is as follows:

Conditioning often makes things better.

This problem explores how the above rule of thumb applies to estimating unknown parameters. Let θ be an unknown parameter that we wish to estimate based on data $X_1, X_2, \ldots, X_n$ (these are r.v.s before being observed, and then after the experiment they "crystallize" into data). In this problem, θ is viewed as an unknown constant, and is not treated as an r.v. as in the Bayesian approach. Let T_1 be an estimator for θ (this means that T_1 is a function of $X_1, \ldots, X_n$ which is being used to estimate θ).

A strategy for improving T_1 (in some problems) is as follows. Suppose that we have an r.v. R such that $T_2 = E(T_1|R)$ is a function of $X_1, \ldots, X_n$ (in general, $E(T_1|R)$ might involve unknowns such as θ but then it couldn't be used as an estimator). Also suppose that $P(T_1 = T_2) < 1$, and that $E(T_1^2)$ is finite.

(a) Use Jensen's inequality to show that T_2 is better than T_1 in the sense that the mean squared error is less, i.e.,
$$E(T_2 - \theta)^2 < E(T_1 - \theta)^2.$$

Hint: Use Adam's law on the right-hand side.

(b) The *bias* of an estimator T for θ is defined to be $b(T) = E(T) - \theta$. An important identity in statistics, a form of the *bias-variance tradeoff*, is that mean squared error is variance plus squared bias:

$$E(T - \theta)^2 = \text{Var}(T) + (b(T))^2.$$

Use this identity and Eve's law to give an alternative proof of the result from (a).

(c) Now suppose that $X_1, X_2, \ldots$ are i.i.d. with mean θ, and consider the special case $T_1 = X_1$, $R = \sum_{j=1}^n X_j$. Find T_2 in simplified form, and check that it has lower mean squared error than T_1 for $n \geq 2$. Also, say what happens to T_1 and T_2 as $n \to \infty$.

Solution:

(a) Let $g_\theta(t) = (t - \theta)^2$. Then g_θ is strictly convex so by Jensen's inequality,

$$E(g_\theta(T_1)) = E(E(g_\theta(T_1)|R)) > E(g_\theta(E(T_1|R))) = E(g_\theta(T_2)).$$

(This technique for improving an estimator is called *Rao-Blackwellization*.)

(b) Note that T_1 and T_2 have the same bias, as $E(T_2) = E(E(T_1|R)) = E(T_1)$. By Eve's Law,

$$\text{Var}(T_1) = E(\text{Var}(T_1|R)) + \text{Var}(E(T_1|R)) \geq \text{Var}(T_2),$$

with strict inequality unless $\text{Var}(T_1|R) = 0$ has probability 1 (which would imply that $P(T_1 = T_2) = 1$). Therefore,

$$E(T_1 - \theta)^2 = \text{Var}(T_1) + (b(T_1))^2 > \text{Var}(T_2) + (b(T_2))^2 = E(T_2 - \theta)^2.$$

(c) As shown in Example 9.3.6,

$$T_2 = E(X_1|X_1 + \cdots + X_n) = \frac{1}{n}\sum_{j=1}^{n} X_j = \bar{X}_n.$$

Let $\sigma^2 = \text{Var}(X_1)$. The mean squared errors of T_1 and T_2 are their variances, since both have bias 0. For $n \geq 2$, $\text{Var}(T_1) = \sigma^2 > \sigma^2/n = \text{Var}(T_2)$. Since T_1 does not depend on n, nothing happens to it as $n \to \infty$. In contrast, by the law of large numbers we have that $T_2 = \bar{X}_n \to \theta$ as $n \to \infty$, with probability 1. This makes sense intuitively: the original estimator T_1 foolishly ignores all the data except for the first observation, and the Rao-Blackwellized estimator T_2 fixes this, allowing more and more information about θ to accumulate as n grows.

34. Each page of an n-page book has a Pois(λ) number of typos, where λ is unknown (but is not treated as an r.v.). Typos on different pages are independent. Thus we have i.i.d. $X_1, \ldots, X_n \sim$ Pois(λ), where X_j is the number of typos on page j. Suppose we are interested in estimating the probability θ that a page has no typos:

$$\theta = P(X_j = 0) = e^{-\lambda}.$$

(a) Let $\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$. Show that $T_n = e^{-\bar{X}_n}$ is biased for estimating θ. (In other words, $E(T_n) \neq \theta$.)

(b) Show that as $n \to \infty$, $T_n \to \theta$ with probability 1.

(c) Show that $W = \frac{1}{n}(I(X_1 = 0) + \cdots + I(X_n = 0))$ is unbiased for θ. Using the fact that $X_1|(X_1 + \cdots + X_n = s) \sim$ Bin($s, 1/n$), find $E(W|X_1 + \cdots + X_n)$. Is $\tilde{W} = E(W|X_1 + \cdots + X_n)$ also unbiased for θ?

(d) Using Eve's law or otherwise, show that $\tilde{W}$ has lower variance than W, and relate this to the previous question.

Solution:

(a) The function $g(x) = e^{-x}$ is convex. So by Jensen's inequality,

$$E(T_n) > \exp(-E(\bar{X}_n)) = \exp(-\lambda) = \theta.$$

(b) By LLN, $\bar{X}_n \to \lambda$ with probability 1. So $e^{-\bar{X}_n} \to e^{-\lambda}$ with probability 1.

(c) By linearity, symmetry, and the fundamental bridge, $E(W) = E(I(X_1 = 0)) = P(X_1 = 0) = \theta$. Since $X_1|(X_1 + \cdots + X_n = s) \sim$ Bin($s, 1/n$), we have

$$E(I(X_j = 0)|X_1 + \cdots + X_n) = \left(1 - \frac{1}{n}\right)^s,$$

so

$$E(W|X_1 + \cdots + X_n) = \left(1 - \frac{1}{n}\right)^{X_1 + \cdots + X_n}.$$

By Adam's law, $E(E(W|X_1 + \cdots + X_n)) = E(W) = \theta$, so $\tilde{W}$ is unbiased for θ.

(d) Let $S_n = \sum_{j=1}^{n} X_j$. By Eve's law,

$$\text{Var}(W) = E(\text{Var}(W|S_n)) + \text{Var}(E(W|S_n)) = E(\text{Var}(W|S_n)) + \text{Var}(\tilde{W}) > \text{Var}(\tilde{W}).$$

Strict inequality holds since W is not a deterministic function of S_n (if $S_n = 0$ occurs we know $W = 1$, and if $S_n = 1$ occurs we know $W = (n-1)/n$, but otherwise knowing the value of S_n does not determine the value of W).

35. A binary sequence is being generated through some process (random or deterministic). You need to sequentially predict each new number, i.e, you predict whether the next number will be 0 or 1, then observe it, then predict the next number, etc. Each of your predictions can be based on the entire past history of the sequence.

(a) Suppose for this part that the binary sequence consists of i.i.d. Bern(p) r.v.s, with p known. What is your optimal strategy (for each prediction, your goal is to maximize the probability of being correct)? What is the probability that you will guess the nth value correctly with this strategy?

(b) Now suppose that the binary sequence consists of i.i.d. Bern(p) r.v.s, with p unknown. Consider the following strategy: say 1 as your first prediction; after that, say "1" if the proportion of 1's so far is at least $1/2$, and say "0" otherwise. Find the limit as $n \to \infty$ of the probability of guessing the nth value correctly (in terms of p).

(c) Now suppose that you follow the strategy from (b), but that the binary sequence is generated by a nefarious entity who knows your strategy. What can the entity do to make your guesses be wrong as often as possible?

Solution:

(a) For $p > 1/2$, your optimal strategy is always to guess 1 since this maximizes the chance of being correct (the probability of being correct is p); likewise, for $p < 1/2$, your optimal strategy is always to guess 0 (the probability of being correct is $1 - p$). For $p = 1/2$, all strategies result in probability $1/2$ of being correct on any particular guess. In general, with the optimal strategy your probability of being correct on the nth guess is $\max(p, 1 - p)$.

(b) Consider the case $p > 1/2$. By the law of large numbers, the probability is 1 that the proportion of 1's will converge to p. In particular, the probability is 1 that from some point onward, the proportions of 1's will all be greater than $1/2$. But then from that point onward, your probability of being correct is p. Similarly, for $p < 1/2$, the probability is 1 that from some point onward, your probability of being correct is $1 - p$. For $p = 1/2$, any strategy results in probability $1/2$ of being correct. Thus, with probability 1 your probability of guessing the nth value correctly converges to $\max(p, 1 - p)$.

(c) By choosing the sequence $0, 1, 0, 1, 0, 1, \ldots$, the nefarious entity will make *all* your guesses wrong if you follow the strategy from (b).

36. Ⓢ Let X and Y be independent standard Normal r.v.s and let $R^2 = X^2 + Y^2$ (where $R > 0$ is the distance from (X, Y) to the origin).

(a) The distribution of R^2 is an example of three of the important distributions we have seen. State which three of these distributions R^2 is an instance of, specifying the parameter values.

(b) Find the PDF of R.

Hint: Start with the PDF $f_W(w)$ of $W = R^2$.

(c) Find $P(X > 2Y + 3)$ in terms of the standard Normal CDF Φ.

(d) Compute $\text{Cov}(R^2, X)$. Are R^2 and X independent?

Solution:

(a) It is χ_2^2, Expo(1/2), and Gamma(1,1/2).

(b) Since $R = \sqrt{W}$ with $f_W(w) = \frac{1}{2}e^{-w/2}$, we have

$$f_R(r) = f_W(w)|dw/dr| = \frac{1}{2}e^{-w/2}2r = re^{-r^2/2}, \text{ for } r > 0.$$

(This is the *Rayleigh distribution*, which was seen in Example 5.1.7.)

(c) We have
$$P(X > 2Y + 3) = P(X - 2Y > 3) = 1 - \Phi\left(\frac{3}{\sqrt{5}}\right)$$

since $X - 2Y \sim \mathcal{N}(0, 5)$.

(d) They are not independent since knowing X gives information about R^2, e.g., X^2 being large implies that R^2 is large. But R^2 and X are uncorrelated:

$$\text{Cov}(R^2, X) = \text{Cov}(X^2 + Y^2, X) = \text{Cov}(X^2, X) + \text{Cov}(Y^2, X) = E(X^3) - (EX^2)(EX) + 0 = 0.$$

37. Ⓢ Let $Z_1, \ldots, Z_n \sim \mathcal{N}(0, 1)$ be i.i.d.

(a) As a function of Z_1, create an Expo(1) r.v. X (your answer can also involve the standard Normal CDF Φ).

(b) Let $Y = e^{-R}$, where $R = \sqrt{Z_1^2 + \cdots + Z_n^2}$. Write down (but do not evaluate) an integral for $E(Y)$.

(c) Let $X_1 = 3Z_1 - 2Z_2$ and $X_2 = 4Z_1 + 6Z_2$. Determine whether X_1 and X_2 are independent (be sure to mention which results you're using).

Solution:

(a) Use Z_1 to get a Uniform and then the Uniform to get X: we have $\Phi(Z_1) \sim \text{Unif}(0, 1)$, and we can then take $X = -\log(1 - \Phi(Z_1))$. By symmetry, we can also use $-\log(\Phi(Z_1))$.

Sanity check: $0 < \Phi(Z_1) < 1$, so $-\log(\Phi(Z_1))$ is well-defined and positive.

(b) Let $W = Z_1^2 + \cdots + Z_n^2 \sim \chi_n^2$, so $Y = e^{-\sqrt{W}}$. We will use LOTUS to write $E(Y)$ using the PDF of W (there are other possible ways to use LOTUS here, but this is simplest since we get a single integral and we know the χ_n^2 PDF). This gives

$$E(Y) = \int_0^\infty e^{-\sqrt{w}} \frac{1}{2^{n/2}\Gamma(n/2)} w^{n/2-1} e^{-w/2} dw.$$

(c) They are uncorrelated:

$$\text{Cov}(X_1, X_2) = 12\text{Var}(Z_1) + 10\text{Cov}(Z_1, Z_2) - 12\text{Var}(Z_2) = 0.$$

Also, (X_1, X_2) is Multivariate Normal since any linear combination of X_1, X_2 can be written as a linear combination of Z_1, Z_2 (and thus is Normal since the sum of two independent Normals is Normal). So X_1 and X_2 are independent.

38. Ⓢ Let $X_1, X_2, \ldots$ be i.i.d. positive r.v.s. with mean μ, and let $W_n = \frac{X_1}{X_1 + \cdots + X_n}$.

(a) Find $E(W_n)$.

Hint: Consider $\frac{X_1}{X_1 + \cdots + X_n} + \frac{X_2}{X_1 + \cdots + X_n} + \cdots + \frac{X_n}{X_1 + \cdots + X_n}$.

(b) What random variable does nW_n converge to (with probability 1) as $n \to \infty$?

(c) For the case that $X_j \sim \text{Expo}(\lambda)$, find the distribution of W_n, preferably without using calculus. (If it is one of the named distributions, state its name and specify the parameters; otherwise, give the PDF.)

Solution:

(a) The expression in the hint equals 1, and by linearity and symmetry its expected value is $nE(W_n)$. So $E(W_n) = 1/n$.

Sanity check: in the case that the X_j are actually constants, $\frac{X_1}{X_1 + \cdots + X_n}$ reduces to $\frac{1}{n}$. Also in the case $X_j \sim \text{Expo}(\lambda)$, Part (c) shows that the answer should reduce to the mean of a $\text{Beta}(1, n-1)$ (which is $\frac{1}{n}$).

(b) By LLN, with probability 1 we have

$$nW_n = \frac{X_1}{(X_1 + \cdots + X_n)/n} \to \frac{X_1}{\mu} \text{ as } n \to \infty.$$

Sanity check: the answer should be a random variable since it's asked what random variable nW_n converges to. It should *not* depend on n since we let $n \to \infty$.

(c) Recall that $X_1 \sim \text{Gamma}(1)$ and $X_2 + \cdots + X_n \sim \text{Gamma}(n-1)$. By the connection between Beta and Gamma (i.e., the bank–post office story), $W_n \sim \text{Beta}(1, n-1)$.

Sanity check: The r.v. W_n clearly always takes values between 0 and 1, and the mean should agree with the answer from (a).

39. Let $X_1, X_2, \ldots$ be i.i.d. Expo(1).

(a) Let $N = \min\{n : X_n \geq 1\}$ be the index of the first X_j to exceed 1. Find the distribution of $N - 1$ (give the name and parameters), and hence find $E(N)$.

(b) Let $M = \min\{n : X_1 + X_2 + \cdots + X_n \geq 10\}$ be the number of X_j's we observe until their sum exceeds 10 for the first time. Find the distribution of $M - 1$ (give the name and parameters), and hence find $E(M)$.
Hint: Consider a Poisson process.

(c) Let $\bar{X}_n = (X_1 + \cdots + X_n)/n$. Find the exact distribution of $\bar{X}_n$ (give the name and parameters), as well as the approximate distribution of $\bar{X}_n$ for n large (give the name and parameters).

Solution:

(a) Each X_j has probability $1/e$ of exceeding 1, so $N - 1 \sim \text{Geom}(1/e)$ and $E(N) = e$.

(b) Interpret $X_1, X_2, \ldots$ as the interarrival times in a Poisson process of rate 1. Then $X_1 + X_2 + \cdots + X_j$ is the time of the jth arrival, so $M - 1$ is the number of arrivals in the time interval $[0, 10]$. Thus, $M - 1 \sim \text{Pois}(10)$ and $E(M) = 10 + 1 = 11$.

(c) We have $X_j/n \sim \text{Expo}(n)$, so $\bar{X}_n \sim \text{Gamma}(n, n)$. In particular, $E(\bar{X}_n) = 1, \text{Var}(\bar{X}_n) = 1/n$. By the CLT, the distribution of $\bar{X}_n$ is approximately $\mathcal{N}(1, 1/n)$ for n large.

11

Markov chains

Markov property

1. Ⓢ Let $X_0, X_1, X_2, \ldots$ be a Markov chain. Show that $X_0, X_2, X_4, X_6, \ldots$ is also a Markov chain, and explain why this makes sense intuitively.

   Solution: Let $Y_n = X_{2n}$; we need to show $Y_0, Y_1, \ldots$ is a Markov chain. By the definition of a Markov chain, we know that $X_{2n+1}, X_{2n+2}, \ldots$ ("the future" if we define the "present" to be time $2n$) is conditionally independent of $X_0, X_1, \ldots, X_{2n-2}, X_{2n-1}$ ("the past"), given X_{2n}. So given Y_n, we have that $Y_{n+1}, Y_{n+2}, \ldots$ is conditionally independent of $Y_0, Y_1, \ldots, Y_{n-1}$. Thus,

   $$P(Y_{n+1} = y | Y_0 = y_0, \ldots, Y_n = y_n) = P(Y_{n+1} = y | Y_n = y_n).$$

2. Ⓢ Let $X_0, X_1, X_2, \ldots$ be an irreducible Markov chain with state space $\{1, 2, \ldots, M\}$, $M \geq 3$, transition matrix $Q = (q_{ij})$, and stationary distribution $\mathbf{s} = (s_1, \ldots, s_M)$. Let the initial state X_0 follow the stationary distribution, i.e., $P(X_0 = i) = s_i$.

   (a) On average, how many of $X_0, X_1, \ldots, X_9$ equal 3? (In terms of $\mathbf{s}$; simplify.)

   (b) Let $Y_n = (X_n - 1)(X_n - 2)$. For $M = 3$, find an example of Q (the transition matrix for the *original* chain $X_0, X_1, \ldots$) where $Y_0, Y_1, \ldots$ is Markov, and another example of Q where $Y_0, Y_1, \ldots$ is not Markov. In your examples, make $q_{ii} > 0$ for at least one i and make sure it is possible to get from any state to any other state eventually.

   Solution:

   (a) Since X_0 has the stationary distribution, all of $X_0, X_1, \ldots$ have the stationary distribution. By indicator random variables, the expected value is $10s_3$.

   (b) Note that Y_n is 0 if X_n is 1 or 2, and Y_n is 2 otherwise. So the Y_n process can be viewed as merging states 1 and 2 of the X_n-chain into one state. Knowing the history of Y_n's means knowing when the X_n-chain is in state 3, without being able to distinguish state 1 from state 2.

   If $q_{13} = q_{23}$, then Y_n is Markov since given Y_n, even knowing the past $X_0, \ldots, X_n$ does not affect the transition probabilities. But if $q_{13} \neq q_{23}$, then the Y_n past history can give useful information about X_n, affecting the transition probabilities. So one example (not the only possible example!) is

   $$Q_1 = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix} \text{ (Markov)} \qquad Q_2 = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 1 & 0 & 0 \end{pmatrix} \text{ (not Markov).}$$

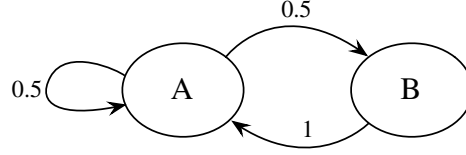   (c) The stationary distribution is uniform over all states:

   $$\mathbf{s} = (1/M, 1/M, \ldots, 1/M).$$

   This is because

   $$\begin{pmatrix} 1/M & 1/M & \ldots & 1/M \end{pmatrix} Q = \frac{1}{M} \begin{pmatrix} 1 & 1 & \ldots & 1 \end{pmatrix} Q = \begin{pmatrix} 1/M & 1/M & \ldots & 1/M \end{pmatrix},$$

where the matrix multiplication was done by noting that multiplying a row vector of 1's times Q gives the column sums of Q.

3. A Markov chain has two states, A and B, with transitions as follows:

Suppose we do not get to observe this Markov chain, which we'll call $X_0, X_1, X_2, \ldots$. Instead, whenever the chain transitions from A back to A, we observe a 0, and whenever it changes states, we observe a 1. Let the sequence of 0's and 1's be called $Y_0, Y_1, Y_2, \ldots$.

For example, if the X chain starts out as

$$A, A, B, A, B, A, A, \ldots$$

then the Y chain starts out as

$$0, 1, 1, 1, 1, 0, \ldots.$$

(a) Show that $Y_0, Y_1, Y_2, \ldots$ is not a Markov chain.

(b) In Example 11.1.3, we dealt with a violation of the Markov property by enlarging the state space to incorporate second-order dependence. Show that such a trick will not work for $Y_0, Y_1, Y_2, \ldots$. That is, no matter how large m is,

$$Z_n = (Y_{n-m+1}, Y_{n-m+2}, \ldots, Y_n), \ \ n = m - 1, m, \ldots$$

is still not a Markov chain.

Solution:

(a) Note that if the X chain is currently at A, then the next Y value is Bern$(1/2)$, but if the X chain is currently at B, then the next Y value will always be 1. So

$$P(Y_{n+1} = 1 | Y_n = 1, Y_{n-1} = 0) = 1,$$

since given that $Y_n = 1, Y_{n-1} = 0$, we have $X_n = A, X_{n+1} = B$, and then we will have $X_{n+2} = A$. But

$$P(Y_{n+1} = 1 | Y_n = 1) < 1$$

since just knowing $Y_n = 1$ isn't enough to determine where the X chain is just after the corresponding transition. So Y_{n+1} and Y_{n-1} are *not* conditionally independent given Y_n, which shows that $Y_0, Y_1, \ldots$ is *not* a Markov chain.

(b) Assume that the X chain starts at A (if it starts at B, then the next step will take it to A and we can study the chain from that time onwards). Decompose the Y chain into runs of 0's and runs of 1's. For example, $0, 1, 1, 1, 1, 0, 0, 0, 1, 1, 0, \ldots$ has a run of one 0, then a run of four 1's, then a run of three 0's, then a run of two 1's, etc.

Note that every run of 1's has *even* length, since a transition from A to B is always immediately followed by a transition from B back to A. No matter what fixed amount of history is incorporated into the state space, the earlier history may still be crucial for determining whether it is possible to have a 0 next. So Z_n is *not* a Markov chain.

4. There are three blocks, floating in a sea of lava. Label the blocks $1, 2, 3$, from left to right. Mark the Kangaroo, a video game character, is standing on block 1. To reach safety, he must get to block 3. He can't jump directly from block 1 to block 3; his only

hope is to jump from block 1 to block 2, then jump from block 2 to block 3. Each time he jumps, he has probability $1/2$ of success and probability $1/2$ of "dying" by falling into the lava. If he "dies", he starts again at block 1.

Let J be the total number of jumps that Mark will make in order to get to block 3.

(a) Find $E(J)$, using the Markov property and conditional expectation.

(b) Explain how this problem relates to a coin tossing problem from Chapter 9.

Solution:

(a) Mark will get to block 3 as soon as he has 2 consecutive successful jumps. Let H denote hopping to the next block and T denote tumbling into the lava. Then J is the waiting time until HH in a sequence of H's and T's, where H and T have probability $1/2$ each. We will follow the strategy of conditioning on the first jump. Let A be the event that the first jump succeeds. The Markov property says that each time Mark is back at block 1, the situation is reset. So

$$E(J) = \frac{1}{2} \cdot E(J|A) + \frac{1}{2} \cdot E(J|A^c) = \frac{1}{2}\left(\frac{1}{2} \cdot 2 + \frac{1}{2} \cdot (2 + E(J))\right) + \frac{1}{2}\left(1 + E(J)\right).$$

Solving this equation for $E(J)$, we have $E(J) = 6$.

(b) The problem amounts to finding the expected waiting time for the pattern HH in a sequence of fair coin flips. This expected waiting time was shown in Chapter 9 to be 6, in agreement with the result of (a).

Stationary distribution

5. Ⓢ Consider the Markov chain shown below, where $0 < p < 1$ and the labels on the arrows indicate transition probabilities.

(a) Write down the transition matrix Q for this chain.

(b) Find the stationary distribution of the chain.
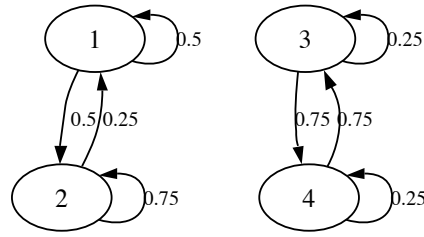
(c) What is the limit of Q^n as $n \to \infty$?

Solution:

(a) The transition matrix is

$$Q = \begin{pmatrix} p & 1-p \\ 1-p & p \end{pmatrix}$$

(b) Because Q is symmetric, the stationary distribution for the chain is the uniform distribution $(1/2, 1/2)$.

(c) The limit of Q^n as $n \to \infty$ is the matrix with the limit distribution $(1/2, 1/2)$ as each row, i.e., $\begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}$.

6. ⓢ Consider the Markov chain shown below, with state space $\{1, 2, 3, 4\}$ and the labels on the arrows indicate transition probabilities.

(a) Write down the transition matrix Q for this chain.

(b) Which states (if any) are recurrent? Which states (if any) are transient?

(c) Find two different stationary distributions for the chain.

Solution:

(a) The transition matrix is

$$
Q = \begin{pmatrix}
0.5 & 0.5 & 0 & 0 \\
0.25 & 0.75 & 0 & 0 \\
0 & 0 & 0.25 & 0.75 \\
0 & 0 & 0.75 & 0.25
\end{pmatrix}
$$

(b) All of the states are recurrent. Starting at state 1, the chain will go back and forth between states 1 and 2 forever (sometimes lingering for a while). Similarly, for any starting state, the probability is 1 of returning to that state.

(c) Solving

$$
(a \quad b) \begin{pmatrix} 0.5 & 0.5 \\ 0.25 & 0.75 \end{pmatrix} = (a \quad b)
$$

$$
(c \quad d) \begin{pmatrix} 0.25 & 0.75 \\ 0.75 & 0.25 \end{pmatrix} = (c \quad d)
$$

shows that $(a, b) = (1/3, 2/3)$, and $(c, d) = (1/2, 1/2)$ are stationary distributions on the 1, 2 chain and on the 3, 4 chain respectively, viewed as separate chains. It follows that $(1/3, 2/3, 0, 0)$ and $(0, 0, 1/2, 1/2)$ are both stationary for Q (as is any mixture $p(1/3, 2/3, 0, 0) + (1 - p)(0, 0, 1/2, 1/2)$ with $0 \leq p \leq 1$).

7. ⓢ Daenerys has three dragons: Drogon, Rhaegal, and Viserion. Each dragon independently explores the world in search of tasty morsels. Let X_n, Y_n, Z_n be the locations at time n of Drogon, Rhaegal, Viserion respectively, where time is assumed to be discrete and the number of possible locations is a finite number M. Their paths $X_0, X_1, X_2, \ldots$; $Y_0, Y_1, Y_2, \ldots$; and $Z_0, Z_1, Z_2, \ldots$ are independent Markov chains with the same stationary distribution **s**. Each dragon starts out at a random location generated according to the stationary distribution.

(a) Let state 0 be home (so s_0 is the stationary probability of the home state). Find the expected number of times that Drogon is at home, up to time 24, i.e., the expected number of how many of $X_0, X_1, \ldots, X_{24}$ are state 0 (in terms of s_0).

(b) If we want to track all 3 dragons simultaneously, we need to consider the vector of positions, (X_n, Y_n, Z_n). There are M^3 possible values for this vector; assume that each is assigned a number from 1 to M^3, e.g., if $M = 2$ we could encode the states

$(0, 0, 0), (0, 0, 1), (0, 1, 0), \ldots, (1, 1, 1)$ as $1, 2, 3, \ldots, 8$ respectively. Let W_n be the number between 1 and M^3 representing (X_n, Y_n, Z_n). Determine whether $W_0, W_1, \ldots$ is a Markov chain.

(c) Given that all 3 dragons start at home at time 0, find the expected time it will take for all 3 to be at home again at the same time.

Solution:

(a) By definition of stationarity, at each time Drogon has probability s_0 of being at home. By linearity, the desired expected value is $25s_0$.

(b) Yes, $W_0, W_1, \ldots$ is a Markov chain, since given the entire past history of the X, Y, and Z chains, only the most recent information about the whereabouts of the dragons should be used in predicting their vector of locations. To show this algebraically, let A_n be the event $\{X_0 = x_0, \ldots, X_n = x_n\}$, B_n be the event $\{Y_0 = y_0, \ldots, Y_n = y_n\}$, C_n be the event $\{Z_0 = z_0, \ldots, Z_n = z_n\}$, and $D_n = A_n \cap B_n \cap C_n$. Then

$$P(X_{n+1} = x, Y_{n+1} = y, Z_{n+1} = z | D_n)$$

$$= P(X_{n+1} = x | D_n)P(Y_{n+1} = y | X_{n+1} = x, D_n)P(Z_{n+1} = z | X_{n+1} = x, Y_{n+1} = y, D_n)$$

$$= P(X_{n+1} = x | A_n)P(Y_{n+1} = y | B_n)P(Z_{n+1} = z | C_n)$$

$$= P(X_{n+1} = x | X_n = x_n)P(Y_{n+1} = y | Y_n = y_n)P(Z_{n+1} = z | Z_n = z_n).$$

(c) The stationary probability for the W-chain of the state with Drogon, Rhaegal, Viserion being at locations x, y, z is $s_x s_y s_z$, since if (X_n, Y_n, Z_n) is drawn from this distribution, then marginally each dragon's location is distributed according to its stationary distribution, so

$$P(X_{n+1} = x, Y_{n+1} = y, Z_{n+1} = z) = P(X_{n+1} = x)P(Y_{n+1} = y)P(Z_{n+1} = z) = s_x s_y s_z.$$

So the expected time for all 3 dragons to be home at the same time, given that they all start at home, is $1/s_0^3$.
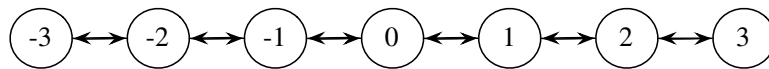
8. Consider the following Markov chain with $52! \approx 8 \times 10^{67}$ states. The states are the possible orderings of a standard 52-card deck. To run one step of the chain, pick 2 different cards from the deck, with all pairs equally likely, and swap the 2 cards. Find the stationary distribution of the chain.

Solution:

The transition probability from state x to state y is $1/\binom{52}{2}$ if x and y differ by a single swap, and 0 otherwise. So the transition matrix is symmetric. Thus, the stationary distribution $\mathbf{s}$ is uniform over the 52! states, i.e., $\mathbf{s} = \frac{1}{52!}(1, 1, \ldots, 1)$.

Reversibility

9. ⓢ A Markov chain $X_0, X_1, \ldots$ with state space $\{-3, -2, -1, 0, 1, 2, 3\}$ proceeds as follows. The chain starts at $X_0 = 0$. If X_n is not an endpoint (-3 or 3), then X_{n+1} is $X_n - 1$ or $X_n + 1$, each with probability $1/2$. Otherwise, the chain gets reflected off the endpoint, i.e., from 3 it always goes to 2 and from -3 it always goes to -2. A diagram of the chain is shown below.

(a) Is $|X_0|, |X_1|, |X_2|, \ldots$ also a Markov chain? Explain.

Hint: For both (a) and (b), think about whether the past and future are conditionally independent given the present; don't do calculations with a 7 by 7 transition matrix!

(b) Let sgn be the sign function: $\text{sgn}(x) = 1$ if $x > 0$, $\text{sgn}(x) = -1$ if $x < 0$, and $\text{sgn}(0) = 0$. Is $\text{sgn}(X_0), \text{sgn}(X_1), \text{sgn}(X_2), \ldots$ a Markov chain? Explain.

(c) Find the stationary distribution of the chain $X_0, X_1, X_2, \ldots$.

(d) Find a simple way to modify some of the transition probabilities q_{ij} for $i \in \{-3, 3\}$ to make the stationary distribution of the modified chain uniform over the states.

Solution:

(a) Yes, $|X_0|, |X_1|, |X_2|, \ldots$ is also a Markov Chain. It can be viewed as the chain on state space $0, 1, 2, 3$ that moves left or right with equal probability, except that at 0 it bounces back to 1 and at 3 it bounces back to 2. Given that $|X_n| = k$, we know that $X_n = k$ or $X_n = -k$, and being given information about $X_{n-1}, X_{n-2}, \ldots$ does not affect the conditional distribution of $|X_{n+1}|$.
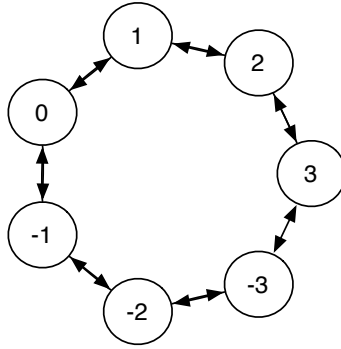
(b) No, this is not a Markov chain because knowing that the chain was at 0 recently affects how far the chain can be from the origin. For example,

$$P(\text{sgn}(X_2) = 1|\text{sgn}(X_1) = 1) > P(\text{sgn}(X_2) = 1|\text{sgn}(X_1) = 1, \text{sgn}(X_0) = 0)$$

since the conditioning information on the righthand side implies $X_1 = 1$, whereas the conditioning information on the lefthand side says exactly that X_1 is $1, 2$, or 3.

(c) Using the result about the stationary distribution of a random walk on an undirected network, the stationary distribution is proportional to the degree sequence, $(1, 2, 2, 2, 2, 2, 1)$. Thus, the stationary distribution is $\frac{1}{12}(1, 2, 2, 2, 2, 2, 1)$.

(d) The uniform distribution will be the stationary distribution if we modify the transition matrix to make it symmetric. Connecting state -3 to state 3 so that the states are arranged in a circle gives the desired symmetry, as illustrated below.

10. Ⓢ Let G be an undirected network with nodes labeled $1, 2, \ldots, M$ (edges from a node to itself are not allowed), where $M \geq 2$ and random walk on this network is irreducible. Let d_j be the degree of node j for each j. Create a Markov chain on the state space $1, 2, \ldots, M$, with transitions as follows. From state i, generate a proposal j by choosing a uniformly random j such that there is an edge between i and j in G; then go to j with probability $\min(d_i/d_j, 1)$, and stay at i otherwise.

(a) Find the transition probability q_{ij} from i to j for this chain, for all states i, j.

(b) Find the stationary distribution of this chain.

Solution:

(a) First let $i \neq j$. If there is no $\{i, j\}$ edge, then $q_{ij} = 0$. If there is an $\{i, j\}$ edge, then
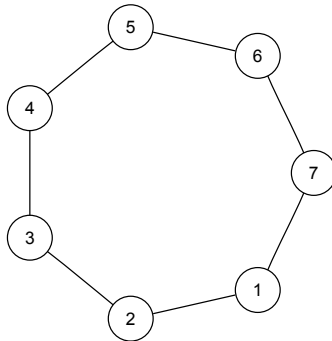
$$q_{ij} = (1/d_i) \min(d_i/d_j, 1) = \begin{cases} 1/d_i \text{ if } d_i \geq d_j, \\ 1/d_j \text{ if } d_i < d_j \end{cases},$$

since the proposal to go to j must be made and then accepted. For $i = j$, we have $q_{ii} = 1 - \sum_{j \neq i} q_{ij}$ since each row of the transition matrix must sum to 1.

(b) Note that $q_{ij} = q_{ji}$ for all states i, j. This is clearly true if $i = j$ or $q_{ij} = 0$, so assume $i \neq j$ and $q_{ij} > 0$. If $d_i \geq d_j$, then $q_{ij} = 1/d_i$ and $q_{ji} = (1/d_j)(d_j/d_i) = 1/d_i$, while if $d_i < d_j$, then $q_{ij} = (1/d_i)(d_i/d_j) = 1/d_j$ and $q_{ji} = 1/d_j$.

Thus, the chain is reversible with respect to the uniform distribution over the states, and the stationary distribution is uniform over the states, i.e., state j has stationary probability $1/M$ for all j. (This is an example of the *Metropolis algorithm*, a Monte Carlo method explored in Chapter 12.)

11. Ⓢ (a) Consider a Markov chain on the state space $\{1, 2, \ldots, 7\}$ with the states arranged in a "circle" as shown below, and transitions given by moving one step clockwise or counterclockwise with equal probabilities. For example, from state 6, the chain moves to state 7 or state 5 with probability $1/2$ each; from state 7, the chain moves to state 1 or state 6 with probability $1/2$ each. The chain starts at state 1.

Find the stationary distribution of this chain.

(b) Consider a new chain obtained by "unfolding the circle". Now the states are arranged as shown below. From state 1 the chain always goes to state 2, and from state 7 the chain always goes to state 6. Find the new stationary distribution.

Solution:

(a) The symmetry of the chain suggests that the stationary distribution should be uniform over all the states. To verify this, note that the reversibility condition is satisfied. So the stationary distribution is $(1/7, 1/7, \ldots, 1/7)$.

(b) By the result about random walk on an undirected network, the stationary probabilities are proportional to the degrees. So we just need to normalize $(1, 2, 2, 2, 2, 2, 1)$, obtaining $(1/12, 1/6, 1/6, 1/6, 1/6, 1/6, 1/12)$.

12. Ⓢ Let X_n be the price of a certain stock at the start of the nth day, and assume that $X_0, X_1, X_2, \ldots$ follows a Markov chain with transition matrix Q. (Assume for simplicity that the stock price can never go below 0 or above a certain upper bound, and that it is always rounded to the nearest dollar.)

(a) A lazy investor only looks at the stock once a year, observing the values on days $0, 365, 2 \cdot 365, 3 \cdot 365, \ldots$. So the investor observes $Y_0, Y_1, \ldots$, where Y_n is the price after n years (which is $365n$ days; you can ignore leap years). Is $Y_0, Y_1, \ldots$ also a Markov chain? Explain why or why not; if so, what is its transition matrix?

(b) The stock price is always an integer between \$0 and \$28. From each day to the next, the stock goes up or down by \$1 or \$2, all with equal probabilities (except for days when the stock is at or near a boundary, i.e., at \$0, \$1, \$27, or \$28).

If the stock is at \$0, it goes up to \$1 or \$2 on the next day (after receiving government bailout money). If the stock is at \$28, it goes down to \$27 or \$26 the next day. If the stock is at \$1, it either goes up to \$2 or \$3, or down to \$0 (with equal probabilities); similarly, if the stock is at \$27 it either goes up to \$28, or down to \$26 or \$25. Find the stationary distribution of the chain.
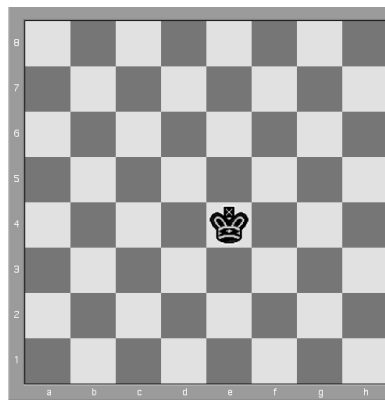
Solution:

(a) Yes, it is a Markov chain: given the whole past history $Y_0, Y_1, \ldots, Y_n$, only the most recent information Y_n matters for predicting Y_{n+1}, because $X_0, X_1, \ldots$ is Markov. The transition matrix of $Y_0, Y_1, \ldots$ is Q^{365}, since the kth power of Q gives the k-step transition probabilities.

(b) This is an example of random walk on an undirected network, so we know the stationary probability of each node is proportional to its degree. The degrees are $(2, 3, 4, 4, \ldots, 4, 4, 3, 2)$, where there are $29 - 4 = 25$ 4's. The sum of these degrees is 110 (coincidentally?). Thus, the stationary distribution is

$$\left(\frac{2}{110}, \frac{3}{110}, \frac{4}{110}, \frac{4}{110}, \ldots, \frac{4}{110}, \frac{4}{110}, \frac{3}{110}, \frac{2}{110} \right),$$

with 25 $\frac{4}{110}$'s.

13. Ⓢ In chess, the king can move one square at a time in any direction (horizontally, vertically, or diagonally).

For example, in the diagram, from the current position the king can move to any of 8 possible squares. A king is wandering around on an otherwise empty 8×8 chessboard, where for each move all possibilities are equally likely. Find the stationary distribution of this chain (of course, don't list out a vector of length 64 explicitly! Classify the 64 squares into types and say what the stationary probability is for a square of each type).

Solution: There are 4 corner squares, 24 edge squares, and 36 normal squares, where by "edge" we mean a square in the first or last row or column, excluding the 4 corners, and by "normal" we mean a square that's not on the edge or in a corner. View the chessboard as an undirected network, where there is an edge between two squares if the king can walk from one to the other in one step.

The stationary probabilities are proportional to the degrees. Each corner square has degree 3, each edge square has degree 5, and each normal square has degree 8. The total degree is $420 = 3 \cdot 4 + 24 \cdot 5 + 36 \cdot 8$ (which is also twice the number of edges in the network). Thus, the stationary probability is $\frac{3}{420}$ for a corner square, $\frac{5}{420}$ for an edge square, and $\frac{8}{420}$ for a normal square.

14. A chess piece is wandering around on an otherwise vacant 8×8 chessboard. At each move, the piece (a king, queen, rook, bishop, or knight) chooses uniformly at random where to go, among the legal choices (according to the rules of chess, which you should look up if you are unfamiliar with them).

    (a) For each of these cases, determine whether the Markov chain is irreducible, and whether it is aperiodic.

    Hint for the knight: Note that a knight's move always goes from a light square to a dark square or vice versa. A *knight's tour* is a sequence of knight moves on a chessboard such that the knight visits each square exactly once. Many knight's tours exist.

    (b) Suppose for this part that the piece is a rook, with initial position chosen uniformly at random. Find the distribution of where the rook is after n moves.

    (c) Now suppose that the piece is a king, with initial position chosen deterministically to be the upper left corner square. Determine the expected number of moves it takes him to return to that square, fully simplified, preferably in at most 140 characters.

    (d) The stationary distribution for the random walk of the king from the previous part is not uniform over the 64 squares of the chessboard. A recipe for modifying the chain to obtain a uniform stationary distribution is as follows. Label the squares as $1, 2, \ldots, 64$, and let d_i be the number of legal moves from square i. Suppose the king is currently at square i. The next move of the chain is determined as follows:

    Step 1: Generate a *proposal square* j by picking uniformly at random among the legal moves from i.

    Step 2: Flip a coin with probability $\min(d_i/d_j, 1)$ of Heads. If the coin lands Heads, go to j. Otherwise, stay at i.

    Show that this modified chain has a stationary distribution that is uniform over the 64 squares.

    Solution:

    (a) For the bishop, the Markov chain is reducible since a bishop can't move from a light square to a dark square or from a dark square to a light square. For all of the other pieces, the chains are irreducible. In fact, a queen or rook can get from anywhere to anywhere in at most 2 moves (go to the desired row and then go to the desired column). The king can get from anywhere to anywhere in at most 7 moves (simply by walking there, one step at a time). By the hint, the knight can get from anywhere to anywhere by hopping around (it turns out that the knight can get from anywhere to anywhere in at most 6 moves).

    For the knight, each state has period 2 since each move goes from a light square to a dark square or vice versa, so an even number of moves is needed to return from any state i to state i. So the knight's walk is periodic. The walks of the other pieces are

aperiodic since from any square, there are paths of length 2 and 3 from the square to itself (so the greatest common divisor of these path lengths is 1).

(b) We can view this Markov chain as a random walk on an undirected network, with one node for each square and an edge between two nodes if they are one Rook move apart. Each node has degree 14 since from any square, the Rook has 7 choices for a move along the same row and 7 choices for a move along the same column. So the stationary distribution is uniform. Since the chain starts out according to the stationary distribution, by definition of "stationary" it will stay in the same distribution forever. So the distribution after n moves is uniform over the 64 squares.
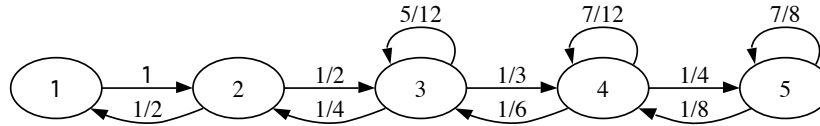
(c) The stationary probability of that corner is its degree, 3, over the total degree, $4 \cdot 3 + 24 \cdot 5 + 36 \cdot 8$. So the mean return time is $420/3 = 140$.

(d) Let q_{ij} be the transition probability of going from i to j for the modified chain. It suffices to show that $q_{ij} = q_{ji}$ for all i, j, since this says that the chain is reversible with respect to the uniform distribution.

For $i = j$, this just says $q_{ii} = q_{ii}$, so let $i \neq j$. We have $q_{ij} = 0$ if and only if $q_{ji} = 0$, so assume $q_{ij} > 0$. Then $q_{ij} = \frac{1}{d_i} \min(d_i/d_j, 1), q_{ji} = \frac{1}{d_j} \min(d_j/d_i, 1)$. If $d_i < d_j$, then $q_{ij} = (1/d_i)(d_i/d_j) = 1/d_j = q_{ji}$. If $d_i \geq d_j$, then $q_{ij} = 1/d_i = (1/d_j)(d_j/d_i) = q_{ji}$. Thus, $q_{ij} = q_{ji}$ for all i, j, which shows that the stationary distribution is uniform.

This algorithm is a special case of the *Metropolis-Hastings algorithm* (see Chapter 12).

15. Ⓢ Find the stationary distribution of the Markov chain shown below, *without using matrices*. The number above each arrow is the corresponding transition probability.

Solution: We will show that this chain is reversible by solving for **s** (which will work out nicely since this is a birth-death chain). Let q_{ij} be the transition probability from i to j, and solve for **s** in terms of s_1. Noting that $q_{ij} = 2q_{ji}$ for $j = i + 1$ (when $1 \leq i \leq 4$), we have that

$$s_1 q_{12} = s_2 q_{21} \text{ gives } s_2 = 2s_1.$$
$$s_2 q_{23} = s_3 q_{32} \text{ gives } s_3 = 2s_2 = 4s_1.$$
$$s_3 q_{34} = s_4 q_{43} \text{ gives } s_4 = 2s_3 = 8s_1.$$
$$s_4 q_{45} = s_5 q_{54} \text{ gives } s_5 = 2s_4 = 16s_1.$$

The other reversibility equations are automatically satisfied since here $q_{ij} = 0$ unless $|i - j| \leq 1$. Normalizing, the stationary distribution is

$$\left(\frac{1}{31}, \frac{2}{31}, \frac{4}{31}, \frac{8}{31}, \frac{16}{31} \right).$$

Sanity check: This chain "likes" going from left to right more than from right to left, so the stationary probabilities should be increasing from left to right. We also know that $s_j = \sum_i s_i q_{ij}$ (since if the chain is in the stationary distribution at time n, then it is also in the stationary distribution at time $n + 1$), so we can check, for example, that $s_1 = \sum_i s_i q_{i1} = \frac{1}{2} s_2$.

16. There are two urns with a total of $2N$ distinguishable balls. Initially, the first urn has N white balls and the second urn has N black balls. At each stage, we pick a ball at random from each urn and interchange them. Let X_n be the number of black balls in the first urn at time n. This is a Markov chain on the state space $\{0, 1, \ldots, N\}$.

(a) Give the transition probabilities of the chain.

(b) Show that $(s_0, s_1, \ldots, s_N)$ where

$$s_i = \frac{\binom{N}{i}\binom{N}{N-i}}{\binom{2N}{N}}$$

is the stationary distribution, by verifying the reversibility condition.

Solution:

(a) Let p_{ij} be the transition probability from i to j. The number of black balls changes by at most 1 at each step, so $p_{ij} = 0$ for $|i - j| > 1$. Note that if urn 1 has i black balls and $N - i$ white balls, then urn 2 has i white balls and $N - i$ black balls. So

$$p_{i,i+1} = \left(\frac{N - i}{N}\right)^2,$$

since to get from state i to state $i + 1$ we need to swap a white ball from urn 1 with a black ball from urn 2. Similarly,

$$p_{i,i-1} = \left(\frac{i}{N}\right)^2.$$

For the number of black balls to stay the same, we need to choose two black balls or two white balls. So

$$p_{i,i} = \frac{2i(N - i)}{N^2}.$$

As a check, note that

$$\left(\frac{N - i}{N}\right)^2 + \left(\frac{i}{N}\right)^2 + \frac{2i(N - i)}{N^2} = \frac{(N - i + i)^2}{N^2} = 1.$$

This Markov chain is called the *Bernoulli-Laplace chain*.

(b) We need to show that $s_i p_{ij} = s_j p_{ji}$ for all states i and j. If $|i - j| > 1$, then both sides are 0. If $i = j$, then both sides are $s_i p_{ii}$. So we just need to verify that both sides are equal for the cases $j = i + 1$ and $j = i - 1$.

Case 1: $j = i + 1$. In this case,

$$s_i p_{ij} = \frac{\binom{N}{i}\binom{N}{N-i}}{\binom{2N}{N}}\left(\frac{N - i}{N}\right)^2 = \frac{\binom{N}{i}^2(N - i)^2}{\binom{2N}{N}N^2}$$

and

$$s_j p_{ji} = \frac{\binom{N}{i+1}\binom{N}{N-(i+1)}}{\binom{2N}{N}}\left(\frac{i + 1}{N}\right)^2 = \frac{\binom{N}{i+1}^2(i + 1)^2}{\binom{2N}{N}N^2}.$$

The denominators are the same, so it suffices to show that

$$\binom{N}{i}(N - i) = \binom{N}{i+1}(i + 1).$$

But this is true, as can be seen algebraically or via the following story: from N people, we wish to select a team of size $i + 1$ and designate one team member as captain. The left-hand side corresponds to choosing the i team members other than the captain,

followed by choosing the captain from the remaining $N - i$ people. The right-hand side corresponds to choosing the $i + 1$ team members and then choosing one of them to be captain. Thus, $s_i p_{ij} = s_j p_{ji}$.

Case 2: $j = i - 1$. In this case,

$$s_i p_{ij} = \frac{\binom{N}{i}\binom{N}{N-i}}{\binom{2N}{N}}\left(\frac{i}{N}\right)^2 = \frac{\binom{N}{i}^2 i^2}{\binom{2N}{N}N^2}$$

and

$$s_j p_{ji} = \frac{\binom{N}{i-1}\binom{N}{N-(i-1)}}{\binom{2N}{N}}\left(\frac{N-(i-1)}{N}\right)^2 = \frac{\binom{N}{i-1}^2(N-(i-1))^2}{\binom{2N}{N}N^2}.$$

Again the denominators are the same, so it suffices to show that

$$\binom{N}{i}i = \binom{N}{i-1}(N-(i-1)).$$

But this is true by the same story as for Case 1, except with a team of size i rather than a team of size $i + 1$. So again we have $s_i p_{ij} = s_j p_{ji}$.

Hence, the chain is reversible, with stationary distribution $(s_0, s_1, \ldots, s_N)$.

17. Find the stationary distribution of a Markov chain $X_0, X_1, X_2, \ldots$ on the state space $\{0, 1, \ldots, 110\}$, with transition probabilities given by

$$P(X_{n+1} = j | X_n = 0) = p, \text{ for } j = 1, 2, \ldots, 110;$$

$$P(X_{n+1} = 0 | X_n = 0) = 1 - 110p;$$

$$P(X_{n+1} = j | X_n = j) = 1 - r, \text{ for } j = 1, 2, \ldots, 110;$$

$$P(X_{n+1} = 0 | X_n = j) = r, \text{ for } j = 1, 2, \ldots, 110,$$

where p and r are constants with $0 < p < \frac{1}{110}$ and $0 < r < 1$.

Solution:

Let q_{ij} be the transition probability from i to j. We will solve for a probability vector $\mathbf{s} = (s_0, s_1, \ldots, s_{110})$ satisfying the reversibility equation $s_i q_{ij} = s_j q_{ji}$ for all i, j. This is automatically satisfied if $i = j$. For $i \neq j$, it reduces to $0 = 0$ unless $i = 0$ or $j = 0$. For $j = 1, 2, \ldots, 110$, the equation $s_0 q_{0j} = s_j q_{j0}$ is equivalent to

$$s_j = s_0 p/r.$$

We then have $\mathbf{s}$ in terms of s_0, and can solve for s_0 by normalizing: setting

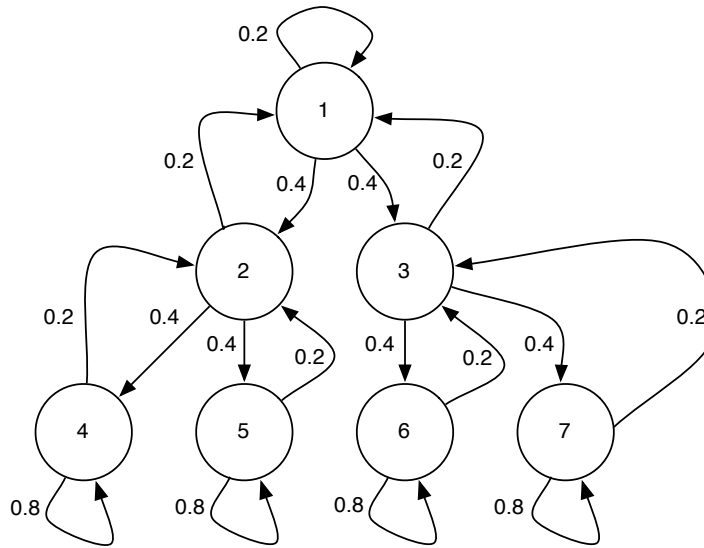$$s_0 + 110 s_0 p/r = 1,$$

we have

$$s_0 = \frac{r}{r + 110p}, \text{ and } s_j = \frac{p}{r + 110p} \text{ for } j = 1, 2, \ldots, 110.$$

Then $\mathbf{s}$ satisfies the reversibility equations, so it is the stationary distribution.

18. Determine whether the Markov chain shown below is reversible, and find the stationary distribution of the chain. The label to the left of an arrow gives the corresponding transition probability.

Solution:

Let q_{ij} be the transition probability from i to j. Let's try to solve for $\mathbf{s}$ such that $s_i q_{ij} = s_j q_{ji}$ for all i, j.

- $s_1 q_{12} = s_2 q_{21}$ gives $s_2 = s_1 q_{12}/q_{21} = 2s_1$. Similarly, $s_3 = 2s_1$.
- $s_2 q_{24} = s_4 q_{42}$ gives $s_4 = 2s_2 = 4s_1$. Similarly, $s_5 = s_6 = s_7 = 4s_1$.

It follows that

$$s_i q_{ij} = s_j q_{ji}$$

for all i, j, so the chain is reversible. Normalizing, the stationary distribution is

$$\mathbf{s} = \frac{1}{21}(1, 2, 2, 4, 4, 4, 4).$$

19. Nausicaa Distribution sells distribution plushies on Etsy. They have two different photos of the Evil Cauchy plushie but do not know which is more effective in getting a customer to purchase an Evil Cauchy plushie. Each visitor to their website is shown one of the two photos (call them Photo A and Photo B), and then the visitor either does buy an Evil Cauchy ("success") or does not buy one ("failure").

Let a and b be the probabilities of success when Photo A is shown and when Photo B is shown, respectively. Even though the Evil Cauchy is irresistible, suppose that $0 < a < 1$ and $0 < b < 1$. Suppose that the following strategy is followed (note that the strategy can be followed without knowing a and b). Show the first visitor Photo A. If that visitor buys an Evil Cauchy, continue with Photo A for the next visitor; otherwise, switch to Photo B. Similarly, if the nth visitor is a "success" then show the $(n + 1)$st visitor the same photo, and otherwise switch to the other photo.

(a) Show how to represent the resulting process as a Markov chain, drawing a diagram and giving the transition matrix. The states are A1, B1, A0, B0 (use this order for the transition matrix and stationary distribution), where, for example, being at state A1 means that the current visitor was shown Photo A and was a success.

(b) Determine whether this chain is reversible.

Hint: First think about which transition probabilities are zero and which are nonzero.

(c) Show that the stationary distribution is proportional to $\left(\frac{a}{1-a}, \frac{b}{1-b}, 1, 1\right)$, and find the stationary distribution.
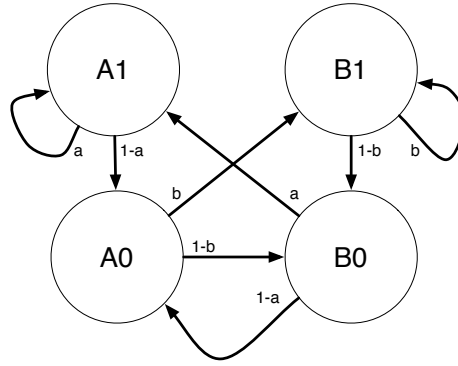
(d) Show that for $a \neq b$, the stationary probability of success for each visitor is strictly better than the success probability that would be obtained by independently, randomly choosing (with equal probabilities) which photo to show to each visitor.

Solution:

(a) The process is a Markov chain since the choice of which photo to show the next visitor depends only on the current state (which is a record of which photo the current visitor was shown, and whether this visitor was a success). The information about which photo the next visitor is shown and whether that visitor is a success is then recorded, determining the next state, etc. The transition matrix is

$$Q = \begin{pmatrix} a & 0 & 1-a & 0 \\ 0 & b & 0 & 1-b \\ 0 & b & 0 & 1-b \\ a & 0 & 1-a & 0 \end{pmatrix}.$$

A diagram of this Markov chain is shown below.

(b) This chain is *not* reversible, since, e.g., the transition probability from A0 to B1 is nonzero but that from B1 to A0 is zero; this is incompatible with having the reversibility condition $s_i q_{ij} = s_j q_{ji}$ hold for all i, j (the stationary probabilities are all positive since the chain is irreducible).

(c) To find the stationary distribution, we can solve the system $\mathbf{v}Q = \mathbf{v}$, where $\mathbf{v}$ is a row vector of length 4, choose a solution with all positive components, and then normalize. Or we can check that the given vector is as desired:

$$\begin{pmatrix} a & 0 & 0 & a \\ 0 & b & b & 0 \\ 1-a & 0 & 0 & 1-a \\ 0 & 1-b & 1-b & 0 \end{pmatrix} \begin{pmatrix} \frac{a}{1-a} \\ \frac{b}{1-b} \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{a^2}{1-a} + a \\ \frac{b^2}{1-b} + b \\ a + 1 - a \\ b + 1 - b \end{pmatrix} = \begin{pmatrix} \frac{a}{1-a} \\ \frac{b}{1-b} \\ 1 \\ 1 \end{pmatrix}.$$

Thus, the stationary distribution is proportional to $\left(\frac{a}{1-a}, \frac{b}{1-b}, 1, 1 \right)$.

Defining $\tilde{a} = 1-a, \tilde{b} = 1-b$, the sum of the components of the above vector is $(\tilde{a}+\tilde{b})/(\tilde{a}\tilde{b})$, so the stationary distribution is $\frac{1}{\tilde{a}+\tilde{b}}(a\tilde{b}, \tilde{a}b, \tilde{a}\tilde{b}, \tilde{a}\tilde{b})$.

(d) Let $a \neq b$. The stationary probability of success is

$$\frac{a\tilde{b} + \tilde{a}b}{\tilde{a} + \tilde{b}} = \frac{a + b - 2ab}{2 - (a + b)}.$$

The success probability for the independent random choices is $\frac{a+b}{2}$. The desired inequality simplifies to the equivalent inequality $a^2 + b^2 - 2ab > 0$, and this inequality is true since $a^2 + b^2 - 2ab = (a - b)^2 > 0$.

20. This exercise considers random walk on a *weighted* undirected network. Suppose that an undirected network is given, where each edge (i, j) has a nonnegative weight w_{ij} assigned to it (we allow $i = j$ as a possibility). We assume that $w_{ij} = w_{ji}$ since the edge from i to j is considered the same as the edge from j to i. To simplify notation, define $w_{ij} = 0$ whenever (i, j) is not an edge.

When at node i, the next step is determined by choosing an edge attached to i with probabilities proportional to the weights. For example, if the walk is at node 1 and there are 3 possible edges coming out from node 1, with weights $7, 1, 4$, then the first of these 3 edges is traversed with probability $7/12$, the second is traversed with probability $1/12$, and the third is traversed with probability $4/12$. If all the weights equal 1, then the process reduces to the kind of random walk on a network that we studied earlier.

(a) Let $v_i = \sum_j w_{ij}$ for all nodes i. Show that the stationary distribution of node i is proportional to v_i.

(b) Show that *every* reversible Markov chain can be represented as a random walk on a weighted undirected network.

Hint: Let $w_{ij} = s_i q_{ij}$, where **s** is the stationary distribution and q_{ij} is the transition probability from i to j.

Solution:

(a) Let p_{ij} be the transition probability from i to j. Let $c = \sum_i v_i$. It suffices to show that $v_i p_{ij} = v_j p_{ji}$ for all nodes i and j, since then the reversibility condition $s_i p_{ij} = s_j p_{ji}$ holds for $s_i = v_i/c$. If $w_{ij} = 0$, then both sides are 0. So assume $w_{ij} > 0$. Then

$$v_i p_{ij} = v_i \frac{w_{ij}}{\sum_k w_{ik}} = v_i \frac{w_{ij}}{v_i} = w_{ij} = w_{ji} = v_j \frac{w_{ji}}{v_j} = v_j p_{ji}.$$

(b) Consider a reversible Markov chain with state space $\{1, 2, \ldots, M\}$, transition probability q_{ij} from i to j, and stationary distribution **s**. As in the hint, let $w_{ij} = s_i q_{ij}$. We have $w_{ij} = w_{ji}$ because the chain is reversible.

Now consider a weighted undirected network with nodes $1, 2, \ldots, M$ and an edge (i, j) with weight w_{ij} for each (i, j) with $w_{ij} > 0$ (and no edge (i, j) if $w_{ij} = 0$). Then random walk on this weighted undirected network has transition probability q_{ij} from i to j, since the probability of going from i to j in one step is

$$\frac{w_{ij}}{\sum_k w_{ik}} = \frac{s_i q_{ij}}{\sum_k s_i q_{ik}} = \frac{q_{ij}}{\sum_k q_{ik}} = q_{ij}.$$

Mixed practice

21. ⑤ A cat and a mouse move independently back and forth between two rooms. At each time step, the cat moves from the current room to the other room with probability 0.8. Starting from room 1, the mouse moves to room 2 with probability 0.3 (and remains otherwise). Starting from room 2, the mouse moves to room 1 with probability 0.6 (and remains otherwise).

(a) Find the stationary distributions of the cat chain and of the mouse chain.

(b) Note that there are 4 possible (cat, mouse) states: both in room 1, cat in room 1 and mouse in room 2, cat in room 2 and mouse in room 1, and both in room 2. Number these cases $1, 2, 3, 4$, respectively, and let Z_n be the number representing the (cat, mouse) state at time n. Is $Z_0, Z_1, Z_2, \ldots$ a Markov chain?

(c) Now suppose that the cat will eat the mouse if they are in the same room. We wish to know the expected time (number of steps taken) until the cat eats the mouse for two initial configurations: when the cat starts in room 1 and the mouse starts in room 2,

and vice versa. Set up a system of two linear equations in two unknowns whose solution is the desired values.

Solution:

(a) The cat chain has transition matrix

$$Q_{\text{cat}} = \begin{pmatrix} \frac{2}{10} & \frac{8}{10} \\ \frac{8}{10} & \frac{2}{10} \end{pmatrix}.$$

The uniform distribution $(\frac{1}{2}, \frac{1}{2})$ is stationary since the transition matrix is symmetric.

The mouse chain has transition matrix

$$Q_{\text{mouse}} = \begin{pmatrix} \frac{7}{10} & \frac{3}{10} \\ \frac{6}{10} & \frac{4}{10} \end{pmatrix}.$$

The stationary distribution is proportional to (x, y) with $7x + 6y = 10x, 3x + 4y = 10y$. This reduces to $x = 2y$. So the stationary distribution is $(\frac{2}{3}, \frac{1}{3})$.

(b) Yes, it is a Markov chain. Given the current (cat, mouse) state, the past history of where the cat and mouse were previously are irrelevant for computing the probabilities of what the next state will be.

(c) Let a and b be the expected values for the two initial configurations, respectively. Conditioning on the first move of the cat and the first move of the mouse, we have

$$a = \underbrace{(0.2)(0.6)}_{\text{both in room 1}} + \underbrace{(0.8)(0.4)}_{\text{both in room 2}} + \underbrace{(0.2)(0.4)(1 + a)}_{\text{cat in room 1, mouse in room 2}} + \underbrace{(0.8)(0.6)(1 + b)}_{\text{cat in room 2, mouse in room 1}},$$

$$b = \underbrace{(0.8)(0.7)}_{\text{both in room 1}} + \underbrace{(0.2)(0.3)}_{\text{both in room 2}} + \underbrace{(0.8)(0.3)(1 + a)}_{\text{cat in room 1, mouse in room 2}} + \underbrace{(0.2)(0.7)(1 + b)}_{\text{cat in room 2, mouse in room 1}}.$$

(The solution to this system works out to $a = 335/169, b = 290/169$.)

22. (a) Alice and Bob are wandering around randomly, independently of each other, in a house with M rooms, labeled $1, 2, \ldots, M$. Let d_i be the number of doors in room i (leading to other rooms, not leading outside). At each step, Alice moves to another room by choosing randomly which door to go through (with equal probabilities). Bob does the same, independently. The Markov chain they each follow is irreducible and aperiodic. Let A_n and B_n be Alice's room and Bob's room at time n, respectively, for $n = 0, 1, 2, \ldots$.

Find $\lim_{n \to \infty} P(A_n = i, B_n = j)$.

(b) With setup as in (a), let p_{ij} be the transition probability for going from room i to room j. Let t_{ik} be the expected first time at which Alice and Bob are in the same room, if Alice starts in room i and Bob starts in room k (note that $t_{ik} = 0$ for $i = k$).

Provide a system of linear equations, the solution of which would yield t_{ik} for all rooms i, k.

Hint: Condition on Alice's first move and Bob's first move.

Solution:

(a) Alice is following a random walk on an undirected network, where room i has degree d_i. So the stationary distribution for her chain is $(d_1, d_2, \ldots, d_M)/\sum_k d_k$. Bob's chain has the same stationary distribution. Therefore,

$$\lim_{n \to \infty} P(A_n = i, B_n = j) = \lim_{n \to \infty} P(A_n = i)P(B_n = j) = \frac{d_i d_j}{\left(\sum_k d_k\right)^2}.$$

(b) Suppose that Alice starts out in room i and Bob starts out in room k, where $i \neq k$. If Alice's first move is to room j and Bob's first move is to room l, then by the Markov property the expected *additional* number of steps until they are in the same room is t_{jm}. So by the law of total expectation,

$$t_{ik} = \sum_{j,l}(1 + t_{jl})p_{ij}p_{kl}, \text{ for } i \neq k.$$

Together with the equations

$$t_{ik} = 0, \text{ for } i = k,$$

we have a system of M^2 linear equations in the M^2 variables t_{jl}.

23. Let $\{X_n\}$ be a Markov chain on states $\{0, 1, 2\}$ with transition matrix
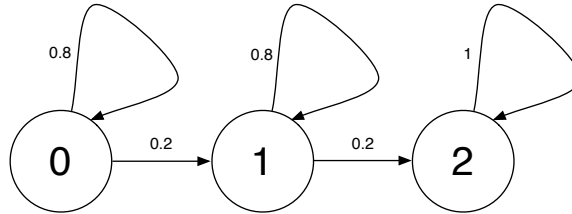
$$\begin{pmatrix} 0.8 & 0.2 & 0 \\ 0 & 0.8 & 0.2 \\ 0 & 0 & 1 \end{pmatrix}$$

The chain starts at $X_0 = 0$. Let T be the time it takes to reach state 2:

$$T = \min\{n : X_n = 2\}.$$

By drawing the Markov chain and telling a story, find $E(T)$ and $\text{Var}(T)$.
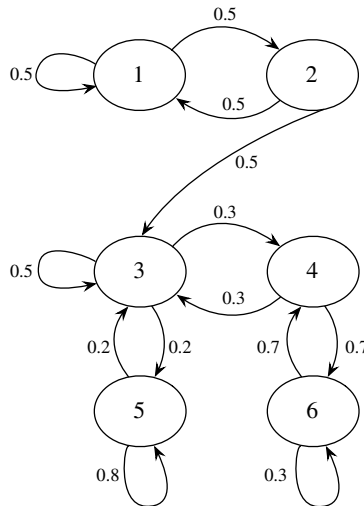
Solution: This Markov chain is illustrated below. To get from state 0 to state 2, the chain needs to get from state 0 to state 1 and then get from state 1 to state 2.

So $T = T_1 + T_2$, where T_1 is the time it takes to reach state 1 and T_2 is the additional time it takes to reach state 2. Then T_1 and T_2 are independent (by the Markov property), with $T_1 \sim \text{FS}(0.2)$ and also $T_2 \sim \text{FS}(0.2)$. So

$$E(T) = 5 + 5 = 10 \text{ and } \text{Var}(T) = 20 + 20 = 40.$$

24. Consider the following Markov chain on the state space $\{1, 2, 3, 4, 5, 6\}$.

(a) Suppose the chain starts at state 1. Find the distribution of the number of times that the chain returns to state 1.

(b) In the long run, what fraction of the time does the chain spend in state 3? Explain briefly.

Solution:

(a) From state 1, the probability of not returning to state 1 is 0.25, since the chain will not return if and only if the next 2 transitions are state 1 to state 2, followed by state 2 to state 3. Considering this sequence of transitions a "success" and returning to state 1 as a "failure", we have that the number of returns is Geom(0.25).

(b) States 1 and 2 are transient, so in the long run the chain will spend all its time in states 3, 4, 5, and 6, never again returning to state 1 or state 2. So we can restrict attention to the 4-state chain with states 3, 4, 5, and 6. The transition matrix for these states is symmetric, so the stationary distribution is uniform on these 4 states. So in the long run, the original chain spends $1/4$ of its time in state 3.

25. Let Q be the transition matrix of a Markov chain on the state space $\{1, 2, \ldots, M\}$, such that state M is an *absorbing state*, i.e., from state M the chain can never leave. Suppose that from any other state, it is possible to reach M (in some number of steps).

(a) Which states are recurrent, and which are transient? Explain.

(b) What is the limit of Q^n as $n \to \infty$?

(c) For $i, j \in \{1, 2, \ldots, M-1\}$, find the probability that the chain is at state j at time n, given that the chain is at state i at time 0 (your answer should be in terms of Q).

(d) For $i, j \in \{1, 2, \ldots, M-1\}$, find the expected number of times that the chain is at state j up to (and including) time n, given that the chain is at state i at time 0 (in terms of Q).

(e) Let R be the $(M-1) \times (M-1)$ matrix obtained from Q by deleting the last row and the last column of Q. Show that the (i, j) entry of $(I - R)^{-1}$ is the expected number of times that the chain is at state j before absorption, given that it starts out at state i.

Hint: We have $I + R + R^2 + \cdots = (I - R)^{-1}$, analogously to a geometric series. Also, if we partition Q as

$$Q = \left(\begin{array}{c|c} R & B \\ \hline 0 & 1 \end{array} \right)$$

where B is a $(M-1) \times 1$ matrix and 0 is the $1 \times (M-1)$ zero matrix, then

$$Q^k = \left(\begin{array}{c|c} R^k & B_k \\ \hline 0 & 1 \end{array} \right)$$

for some $(M-1) \times 1$ matrix B_k.

(a) State M is recurrent since starting there, the chain is not only guaranteed to visit again, but in fact it will stay there forever. The other $M-1$ states are transient since the probability is 1 that the chain will eventually get absorbed at M. To see this, note that there are finitely many states and it is possible to get from anywhere to M, so eventually the chain will wander into M. If one of the other $M-1$ states were recurrent, there would be infinitely many opportunities to go from that state to M, so eventually it *would* go to M. So the other $M-1$ states are transient, and state M is the last refuge of the chain.

(b) The matrix Q^n gives the n-step transition probabilities for the chain. The probability

is 1 that the chain will eventually get absorbed at M, so Q^n converges to the matrix with all 0's except that the last column is all 1's.

(c) Since Q^n gives the n-step transition probabilities, this is the (i,j) entry of Q^n.

(d) Let J_k be the indicator r.v. of the chain being at state j at time k, for $0 \le k \le n$. By linearity, the desired expectation is $E(J_0) + E(J_1) + \cdots + E(J_n)$. By the previous part, this is the (i,j) entry of $I + Q + Q^2 + \cdots + Q^n$.

(e) For $i,j \in \{1, 2, \ldots, M-1\}$ and any $k \ge 0$, the (i,j) entry in Q^k is the same as the (i,j) entry in R^k, since we can partition Q as

$$Q = \left(\begin{array}{c|c} R & B \\ \hline 0 & 1 \end{array} \right)$$

where B is a $(M-1) \times 1$ matrix and the 0 is the $1 \times (M-1)$ zero matrix, and then

$$Q^k = \left(\begin{array}{c|c} R^k & B_k \\ \hline 0 & 1 \end{array} \right)$$

for some $(M-1) \times 1$ matrix B_k. So by the previous part, for a chain starting at i the expected number of times the chain is at state j up to (and including) time n is the (i,j) entry of $I + R + R^2 + \cdots + R^n$. Letting $n \to \infty$, the expected number of times the chain is at state j before absorption is the (i,j) entry of

$$I + R + R^2 + \cdots = (I - R)^{-1}.$$

26. In the game called *Chutes and Ladders*, players try to be first to reach a certain destination on a board. The board is a grid of squares, numbered from 1 to the number of squares. The board has some "chutes" and some "ladders", each of which connects a pair of squares. Here we will consider the one player version of the game (this can be extended to the multi-player version without too much trouble, since with more than one player, the players simply take turns independently until one reaches the destination).

On each turn, the player rolls a fair die, which determines how many squares forward to move on the grid, e.g., if the player is at square 5 and rolls a 3, then they advance to square 8. If the resulting square is the base of a ladder, the player gets to climb the ladder, instantly arriving at a more advanced square. If the resulting square is the top of a chute, the player instantly slides down to the bottom of the chute.
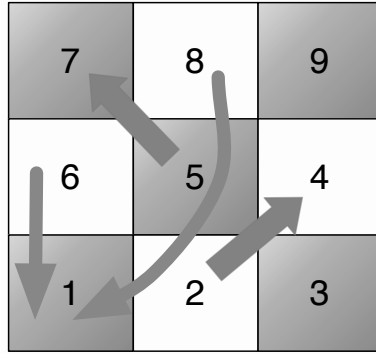
This game can be viewed naturally as a Markov chain: given where the player currently is on the board, the past history does not matter for computing, for example, the probability of winning within the next 3 moves.

Consider a simplified version of Chutes and Ladders, played on the 3×3 board shown below. The player starts out at square 1, and wants to get to square 9. On each move, a fair coin is flipped, and the player gets to advance 1 square if the coin lands Heads and 2 squares if the coin lands Tails. However, there are 2 ladders (shown as upward-pointing arrows) and 2 chutes (shown as downward-pointing arrows) on the board.

(a) Explain why, despite the fact that there are 9 squares, we can represent the game using the 5×5 transition matrix

$$Q = \begin{pmatrix} 0 & 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0 & 0.5 & 0 \\ 0.5 & 0 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

(b) Find the mean and variance for the number of times the player will visit square 7, *without* using matrices or any messy calculations.

The remaining parts of this problem require matrix calculations that are best done on a computer. You can use whatever computing environment you want, but here is some information for how to do it in R. In any case, you should state what environment you used and include your code. To create the transition matrix in R, you can use the following commands:

```
a <- 0.5
Q <- matrix(c(0,0,a,a,0,a,0,0,0,0,a,a,0,0,0,0,a,a,0,0,0,0,0,a,1),nrow=5)
```

Some useful R commands for matrices are in Appendix B.2. In particular, `diag(n)` gives the $n \times n$ identity matrix, `solve(A)` gives the inverse A^{-1}, and `A %*% B` gives the product AB (note that `A*B` does *not* do ordinary matrix multiplication). Matrix powers are not built into R, but you can compute A^k using `A %^% k` after installing and loading the `expm` package.

(c) Find the median duration of the game (defining duration as the number of coin flips).

Hint: Relate the CDF of the duration to powers of Q.

(d) Find the mean duration of the game (with duration defined as above).

Hint: Relate the duration to the total amount of time spent in transient states, and apply Part (e) of the previous problem.

Solution:

(a) If the player reaches the top of a chute or a base of a ladder, then the player is instantly transported to a new square, so we can exclude chute-tops and ladder-bases as states. This leaves 5 states for the Markov chain, after eliminating squares $2, 5, 6$, and 8 as states. The transition matrix is as above, if the states are kept in increasing numerical order: it is

$$
\begin{array}{c@{}c}
 & \begin{array}{ccccc} 1 & 3 & 4 & 7 & 9 \end{array} \\
\begin{array}{c} 1 \\ 3 \\ 4 \\ 7 \\ 9 \end{array} &
\left(\begin{array}{ccccc}
0 & 0.5 & 0.5 & 0 & 0 \\
0 & 0 & 0.5 & 0.5 & 0 \\
0.5 & 0 & 0 & 0.5 & 0 \\
0.5 & 0 & 0 & 0 & 0.5 \\
0 & 0 & 0 & 0 & 1
\end{array} \right),
\end{array}
$$

since, e.g., from square 7 the chain will go to square 1 or square 9 with equal probabilities.

(b) Let N be the number of visits to square 7. There is no way to reach square 9 except via square 7, so $N \geq 1$. Each time the chain is at square 7 we have a Bern(1/2) trial: define "success" to be reaching square 9 on the next move, and "failure" to be going back to square 1. The trials are independent by the Markov property, so N has a

First Success distribution with $p = 1/2$. (Alternatively, apply Proposition 11.2.2.) Thus, $E(N) = 2, \text{Var}(N) = 2$.

(c) Let T be the duration of the game. The $(1, 5)$ entry of Q^n is the probability that the chain is at square 9 after n steps, which is $P(T \leq n)$. Using the R commands given above and computing some powers of Q, we find that

$$((1, 5) \text{ entry of } Q^6) = P(T \leq 6) \approx 0.484,$$

$$((1, 5) \text{ entry of } Q^7) = P(T \leq 7) \approx 0.539,$$

so the median duration is 7 moves.

(d) This Markov chain is as in the previous problem, with square 9 the absorbing state. Using the notation and result of Part (e) of the previous problem, the sum of the top row of $(I - R)^{-1}$ is the expected total amount of time spent in the transient states, which is also the expected duration since 1 coin flip is made for each time at which the chain is in a transient state. Entering

```
R <- Q[1:4,1:4]
solve(diag(4)-R)
```

in R, we get a matrix whose first row is $(3.2, 1.6, 2.4, 2)$. This sums to $E(T) = 9.2$.

As a check, note that the 2 in the first row agrees with the result of (b). To check $E(T)$, we can apply Theorem 4.4.8 and the fact that $1 - P(T > n)$ is the $(1, 5)$ entry of Q^n.

12

Markov chain Monte Carlo

1. Let $p(x, y)$ be the joint PMF of two discrete r.v.s X and Y. Using shorthand notation as we used with the Gibbs sampler, let $p(x)$ and $p(y)$ be the marginal PMFs of X and Y, and $p(x|y)$ and $p(y|x)$ be the conditional PMFs of X given Y and Y given X. Suppose the support of Y is the same as the support of the conditional distribution of $Y|X$.

   (a) Use the identity $p(x)p(y|x) = p(y)p(x|y)$ to find an expression for the marginal PMF $p(y)$ in terms of the conditional PMFs $p(x|y)$ and $p(y|x)$.

   Hint: Rewrite the identity as $p(x)/p(y) = p(x|y)/p(y|x)$ and take a sum.

   (b) Explain why the two conditional distributions $p(x|y)$ and $p(y|x)$ determine the joint distribution $p(x, y)$, and how this fact relates to the Gibbs sampler.

   Solution:

   (a) Summing
   $$\frac{p(x|y)}{p(y|x)} = \frac{p(x)}{p(y)}$$
   over all x in the support of X, we have
   $$\sum_x \frac{p(x|y)}{p(y|x)} = \sum_x \frac{p(x)}{p(y)} = \frac{\sum_x p(x)}{p(y)} = \frac{1}{p(y)}.$$
   Thus,
   $$p(y) = \frac{1}{\sum_x \frac{p(x|y)}{p(y|x)}}.$$

   (b) If we know the conditional distributions $p(x|y)$ and $p(y|x)$, then by (a) we know the marginal distribution $p(y)$. Then we can obtain the joint PMF using
   $$p(x, y) = p(x|y)p(y).$$

   The method just discussed is a *mathematical* way of obtaining the joint PMF of X and Y in terms of the two conditional distributions $p(x|y)$ and $p(y|x)$; the sum from (a) may be very difficult to compute though. The Gibbs sampler is a *computational* way of *simulating* draws from the joint PMF of X and Y using the two conditional distributions $p(x|y)$ and $p(y|x)$.

2. We have a network G with n nodes and some edges. Each node of G can either be vacant or occupied. We want to place particles on the nodes of G in such a way that the particles are not too crowded. Thus, define a feasible configuration as a placement of particles such that each node is occupied by at most one particle, and no neighbor of an occupied node is occupied.

   Construct a Markov chain whose stationary distribution is uniform over all feasible configurations. Clearly specify the transition rule of your Markov chain, and explain why its stationary distribution is uniform.

   Solution: Starting from any feasible configuration, pick a node v uniformly at random. If

v has any occupied neighbors, make no changes; otherwise, flip v's status with probability $1/2$. Note that this procedure does result in a feasible configuration.

Label the feasible configurations from 1 to M in some way, where M is the number of feasible configurations, and let q_{ij} be the transition probability from state i to state j. Then $q_{ij} = 0$ if i and j differ at more than one node, and

$$q_{ij} = \frac{1}{2n}, \text{ if } i \text{ and } j \text{ differ at exactly one node,}$$

since to get from i to j we need to choose the node at which they differ and then flip it. Therefore, $q_{ij} = q_{ji}$ for all i and j, which shows that the stationary distribution is uniform over all feasible configurations.

3. This problem considers an application of MCMC techniques to image analysis. Imagine a 2D image consisting of an $L \times L$ grid of black-or-white pixels. Let Y_j be the indicator of the jth pixel being white, for $j = 1, \ldots, L^2$. Viewing the pixels as nodes in a network, the neighbors of a pixel are the pixels immediately above, below, to the left, and to the right (except for boundary cases).

Let $i \sim j$ stand for "i and j are neighbors". A commonly used model for the joint PMF of $\mathbf{Y} = (Y_1, \ldots, Y_{L^2})$ is

$$P(\mathbf{Y} = \mathbf{y}) \propto \exp\left(\beta \sum_{(i,j):i\sim j} I(y_i = y_j) \right).$$

If β is positive, this says that neighboring pixels prefer to have the same color. The normalizing constant of this joint PMF is a sum over all 2^{L^2} possible configurations, so it may be very computationally difficult to obtain. This motivates the use of MCMC to simulate from the model.

(a) Suppose that we wish to simulate random draws from the joint PMF of $\mathbf{Y}$, for a particular known value of β. Explain how we can do this using Gibbs sampling, cycling through the pixels one by one in a fixed order.

(b) Now provide a Metropolis-Hastings algorithm for this problem, based on a proposal of picking a uniformly random site and toggling its value.

Solution:

(a) For each pixel i, let $Y_{-i} = (Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_{L^2})$ consist of all the pixels other than i. The full conditional distribution of Y_i given the rest of the pixels is

$$
\begin{aligned}
P(Y_i = y_i | Y_{-i} = y_{-i}) &= \frac{P(Y_i = y_i, Y_{-i} = y_{-i})}{P(Y_{-i} = y_{-i})} \\
&= \frac{P(Y_i = y_i, Y_{-i} = y_{-i})}{P(Y_i = 1, Y_{-i} = y_{-i}) + P(Y_i = 0, Y_{-i} = y_{-i})} \\
&= \frac{\exp\left(2\beta \sum_{j:j\sim i} I(y_j = y_i) \right)}{\exp\left(2\beta \sum_{j:j\sim i} I(y_j = 1) \right) + \exp\left(2\beta \sum_{j:j\sim i} I(y_j = 0) \right)},
\end{aligned}
$$

since the normalizing constant and the terms (j, k) where $j \neq i, k \neq i$ cancel out.

For any i, the above quantity can be computed quickly since the normalizing constant disappeared and there are at most 4 pixels j with $j \sim i$.

A Gibbs sampler can then proceed as follows. Cycle through the pixels in a fixed order, updating them one at a time. When it is pixel i's turn to be updated, let the new value of Y_i be a draw from the conditional distribution of Y_i given the current value of Y_{-i}. A step of the chain is complete when all the pixels have been cycled through.

(b) A Metropolis-Hastings chain can proceed as follows. Let's describe one transition:

(1) Propose to toggle the value of pixel J, where J is Discrete Uniform on $1, 2, \ldots, L^2$. Let $\mathbf{y}$ be the current state and $\tilde{\mathbf{y}} = (y_1, \ldots, y_{i-1}, 1 - y_i, y_{i+1}, \ldots, y_{L^2})$ be the proposed state. (Note that the probability of proposing $\tilde{\mathbf{y}}$ when at $\mathbf{y}$ is the same as the probability of proposing $\mathbf{y}$ when at $\tilde{\mathbf{y}}$.)

(2) Accept the proposal with probability

$$a(\mathbf{y}, \tilde{\mathbf{y}}) = \min\left(\frac{\exp\left(2\beta \sum_{j:j\sim i} I(y_j = 1 - y_i)\right)}{\exp\left(2\beta \sum_{j:j\sim i} I(y_j = y_i)\right)}, 1 \right).$$

If the proposal is accepted, toggle the value of pixel J; otherwise, stay at the current state.

13

Poisson processes

1. Passengers arrive at a bus stop according to a Poisson process with rate λ. The arrivals of buses are exactly t minutes apart. Show that on average, the sum of the waiting times of the riders on one of the buses is $\frac{1}{2}\lambda t^2$.

   Solution: Let T be the total of the waiting times of the riders on one of the buses, and let $N \sim \text{Pois}(\lambda t)$ be the number of people on that bus. Given N, the arrival times of the people on that bus are i.i.d. Uniform over the t minutes preceding the arrival of that bus, so each of those N people has to wait $t/2$ minutes on average. So

   $$E(T) = E(E(T|N)) = E(Nt/2) = \frac{1}{2}\lambda t^2.$$

2. Earthquakes occur over time according to a Poisson process with rate λ. The jth earthquake has intensity Z_j, where the Z_j are i.i.d. with mean μ and variance σ^2. Find the mean and variance of the cumulative intensity of all the earthquakes up to time t.

   Solution: Let $N \sim \text{Pois}(\lambda t)$ be the number of earthquakes up to time t, and $Z = Z_1 + Z_2 + \cdots + Z_N$ be the total intensity of these earthquakes. Then

   $$E(Z) = E(E(Z|N)) = E(N\mu) = \mu\lambda t,$$
   $$\text{Var}(Z) = E(\text{Var}(Z|N)) + \text{Var}(E(Z|N)) = E(N\sigma^2) + \text{Var}(N\mu) = \sigma^2\lambda t + \mu^2\lambda t.$$

3. Alice receives phone calls according to a Poisson process with rate λ. Unfortunately she has lost her cell phone charger. The battery's remaining life is a random variable T with mean μ and variance σ^2. Let $N(T)$ be the number of phone calls she receives before the battery dies; find $E(N(T))$, $\text{Var}(N(T))$, and $\text{Cov}(T, N(T))$.

   Solution: Note that $N(T)|T \sim \text{Pois}(\lambda T)$. So by Adam's law,

   $$E(N(T)) = E(E(N(T)|T)) = E(\lambda T) = \lambda\mu.$$

   By Eve's law,

   $$\text{Var}(N(T)) = E(\text{Var}(N(T)|T)) + \text{Var}(E(N(T)|T)) = E(\lambda T) + \text{Var}(\lambda T) = \lambda\mu + \lambda^2\sigma^2.$$

   Again by Adam's law,

   $$E(TN(T)) = E(E(TN(T)|T)) = E(TE(N(T)|T)) = \lambda E(T^2) = \lambda(\sigma^2 + \mu^2).$$

   So

   $$\text{Cov}(T, N(T)) = E(TN(T)) - E(T)E(N(T)) = \lambda(\sigma^2 + \mu^2) - \lambda\mu^2 = \lambda\sigma^2.$$

4. Emails arrive in Bob's inbox according to a Poisson process with rate λ, measured in emails per hour; each email is work-related with probability p and personal with probability $1 - p$. The amount of time it takes to answer a work-related email is a random variable with mean μ_W and variance σ_W^2, the amount of time it takes to answer a personal email has mean μ_P and variance σ_P^2, and the response times for different emails are independent.

What is the average amount of time Bob has to spend answering all the emails that arrive in a t-hour interval? What about the variance?

Solution: Let T_W and T_P be the amounts of time that Bob spends on answering work-related and personal emails in a t-hour interval, respectively. Let N_W and N_P be the numbers of work-related and personal emails in that time interval, respectively. By Poisson thinning, $N_W \sim \text{Pois}(\lambda t p)$ and $N_P \sim \text{Pois}(\lambda t(1-p))$, with N_W and N_P independent. So T_W and T_P are also independent. We have

$$E(T_W) = E(E(T_W|N_W)) = E(\mu_W N_W) = \mu_W \lambda t p$$

and

$$\text{Var}(T_W) = E(\text{Var}(T_W|N_W)) + \text{Var}(E(T_W|N_W)) = E(\sigma_W^2 N_W) + \text{Var}(\mu_W N_W) = (\sigma_W^2 + \mu_W^2)\lambda t p.$$

The analogous results hold for T_P. Therefore, the expected total time is

$$E(T_W) + E(T_P) = \mu_W \lambda t p + \mu_P \lambda t(1-p)$$

and the variance of the total time is

$$\text{Var}(T_W) + \text{Var}(T_P) = (\sigma_W^2 + \mu_W^2)\lambda t p + (\sigma_P^2 + \mu_P^2)\lambda t(1-p).$$

5. In an endless soccer match, goals are scored according to a Poisson process with rate λ. Each goal is made by team A with probability p and team B with probability $1 - p$. For $j > 1$, we say that the jth goal is a turnaround if it is made by a different team than the $(j-1)$st goal; for example, in the sequence AABBA..., the 3rd and 5th goals are turnarounds.

(a) In n goals, what is the expected number of turnarounds?

(b) If an A-to-B turnaround has just occurred, what is the expected time until the next B-to-A turnaround? If a B-to-A turnaround has just occurred, what is the expected time until the next A-to-B turnaround?

Solution:

(a) For $2 \leq j \leq n$, the jth goal is a turnaround with probability $2p(1-p)$. For each of these goals, create an indicator r.v. for that goal being a turnaround. By linearity, the expected number of turnarounds is $2p(1-p)(n-1)$.

(b) By Poisson process thinning, we can split the goal-scoring process into two independent Poisson processes: one whose arrivals are goals for team A, and one whose arrivals are goals for team B. These processes have rate λp and rate $\lambda(1-p)$, respectively.

If an A-to-B turnaround has just occurred, the next turnaround will occur the next time that A has a goal. So by the memoryless property and independence of the processes, the time until the next turnaround after an A to B turnaround is $\text{Expo}(\lambda p)$, with mean $1/(\lambda p)$. Similarly, the time until the next turnaround after an B to A turnaround is $\text{Expo}(\lambda(1-p))$, with mean $1/(\lambda(1-p))$.

6. Let N_t be the number of arrivals up until time t in a Poisson process of rate λ, and let T_n be the time of the nth arrival. Consider statements of the form

$$P(N_t \lesseqgtr_1 n) = P(T_n \lesseqgtr_2 t),$$

where $\lesseqgtr_1$ and $\lesseqgtr_2$ are replaced by symbols from the list $<, \leq, \geq, >$. Which of these statements are true?

Solution: Note that $N_t \geq n$ is the same event as $T_n \leq t$ (this is a form of the count-time duality). So

$$P(N_t \geq n) = P(T_n \leq t).$$

Since T_n is a continuous r.v., we also have

$$P(N_t \geq n) = P(T_n < t).$$

Taking complements, $N_t < n$ is the same event as $T_n > t$, so

$$P(N_t < n) = P(T_n > t).$$

Again since T_n is a continuous r.v., we also have

$$P(N_t < n) = P(T_n \geq t).$$

The other statements are false since, for example,

$$P(N_t > n) = P(N_t \geq n+1) = P(T_{n+1} \leq t).$$

7. Claims against an insurance company follow a Poisson process with rate $\lambda > 0$. A total of N claims were received over two disjoint time periods of combined length $t = t_1 + t_2$, with t_i the length of period i.

(a) Given this information, derive the conditional probability distribution of N_1, the number of claims made in the first period, given N.

(b) The amount paid for the ith claim is X_i, with $X_1, X_2, \ldots$ i.i.d. and independent of the claims process. Let $E(X_i) = \mu$, $\mathrm{Var}(X_i) = \sigma^2$, for $i = 1, \ldots, N$. Given N, find the mean and variance of the total claims paid in period 1. That is, find these two conditional moments of the quantity

$$W_1 = \sum_{i=1}^{N_1} X_i,$$

where (by convention) $W_1 = 0$ if $N_1 = 0$.

Solution:

(a) The conditional distribution of $N_1|N$ is $\mathrm{Bin}\left(N, \frac{t_1}{t_1+t_2}\right)$, by Theorem 13.2.1.

(b) By Adam's law, Eve's law, and the result from (a),

$$\begin{aligned}
E(W_1|N) &= E(E(X_1 + \cdots + X_{N_1}|N_1, N)|N) \\
&= \mu E(N_1|N) \\
&= \mu N \frac{t_1}{t_1 + t_2},
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Var}(W_1|N) &= E(\mathrm{Var}(W_1|N_1, N)|N) + \mathrm{Var}(E(W_1|N_1, N)|N) \\
&= E(N_1\sigma^2|N) + \mathrm{Var}(N_1\mu|N) \\
&= \sigma^2 N \frac{t_1}{t_1 + t_2} + \mu^2 N \frac{t_1 t_2}{(t_1 + t_2)^2}.
\end{aligned}$$

8. On a certain question-and-answer website, $N \sim \mathrm{Pois}(\lambda_1)$ questions will be posted to-morrow, with λ_1 measured in questions/day. Given N, the post times are i.i.d. and uniformly distributed over the day (a day begins and ends at midnight). When a question is posted, it takes an $\mathrm{Expo}(\lambda_2)$ amount of time (in days) for an answer to be posted, independently of what happens with other questions.

(a) Find the probability that a question posted at a uniformly random time tomorrow will not yet have been answered by the end of that day.

(b) Find the joint distribution of how many answered and unanswered questions posted tomorrow there will be at the end of that day.

Solution:

(a) Let $X \sim \text{Unif}(0,1)$ be the time that the question is posted (treating time 0 as the start of tomorrow), and T be how long it takes until an answer is posted. Then

$$
\begin{aligned}
P(X + T > 1) &= \int_0^1 \int_{1-x}^\infty \lambda_2 e^{-\lambda_2 t} dt dx \\
&= \int_0^1 e^{-\lambda_2 (1-x)} dx \\
&= e^{-\lambda_2} \int_0^1 e^{\lambda_2 x} dx \\
&= \frac{1 - e^{-\lambda_2}}{\lambda_2}.
\end{aligned}
$$

(b) By the chicken-egg story, these are *independent* with $\text{Pois}(\lambda_1 p)$ answered questions and $\text{Pois}(\lambda_1 q)$ unanswered questions, where q is the answer to (a) and $p = 1 - q$.

9. An *inhomogeneous Poisson process* in one dimension is a Poisson process whose rate, instead of being constant, is a nonnegative function $\lambda(t)$ of time. Formally, we require that the number of arrivals in the interval $[t_1, t_2)$ be Poisson-distributed with mean $\int_{t_1}^{t_2} \lambda(t) dt$ and that disjoint intervals be independent. When $\lambda(t)$ is constant, this reduces to the definition of the ordinary or *homogeneous* Poisson process.

(a) Show that we can generate arrivals from an inhomogeneous Poisson process in the interval $[t_1, t_2]$ using the following procedure.

1. Let $\lambda_{\max}$ be the maximum value of $\lambda(t)$ in the interval $[t_1, t_2]$. Create a 2D rectangle $[t_1, t_2] \times [0, \lambda_{\max}]$, and plot the function $\lambda(t)$ in the rectangle.
2. Generate $N \sim \text{Pois}(\lambda_{\max}(t_2 - t_1))$, and place N points uniformly at random in the rectangle.
3. For each of the N points: if the point falls below the curve $\lambda(t)$, accept it as an arrival in the process, and take its horizontal coordinate to be its arrival time. If the point falls above the curve $\lambda(t)$, reject it.

Hint: Verify that the two conditions in the definition are satisfied.

(b) Suppose we have an inhomogeneous Poisson process with rate function $\lambda(t)$. Let $N(t)$ be the number of arrivals up to time t and T_j be the time of the jth arrival. Explain why the hybrid joint PDF of $N(t)$ and $T_1, \ldots, T_{N(t)}$, which constitute all the data observed up to time t, is given by

$$
f(n, t_1, \ldots, t_n) = \frac{e^{-\lambda_{\text{total}}} \lambda_{\text{total}}^n}{n!} \cdot n! \frac{\lambda(t_1) \ldots \lambda(t_n)}{\lambda_{\text{total}}^n} = e^{-\lambda_{\text{total}}} \lambda(t_1) \ldots \lambda(t_n)
$$

for $0 < t_1 < t_2 < \cdots < t_n$ and nonnegative integer n, where $\lambda_{\text{total}} = \int_0^t \lambda(u) du$.

Solution:

(a) The procedure generates a 2D Poisson process of rate 1 in the rectangle $[t_1, t_2) \times [0, \lambda_{\max}]$ and then extracts the horizontal axis coordinates of the points below the curve $\lambda(t)$. So the numbers of arrivals in disjoint subintervals of $[t_1, t_2)$ are independent, since the numbers of points in the corresponding subrectangles of $[t_1, t_2) \times [0, \lambda_{\max}]$ are independent and the accept-reject results for different points are independent.

Let (a, b) be a subinterval (a, b) of $[t_1, t_2)$, and $A = \int_a^b \lambda(t) dt$ be the area under the curve $\lambda(t)$ from a to b. The number of arrivals in (a, b) is the number of generated points below

the curve $\lambda(t)$ from a to b. But since the generated points follow a 2D Poisson process of rate 1, there are $\text{Pois}(A)$ such points. Thus, the number of arrivals in (a, b) is $\text{Pois}(A)$, as desired.

(b) We have $N(t) \sim \text{Pois}(\lambda_{\text{total}})$ by definition of inhomogeneous Poisson process.

Given $N(t) = n$, the arrival times can be generated as the order statistics of n i.i.d. times $V_1, \ldots, V_n$ in $(0, t)$ with PDF at s proportional to $\lambda(s)$ for $s \in (0, t)$. To see this, imagine generating the process according to the procedure from (a). The generated points that get accepted are independent and uniformly random in the area below the curve $\lambda(s)$, $0 < s < t$. And a uniformly random point in the area below the curve $\lambda(s)$, $0 < s < t$, has a horizontal coordinate with PDF proportional to $\lambda(s)$ since the probability of that horizontal coordinate being in a subinterval (t_1, t_2) of $(0, t)$ is proportional to the area under the curve $\lambda(s)$ from $s = t_1$ to $s = t_2$.

So the conditional PDF of one arrival time V_j, given $N(t)$, is $\lambda(s)/\text{Pois}(\lambda_{\text{total}})$ for $0 < s < t$ (and 0 otherwise). Taking the order statistics since by definition the T_j are ordered, the conditional joint PDF of $T_1, \ldots, T_{N(t)}$, given $N(t) = n$, is

$$f(t_1, \ldots, t_n) = n! \frac{\lambda(t_1) \ldots \lambda(t_n)}{\lambda_{\text{total}}^n},$$

for $0 < t_1 < \cdots < t_n$, where the $n!$ accounts for the fact that if the V_j are permuted in any way in the event $V_1 = t_1, V_2 = t_2, \ldots, V_n = t_n$, the same values for $T_1, T_2, \ldots, T_n$ are obtained.

The hybrid joint PDF of $N(t)$ and $T_1, \ldots, T_{N(t)}$ is the PMF of $N(t)$ times the conditional PDF of $T_1, \ldots, T_{N(t)}$ given $N(t)$. Thus, the hybrid joint PDF is as claimed.

10. A *Cox process* is a generalization of a Poisson process, where the rate λ is a random variable. That is, λ is generated according to some distribution on $(0, \infty)$ and then, given that value of λ, a Poisson process with that rate is generated.

(a) Explain intuitively why disjoint intervals in a 1D Cox process are *not* independent.

(b) In a 1D Cox process where $\lambda \sim \text{Gamma}(\alpha, \beta)$, find the covariance between the number of arrivals in $[0, t)$ and the number of arrivals in $[t, t + s)$.
Hint: Condition on λ.

Solution:

(a) Disjoint intervals in a 1D Cox process are not independent (except in the degenerate case where λ is constant, in which case we have a Poisson process). This is because observing the number of arrivals in one interval gives information about λ, which in turn gives us information about the numbers of arrivals in other intervals.

(b) Let $N_{a,b}$ be the number of arrivals in $[a, b)$ for any a, b. By Adam's law,

$$E(N_{0,t}) = E(E(N_{0,t}|\lambda)) = E(\lambda t) = \frac{t\alpha}{\beta}.$$

Similarly,

$$E(N_{t,t+s}) = \frac{s\alpha}{\beta}.$$

By Adam's law and the fact that in a Poisson process the numbers of arrivals in disjoint intervals are independent, we have

$$E(N_{0,t}N_{t,t+s}) = E(E(N_{0,t}N_{t,t+s}|\lambda)) = E(E(N_{0,t}|\lambda)E(N_{t,t+s}|\lambda)) = stE(\lambda^2) = \frac{st\alpha(\alpha+1)}{\beta^2}.$$

Therefore,

$$\text{Cov}(N_{0,t}, N_{t,t+s}) = E(N_{0,t}N_{t,t+s}) - E(N_{0,t})E(N_{t,t+s}) = \frac{st(\alpha + \alpha^2)}{\beta^2} - \frac{st\alpha^2}{\beta^2} = \frac{st\alpha}{\beta^2}.$$

11. In a *Yule process* with rate λ, the rate of arrivals increases after each new arrival, so that the time of the first arrival is $\text{Expo}(\lambda)$ and the time between the $(j-1)$st and jth arrivals is $\text{Expo}(j\lambda)$ for $j = 2, 3, \ldots$. So interarrival times are independent but not i.i.d.

(a) Show that the superposition of two independent Yule processes with the same rate λ is a Yule process, except shifted so that the interarrival times are $\text{Expo}(2\lambda), \text{Expo}(3\lambda), \text{Expo}(4\lambda), \ldots$ rather than $\text{Expo}(\lambda), \text{Expo}(2\lambda), \text{Expo}(3\lambda), \ldots$.

(b) Show that if we project the process from Part (a) into discrete time, the resulting sequence of type-1 and type-2 arrivals is equivalent to the following discrete-time process:

1. Start with two balls in an urn, labeled 1 and 2.
2. Draw a ball out of the urn at random, note its number, and replace it along with another ball with the same number.
3. Repeat step 2 over and over.

Solution:

(a) Let $X_1, X_2, \ldots$ be the interarrival times of the first Yule process and $Y_1, Y_2, \ldots$ be the interarrival times of the second Yule process, so X_j and Y_j are $\text{Expo}(j\lambda)$ and the X_i's and the Y_j's are all independent. Let $Z_1, Z_2, \ldots$ be the interarrival times of the superposed process, and call arrivals that came from the first process "type-1" and arrivals that came from the second process "type-2".

Recall from Example 5.6.3 that the minimum of independent $X \sim \text{Expo}(\lambda_1)$ and $Y \sim \text{Expo}(\lambda_2)$ is $\text{Expo}(\lambda_1 + \lambda_2)$. Then

$$Z_1 = \min(X_1, Y_1) \sim \text{Expo}(2\lambda).$$

Given that $X_1 < Y_1$, by the memoryless property the conditional distribution of Z_2 is $\text{Expo}(3\lambda)$ (since given that $X_1 < Y_1$, the interarrival time Z_2 is the minimum of an $\text{Expo}(2\lambda)$ r.v. and an $\text{Expo}(\lambda)$ r.v.). The same conditional distribution holds given that $Y_1 < X_1$, so it is also true unconditionally that $Z_2 \sim \text{Expo}(3\lambda)$.

More generally, given that there have been n_1 type-1 arrivals and n_2 type-2 arrivals, the time until the next arrival is $\min(X_{n_1+1}, Y_{n_2+1}) \sim \text{Expo}((n_1 + n_2 + 2)\lambda)$. This conditional distribution depends on n_1 and n_2 only through the sum $n_1 + n_2$, so

$$Z_n \sim \text{Expo}((n+2)\lambda).$$

(b) Consider the sequence of *labels* for the arrivals in the superposed process from (a). For example, $2221121111\ldots$ says that the first three arrivals in the superposed process are of type-2, then the next two are of type-1, then the next one is of type-2, then the next four are of type-1, etc. The urn scheme also creates a sequence of 1's and 2's: the sequence recording the numbers of the balls that get drawn, by order of draw.

Recall from Example 7.1.23 or the proof of Theorem 13.2.7 that if $X \sim \text{Expo}(\lambda_1)$ and $Y \sim \text{Expo}(\lambda_2)$ are independent, then $P(X < Y) = \lambda_1/(\lambda_1 + \lambda_2)$; we will apply this fact to the superposed process.

In the urn scheme, if previously there have been exactly n_1 1's and exactly n_2 2's, then the probability that the next draw is a 1 is $(n_1 + 1)/(n_1 + n_2 + 2)$. In the superposed process, if previously there have been exactly n_1 arrivals of type-1 and exactly n_2 arrivals of type-2, then the probability that the next arrival is of type-1 is

$$\frac{(n_1 + 1)\lambda}{(n_1 + 1)\lambda + (n_2 + 1)\lambda} = \frac{n_1 + 1}{n_1 + n_2 + 2}.$$

Thus, the urn scheme sequence and the label sequence for the superposed process are equivalent; they have exactly the same probabilistic structure.

12. Consider the coupon collector problem: there are n toy types, and toys are collected one by one, sampling with replacement from the set of toy types each time. We solved this problem in Chapter 4, assuming that all toy types are equally likely. Now suppose that at each stage, the jth toy type is collected with probability p_j, where the p_j are not necessarily equal. Let N be the number of toys needed until we have a full set; we wish to find $E(N)$. This problem outlines an embedding method for calculating $E(N)$.

(a) Suppose that the toys arrive according to a Poisson process with rate 1, so that the interarrival times between toys are i.i.d. $X_j \sim \text{Expo}(1)$. For $j = 1, \ldots, n$, let Y_j be the waiting time until the first toy of type j. What are the distributions of the Y_j? Are the Y_j independent?

(b) Explain why $T = \max(Y_1, \ldots, Y_n)$, the waiting time until all toy types are collected, can also be written as $X_1 + \cdots + X_N$, where the X_j are defined as in Part (a). Use this to show that $E(T) = E(N)$.

(c) Show that $E(T)$, and hence $E(N)$, can be found by computing the integral

$$\int_0^\infty \left(1 - \prod_{j=1}^n (1 - e^{-p_j t})\right) dt.$$

Hint: Use the identity $E(T) = \int_0^\infty P(T > t)dt$, which was shown in Example 5.3.8.

Solution:

(a) By thinning the Poisson process based on toy type, we have n independent Poisson processes, where an arrival for the jth of these processes occurs whenever a toy of type j is collected. The Poisson process of type j toys has rate p_j. Therefore, the Y_j are independent, with $Y_j \sim \text{Expo}(p_j)$.

(b) The waiting time T until all toy types are collected can be expressed as $\max(Y_1, \ldots, Y_n)$ since the longest waiting time corresponds to when the last new toy type is obtained. But T can also be expressed as $X_1 + X_2 + \cdots + X_N$, since N is the number of toys collected and $X_1 + \cdots + X_N$ is the time at which the Nth toy is collected. Therefore,

$$E(T) = E(X_1 + \cdots + X_N) = E(E(X_1 + \cdots + X_N|N)) = E(NE(X_1)) = E(N),$$

since N is independent of the interarrival times between toys.

(c) The CDF of T is

$$P(T \le t) = P(Y_1 \le t, \ldots, Y_n \le t) = (1 - e^{-p_1 t}) \cdots (1 - e^{-p_n t}),$$

since $Y_1, \ldots, Y_n$ are independent with $Y_j \sim \text{Expo}(p_j)$. Hence,

$$E(N) = E(T) = \int_0^\infty P(T > t)dt = \int_0^\infty \left(1 - \prod_{j=1}^n (1 - e^{-p_j t})\right) dt.$$