STATS 116: Summer 2023

Harrison Li, Stanford University

Contents

Preface

These are course notes for the class "Introduction to Probability" (STATS 116) at Stanford University for the Summer 2023 quarter. STATS 116 is a standard undergraduate level probability class. It is designed to provide a rigorous overview of probability theory, emphasizing both intuition and mathematical precision.

This iteration of the course, including these notes, are based on the textbook `Introduction to Probability` by Joseph K. Blitzstein and Jessica Hwang (hereafter BH). That book itself arose from the course "STAT 110" at Harvard University, for which lecture notes and materials have been made freely available by Prof. Blitzstein. However, STATS 116 in Summer 2023 is a substantially shorter course, containing four 50-minute lectures per week for a condensed 8-week quarter, whereas STAT 110 at Harvard, already considered a rather challenging and fast-paced course, is typically taught in two 80-minute lectures per week throughout a 13-week semester. Thus, the content coverage in these notes is a strict subset of the STAT 110 curriculum. Emphasis is placed on those topics that are, in my view, the most important for further study in statistics and related areas such as machine learning.

Some of the content in these notes is copied or paraphrased directly from BH, so I do not claim proper authorship of these notes. On the other hand, any errors or omissions in these notes are solely my responsibility. The main purpose of these notes, for a student taking STATS 116, is to provide a condensed yet self-contained overview of the material in BH relevant to each day's lecture, along with some of my personal insights and examples which I hope may be of use. Each lecture is separated by a numbered heading. The student is encouraged to refer to BH for an excellent, more discursive exposition of all topics.

The mathematical background assumed for these notes is primarily a solid command of high school algebra and single variable calculus. Some familiarity with partial differentiation and double integration is assumed briefly in the sections on joint distributions, though a full course in multivariable calculus should not be considered necessary. Basic knowledge of sets and set notation is assumed, and these notes are written in a mathematical "definition/theorem/proof" format, with the goal of encouraging rigorous mathematical thinking. In particular, the reader is also encouraged to think carefully about the logical relationship between different ideas. Is concept X a definition? Or is it a theorem; if so, how does it follow from first principles? Some proofs are omitted, however, to maintain focus on the ideas that I view as most important for holistic understanding, and to keep the mathematical level of these notes moderate.

1. Counting: Naive definition of probability, multiplication rule

Probability is the quantitative study of *uncertainty*. Mathematicians and math-adjacent scholars have worked for centuries to try to formalize this in a mathematical way. The modern framework we will be studying in this class is due to Andrey Kolmogorov, an influential Soviet mathematician who lived in the 20th century. We begin with some important fundamental definitions.

Definition 1.1 (Sample space)**.** The **sample space** S of an experiment is the set of all possible outcomes. An **event** A is any subset of the sample space S, i.e. a collection of any number of outcomes in S.

Example 1.2 (Die roll)**.** Suppose I roll a single six-sided die with faces numbered $1, \ldots, 6$. Then the outcome refers to the number rolled, so the sample space is $S = \{1, \ldots, 6\}$. Some events $A_1, \ldots, A_4$ on this sample space are: $A_1 = S$, $A_2 = \{2, 3\}$, $A_3 = \emptyset$, and $A_4 = \{1\}$.

Remark. An event containing one element, like A_4 in Example 1.2, is a set containing one element (the outcome "1"), which is to be distinguished from the outcome itself (which is not a set).

Sample spaces can be finite or infinite. In the *naive definition of probability*, S is assumed finite.

Definition 1.3 (Naive definition of probability)**.** Under the **naive definition of probability**, the probability of the event A (within some finite sample space S), is given by $P(A) = \frac{|A|}{|S|}$, where for any set M, $|M|$ denotes the cardinality of M, i.e. the number of elements in M. Note $0 \leq P(A) \leq 1$ for any event A, at least under the naive definition.

Example 1.4. In the setting of Example 1.2, we have $P(A_2) = \frac{2}{6} = \frac{1}{3}$.

Remark. Probabilities are defined on events, which are *sets*. It does not make (strict mathematical) sense to compute the probability of an outcome, or to compute the probability of anything that is not an event.

The naive definition of probability is naive because it inherently assigns all outcomes in S "equal weight." Intuitively, we know this must not always be true. Later on, we will develop a more refined notion of probability to address this. But for today, we will work with the naive definition, which is a useful starting point in experiments where all outcomes are indeed equally likely. The naive definition reduces computation of probability to counting the number of elements in two sets and dividing them. Thus, to apply it, we need some techniques for counting.

Proposition 1.5 (Multiplication rule)**.** *Suppose I run a compound experiment consisting of two sub-experiments with sample spaces A and B. If, for each of the a possible outcomes in experiment A, there are b possible outcomes in experiment B, then there are ab total outcomes for the compound experiment.*

Proof. Immediate by the definition of multiplication; see the tree diagram, Figure 1.3 of BH, for a pictoral representation. □

Example 1.6 (License plates)**.** Standard license plates in California start with a digit, followed by 3 letters, followed by 3 more digits. Assuming no additional restrictions, the multiplication rule indicates there are $10 \cdot 26^3 \cdot 10^3$ total possible license plates.

Example 1.7 (Sampling with replacement)**.** Suppose I have n items in a hat. I do the following k times: select one item and put it back. Then there are n^k possible sequences of items.

Proof. Immediate from the multiplication rule: I have a compound experiment of k sub-experiments, and regardless of the outcome of the other experiments, I have n outcomes for each experiment. □

Example 1.8 (Sampling without replacement)**.** Suppose I have n items in a hat. For $k \leq n$ iterations, I take one item out of the hat and leave it outside, so that for the next iteration I am only selecting from the items remaining in the hat. Then there are $n \cdot (n-1) \cdot \ldots \cdot (n-k+1) = \frac{n!}{(n-k)!}$ possible sequences of items (here $n!$ is read as "n factorial", and is short-hand for the quantity $n \cdot (n-1) \cdot \ldots \cdot 1$ for each integer $n \geq 0$, with $0! = 1$ for convenience). In particular, there are $n!$ ways to order n objects in a line (take $k = n$).

Proof. There are n choices for the first experiment, $n-1$ choices for the second experiment (regardless of which item I drew in the first experiment), etc. Apply the multiplication rule. □

Remark. The previous example shows that the multiplication rule does not require the same outcomes in sub-experiment B to be possible for every outcome in sub-experiment A. If I took

out item 1 in the first sub-experiment, then I can only get one of the $n-1$ items $2, \ldots, n$ in the next sub-experiment. By contrast, if I took out item n in the first sub-experiment, I can only get items $1, \ldots, n-1$ in the next sub-experiment. That is, the possible outcomes in sub-experiment B can differ depending on the outcome in sub-experiment A. But the multiplication rule is still valid so long as the *number* of possible outcomes in sub-experiment B is the same, regardless of the outcome of sub-experiment A.

One complication in applying the multiplication rule is that if we perform multiple experiments on the same sample space, we might not care about the order of outcomes. Implicit in the statement of the multiplication rule is that the outcomes in the compound experiment correspond to *ordered* pairs of outcomes in the sub-experiments. This is fine in Examples 1.7 and 1.8, we were counting sequences, meaning we cared about the outcome order. But if I want to select a pair of students from a classroom, I don't care which student I selected "first." Our next example resolves this.

Example 1.9 (Binomial coefficient)**.** There are $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ ways to choose a subset of k items out of n. $\binom{n}{k}$ is read as "n choose k" and is known as a binomial coefficient.

Proof. We consider Example 1.8 as a compound experiment with two sub-experiments. For the first sub-experiment, we choose which k (distinct) items are chosen at any point, and for the second sub-experiment we pick the order of those k items. For the second sub-experiment, we note that for any set of k items chosen in the first sub-experiment, there are $k!$ ways to arrange them (last sentence of Example 1.8). Yet Example 1.8 also shows that the total number of outcomes in the compound experiment $\frac{n!}{(n-k)!}$. Thus, by the multiplication rule, the number of outcomes for the first sub-experiment is $\frac{n!}{k!(n-k)!} = \binom{n}{k}$, as desired. □

Example 1.10 (Counting subsets)**.** Suppose the sample space S has size n. How many possible events are there on S? One way to solve this is to note that the number of events with a given cardinality $k \leq n$ is $\binom{n}{k}$ by the previous example (a set is a collection of distinct elements, where the order of those elements does not matter). As k can range from 0 up to n, the answer is $\sum_{k=0}^{n} \binom{n}{k}$. On the other hand, we could consider making a binary decision for each of the n elements of the set: whether to include it or not in the event. By the multiplication rule there are 2^n possible subsets. Thus we have shown the mathematical identity $\sum_{k=0}^{n} \binom{n}{k} = 2^n$ via a "story proof."

2. Counting: Examples, Bose-Einstein, and inclusion-exclusion

We begin with some additional problems reviewing the multiplication rule. Defining an appropriate compound experiment for which the multiplication rule applies takes some practice.

Exercise 2.1 (Three of a kind)**.** Suppose we randomly select 3 cards from a standard deck of 52 cards. What is the probability they all have the same rank?

Solution. We provide two equally valid solutions based on two different sample spaces. One considers an experiment where the outcome is a(n ordered) sequence of 3 cards; the other considers an experiment where the outcome is the unordered collection of 3 cards selected. By symmetry, in both experiments all outcome are equally likely, so we can apply the naive definition of probability.

In the first experiment where order matters, by Example 1.7 there are $52 \cdot 51 \cdot 50$ total outcomes. Letting A be the event consisting of all outcomes composed of cards with the same rank, $|A| = 52 \cdot 3 \cdot 2$ by the multiplication rule (for any first card we have 3 choices of cards with the same rank for the second card, and then 2 cards of the same rank as the first two cards for the third card). Thus, by the naive definition we have $P(A) = \frac{52 \cdot 3 \cdot 2}{52 \cdot 51 \cdot 50} = \boxed{\frac{1}{425}}$.

Alternatively, in the second experiment where order doesn't matter, by Example 1.9 there are $\binom{52}{3}$ total outcomes. There are 13 card ranks, and for each rank there are $\binom{4}{3} = 4$ choices of 3 cards with that rank (by 1.9 again). So by the multiplication rule, there are $13 \cdot 4 = 52$ choices of 3 cards that have the same rank. Then by the naive definition, we have $P(A) = \frac{52}{\binom{52}{3}} = \frac{52}{\frac{52!}{3!49!}} = \frac{52}{\frac{52 \cdot 51 \cdot 50}{3!}} = \boxed{\frac{1}{425}}$. □

Exercise 2.2 (Birthday problem)**.** There are k people in a room. Assume each person's birthday is equally likely to be any of 365 days of the year (ignore leap years). Also assume people's birthdays are independent (we will make independence more precise in a future lecture, but this means all possible sequences of birthdays are equally likely). What is the probability that any two of the k people have the same birthday?

Solution. For this example we define the experiment as the ordered sequence of birthdays of the k people. By Example 1.7 there are 365^k possible such sequences. By Example 1.8, assuming $k \leq 365$, there are $365 \cdot 364 \cdot \ldots \cdot (365 - k + 1)$ sequences where **no** two people have the same birthday (if $k > 365$, then we are guaranteed to have at least 2 people with the same birthday). Thus, there are $365^k - 365 \cdot 364 \cdot \ldots \cdot (365 - k + 1)$ outcomes in the event A that any two of the k people have the same birthday. By the naive definition, the desired probability is

$$P(A) = \frac{365^k - 365 \cdot 364 \cdot \ldots \cdot (365 - k + 1)}{365^k} = \boxed{1 - \frac{365 \cdot 364 \cdot \ldots \cdot (365 - k + 1)}{365^k}}$$

It can be shown that $P(A) \geq 0.5$ for $k \geq 23$, which most people find counterintuitive. □

Last lecture, we discussed how to count the number of ways to select k items out of n with or without replacement when order matters (Examples 1.7 and 1.8). We also discussed binomial coefficients, which pertains to sampling without replacement when order does not matter (Example 1.9). We conclude by computing the number of ways to sample with replacement when order does not matter.

Theorem 2.3 (Bose-Einstein coefficient)**.** *Suppose I have n items in a hat. I do the following k times: select one item and put it back, recording each item I selected on a piece of paper. Then there are $\binom{n+k-1}{k}$ distinct collections of k items I could have recorded.*

Proof. Consider an arrangement of n dots and k bars in some order where the first (leftmost) item is a dot. For example, with $n = 4$ and $k = 3$ we might have

$$\cdot || \cdot \cdot \cdot |$$

Viewing the n dots as the n items, ordered from left to right, each such arrangement corresponds to a unique collection of items from k draws with replacement, where the number of times a particular item (dot) is selected corresponds to the number of bars immediately to the right of that item (dot). The example above corresponds to the collection containing two of the first item and one of the fourth item. Conversely, for any item collection we can draw out a unique ordering of dots and bars

— with a dot in the leftmost position — based on the number of each item in the collection. Thus, the number of possible collections is equal to the number of possible orderings of n dots and k bars where the first item is a dot. This amounts to choosing k positions for the bars among $n+k-1$ possible positions, for which there are $\binom{n+k-1}{k}$ ways. □

The final counting tool we will cover in this class is known as the principle of inclusion-exclusion. It expresses the size of the union of some collection of sets in terms of the size of the individual sets and their intersections, which are sometimes easier to deal with.

Theorem 2.4 (Inclusion-exclusion with two sets)**.** *For any finite sets A and B, we have $|A \cup B| = |A| + |B| - |A \cap B|$.*

Proof. We can separate the set $A \cup B$ into three mutually disjoint sets: $A \setminus B$, $A \cap B$, and $B \setminus A$. Thus

$$|A \cup B| = |A \setminus B| + |A \cap B| + |B \setminus A|$$

Yet A can be decomposed into the disjoint sets $A \setminus B$ and $A \cap B$, so

$$|A| = |A \setminus B| + |A \cap B| \iff |A \setminus B| = |A| - |A \cap B|$$

Similarly we have $|B \setminus A| = |B| - |A \cap B|$. Plugging in to the first equation gives

$$|A \cup B| = |A| - |A \cap B| + |A \cap B| + |B| - |A \cap B| = |A| + |B| - |A \cap B|$$

as desired. □

We can generalize the result to more than two events.

Theorem 2.5 (Inclusion-exclusion)**.** *For any finite sets $A_1, \ldots, A_n$, we have*

$$\left|A_1 \cup \cdots \cup A_n\right| = \sum_{i=1}^{n} |A_i| - \sum_{i<j} |A_i \cap A_j| + \sum_{i<j<k} |A_i \cap A_j \cap A_k| - \ldots + (-1)^{n+1} |A_1 \cap \cdots \cap A_n|$$

Proof. By induction; omitted. □

Example 2.6 (Matching)**.** If we select a random ordering of $(1, 2, 3, 4)$, what is the probability that at least one of the numbers is in its original position?

Proof. The sample space has $4! = 24$ possible orderings (outcomes) so this wouldn't be too hard to solve by exhaustive listing, but inclusion-exclusion provides us with a more systematic way to

compute the numerator in the naive definition. Let $A_i, i = 1, 2, 3, 4$ be the event that the number i is in the i-th position. Then the desired event is $A = A_1 \cup A_2 \cup A_3 \cup A_4$. By inclusion-exclusion we have

$$|A| = \sum_{i=1} 4|A_i| - \sum_{i<j} |A_i \cap A_j| + \sum_{i<j<k} |A_i \cap A_j \cap A_k| - |A_1 \cap A_2 \cap A_3 \cap A_4|$$

For each i we have $|A_i| = 3! = 6$ (we can order the remaining 3 numbers in any order), while $|A_i \cap A_j| = 2$ for any $i < j$, $|A_i \cap A_j \cap A_k| = 1$ for any $i < j < k$, and evidently $|A_1 \cap A_2 \cap A_3 \cap A_4| = 1$. There are $\binom{4}{2} = 6$ pairs of distinct indices (i, j) with $i < j$ and $\binom{4}{3} = 4$ triples of distinct indices (i, j, k) with $i < j < k$. So

$$|A| = 4 \cdot 6 - 6 \cdot 2 + 4 \cdot 1 - 1 = 15$$

and the desired probability is $\frac{15}{24} = \boxed{\frac{5}{8}}$ by the naive definition. □

3. A general definition of (conditional) probability

Counting can be quite useful, but the naive definition of probability is quite limited. For instance, typically one cannot use the naive definition of probability with the Bose-Einstein coefficient, because the collections of items are, in general, not equally likely when each possible *ordered sequence* of items orderings is equally likely. For example, consider the setting of Theorem 2.3 with $n = k = 2$, and suppose we label the two items in the hat as "H" and "T," since this setting is equivalent to flipping a fair coin twice. Then there is a $2/4$ chance of getting the collection $\{H, T\}$ in the two draws/flips, but only a $1/4$ chance of getting $\{H, H\}$, since there are two ways to get the former (H then T or T then H), but only one way to get the latter. This demonstrates the need for a more general definition of probability.

Definition 3.1 (General definition of probability)**.** A **probability space** consists of a sample space S and a **probability function** P mapping each[1] event $A \subseteq S$ to a number $P(A) \in [0, 1]$, where P satisfies the following properties (axioms):

1. $P(\emptyset) = 0$, $P(S) = 1$

2. (Countable additivity) If $A_1, A_2, \ldots$, are disjoint events (i.e. $A_i \cap A_j = \emptyset$ for any $i \neq j$) then

$$P\left(\cup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} P(A_j)$$

[1] If the sample space S is uncountable, e.g. the real line, it turns out that it is impossible to construct a probability function that satisfies the axioms in this definition. This is known as the Banach-Tarski paradox. Kolmogorov's solution is to restrict the domain of P, i.e. the set of events on which probability is well defined, to not include every possible subset of S. However, this restricted domain can still be made sufficiently large so that event encountered in real life is almost surely within this restricted domain. Thus, we do not concern ourselves with these measurability issues in the remainder of this course, and assume every event can be assigned a probability.

Note that by taking $A_{n+1}, A_{n+2}, ...$ to be the empty set, countable additivity implies finite additivity $P(A_1 \cup \ldots \cup A_n) = \sum_{i=1}^{n} P(A_i)$ for any disjoint events $A_1, \ldots, A_n$.

The naive definition is a special case of this general definition of probability, where the sample space S is finite and the probability function P satisfies $P(\{s\}) = 1/|S|$ for all outcomes $s \in S$. The general definition allows for infinite sample spaces and unequal probabilities assigned to each outcome. It also implies several important and intuitive properties:

Theorem 3.2. *Suppose (S, P) is an arbitrary probability space. Then the following are true for any events $A_1, \ldots, A_n \subseteq S$:*

1. *(Complement) $P(A_1^c) = 1 - P(A_1)$, where $A^c = S \setminus A$ for any event A.*
2. *(Monotonicity) If $A_1 \subseteq A_2$, then $P(A_1) \leq P(A_2)$.*
3. *(Inclusion-exclusion) $P\left(\sum_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{i \leq j} P(A_i \cap A_j) + \ldots + (-1)^{n+1} P(A_1 \cap \cdots \cap A_n)$*

Proof. We prove each statement separately:

1. Note A_1 and A_1^c are disjoint with $A_1 \cup A_1^c = S$. Hence by countable additivity and the first axiom, $1 = P(S) = P(A_1 \cup A_1^c) = P(A_1) + P(A_1^c)$.
2. Note A_1 and $B = A_2 \setminus A_1 = A_2 \cap A_1^c$ are disjoint with $A_1 \cup B = A_2$. Hence by countable additivity, $P(A_2) = P(A_1) + P(B) \geq P(A_1)$.
3. This is similar in spirit to Theorem 2.5. We only prove the case of two events, for simplicity. We can write $A_1 \cup A_2$ as the union of three disjoint sets $A_1 \setminus A_2$, $A_1 \cap A_2$, and $A_2 \setminus A_1$. We have $P(A_1 \setminus A_2) = P(A_1) - P(A_1 \cap A_2)$ by countable additivity, after writing A as the disjoint union of $A_1 \cap A_2$ and $A_1 \setminus A_2$. Similarly $P(A_2 \setminus A_1) = P(A_2) - P(A_1 \cap A_2)$. We conclude with one more application of countable additivity, which shows

$$P(A_1 \cup A_2) = P(A_1 \setminus A_2) + P(A_1 \cap A_2) + P(A_2 \setminus A_1) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$$

□

A very common situation when reasoning about probability is to have external information that certain outcomes can be ruled out. If we know this information, it seems like we should not assign any probability to those outcomes. The notion of **conditional probability** formalizes this idea.

Definition 3.3. Let A and B be events with $P(B) > 0$. Then we define $P(A \mid B)$, the conditional probability of the event A *given* B, by

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Remark. The expression $A \mid B$ should not be interpreted as an event. After all, $A \mid B$ is not a set of outcomes. Rather, $P(\cdot \mid B)$ should be viewed as a new probability function defined in terms of the probability function $P(\cdot)$. One can check that so long as $P(B) > 0$, the function $P(\cdot \mid B)$ is a valid probability function in that it satisfies the properties in Definition 3.1 with the same sample space S as the original probability function $P(\cdot)$. Thus, all properties of probability functions (e.g. Theorem 3.2) hold for conditional probability functions, too.

The idea in the definition of $P(A \mid B)$ is to encode the chance that the event A occurred, given knowledge of the fact that the event B occurred. Thinking back to the naive definition of probability should provide some intuition about this definition. If the probabilities in Definition 3.3 were naive probabilities, then we could multiply the numerator and denominator of $P(A \mid B)$ by $|S|$ to get the identity $P(A \mid B) = \frac{|A \cap B|}{|B|}$. This corresponds to the naive definition of probability for the set of outcomes in A in a world where we restrict our sample space to be B.

Example 3.4. Suppose I flip two independent fair coins. What is the (conditional) probability that both coins land heads, given that at least one of them lands heads? What is the (conditional probability) that both coins land heads, given that the first coin lands heads?

Solution. Independence will be defined more precisely later on; here it means that the 4 possible coin flip outcomes $\{HH, HT, TH, TT\}$ are equally likely, so we can apply the naive definition. Let $A = \{HH\}$ be the event that both coins land heads, $B = \{HH, HT, TH\}$ be the event that at least one of the coins lands heads, and $C = \{HH, HT\}$ be the event that the first coin lands heads. Then by the definition of conditional probability, we have

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(\{HH\})}{P(\{HH, HT, TH\})} = \frac{1/4}{3/4} = \boxed{1/3}$$

and

$$P(A \mid C) = \frac{P(A \cap C)}{P(C)} = \frac{P(\{HH\})}{P(\{HH, HT\})} = \frac{1/4}{2/4} = \boxed{1/2}$$

□

4. Bayes' rule and the law of total probability

Recall that we defined conditional probability last time. A useful tool in computing conditional probabilities is Bayes' rule. It relates $P(A \mid B)$, the probability of the event A given B has occured, to $P(B \mid A)$, the probability of B given A. In general, they are NOT the same thing; confusing the two is often known as the prosecutor's fallacy.

Theorem 4.1 (Bayes' rule)**.** *For any events A and B with $P(A) > 0$ and $P(B) > 0$, we have*

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Proof. By the definition of conditional probability, we have

$$P(A \cap B) = P(A \mid B)P(B), \quad P(B \cap A) = P(B \mid A)P(A)$$

Of course, $A \cap B$ and $B \cap A$ are the same event, so we can equate the right-hand side of these two equations:

$$P(A \mid B)P(B) = P(B \mid A)P(A) \implies P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

□

Bayes' rule is often useful in conjunction with the law of total probability (LOTP) to expand the denominator. LOTP expresses the probability of an event in terms of the conditional probability of that event, given other events. These conditional probabilities are often easier to reason about.

Theorem 4.2 (Law of total probability (LOTP))**.** *Suppose $B_1, B_2, \ldots$ is a* partition *of the sample space S, meaning $B_1, B_2, \ldots$ are disjoint events with $\cup_{i=1}^{\infty} B_i = S$. Then*

$$P(A) = \sum_{i=1}^{\infty} P(A \cap B_i) = \sum_{i=1}^{\infty} P(A \mid B_i)P(B_i)$$

Proof. We write $A = \cup_{i=1}^{\infty}(A \cap B_i)$. The events $A \cap B_i$ are disjoint so the first equality follows by countable additivity. The second equality follows directly from the definition of the conditional probabilities $P(A \mid B_i)$. □

We now work through an important example of how to use conditioning to reason about probabilities. It is a good idea to always start by precisely defining the events of interest. Conditioning can be useful because often, conditional probabilities are known but unconditional ones are not. Then one can use some combination of the definition of conditional probability, Bayes' rule, and LOTP to compute the desired probability in terms of the known conditional probabilities.

Exercise 4.3. Suppose a disease affects 1% of the population, and a random individual from the population is selected to take a test billed as 95% accurate. This means that for any infected individual, the test has a 95% chance of returning positive, while for any healthy individual, the test has a 95% chance of returning negative. Find the probability that the individual is infected, conditional on testing positive.

Solution. Let T be the event that the individual tests positive and D be the event that the random individual is infected. From the information in the problem, we have $P(D) = 0.01$, $P(T \mid D) = 0.95$, and $P(T \mid D^c) = 0.05$. We see $P(D \mid T)$; by Bayes' rule and LOTP, noting that D, D^c partition the sample space, we have

$$P(D \mid T) = \frac{P(T \mid D)P(D)}{P(T)} = \frac{P(T \mid D)P(D)}{P(T \mid D)P(D) + P(T \mid D^c)P(D^c)} = \frac{0.95 \cdot 0.01}{0.95 \cdot 0.01 + 0.05 \cdot 0.99} \approx \boxed{0.16}$$

□

The result of the previous example might be surprising; it is likely much lower than what most people might naively expect, which is something closer to 0.95. The key is that without observing the test result, there is a very low chance of selecting someone who tests positive, because the disease is so rare. Seeing the positive test raises the conditional probability well above the "prior" probability of 0.01 that a random person tests positive, but not as high as 0.95, because most of those people who test positive will in fact be healthy (to convince yourself, consider an extreme case where nobody has the disease, in which case the conditional probability of getting the disease given a positive result remains 0).

Next, we describe how to condition on more than one event.

Definition 4.4 (Multiple conditioning)**.** Let A, B, and E be events with $P(B \cap E) > 0$. Then we define the conditional probability of A given B and E as the conditional probability of A given B under the probability function $P(\cdot \mid E)$ (note this is a valid probability function, as $P(E) \geq P(B \cap E) > 0$ by monotonicity):

$$P(A \mid B, E) = \frac{P(A \cap B \mid E)}{P(B \mid E)}$$

Since multiple conditional probabilities are just conditional probabilities under a conditional probability function, which is itself a valid probability function by the remark after Definition 3.3, we immediately have conditional versions of Bayes' rule and LOTP.

Theorem 4.5 (Conditional Bayes' rule)**.** *For any events A, B, and E with $P(A \cap E) > 0$ and $P(B \cap E) > 0$, we have*

$$P(A \mid B, E) = \frac{P(B \mid A, E)P(A \mid E)}{P(B \mid E)}$$

Proof. Note that monotonicity (Theorem 3.2) implies $P(E) \geq P(A \cap E) > 0$. Then $P(\cdot \mid E)$ is a valid probability function, and the result follows from applying Bayes' rule to it. □

Theorem 4.6 (Conditional LOTP)**.** *Suppose $A_1, \ldots, A_n$ is a partition of the sample space S on which events B and E are defined with $P(A_i \cap E) > 0$ for each i. Then*

$$P(B \mid E) = \sum_{i=1}^{n} P(B \mid A_i, E)P(A_i \mid E)$$

Proof. Again, monotonicity (Theorem 3.2) implies $P(E) \geq P(A_i \cap E) > 0$ for each i. Then $P(\cdot \mid E)$ is a valid probability function, and the result follows from applying LOTP to it. □

The definition of multiple conditional probabilities enjoys some nice *consistency* properties: the order of the events to the right of the conditioning bar "—" doesn't matter, and we can interpret all those events as being intersected:

Theorem 4.7. *For any events A, B, and E with $P(B \cap E) > 0$, we have $P(A \mid B, E) = P(A \mid E, B) = P(A \mid (B \cap E))$.*

Proof. We compute

$$P(A \mid E, B) = \frac{P(A \cap E \mid B)}{P(E \mid B)} = \frac{P(A \cap B \cap E)}{P(B)P(E \mid B)} = \frac{P(A \cap B \cap E)}{P(B \cap E)} = P(A \mid [B \cap E])$$

where the first equality is by Definition 4.4 and the remaining equalities apply Definition 3.3. Swapping the roles of B and E in the preceding display, we get $P(A \mid B, E) = P(A \mid [E \cap B]) = P(A \mid [B \cap E])$ as well since of course $E \cap B = B \cap E$. □

5. Independence and conditional independence

We have alluded several times in this course to the idea of "independence." For example, in the birthday problem (Exercise 2.2), we assumed that different people's birthdays were independent. Intuitively, this meant that knowledge about person A's birthday does not give any information about person B's birthday. We now formalize this idea mathematically.

Definition 5.1. Two events A and B are **independent** if $P(A \cap B) = P(A)P(B)$. For short, we may write $A \amalg B$.

The following characterization shows that the definition of independence encodes the intuition described above:

Proposition 5.2. *If $P(A) > 0$, then $A \amalg B$ if and only if $P(B \mid A) = P(B)$. Similarly, if $P(B) > 0$, then A and B are independent if and only if $P(A \mid B) = P(A)$.*

Proof. Assuming $P(A) > 0$, $A \amalg B$ implies (by the definition of conditional probability) that

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A)P(B)}{P(A)} = P(B)$$

Conversely, if $P(B \mid A) = P(B)$, then we reverse this reasoning: $P(A \cap B) = P(B \mid A)P(A) = P(B)P(A)$ where the first equality is again by the definition of conditional probability. The result in the second sentence of the proposition follows immediately by swapping the roles of A and B. $\square$

Recall $P(A \mid B)$ is interpreted as the probability of the event A, given knowledge that B occurred. Thus, Proposition 5.2 shows that saying $A \amalg B$ is equivalent to saying that knowledge of B does not change the probability of A. The characterization of Proposition 5.2 thus provides insight into why e.g. we model repeated coin flips or die rolls as independent.

A related, but very much distinct notion is that of *conditional independence*.

Definition 5.3. Events A and B are **conditionally independent** given a third event E with $P(E) > 0$ if

$$P(A \cap B \mid E) = P(A \mid E)P(B \mid E)$$

The definition of conditional independence is equivalent to that of independence, but applied to the probability function $P(\cdot \mid E)$ instead of $P(\cdot)$. So we can immediately apply Proposition 5.2 to show that Definition 5.3 is equivalent to $P(A \mid B, E) = P(A \mid E)$ whenever $P(B \cap E) > 0$. However, it is important to realize that independence does not imply conditional independence given an arbitrary event E and likewise conditional independence given some event E does not imply (unconditional) independence.

Example 5.4 (Independence does not imply conditional independence)**.** Suppose I flip a fair coin twice. Let A be the event that the first flip lands heads, B be the event that the second flip lands heads, and E be the event that at least one flip lands heads. Events A and B are clearly independent, since the outcome of the first flip does not affect the probability the second flip lands heads. But by Example 3.4, we know $P(A \cap B \mid E) = 1/3$, yet $P(A \mid E) = P(B \mid E) = 2/3$ (why?), so A and B are not conditionally independent given E.

Example 5.5 (Conditional independence does not imply independence)**.** Chris has one fair coin and one double-headed coin. He chooses one coin at random without looking and flips it twice. Conditional on the fair coin being chosen, the coin flips are independent. However, Chris does not know this information, so the first and second flips are intuitively **not** independent because the first flip landing heads makes it more likely that the double-headed coin was chosen, and thus that the next flip will land heads. To see this, let A and B be the events that the first and second flips, respectively, land heads. Let D be the event the double-headed coin was chosen. Then by LOTP

$$P(A) = P(A \mid D)P(D) + P(A \mid D^c)P(D^c) = 1 \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{3}{4}$$

and similarly $P(B) = \frac{3}{4}$ as well. However, by conditional LOTP we have

$$P(B \mid A) = P(B \mid A, D)P(D \mid A) + P(B \mid A, D^c)P(D^c \mid A) = P(D \mid A) + \frac{1}{2} \cdot (1 - P(D \mid A))$$

where we used conditional independence to reason $P(B \mid A, D^c) = P(B \mid D^c) = \frac{1}{2}$. By Bayes' rule we have

$$P(D \mid A) = \frac{P(A \mid D)P(D)}{P(A)} = \frac{1 \cdot \frac{1}{2}}{\frac{3}{4}} = \frac{2}{3}$$

Thus we conclude $P(B \mid A) = \frac{5}{6} > P(B)$, so A and B are not independent.

Exercise 5.6. We return to the disease testing setting of Exercise 4.3. Under the same assumptions as Exercise 4.3, suppose the individual takes two tests, and we assume that the results of the two tests are conditionally independent given their disease status. Note this might not be a reasonable assumption if e.g. the two tests are from the same box, and a significant source of inaccuracies in the test is from box-level contamination. What is the probability the individual is infected, given they tested positive on both tests?

Solution. Let T_i be the event that the individual tests positive on test i and D be the event the individual is infected. Then using Bayes' rule and LOTP, we have

$$\begin{aligned} P(D \mid T_1, T_2) = \frac{P(T_1, T_2 \mid D)P(D)}{P(T_1, T_2)} &= \frac{P(T_1, T_2 \mid D)P(D)}{P(T_1, T_2 \mid D)P(D) + P(T_1, T_2 \mid D^c)P(D^c)} \\ &= \frac{P(T_1 \mid D)P(T_2 \mid D)P(D)}{P(T_1 \mid D)P(T_2 \mid D)P(D) + P(T_1 \mid D^c)P(T_2 \mid D^c)P(D^c)} \\ &= \frac{0.95 \cdot 0.95 \cdot 0.01}{0.95 \cdot 0.95 \cdot 0.01 + 0.05 \cdot 0.05 \cdot 0.99} \approx 0.78 \end{aligned}$$

where the penultimate equality uses conditional independence. Comparing our answer to 0.16 from Exercise 4.3, we see that having two positive tests makes us much more confident that the individual is sick. □

Finally, we extend the definition of independence to more than two events.

Definition 5.7. Events $A_1, \ldots, A_n$ are independent if for any $1 \leq i_1 < \cdots < i_k \leq n$, we have

$$P\left(\cap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} P(A_{i_j})$$

Independence of more than two events implies that every pair of events is independent, but not vice versa.

Example 5.8. Once again, consider two coin tosses. Let A and B be the events that the first and second coin tosses, respectively, land heads. Let C be the event that the two coin flips are the same. Then one can verify that A, B and C are pairwise independent, but $P(A \cap B \cap C) = 1/4$ yet $P(A)P(B)P(C) = 1/8$.

6. Problems and paradoxes in conditioning

Problem solving with conditioning is extremely useful but challenging, so we go through some additional examples in this lecture, starting with the famous Monty Hall problem which has stumped many.

Exercise 6.1 (Monty Hall)**.** A contestant on a game show must pick one of three identical-looking doors. Behind one of the doors is a car; behind the other two doors is a goat. The contestant begins by choosing one of three doors. Monty Hall, the game show host, knows the location of the car and reveals a goat behind one of the two other doors. If both doors are goats, he picks a door at random. The contestant is then allowed to either stay with the original door they chose, or to switch to the remaining door. What should they do?

Solution. The vast majority of people's intuition suggests that switching or staying should lead to an equal probability of getting the car. However, this turns out not to be the case. For $i = 1, 2, 3$, let C_i be the event that the car is behind door i. Let G_i be the event that Monty reveals a goat behind door i. Without loss of generality let door 1 be the one the contestant originally chose. Then by Bayes' rule and LOTP we have

$$P(C_1 \mid G_2) = \frac{P(G_2 \mid C_1)P(C_1)}{\sum_{i=1}^{3} P(G_2 \mid C_i)P(C_i)} = \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2} \cdot \frac{1}{3} + 0 + 1 \cdot \frac{1}{3}} = \frac{1}{3}$$

An identical argument shows $P(C_1 \mid G_3) = \frac{1}{3}$. Thus, regardless of what door Monty picks, the contestant is twice as likely to win the car by always switching! □

The incorrect "intuition" that switching and staying have equal probabilities of success comes from a misapplication of the naive definition of probability, and incorrect use of conditioning. The fact that Monty is using the information he knows about the location of the car and giving you a "hint"

essentially collapses the prior 2/3 probability that the car is not behind the door the contestant initially chose onto the single outside door.

Another important conditioning strategy is that of *first-step analysis*, which is relevant when the problem has some sort of self-similar or recursive structure, making it natural to condition on the outcome of some "first step." We illustrate first-step analysis using the famous *Gambler's ruin* problem.

Exercise 6.2 (Gambler's ruin)**.** A gambler makes a series of independent \$1 bets, each with probability p of succeeding. They start with a bank roll of $i > 0$ dollars and keeps betting until they either have $N > i$ dollars or 0 dollars. What is the probability the gambler will win (have an N dollar bankroll before going bankrupt, i.e. getting "ruined")?

Solution. Let p_i be the probability that the gambler is ruined starting with a bankroll of i dollars, with $p_0 = 1$ and $p_N = 0$. Let A be the event the gambler wins the first bet. Then after the first bet, the gambler's bankroll is either $i + 1$ or $i - 1$ dollars, depending on if they won. With all the bets independent, their probability of going bankrupt given A is p_{i+1}, and the probability of going bankrupt given A^c is p_{i-1}. This gives us the equation

$$p_i = p_{i+1}P(A) + p_{i-1}(1 - P(A)) = p_{i-1} + p(p_{i+1} - p_{i-1})$$

Considering this equation for all $i = 1, \ldots, N-1$, we get a series of $N - 1$ equations in the $N - 1$ unknowns $p_1, \ldots, p_{N-1}$, and can solve them. The result is as follows:

$$p_i = \begin{cases} \frac{1-\left(\frac{1-p}{p}\right)^i}{1-\left(\frac{1-p}{p}\right)^N} & p \neq \frac{1}{2} \\ \frac{i}{N} & p = \frac{1}{2} \end{cases}$$

□

Example 6.3 (Defense attorney's fallacy)**.** A husband is accused of murdering his spouse on the basis of a proven history of abuse. Suppose 1 out of 10,000 husbands who abuse their spouses subsequently go on to murder them; on this basis the defense attorney argues the evidence against his client is slim. Further assume that among people with husbands who are murdered, 1 out of 4 are abused, 1 out of 5 are murdered by their husbands, and 1 out of 2 of those murdered by their husbands were previously abused by their husbands.

The defense attorney is making an insidious mistake: they have failed to account for the information that the spouse has indeed been murdered! (Of course, there's lots of other information to condition too in real life, but we ignore it for simplicity). Let G be the event the husband is guilty of murdering his spouse, A be the event the husband abused his spouse, and M be the event the spouse is murdered (by anyone). We have $P(G \mid A) = 1/10000$, $P(A \mid M) = 1/4$, $P(G \mid M) = 1/5$, and $P(A \mid G, M) = 1/2$. Note all probabilities are implicitly conditioning on the victim having a husband. By Bayes' rule

$$P(G \mid A, M) = \frac{P(A \mid G, M)P(G \mid M)}{P(A \mid M)} = \frac{1/2 \cdot 1/5}{1/4} = \frac{2}{5}$$

which is much higher than $1/10000$.

Our next example is a well-known result called *Simpson's paradox*.

Example 6.4 (Simpson's paradox)**.** The statistics department received applications from 5 tall students and 7 short students. They admitted all 5 tall students and 5 short students. The English department got applications from 20 tall students and 5 short students. They admitted 5 tall students and 1 short student. Thus, within each department, tall students were admitted at a greater rate than short students. Yet across both departments, 10 out of 25 tall students got into either department, which is a lower proportion than the 6 out of 12 short students. Mathematically, let A be the event that a randomly chosen student (among those who applied to either statistics or English) is admitted, S be the event that the student applied to statistics, and T be the event that the student is tall. We have $P(A \mid T, S) > P(A \mid T^c, S)$ and $P(A \mid T, S^c) > P(A \mid T^c, S^c)$, yet $P(A \mid T) < P(A \mid T^c)$. To gain some intuition, note by conditional LOTP that

$$P(A \mid T) = P(A \mid T, S)P(S \mid T) + P(A \mid T, S^c)P(S^c \mid T)$$
$$P(A \mid T^c) = P(A \mid T^c, S)P(S \mid T^c) + P(A \mid T^c, S^c)P(S^c \mid T^c)$$

A randomly chosen tall student has a higher chance of being an English applicant than a randomly chosen short student, and English is a much more competitive department than statistics (as evidenced by the lower admission rates). Thus, even though admissions seems to favor the tall students within each department, in aggregate the tall students have lower admissions rates. If admissions is done on the department level, it can thus be very misleading to draw conclusions based on aggregate statistics.

7. Random variables: Definitions, distributions, and PMFs

Until now, we have worked directly with outcomes and events to reason about probability. This can become unwieldy, especially in settings where the sample space is very large and varied. Often, we care about attributes of the outcome that can be summarized *numerically*. Having a numerical summary will enable us to reason about the outcome using arithmetic operations, which is substantially more convenient. The concept of a random variable addresses this.

Definition 7.1. A **random variable** (r.v.) X is a *function* from the sample space S to the real numbers $\mathbb{R}$. That is, for each outcome $s \in S$, the random variable X returns a number $X(s)$.

Remark. The function $X : S \to \mathbb{R}$ is not itself random. For any fixed outcome $s \in S$, $X(s)$ is a well-defined, single number (otherwise X would not be a function, in the mathematical sense). The randomness in X comes solely through the randomness in the input outcome s.

Example 7.2. Suppose we flip a coin twice and record the outcome of the two flips. Then, as before, we have a sample space $S = \{HH, HT, TH, TT\}$. We can define a random variable X that returns the number of heads in the two flips. Then $X(HH) = 2$, $X(HT) = X(TH) = 1$, and $X(TT) = 0$. We can thus view X as a particular numerical summary of the outcome.

Recall from Lecture 1 that probability is only defined on events. Thus it does not make sense to compute $P(X)$ for a r.v. X, as a random variable is not an event (a collection of outcomes in S), but rather a function. However, what does make sense is to compute probabilities that X takes on certain values. For example, in our coin flipping example, the probability that we get at least one head can be written as $P(X \geq 1)$. Formally, "$X \geq 1$" here should be viewed as convenient shorthand for the event $\{s : X(s) \geq 1\}$, i.e. the set of outcomes in the sample space for which the numerical summary X returns a value of at least 1. In previous lectures we explicitly used a letter

like A to define the event "$X \geq 1$". The advantage of defining these events in terms of random variables will become clearer as we begin to deal with increasingly complex situations.

A fundamental attribute of a random variable is to understand the probability that it takes on certain values. This concept is formalized by the idea of a distribution.

Definition 7.3. The **distribution** $\mathcal{L}_X$ of a random variable X is a function from the set of subsets[1] of $\mathbb{R}$ to $[0, 1]$ specifying the probability that X takes on a value inside that subset. That is, for any subset $B \subseteq \mathbb{R}$, we have $\mathcal{L}_X(B) = P(X \in B)$, where "$X \in B$" is shorthand for the event $\{s \in S \mid X(s) \in B\}$; here S is the sample space that forms the domain of X.

Remark. A random variable X and its distribution $\mathcal{L}_X$ are both functions, but they are very different. For one thing, they don't have the same domains and codomains. X takes as input an outcome from the sample space, and returns a number based on that outcome, according to some pre-specified rule that is designed to capture some aspect of the outcome. By contrast, $\mathcal{L}_X$ takes in a *subset* of the *real line* and returns a probability (between 0 and 1).

Remark. If $g : \mathbb{R} \to \mathbb{R}$ is a function and X is a r.v. on a sample space S then $Y = g(X)$ is also a r.v. on S. Formally, $Y(s) = g(X(s))$ for each $s \in S$. Y is often called a *transformation* of X.

As you can imagine, it would be quite unwieldy to specify the distribution $\mathcal{L}_X$ of a random variable X by enumerating its value on all subsets of $\mathbb{R}$. Luckily, there are more convenient representations of distributions that *determine* the distribution. This means that two random variables for which these representations match must have the same distribution. The easiest case to reason about is when the set of values that X can take on is finite, or countably infinite (i.e. can be listed out in an infinite sequence). This excludes those random variables that can take on any value in $\mathbb{R}$, which we will study at a later date.

Definition 7.4. A random variable X is **discrete** if there exists a finite or countably infinite list of values $a_1, a_2, \ldots$ for which $P(X = a_j \text{ for some } j) = 1$.

Definition 7.5. Let X be a *discrete* random variable. Then its **probability mass function (PMF)** is a function $p_X : \mathbb{R} \to [0, 1]$ mapping each $x \in \mathbb{R}$ to $P(X = x)$. In other words, the PMF of X specifies the probability that X takes on any given value. The **support** of X is the set of

[1] Technically, the distribution can only be defined on a collection of "measurable" subsets of $\mathbb{R}$, which excludes certain pathological ones of no practical concern.

values that X can take on with positive probability, i.e. $\text{supp}(X) = \{x \mid p_X(x) > 0\}$. Note that by Definition 7.4, $\text{supp}(X)$ is finite or countably infinite and $P(X \notin \text{supp}(X)) = 0$.

It turns out that the PMF determines the distribution, so that when asked to specify the distribution of a discrete r.v., it suffices to specify its PMF. This is simpler than specifying the whole distribution, since we do not need to enumerate every subset. But it still requires specifying $P(X = x)$ for all $x \in \text{supp}(X)$.

Theorem 7.6 (PMF determines the distribution)**.** *Suppose X and Y are discrete random variables with the same PMFs, i.e. $p_X(a) = p_Y(a)$ for all $a \in \mathbb{R}$. Then they have the same distributions, i.e. $\mathcal{L}_X(B) = \mathcal{L}_Y(B)$ for all $B \subseteq \mathbb{R}$.*

Proof. Fix $B \subseteq \mathbb{R}$. Since X and Y have the same PMF they must have a common support $\mathcal{S}$ that is at most countably infinite. By LOTP we have

$$\begin{aligned}
\mathcal{L}_X(B) = P(X \in B \cap \mathcal{S}) + P(X \in B \cap \mathcal{S}^c) &= P(X \in B \cap \mathcal{S}) \\
&= \sum_{a \in B \cap \mathcal{S}} p_X(a) \text{ by countable additivity} \\
&= \sum_{a \in B \cap \mathcal{S}} p_Y(a) \text{ since } p_X = p_Y \\
&= P(Y \in B \cap \mathcal{S}) \text{ by countable additivity again} \\
&= \mathcal{L}_Y(B) \text{ repeating the first line}
\end{aligned}$$

□

We caution that equality of random variables is much stronger than equality of distributions. Let X and Y be random variables for which $X = Y$. This means that X and Y are defined on the same sample space S with $X(s) = Y(s)$ for *all* $s \in S$. Then for any $B \subseteq \mathbb{R}$, the event $X(s) \in B$ is identical to the event $Y(s) \in B$ (i.e. they consist of the exact same set of outcomes), so $\mathcal{L}_X(B) = P(X(s) \in B) = P(Y(s) \in B) = \mathcal{L}_Y(B)$, i.e. X and Y have the same distribution. However, two random variables with the same distribution may not be equal. In fact, they don't even have to be defined on the same sample space.

Example 7.7. Suppose X is defined on the sample space $S_X = \{H, T\}$ specifying the outcome of a single fair coin flip, while Y is defined on the sample space $S_Y = \{B, R\}$ specifying whether a randomly drawn card from a standard 52-card deck is black (B) or red (R). We could have

$X(H) = 1 - X(T) = 1$ and $Y(B) = 1 - Y(R) = 1$ so that $p_X(1) = p_Y(1) = p_X(0) = p_Y(0) = 1/2$, so $\mathcal{L}_X = \mathcal{L}_Y$ by Theorem 7.6.

Finally, we show that distributional equality is closed under transformations g.

Proposition 7.8. *For any $g : \mathbb{R} \to \mathbb{R}$, $\mathcal{L}_X = \mathcal{L}_Y$ implies $\mathcal{L}_{g(X)} = \mathcal{L}_{g(Y)}$.*

Proof. Fix $A \subseteq \mathbb{R}$. We have

$$P(g(X) \in A) = P(X \in g^{-1}(A)) = P(Y \in g^{-1}(A)) = P(g(Y) \in A)$$

where $g^{-1}(A) \subseteq \mathbb{R}$ denotes the preimage of A under g, and the second equality uses the fact $\mathcal{L}_X = \mathcal{L}_Y$. □

8. Coin flipping distributions

This lecture and next, we will discuss several common families of discrete distributions. While Theorem 7.6 shows that mathematically, we could define any discrete distribution by its PMF, it will be much more informative and useful to define these distribution families with a "story," that is, a description of how the distributions may arise in nature.

Each of these distribution families is indexed by one or more "parameters" that describe certain aspects of the distribution. For each possible set of parameter values we get a specific distribution, but the set of distributions across all parameter values is unified by a story.

Definition 8.1. Suppose I flip a weighted coin that lands heads with probability $p \in (0, 1)$. Then the distribution of the random variable X that is equal to 1 when I flip heads and 0 otherwise is called a **Bernoulli** distribution with parameter p, abbreviated $\text{Bern}(p)$.

From the definition of a Bernoulli random variable, it is clear that if $X \sim \text{Bern}(p)$, its support is $\{0, 1\}$ and its PMF is

$$p_X(x) = P(X = x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

An important example of a random variable with a Bernoulli distribution is known as an indicator random variable. Indicators will be tremendously useful in our later discussion of expectation.

Definition 8.2. Let A be any event. Then the random variable I_A equal to 1 when A occurs and 0 otherwise is called the **indicator random variable** for A. Clearly, $I_A \sim \text{Bern}(P(A))$.

The Bernoulli distribution is very simple and limiting, being supported on only 2 points at 0 and 1. We now describe some more complex distributions that arise from coin tossing stories.

Definition 8.3. Suppose I flip a weighted coin that lands heads with probability $p \in (0, 1)$, and I do this $n \in \{1, 2, \dots, \}$ times. Then the distribution of the random variable X that counts the

number of heads I flip is called a **Binomial** distribution with parameters n and p, abbreviated $\text{Bin}(n, p)$.

The PMF of a binomial distribution is a bit harder to derive, so we state it as a proposition.

Proposition 8.4. *If $X \sim \text{Bin}(n, p)$, then $\text{supp}(X) = \{0, 1, \ldots, n\}$ and its PMF is given by*

$$p_X(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n$$

Proof. That $\text{supp}(X) = \{0, 1, \ldots, n\}$ is immediate from the story definition. Now fix $x \in \{0, 1, \ldots, n\}$ and consider a fixed sequence of n coin flips containing x heads and $n - x$ tails. That sequence has probability $p^x(1-p)^{n-x}$ for the weighted coin in Definition 8.3, since each coin flip is independent. There are $\binom{n}{x}$ possible flip sequences with x heads and $n-x$ tails (consider choosing the x positions of the heads among $\{1, \ldots, n\}$). Thus for $X \sim \text{Bin}(n, p)$, we have $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$. □

Remark. The $\text{Bin}(1, p)$ distribution is the same as the $\text{Bern}(p)$ distribution. This is immediate from comparing the stories ("number of heads in 1 flip" is equivalent to recording 1 when a flip is heads, 0 otherwise). It is also immediate from comparing the PMFs.

Next, we consider some distribution families that arise when the number of coin flips is not fixed, but rather dependent on what we observe.

Definition 8.5. Suppose I repeatedly flip a weighted coin that lands heads with probability $p \in (0, 1)$. Let X be the number of tails I observe before flipping heads for the first time. The distribution of X is called **Geometric** with parameter p, abbreviated $\text{Geom}(p)$. Furthermore, the distribution of $X + 1$ (i.e. the number of flips including the final heads flip) is called **First Success** with parameter p, abbreviated $\text{FS}(p)$.

The geometric and first success distributions are the first example of distributions with *infinite* support. However, they are still discrete, since their supports are countable.

Proposition 8.6. *If $X \sim \text{Geom}(p)$, then $\text{supp}(X) = \{0, 1, \ldots, \}$ and its PMF is given by $p_X(x) = (1-p)^x p$ for $x \in \text{supp}(X)$. If $Y \sim \text{FS}(p)$, then its PMF is given by $p_Y(y) = (1-p)^{y-1} p$ for $y = 1, 2, \ldots,$.*

Proof. If $X \sim \text{Geom}(p)$ according to the story definition, then for any nonnegative integer x, the event $X = x$ is equivalent to the event that among the first $x + 1$ coin flips we get a sequence of x

tails followed by one head. This has probability $(1-p)^x p$ by independence, showing the expression for $p_X(x)$. If $Y \sim \text{FS}(p)$, then $Y - 1 \sim \text{Geom}(p)$, so for each $y \in \{1, 2, \dots, \}$ we have

$$p_Y(y) = P(Y = y) = P(Y - 1 = y - 1) = (1-p)^{y-1}p$$

□

Remark. If X is a discrete random variable, by countable additivity we must have

$$\sum_{x \in \text{supp}(X)} p_X(x) = \sum_{x \in \text{supp}(X)} P(X = x) = P(X \in \text{supp}(X)) = 1$$

Our PMF derivations above thus show combinatorial identities that are not necessarily the easiest to prove using only algebra, e.g. for all $p \in (0, 1)$ and $n = \{0, 1, \dots, \}$ we have

$$\sum_{i=0}^{\infty} (1-p)^i p = 1, \quad \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} = 1$$

Our final distribution for today is a generalization of the geometric distribution.

Definition 8.7. Suppose I repeatedly flip a weighted coin that lands heads with probability $p \in (0, 1)$. Let X be the number of tails I observe before flipping the r-th head, $r = 1, 2, \dots,$. Then X follows a **Negative Binomial distribution** with parameters r and p, abbreviated $\text{NBin}(r, p)$.

Proposition 8.8. *If $X \sim \text{NBin}(r, p)$, then $\text{supp}(X) = \{0, 1, \dots, \}$ and its PMF is given by $p_X(x) = \binom{x+r-1}{r-1} p^r (1-p)^x$.*

Proof. If $X \sim \text{NBin}(r, p)$ according to the story definition, then $P(X = x)$ is the probability that, among the first $x + r - 1$ flips, we have exactly $r - 1$ tails, and then on the next flip (the $x + r$-th) we have heads. Let Y be the number of tails in the first $x + r - 1$ flips. By the story of the binomial, $Y \sim \text{Bin}(x + r - 1, 1 - p)$. Let A be the event that the $x + r$-th flip is heads. Then

$$p_X(x) = P(Y = y, A) = P(Y = y)P(A) = \binom{x+r-1}{r-1} p^{r-1}(1-p)^x \cdot p = \binom{x+r-1}{r-1} p^r (1-p)^x$$

by Proposition 8.4 and independence of the events $Y = y$ and A. □

In actual problems, you typically will not be dealing with actual coin flips. But they provide a useful starting point for lots of problems. One of the most important skills to gain from this course should be the ability to perform *pattern matching*, relating new probabilistic scenarios and problems to more familiar ones like coin flips. Problems in this course are designed with this learning objective in mind.

9. Discrete Uniform, Hypergeometric, and Poisson

Last lecture, we covered several named distribution families that arise from coin flipping. Today, we consider several more important discrete distribution families that arise in other settings.

Definition 9.1. Suppose C is a finite and nonempty set of numbers, and let X be a number drawn uniformly at random from C (i.e. each number has equal probability of being drawn). Then X has the **Discrete Uniform distribution** with parameter C, abbreviated as $\text{DUnif}(C)$.

Remark. It is immediate from the definition that if $X \sim \text{DUnif}(X)$, then $\text{supp}(X) = C$ and its PMF is given by $p_X(x) = 1/|C|$ for all $x \in C$. Also, note that the parameter C is a *set*.

Several of the distributions from last lecture considered a setting where coins were flipped repeatedly, with the coin tosses being independent. Our next named distribution family considers a setting where draws are done without replacement, and hence not independent.

Definition 9.2. Let w, b, and n be positive integers with $n \leq w + b$, and suppose we have an urn containing w white balls and b black balls. Let X be the number of white balls in a random sample of n of these balls drawn *without replacement*. Then the distribution of X is **Hypergeometric** with parameters w, b, and n, abbreviated as $\text{HGeom}(w, b, n)$.

We can derive the PMF of the hypergeometric using counting arguments.

Proposition 9.3. *Suppose* $X \sim \text{HGeom}(w, b, n)$. *Then* $\text{supp}(X) = \{\max(0, n-b), \ldots, \min(w, n)\}$ *and the PMF of* X *is given by*

$$p_X(x) = P(X = x) = \frac{\binom{n}{x}\binom{w+b-n}{w-x}}{\binom{w+b}{w}} = \frac{\binom{w}{x}\binom{b}{n-x}}{\binom{w+b}{n}}$$

for all $x \in \text{supp}(X)$.

Proof. Consider $X \sim \text{HGeom}(w, b, n)$ as in the story definition. The number of white balls in the sample upper bounded by both n (the size of the sample) and w (the total number of white balls in the urn). Conversely, the number of white balls in the sample is trivially lower bounded by 0 and also by $n - b$, since the number of black balls in the sample is upper bounded by $\min(b, n)$. Thus $\text{supp}(X) = \{\max(0, n - b), \ldots, \min(w, n)\}$.

Now we show the two different representations of the PMF by two different stories. First, we view balls of the same color as indistinguishable. Suppose the $w + b$ balls in the urn are arranged in a random order and the n balls drawn correspond to the first n of these in the order. For any $x \in \text{supp}(X)$, the event $X = x$ is equivalent to the event that there are x white balls and $n - x$ black balls among the first n balls. There are $\binom{w+b}{w}$ total ways to order all the balls (choose the positions of the white balls). Out of these, by the multiplication rule there are $\binom{n}{x}\binom{w+b-n}{w-x}$ orderings containing both x white balls in the first n positions and $w - x$ white balls in the last $w + b - n$ positions. With all orderings equally likely, by the naive definition of probability we conclude

$$P(X = x) = \frac{\binom{n}{x}\binom{w+b-n}{w-x}}{\binom{w+b}{w}}$$

Alternatively, we could view all the balls as distinguishable. Then there are a total of $\binom{w+b}{n}$ ways to select a collection of n balls from the urn. Of these, there are $\binom{w}{x}\binom{b}{n-x}$ selections containing x white balls and $n - x$ black balls. This shows

$$P(X = x) = \frac{\binom{w}{x}\binom{b}{n-x}}{\binom{w+b}{n}}$$

as well. Note that the equivalence of our two expressions for the PMF could have been shown using algebra. However, this "story proof" provides a different probabilistic insight. □

The hypergeometric distribution can be quite challenging to grasp fully upon a first (or second, or third) reading, so we provide some examples.

Example 9.4 (Capture-recapture)**.** Suppose a forest contains N elk. We randomly see and tag m of them. At a later date, we see n of these elk; assume that all $\binom{N}{n}$ of the elk are equally likely to appear. Then the number of tagged elk follows a $\text{HGeom}(m, N - m, n)$ distribution (the tagged elk correspond to the "white balls" and the untagged elk correspond to the "black balls"). This idea of capture-recapture is widely used in ecology to infer the total population size N when it is unknown.

Example 9.5 (Aces in a poker hand)**.** The distribution of the number of aces in a 5-card hand drawn at random from a standard 52-card deck is HGeom$(4, 48, 5)$.

Proposition 9.6. *Suppose* $X \sim \text{HGeom}(w, b, n)$ *and* $Y \sim \text{HGeom}(n, w + b - n, w)$*. Then* X *and* Y *have the same distribution.*

Proof. Immediate after noting that X and Y have the same PMF, by Proposition 9.3. The "story" argument comes from the equivalence of the two stories in the proof of that result. □

The final named discrete distribution family we will study is commonly used in modeling generic count data, due to a few nice mathematical properties, some of which we will come to discover later in the quarter.

Definition 9.7. A random variable X follows a **Poisson** distribution with parameter $\lambda > 0$, abbreviated Pois(λ), if its PMF satisfies $p_X(x) = P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!}$.

The Poisson distribution arises from a fairly technical result known as the law of rare events, which we won't prove in this course; instead we state an informal version.

Theorem 9.8 (Law of rare events)**.** *Suppose we have a series of* n *events* $A_1, \ldots, A_n$*, occurring with probabilities* $p_1, \ldots, p_n$*, respectively, and assume the events are independent or nearly so. Then if* $\sum_{i=1}^n p_i$ *converges to some* $\lambda > 0$ *as* n *gets large, the distribution of* $\sum_{i=1}^n I_{A_i}$*, the number of events among* $A_1, \ldots, A_n$ *that occur, is approximately* Pois(λ)*.*

Example 9.9. Suppose 10,000 cars pass by a curved stretch of Highway 17 on a given day, and each of those cars has probability 0.00005 of crashing into the guardrail. Then the distribution of the number of cars that crash into the guardrail that day is approximately Pois(0.5) (note $0.5 = 10,000 \cdot 0.00005$), and the probability of at least one accident is about $1 - \exp(-0.5)$, by the Poisson PMF. Note that we might not expect each car to crash into the guardrail independently; one car crashing might encourage a cascade of crashes due to distracted driving, etc. But as long as this dependence is sufficiently "weak," the law of rare events ensures the Poisson is a good approximation for the distribution of the number of accidents.

Remark. In the setup of the law of rare events, if the events are all completely independent and $p_1 = \ldots = p_n = p$, then the exact distribution of $\sum_{i=1}^n I_{A_i}$ is Bin(n, p). The law of rare events says that if $p \to 0$ as $n \to \infty$ in a way such that $np \to \lambda \in (0, \infty)$, then this distribution is approximately equivalent to Pois(λ) in the limit.

10. CDFs, transformations, and independence of random variables

Besides the PMF, another important function that determines the distribution of a random variable is the cumulative distribution function, or CDF. Unlike PMFs, CDFs are defined for any random variables (including non-discrete ones), and we will study them extensively later on in the course.

Definition 10.1. The **cumulative distribution function (CDF)** of a random variable X is a function $F_X : \mathbb{R} \to [0, 1]$ mapping each $x \in \mathbb{R}$ to $P(X \leq x)$.

Example 10.2. If $X \sim \text{Bern}(p)$, then the CDF of X is given by

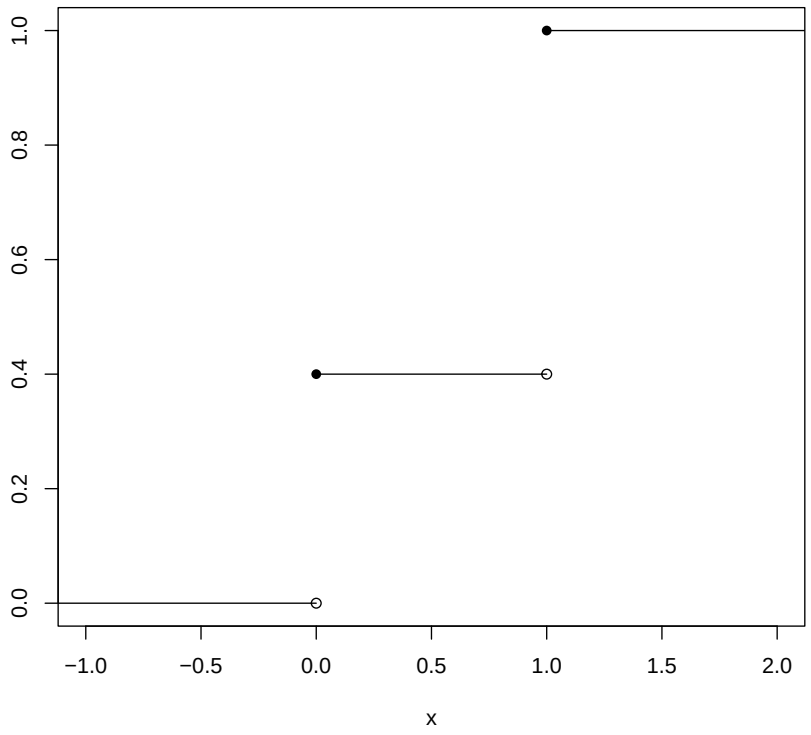
$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1 - p & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases}$$

See Fig. 10.1 for a graphical representation in the case $p = 0.6$.

It is easy to show that for discrete random variables, the CDF determines the PMF, showing that the CDF determines the distribution by Theorem 7.6.

Theorem 10.3 (CDF determines the distribution)**.** *Suppose X and Y are discrete random variables with the same CDFs, i.e. $F_X(a) = F_Y(a)$ for all $a \in \mathbb{R}$. Then they have the same distributions, i.e. $\mathcal{L}_X(B) = \mathcal{L}_Y(B)$ for all $B \subseteq \mathbb{R}$.*

Figure 10.1: The CDF of the Bern(0.6) distribution



Proof. Fix $x \in \text{supp}(X)$. With X discrete, there exists $x_0 < x$ so that $\text{supp}(X) \cap (x_0, x) = \emptyset$. Then

$$\begin{aligned} P(X = x) = P(X \leq x) - P(X \leq x_0) &= F_X(x) - F_X(x_0) \\ &= F_Y(x) - F_Y(x_0) \text{ since } X \text{ and } Y \text{ have the same CDF} \\ &= P(Y \leq x) - P(Y \leq x_0) \\ &= P(Y = x) \end{aligned}$$

where the last equality follows from the fact that X and Y must have the same support (if not, F_Y would have a "jump" where F_X does not). Thus X and Y have the same PMF. □

Next, we define the notion of independence of random variables, related to but distinct from the notion of independence of events.

Definition 10.4. Random variables $X_1, \ldots, X_n$ are independent if $P\left(\cap_{i=1}^n X_i \in B_i\right) = \prod_{i=1}^n P(X_i \in B_i)$ for any sets $B_1, \ldots, B_n$ of real numbers.

Remark. If $X_1, \ldots, X_n$ are independent, then for any sets $B_1, \ldots, B_n$, the events $X_1 \in B_1$,...,$X_n \in B_n$ are independent (to check e.g. pairwise independence of $X_1 \in B_1$ and $X_2 \in B_2$, apply Definition 10.4 with $B_3 = \cdots = B_n = \mathbb{R}$).

Definition 10.4 is quite cumbersome to check in practice, since it suggests we need to check some probabilities for every possible set of real subsets $B_1, \ldots, B_n$. Luckily, it turns out this is not necessary (we will not prove this).

Proposition 10.5. *Random variables $X_1, \ldots, X_n$ are independent if and only if $P(\cap_{i=1}^n X_i \leq x_i) = \prod_{i=1}^n P(X_i \leq x_i)$ for all real numbers $x_1, \ldots, x_n$. If $X_1, \ldots, X_n$ are discrete, then they are independent if and only if $P(\cap_{i=1}^n X_i = x_i) = \prod_{i=1}^n P(X_i = x_i)$ for all $x_1 \in \text{supp}(X_1), \ldots, x_n \in \text{supp}(X_n)$.*

Intuitively, independence of X and Y means that knowing anything about the distribution of X does not change your belief about the distribution of Y. This can be seen if we rewrite the definition of independence in terms of conditional probabilities; if $P(X \in B_1) > 0$, we have by the definition of conditional probability that

$$P(X \in B_1 \cap Y \in B_2) = P(X \in B_1)P(Y \in B_2) \iff P(Y \in B_2 \mid X \in B_1) = P(Y \in B_2)$$

An important special case of independent random variables, commonly encountered in statistics, is when all those random variables have the same distribution.

Definition 10.6. If $X_1, \ldots, X_n$ are independent random variables with the same distribution, then they are said to be **independent and identically distributed (i.i.d.)**.

Armed with the concept of independence of random variables, we are ready to state a few results about sums of independent random variables.

Theorem 10.7. *Suppose $I_1, \ldots, I_n$ are i.i.d.* Bern(p) *random variables. Then $I_1 + \cdots + I_n \sim$* Bin(n, p).

Proof. We present a "story proof." A more analytical proof of a similar result is presented in Theorem 10.9. View each I_i as the indicator of whether the i-th flip of a coin that lands heads with probability p indeed lands heads. Note this is valid because we have assumed the random variables are independent (if they weren't, we could not model the setting with independent coin flips). Then $I_1 + \cdots + I_n$ is the number of heads in n flips of the coin, so must have the Bin(n, p) distribution by the definition of the binomial. □

Corollary 10.8. *If $X \sim \text{Bin}(n,p)$ and $Y \sim \text{Bin}(m,p)$ are independent, then $X+Y \sim \text{Bin}(n+m,p)$.*

Proof. By the previous theorem we can represent $X = I_1 + \cdots + I_n$ and $Y = J_1 + \cdots + J_m$ where $I_1, \ldots, I_n, J_1, \ldots, J_m$ are i.i.d. Bern(p) random variables (independence of X and Y ensures we can view the J_j as independent of the I_i). Thus $X + Y \sim \text{Bin}(n+m,p)$ by applying the previous theorem once again. □

Now we show that the sum of independent Poisson random variables also has a Poisson distribution.

Theorem 10.9. *Suppose $X \sim \text{Pois}(\lambda)$ is independent of $Y \sim \text{Pois}(\mu)$. Then $X+Y \sim \text{Pois}(\lambda+\mu)$.*

Proof. We directly compute the PMF of $Z = X + Y$. Let $\mathbb{Z}_+ = \{0, 1, \ldots\}$ and fix $z \in \mathbb{Z}_+$. Then

$$\begin{aligned}
P(Z=z) = P(X+Y=z) &= \sum_{x=0}^{\infty} P(X+Y=z \cap X=x) \text{ by countable (finite) additivity} \\
&= \sum_{x=0}^{\infty} P(X=x \cap Y=z-x) \\
&= \sum_{x=0}^{z} P(X=x)P(Y=z-x) \text{ as } X \amalg Y \text{ and } \text{supp}(Y) = \mathbb{Z}_+ \\
&= \sum_{x=0}^{z} \frac{e^{-\lambda}\lambda^x}{x!}\frac{e^{-\mu}\mu^{z-x}}{(z-x)!} \text{ by the Poisson PMF} \\
&= e^{-(\lambda+\mu)}(\lambda+\mu)^z \sum_{x=0}^{z} \frac{1}{x!(z-x)!}\left(\frac{\lambda}{\lambda+\mu}\right)^x \left(\frac{\mu}{\lambda+\mu}\right)^{z-x}
\end{aligned}$$

Since the PMF of a Bin $\left(z, \frac{\lambda}{\lambda+\mu}\right)$ random variable must sum to 1, we have

$$1 = \sum_{x=0}^{z} \binom{z}{x}\left(\frac{\lambda}{\lambda+\mu}\right)^x \left(1-\frac{\lambda}{\lambda+\mu}\right)^{n-x} = z! \sum_{x=0}^{z} \frac{1}{x!(z-x)!}\left(\frac{\lambda}{\lambda+\mu}\right)^x \left(\frac{\mu}{\lambda+\mu}\right)^{n-x}$$

We conclude from the above that $P(Z=z) = e^{-(\lambda+\mu)}\frac{(\lambda+\mu)^z}{z!}$, showing $Z \sim \text{Pois}(\lambda+\mu)$. □

Proposition 10.10. *If $X_1, \ldots, X_n$ are i.i.d.* Geom(p) *random variables, then $X_1 + \cdots + X_n \sim$ NBin(n,p).*

Proof. As in the Negative Binomial story, repeatedly flip a coin that lands heads with probability p until you've seen n heads. Let X_1 be the number of flips before the first head, and X_i between the $(i-1)$-th head and the i-th head for $i \geq 2$. Then $X_1, \ldots, X_n$ are independent since they correspond to distinct flips, while they all have a Geom(p) distribution by the story of the Geometric. □

11. Expectation: Definitions, linearity, and LOTUS

Today we begin our discussion of *expectation*. The expectation of a random variable is a single number that corresponds to the average value it takes on. For a discrete random variable, it can be viewed as a weighted sum of the possible values of the random variable, where the weight for a given value is equal to the probability the random variable takes on that value.

Definition 11.1. Let X be a discrete random variable with PMF p_X. Then its **expectation** $\mathbb{E}[X]$ is given by

$$\mathbb{E}[X] = \sum_{x \in \text{supp}(X)} x \cdot p_X(x)$$

Note that if $\text{supp}(X)$ is infinite, then $\mathbb{E}[X]$ is defined by an infinite sum which may not converge. In that case, we say $\mathbb{E}[X]$ *does not exist*.

Remark. The expectation of a (discrete) random variable is a deterministic function of its PMF, so any random variables with the same distribution must have the same expectation.

Example 11.2. Suppose X is a constant random variable, i.e. $P(X = c) = 1$ for some constant c. Then $\mathbb{E}[X] = c \cdot 1 = c$.

Example 11.3. Suppose $X \sim \text{Bern}(p)$. Then $\mathbb{E}[X] = 0 \cdot (1-p) + 1 \cdot p = p$.

Example 11.4. Suppose $X \sim \text{Pois}(\lambda)$. Then

$$\begin{aligned}
\mathbb{E}[X] = \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!} = \sum_{x=1}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!} &= e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^x}{(x-1)!} \\
&= \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \\
&= \lambda \text{ since } \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{\lambda} \text{ by Taylor series}
\end{aligned}$$

An extremely useful property of expectation is *linearity*. The utility comes from the fact that deriving the distribution of a sum of random variables (with possibly complex dependence structure) can be quite tedious. However, if each of the individual random variables has an expectation that's easier to calculate (e.g. because its PMF is known or easy to work with), then it's easy to compute the expectation of the sum by linearity. Note that linearity does not require *any* assumptions about the dependence structure between the random variables in the sum.

Theorem 11.5 (Linearity of expectation)**.** *Let X and Y be random variables and a, b be real numbers. Then*

1. $\mathbb{E}[aX] = a\mathbb{E}[X]$
2. $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$

and hence $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$.

Proof. We will only prove linearity for the case that X and Y are discrete. Let p_X and p_Y be the PMFs of X and Y, respectively. First, we show item 1. Let $Z = aX$ and first suppose $a \neq 0$, so that we can uniquely write each $z \in \text{supp}(Z)$ as $z = ax$ for some $x \in \text{supp}(X)$. Then

$$\mathbb{E}[Z] = \sum_{z \in \text{supp}(Z)} zP(Z = z) = \sum_{x \in \text{supp}(X)} (ax)P(Z = ax) = \sum_{x \in \text{supp}(X)} axP(X = x) = a\mathbb{E}[X]$$

by applying the definition of $\mathbb{E}[Z]$, the fact $Z = ax \iff aX = ax \iff X = x$, and then the definition of $\mathbb{E}[Z]$. For $a = 0$ we have $Z = 0$ and hence $\mathbb{E}[Z] = 0 = 0 \cdot \mathbb{E}[X]$ by Example 11.2.

Now for item 2, we write $Z = X + Y$ and note that

$$\begin{aligned}
\mathbb{E}[Z] &= \sum_{z \in \text{supp}(Z)} zP(Z = z) \\
&= \sum_{z \in \text{supp}(Z)} z \sum_{x \in \text{supp}(X)} P(X = x, Z = z) \text{ by countable additivity} \\
&= \sum_{x \in \text{supp}(X)} \sum_{z \in \text{supp}(Z)} zP(X = x, Z = z) \\
&= \sum_{x \in \text{supp}(X)} \sum_{z \in \{z' \in \text{supp}(Z) : z' - x \in \text{supp}(Y)\}} zP(X = x, Y = z - x) \\
&= \sum_{x \in \text{supp}(X)} \sum_{y \in \text{supp}(Y)} (x + y)P(X = x, Y = y) \\
&= \sum_{x \in \text{supp}(X)} x \sum_{y \in \text{supp}(Y)} P(X = x, Y = y) + \sum_{y \in \text{supp}(Y)} y \sum_{x \in \text{supp}(X)} P(X = x, Y = y) \\
&= \sum_{x \in \text{supp}(X)} xP(X = x) + \sum_{y \in \text{supp}(Y)} yP(Y = y) \text{ by countable additivity, twice} \\
&= \mathbb{E}[X] + \mathbb{E}[Y]
\end{aligned}$$

The fourth equality above follows because $X = x$ and $Z = z$ if and only if $X = x$ and $Y = z - x$, and if $z - x \notin \text{supp}(Y)$ we have $P(X = x, Y = z - x) \leq P(Y = z - x) = 0$. Finally, we conclude $\mathbb{E}[aX + bY] = \mathbb{E}[aX] + \mathbb{E}[bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$ by applying item 2 followed by item 1. □

Corollary 11.6 (Monotonicity of expectation)**.** *If $P(X \geq Y) = 1$ then $\mathbb{E}[X] \geq \mathbb{E}[Y]$.*

Proof. If $P(X \geq Y) = 1$ then $X - Y \geq 0$ with probability 1, hence $\mathbb{E}[X - Y]$ is a sum of nonnegative numbers, thus nonnegative. The result follows by linearity. □

Proposition 11.7. *Suppose $X \sim \text{Bin}(n, p)$. Then $\mathbb{E}[X] = np$.*

Proof. X has the same distribution as $I_1 + \cdots + I_n$ where $I_1, \ldots, I_n$ are i.i.d. Bern(p) r.v.'s, so $\mathbb{E}[I_i] = p$ for each i. Then by linearity $\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[I_i] = np$. Note the independence is not needed. □

We will see several more examples of the power of linearity in the next lecture. For today, we conclude with one other tool that allows us to compute expectations of a transformation (function) of a random variable in terms of the distribution (PMF) of the *original* random variable.

Theorem 11.8 (LOTUS)**.** *Let $g : \mathbb{R} \to \mathbb{R}$ be a function and X be a discrete random variable with PMF p_X. Then $\mathbb{E}[g(X)] = \sum_{x \in \text{supp}(X)} g(x)p_X(x)$.*

Proof. Let $Y = g(X)$.

$$\begin{aligned}
\mathbb{E}[Y] &= \sum_{y \in \text{supp}(Y)} yP(Y = y) \\
&= \sum_{y \in \text{supp}(Y)} yP(X \in \{x \in \text{supp}(X) \mid g(x) = y\}) \\
&= \sum_{y \in \text{supp}(Y)} y \sum_{x \in \text{supp}(X), g(x) = y} P(X = x) \text{ by countable additivity} \\
&= \sum_{y \in \text{supp}(Y)} \sum_{x \in \text{supp}(X) : g(x) = y} g(x)P(X = x) \\
&= \sum_{x \in \text{supp}(X)} g(x)p_X(x)
\end{aligned}$$

where the last equality follows because $\text{supp}(Y) = \{g(x) \mid x \in \text{supp}(X)\}$, so that □

The preceding theorem is called the **law of the unconscious statistician (LOTUS)**, because it seems like something a statistician might do in their sleep: replace x in the definition of expectation with $g(x)$ but not touch the $p_X(x)$ part. However, it works!

Exercise 11.9. Suppose $X \sim \text{Pois}(\lambda)$. Find $\mathbb{E}[e^{-X}]$.

Solution. By LOTUS and the Taylor expansion of the exponential function, we have

$$\mathbb{E}[e^{-X}] = \sum_{x=0}^{\infty} e^{-x} e^{-\lambda} \frac{\lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\left(\frac{\lambda}{e}\right)^x}{x!} = \boxed{e^{\lambda(1/e-1)}}$$

□

12. Indicator random variables

Recall that for any event A, the corresponding indicator random variable I_A equals 1 whenever A is true, and 0 otherwise. Thus $I_A \sim \text{Bern}(P(A))$ which immediately implies the following property:

Proposition 12.1 (Fundamental bridge)**.** *For I_A the indicator random variable of an event A, we have $\mathbb{E}[I_A] = P(A)$.*

The fundamental bridge is incredibly powerful when paired with linearity. Frequently, some random variable X of interest can be written as a sum of indicators of various events, i.e. $X = \sum_{i=1}^{n} I_{A_i}$ for some events $A_1, \ldots, A_n$. These indicators are often not independent, which can make the full distribution of X challenging to derive. However, since linearity of expectation holds for any collection of random variables regardless of their dependence structure, we have

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[I_{A_i}] = \sum_{i=1}^{n} P(A_i)$$

Exercise 12.2. The numbers $1, \ldots, n$ are randomly ordered. Define a local maximum in the ordering to be any number in the ordering that is larger than the adjacent numbers. For instance, the number in the first position is a local maximum if it is larger than the number in second position, the number in the last position is a local maximum if it is larger than the number in the penultimate position, and all other numbers are local maxima if they are larger than both of the numbers adjacent. What is the average number of local maxima in the ordering?

Solution. Let X be the number of local maxima in the ordering. We can write $X = I_1 + \cdots + I_n$ where for each $i = 1, \ldots, n$, $I_i = I_{A_i}$ is the indicator for the event A_i that the number in position i is a local maximum. By symmetry we have $P(A_1) = P(A_n) = 1/2$ while $P(A_i) = 1/3$ for all $i = 2, \ldots, n-1$. Hence by linearity and the fundamental bridge,

$$\mathbb{E}[X] = \sum_{i=1}^{n} P(A_i) = 1 + \frac{1}{3}(n-2) = \boxed{\frac{n+1}{3}}$$

□

Exercise 12.3. Thomas draws cards one at a time, without replacement, from a standard deck of 52 cards. On average, how many cards does he draw before getting an ace for the first time?

Solution. Number the non-ace cards from 1 through 48, and let $I_i = I_{A_i}$ be the indicator of the event A_i that the i-th non-ace card appears in the deck before the first ace. Note $P(A_i) = 1/5$ since by symmetry, all five orderings of the four aces and the i-th non-ace card are equally likely. But then $X = \sum_{i=1}^{48} I_i$ is the number of cards drawn before getting an ace for the first time, so by linearity and the fundamental bridge

$$\mathbb{E}[X] = \sum_{i=1}^{n} P(A_i) = \boxed{\frac{48}{5}}$$

Another solution is to create indicators $J_1, \ldots, J_{52}$ where J_i is the indicator that cards $1, \ldots, i$ are non-aces. Then $\mathbb{E}[J_i] = \frac{48 \cdot \cdots \cdot (48-i+1)}{52 \cdot \cdots \cdot (52-i+1)}$ for $i = 1, \ldots, 48$ (with $J_{49} = \cdots = J_{52} = 0$). This will give the right answer but the sum $\sum_{i=1}^{48} \mathbb{E}[J_i]$ is much harder to compute without a calculator/computer. □

Indicator random variables will also allow us to compute the mean of geometric, hypergeometric, and negative binomial random variables.

Theorem 12.4. *Suppose $X \sim \text{Geom}(p)$, $Y \sim \text{NBin}(r, p)$, and $Z \sim \text{HGeom}(w, b, n)$. Then*

$$\mathbb{E}[X] = \frac{1-p}{p}, \quad \mathbb{E}[Y] = \frac{nw}{w+b}, \quad \mathbb{E}[Z] = \frac{r(1-p)}{p}$$

Proof. As in the definition of the Geometric distribution, we can write $X = \sum_{i=1}^{\infty} I_i$ where for each $i \geq 1$, $I_i = I_{A_i}$ is the indicator of the event A_i that the first i flips are all tails. Since the coin flips are independent, we have $P(A_i) = (1-p)^i$, so by linearity and the fundamental bridge

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} P(A_i) = \sum_{i=1}^{\infty} (1-p)^i = \frac{1-p}{p}$$

by the formula for the sum of a geometric series.

Since we can represent Y as the sum of r (independent) $\text{Geom}(p)$ random variables, linearity immediately implies $\mathbb{E}[Z] = \frac{r(1-p)}{p}$.

Finally, as in the definition of the Hypergeometric, for $i = 1, \ldots, n$ let I_i be the indicator of the event A_i that the i-th ball drawn is white. Then $Z = \sum_{i=1}^{n} I_i$. As $P(A_i) = \frac{w}{w+b}$ for all i, by linearity and the fundamental bridge

$$\mathbb{E}[Z] = \sum_{i=1}^{n} P(A_i) = \frac{nw}{w+b}$$

□

Remark. If the cards in Exercise 12.3 were drawn with replacement, then we'd have $X \sim \text{Geom}(1/13)$ and $\mathbb{E}[X] = 12$ by the previous theorem, which is larger. On the other hand, if in the Hypergeometric story of the previous proof, we drew the n balls with replacement, the number of white balls would be $\text{Bin}(n, w/w+b)$ and hence have the same mean as $Z \sim \text{HGeom}(w, b, n)$. To summarize: In a fixed number of draws from a bifurcated population, the expected number of individuals of one type is the same whether sampling is done with or without replacement. But when drawing until the *first* individual of that type, on average it takes fewer draws to do so when sampling without replacement than when sampling with replacement.

Example 12.5. In the setting of the Birthday problem (Exercise 2.2), what is the expected number of distinct birthdays of the k people? What is the expected number of pairs of people with the same birthday?

Solution. Let X be the number of distinct birthdays. We write $X = \sum_{i=1}^{365} I_i$ where I_i is the indicator of the event A_i that someone in the room has a birthday on day i of the year. We have

$$\begin{aligned} P(A_i^c) &= P(\cap_{j=1}^{k} \text{ person } j \text{ does not have birthday on day } i) \\ &= \left(\frac{364}{365}\right)^k \text{ as birthdays are assumed independent} \end{aligned}$$

Hence by linearity and the fundamental bridge

$$\mathbb{E}[X] = \sum_{i=1}^{k} P(A_i) = \boxed{365\left(1 - \left(\frac{364}{365}\right)^k\right)}$$

Now let Y be the number of pairs of people with the same birthday. There are $\binom{k}{2}$ pairs of people in the room. Changing notation a bit to not conflict with the previous part, let J_i be the indicator of the event B_i that the i-th pair has the same birthday. We have $P(B_i) = 1/365$ and $Y = \sum_{i=1}^{\binom{k}{2}} J_i$, so

$$\mathbb{E}[Y] = \sum_{i=1}^{\binom{k}{2}} P(B_i) = \boxed{\frac{\binom{k}{2}}{365}}$$

□

13. Variance

Expectation is a measure of the *average* value of a random variable. Another fundamental aspect of a distribution we might care about is its variability about its average value. The concept of variance quantifies this.

Definition 13.1. For a random variable X, its **variance** is given by $\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$, and its **standard deviation** is $\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}$.

Remark. The random variable $(X - \mathbb{E}[X])^2$ is the squared difference between X and its mean $\mathbb{E}[X]$ (which is non-random, i.e. a number). Thus, $\mathrm{Var}(X)$ can be viewed as the "average squared deviation" of X from its mean.

The definition of variance is a bit clumsy to work with. It is often much easier to use the following identity.

Proposition 13.2. *For any random variable X, we have* $\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.

Proof. We compute

$$\begin{aligned}
\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] &= \mathbb{E}[X^2 - 2X\mathbb{E}[X] + (\mathbb{E}[X])^2] \\
&= \mathbb{E}[X^2] - \mathbb{E}[2X\mathbb{E}[X]] + \mathbb{E}[(\mathbb{E}[X])^2] \text{ by linearity of expectation} \\
&= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + (\mathbb{E}[X])^2 \text{ by linearity again} \\
&= \mathbb{E}[X^2] - (\mathbb{E}[X])^2
\end{aligned}$$

□

Example 13.3. If $X \sim \mathrm{Bern}(p)$, then $\mathrm{Var}(X) = p(1-p)$. To see this, by LOTUS we have $\mathbb{E}[X^2] = p \cdot 1^2 + (1-p) \cdot 0^2 = p$. Recalling $\mathbb{E}[X] = p$ we have $\mathrm{Var}(X) = p - p^2 = p(1-p)$ as desired.

Example 13.4 (Poisson variance)**.** Suppose $X \sim \text{Pois}(\lambda)$. Then

$$\begin{aligned}
\mathbb{E}[X^2] &= \sum_{x=0}^{\infty} x^2 \exp(-\lambda)\frac{\lambda^x}{x!} \\
&= \lambda \exp(-\lambda) + \sum_{x=2}^{\infty} x^2 \exp(-\lambda)\frac{\lambda^x}{x!} \\
&= \lambda \exp(-\lambda) + \sum_{x=2}^{\infty} \exp(-\lambda) x \frac{\lambda^x}{(x-1)!} \\
&= \lambda \left[\exp(-\lambda) + \sum_{x=1}^{\infty} \exp(-\lambda)(x+1)\frac{\lambda^x}{x!} \right] \\
&= \lambda[\exp(-\lambda) + \mathbb{E}[X] + (1 - \exp(-\lambda))] = \lambda(\lambda+1)
\end{aligned}$$

Thus $\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \lambda(\lambda+1) - \lambda^2 = \boxed{\lambda}$.

Example 13.5 (Geometric variance)**.** Computing the geometric variance requires some wizardry. Let $X \sim \text{Geom}(p)$; then the fact that the PMF of a Geometric distribution sums to 1 yields

$$1 = \sum_{x=0}^{\infty} (1-p)^x p = p + (1-p)p + \sum_{x=2}^{\infty} (1-p)^x p$$

Taking the derivative with respect to p on both sides, we obtain

$$0 = 2 - 2p - \sum_{x=2}^{\infty} x(1-p)^{x-1} p + \sum_{x=2}^{\infty} (1-p)^x$$

Taking another derivative yields

$$\begin{aligned}
0 &= -2 + \sum_{x=2}^{\infty} x(x-1)(1-p)^{x-2} p - \sum_{x=2}^{\infty} x(1-p)^{x-1} - \sum_{x=2}^{\infty} x(1-p)^{x-1} \\
&= -2 + \sum_{x=0}^{\infty} (x^2 + 3x + 2)(1-p)^x p - 2\sum_{x=0}^{\infty} (x+2)(1-p)^{x+1} \\
&= -2 + \mathbb{E}[X^2 + 3X + 2] - \frac{2(1-p)}{p}\mathbb{E}[X+2]
\end{aligned}$$

With $\mathbb{E}[X] = (1-p)/p$ as computed above, we conclude

$$\mathbb{E}[X^2] = \frac{1-p}{p}\left(4 - 3 + \frac{2(1-p)}{p}\right) = \frac{(1-p)(2-p)}{p^2}$$

and

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{(1-p)(2-p)}{p^2} - \frac{(1-p)^2}{p^2} = \boxed{\frac{1-p}{p^2}}$$

Here are some important properties of the variance:

Theorem 13.6 (Properties of variance)**.** *Suppose $X_1, \ldots, X_n$ are independent random variables. Then for any constant c, we have*

1. $\mathrm{Var}(X_1 + c) = \mathrm{Var}(X_1)$
2. $\mathrm{Var}(cX_1) = c^2\mathrm{Var}(X_1)$
3. $\mathrm{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathrm{Var}(X_i)$

Proof. We only prove the first two identities; the third will be proven later in the course. That first identity follows from the decomposition $\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$:

$$\mathrm{Var}(cX_1) = \mathbb{E}[(cX_1)^2] - (\mathbb{E}[cX_1])^2 = c^2\mathbb{E}[X_1^2] - (c\mathbb{E}[X_1])^2 = c^2\mathrm{Var}(X_1)$$

The second is easiest to show by the definition of variance. Let $Y = X_1 + c$, so

$$\mathrm{Var}(Y) = \mathbb{E}[(Y - \mathbb{E}[Y])^2] = \mathbb{E}[([X_1 + c] - \mathbb{E}[X_1 + c])^2] = \mathbb{E}[(X_1 - \mathbb{E}[X_1])^2] = \mathrm{Var}(X_1)$$

□

Remark. The previous result shows a sort of "linearity of variance" for *independent* random variables. It does not hold for random variables with general dependence structure; for instance $\mathrm{Var}(X + X) = \mathrm{Var}(2X) = 4\mathrm{Var}(X)$ which is not the same as $\mathrm{Var}(X) + \mathrm{Var}(X) = 2\mathrm{Var}(X)$.

Example 13.7 (Binomial variance)**.** Since a $\mathrm{Bin}(n, p)$ random variable has the same distribution as the sum of n independent $\mathrm{Bern}(p)$ random variables, if $X \sim \mathrm{Bin}(n, p)$ then $\mathrm{Var}(X) = \boxed{np(1-p)}$. We can view X/n as modeling the proportion of successes in n independent trials, each with probability p of success; note $\mathrm{Var}(X/n) = (1/n^2)\mathrm{Var}(X) = (p(1-p))/n$ gets smaller as n gets large.

Example 13.8 (Negative Binomial variance)**.** Since a $\mathrm{NBin}(r, p)$ random variable has the same distribution as the sum of r independent $\mathrm{Geom}(p)$ random variables, if $X \sim \mathrm{NBin}(r, p)$ then $\mathrm{Var}(X) = \boxed{\dfrac{r(1-p)}{p^2}}$.

Example 13.9 (First success variance)**.** If $X \sim \mathrm{FS}(p)$, then $X - 1 \sim \mathrm{Geom}(p)$. Thus $\mathrm{Var}(X) = \mathrm{Var}(X - 1) = \boxed{\dfrac{1-p}{p^2}}$.

14. Continuous random variables: PDFs and the Uniform

Definition 14.1. A random variable X is **continuous** if its CDF $F_X(x) = P(X \leq x)$ is continuous everywhere, and differentiable at all but countably many points. In that case, $f_X(x) = F'_X(x)$ is called the **probability density function (PDF)** of X, and the support of X, supp(X), is defined as the interior of $\{x \mid f(x) > 0\}$. [1]

Proposition 14.2. *If X is continuous, then $P(X = a) = 0$ for any constant a.*

Proof. As usual, let F_X be the CDF of X. Then

$$P(X = a) \leq P(a - h < x \leq a) = F_X(a) - F_X(a - h), \quad \text{for all } h > 0$$

The result follows by continuity of F_X upon letting $h \downarrow 0$. □

We can compute the CDF from the PDF by the fundamental theorem of calculus, and using the fact that $\lim_{x\to-\infty} F_X(x) = \lim_{x\to-\infty} P(X \leq x) = 0$.

Proposition 14.3. *If f_X is the PDF of a continuous random variable X, then its CDF F_X is given by*

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

Remark. It can be shown that Theorem 10.3 holds for continuous random variables as well. Then by Proposition 14.3, the PDF determines the distribution of a continuous random variable. So when asked to specify the distribution of a continuous random variable, it suffices to either specify the PDF or CDF.

[1] In this class, for simplicity we will assume the support of any continuous r.v. is an open interval, including possibly the entire real line.

Proposition 14.3 shows us how to compute the probability that X lies in any interval using integration:

$$P(X \in [a,b]) = \int_a^b f(x)dx$$

Note that it doesn't matter whether we include either of the endpoints a or b in the interval, by Proposition 14.2. It also shows that PDFs have to integrate to 1: $\int_{-\infty}^{\infty} f_X(t)dt = 1$. We also remark without proof that we can generalize the preceding display to

$$\mathcal{L}_X(B) = P(X \in B) = \int_B f(x)dx$$

for all $B \subseteq \mathbb{R}$. The preceding display shows indeed that the PDF determines the distribution.

Remark. While a PMF of a discrete random variable evaluated at any given value x is the probability of an event, PDFs should *not* be interpreted as probabilities. For one thing, PDFs can be larger than 1, as we will see below. The proper way to interpret PDFs is to note that for small $\Delta > 0$ we have

$$\frac{P(X \in [x, x+\Delta])}{\Delta} \approx f_X(x)$$

We now provide our first example of a continuous distribution family.

Definition 14.4. The **Uniform** distribution with parameters $a < b$, abbreviated as $\text{Unif}(a,b)$, is the continuous distribution supported on $[a,b]$ whose PDF is constant on that interval. That is, $X \sim \text{Unif}(a,b)$ then $f_X(x) = (b-a)^{-1}$ for all $a < x < b$ (with $f_X(x) = 0$ otherwise).

Remark. If $X \sim \text{Unif}(a,b)$, the constant PDF $(b-a)^{-1}$ can be easily derived from the requirement that PDFs integrate to 1. Also, by Proposition 14.3, its CDF is

$$F_X(x) = \int_a^x (b-a)^{-1} = \frac{x-a}{b-a} \quad x \in (a,b)$$

Remark. The PDF of a $\text{Unif}(0,1/4)$ random variable is equal to 4 on the interval $(0,1/4)$, emphatically reminding us that a PDF does not directly correspond to a probability. To derive PDFs, it is often useful to start with the CDF and then differentiate, because the CDF corresponds to the probability of an actual event.

A scaled and shifted Uniform random variable is still Uniform, with appropriately scaled and shifted parameters:

Proposition 14.5. *Suppose $U \sim \text{Unif}(0,1)$. Then for any $a < b$, we have $(b-a)U + a \sim \text{Unif}(a,b)$.*

Proof. Let $X = (b-a)U + a$. Then for each $x \in [a, b]$ we have

$$F_X(x) = P((b-a)U + a \leq x) = P\left(U \leq \frac{x-a}{b-a}\right) = \frac{x-a}{b-a}$$

which matches the Unif(a, b) CDF. □

Uniform random variables are useful because of the following property which shows that a uniform random variable can be elegantly transformed into any other continuous random variable by means of a (deterministic) transformation:

Theorem 14.6 (Universality of the uniform)**.** *Suppose F is a continuous CDF that is strictly increasing on the support of the corresponding distribution. This implies the existence of the inverse function $F^{-1} : (0, 1) \to \mathbb{R}$, which is also increasing. Then if $U \sim$ Unif$(0, 1)$, $F^{-1}(U)$ is a random variable with CDF F. Conversely, if X is a random variable with CDF F, then $F(X) \sim$ Unif$(0, 1)$.*

Proof. First suppose $U \sim$ Unif$(0, 1)$ so its PDF is $f_U(u) = 1$ on $(0, 1)$. Then for any x we have

$$P(F^{-1}(U) \leq x) = P(U \leq F(x)) = \int_0^{F(x)} f_U(u)du = F(x)$$

where the first inequality uses the fact that F is increasing and thus can be applied to both sides of the inequality $F^{-1}(U) \leq x$. This shows that $F^{-1}(U)$ has CDF F. Conversely, if X has CDF F, then for each $u \in (0, 1)$ we have

$$P(F(X) \leq u) = P(X \leq F^{-1}(u)) = F(F^{-1}(u)) = u$$

where the first inequality uses the fact that F^{-1} is increasing. □

In the previous theorem, $F(X)$ and $F^{-1}(U)$ may seem like a strange objects — plugging random variables into CDFs and their inverses. But remember that F and F^{-1} are just (non-random) functions, so $F(X)$ and $F^{-1}(U)$ are just transformations of random variables, hence themselves random variables (recall Lecture 10). The utility of universality of the uniform in computation is that any continuous distribution (with known inverse CDF F^{-1}) can be simulated by simply first drawing a Unif$(0, 1)$ random variable, then applying F^{-1} to it.

We conclude by defining expectation for continuous random variables. It is analogous to the discrete definition, except the PMF is replaced with the PDF, and the sum replaced by an integral:

Definition 14.7. If X is a continuous random variable with PDF f_X, its expectation $\mathbb{E}[X]$ is

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx$$

It can be shown that linearity also holds for continuous random variables, along with the following statement of LOTUS, making it possible to compute the expectation of $g(X)$ using only the PDF of X. We do not prove this.

Theorem 14.8 (Continuous LOTUS)**.** *Let $g : \mathbb{R} \to \mathbb{R}$ be a function and X be a continuous random variable with PDF f_X. Then*

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

Example 14.9. Suppose $U \sim \text{Unif}(0, 1)$. Then using the definition of expectation and LOTUS:

$$\mathbb{E}[U] = \int_0^1 u du = \frac{1^2}{2} - \frac{0^2}{2} = \boxed{\frac{1}{2}}$$
$$\mathbb{E}[U^2] = \int_0^1 u^2 du = \frac{1^3}{3} - \frac{0^3}{3} = \frac{1}{3}$$

Hence

$$\text{Var}(U) = \mathbb{E}[U^2] - (\mathbb{E}[U])^2 = \frac{1}{3} - \left(\frac{1}{2}\right)^2 = \boxed{\frac{1}{12}}$$

by Proposition 13.2. Also, by Proposition 14.5, for $X \sim \text{Unif}(a, b)$ we know X has the same distribution as $(b - a)U + a$, so by the properties of expectation and variance we have

$$\mathbb{E}[X] = \frac{b-a}{2} + a = \frac{a+b}{2}, \qquad \text{Var}(X) = \frac{(b-a)^2}{12}$$

15. The Normal distribution

One of the most ubiquitous continuous distribution families is the Normal distribution. The reason the Normal distribution family arises so frequently is due to something called the central limit theorem, which we will study at the end of the course. For now, we define the Normal distribution family by the PDF and study some of its mathematical properties.

Definition 15.1 (Normal distribution)**.** A random variable Z has the **Normal distribution** with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$, abbreviated as $Z \sim \mathcal{N}(\mu, \sigma^2)$, if its PDF is given by

$$f_Z(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right)$$

for $z \in \mathbb{R}$. The $\mathcal{N}(0,1)$ distribution is called the **standard Normal distribution**.

Definition 15.2. Let φ denote the standard Normal PDF and let Φ denote the standard Normal CDF:

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right), \quad \Phi(z) = \int_{-\infty}^{z} \varphi(t)dt$$

Note that there is no closed-form expression for Φ.

Figure 15.1 plots φ and Φ. The shape of the standard Normal PDF φ is the reason the normal distribution is often called a "bell curve." The functions φ and Φ have some useful symmetry properties:

Proposition 15.3. *For all $z \in \mathbb{R}$, we have $\varphi(z) = \varphi(-z)$ and $\Phi(z) = 1 - \Phi(-z)$.*

Proof. That $\varphi(z) = \varphi(-z)$ is immediate from the definition. Next

$$\Phi(z) = \int_{-\infty}^{z} \varphi(x)dx = \int_{-\infty}^{z} \varphi(-x)dx = \int_{-z}^{\infty} \varphi(y)dy = 1 - \Phi(-z)$$

by the substitution $y = -x$. □

An important property of the normal distribution family is that it is closed under translation and scalar multiplication. To show this, we begin by writing the CDF and PDF of a translated and scaled random variable in terms of the CDF and PDF of the original random variable.

Proposition 15.4. *Let X be a random variable with CDF F. Then for any $\mu \in \mathbb{R}$ and $\sigma > 0$, the CDF of $Y = \mu + \sigma X$ is given by $F_Y(y) = F\left(\frac{y-\mu}{\sigma}\right)$. Furthermore, if X is continuous with PDF f, then Y is continuous with PDF $f_Y(y) = \frac{1}{\sigma} f\left(\frac{y-\mu}{\sigma}\right)$.*

Proof. By the definition of the CDF, we have

$$F_Y(y) = P(Y \leq y) = P(\mu + \sigma X \leq y) = P\left(X \leq \frac{y-\mu}{\sigma}\right) = F\left(\frac{y-\mu}{\sigma}\right)$$

Note we used the fact that $\sigma > 0$ to ensure that $\mu + \sigma X \leq y$ is equivalent to $X \leq \frac{y-\mu}{\sigma}$. The second assertion $f_Y(y) = \frac{1}{\sigma} f\left(\frac{y-\mu}{\sigma}\right)$ follows by differentiating both sides of the preceding display. □

Exercise 15.5. If $Z \sim \mathcal{N}(0,1)$, then $Y = \mu + \sigma Z \sim \mathcal{N}(\mu, \sigma^2)$ for any $\mu \in \mathbb{R}$, $\sigma \neq 0$.

Solution. First suppose $\sigma > 0$. Then by the previous proposition, Y has PDF

$$f_Y(y) = \frac{1}{\sigma}\varphi\left(\frac{y-\mu}{\sigma}\right) = \frac{1}{\sigma} \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\left(\frac{y-\mu}{\sigma}\right)^2/2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

for all $y \in \mathbb{R}$. As this is the PDF of a $\mathcal{N}(\mu, \sigma^2)$ random variable, the result follows since the PDF determines the distribution.

If $\sigma < 0$, first we note that $-Z \sim \mathcal{N}(0,1)$, since

$$P(-Z \leq z) = P(Z \geq -z) = 1 - \Phi(-z) = \Phi(z)$$

by Proposition 15.3. Then $Y = \mu + (-\sigma)(-Z)$ and the result follows from the case $\sigma > 0$. □

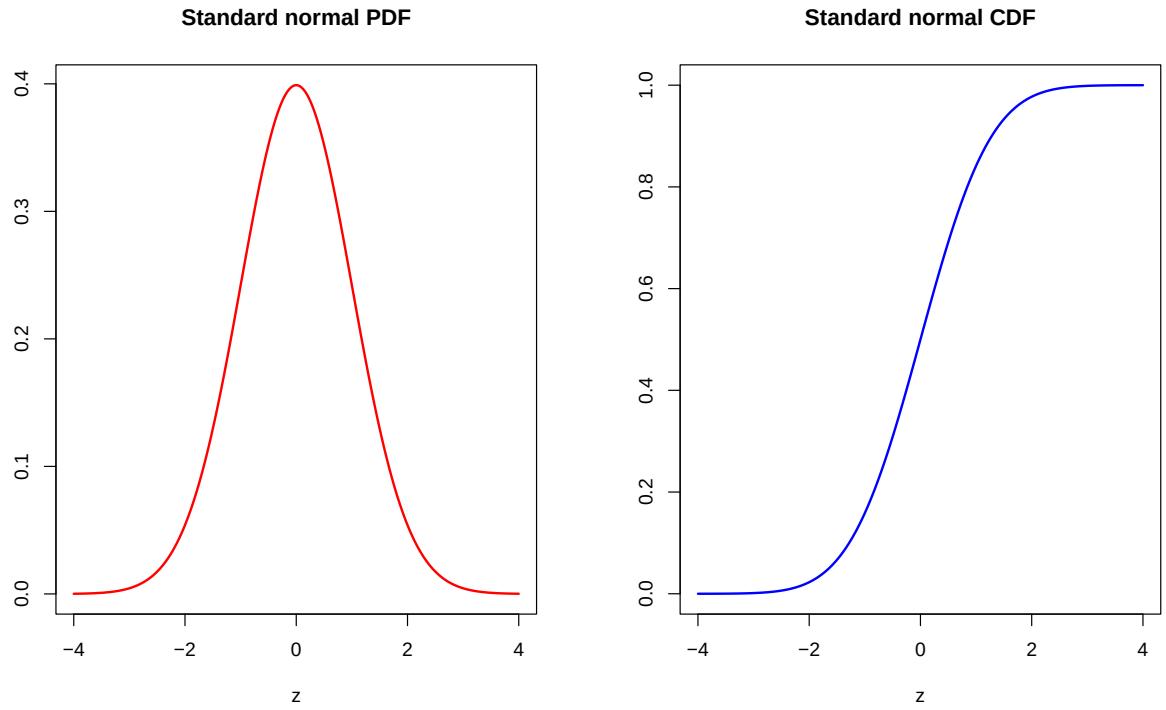
The Normal distribution family has two parameters μ and σ^2, which we called the mean and variance, respectively. It's a useful exercise to verify that a random variable $Z \sim \mathcal{N}(\mu, \sigma^2)$ indeed has mean μ and variance σ^2:

Proposition 15.6. *Suppose $Z \sim \mathcal{N}(\mu, \sigma^2)$ for any $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Then $\mathbb{E}[Z] = \mu$ and $\mathrm{Var}(Z) = \sigma^2$.*

Proof. Take $Y \sim \mathcal{N}(0,1)$. By definition of expectation and the fundamental theorem of calculus, we have

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} y\varphi(y)dy = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y \exp\left(-\frac{y^2}{2}\right) dy = \lim_{L\to\infty} -\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)\Big|_{y=-L}^{y=L} = 0$$

Figure 15.1: The standard Normal PDF φ (left) and the standard Normal CDF Φ (right) on the interval $[-4, 4]$:



By LOTUS we have

$$\begin{aligned}\mathbb{E}[Y^2] = \int_{-\infty}^{\infty} y^2 \varphi(y) dy &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 \exp\left(-\frac{y^2}{2}\right) dy \\ &= \frac{1}{\sqrt{2\pi}} \lim_{L \to \infty} \left[-y \exp\left(-\frac{y^2}{2}\right) \Big|_{y=-L}^{y=L} + \int_{-\infty}^{\infty} \exp\left(-\frac{y^2}{2}\right) dy \right] \\ &= \frac{1}{\sqrt{2\pi}} \cdot \sqrt{2\pi} = 1\end{aligned}$$

where the penultimate equality uses integration by parts. Since Z has the same distribution as $\mu + \sigma Y$ by the previous exercise, we conclude by linearity and Theorem 13.6 that

$$\mathbb{E}[Z] = \mathbb{E}[\mu + \sigma Y] = \mu + \sigma \mathbb{E}[Y] = \mu$$

$$\text{Var}(Z) = \text{Var}(\mu + \sigma Y) = \sigma^2 \text{Var}(Y) = \sigma^2$$

□

We end with an example of computing probabilities involving normal random variables, leaving final answers in terms of Φ. The key insight is that Exercise 15.5 implies that if $Y \sim \mathcal{N}(\mu, \sigma^2)$, then $Z = \frac{Y-\mu}{\sigma} \sim \mathcal{N}(0, 1)$. In general, taking a random variable, subtracting its mean, and dividing by

its standard deviation gives a random variable that has mean 0 and variance 1, by the properties of mean and variance. We've shown above, however, that standardization keeps a Normally distributed random variable within the Normal family, which will not generally be true for other distribution families.

Exercise 15.7. Suppose $X \sim \mathcal{N}(-1, 4)$. Compute $P(|X| \leq 3)$ in terms of Φ.

Solution. The key trick is "standardization"; with $X \sim \mathcal{N}(\mu, \sigma^2)$ we know from Exercise 15.5 that $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$. With $\mu = -1$, $\sigma = 2$ we have

$$\begin{aligned} P(|X| \leq 3) &= P(X \leq 3) - P(X < -3) \\ &= P\left(Z \leq \frac{3-\mu}{\sigma}\right) - P\left(Z < \frac{-3-\mu}{\sigma}\right) \\ &= P(Z \leq 2) - P(Z < -1) \\ &= \boxed{\Phi(2) - \Phi(-1)} \approx 0.819 \end{aligned}$$

Alternatively one could use Proposition 15.4 directly. However, the method above avoids having to memorize Proposition 15.4. □

16. The Exponential distribution

Today, we will study another important continuous distribution family known as the Exponential.

Definition 16.1. A random variable X has an **Exponential distribution** with rate $\lambda > 0$, abbreviated $X \sim \text{Expo}(\lambda)$, if it has PDF f_X given by

$$f_X(x) = \lambda \exp(-\lambda x), \quad x > 0$$

As in our discussion of the normal distribution, we begin with a few nice properties of the exponential:

Proposition 16.2. *If $X \sim \text{Expo}(\lambda)$ and $Y = cX$ for some $c > 0$, then $Y \sim \text{Expo}(\lambda/c)$.*

Proof. By Proposition 15.4, for all $y > 0$ the PDF of Y is given by

$$f_Y(y) = \frac{1}{c} f_X\left(\frac{y}{c}\right) = \frac{\lambda}{c} \exp\left(-\frac{\lambda}{c} y\right)$$

which matches the PDF of an $\text{Expo}(\lambda/c)$ distribution. □

Now we compute the CDF, mean, and variance of an exponential distribution. This provides some additional practice of the concepts developed in the previous lectures.

Proposition 16.3. *If $X \sim \text{Expo}(\lambda)$, then $\mathbb{E}[X] = \lambda^{-1}$ and $\text{Var}(X) = \lambda^{-2}$, and the CDF of X is given by $F_X(x) = 1 - \exp(-\lambda x)$ for all $x \geq 0$.*

Proof. First we use integration by parts to compute some antiderivatives:

$$\int x \exp(-x) = -x \exp(-x) + \int \exp(-x) dx = -(1 + x) \exp(-x) + C$$
$$\int x^2 \exp(-x) = -x(1 + x) \exp(-x) + \int (1 + x) \exp(-x) dx = -(x^2 + 2x + 2) \exp(-x) + C$$

Then by the definition of expectation and LOTUS, for $Y \sim \text{Expo}(1)$ we have

$$\mathbb{E}[Y] = \int_0^\infty y \exp(-y) dy = \lim_{L\to\infty} -(1+y)\exp(-y)\Big|_{y=0}^{y=L} = 1$$
$$\mathbb{E}[Y^2] = \int_0^\infty y^2 \exp(-y) dy = \lim_{L\to\infty} -(y^2+2y+2)\exp(-y)\Big|_{y=0}^{y=L} = 2$$

So by Proposition 13.2, we have $\text{Var}(Y) = \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2 = 2 - 1^2 = 1$. As X has the same distribution as Y/λ by the previous proposition, by linearity and Theorem 13.6 we have

$$\mathbb{E}[X] = \mathbb{E}\left[\frac{Y}{\lambda}\right] = \lambda^{-1}\mathbb{E}[Y] = \lambda^{-1}$$
$$\text{Var}(X) = \text{Var}\left(\frac{Y}{\lambda}\right) = \lambda^{-2}\text{Var}(Y) = \lambda^{-2}$$

Finally, to compute the CDF we use Proposition 14.3 so that for all $x > 0$, we have

$$F_X(x) = \int_0^x \lambda \exp(-\lambda t) dt = -\exp(-\lambda t)\Big|_{t=0}^{t=x} = 1 - \exp(-\lambda x)$$

□

What distinguishes the Exponential distribution is something called the *memorylessness* property. The intuition is as follows: suppose the waiting time for a bus in a dysfunctional city "Blotchville" follows an Expo(λ) distribution. Then conditional on having waited x minutes already, the distribution of the *additional* time that you need to wait is still Expo(λ).

Theorem 16.4 (Memorylessness)**.** *Suppose $X \sim \text{Expo}(\lambda)$. Then for all $s, t > 0$ we have*

$$P(X \geq t+s \mid X \geq s) = P(X \geq t)$$

Proof. We compute

$$\begin{aligned}
P(X \geq t+s \mid X \geq s) &= \frac{P(X \geq t+s, X \geq s)}{P(X \geq s)} \text{ by the definition of conditional probability} \\
&= \frac{P(X \geq t+s)}{P(X \geq s)} \\
&= \frac{1 - F_X(t+s)}{1 - F_X(s)} \\
&= \frac{\exp(-\lambda(t+s))}{\exp(-\lambda s)} \\
&= \exp(-\lambda t) = 1 - F_X(t) = P(X \geq t)
\end{aligned}$$

□

It turns out that the Exponential distributions are the only continuous distributions that are memoryless. Meanwhile, the Geometric distributions are the only discrete distributions that are memoryless. So the Exponential can be viewed as a continuous version of the Geometric. Indeed, the following result makes this connection more explicit. Recall $\lfloor x \rfloor$ denotes the largest integer that is smaller than or equal to a real number x, while $\lceil x \rceil$ is the smallest integer that is larger or equal to x.

Proposition 16.5. *Suppose* $X \sim \text{Expo}(\lambda)$. *Then* $\lfloor X \rfloor \sim \text{Geom}(1 - \exp(-\lambda))$ *and* $\lceil X \rceil \sim \text{FS}(1 - \exp(-\lambda))$.

Proof. Let $Y = \lfloor X \rfloor$. Clearly $\text{supp}(Y) = \{0, 1, \dots, \}$. For each $y \in \text{supp}(Y)$, we have

$$\begin{aligned} P(Y = y) = P(y \leq X < y + 1) = F_X(y + 1) - F_X(y) &= (1 - \exp(-\lambda(y + 1))) - (1 - \exp(-\lambda y)) \\ &= \exp(-\lambda y)(1 - \exp(-\lambda)) \end{aligned}$$

which matches the PMF of the $\text{Geom}(1 - \exp(-\lambda))$ distribution. Next note that $\lceil X \rceil = \lfloor X \rfloor + 1$ unless X is an integer. But $P(X \text{ is an integer}) = 0$ since X is continuous, so indeed $\lceil X \rceil$ has the same distribution as $\lfloor X \rfloor + 1$. Formally, this shows that for any set $B \subseteq \mathbb{R}$ we have

$$\begin{aligned} \mathcal{L}_{\lceil X \rceil}(B) = P(\lceil X \rceil \in B) &= P(\lceil X \rceil \in B \cap A) + P(\lceil X \rceil \in B \cap A^c) \\ &= 0 + P(\lfloor X \rfloor + 1 \in B) = \mathcal{L}_{\lfloor X \rfloor + 1}(B) \end{aligned}$$

where A is the event that X is an integer. □

Proposition 16.6 (Minimum of independent Exponentials)**.** *Suppose* $X_1, \dots, X_n$ *are independent with* $X_i \sim \text{Expo}(\lambda_i)$ *for* $i = 1, \dots, n$. *Then* $X = \min(X_1, \dots, X_n) \sim \text{Expo}(\sum_{i=1}^n \lambda_i)$.

Proof. For each $x > 0$ we have

$$\begin{aligned} P(X \leq x) &= P\left(\cup_{i=1}^n \{X_i \leq x\}\right) \\ &= 1 - P\left(\cap_{i=1}^n \{X_i > x\}\right) \\ &= 1 - \prod_{i=1}^n P(X_i > x) \text{ by independence} \\ &= 1 - \prod_{i=1}^n \exp(-\lambda_i x) \\ &= 1 - \exp\left(-\sum_{i=1}^n \lambda_i x\right) \end{aligned}$$

which matches the CDF of an $\text{Expo}\left(\sum_{i=1}^n \lambda_i\right)$ random variable (Proposition 16.3). □

17. Joint and marginal distributions

Up until now, we have studied the distribution of one random variable at a time. Today, we will discuss the distribution of two (or more) random variables *simultaneously*, known as a *joint* distribution. To simplify the exposition, we'll focus on the case of two r.v.'s, though the concepts extend generally to any (finite) number of random variables.

Definition 17.1. The **joint CDF** of two random variables X and Y is the function $F_{XY} : \mathbb{R}^2 \to [0, 1]$ specifying

$$F_{XY}(x, y) = P(X \leq x, Y \leq y)$$

Just as the CDF determines the distribution of a single r.v., it can be shown that the joint CDF determines the *joint distribution* $\mathcal{L}_{XY}$ given by $\mathcal{L}_{XY}(B) = P((X, Y) \in B)$ for all subsets $B \subseteq \mathbb{R}^2$. Joint distributions can be also specified by a joint PMF or PDF:

Definition 17.2. If X and Y are discrete, their **joint PMF** is the function $p_{XY} : \mathbb{R}^2 \to [0, 1]$ specifying $p_{XY}(x, y) = P(X = x, Y = y)$. If X and Y are continuous, their **joint PDF** is the function $f_{XY} : \mathbb{R}^2 \to [0, \infty)$ with $f_{XY}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{XY}(x, y)$. The **support** $\text{supp}(X, Y)$ of (X, Y) is the set of all $(x, y) \in \mathbb{R}^2$ where the joint PMF/PDF is strictly positive.

If X and Y are discrete, we convert the joint PMF to the joint CDF using countable additivity:

$$\begin{aligned} F_{XY}(x, y) = P(X \leq x, Y \leq y) &= \sum_{x_0 \in (-\infty, x] \cap \text{supp}(X)} \sum_{y_0 \in (-\infty, y] \cap \text{supp}(Y)} P(X = x_0, Y = y_0) \\ &= \sum_{x_0 \in (-\infty, x] \cap \text{supp}(X)} \sum_{y_0 \in (-\infty, y]} p_{XY}(x_0, y_0) \end{aligned}$$

Similarly, we can convert from the joint PDF to the joint CDF by the (multivariate) fundamental theorem of calculus:

$$F_{XY}(x, y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(a, b) dbda$$

This generalizes to general sets $B \subseteq \mathbb{R}^2$:

$$\mathcal{L}_{XY}(B) = P((X,Y) \in B) = \int\int_B f_{XY}(x,y)dydx$$

Thus, the joint PMF/PDF determine the joint distribution as well.

The joint distribution of two random variables, in turn, determines the distribution of each component random variable, known as the **marginal** distribution:

Proposition 17.3 (Marginalization)**.** *If X and Y are discrete with joint PMF p_{XY}, the (marginal) PMF p_X of X is given by*

$$p_X(x) = \sum_{y \in \text{supp}(Y)} p_{XY}(x,y)$$

If X and Y are continuous with joint PDF f_{XY}, the (marginal) PDF f_X of X is given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)dy$$

Proof. The discrete case follows immediately from LOTP:

$$p_X(x) = P(X = x) = \sum_{y \in \text{supp}(Y)} P(X = x, Y = y) = \sum_{y \in \text{supp}(Y)} p_{XY}(x,y)$$

The continuous case follows from first considering the CDF:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x}\int_{-\infty}^{\infty} f_{XY}(a,y)dyda$$

Differentiate both sides and apply the fundamental theorem of calculus. □

Example 17.4 (Uniform)**.** Suppose U is uniform on some region $B \subseteq \mathbb{R}^2$, meaning it is supported on B and has a constant joint PDF on B. Let c be this constant; then we must have $1 = P(U \in B) = \int\int_B c dydx$ so we must have $c = |B|^{-1}$ where $B = \int\int_B 1 dydx$ is the area of B. Then for any $S \subseteq B$ we have

$$P(U \in S) = \int\int_S |B|^{-1} dydx = \frac{|S|}{|B|}$$

Example 17.5. Suppose $U = (X,Y)$ is uniform on the unit circle in $\mathbb{R}^2$. Then the marginal density of Y is

$$f_Y(y) = \int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} \pi^{-1} dx = \boxed{\frac{2}{\pi}\sqrt{1-y^2}}, \quad -1 < y < 1$$

and by symmetry, the marginal density of X is $f_X(x) = \boxed{\frac{2}{\pi}\sqrt{1-x^2}}, \quad -1 < x < 1$

Proposition 17.3 shows that the joint distribution determines the marginal distributions. However, marginal distributions do not determine the joint distribution. For example, suppose we know both X and Y (marginally) both have a Bern(0.5) distribution. We could have $X = Y$, in which case $P(X = 1, Y = 0) = 0$. Conversely, we could have $Y = 1 - X$, in which case $P(X = 1, Y = 0) = P(X = 1) = 1/2$.

However, if we know the marginal distributions *and* that the random variables are *independent*, then we have the joint distribution. The easiest way to see this is to note that when X and Y are independent, their joint CDF is determined by the marginal CDFs by Proposition 10.5:

$$F_{XY}(x, y) = F_X(x)F_Y(y)$$

When X and Y are continuous, taking partial derivatives with respect to x and y in the preceding display also reveals the joint PDF is determined by the marginal PDFs:

$$f_{XY}(x, y) = f_X(x)f_Y(y)$$

Thus, one strategy to show two continuous random variables X and Y are independent is to show their joint PDF $f_{XY}(x, y)$ factors into a product $f_X(x)f_Y(y)$. Analogously, we already know that we can show discrete X and Y are independent if their joint PMF factors into a product of marginal PMFs (Proposition 10.5).

Exercise 17.6. Suppose $T_1 \sim \text{Expo}(\lambda_1)$ and $T_2 \sim \text{Expo}(\lambda_2)$ are independent. Find $P(T_1 < T_2)$.

Solution. We integrate the joint PDF of (T_1, T_2), which is the product of the marginal PDFs:

$$\begin{aligned} P(T_1 < T_2) &= \int_0^\infty \int_x^\infty \lambda_1\lambda_2 \exp(-\lambda_1 x) \exp(-\lambda_2 y) dy dx \\ &= \int_0^\infty (\lambda_1 \exp(-\lambda_1 x)) \left(\lim_{L\to\infty} -\exp(-\lambda_2 y)\Big|_{y=x}^{y=L} \right) dx \\ &= \int_0^\infty \lambda_1 \exp(-(\lambda_1 + \lambda_2)x) dx \\ &= \boxed{\frac{\lambda_1}{\lambda_1 + \lambda_2}} \end{aligned}$$

where the fastest way to see the last line is to note the $\text{Expo}(\lambda_1 + \lambda_2)$ PDF integrates to 1. □

We conclude by noting independent and identically distributed (i.i.d.) random variables satisfy a nice symmetry property:

Definition 17.7. Random variables $X_1, \ldots, X_n$ are **exchangeable** if $(X_1, \ldots, X_n)$ has the same joint distribution as $(X_{\pi(1)}, \ldots, X_{\pi(n)})$ for all permutations π of $\{1, \ldots, n\}$.

Proposition 17.8. *If $X_1, \ldots, X_n$ are i.i.d., then they are exchangeable.*

Proof. Let F be the common marginal CDF of the X_i. By Proposition 10.5, for any permutation π, the joint CDF of $(X_{\pi(1)}, \ldots, X_{\pi(n)})$ is given by

$$P(X_{\pi(1)} \leq x_1, \ldots, X_{\pi(n)} \leq x_n) = \prod_{i=1}^{n} P(X_{\pi(i)} \leq x_i) = \prod_{i=1}^{n} F(X_i)$$

□

Example 17.9. Suppose X and Y are exchangeable. Then $P(X < Y) = P(Y < X)$. If X and Y are continuous, this immediately implies $P(X < Y) = 1/2$ as $P(X = Y) = 0$ (by integrating the joint PDF, it is easy to see that the difference or sum of continuous r.v.'s is continuous). More generally, exchangeability allows us to swap X and Y in any probability statement involving those two random variables, as by definition, (X, Y) and (Y, X) having the same joint distribution means $P((X, Y) \in B) = P((Y, X) \in B)$ for any set $B \subseteq \mathbb{R}^2$.

18. Conditional distributions, multivariate LOTUS

We've seen that conditional probability allows us to update our belief about uncertainty based on knowledge that some event occurred. Today, we extend this notion to distributions.

Definition 18.1 (Conditioning on a discrete random variable)**.** Suppose Y is a discrete random variable with PMF p_Y. If X is also discrete, for any $y \in \text{supp}(Y)$, the **conditional PMF** of X given $Y = y$ is

$$p_{X|Y}(x \mid y) := P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{p_{XY}(x, y)}{p_Y(y)}$$

where p_{XY} is the joint PMF of (X, Y). If X is continuous, the **conditional PDF** of X given $Y = y$ is

$$f_{X|Y}(x \mid y) = \frac{d}{dx} P(X \leq x \mid Y = y)$$

When Y is continuous, the event $Y = y$ has probability 0 for *all* y. However, it is still possible to define a coherent conditional distribution given $Y = y$, if $y \in \text{supp}(Y)$. The definition is motivated by conditioning on the event $Y \in [y, y + \Delta]$ for small Δ, and taking the limit as $\Delta \downarrow 0$.

Definition 18.2 (Conditioning on a continuous random variable)**.** Suppose Y is a continuous random variable with PDF f_Y. Then if X is discrete with PMF p_X, for each $x \in \text{supp}(X)$ and $y \in \text{supp}(Y)$ the **conditional PMF** of X given $Y = y$ is defined as

$$p_{X|Y}(x \mid y) = \frac{f_{Y|X}(y \mid x) p_X(x)}{f_Y(y)}$$

If X is continuous, the **conditional PDF** of X given $Y = y$ is

$$f_{X|Y}(x \mid y) = \frac{f_{XY}(x, y)}{f_Y(y)}$$

Remark. Using the definition of independence, we can show that independence of X and Y is equivalent to $p_{X|Y}(x \mid y) = p_X(x)$ for all x, y when X is discrete and $f_{X|Y}(x \mid y) = f_X(x)$ for all x, y when X is continuous.

It can be shown that no matter if X and Y are continuous or discrete, the conditional PMF $p_{X|Y}$ and the conditional PDF $f_{X|Y}$ are valid, in the sense that they determine valid (univariate) distributions on X for each $y \in \text{supp}(Y)$. Just like conditional probabilities are probabilities, conditional distributions are distributions.

Exercise 18.3 (Chicken-egg story)**.** Suppose a chicken lays $N \sim \text{Pois}(\lambda)$ eggs. Each egg hatches independently with probability p. Let X be the number of eggs that hatch and Y be the number of eggs that don't hatch. Then $X \sim \text{Pois}(\lambda p)$, $Y \sim \text{Pois}(\lambda(1-p))$, and X and Y are independent.

Solution. We compute the joint PMF of X and Y for arbitrary $x, y \in \{0, 1, 2, \ldots\}$, noting that $N = X + Y$ and that by the story of the Binomial, $X \mid N = n \sim \text{Bin}(n, p)$:

$$\begin{aligned} p_{XY}(x,y) = P(X = x, Y = y) &= P(X = x, N = x + y) \\ &= P(X = x \mid N = x + y)P(N = x + y) \\ &= \binom{x+y}{x} p^x (1-p)^y \exp(-\lambda) \frac{\lambda^{x+y}}{(x+y)!} \\ &= \exp(-\lambda p) \frac{(\lambda p)^x}{x!} \exp(-\lambda(1-p)) \frac{(\lambda(1-p))^y}{y!} \end{aligned}$$

This is the product of the $\text{Pois}(\lambda p)$ PMF evaluated at x and the $\text{Pois}(\lambda(1-p))$ PMF evaluated at y, so the result follows. □

Remark. It is perhaps quite unintuitive that the number of eggs that hatch and don't hatch are independent. One might naively expect knowing that a large number of eggs hatched makes it more likely that a large number of eggs were laid overall, and hence the number of unhatched eggs is also large. The previous calculation shows, however, that for the Poisson, this intuition is not quite right.

If X and Y are discrete, the definition of conditional PMFs immediately gives the following restatements of Bayes' rule and LOTP:

$$p_{X|Y}(x \mid y) = \frac{p_{Y|X}(y \mid x) p_X(x)}{p_Y(y)}, \quad p_X(x) = \sum_y p_{X|Y}(x \mid y) p_Y(y)$$

We can generalize these results when X and Y are continuous by replacing conditional and marginal PMFs with conditional and marginal PDFs:

$$f_{X|Y}(x \mid y) = \frac{f_{Y|X}(y \mid x) f_X(x)}{f_Y(y)}, \quad f_X(x) = \int f_{X|Y}(x \mid y) f_Y(y) dy$$

We can also deal with the case where one of X and Y is continuous and the other is discrete, mixing (conditional) PMFs and PDFs as appropriate. These generalized results follow immediately from the definitions of conditional PMFs/PDFs. One important example will follow, but first we'll need to introduce a new distribution family, the Beta.

Definition 18.4. The *Gamma function* is defined as $\Gamma(x) = \int_0^\infty t^{x-1} \exp(-t) dt$ for all $x > 0$.

Proposition 18.5. *The Gamma function has the following properties.*

- $\Gamma(a+1) = a\Gamma(a)$ *for all* $a > 1$
- *If* n *is a positive integer, then* $\Gamma(n) = (n-1)!$

Proof. Induction and integration by parts (skipped). □

Definition 18.6. A random variable X has a **Beta distribution** with parameters $a > 0$ and $b > 0$, abbreviated $X \sim \text{Beta}(a, b)$, if it has PDF

$$f_X(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \quad 0 < x < 1$$

Remark. By matching PDFs, we see that the $\text{Beta}(1,1)$ and $\text{Unif}(0,1)$ distributions are the same.

Example 18.7. (Beta-binomial conjugacy) Suppose we have a coin that lands heads with probability $p \in (0,1)$. If p is unknown, we might model this uncertainty by placing a distribution on p, e.g. $p \sim \text{Beta}(a, b)$. This distribution is called a **prior** distribution, as it is based on our beliefs prior to collecting any data. By flipping the coin, however, we can hope to learn more about p. For example, suppose we flip the coin n times and let X be the number of heads we observe. Since p is considered random, we cannot say that $X \sim \text{Bin}(n, p)$ marginally; rather $X \mid p \sim \text{Bin}(n, p)$. We can then use Bayes' rule to compute the conditional PDF of p given $X = x$, which corresponds to the **posterior** distribution of p given $X = x$:

$$\begin{aligned} f_{p|X}(p \mid x) = \frac{p_{X|p}(x \mid p) f_p(p)}{p_X(x)} \propto p_{X|p}(x \mid p) f_p(p) &= \binom{n}{x} p^x (1-p)^{n-x} p^{a-1}(1-p)^{b-1} \\ &\propto p^{x+a-1}(1-p)^{(n-x)+b-1} \end{aligned}$$

Here the $\propto$ symbol means *proportional to*, indicating that the left-hand side and right-hand side are equivalent up to a multiplicative constant that does not depend on p (but may depend on x). Since PDFs (including conditional PDFs) must integrate to 1, knowing the conditional PDF of p given $X = x$ up to such a constant determines the PDF completely. By pattern matching, we conclude $p \mid X = x \sim \text{Beta}(a + x, b + n - x)$.

Remark. In general, the posterior and prior distributions will not belong to the same family of distributions. In this case, they do, which is a property known as conjugacy (one says the Beta distribution is the *conjugate prior* for the Binomial distribution).

We conclude this lecture with a generalization of LOTUS, which shows that the expectation of any transformation of two random variables can be computed in terms of the *joint* distribution of the *original* random variables. We do not prove this.

Theorem 18.8 (Multivariate LOTUS)**.** *For any function $g : \mathbb{R}^2 \to \mathbb{R}$ and random variables (X, Y) with joint PDF f_{XY} or joint PMF p_{XY}, we have*

$$\mathbb{E}[g(X,Y)] = \int\int g(x,y) f_{XY}(x,y)\,dy\,dx, \quad \mathbb{E}[g(X,Y)] = \sum_{(x,y)\in \text{supp}(X,Y)} g(x,y) p_{XY}(x,y)$$

19. Covariance and correlation

Definition 19.1. Continuing our discussion of joint distributions, today we discuss covariance and correlation, which pertain to the extent to which two random variables "vary together." The **covariance** between two random variables X and Y is defined as

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

The **correlation** between X and Y is a version of the covariance that normalizes by the product of the standard deviations:

$$\mathrm{Cor}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\mathrm{SD}(X)\mathrm{SD}(Y)}$$

We will show later in the course that $-1 \leq \mathrm{Cor}(X, Y) \leq 1$ for any X, Y.

Here are some important properties of covariance:

Theorem 19.2. *The following statements are true for any random variables W, X, Y, and Z and constants a, b, c, and d:*

1. *(Covariance to variance formula)* $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$
2. *(No covariance with constants)* $\mathrm{Cov}(X, a) = 0$
3. *(Symmetry)* $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$
4. *(Bilinearity)* $\mathrm{Cov}(aX+bY, cZ+dW) = ac\mathrm{Cov}(X, Z)+ad\mathrm{Cov}(X, W)+bc\mathrm{Cov}(Y, Z)+bd\mathrm{Cov}(Y, W)$

Proof. Properties 1 through 3 are immediate from the definition of covariance and linearity of expectation. Bilinearity can be shown in two steps: first show $\mathrm{Cov}(aX, Z) = a\mathrm{Cov}(X, Z)$ (immediate from the definition) and then show $\mathrm{Cov}(X + Y, Z) = \mathrm{Cov}(X, Z) + \mathrm{Cov}(Y, Z)$ using linearity of expectation:

$$\begin{aligned}\mathrm{Cov}(X + Y, Z) = \mathbb{E}[(X + Y - \mathbb{E}[X + Y])(Z - \mathbb{E}[Z])] &= \mathbb{E}[(X - \mathbb{E}[X])(Z - \mathbb{E}[Z])] + \mathbb{E}[(Y - \mathbb{E}[Y])(Z - \mathbb{E}[Z])] \\ &= \mathrm{Cov}(X, Z) + \mathrm{Cov}(Y, Z)\end{aligned}$$

Chaining these two results (with Symmetry) shows the full statement of bilinearity. □

For computing *variance*, we recall Proposition 13.2 gave an alternative formula from Definition 13.1 that is typically easier to work with in practice. There is an analogous result for covariance, which holds by a similar argument (expanding out terms and applying linearity of expectation):

Proposition 19.3. *For any random variables X and Y, we have* $\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.

Indeed, with $X = Y$, Proposition 19.3 reduces to Proposition 13.2 by statement 1 in Theorem 19.2. Our next result is quite useful, showing that independent random variables are uncorrelated, meaning their covariance (and correlation) is zero.

Theorem 19.4. *Suppose X and Y are independent random variables. Then they are uncorrelated.*

Proof. We assume X and Y are discrete for simplicity. We compute

$$\begin{aligned}
\mathbb{E}[XY] &= \sum_{(x,y)\in\mathrm{supp}(X,Y)} xyp_{X,Y}(x,y) \text{ by multivariate LOTUS} \\
&= \sum_{(x,y)\in\mathrm{supp}(X,Y)} xyp_X(x)p_Y(y) \text{ by independence} \\
&= \sum_{x\in\mathrm{supp}(X)} \sum_{y\in\mathrm{supp}(Y)} xyp_X(x)p_Y(y) \\
&= \sum_{x\in\mathrm{supp}(X)} xp_X(x) \sum_{y\in\mathrm{supp}(Y)} yp_Y(y) = \mathbb{E}[X]\mathbb{E}[Y]
\end{aligned}$$

where the third equality uses the fact that $\mathrm{supp}(X)\times\mathrm{supp}(Y)$ contains $\mathrm{supp}(X,Y)$, and $p_X(x)p_Y(y) = 0$ for all (x,y) not in $\mathrm{supp}(X,Y)$. The result follows by Proposition 19.3. □

Corollary 19.5. *If X and Y are independent random variables, then $\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$ for any functions g and h, whenever these expectations exist.*

Proof. Since X and Y are independent, $g(X)$ and $h(Y)$ are independent: For $A, B \subseteq \mathbb{R}$ note

$$\begin{aligned}
P(g(X)\in A, h(Y)\in B) &= P(X\in g^{-1}(A), Y\in h^{-1}(B)) \\
&= P(X\in g^{-1}(A))P(Y\in h^{-1}(B)) \\
&= P(g(X)\in A)P(h(Y)\in B)
\end{aligned}$$

where $g^{-1}(A)$ is the preimage of A under g and similarly for $h^{-1}(B)$. But by the previous result

$$0 = \mathrm{Cov}(g(X),h(Y)) = \mathbb{E}[g(X)h(Y)] - \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$$

□

On the other hand, uncorrelated random variables can be very much dependent.

Example 19.6. Suppose $X \sim \mathcal{N}(0,1)$. Then

$$\mathrm{Cov}(X, X^2) = \mathbb{E}[X^3] - \mathbb{E}[X]\mathbb{E}[X^2] = 0$$

where we use the fact that X^3 and $(-X)^3$ have the same distribution, since X and $-X$ are both Standard Normal by Exercise 15.5, implying $\mathbb{E}[X^3] = \mathbb{E}[(-X)^3] = -\mathbb{E}[X^3]$, and hence $\mathbb{E}[X^3] = 0$. But clearly, X and X^2 are not independent - knowing X means you know X^2 deterministically.

Theorem 19.4 and bilinearity allow us to finish the proof of Theorem 13.6. In particular for any independent random variables $X_1, \ldots, X_n$ we have

$$\begin{aligned}
\mathrm{Var}\left(\sum_{i=1}^n X_i\right) &= \mathrm{Cov}\left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i\right) \text{ by covariance to variance formula} \\
&= \sum_{i=1}^n \sum_{j=1}^n \mathrm{Cov}(X_i, X_j) \text{ by bilinearity} \\
&= \sum_{i=1}^n \mathrm{Cov}(X_i, X_i) \text{ as } \mathrm{Cov}(X_i, X_j) = 0 \text{ for all } i \neq j \text{ by independence} \\
&= \sum_{i=1}^n \mathrm{Var}(X_i) \text{ by covariance to variance formula}
\end{aligned}$$

Even when random variables are not independent, the covariance to variance formula along with bilinearity can be quite useful for computing the *variance* of their sum, by using similar argument as the preceding display.

Example 19.7. Recall the Hypergeometric distribution. Suppose $X \sim \mathrm{HGeom}(w, b, n)$ so that we can represent $X = \sum_{i=1}^n I_i$ where I_i is the indicator of the event A_i that the i-th ball drawn from the urn is white. Then

$$\mathrm{Var}(X) = \mathrm{Cov}(X, X) = \sum_{i=1}^n \sum_{j=1}^n \mathrm{Cov}(I_i, I_j)$$

For each $i = 1, \ldots, n$ we have $I_i \sim \mathrm{Bern}\left(\frac{w}{w+b}\right)$ so $\mathbb{E}[I_i] = \frac{w}{w+b}$ and $\mathrm{Cov}(I_i, I_i) = \mathrm{Var}(I_i) = \frac{wb}{(w+b)^2}$. For $i \neq j$ note $I_i I_j = I_{A_i \cap A_j}$. We compute

$$P(A_i \cap A_j) = P(A_j \mid A_i) P(A_i) = \frac{w-1}{w+b-1} \cdot \frac{w}{w+b}$$

Thus for all $i \neq j$ we have

$$\mathrm{Cov}(I_i, I_j) = \mathbb{E}[I_i I_j] - \mathbb{E}[I_i]\mathbb{E}[I_j] = \frac{w(w-1)}{(w+b)(w+b-1)} - \left(\frac{w}{w+b}\right)^2$$

Some algebra gives the result $\mathrm{Var}(X) = \frac{w+b-n}{w+b-1} \cdot n \frac{wb}{(w+b)^2}$.

20. Multinomial distribution, convolution

Today we begin by introducting an important *multivariate* discrete distribution.

Definition 20.1. Suppose I throw n balls into k bins. Each ball independently has probability $p_i \in (0,1)$ of landing in bin $i = 1, \ldots, k$. Let X_i be the number of balls that land in bin i. Then the random *vector* $\boldsymbol{X} = (X_1, \ldots, X_k)$ has the k-dimensional **Multinomial distribution** with parameters n and $\boldsymbol{p} = (p_1, \ldots, p_k)$, abbreviated $\boldsymbol{X} \sim \text{Mult}_k(n, \boldsymbol{p})$.

Remark. The distribution of a random vector $\boldsymbol{X} = (X_1, \ldots, X_k)$ means the same thing as the joint distribution of the random variables $(X_1, \ldots, X_k)$.

We can derive the PMF of the multinomial using a counting argument.

Proposition 20.2. *If $\boldsymbol{X} \sim \text{Mult}_k(n, \boldsymbol{p})$, then the PMF of $\boldsymbol{X}$ is given by*

$$p_X(\boldsymbol{x}) = P(X = \boldsymbol{x}) = \frac{n!}{\prod_{i=1}^k x_i!} p_i^{x_i}$$

with $\text{supp}(\boldsymbol{X})$ *the set of vectors $\boldsymbol{x} = (x_1, \ldots, x_k)$ of nonnegative integers with $\sum_{i=1}^k x_i = n$.*

Proof. Consider a particular ordering of the results of the n ball throws in the Multinomial story where x_i of them land in bin i for all $i = 1, \ldots, k$. The probability of this ordering is $\prod_{i=1}^k p_i^{x_i}$ by independence. There are $\frac{n!}{\prod_{i=1}^k x_i!}$ such orderings by the multiplication rule. □

It follows immediately from the story of the Multinomial and Binomial that if $\boldsymbol{X} \sim \text{Mult}_k(n, \boldsymbol{p})$, then $X_i \sim \text{Bin}(n, p_i)$ for each i. More generally, for any $\mathcal{I} \subseteq \{1, \ldots, k\}$ we have $\sum_{i \in \mathcal{I}} X_i \sim \text{Bin}(n, \sum_{i \in \mathcal{I}} p_i)$. Writing this out explicitly, $\sum_{i \in \mathcal{I}} X_i$ counts how many balls landed in one of the bins in $\mathcal{I}$; each of the n balls independently has probability $\sum_{i \in \mathcal{I}} p_i$ of landing in one of these bins. Indeed, we have the following "lumping" result:

Proposition 20.3. *If* $\boldsymbol{X} \sim \text{Mult}_k(n, \boldsymbol{p})$ *for some* $k \geq 2$*, then*

$$(X_1 + X_2, X_3, \ldots, X_k) \sim \text{Mult}_{k-1}(n, (p_1 + p_2, p_3, \ldots, p_k))$$

We can also condition on a component in the multinomial. The remaining components will follows a $k-1$ dimensional Multinomial distribution. Note we define the conditional distribution of X given Y as the distribution of X given $Y = y$, then replacing the number y with the r.v. Y.

Proposition 20.4. *Suppose* $\boldsymbol{X} \sim \text{Mult}_k(n, \boldsymbol{p})$*. Then* $(X_2, \ldots, X_k) \mid X_1 \sim \text{Mult}_{k-1}(n - X_1, \boldsymbol{r})$ *where* $\boldsymbol{r} = (p_2/(1-p_1), \ldots, p_k/(1-p_1)) \in \mathbb{R}^{k-1}$*.*

Proof. Fix $x_1 \in \{0, 1, \ldots, n\}$. Then given any vector of nonnegative integers $(x_2, \ldots, x_k)$ with $\sum_{i=2}^k x_i = n - x_1$, letting $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)$, by the definition of conditional probability and the fact that $X_1 \sim \text{Bin}(n, p_1)$, we have

$$\begin{aligned} P((X_2, \ldots, X_k) = (x_2, \ldots, x_k) \mid X_1 = x_1) &= \frac{P(\boldsymbol{X} = \boldsymbol{x})}{P(X_1 = x_1)} \\ &= \frac{\frac{n!}{\prod_{i=1}^k x_i!} p_i^{x_i}}{\frac{n!}{x_1!(n-x_1)!} p_1^{x_1}(1-p_1)^{n-x_1}} \\ &= \frac{(n-x_1)!}{\prod_{i=2}^k x_i!} \prod_{i=2}^n \left(\frac{p_i}{1-p_1}\right)^{x_i} \end{aligned}$$

which corresponds precisely to the $\text{Mult}_{k-1}(n - x_1, \boldsymbol{r})$ PMF. □

Remark. The intuition behind the preceding result is that given that x_1 balls landed in bin 1, the remaining $n - x_1$ balls independently land in bin j with probability equal to the conditional probability of landing there given they do not land in ball 1.

Finally, we compute the covariance between components of the multinomial

Proposition 20.5. *If* $\boldsymbol{X} \sim \text{Mult}_k(n, \boldsymbol{p})$*, then for* $i, j \in \{1, \ldots, k\}$ *we have*

$$\text{Cov}(X_i, X_j) = \begin{cases} np_i(1-p_i) & i = j \\ -np_ip_j & i \neq j \end{cases}$$

Proof. If $i = j$ then $\text{Cov}(X_i, X_j) = \text{Cov}(X_i, X_i) = \text{Var}(X_i) = np_i(1-p_i)$ as $X_i \sim \text{Bin}(n, p_i)$. If $i \neq j$ then we return to the story and let I_m and J_m be the the indicators that the m-th ball lands in bin i and bin j, respectively. Then $X_i = \sum_{m=1}^n I_m$ and $X_j = \sum_{m=1}^n J_m$. By bilinearity

$$\text{Cov}(X_i, X_j) = \sum_{m=1}^n \sum_{\ell=1}^n \text{Cov}(I_m, J_\ell) = \sum_{m=1}^n \text{Cov}(I_m, J_m)$$

since for $m \neq \ell$, I_m and J_ℓ are independent, hence uncorrelated. But

$$\text{Cov}(I_m, J_m) = \mathbb{E}[I_m J_m] - \mathbb{E}[I_m]\mathbb{E}[J_m] = -p_i p_j$$

for each m, since $I_m J_m = 0$ (it is impossible for the m-th ball to land in both bin i and bin j). □

Our final topic for today is known as convolution: a formal way of deriving the distribution of a sum of independent random variables. We have already seen some convolutions, such as proving that the sum of independent Poissons is still Poisson (Theorem 10.9). We now generalize that construction, also handling the continuous case.

Theorem 20.6 (Convolution)**.** *If X and Y are independent discrete random variables with PMFs p_X and p_Y, respectively, then their sum $T = X + Y$ has PMF*

$$p_T(t) = P(T = t) = \sum_x p_X(x) p_Y(t - x) = \sum_y p_Y(y) p_X(t - y)$$

If X and Y are independent continuous random variables with PDFs f_X and f_Y, respectively, then their sum $T = X + Y$ has PDF

$$f_T(t) = \int_{-\infty}^{\infty} f_X(x) f_Y(t - x) dx = \int_{-\infty}^{\infty} f_Y(y) f_X(t - y) dy$$

Proof. For the discrete case, note that by countable additivity and independence we have

$$P(T = t) = \sum_x P(T = t \cap X = x) = \sum_x P(X = x \cap Y = t - x) = \sum_x p_X(x) p_Y(t - x)$$

as in the proof of Theorem 10.9. By symmetry we can swap the roles of X and Y to show $P(T = t) = \sum_y p_X(x) p_Y(t - x)$ as well. For the continuous case we start by deriving the joint CDF by integrating the joint PDF:

$$F_T(t) = P(X + Y \leq t) = \int_{-\infty}^{\infty} \int_{-\infty}^{t-x} f_X(x) f_Y(y) dy dx = \int_{-\infty}^{\infty} f_X(x) F_Y(t - x) dx$$

Differentiating both sides with respect to t (and swapping integration with differentiation) gives

$$f_T(t) = \int_{-\infty}^{\infty} f_X(x) f_Y(t - x) dx$$

Swapping the roles of X and Y gives $f_T(t) = \int_{-\infty}^{\infty} f_Y(y) f_X(t - y) dy$. □

Convolution will allow us to prove the well-known result that the sum of independent Normal random variables also has a Normal distribution.

Theorem 20.7 (Sum of independent normals is normal)**.** *Suppose X, Y are i.i.d. $\mathcal{N}(0,1)$. Then $X+Y \sim \mathcal{N}(0,2)$.*

Proof. Let $T = X + Y$ and apply the convolution formula:

$$\begin{aligned} f_T(t) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(t-x)^2}{2}\right) dx \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2 - tx + t^2/2}{2 \cdot (1/2)}\right) dx \\ &= \exp\left(-\frac{t^2}{4}\right) \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp\left(-\frac{(x-t/2)^2}{2 \cdot (1/2)}\right) dx \end{aligned}$$

By considering that the $\mathcal{N}(t/2, 1/2)$ PDF integrates to 1 we know $\int_{-\infty}^{\infty} \exp\left(-\frac{(x-t/2)^2}{2\cdot(1/2)}\right) = \sqrt{\pi}$ so f_T is the $\mathcal{N}(0,2)$ PDF, as desired. □

The previous proof can be generalized (with messier notation) to the following result:

Corollary 20.8. *If $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ are independent, then $X+Y \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$.*

The fact that the mean and variance parameters for $X+Y$ are $\mu_X + \mu_Y$ and $\sigma_X^2 + \sigma_Y^2$, respectively, should be easy to remember by linearity of expectation and Theorem 13.6, part 3. The real content of the result is that $X + Y$ is in fact Normal.

21. Change of variables, Gamma distribution

We have seen several examples of dealing with transformations (functions) of random variables. (Multivariate) LOTUS allows us to compute *expectations* of transformations of random variable(s) in terms of the (joint) distribution of the original random variable(s). However, sometimes we want to understand the full distribution of a transformation.

The "fail-safe" way to understand the distribution of a transformation is to work with (joint) CDFs or PMFs. This is because both CDFs and PMFs correspond to probabilities, so we can use the tools of probability to manipulate them (e.g. rewriting events). However, CDFs can sometimes be challenging to work with for continuous random variables (need to integrate PDFs), for which working with the PDF may be more natural. Thus, we begin by deriving some results for dealing with transformations of continuous random variables directly in terms of (joint) PDFs. It is important to remember, however, that they only apply when the transformations are **invertible**. For non-invertible transformations, you will need to use another approach, such as by manipulating the CDF directly.

Theorem 21.1 (1-D change of variables)**.** *Suppose X is a continuous random variable with PDF f_X and g is a real function that is differentiable and strictly increasing or decreasing on* $\text{supp}(X)$*. Then $Y = g(X)$ is continuous with PDF*

$$f_Y(y) = f_X(g^{-1}(y))\left|\frac{d}{dy}g^{-1}(y)\right| = f_X(g^{-1}(y))\left|g'(g^{-1}(y))\right|^{-1}$$

for all $y \in \{g(x) \mid x \in \text{supp}(X)\}$

Proof. That g is differentiable and strictly increasing or decreasing implies that the inverse function g^{-1} exists and is itself differentiable. In the case g is strictly decreasing, fix $y \in \{g(x) \mid x \in \text{supp}(x)\}$

so that $g^{-1}(y) \in \text{supp}(X)$, and note

$$P(Y \leq y) = P(g(X) \leq y) = P(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y))$$

Taking the derivative on both sides yields the first equation by the chain rule (note g strictly decreasing implies g^{-1} strictly decreasing, hence $\frac{d}{dy}g^{-1}(y) < 0$). The second equation follows by the formula for the derivative of an inverse function, which is derived by differentiating both sides of the identity $g(g^{-1}(y)) = y$. The strictly increasing case is analogous. □

Example 21.2. Suppose $U \sim \text{Unif}(0,1)$. Then for all $y \in (0,1)$, the PDF of $Y = U^n$ satisfies

$$f_Y(y) = f_U(g^{-1}(y))\left|\frac{d}{dy}g^{-1}(y)\right| = \frac{1}{n}y^{1/n-1}$$

where $g(u) = u^n \implies g^{-1}(y) = y^{1/n}$. By pattern matching PDFs (recall we can ignore normalizing constants like $1/n$), we have $Y \sim \text{Beta}(1/n, 1)$.

There is a multivariate analogue of the change of variables formula, which we state without proof.

Theorem 21.3. *Suppose $\boldsymbol{X} = (X_1, \ldots, X_n)$ is a continuous random vector with joint PDF $f_{\boldsymbol{X}}$, and suppose g is an invertible function from $\text{supp}(X)$ into $\mathbb{R}^n$. For each $\boldsymbol{x} \in \text{supp}(X)$, let $\boldsymbol{y} = g(\boldsymbol{x})$ (so that $\boldsymbol{x} = g^{-1}(\boldsymbol{y}) = (g_1^{-1}(\boldsymbol{y}), \ldots, g_n^{-1}(\boldsymbol{y}))$). Suppose all partial derivatives $\frac{\partial x_i}{\partial y_j}$ exist for $i = 1, \ldots, n$, $j = 1, \ldots, n$, so that we can define the Jacobian matrix*

$$\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{y}} = \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_n} \end{bmatrix}$$

Then for any $\boldsymbol{x} \in \text{supp}(X)$, the joint PDF of $\boldsymbol{Y} = g(\boldsymbol{X})$ is given by

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}(\boldsymbol{x})\left|\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{y}}\right| = f_{\boldsymbol{X}}(\boldsymbol{x})\left|\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right|^{-1}$$

where $|A|$ denotes the absolute value of the determinant of a square matrix A.

Remark. A useful mnemonic for the change of variables formula is "$f_{\boldsymbol{Y}}(\boldsymbol{y})\partial \boldsymbol{y} = f_{\boldsymbol{X}}(\boldsymbol{x})\partial \boldsymbol{x}$"

To simplify the application of the multivariate change of variables formula for specific choices of $f_{\boldsymbol{X}}$ and g, it is helpful to break up the process into three steps:

1. Compute $\boldsymbol{x} = g^{-1}(\boldsymbol{y})$ by writing down the system $\boldsymbol{y} = g(\boldsymbol{x})$ and solving for $\boldsymbol{x}$. Note g is invertible if and only if you obtain a unique solution for $\boldsymbol{x} \in \text{supp}(\boldsymbol{X})$.

2. Compute $|\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}|^{-1}$ in terms of $\boldsymbol{x}$, or $|\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{y}}|$ directly in terms of $\boldsymbol{y}$.

3. Using the results of the previous steps and deriving $f_{\boldsymbol{X}}$, compute either

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}(\boldsymbol{x})\Big|\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\Big|^{-1}, \qquad f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}(\boldsymbol{x})\Big|\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{y}}\Big|$$

depending on which quantity you computed in step 2. Plug in the expression for $\boldsymbol{x}$ in terms of $\boldsymbol{y}$ from the first step on the right-hand side to get $f_{\boldsymbol{Y}}(\boldsymbol{y})$ fully in terms of $\boldsymbol{y}$.

Exercise 21.4 (Box-Muller)**.** Suppose $U \sim \text{Unif}(-\pi, \pi)$ is independent of $T \sim \text{Expo}(1)$. Define $X = \sqrt{2T}\cos(U)$ and $Y = \sqrt{2T}\sin(U)$. What is the joint distribution of (X, Y)?

Solution. We evaluate the joint PDF of (X, Y) at some $(x, y) \in \mathbb{R}^2$ with $x, y \neq 0$ by applying the change of variables formula.

1. We write $(x, y) = (\sqrt{2t}\cos(u), \sqrt{2t}\sin(u))$. Solving for (u, t) we get the following unique solution in $\text{supp}(U, T) = (-\pi, \pi) \times (0, \infty)$:

$$t = \frac{x^2 + y^2}{2}, \quad u = \begin{cases} \tan^{-1}(y/x) & x > 0 \\ \tan^{-1}(y/x) + \pi & x < 0, y > 0 \\ \tan^{-1}(y/x) - \pi & x < 0, y < 0 \end{cases}$$

(this is essentially just the transformation to polar coordinates).

2. We compute

$$\frac{\partial(x, y)}{\partial(u, t)} = \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial t} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial t} \end{bmatrix} = \begin{bmatrix} -\sqrt{2t}\sin(u) & (2t)^{-1/2}\cos(u) \\ \sqrt{2t}\cos(u) & (2t)^{-1/2}\sin(u) \end{bmatrix}$$

so that $\left|\frac{\partial(x,y)}{\partial(u,t)}\right|^{-1} = |-\sin^2(u) - \cos^2(u)|^{-1} = 1$

3. We obtain

$$f_{XY}(x, y) = f_U(u)f_T(t) = \frac{1}{2\pi} \cdot f_T\left(\frac{x^2 + y^2}{2}\right) = \frac{1}{2\pi}\exp\left(-\frac{x^2 + y^2}{2}\right) = \varphi(x)\varphi(y)$$

where φ is the Standard Normal PDF. Thus X, Y are i.i.d. standard Normal!

□

Another important example of the change of variables formula links the Beta distribution to a widely used generalization of the Exponential distribution called the Gamma distribution.

Definition 21.5. A random variable Y has the **Gamma distribution** with parameters $a > 0$ and $\lambda > 0$, abbreviated $Y \sim \text{Gamma}(a, \lambda)$, if its PDF f_Y is given by

$$f_Y(y) = \frac{1}{y\Gamma(a)}(\lambda y)^a \exp(-\lambda y), \quad y > 0$$

where $\Gamma(\cdot)$ is the Gamma function from Definition 18.4.

The following result shows why the Gamma distribution is a generalization of the Exponential.

Theorem 21.6. *Suppose $X_1, \ldots, X_n$ are i.i.d.* Expo(λ) *r.v.'s. Then $T = \sum_{i=1}^n X_i \sim \text{Gamma}(n, \lambda)$.*

Proof. We prove the $n = 2$ case using convolution; the result follows for larger n by induction. Let f be the Expo(λ) PDF. For any $t > 0$, the PDF of T satisfies

$$f_T(t) = \int_0^\infty f(x)f(t-x)dx = \int_0^t \lambda^2 \exp(-\lambda x)\exp(-\lambda(t-x))dx = t\lambda^2 \exp(-\lambda t)$$

which matches the Gamma$(2, \lambda)$ PDF. □

Theorem 21.7 (Bank/post-office story)**.** *Suppose $X \sim \text{Gamma}(a, \lambda)$ independently of $Y \sim \text{Gamma}(b, \lambda)$. Then $T = X + Y \sim \text{Gamma}(a + b, \lambda)$ and $W = \frac{X}{X+Y} \sim \text{Beta}(a, b)$. Furthermore $T \amalg W$.*

Remark. One could think of X and Y as modeling waiting times for the bank and the post office, respectively. Then the result states that the total time waited, T, is independent of the proportion of the total time spent waiting for the bank, W.

Remark. That $T \sim \text{Gamma}(a + b, \lambda)$ marginally follows immediately from the representation as a sum of i.i.d. Expo(λ) r.v.'s, when a and b are integers.

Proof. We derive the joint PDF $f_{T,W}$ of (T, W) at some $(t, w) \in (0, \infty) \times (0, 1)$

1. We write $(t, w) = (x + y, x/(x + y))$ which has unique solution $(x, y) = (tw, t(1 - w))$ in $(0, \infty)^2$.

2. We compute
$$\frac{\partial(x, y)}{\partial(t, w)} = \begin{bmatrix} \frac{\partial x}{\partial t} & \frac{\partial x}{\partial w} \\ \frac{\partial y}{\partial t} & \frac{\partial y}{\partial w} \end{bmatrix} = \begin{bmatrix} w & t \\ 1 - w & -t \end{bmatrix}$$
which has absolute determinant $|-wt - t(1 - w)| = |-t| = t$

3. Conclude that

$$\begin{aligned} f_{TW}(t,w) &= f_{XY}(x,y)\left|\frac{\partial(x,y)}{\partial(t,w)}\right| \\ &= \frac{1}{x\Gamma(a)}(\lambda x)^a \exp(-\lambda x)\cdot\frac{1}{y\Gamma(b)}(\lambda y)^b\exp(-\lambda y)\cdot t \\ &= \frac{1}{(tw)\Gamma(a)}(\lambda tw)^a \exp(-\lambda tw)\cdot\frac{1}{t(1-w)\Gamma(b)}(\lambda t(1-w))^b\exp(-\lambda t(1-w))\cdot t \\ &= \frac{1}{t\Gamma(a+b)}\exp(-\lambda t)(\lambda t)^{a+b}\cdot\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}w^{a-1}(1-w)^{b-1} \end{aligned}$$

This is the product of the Gamma$(a+b,\lambda)$ PDF evaluated at t with the Beta(a,b) PDF evaluated at w.

□

22. Conditional expectation: Events

Today, we begin our study of conditional expectation. We've seen that (unconditional) expectation is a useful and interpretable quantity that can be computed and studied without knowledge of the entire distribution of a random variable (e.g. using linearity and other tricks). Conditional expectation is an important extension that allows us to account for known information, just like conditional probability allows us to measure probabilities in a way that accounts for such information.

Definition 22.1 (Conditional expectation given an event)**.** Suppose A is an event with $P(A) > 0$ and X is a random variable. Then the **conditional distribution** of X given A is given by $\mathcal{L}_{X|A}(B) = P(X \in B \mid A)$, and the **conditional expectation** of X given A is the expectation of a random variable with the distribution $\mathcal{L}_{X|A}$. In other words, for X discrete we have

$$\mathbb{E}[X \mid A] = \sum_{x \in \text{supp}(X)} xP(X = x \mid A)$$

while for X continuous, we define its conditional PDF $f_X(x \mid A) = \frac{d}{dx}P(X \leq x \mid A)$ and

$$\mathbb{E}[X \mid A] = \int_{-\infty}^{\infty} x f_X(x \mid A)dx$$

The following is an alternative, equivalent definition of $\mathbb{E}[X \mid A]$.

Proposition 22.2. *For any event A with $P(A) > 0$, we have $\mathbb{E}[X \mid A] = \frac{\mathbb{E}[XI_A]}{P(A)}$.*

Proof. We assume X is discrete for technical simplicity. Then

$$\begin{aligned}
\mathbb{E}[XI_A] &= \sum_{x \in \text{supp}(X)} \sum_{a=0}^{1} xaP(X = x, I_A = a) \text{ multivariate LOTUS} \\
&= \sum_x xP(X = x, I_A = 1) \\
&= \sum_x xP(X = x \mid A)P(A) \text{ by definition of conditional probability} \\
&= P(A)\mathbb{E}[X \mid A]
\end{aligned}$$

□

One type of event that can be conditioned on is that the random variable lies within some subset of values within its support.

Example 22.3. Suppose $T \sim \text{Expo}(\lambda)$. For any $t > 0$, the conditional distribution of $T - t$ given $T \geq t$ is $\text{Expo}(\lambda)$. To see this, note that for all $s > 0$ we have

$$P(T - t \leq s \mid T \geq t) = 1 - P(T - t > s \mid T \geq t) = 1 - P(T > s) = P(T \leq s)$$

where the second equality follows by memorylessness; then use the fact that the CDF determines the distribution. We conclude $\mathbb{E}[T \mid T \geq t] = t + \mathbb{E}[T] = t + \lambda^{-1}$.

For computing the probability of an event, the law of total probability is a very useful tool as it allows us to split the event based on a partition of the sample space. There is an extension known as the law of total expectation (LOTE).

Theorem 22.4 (Law of total expectation (LOTE))**.** *Suppose X is a random variable and $B_1, \ldots, B_n$ partition the sample space (as defined in Theorem 4.2), with $P(B_i > 0)$ for all i. Then*

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X \mid B_i] P(B_i)$$

Proof. Since $B_1, \ldots, B_n$ are a partition, we have $X = \sum_{i=1}^{n} X I_{B_i}$. Then by linearity and Proposition 22.2

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X I_{B_i}] = \sum_{i=1}^{n} \mathbb{E}[X \mid B_i] P(B_i)$$

□

Exercise 22.5. Suppose you are bidding on a prize with unknown value $V \sim \text{Unif}(0, 1)$. The rules are as follows: If you bid less than $2V/3$, nothing happens. Otherwise, you win the prize but pay your bid. What bid maximizes your expected winnings?

Solution. Suppose the bid is b. Note it never makes sense to bid more than $2/3$, since you will always get the prize with $b \geq 2/3$ and simply pay extra by bidding higher than $2/3$. By bidding 0 or less, clearly we never win the prize and so our winnings are always 0. Thus we analyze the case $0 < b \leq 2/3$. Let W be the winnings. Then by LOTE

$$\begin{aligned} \mathbb{E}[W] &= \mathbb{E}[W \mid b < 2V/3] P(b < 2V/3) + \mathbb{E}[W \mid b \geq 2V/3] P(b \geq 2V/3) \\ &= \mathbb{E}[V - b \mid b \geq 2V/3] P(b \geq 2V/3) \end{aligned}$$

Note b is a constant and the event $b \geq 2V/3$ can be rewritten as $V \leq (3/2)b$, so for any $v \in (0,1)$

$$P(V \leq v \mid V \leq (3/2)b) = \frac{P(V \leq \min(v, (3/2)b))}{P(V \leq (3/2)b)} = \frac{\min(v, (3/2)b)}{(3/2)b}$$

showing that $V \mid b \geq 2V/3 \sim \text{Unif}(0, (3/2)b)$. Thus

$$\mathbb{E}[W] = \left(\frac{3}{4}b - b\right) \cdot (3/2)b = -\frac{3}{8}b^2 < 0, \quad 0 < b \leq 2/3$$

So in fact it's best to bid 0 (or a negative value), so that you never win the prize. □

Recall that first-step analysis was an important strategy for computing probabilities in settings with a self-similar or recursive structure. The idea was to consider conditional probabilities given all possible initial outcomes, which partitions the sample space. An analogous concept can be useful for computing expectations.

Example 22.6. Suppose I keep flipping a fair coin until I obtain two heads in a row. On average, how many total flips will I need to perform? What if I instead kept flipping until I obtained heads followed immediately by tails?

Solution. Let H_i be the event that the i-th flip lands heads and X be the number of flips I perform before getting two heads in a row. Given H_1^c (the first flip lands tails), the conditional distribution of X is the same as the marginal distribution of $1 + X$, since we made no progress in the first flip and the flips are independent, so the number of additional flips (after the first) to get two heads in a row has the same distribution as the marginal distribution of X. Hence $\mathbb{E}[X \mid H_1^c] = 1 + \mathbb{E}[X]$. Similarly, given H_1 and H_2^c, the progress "resets" after 2 flips and $\mathbb{E}[X \mid H_1, H_2^c] = 2 + \mathbb{E}[X]$. Of course, $X = 2$ with probability 1 given H_1, H_2. Thus by LOTE

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}[X \mid H_1, H_2]P(H_1, H_2) + \mathbb{E}[X \mid H_1, H_2^c]P(H_1, H_2^c) + \mathbb{E}[X \mid H_1^c]P(H_1^c) \\ &= 2 \cdot \frac{1}{4} + (2 + \mathbb{E}[X]) \cdot \frac{1}{4} + (1 + \mathbb{E}[X]) \cdot \frac{1}{2}\end{aligned}$$

Solving gives $\mathbb{E}[X] = \boxed{6}$.

Now let Y be the number of flips to perform before getting heads followed by a tails. Given H_1^c the progress resets as before. But $Y \mid H_1 \sim 1 + \text{FS}(1/2)$, since when the first flip is Heads we will be done the next time a tails comes up. Thus $\mathbb{E}[Y \mid H_1] = 3$ (recalling the First Success expectation) and

$$\mathbb{E}[Y] = \mathbb{E}[Y \mid H_1]P(H_1) + \mathbb{E}[Y \mid H_1^c]P(H_1^c) = 3 \cdot \frac{1}{2} + (1 + \mathbb{E}[Y]) \cdot \frac{1}{2}$$

Solving gives $\mathbb{E}[Y] = \boxed{4}$. □

Remark. The previous result may seem very counterintuitive, since the sequences HH and HT are equally likely for any two coin flips. Indeed, by linearity we can show that for any fixed sequence of n coin flips, the expected number of occurrences of HH and HT are the same (create an indicator for each of the $n-1$ consecutive pairs of coin flips). However, the HH sequences tend to be more clumped together (we can get two HH sequences in three flips by HHH, but this is not possible for HT). Thus, on average, the number of flips until the *first* occurrence of HH needs to be greater to compensate.

23. Conditional expectation: Random variables, iterated expectation

Last lecture, we introduced the concept of conditional expectation given an event. Today we generalize this definition to define the conditional expectation of a random variable *given another random variable*. This will have some useful properties that will simplify problems. It is critical to keep in mind that whereas $\mathbb{E}[X \mid A]$ is a number for any event A, for any r.v. Y, $\mathbb{E}[X \mid Y]$ itself a *random variable*, in particular a transformation of Y.

Definition 23.1 (Conditional expectation given a random variable)**.** Suppose X and Y are random variables. Let $g(y) = \mathbb{E}[X \mid Y = y]$ for all $y \in \text{supp}(Y)$, which refers to the expectation of a random variable with the distribution induced by the conditional PMF or PDF of X given $Y = y$, as in Definitions 18.1 and 18.2. Then the **conditional expectation** of X given Y is the *random variable* $g(Y)$.

Example 23.2. Let I_i be the indicator that the i-th flip of a fair coin lands heads. Then with $Z = I_1 + I_2$, we have $\mathbb{E}[Z \mid I_1 = 1] = 3/2$ and $\mathbb{E}[Z \mid I_1 = 0] = 1/2$, so we can write $\mathbb{E}[Z \mid I_1] = 1/2 + I_1$ since $I_1 \in \{0, 1\}$.

Example 23.3. Suppose we have a stick of length 1 and break the stick at a point X chosen uniformly at random on $[0, 1]$. Given that $X = x$, we then choose another breakpoint Y uniformly on the interval $[0, x]$. Then $Y \mid X = x \sim \text{Unif}(0, x)$, so $\mathbb{E}[Y \mid X = x] = x/2$ and $\mathbb{E}[Y \mid X] = X/2$.

Here are some properties of conditional expectation.

Proposition 23.4. *Suppose X, Y, and Z are random variables. Then*

1. *(Takeout)* $\mathbb{E}[h(X)Y \mid X] = h(X)\mathbb{E}[Y \mid X]$ *for any function h with* $\mathbb{E}[|h(X)|] < \infty$
2. *(Dropping what's independent) If X and Y are independent,* $\mathbb{E}[Y \mid X] = \mathbb{E}[Y]$

3. *(Linearity)* $\mathbb{E}[aX + bY \mid Z] = a\mathbb{E}[X \mid Z] + b\mathbb{E}[Y \mid Z]$

Proof. The results follow quite simply by noting conditional distributions are distributions. For instance, fixing $x \in \text{supp}(X)$ we have

$$\mathbb{E}[h(X)Y \mid X = x] = \mathbb{E}[h(x)Y \mid X = x] = h(x)\mathbb{E}[Y \mid X = x]$$

where the second equality is by linearity of expectation. Replacing x with X proves Takeout. □

We now turn to an important generalization of LOTE, this time stated succinctly in terms of conditioning on a random variable.

Theorem 23.5 (Law of iterated expectations)**.** *For any random variables X and Y, we have* $\mathbb{E}[\mathbb{E}[Y \mid X]] = \mathbb{E}[Y]$

Remark. To parse the law of iterated expectations, recall $\mathbb{E}[Y \mid X]$ is a r.v. (indeed, a transformation of X). Then the outer expectation is a constant, which the law claims is equal to $\mathbb{E}[Y]$.

Proof. In the case that X is discrete, we have

$$\mathbb{E}[\mathbb{E}[Y \mid X]] = \sum_x \mathbb{E}[Y \mid X = x]P(X = x) = \mathbb{E}[Y]$$

where the first equality is by LOTUS and the second equality is by LOTE. If both X and Y are continuous,

$$\begin{aligned}\mathbb{E}[\mathbb{E}[Y \mid X]] &= \int_{-\infty}^{\infty} \mathbb{E}[Y \mid X = x] f_X(x)dx \\ &= \int_{-\infty}^{\infty} \left(\int y f_{Y|X}(y \mid x)dy \right) f_X(x)dx \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{X,Y}(x, y)dydx = \mathbb{E}[Y]\end{aligned}$$

The case of X continuous, Y discrete follows analogously. □

Example 23.6. Continuing Example 23.2, we verify $\mathbb{E}[\mathbb{E}[Z \mid I_1]] = \mathbb{E}[1/2 + I_1] = 1 = \mathbb{E}[Z]$.

Example 23.7. Continuing Example 23.3, we verify $\mathbb{E}[\mathbb{E}[Y \mid X]] = \mathbb{E}[X/2] = 1/4$. We could alternatively check this by deriving the joint PDF of (X, Y) and then marginalizing out X to get the marginal PDF of Y, before computing the expectation of Y directly. However, iterated expectations gives a much quicker solution when we don't need the full distributional information.

Example 23.8 (Random sum)**.** Consider a game where a fair coin is flipped repeatedly. Each time the the coin lands heads, the player is given a reward R_i. Suppose $R_1, R_2, \ldots$ are i.i.d. with mean \$1. Once the coin lands tails, the game is over. Then the total reward is $R = \sum_{i=1}^{N} R_i$ where $N \sim \text{Geom}(1/2)$. With N independent of the rewards, we have

$$\mathbb{E}[R \mid N = n] = \mathbb{E}\left[\sum_{i=1}^{N} R_i \mid N = n\right] = \mathbb{E}\left[\sum_{i=1}^{n} R_i \mid N = n\right] = \sum_{i=1}^{n} \mathbb{E}[R_i \mid N = n] = n\mathbb{E}[R_1] = n$$

and so $\mathbb{E}[R \mid N] = N$ and by iterated expectations we have

$$\mathbb{E}[R] = \mathbb{E}[\mathbb{E}[R \mid N]] = \mathbb{E}[N] = 1$$

We now show that just like there are versions of Bayes' rule and LOTP when conditioning on multiple events — because conditional probabilities are probabilities — there is also a version of iterated expectations based on conditioning on multiple random variables, because conditional expectations are expectations.

Theorem 23.9 (Iterated expectations with extra conditioning)**.** *For any random variables X, Y and Z, we have*

$$\mathbb{E}[\mathbb{E}[X \mid Y, Z] \mid Z] = \mathbb{E}[X \mid Z]$$

Note a conditional expectation given Y and Z jointly is to be interpreted as the conditional expectation given the random vector (Y, Z).

The proof of the above is immediate from applying iterated expectations to $\mathbb{E}[\cdot \mid Z]$.

Example 23.10. For any random variables X and Y, we have the identity

$$\mathbb{E}[X \mid \mathbb{E}[X \mid Y]] = \mathbb{E}[X \mid Y]$$

To see this, let $Z = \mathbb{E}[X \mid Y]$. Recall Z is a random variable, indeed a deterministic transformation of Y, i.e. $Z = h(Y)$. Thus, if we already know Y, conditioning on Z doesn't give any extra information. Formally, each point in $\text{supp}(Y, Z)$ takes the form $(y, h(y))$ for some $y \in \text{supp}(Y)$, yet

$$\mathbb{E}[X \mid Y = y, Z = h(y)] = \mathbb{E}[X \mid Y = y], \forall y \in \text{supp}(Y)$$

Thus $\mathbb{E}[X \mid Y, Z] = \mathbb{E}[X \mid Y]$, and by iterated expectations with extra conditioning, we have

$$\mathbb{E}[X \mid Z] = \mathbb{E}[\mathbb{E}[X \mid Y, Z] \mid Z] = \mathbb{E}[\mathbb{E}[X \mid Y] \mid Z] = \mathbb{E}[X \mid Y]$$

using takeout for the last equality. Intuitively, this result states that knowing $\mathbb{E}[X \mid Y]$ gives all the possible information about the mean of X that you could get from knowing Y. Any additional information about Y won't update your probability beliefs in a way that affects the mean of X.

24. Conditional variance, tail bounds

To understand the variability of a conditional distribution, we introduce the concept of *conditional variance*.

Definition 24.1. The **conditional variance** of a random variable X given another random variable Y is defined as

$$\operatorname{Var}(X \mid Y) = \mathbb{E}[(X - \mathbb{E}[X \mid Y])^2 \mid Y] = \mathbb{E}[X^2 \mid Y] - (\mathbb{E}[X \mid Y])^2$$

where the second equality follows by Proposition 23.4.

Theorem 24.2 (Law of total variance)**.** *For any random variables X and Y, we have*

$$\operatorname{Var}(X) = \mathbb{E}[\operatorname{Var}(X \mid Y)] + \operatorname{Var}(\mathbb{E}[X \mid Y])$$

Proof. We have by linearity and iterated expectations that

$$\mathbb{E}[\operatorname{Var}(X \mid Y)] = \mathbb{E}[\mathbb{E}[X^2 \mid Y] - (\mathbb{E}[X \mid Y])^2) = \mathbb{E}[X^2] - \mathbb{E}[(\mathbb{E}[X \mid Y])^2]$$

By the definition of variance and iterated expectations, we have

$$\operatorname{Var}(\mathbb{E}[X \mid Y]) = \mathbb{E}[(\mathbb{E}[X \mid Y])^2] - (\mathbb{E}[\mathbb{E}[X \mid Y]])^2 = \mathbb{E}[(\mathbb{E}[X \mid Y])^2] - (\mathbb{E}[X])^2$$

Adding together gives

$$\mathbb{E}[\operatorname{Var}(X \mid Y)] + \operatorname{Var}(\mathbb{E}[X \mid Y]) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \operatorname{Var}(X)$$

as desired. □

Example 24.3 (Random sum, continued)**.** We recall the setting of Example 23.8 to compute $\operatorname{Var}(R)$ using the law of total variance. Assume additionally that $\operatorname{Var}(R_i) = 1$. Then

$$\operatorname{Var}(R \mid N = n) = \operatorname{Var}\left(\sum_{i=1}^{N} R_i \mid N = n\right) = \sum_{i=1}^{n} \operatorname{Var}(R_i) = n$$

where the second equality uses the fact that $(R_1, \ldots, R_n)$ are independent of the event $N = n$, and furthermore the R_i are independent so their variances add (Theorem 13.6). Recalling $\mathbb{E}[R \mid N] = N$ from Example 23.8, we conclude by Example 13.9 that

$$\begin{aligned}\mathrm{Var}(R) &= \mathbb{E}[\mathrm{Var}(R \mid N)] + \mathrm{Var}(\mathbb{E}[R \mid N]) \\ &= \mathbb{E}[N] + \mathrm{Var}(N) = 1 + 2 = \boxed{3}\end{aligned}$$

Example 24.4 (Cluster sampling)**.** A plague is affecting a large number of cities. Suppose a city is chosen at random and then the disease statuses of n individuals from that city (chosen with replacement) are noted. Let S be the number of sick individuals in that sample. Assume we do not have any additional information about each city so we model the *illness propensity* in each city as $U \sim \mathrm{Unif}(0, 1)$ (the randomness comes from some unknown variation in illness propensities across cities). Given U, suppose each city resident independently has probability U of being sick. Then by iterated expectations

$$\mathbb{E}[S] = \mathbb{E}[\mathbb{E}[S \mid U]] = \mathbb{E}[nU] = n/2$$

since $S \mid U \sim \mathrm{Bin}(n, U)$. By the law of total variance, we have

$$\begin{aligned}\mathrm{Var}(S) &= \mathbb{E}[\mathrm{Var}(S \mid U)] + \mathrm{Var}(\mathbb{E}[S \mid U]) = \mathbb{E}[nU(1-U)] + \mathrm{Var}(nU) \\ &= n(\mathbb{E}[U] - \mathbb{E}[U^2]) + n^2\mathrm{Var}(U) = n(1/2 - 1/3) + n^2(1/12) \\ &= \frac{n}{6} + \frac{n^2}{12}\end{aligned}$$

For the remainder of the lecture, we will discuss some inequalities involving random variables. Today we focus on tail bounds, which give upper bounds on the probability that a random variable takes on extreme values in terms of expectations. The most basic is an important result called Markov's inequality, which bounds the probability that a nonnegative random variable is large in terms of its expectation.

Theorem 24.5 (Markov's inequality)**.** *Suppose X is a random variable with $\mathbb{E}[|X|] < \infty$. Then for any $a > 0$ we have*

$$P(|X| \geq a) \leq \frac{\mathbb{E}[|X|]}{a}$$

Proof. First we show the inequality $aI_{|X| \geq a} \leq |X|$. On the event $|X| < a$, the LHS is 0 and the RHS is nonnegative (absolute values are nonnegative). On the event $X \geq a$, the LHS in a and the RHS is greater than or equal to a. Then by monotonicity of expectation, linearity, and the fundamental bridge, we conclude $aP(|X| \geq a) \leq \mathbb{E}[|X|]$; then divide by a on both sides. □

Example 24.6. Markov's inequality shows that the probability *any* nonnegative random variable is more than n times its mean (when finite) is at most $1/n$.

Markov's inequality does not assume anything about the distribution of X except a finite first moment, so the bound is often quite crude. However, Markov's inequality is an important building block towards other inequalities that can be more useful (i.e. tighter). A first generalization is Chebyshev's inequality, which accounts for the second moment (i.e. variance).

Theorem 24.7 (Chebyshev's inequality)**.** *For any random variable X with $\mathbb{E}[|X|] < \infty$ and $\mathrm{Var}(X) > 0$, for all $a > 0$ we have*

$$P(|X - \mathbb{E}[X]| \geq a\mathrm{SD}(X)) \leq \frac{1}{a^2}$$

Proof. We write

$$P(|X - \mathbb{E}[X]| \geq a\mathrm{SD}(X)) = P((X - \mathbb{E}[X])^2 \geq a^2\mathrm{Var}(X)) \leq \frac{\mathrm{Var}(X)}{a^2\mathrm{Var}(X)} = \frac{1}{a^2}$$

where the inequality is by Markov's inequality applied to the random variable $Y = (X - \mathbb{E}[X])^2$. □

Example 24.8. Chebyshev's inequality states that the probability any random variable is more than 3 standard deviations from its mean is at most $1/9 \approx 0.11$. Again, this is often crude for specific distributions; if X is Normal then $(X - \mathbb{E}[X])/\mathrm{SD}(X) \sim \mathcal{N}(0, 1)$, so for $Z \sim \mathcal{N}(0, 1)$

$$P(|X - \mathbb{E}[X]| \geq 3\mathrm{SD}(X)) = P(|Z| \geq 3) = \Phi(-3) + (1 - \Phi(3)) = 2\Phi(-3) \approx 0.003$$

Chebyshev's inequality is derived from Markov's inequality using a particular transformation, standardization. Using exponential transformations instead sometimes yields tighter bounds.

Theorem 24.9 (Chernoff's bound)**.** *For any random variable X and constants $t > 0$ and $a > 0$, we have*

$$P(X \geq a) \leq \frac{\mathbb{E}[\exp(tX)]}{\exp(ta)}$$

Proof. Note $x \mapsto \exp(tx)$ is an increasing function, so the event $X \geq a$ is equivalent to the event $\exp(tX) \geq \exp(ta)$. Then

$$P(X \geq a) = P(\exp(tX) \geq \exp(ta)) \leq \frac{\mathbb{E}[\exp(tX)]}{\exp(ta)}$$

by Markov's inequality (note $\exp(tX) > 0$). □

25. Inequalities: Cauchy-Schwarz and Jensen

Today we discuss some additional inequalities that appear frequently in probability and statistics.

Theorem 25.1 (Cauchy-Schwarz)**.** *Suppose X and Y are random variables with finite second moments (the k-th moment of a random variable refers to $\mathbb{E}[X^k]$). Then*

$$|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$$

Proof. For any real number t we have

$$0 \leq \mathbb{E}[(Y - tX)^2] = \mathbb{E}[Y^2] - 2t\mathbb{E}[XY] + t^2\mathbb{E}[X^2]$$

Since this holds for *any* t, it must hold for the particular choice $t = \mathbb{E}[XY]/\mathbb{E}[X^2]$ (which minimizes the right-hand side). Plugging in gives

$$0 \leq \mathbb{E}[Y^2] - \frac{(\mathbb{E}[XY])^2}{\mathbb{E}[X^2]} \iff |\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$$

□

Cauchy-Schwarz is quite useful because it gives an upper bound on the expectation of the product of any two random variables (which may have an arbitrarily complex dependence structure) in terms of a product of second moments of the individual random variables. An important application of Cauchy-Schwarz is the correlation inequality, which states that the correlation of any two random variables lies between 0 and 1:

Corollary 25.2. *For any r.v.'s X and Y with finite second moments, $-1 \leq \mathrm{Cor}(X, Y) \leq 1$.*

Proof. We have

$$|\mathrm{Cor}(X, Y)| = \frac{|\mathrm{Cov}(X, Y)|}{\mathrm{SD}(X)\mathrm{SD}(Y)} = \frac{|\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]|}{\sqrt{\mathbb{E}[(X - \mathbb{E}[X])^2]\mathbb{E}[(Y - \mathbb{E}[Y])^2]}}$$

which we upper bound by 1 by applying Cauchy-Schwarz to the r.v.'s $X - \mathbb{E}[X]$ and $Y - \mathbb{E}[Y]$. □

A creative application of Cauchy-Schwarz also enables a certain *lower bound* on the probability that a random variable is large. By contrast, tail bounds from last lecture *upper bound* such a probability.

Theorem 25.3 (Paley-Zygmund)**.** *Suppose X is a random variable with $\mathbb{E}[X] \geq 0$ and finite second moment. Then for any $0 \leq t \leq 1$, we have*

$$P(X > t\mathbb{E}[X]) \geq (1-t)^2 \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]}$$

Proof. By linearity we can write $X = XI_{X \leq t\mathbb{E}[X]} + XI_{X > t\mathbb{E}[X]}$. Note that $XI_{X \leq t\mathbb{E}[X]} \leq t\mathbb{E}[X]$ always, so by monotonicity of expectation we have $\mathbb{E}[XI_{X \leq t\mathbb{E}[X]}] \leq t\mathbb{E}[X]$. By Cauchy-Schwarz we have

$$\mathbb{E}[XI_{X > t\mathbb{E}[X]}] \leq |\mathbb{E}[XI_{X > t\mathbb{E}[X]}]| \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[I^2_{X > t\mathbb{E}[X]}]} = \sqrt{\mathbb{E}[X^2]P(X > t\mathbb{E}[X])}$$

By linearity we conclude

$$\mathbb{E}[X] \leq t\mathbb{E}[X] + \sqrt{\mathbb{E}[X^2]P(X > t\mathbb{E}[X])}$$

Rearranging and squaring both sides gives the desired result. □

Another fundamental inequality is known as Jensen's inequality. It deals with swapping transformations and expectations when the transformation is *convex* or "concave up."

Definition 25.4. A function g is **convex** on some interval I if

$$g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y)$$

for all x, $y \in I$ and $\lambda \in [0,1]$. A function h is **concave** on I if $-h$ is convex on I.

Remark. If g is twice differentiable, it can be shown that g is convex on I if and only if $g''(x) \geq 0$ for all $x \in I$. However, non-differentiable functions can also be convex, such as $g(x) = |x|$.

Theorem 25.5 (Jensen's inequality)**.** *Suppose X is a random variable and $g : \text{supp}(X) \to \mathbb{R}$ is* convex*. Then $\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$. If $h : \text{supp}(X) \to \mathbb{R}$ is concave, then $\mathbb{E}[h(X)] \leq h(\mathbb{E}[X])$.*

Proof. Suppose g is convex. Then by the supporting hyperplane theorem, for any point $(x_0, g(x_0))$ on the graph of g there exists a line passing through that lying below the entire graph of g[1]. Let the equation of this line, at the point $(\mathbb{E}[X], g(\mathbb{E}[X]))$, be $t(x) = a + bx$ for some constants a and b,

[1]If g is twice differentiable, this line is the tangent line $t(x)$ to the graph of g at $(x_0, g(x_0))$, and the result follows immediately by showing $h(x) = g(x) - t(x)$ is minimized at $x = x_0$ via calculus and the fact that $g''(x) \geq 0$.

so that $t(\mathbb{E}[X]) = a + b\mathbb{E}[X] = g(\mathbb{E}[X])$. Then $g(X) \geq a + bX$, and by monotonicity of expectation we conclude

$$\mathbb{E}[g(X)] \geq \mathbb{E}[a + bX] = a + b\mathbb{E}[X] = g(\mathbb{E}[X])$$

The result for h concave follows immediately by applying the result in the preceding display to the convex function $-h$. □

By applying Jensen with the convex function $g(x) = x^2$, we obtain the inequality $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$ which is equivalent to the statement $\text{Var}(X) \geq 0$. We already knew that, but Jensen allows us to generalize further.

Example 25.6 (Sample standard deviation)**.** Suppose $X_1, \ldots, X_n$ are i.i.d. with finite variance σ^2. Then

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

satisfies $\mathbb{E}[S^2] = \sigma^2$. The quantity S^2 is known as the *sample variance*, as it is a function of some observations $X_1, \ldots, X_n$ equal to the (typically unknown) true variance on average. The sample standard deviation is given by $S = \sqrt{S^2}$. With $g(x) = \sqrt{x}$ concave, we conclude that $\mathbb{E}[S] \leq \sqrt{\mathbb{E}[S^2]} = \sigma$ which suggests the sample standard deviation is *negatively biased* for the true standard deviation σ.

Example 25.7 (KL divergence)**.** The **Kullback-Leibler divergence** between two distributions is a measure of how "different" two distributions are. Suppose for simplicity that we have two discrete distributions supported on a common set of n points, with PMFs specified by vectors $p = (p_1, \ldots, p_n)$ and $q = (q_1, \ldots, q_n)$. Then the KL divergence between p and q is defined as

$$d_{KL}(p, q) = \sum_{j=1}^{n} p_j \log_2(1/q_j) - \sum_{j=1}^{n} p_j \log_2(1/p_j) = -\sum_{j=1}^{n} p_j \log_2(q_j/p_j)$$

Since $g(x) = -\log_2(x)$ is convex, we have

$$d_{KL}(p, q) = \sum_{j=1}^{n} p_j g(q_j/p_j) \geq g\left(\sum_{j=1}^{n} p_j \cdot q_j/p_j\right) = g\left(\sum_{j=1}^{n} q_j\right) = g(1) = 0$$

showing that the KL divergence is nonnegative, with equality when $p = q$. Note the inequality holds by Jensen since we can view $d_{KL}(p, q)$ as the expectation of a random variable taking on the value $g(q_j/p_j)$ with probability p_j for $j = 1, \ldots, n$.

26. Laws of large numbers, central limit theorem

Laws of large numbers provide conditions under which the *sample mean* $\bar{X}_n$ of random variables $X_1, \ldots, X_n$ converges to the nonrandom quantity $\mathbb{E}[X_1] = \mu$. The random variable $\bar{X}_n$ is defined as

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Convergence of random variables is not something we have formally defined; we will make this notion precise shortly. First however, we note some important properties of sample averages when the random variables $X_1, \ldots, X_n$ are i.i.d.

Proposition 26.1. *Suppose $X_1, \ldots, X_n$ are i.i.d. from some distribution with mean μ and variance σ^2. Then $\mathbb{E}[\bar{X}_n] = \mu$ and $\mathrm{Var}(\bar{X}_n) = \sigma^2/n$.*

Proof. By linearity, we have

$$\mathbb{E}[\bar{X}_n] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{1}{n} \sum_{i=1}^{n} \mu = \mu$$

With $X_1, \ldots, X_n$ independent, by Theorem 13.6 we have

$$\mathrm{Var}(\bar{X}_n) = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}$$

□

The previous proposition indicates that while the sample average of i.i.d. random variables with common mean μ has expectation μ for all n, as n gets large, the variance of the sample mean tends to 0. This provides the intuition for one version of the Law of Large Numbers, which we formalize using Chebyshev's inequality.

Theorem 26.2 (Weak law of large numbers)**.** *Suppose $X_1, X_2, \ldots$ are i.i.d. random variables with common mean μ and variance $\sigma^2 < \infty$. For any $\epsilon > 0$, we have*

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| \geq \epsilon) = 0$$

Proof. By Chebyshev's inequality and Proposition 26.1, we have

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\mathrm{Var}(\bar{X}_n)}{\epsilon} = \frac{\sigma^2}{n\epsilon^2} \to 0 \text{ as } n \to \infty$$

□

Theorem 26.2 is known as the *weak* law of large numbers because there is a stronger law of large numbers that enables the proof of a stronger result under weaker assumptions. In particular, the strong law does not require a finite variance (note a finite variance implies a finite mean by Jensen's inequality), and shows a stronger form of convergence than in the weak law. However, it is much harder to prove, so we omit that proof here, and do not dwell on the technical distinctions between the weak and strong laws.

Theorem 26.3 (Strong law of large numbers)**.** *Suppose $X_1, \ldots, X_n$ are i.i.d. with finite mean μ. Then $\bar{X}_n$ converges to μ pointwise with probability 1. That is, the event $\bar{X}_n \to \mu$ as $n \to \infty$ has probability 1.*

Example 26.4. Given i.i.d. random variables $X_1, \ldots, X_n$, for each real number t define

$$\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \leq t}$$

By applying the (weak) law of large numbers to the i.i.d. random variables $Y_i = I_{X_i \leq t}$ (which satisfy $\mathbb{E}[Y_i] = P(X_i \leq t) = F(t)$, where F is the CDF of the X_i), we conclude $P(|\hat{F}_n(t) - F(t)| > \epsilon) \to 0$ as $n \to \infty$ for all $\epsilon > 0$. The function $\hat{F}_n$ is often known as the *empirical CDF*, and is viewed as a natural estimate of the true CDF F that generated the observations.

Example 26.5 (Monte Carlo integration)**.** Suppose f is a continuous, nonnegative function on $[a, b]$. Note this implies that for some finite constant c, we have $f(x) \leq c$ for all $x \in [a, b]$. Now suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ are the coordinates of n points chosen independently and uniformly at random in the square $[a, b] \times [0, c]$. Define

$$\bar{p}_n = \frac{1}{n}\sum_{i=1}^{n} I_{Y_i \leq f(X_i)}$$

to be the proportion of these points lying beneath the graph of f. Note

$$\mathbb{E}(\bar{p}_n) = P(Y_1 \leq f(X_1)) = \int_a^b \int_0^{f(x)} \frac{1}{c(b-a)} dy dx = \frac{1}{c(b-a)} \int_a^b f(x) dx$$

so by applying the (weak) LLN to the random variables $Z_i = c(b-a)I_{Y_i \leq f(X_i)}$, we have $P(|c(b-a)\bar{p}_n - I| > \epsilon) \to 0$ as $n \to \infty$ for all $\epsilon > 0$, where

$$I = \int_a^b f(x) dx$$

This shows that for large n, $\bar{p}_n$ can be used to approximate the integral I.

The law of large numbers is a useful result, but it does not tell us the *rate* at which the sample mean converges to the underlying "true mean." A much stronger result known as the central limit theorem (CLT) does this for us. In fact, the CLT tells us not only the rate of convergence (roughly $1/\sqrt{n}$), but also that the distribution of the sample mean for large n is approximately Normal. The CLT is truly remarkable in that it holds for any i.i.d. random variables with finite variance. Even if the distribution of the individual random variables X_i is discrete, highly skewed or otherwise very different from a Normal distribution, the sample mean $\bar{X}_n$ will have an approximately normal distribution in large samples.

Theorem 26.6 (Central limit theorem)**.** *Suppose $X_1, X_2, \ldots,$ are i.i.d. random variables with mean μ and finite variance σ^2. Define the standardized sample mean*

$$S = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right)$$

Then F_S, the CDF of S, converges pointwise to Φ, the Standard Normal CDF. That is, for each real t we have $F_S(t) \to \Phi(t)$ as $n \to \infty$.

Remark. The CLT implies the weak law of large numbers. To see this, note that for each $\epsilon > 0$ we have

$$P(|\bar{X}_n - \mu| > \epsilon) = P(|S| > \sqrt{n}\epsilon/\sigma) \leq P(S > \sqrt{n}\epsilon/\sigma) + P(S \leq -\sqrt{n}\epsilon/\sigma)$$

For any $M < \infty$, we have $\epsilon\sigma/\sqrt{n} > M$ for all sufficiently large n and hence for all such n

$$P(|\bar{X}_n - \mu| > \epsilon) \leq P(S > M) + P(S \leq -M)$$

The RHS converges to $1 - \Phi(M) + \Phi(-M) = 2\Phi(-M)$ by the CLT. This shows that

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| > \epsilon) \leq 2\Phi(-M)$$

for *any* $M < \infty$. Letting $M \uparrow \infty$, we conclude that in fact $P(|\bar{X}_n - \mu| > \epsilon) \to 0$.

27. Normal and Poisson approximation

Last lecture, we introduced the central limit theorem, which states that the sample average of n i.i.d. random variables (with finite variance) is approximately Normal as n gets large, *no matter the distribution of the individual r.v.'s.* This is not actually the first limit theorem we've seen in this course; recall that the Poisson distribution arose as a limit of the number of occurrences of a large number of rare events. Today, we will present various examples that apply these limit theorems to show how we can approximate complex probabilities. We begin with the well-known example of approximating Binomial probabilities with the Normal distribution.

Example 27.1. Suppose $n = 2,000$ fair coins are flipped. Let's give an approximation to the probability that more than 51% of them land heads in terms of Φ, the CDF of the Standard Normal distribution. Define W to be the number of heads we obtain; we know $W \sim \text{Bin}(n, 0.5)$. Then the exact probability we're looking for can be written as

$$P(W/2000 > 0.51) = P(W > 1020) = 0.5^{2000} \sum_{k=1021}^{2000} \binom{2000}{k}$$

using the Binomial PMF. However, this is quite an unwieldy sum that cannot even be calculated on most computers due to overflow, without some special tricks. An alternative is to write $W = I_1 + \cdots + I_{2000}$ where I_i is the indicator r.v. for the event A_i that the i-th coin lands heads. Then the $I_i \sim \text{Bern}(0.5)$ are i.i.d. Recall $\mathbb{E}[I_i] = 0.5$ and $\text{Var}(I_i) = 0.5(1 - 0.5) = 0.25$. As $W/2000$ is nothing more than the sample mean of the I_i, by the CLT we know $Z = \sqrt{2000}\left(\frac{W/2000 - 0.5}{\sqrt{0.25}}\right)$ has approximately a Standard Normal distribution. Thus

$$P(W/2000 > 0.51) = P(Z > 0.01\sqrt{8000}) \approx \boxed{1 - \Phi(2/\sqrt{5})} = 0.186$$

An equivalent, alternative way of doing the approximation is to recall Exercise 15.5, which shows that the unscaled W is itself approximately Normal. The mean and variance of this Normal

approximation should match those of W. Generalizing slightly, we conclude that if $X \sim \text{Bin}(n,p)$ then for n sufficiently large, we have $X \dot{\sim} \mathcal{N}(np, np(1-p))$, where the symbol $\dot{\sim}$ should be read "is approximately distributed as." We emphasize that the validity of the approximation stems from X being a rescaled/recentered sample average of i.i.d. random variables. Returning to our example, we have $n = 2000$ and $p = 0.5$, hence for $Y \sim \mathcal{N}(1000, 500)$ we have

$$P(W > 1020) \approx P(Y > 1020) = 1 - \Phi(2/\sqrt{5})$$

where the equality above can be derived by standardizing Y.

Remark (Continuity correction)**.** Since the Binomial distribution is discrete while the Normal distribution is continuous, it is customary to apply the following "continuity correction" to approximate a Binomial probability. For each integer k in the support of a r.v. $X \sim \text{Bin}(n,p)$, we approximate

$$\begin{aligned} P(X = k) &\approx P(k - 1/2 \leq Y \leq k + 1/2) \\ &= P\left(\frac{k - 1/2 - np}{\sqrt{np(1-p)}} \leq \frac{Y - np}{\sqrt{np(1-p)}} \leq \frac{k + 1/2 - np}{\sqrt{np(1-p)}}\right) \\ &\approx \Phi\left(\frac{k + 1/2 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k - 1/2 - np}{\sqrt{np(1-p)}}\right) \end{aligned}$$

for $Y \sim \mathcal{N}(np, np(1-p))$. Applying this to the previous example, we'd get the approximation

$$P(W/2000 > 0.51) = \sum_{k=1021}^{2000} P(W = k) \approx \Phi\left(\frac{2000.5 - 1000}{\sqrt{2000 \cdot 0.5 \cdot (1 - 0.5)}}\right) - \Phi\left(\frac{1020.5 - 1000}{\sqrt{2000 \cdot 0.5 \cdot (1 - 0.5)}}\right)$$

which comes out to 0.17963, an accurate approximation to 5 decimal places!

Here are some other examples of Normal approximations:

Example 27.2. If $X \sim \text{Pois}(n)$, then for large n, $X \dot{\sim} \mathcal{N}(n, n)$, since X has the same distribution as the sum of n i.i.d. Pois(1) r.v.'s.

Example 27.3. If $X \sim \text{Gamma}(n, \lambda)$, then for large n, $X \dot{\to} \mathcal{N}(n/\lambda, n/\lambda^2)$ since X has the same distribution as the sum of n i.i.d. Expo(λ) r.v.'s.

Now we consider Poisson approximations. Recall that in the regime where $n \to \infty$ and $p \to 0$ with $np \to \lambda$, if $X \sim \text{Bin}(n,p)$ then $X \dot{\sim} \text{Pois}(np)$. By contrast, the Normal approximation for the Binomial is most effective if p is fixed (or at least bounded away from 0) as $n \to \infty$. We note that if

we tried to use the Poisson approximation for np large then by Example (27.2) we shouldn't do too poorly (though the Normal will generally still be better). By contrast, the Normal approximation can fail spectacularly if np (or $n(1-p)$) are small. Classical textbooks typically advise having $\min(np, n(1-p)) \geq 10$ to use a Normal approximation.

Example 27.4. Suppose $X \sim \text{Bin}(2000, 0.001)$. By Poisson approximation, for $Y \sim \text{Pois}(2)$ we have

$$P(X = 0) \approx P(Y = 0) = \exp(-2) \approx 0.1353$$

The true answer is $(1-0.001)^{2000} = 0.1352$ (we can see mathematically that this should be close to $\exp(-2)$ by recalling that $e^x = \lim_{n\to\infty}(1 + x/n)^n$). On the other hand, the normal approximation with continuity correction gives

$$\Phi\left(\frac{0.5-2}{\sqrt{2000 \cdot 0.001 \cdot (1-0.001)}}\right) - \Phi\left(\frac{-0.5-2}{\sqrt{2000 \cdot 0.001 \cdot (1-0.001)}}\right) = 0.1058$$

We conclude with some additional examples of Poisson approximation.

Example 27.5. Suppose m people are in a room. To estimate the approximate probability that exactly 2 people have the same birthday, we create indicators $I_1, \ldots, I_{\binom{m}{2}}$ for whether each of the $\binom{m}{2}$ pairs of people have the same birthday. Note $\mathbb{E}(I_i) = 1/365$ (conditional on the birthday of the first person in any pair, the second person has probability $1/365$ of matching that birthday). Letting $X = I_1 + \cdots + I_{\binom{m}{2}}$, we note that for m moderately large, we have $X \stackrel{\cdot}{\sim} \text{Pois}\left(\binom{m}{2} \cdot \frac{1}{365}\right)$. Note the indicators are not independent since e.g. if I_i corresponds to persons A and B while I_j corresponds to persons A and C, then $I_i = I_j = 1$ implies $I_k = 1$ where I_k is the indicator for persons B and C. But we can ignore this dependence and still get a reasonably good approximation using the Poisson PMF:

$$P(X = 1) \approx \exp\left(-\binom{m}{2} \cdot \frac{1}{365}\right) \cdot \binom{m}{2} \cdot \frac{1}{365}$$

Example 27.6. A random survey of 1000 people is taken from a city of 10^6 people *with replacement*, and we want to approximate the probability that at least one person is surveyed two or more times. We can create $\binom{1000}{2}$ indicators $I_1, \ldots, I_{\binom{1000}{2}}$ for whether each pair of selected people contains two of the same person. Each of these indicators has expectation $1/10^6$. Then for $X = I_1 + \cdots + I_{\binom{1000}{2}}$ we desire

$$P(X \geq 1) = 1 - P(X = 0) \approx 1 - \exp\left(-\frac{\binom{1000}{2}}{10^6}\right) \approx 1 - \exp(-1/2)$$

which is correct up to 3 decimal places. Note you could also consider 10^6 indicators, one for each person in the city on whether they're selected twice, but this leads to a slightly messier calculation.

28. Moment generating functions

Moment generating functions are a powerful mathematical tool in probability, often simplifying the proofs of many results.

Definition 28.1. Suppose X is a random variable. Then its **moment generating function (MGF)** is $M_X(t) = \mathbb{E}[\exp(tX)]$. The MGF is said to exist if for some $a > 0$, $\mathbb{E}[\exp(tX)] < \infty$ for all $t \in (-a, a)$.

Example 28.2. If $X \sim \text{Bern}(p)$, then for all $t \in \mathbb{R}$ we have

$$M_X(t) = \exp(t) \cdot p + \exp(0) \cdot (1-p) = 1 - p(1 - \exp(t))$$

Example 28.3. If $X \sim \mathcal{N}(0,1)$, then for all $t \in \mathbb{R}$

$$\begin{aligned} M_X(t) = \int_{-\infty}^{\infty} \exp(tx)\varphi(x)dx &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(tx - \frac{x^2}{2}\right) dx \\ &= \frac{\exp\left(\frac{t^2}{2}\right)}{\sqrt{2\pi}} \left[\int_{-\infty}^{\infty} \exp\left(-\frac{(x-t)^2}{2}\right) dx\right] \\ &= \exp\left(\frac{t^2}{2}\right) \text{ since the } \mathcal{N}(t,1) \text{ PDF integrates to 1} \end{aligned}$$

Proposition 28.4. *Let X be a r.v. whose MGF M_X exists. Then for any constants a and b, the MGF of $Y = a + bX$ is given by $M_Y(t) = \exp(at)M_X(bt)$ whenever $M_X(bt)$ is finite.*

Proof. We use linearity of expectation and the definition of the MGF:

$$M_Y(t) = \mathbb{E}[\exp((a+bX)t)] = \exp(at)\mathbb{E}[\exp(bXt)] = \exp(at)M_X(bt)$$

□

Example 28.5. Recall that if $X \sim \mathcal{N}(\mu, \sigma^2)$, then X has the same distribution as $\mu + \sigma Z$ for $Z \sim \mathcal{N}(0,1)$. By Example 28.3 and the preceding proposition, we conclude that

$$M_X(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)$$

Proposition 28.6. *Suppose $X_1, \ldots, X_n$ are independent and let $Y = X_1 + \cdots + X_n$. Then for any t for which $M_{X_i}(t)$ is finite for all i, we have*

$$M_Y(t) = \prod_{i=1}^{n} M_{X_i}(t)$$

Proof. Since $X_1, \ldots, X_n$ are independent, so are $\exp(tX_1), \ldots, \exp(tX_n)$. But then $\exp(tX_1), \ldots, \exp(tX_n)$ are uncorrelated. Hence the expectation of their product is the product of their expectations:

$$M_Y(t) = \mathbb{E}[\exp(t(X_1 + \cdots + X_n))] = \mathbb{E}\left[\prod_{i=1}^{n} \exp(tX_i)\right] = \prod_{i=1}^{n} \mathbb{E}[\exp(tX_i)] = \prod_{i=1}^{n} M_{X_i}(t)$$

□

Example 28.7. If $X \sim \text{Bin}(n, p)$, then $M_X(t) = (1 - p(1 - \exp(t)))^n$ for all $t \in \mathbb{R}$.

What makes the MGF so powerful is that it determines the distribution, just like the CDF or PDF/PMF. One caveat is that the MGF does not always exist. However, it does for most common distributions whose tails aren't too heavy. In that case, MGFs are often simpler computationally.

Theorem 28.8 (MGF determines the distribution)**.** *Let X and Y be r.v.'s such that $M_X(t) = M_Y(t)$ for all $t \in (-a, a)$ for some $a > 0$. Then X and Y have the same distribution.*

The proof is omitted, as it is quite technical. However, we can now use MGFs to provide a simpler proof of something we already knew.

Example 28.9. Suppose $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ are independent. Using convolutions, we showed in Corollary 20.8 that $X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$. We could alternatively give a one-line proof with MGFs, using the preceding theorem and our computation of the Normal MGF; for all $t \in \mathbb{R}$ we have

$$M_{X+Y}(t) = M_X(t)M_Y(t) = \exp\left(\mu_X t + \frac{\sigma_X^2 t^2}{2}\right)\exp\left(\mu_Y t + \frac{\sigma_Y^2 t^2}{2}\right) = \exp\left((\mu_X + \mu_Y)t + \frac{(\sigma_X^2 + \sigma_Y^2)t^2}{2}\right)$$

which matches the $\mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$ MGF.

We now describe another use of MGFs, as its name suggests: to compute moments (the moments of a r.v. X are the quantities $\mathbb{E}[X^a]$ for positive integers a).

Proposition 28.10. *For any positive integer n and r.v. X whose MGF exists, we have $\mathbb{E}[X^n] = M_X^{(n)}(0)$, where $M_X^{(n)}$ denotes the n-th derivative of M_X. Equivalently, $\mathbb{E}[X^n]$ is the coefficient of $t^n/n!$ in the Taylor expansion of $M_X(t)$ about $t = 0$.*

Proof. The trick is to recall the Taylor expansion of the exponential, so that

$$M_X(t) = \mathbb{E}[\exp(tX)] = \mathbb{E}\left[\sum_{n=0}^{\infty} \frac{(tX)^n}{n!}\right] = \sum_{n=0}^{\infty} \frac{t^n}{n!}\mathbb{E}[X^n]$$

On the other hand, we can Taylor expand $M_X(t)$ directly about $t = 0$:

$$M_X(t) = \sum_{n=0}^{\infty} \frac{t^n}{n!} M_X^{(n)}(0)$$

By matching coefficients we have $\mathbb{E}[X^n] = M_X^{(n)}(0)$ for each n.[1] □

Example 28.11 (Standard Normal moments)**.** For $X \sim \mathcal{N}(0,1)$ we can write

$$M_X(t) = \exp(t^2/2) = \sum_{k=0}^{\infty} \frac{t^{2k}}{2^k k!} = \sum_{k=0}^{\infty} \frac{t^{2k}}{(2k)!} \cdot \frac{(2k)!}{2^k k!}$$

which shows $\mathbb{E}[X^{2k}] = (2k)!/(2^k k!)$ for every positive integer k, and $\mathbb{E}[X^n] = 0$ for any odd n.

Example 28.12 (Exponential moments)**.** If $X \sim \text{Expo}(1)$ then its MGF is

$$M_X(t) = \int_0^{\infty} \exp(tx)\exp(-x)dx = \int_0^{\infty} \exp(-(1-t)x)dx = \frac{1}{1-t}, \quad t < 1$$

By the formula for the sum of a geometric series, we have

$$\frac{1}{1-t} = 1 + t + t^2 + \cdots = \sum_{k=0}^{\infty} \frac{t^k}{k!} \cdot k!, \quad -1 < t < 1$$

Thus $\mathbb{E}[X^n] = n!$ for all positive integers n, and for $Y \sim \text{Expo}(\lambda)$ we have $\mathbb{E}[Y^n] = n!/\lambda^n$, since $\lambda Y \sim \text{Expo}(1)$.

There is one additional powerful result concerning MGFs that allows for a proof of the central limit theorem under regularity conditions:

Theorem 28.13 (Levy's continuity theorem)**.** *Suppose M is the CDF of a continuous random variable X, and let $M_1, M_2, \ldots$ be the MGFs of a sequence of random variables $X_1, X_2, \ldots$, assumed to all exist on $(-a, a)$ for some $a > 0$. Then if $\lim_{n\to\infty} M_n(t) = M(t)$ for all $t \in (-a, a)$, the CDFs $F_{X_1}, F_{X_2}, \ldots$ of $X_1, X_2, \ldots$, converge pointwise to the CDF F_X of X.*

[1] We have omitted some technical arguments to guarantee M_X is infinitely differentiable and that the infinite sum and expectation can be interchanged. The ability to match coefficients of infinite series that agree on an interval follows for instance by Theorem 8.5 of Walter Rudin's *Principles of Mathematical Analysis*.

Example 28.14. Suppose $X_1, X_2, \ldots$ are i.i.d. mean 0, variance 1 r.v.s with common MGF M. Then by the fact that the X_i are independent, the MGF of $Y_n = \sqrt{n}\bar{X}_n = \frac{1}{\sqrt{n}}(X_1 + \cdots + X_n)$ is

$$M_{Y_n}(t) = \mathbb{E}[\exp(t/\sqrt{n}(X_1 + \cdots + X_n))] = \prod_{i=1}^{n} \mathbb{E}[\exp(tX_i/\sqrt{n})] = (M(t/\sqrt{n}))^n$$

Now by Taylor's theorem we have

$$M(t/\sqrt{n}) = M(0) + M'(0) \cdot \frac{t}{\sqrt{n}} + \frac{M''(0)}{2} \cdot \frac{t^2}{n} + R_n = 1 + \frac{t^2}{2n} + R_n$$

since $M'(0) = \mathbb{E}[X_1] = 0$ and $M''(0) = \mathbb{E}[X_1^2] = 1$. Above, R_n is a remainder term such that $nR_n \to 0$ as $n \to \infty$. Then

$$\log M_{Y_n}(t) = n \log(M(t/\sqrt{n})) = n \log\left(1 + \frac{t^2}{2n} + R_n\right) = n\left(\frac{t^2}{2n} + R_n'\right)$$

where R_n' is another remainder term satisfying $nR_n' \to 0$ as $n \to \infty$, and for the last equality we've used Taylor's theorem on $f(x) = \log(1 + x)$ to argue that $\log(1 + x) = x + R(x)$ where $xR(x) \to 0$ as $x \to 0$. Taking limits on both sides, we get

$$\lim_{n\to\infty} \log M_{Y_n}(t) = \frac{t^2}{2}$$

Exponentiating and recalling the Standard Normal MGF $M_Z(t) = \exp(t^2/2)$ gives the desired result.