

Lecture 9: A/B testing

Stats 32: Introduction to R for Undergraduates

Harrison Li

April 30, 2024

Agenda

- 1 Null and alternative hypotheses
- 2 t-tests of averages
- 3 Tests of proportions

Reading:

<https://machinelearningmastery.com/statistical-hypothesis-tests/>

Null and alternative hypotheses

Null and alternative hypotheses

A credit card company is running an online advertising campaign. They have a proposed new ad featuring an athlete (version B) and want to know whether it leads to more purchases than the existing ad (version A).

A/B test: Recruit 10,000 credit card users, randomly assign 5,000 to get new ad (treatment group), other 5,000 to get existing ad (control group)

Those who saw version B averaged 3.2 more purchases than those who saw version A. Is version B better?

Null and alternative hypotheses

A/B testing is based on the standard statistical framework of hypothesis testing. There are two hypotheses, which are conflicting statements about the world that we want to distinguish between using data.

The **null hypothesis** (typically denoted H_0) corresponds to the “less interesting” or “default” state of the world, while the **alternative hypothesis** (typically denoted H_1 or H_A) corresponds to the “more interesting” state of the world.

By default, we assume the null hypothesis is true. However, if the data we observe is **unlikely** to occur due to chance *under the assumption that the null hypothesis is true*, then we “reject” the null hypothesis in favor of the alternative hypothesis.

Null and alternative hypotheses

Examples of null hypotheses:

- Ads A and B lead to the same number of clicks, on average
- Diabetes drug D does not have any effect on insulin levels
- Global mean temperatures are not changing over time

Corresponding alternative hypotheses:

- Ads A and B do not lead to the same number of clicks, on average
- Diabetes drug D changes insulin levels
- Global mean temperatures are changing

One-sided vs. two-sided alternatives

The alternative hypotheses on the previous slide are all **two-sided alternatives**. Contrast these with the following **one-sided alternatives**:

- Ad B leads to more clicks than ad A
- Diabetes drug D reduces insulin levels
- The Earth is warming

Data-generating processes

Note hypotheses are not statements about your data, but rather statements about the underlying **data-generating process** (“state of the world”)

For example, we (implicitly) assume that the 10,000 users in our A/B test are randomly sampled from a much larger population. We are using the limited information from these users to try and figure out the true effectiveness of ad B for the average person in the “infinite” population of all users.

Structure of a hypothesis test

A *hypothesis test* aims to distinguish between H_0 and H_1 using a *test statistic*, a numerical summary of the data. A common example of a test statistic is the sample average (the mean of all observations).

- 1 Determine a desired *significance level* α , e.g. 0.01 or 0.05.
- 2 Collect some data.
- 3 Compute the test statistic from the data.
- 4 Calculate a *p-value* using the test statistic.
- 5 Reject H_0 if the *p-value* is less than α , otherwise fail to reject H_0 .

Type I vs. type II error

Suppose H_0 is true. If a test (falsely) rejects H_0 , it's said to commit a *Type I error*. Example: Ads A and B actually have the same performance, but we reject the null hypothesis that this is the case.

Conversely, suppose H_0 is false. If a test (falsely) fails to reject H_0 , it's said to commit a *Type II error*. Example: Ad B actually leads to more clicks than ad A (on average over the population of all users), but the test fails to reject the null hypothesis

It's important to emphasize that H_0 is either true or not — it's not random. Also, it can't be both true and false at the same time. The core problem is, we don't know which it is. We can only make an educated guess using a hypothesis test.

Significance levels

The *significance level* α of a test is the probability the test makes a Type I error. It is under the control of the scientist for a given testing procedure, and typically set (arbitrarily) to something like 0.05 or 0.01.

It's important to note it only makes sense to speak of the probability of a type I error when H_0 is true. When H_0 is false, it's impossible to make a Type I error, by definition!

Power

A test that never rejects H_0 will never make a Type I error, but is also not useful when H_0 is false — it will never find the good ads!

Thus, a good test also has a small Type II error rate. Statisticians use the term *power* to denote the chance of (correctly) rejecting H_0 when it is false.

power + type II error rate = 1 or 100%, by definition.

p-values

The **p-value** for a hypothesis test is the *probability of observing a test statistic more extreme than the observed value, assuming H_0 is true*.

The p-value is a mathematical computation based on the value of the test statistic and the nature of the null hypothesis. More extreme test statistics, relative to H_0 , give smaller p-values.

Suppose we obtain a p-value of 0.03 based on the 200 participants in our trial. If H_0 were true, we should only see a p-value smaller than 0.03, i.e. a test statistic more extreme than the one we actually got, 3% of the time. At a significance level threshold of 0.05, this is sufficiently small to reject H_0 (but not if we chose $\alpha = 0.01$).

What a p-value is NOT

A p-value is NOT the probability H_0 is true. The truth of H_0 is not random, it is just unknown.

What a p-value is NOT

A small p-value also does not tell you that an effect is actually meaningful to a human.

In fact, it's a bit of straw man, just telling you whether the data is consistent with a counterfactual that the true effect is 0.

Let's say ad B increases watch time by 0.002 purchases/day on average. Technically this is not 0 so with a large enough sample size you will frequently (correctly) reject the null. But in this case it may not be worth paying the athlete for this more expensive ad.

What a p-value is NOT

Conversely, maybe the true effect is huge but your sample size is too small for your test to have reasonable power.

Then you will likely make a type II error. Thus you should not construe a failure to reject as evidence **for** the null.

t-tests of averages

t-tests of averages

We will not get into the mathematical details of how to design good test statistics or compute p-values (take Stats 60 for that), but seek instead to understand some common examples and their R implementations.

One sample t-test

In a **one-sample t-test**, we have n independent, quantitative observations, and H_0 is that the true mean of the data-generating process is equal to a particular value. The test statistic used, as you might expect, is the sample average of all the observations.

For example, we might want to test whether the average abalone length is equal to 0.5.

Note the sample mean here is 0.5272. Is this extreme under H_0 ? The p -value will tell us.

```
library(tidyverse)
abalone <- read_csv("abalone.csv")
mean(abalone$Length)

## [1] 0.5272167
```

One sample t-test

Let μ be the average abalone length (in the population our data comes from). Remember μ is a characteristic of the data-generating process, not our particular sample.

Our hypotheses (note the two-sided alternative) are then

$$H_0 : \mu = 0.5$$

$$H_1 : \mu \neq 0.5$$

One sample t-test

$$H_0 : \mu = 0.5$$

$$H_1 : \mu \neq 0.5$$

To get the p-value for these hypotheses on the data, we use `t.test()`. It is the probability, assuming H_0 were true, of observing a sample mean greater than the observed 0.5272 OR less than $0.5 - (0.5272 - 0.5) = 0.4728$ (for a two-sided alternative, “more extreme” means “farther away in absolute value”).

Note the mathematical validity of the p -value depends on an assumption that the different abalone's lengths are independent and random.

One sample t-test

```
t.test(x=abalone$Length, mu=0.5, alternative="two.sided")  
  
##  
## One Sample t-test  
##  
## data: abalone$Length  
## t = 3.7817, df = 299, p-value = 0.0001881  
## alternative hypothesis: true mean is not equal to 0.5  
## 95 percent confidence interval:  
## 0.5130535 0.5413798  
## sample estimates:  
## mean of x  
## 0.5272167
```

With such a small p-value of 0.00019 on our two-sided test, we emphatically reject H_0 at any reasonable significance level.

One-sided test

We could also perform a one-sided t -test, for the alternative $H_1 : \mu > 0.5$:

```
t.test(x=abalone$Length, alternative="greater", mu=0.5)
##
## One Sample t-test
##
## data:  abalone$Length
## t = 3.7817, df = 299, p-value = 9.405e-05
## alternative hypothesis: true mean is greater than 0.5
## 95 percent confidence interval:
##  0.5153419      Inf
## sample estimates:
## mean of x
## 0.5272167
```

Now the p -value of $9.405e-05$, which means $9.405 \cdot 10^{-5}$, is the probability (assuming H_0 is true) of observing a sample mean **greater** than 0.5272. With such a small p -value, we again reject H_0 emphatically.

One-sided test

What if we had the other one-sided alternative hypothesis
 $H_1 : \mu < 0.5$?

```
t.test(x=abalone$Length, alternative="less", mu=0.5)
##
## One Sample t-test
##
## data:  abalone$Length
## t = 3.7817, df = 299, p-value = 0.9999
## alternative hypothesis: true mean is less than 0.5
## 95 percent confidence interval:
##      -Inf 0.5390914
## sample estimates:
## mean of x
## 0.5272167
```

Now the p -value of 0.999 suggests it is very likely to observe a sample mean **less** than 0.5272 when the true mean is 0.5. We certainly do not reject our null hypothesis here.

Two sample t-test

Another common testing scenario is to check if *two* groups of independent observations come from data-generating processes with different means. For example, do history majors have a higher average IQ than English majors?

As another example let μ_1 be the average `arr_delay` for flights from JFK, and μ_2 be the average `arr_delay` for flights from LGA. How would you interpret the following set of hypotheses, in words?

$$\mu_1 - \mu_2 = 0$$

$$\mu_1 - \mu_2 > 0$$

Two sample t-test

The two sample t-test, also using `t.test()`, can be used to test the hypotheses on the previous slide, *assuming all the flights are independent*:

```
library(nycflights13)
jfk_flights <- flights %>%
  filter(origin=="JFK")
lga_flights <- flights %>%
  filter(origin=="LGA")
t.test(x=jfk_flights$arr_delay, y=lga_flights$arr_delay, mu=0)

##
## Welch Two Sample t-test
##
## data:  jfk_flights$arr_delay and lga_flights$arr_delay
## t = -1.2062, df = 209301, p-value = 0.2277
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.6089919  0.1449775
## sample estimates:
## mean of x mean of y
##  5.551481  5.783488
```

Two-sample t-test

The p -value of 0.2278 means that there is a 0.2278 probability that the sample mean JFK delay minus the sample mean LGA delay would be more than $5.76 - 5.55 = 0.23$ or less than -0.23 (recall our alternative is two-sided).

With such a high p -value, we fail to reject H_0 , since if H_0 were true, it'd be plausible to see a difference in mean arrival delays as extreme as we did.

Note the independence assumption is probably not valid here. If New York City had bad weather on a given day, you'd likely see delays at both JFK and LGA.

Paired t-test

Since we have the same users before and after the change, the watchtime before and after the change are not independent.

However, it is typically assumed that different users are independent (not a perfect assumption in practice).

In that case, we can recover independence by taking differences: Let D_i be user i 's watchtime after the change minus their watch time before the change.

Then perform a **one-sample** t-test on the differences D_i .

Tests of proportions

Tests of proportions

The t -tests we discussed are designed to test hypotheses about averages of continuous, quantitative variables.

Sometimes, we instead want to test hypotheses about proportions.

For example, given a poll of 600 randomly chosen registered voters in Wisconsin, we may want to test if the true support of presidential candidate A is 50%.

One sample test of proportions

Suppose 310 of the WI voters polled said they would support candidate A. We can test whether this provides enough evidence to reject the null that the candidate has 50% support using `prop.test()` in R:

```
prop.test(x=310, n=600, p=0.5, alternative="two.sided")  
  
##  
## 1-sample proportions test with continuity correction  
##  
## data: 310 out of 600, null probability 0.5  
## X-squared = 0.60167, df = 1, p-value = 0.4379  
## alternative hypothesis: true p is not equal to 0.5  
## 95 percent confidence interval:  
## 0.4758723 0.5572445  
## sample estimates:  
## p  
## 0.5166667
```

With a p -value of 0.43, we fail to reject H_0 .

Two sample test of proportions

We can also test a difference in proportions in two different populations.

For example, suppose 800 voters in Michigan were also polled, with only 385 of them supporting candidate A.

We want to test the null that the proportion of support in Michigan and Wisconsin is the same.

Two-sample test of proportions

```
prop.test(x=c(310, 385), n=c(600, 800), alternative="two.sided")  
  
##  
## 2-sample test for equality of proportions with continuity correction  
##  
## data:  c(310, 385) out of c(600, 800)  
## X-squared = 1.5816, df = 1, p-value = 0.2085  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## -0.01893398  0.08976732  
## sample estimates:  
##      prop 1      prop 2  
## 0.5166667 0.4812500
```

Note: The validity of the p-value in a two sample test of proportions hinges on the samples being independent.