Permutation Tests

Hypothesis Testing

Suppose I flip a coin 1000 times.

If the coin were fair, we would expect it to come up heads about 500 times.

The coin comes up heads 461 times.

Is the coin fair?

When collecting data, there is almost always some random variability.

Hypothesis testing is a process by which we determine whether the disparity between what was observed and what was expected could reasonably be explained by random variability, or whether it indicates our initial assumption may be incorrect. We use sample data to compare two competing hypotheses regarding a population parameter.

We compare two hypotheses: the **null hypothesis** and the **alternative hypothesis**.

The null hypothesis is some starting assumption. Most often, this is a claim that the population parameter is equal to some particular value.

The alternative hypothesis is a claim that the null hypothesis is wrong in some way. Most often, this is a claim that the population parameter is different in some way from the value claimed in the null hypothesis.

We tend to abbreviate the phrase "the null hypothesis" as H_0, and we abbreviate the phrase "the alternative hypothesis" as H_a.

In our example,

Null hypothesis: The coin is fair, which would mean that for any given flip, P(heads) is 0.5.

Alternative hypothesis: The coin is not fair, which would mean that for any given flip, P(heads) = something other than 0.5.

H_0: $p = 0.5$

H_a: $p \neq 0.5$

As another example, suppose we are studying the neck girths of polar bears. We know from past experience that polar bears living in the wild have a mean neck girth of 35". We are curious to see if this holds true for polar bears living in captivity as well, or if they tend to be smaller than polar bears living in the wild. We could set up our hypotheses:

There are a few general patterns we should note with how the null and alternative hypotheses are set up.

- Both H_0 and H_a are statements about the same parameter
- Both H_0 and H_a compare the parameter to the same numerical value
- H_0 is always a statement that the parameter is equal to that value
- H_a will be one of three options:
  - a statement that the parameter is greater than the specified value
  - a statement that the parameter is less than the specified value
  - a statement that the parameter is not equal to the specified value

Also, note that the hypotheses are not based on any sample data collected. The null and alternative hypotheses are determined based upon the question of interest, before we examine the sample data.

Conducting a Hypothesis Test

If the coin were, in fact, fair, how unusual is our observed result?

If our result isn't that unusual, then it's plausible the coin could be fair – what we observed is consistent with what we would expect to observe if the coin were fair.

What if our result is unusual? If what we observed is not consistent with what we would expect to observe if the coin were fair, then we can conclude that the most reasonable explanation is that the coin is likely not fair.

How do we measure "unusual"?

If you flip a fair coin 1000 times, the probability of getting 461 heads is 0.0012.

But if you flip a fair coin 1000 times, the probability of getting 500 heads is 0.0252.

Even the most likely outcome has a very low probability.

What seems unusual is not the exact value of 461, but rather the fact that it's far away from what we would expect.

Rather than asking "What is the probability of getting 461 heads?" we can ask "What is the probability of being at least that far away from what we expect?"

A **p-value** refers to the probability of observing data at least as far away from the expected value as what we actually observed.

We start by assuming the null hypothesis is true.

We then calculate how unusual our sample data would be, by finding the probability of observing sample data at least as extreme as ours, if the null hypothesis were true.

If the sample data would have been unusual when we assumed the null hypothesis is true, then we decide it was incorrect to assume the null hypothesis was true.

Often we use a probability of 0.05 as a cutoff for "unusual" (we will get into a deeper discussion of appropriate cutoffs in a later chapter). More generally, we will use to refer to our cutoff for "unusual". This cutoff is often referred to as the **significance level** of a test. The most common significance level to use is 0.05. If we want to require stronger evidence to convince us the null hypothesis is wrong, we could pick a smaller significance level. If we are okay with weaker evidence still being convincing, we could pick a larger significance level.

In general, we can think of three key steps in conduction a hypothesis test:

1. State our null and alternative hypotheses
2. Find a p-value
3. State our conclusions

Step 1 is a process of formalizing the question we are interested in.

Step 2 is where all of the math and / or coding comes into play.

Step 3 is interpreting and explaining the results.

Once we have found our p-value, we compare it to . We make one of two conclusions:

- If our p-value is less than or equal to , we reject the null hypothesis in favor of the alternative hypothesis. We make a conclusion like "There is evidence that the mean neck girth for polar bears living in captivity is less than 35 inches."
- If our p-value is greater than , we do not reject the null hypothesis. We make a conclusion like "There is not evidence that the mean neck girth for polar bears living in captivity is less than 35 inches."

461 heads is 39 heads fewer than what we would expect if the coin were fair.

Our p-value will be the probability, if the coin really were fair, of being at least 39 away from 500.

That is, we want P($X \leq 461$ or $X \geq 539$).

We can use the binomial distribution to find this. The probability comes out to 0.0149.

```
2*pbinom(461, 1000, .5, lower.tail=TRUE)
```

That is, if we had a fair coin, there's only about a 1.5% chance that, in 1000 flips, we'd be off from our expected value by at least 39 heads.

This means it would be very unusual to have observed 461 heads in 1000 flips, if the coin were fair (most often we use 5% as a threshold for "unusual").

Therefore, we reject the null hypothesis in favor of the alternative hypothesis. The most reasonable explanation for what we have observed is that the coin is not fair.

One-Sided and Two-Sided Tests

How we find a p-value depends upon the specific alternative hypothesis we are considering. In the above example, our alternative hypothesis was H_a: $p \neq 0.5$. That is, we were looking for evidence that the probability of heads was different from .5 in either direction. Consequently, when we found our p-value, we found the probability of being at least 39 away from 500 in either direction. We did so by finding the probability in one direction, and multiplying it by 2 to include an equal probability in the other direction as well.

When the alternative hypothesis involves ≠, we refer to it as a two-sided test. When we have a two-sided test, the p-value will be found by finding the probability of being at or beyond our observed value, and then doubling it to include an equal probability in the other direction as well.

On the other hand, some alternative hypotheses involve > or <. We refer to these as one-sided tests. In these situations, we are only interested in the probability of being away from our expected value in one direction. Consequently, we do not double the probability we calculate when finding a p-value for a one-sided test. We will look at an example of a one-sided test in our next example.

Permutation & Randomization Tests

Rather than rely on a probability distribution (like the binomial in this last example, or like normal distributions in many classical statistics methods), we can often find a p-value by thinking about our null hypothesis in terms

of **interchangeability**.

Suppose we want to test the effectiveness of cloud seeding on producing rain.

We will look at a study from 1975 by Joanne Simpson, Anthony Olsen, and Jane C. Eden, published in “A Bayesian Analysis of a Multiplicative Treatment Effect in Weather Modification.”

Between 1968 and 1972, experiments were conducted in which cumulus clouds were seeded with silver nitrate in the hopes of increasing cloud growth and prolonging cloud life, with the intention of leading to increased rainfall. Total rainfall (in acre-feet) was measured for 26 seeded clouds and 26 unseeded clouds.

We start with exploratory data analysis. Let's load in the data set, and examine it visually.

The cloud seeding data set can be downloaded here .

The data is stored as an Excel file. We can either use the import options we saw before, or we can use the readxl package (this is part of tidyverse, so you won't need to install it, but you will need to load it separately, because it is not considered a "core" package in tidyverse in the way that things like ggplot2 or dplyr are).

When reading the file in via the read_excel() function, we could type out the path name for the file. Or we could use the file.choose() function, which will pop up a window to let us browse and find the file. Using file.choose() is convenient when you're doing a quick task, or exploring a new data set. But if you were designing production-ready code, or something that you would want to run repeatedly, it would be better to give the file path, to avoid having to navigate through the window that pops up each time file.choose() is run.

We will create histograms for the rainfall amounts for both the treatment and control groups.

```
library(tidyverse)
library(readxl)
clouds <- read_excel(file.choose())
glimpse(clouds)

ggplot(data=clouds, mapping=aes(x=Rainfall)) +
  geom_histogram() +
  facet_wrap(~Treatment, nrow=2)
```

We can see that there's some evidence we might have increased rainfall for seeded clouds. But is the difference that we see within the realm of what might just be random noise? Or would we only expect to see a difference like this if there were some true underlying effect from cloud seeding? We will use a hypothesis test to determine this.

We can also see that the distributions of rainfall amounts are very skewed, suggesting that medians, rather than means, are more appropriate for summarizing this data. Our null hypothesis, then, will be that the medians for the two groups are equal, while our alternative hypothesis will be that the median rainfall is greater for seeded clouds than for unseeded clouds.

Another way to state the null hypothesis would be to say that the treatment (seeded versus unseeded clouds) is unrelated to the amount of rainfall recorded, or that the labels of “seeded” or “unseeded” would be arbitrary, or interchangeable.

We can record the observed magnitude of difference between seeded and unseeded clouds by finding the difference in median rainfalls for seeded versus unseeded clouds.

In our previous example, we conducted a hypothesis test directly based on our observed data. But in this example, we first have to make a choice about what sort of summary measurement we will base our hypothesis test on. In this case, we are basing it on the difference in median rainfalls for the two groups. When we base a hypothesis test on a summary measurement, we call that value a **test statistic**. Most, though not all, hypothesis tests involve having to choose and calculate an appropriate test statistic.

To think of how unusual this is, we can randomly simulate what sorts of differences we could expect to see if seeding status was unrelated to rainfall. We can randomly shuffle the labels of “seeded” and “unseeded” in our data, and see what sorts of differences in median rainfall we could expect to see if those labels are arbitrary. We can examine the distribution of these simulated differences. We refer to this distribution as the **null distribution**, since it shows us a distribution of the sorts of values we could expect to see if the null hypothesis were true.

If our observed difference is in line with what we could expect if the labels were arbitrary, then there’s no evidence that seeding is related to rainfall amount.

If our observed difference would be unusual if the labels were arbitrary, that suggests that seeding is related to rainfall amount.

```
#calculate and store the observed difference in the sample
observed <- median(clouds$Rainfall[clouds$Treatment=="Seeded"]) - median(clouds$Rainfall[clouds$Treatment=
="Unseeded"])

#N = number of simulations we will use
N <- 10^4-1

#sample.size = the number of observations in our sample
sample.size = nrow(clouds)

#group.1.size = the number of observations in the first group
group.1.size = nrow(clouds[clouds$Treatment=="Seeded"])

#create a blank vector to store the simulation results
result <- numeric(N)

#use a for loop to cycle through values of i ranging from 1 to N
for(i in 1:N)
{
  #each iteration, randomly sample index values
  #sample.size gives the total number of index values to sample from
  #group.1.size gives the number of index values to sample
  #sample without replacement
  #indexes sampled will be treated as the "seeded" group, indexes not sample as "unseeded"
  index = sample(sample.size, size=group.1.size, replace = FALSE)

  #calculate and store the difference in
  #median rainfall between the index and non-index groups
  result[i] = median(clouds$Rainfall[index]) - median(clouds$Rainfall[-index]))
}

#plot a histogram of the simulated differences
#add a vertical line at the observed difference
ggplot(data=tibble(result), mapping = aes(x=result)) +
  geom_histogram(breaks=seq(-300,300,by=25)) +
  geom_vline(xintercept = observed, color = "red")

#Calculate the p-value
(sum(result >= observed) + 1) / (N + 1)
```

We find our p-value by taking the total number of simulations with a result as extreme as our observed difference, divided by the total number of simulated outcomes (we add a 1 to both the numerator and denominator to include out actual observation alongside the simulations).

In this case, our p-value came out quite small. This suggests that our observed difference is very unusual relative to the sorts of differences we could expect to see just via random chance. This would indicate that the cloud seeding does have an effect. We would reject our null hypothesis in favor of the alternative hypothesis, and conclude that there is evidence that median rainfall is higher for seeded than unseeded clouds.

Permutation vs. Randomization

The terms **permutation test** and **randomization test** are often used interchangeably. There is, strictly speaking, a small difference between them. In our example above, what we conducted was a randomization test - we randomly simulated a number of possible outcomes to establish a null distribution. In a permutation test, instead, every

possible permutation of labels is considered, and the null distribution is based on this exhaustive list of possible permutations.

It is most often the case that it would be computationally prohibitive to calculate results for every possible permutation, and so randomization tests are more commonly used as approximations to a full permutation test.

Choice of Test Statistic

In the cloud seeding example, we used a difference in medians as our test statistic, because medians seemed like the appropriate numerical summaries to use given how skewed the data was. In many situations, where data is not so skewed, it is more common to see test statistics based on means. Other quantities could be considered as well. For example, if the goal was not to compare how values are centered in two groups, but instead how the variability in two groups compare, then a test statistic could be calculated based on standard deviations.

Assumptions

Unlike many traditional hypothesis testing methods (which you may have encountered before, and which we will discuss in later sections), a permutation test does not require any assumptions about the underlying distribution of the data. They also do not require an assumption that the data come from a random sample. This means that permutation tests can be used in far more situations than more traditional hypothesis testing methods, which is a large part of why they are becoming an increasingly common way to conduct tests.

The only major assumption necessary for a permutation test is that, under the null hypothesis, the data in both groups come from the same distribution. That is, if the null hypothesis were true, it is not enough that the two groups just have the same center - the shapes and spreads of their distributions would need to be the same as well.

When we consider assumptions, we often talk about **robustness**. Robustness refers to how much our results can be impacted if an assumption is not met. If an assumption not being met does not have a large impact on results, we say that a particular testing method is robust to that assumption.

In the case of permutation tests, the tests are generally robust to the assumption of equal distributions under the null hypothesis. If the centers are the same under the null, but the shape and/or spread are not, it will generally not have a large impact on our results. The one situation where the permutation test is not robust to this assumption is if there are substantial differences in spread or variability between the two groups, as well as substantial differences in sample sizes. When this occurs, the null distribution will end up mostly reflecting the spread of whichever group had the larger sample size. In this case, some approach other than a randomization test may be more appropriate.

Matched Pairs

In the cloud seeding example, we were considering data from a study in which there were two independent groups being analyzed, a seeded group and an unseeded group.

Another common study design is a matched pairs design. In this type of study, rather than collecting data from two independent groups, the study is designed so that each observation in one condition has a corresponding, or paired, observation in the other condition.

This might be achieved by pairing subjects who are most similar to one another based on some traits of interest. In some medical studies, identical twins are recruited. Or, we might look at the same subject twice - perhaps a study looking to see if a topical ointment causes rashes could make use of both the left and right forearms of the same individual, randomly selecting for each individual which arm receives the real ointment and which arm receives a placebo. Or a subject might receive one treatment at one time, and another treatment at another time, so that they could be compared.

At the end of Chapter 3, our textbook shows how to extend the idea of permutation tests to a matched pairs design. Look over this section on your own, and if you have any questions about it, check in with me.