Notes 12 - Hypothesis Testing and Null Distributions

STS 2300 Introduction to Data Analytics

Updated: 2025-04-21

Table of Contents

[Reading for Notes 12 1](#_Toc196158744)

[Learning Goals for Notes 12 1](#_Toc196158745)

[Background example 2](#_Toc196158746)

[Hypothesis testing overview 3](#_Toc196158747)

[Step 1 - Our Two Hypotheses 3](#_Toc196158748)

[Step 2 - Gathering and summarizing data 4](#_Toc196158749)

[Step 3 - Comparing our data to 6](#_Toc196158750)

[Step 4 - Reaching conclusions 7](#_Toc196158751)

[Using the infer package 9](#_Toc196158752)

[Revisiting the Learning Goals for Notes 12 12](#_Toc196158753)

# Reading for Notes 12

Read [Chapter 9 (Sections 9.1 and 9.2) of the Modern Dive textbook](https://moderndive.com/9-hypothesis-testing.html) to supplement Notes 12.

# Learning Goals for Notes 12

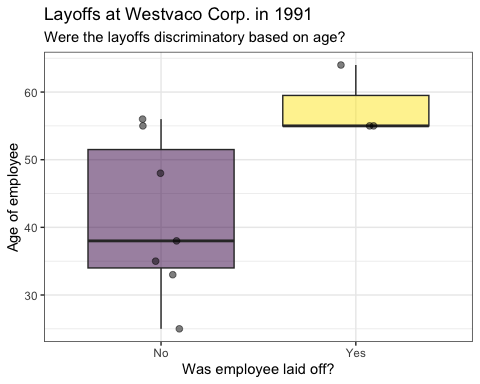
* Be able to describe the general process of hypothesis testing and how it is used to answer scientific questions
* Understand the concept of a null distribution and how it is used in hypothesis testing
* Understand the general steps of using the infer package for creating null distributions for hypothesis testing.

I’ll be using the following packages in this set of notes, so I’ll load them before I get started.

library(infer)  
library(ggplot2)

# Background example

Let’s consider an example similar to one in the textbook. This is based on a real life example that happened in 1991. A man named Robert Martin was laid off from his company (Westvaco Corporation). He believed that the company was discriminating based on age in the layoffs. He filed a lawsuit against the company, and one piece of the lawsuit was the data below. In the round of layoffs that Martin was part of, there were 10 employees eligible to be laid off. Three of the employees were laid off (ages 55, 55, and 64), and seven were not laid off (ages 25, 33, 35, 38, 48, 55, and 56). Below is a graph comparing the ages of the two groups. The company claimed that these layoffs were done randomly.



At this point, we have two competing explanations for the layoffs.

1. (Westvaco) The layoffs were done randomly
2. (Robert Martin) Age (or a related characteristic) was used in the layoffs

In a court of law, we assume someone is “innocent until proven guilty”, so let’s give Westvaco the benefit of the doubt and assume that they’re telling the truth. Then we will gather some data/evidence (the ages of those laid off) and see if it changes our mind. In other words, hypothesis testing is a lot like a trial in a court room.

The **basic steps of a hypothesis test** are:

1. Come up with our two competing hypotheses (one of which we will assume to be true for now)
2. Gather data and summarize it
3. Consider how likely or unlikely our data would be in a world where our assumed hypothesis is actually true
4. Reach a conclusion based on this information

In this case, our main question to answer might be: “how often will the difference in average ages be this large if layoffs are truly random”? One way to investigate this would be to “shuffle” all of the ages up and randomly choose three people to “lay off”. If we did this many, many times, we would start to see how often what actually happened would occur without age discrimination.

# Hypothesis testing overview

Hypothesis tests, like confidence intervals, are used when we want to learn more about a population or a process. Let’s review some important terminology to help us better understand the four steps mentioned above.

## Step 1 - Our Two Hypotheses

We will come up with two hypotheses that describe possible states of the world. In other words, these hypotheses should be phrased in terms of a population (or a process that created the data), not in terms of our data. This means the hypotheses can often be written in terms of parameters like or .

* The **null hypothesis** (written as ) is a hypothesis based on the idea that “nothing unusual is going on”. This can often be thought of as the status quo, and it is the hypothesis that we will assume is true for the sake of our exploration. In the example above, if “nothing is going on” then the company really did lay people off randomly. The court will assume this is true until they have been convinced otherwise.
* The **alternative hypothesis** (written as or ) is a hypothesis based on our question of interest. This is usually the question that leads to us conducting a hypothesis test in the first place. In the example above, there wouldn’t be a need for a trial if Robert Martin didn’t think age discrimination took place. Notice that the alternative hypothesis directly contradicts our null hypothesis.

We could write our two hypotheses like this:

: The company laid people off at random.

: Older employees were discriminated against in layoffs (i.e., company did not lay people off at random).

These hypotheses are written in terms of the process that was used, *not* in terms of our data.

## Step 2 - Gathering and summarizing data

To test our hypotheses, we gather evidence and then summarize it. We will summarize our evidence by calculating some sort of statistic that takes all of our data and condenses it into one number. This could be something like a *sample mean*, *sample proportion*, *difference in sample means*, or *difference in sample proportions*.

In the age discrimination example, we might be interested in comparing the ages of those laid off versus those not laid off. Specifically, we could look at how far apart our two sample means are. Let’s do that for the sample mean age of those laid off minus the sample mean age of those not laid off.

layoffs <- data.frame(age = c(25, 33, 35, 38, 48, 55, 55, 55, 56, 64),  
 laidoff = c("No", "No", "No", "No", "No",   
 "Yes", "Yes", "No", "No", "Yes"))  
obs\_xbardiff <- layoffs |>   
 specify(formula = age ~ laidoff) |>   
 calculate(stat = "diff in means", order = c("Yes", "No"))  
obs\_xbardiff

## Response: age (numeric)  
## Explanatory: laidoff (factor)  
## # A tibble: 1 × 1  
## stat  
## <dbl>  
## 1 16.6

Those who were laid off were an average of 16.6 years older than those not laid off. Mathematically, that would look like: .

Most theory-based tests will use a statistic like this to calculate a **test statistic**. Test statistics summarizes our data *in relation to a null hypothesis*. For example, in a case where we are interested in a difference in population means, we might use a test statistic of the following form:

To calculate this number, I need to know . While I won’t *actually* know or , I can use my null hypothesis (which I’m *assuming* is true until convinced otherwise) to determine that the difference in those values should be 0 (because the process is assumed fair on average).

Everything else in the formula is based on my sample, so I could calculate by plugging in values from my data. We can also use the infer package to calculate this. Similar to how we calculated the difference in means above, we can use the stat = "\_\_" argument in the calculate() function to calculate a t statistic.

obs\_t <- layoffs |>   
 specify(formula = age ~ laidoff) |>   
 calculate(stat = "t", order = c("Yes", "No"))  
obs\_t

## Response: age (numeric)  
## Explanatory: laidoff (factor)  
## # A tibble: 1 × 1  
## stat  
## <dbl>  
## 1 3.09

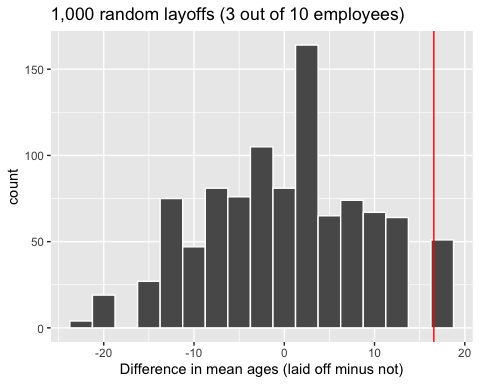
For this data, my test statistic is . However, neither the difference in means above or my test statistic here means much without knowing what sort of values I should expect to get *in a world where my null hypothesis really is true*.

## Step 3 - Comparing our data to

To add context to the statistic(s) we calculated, we need to consider what type of statistic or test statistic I would expect to see if the null hypothesis were true. Then I can compare what I actually saw (my observed statistic) to what I “expected”.

To do this, I need something called a null distribution. A **null distribution** is essentially a sampling distribution that comes from a world where the null hypothesis is true. In other words, it is a distribution of values that I would get for my sample statistic if I repeatedly took samples from a population where the null hypothesis were true. We’ll see how to generate these later, but here’s an example for the differences in mean ages that I’d expect if the layoffs were truly random.

Below are the results of randomly laying off 3 out of 10 employees 1,000 times. I’ve plotted the difference in sample mean ages for those laid off minus those not laid off for each of the 1,000 simulated layoffs. The result that actually occurred (often called our observed result) is in red.



Notice that what actually occurred is a fairly rare result if layoffs are random. We can quantify just how rare or unusual these layoffs with a p-value.

A **p-value** is the probability of observing results as or more extreme than the results that actually occurred *if the null hypothesis was true*. What counts as “extreme” are things that support our alternative hypothesis even more than what we observed in the data.

Another way to think of this: In a world where is true, how often would something this extreme happen?

Even if you’ve used p-values before, it can be a tricky concept, so let’s apply it to our example. The p-value would be the probability of seeing a difference in sample means of 16.6 or bigger if the layoffs were truly random. We are considering differences in sample means *bigger* than what we observed because the alternative hypothesis said there was age discrimination, which would result it those laid off having big sample means compared to those not laid off.

**Note**: One of the most common misconceptions in statistics is to think that the p-value is the probability our null hypothesis is true. This is **not** the case. It is the probability of data as extreme as what we observed *IF* the null hypothesis were true. You can also think of the p-value as a measure of how compatible your data is with the null hypothesis. We will revisit this idea in step 4.

Let’s estimate the p-value for the age discrimination example. Just eyeballing the graph, it looks like about 50 of the 1,000 simulated layoffs resulted in an average age difference as big as what really occurred. This means the probability of a difference at least this extreme is around 0.05.

## Step 4 - Reaching conclusions

Once we have our p-value, we use that to reach a conclusion related to our two original hypotheses. To do this, we need a significance level (Note: This should typically be specified before you gather data).

The **significance level** (written as , the Greek letter alpha) is a cutoff for how strong of evidence will be required for us to abandon our assumption that the null hypothesis is true. The most common value used is , but what we choose will depend on the situation. and are common alternatives. The significance level is often based on how much of a problem it will be if we’re wrong when we abandon the null hypothesis. For example, [particle physicists](https://blogs.scientificamerican.com/observations/five-sigmawhats-that/) typically uses because they don’t want to reject the status quo and publish a new discovery unless they are VERY confident it is correct. At the beginning of the semester, we did the Lady Tasting Tea activity. In that example, it’s not a very big deal if we’re wrong, so we might have used a larger p-value like .

Once we have a p-value and a significance level, there are two possible conclusions we will make:

1. If our -> We reject the null hypothesis. In this case, there is sufficient evidence to convince us that the alternative hypothesis is true. You can think of this as our p-value telling us that our data is not very compatible with a world where the null hypothesis is true.
2. If our -> We fail to reject the null hypothesis. In this case, there is insufficient evidence to convince us the alternative hypothesis is true. You can think of this as our p-value telling us that our data *is* compatible with a world where the null hypothesis is true.

Notice that in the first case I’m essentially saying “I have seen enough evidence to believe my null hypothesis is wrong and my alternative hypothesis is correct”.

In the second case I’m saying “I didn’t see enough evidence to abandon my null hypothesis or to believe my alternative hypothesis is true.” This second option does **NOT** mean that I’ve shown the null hypothesis is correct. It really means that I’m not sure which hypothesis is correct. There’s a good chance my data is compatible with ***both*** of my hypotheses in these situations.

How does this all apply to the age discrimination example? Let’s say the court decides to use because they don’t want to falsely convict the company of age discrimination if the company didn’t discriminate based on age.

**Step 1:** Our two hypotheses

: The company laid people off at random.

: Older employees were discriminated against in layoffs (i.e., company did not lay people off at random).

**Step 2:** Summarizing our data

Those laid off were 16.6 years older on average than those not laid off

**Step 3:** Comparing our data to

p-value

In a world where the layoffs were random ( is true), we would see a result favoring age discrimination at least this strongly around 5% of the time.

**Step 4:** Reaching a conclusion

Our p-value of 0.05 was greater than our significance level of 0.01. We fail to reject the null hypothesis. There is insufficient evidence to claim that the layoffs were due to age discrimination.

**Final note on this example:** This would mean the jury would find Westvaco “not guilty” because the evidence was consistent with a world where they laid people off at random. This does **not** mean the company laid people off at random. It just means that this data alone (the ages of the 10 people eligible to be laid off) isn’t enough to convince us of age discrimination. If Robert Martin had additional evidence, that might be enough to then convince us.

# Using the infer package

Just like when we made confidence intervals, we will use the infer package to conduct our hypothesis tests. Most of this process will be very similar to what we’ve already seen. The main differences will be that we will add a step for a hypothesize() function that lets us say what our hypotheses are. We will also generate our null distribution a bit differently than we generated a bootstrap distribution because now we’re basing things on a null hypothesis, not just our data.

1. specify()

This step will be identical to our confidence interval examples. We need to specify what variables we’re studying. In the age discrimination example, we’re comparing the ages among the groups that were laid off versus not laid off.

layoff\_perm <- layoffs |>   
 specify(formula = age ~ laidoff)

1. hypothesize()

In the hypothesize() step we need to specify what our null hypothesis is. We’ll focus on this more when we look at examples in upcoming sections of the notes. For now, I’m specifying that I think age and whether someone is laid off are independent of one another. In other words, layoffs are completely random and age plays no role. This was what our null hypothesis said above.

layoff\_perm <- layoffs |>   
 specify(formula = age ~ laidoff) |>   
 hypothesize(null = "independence")

1. generate()

The generate() step will be similar, but we don’t want a bootstrap distribution. With a boostrap distribution, each person’s age and whether they were laid off is linked together, and we sampled entire people at random. However, if age and whether someone is laid off aren’t related, then a 65 year old who was laid off could just as easily have not been laid off. With type = "permute", we shuffle the ages among our groups of those laid off and not (like would happen if layoffs were random). This step will generate 1,000 samples from a world where layoffs were completely random for these 10 people.

layoff\_perm <- layoffs |>   
 specify(formula = age ~ laidoff) |>   
 hypothesize(null = "independence") |>   
 generate(reps = 1000, type = "permute")

1. calculate()

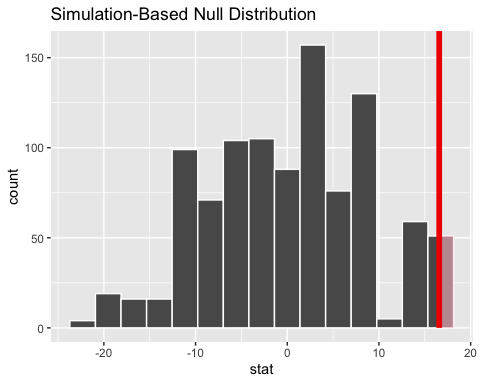
Now that we have 1,000 samples from our hypothetical world where the null hypothesis is true, we need to calculate a statistic for each sample to make our null distribution. Again, this step will be identical to what we saw with our confidence interval approach.

set.seed(201005)  
  
layoff\_perm <- layoffs |>   
 specify(formula = age ~ laidoff) |>   
 hypothesize(null = "independence") |>   
 generate(reps = 1000, type = "permute") |>   
 calculate(stat = "diff in means", order = c("Yes", "No"))

1. visualize()

We can use the visualize() function to see our null distribution and the shade\_pvalue() function to include our p-value on the graph. The two arguments we need in shade\_pvalue() are obs\_stat = \_\_ and direction = \_\_. Remember that a p-value is the probability of seeing something as or more extreme than what I saw if the null hypothesis were true. My observed statistic is “what I saw”, so we use the obs\_stat argument to mark that value on the graph. In this example, age discrimination () would show up by laying off older employees, so anything “more extreme” than what occurred would be values showing even higher differences in sample means. We can specify this by direction = "right" to shade everything to the right of what we observed.

visualize(layoff\_perm) +  
 shade\_pvalue(obs\_stat = obs\_xbardiff,   
 direction = "right")



To calculate my p-value, I use similar syntax to the shade\_p\_value() function.

layoff\_perm |>   
 get\_p\_value(obs\_stat = obs\_xbardiff,   
 direction = "right")

## # A tibble: 1 × 1  
## p\_value  
## <dbl>  
## 1 0.051

Our p-value of 0.051 matches what we saw on the graph. If layoffs were truly random, the probability is only around 0.051 of ending up with a gap in sample mean ages at least this big between those laid off and not laid off.

# Revisiting the Learning Goals for Notes 12

* Be able to describe the general process of hypothesis testing and how it is used to answer scientific questions
* Understand the concept of a null distribution and how it is used in hypothesis testing
* Understand the general steps of using the infer package for creating null distributions for hypothesis testing.