Hypothesis testing (NHST)

What Is A Hypothesis?

In statistics, a **hypothesis** is a statement about the population. It is usually a prediction that a parameter describing some characteristic of a variable takes a particular numerical value, or falls into a certain range of values.

Ex: given equal qualifications, male candidates are viewed more favorably and more likely to be hired than female candidates (Moss-Rascusin et al., PNAS, 2012). **Research hypothesis.**

Numerically, a **statistical hypothesis**:

$$\text{Proportion}_{\text{Male applicants hired}} > \text{Proportion}_{\text{Females applicants hired}}$$

The null hypothesis

In Null Hypothesis Significance Testing, we... test a null hypothesis.

A **null hypothesis** (H_0) is a statement of no effect. The *research hypothesis* states that there is no relationship between X and Y, or our intervention has no effect on the outcome.

- The *statistical hypothesis* is either that the population parameter is a single value, like 0, or that a range, like 0 or smaller.

The alternative hypothesis

According to probability theory, our sample space must cover all possible elementary events. Therefore, we create an **alternative hypothesis** (H_1) that is every possible event ***not*** represented by our null hypothesis.

$$H_0 : \mu = 4 \qquad H_0 : \mu < -4$$

$$H_1 : \mu \neq 4 \qquad H_1 : \mu \geq -4$$

The tortured logic of NHST

We create two hypotheses, H_0 and H_1. Usually, we care about H_1, not H_0.

What we want to know:

$$P(H_1|D)$$

Instead, we're going to assume our null hypothesis is true, and test how likely we would be to get these data.

$$P(D|H_0)$$

Example #1

Consider the example of possible gender discrimination in hiring a research technician at University X.

Let Π (pi) be the probability any particular selection is male.

University X claims that $\Pi = .5$. This can be the null hypothesis for two reasons:

1. If this is true, than an equal number of men and women would be hired, and there would be no bias.
2. Regardless of whether this matches the population, this is the University's claim and we want to test it.

Example #1

As a critical and suspicious graduate student who is well informed about unconscious bias in STEM, you're skeptical that there is no gender bias in hiring, and you have an alternative hypothesis that the probability of men being hired is different than .5.

$$H_0 : \Pi = .5$$

$$H_1 : \Pi \neq .5$$

Example #1

To test the null hypothesis, you look at the hiring practices of University X over the last year, and find that they had 10 job openings, for which they hired 9 men.

The question you're going to ask is:

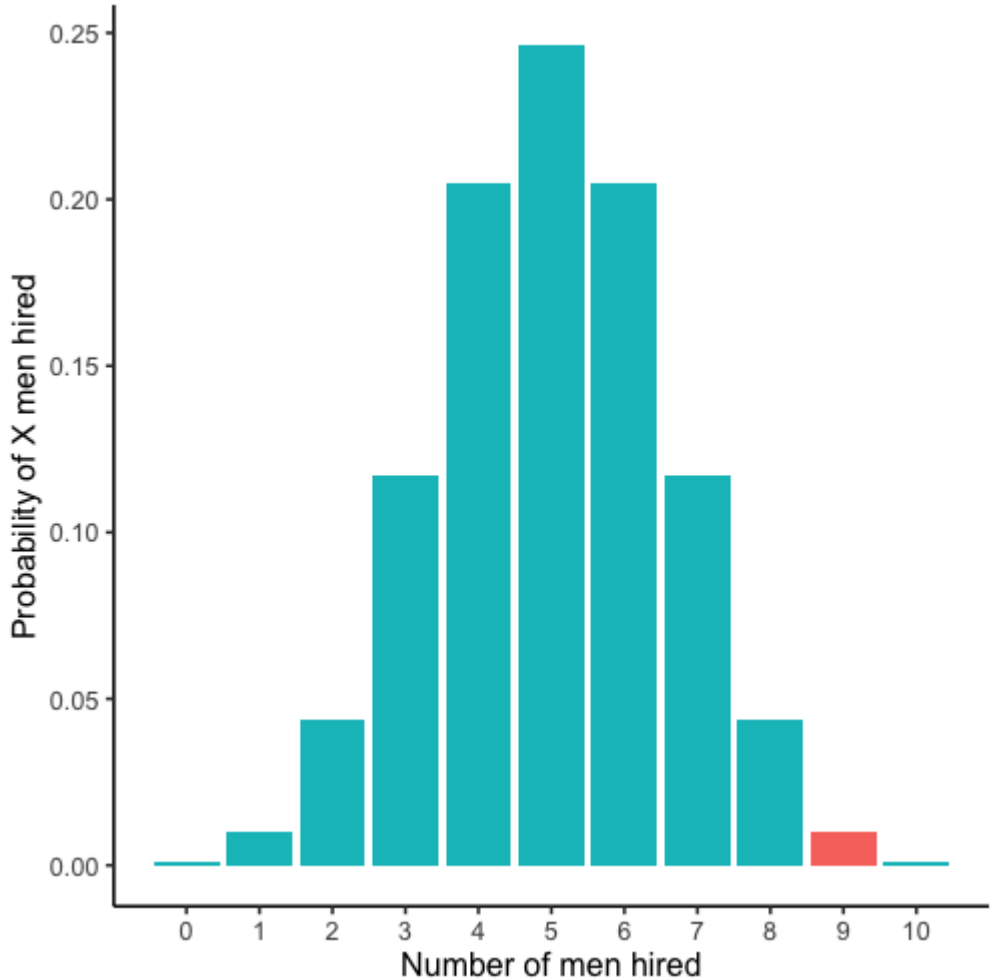
- "How likely is it that 9 out of 10 hires were men, if the probability of hiring a man is .5?"

This is the essence of NHST.

You can already test this using what you know about the binomial distribution.

```r
dbinom(9, size = 10, prob = .5)
```

```
## [1] 0.009765625
```



Complications with the binomial

The likelihood of hiring a man 9 times out of 10 *if the true probability of hiring a man is .5* is 0.01. That's pretty low! That's so low that we might begin to suspect that the true probability is not .5.

But there's a problem with this example. Sometimes University X has 10 job openings, but sometimes it has 1 and sometimes it has 30, and on and on. The binomial doesn't really apply to University X's hiring practices because N can change (and it's an assumption of the binomial that we know N).

What we really want is not to assess 9 out of 10 times, but a proportion, like .9. How many different proportions could result year to year?

Our statistic is usually continuous

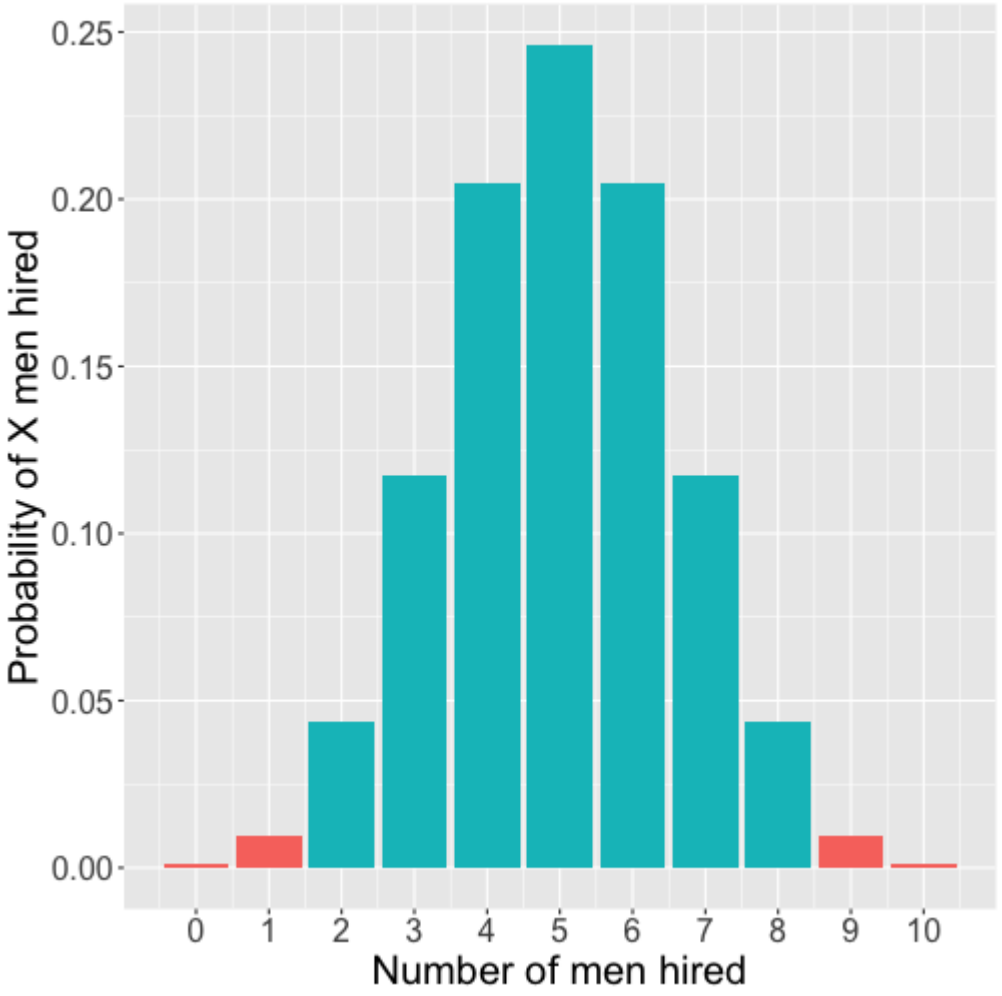
When we estimate a statistic for our sample -- like the proportion of males hired, or the average IQ score, or the relationship between age in months and second attending to a new object -- that statistic is nearly always **continuous**.

So we have to assess the probability of that statistic using a probability distribution for continuous variables, like the normal distribution (or t, etc.).

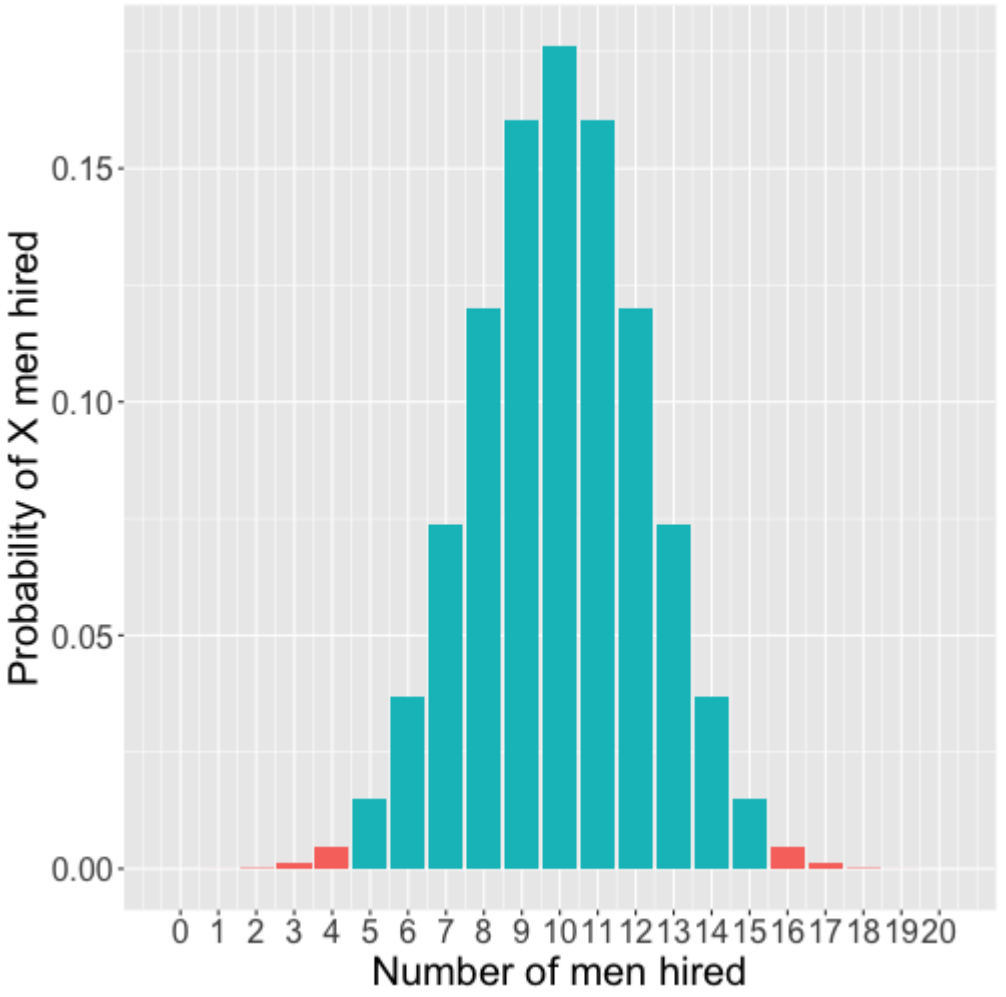
What is the probability of any value in a continuous distribution?

Instead of calculating the probability of our statistic, we calculate the probability of our statistic *or more extreme* under the null.

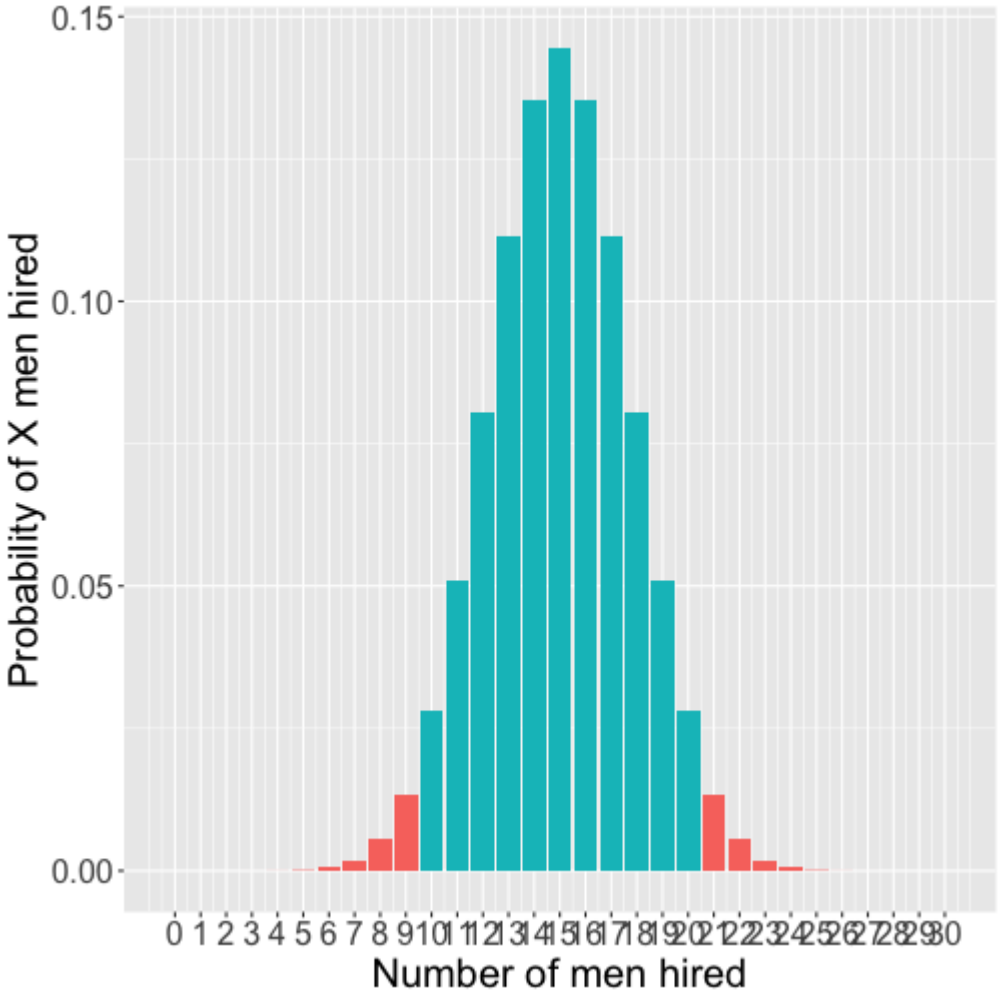
The probability of hiring 9 men out of 10 or more extreme is 0.02.



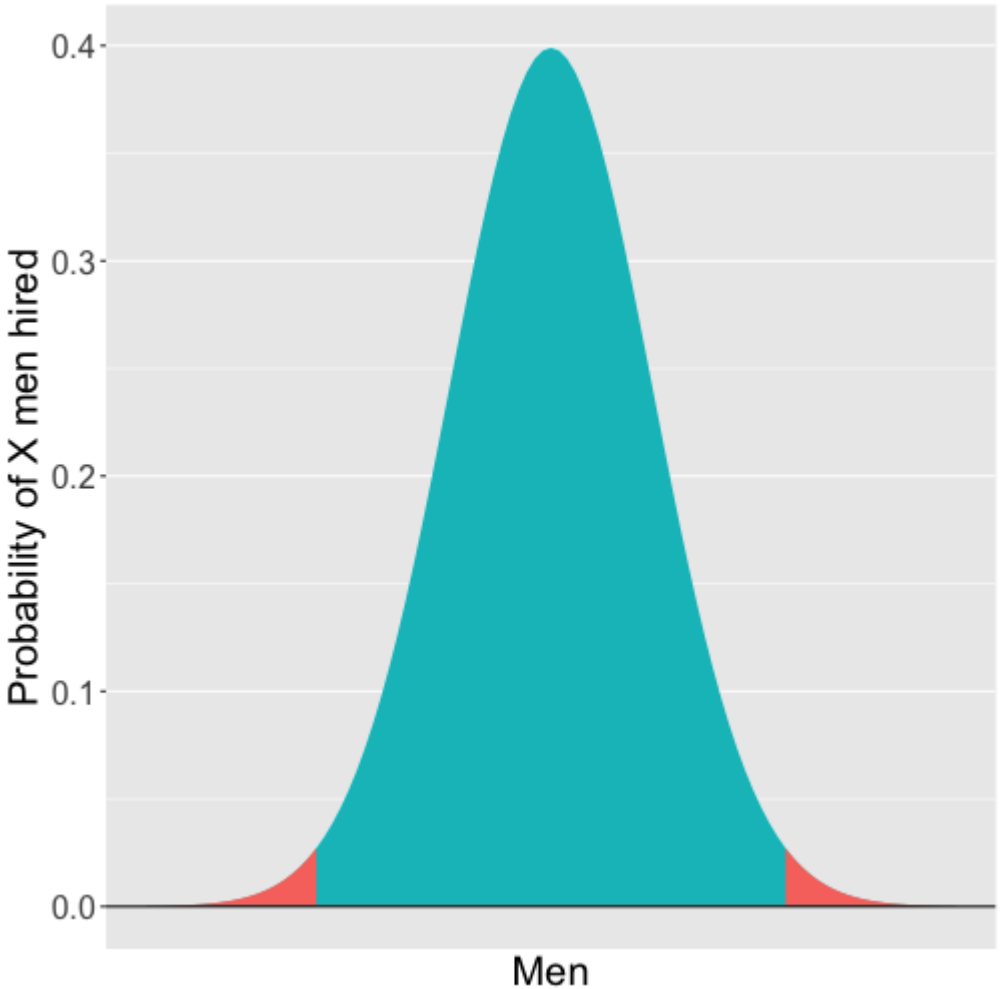
As we have more trials...



... and more trials...



If our measure was continuous, it would look something like this.



Quick recap

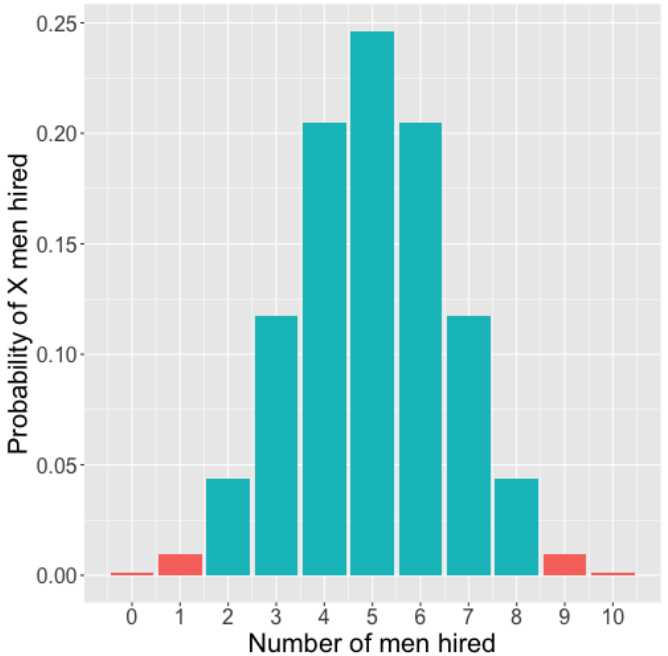
For any NHST test, we:

1. Identify the null hypothesis (H_0), which is usually the opposite of what we think to be true.
2. Collect data.
3. Determine how likely we are to get these data or more extreme if the null is true.

What's missing?

1. How do we determine what the distribution looks like if the null hypothesis is true?
2. How unlikely do the data have to be to "reject" the null?

Enter sampling distributions



When we were analyzing the gender problem, we built the distribution under the null using the binomial.

This is our **sampling distribution.**

What do we need to know to build a sampling distribution based on the binomial?

But as we said before, we're not really going to use the binomial much to make inferences about statistics, because the vast majority of our statistics are continuous, not discrete. Instead, we'll use other distributions (like the normal or t etc.) to create our sampling distributions.

For now, we'll work through an example using the standard normal distribution.

Example #2

University X has been around for 150 years, and so has 150 years worth of ratings of applicants. You pay an undergrad dig through all the old university files and calculate the average rating of productivity for the job applicants (5.3 out of 10) and also the standard deviation of those ratings (3.3).

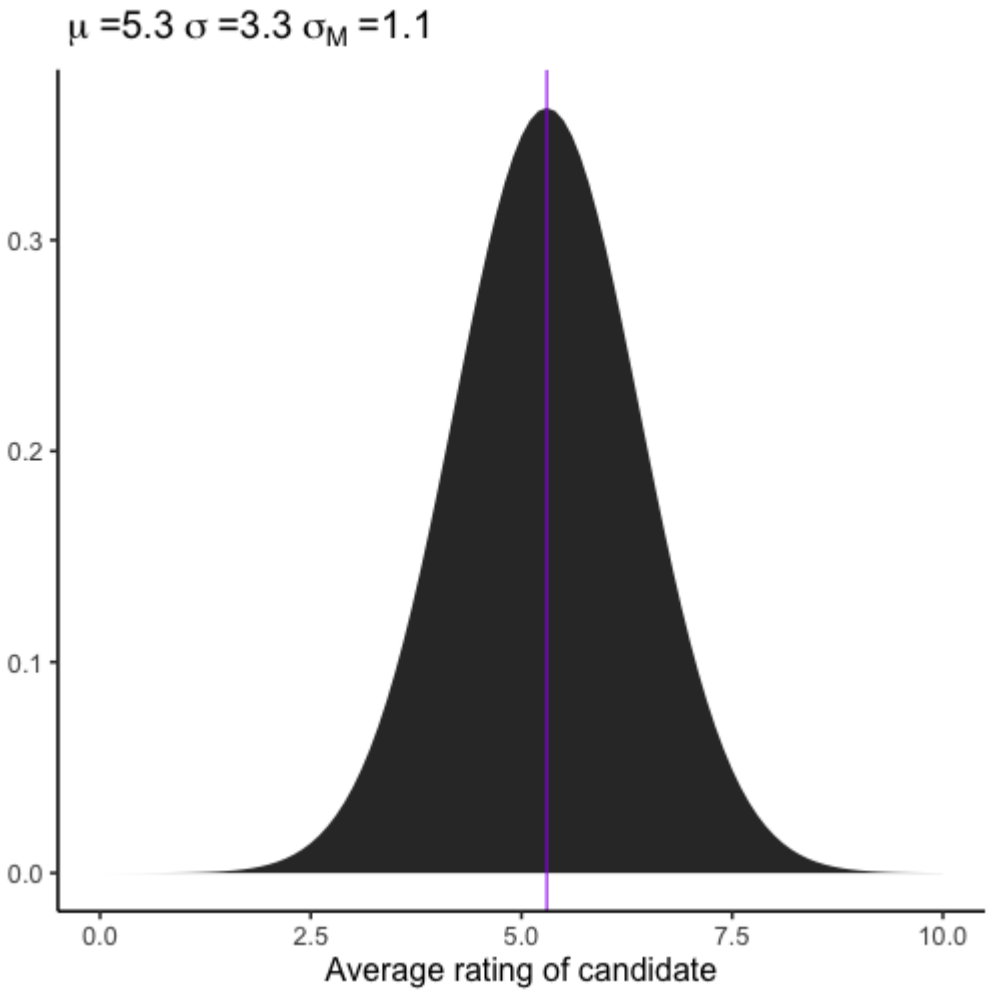
You are interested in if the pandemic contributed to productivity declines. You then collect the ratings of 9 very recent applicants from 2020, and you calculate their average rating (2.9).

How do you generate the sampling distribution around the null?

The mean of the sampling distribution = the mean of the null hypothesis

The standard deviation of the sampling distribution:

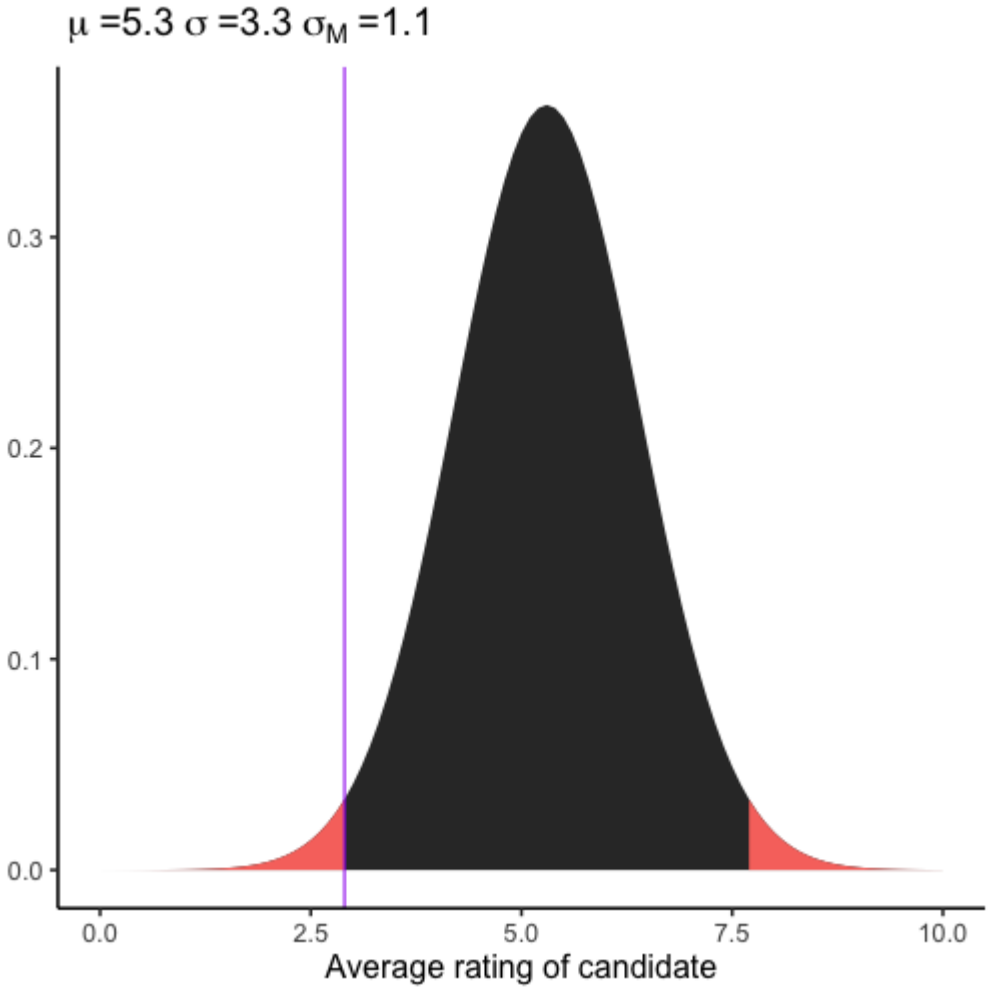
(Assume random sampling in order to use SEM)

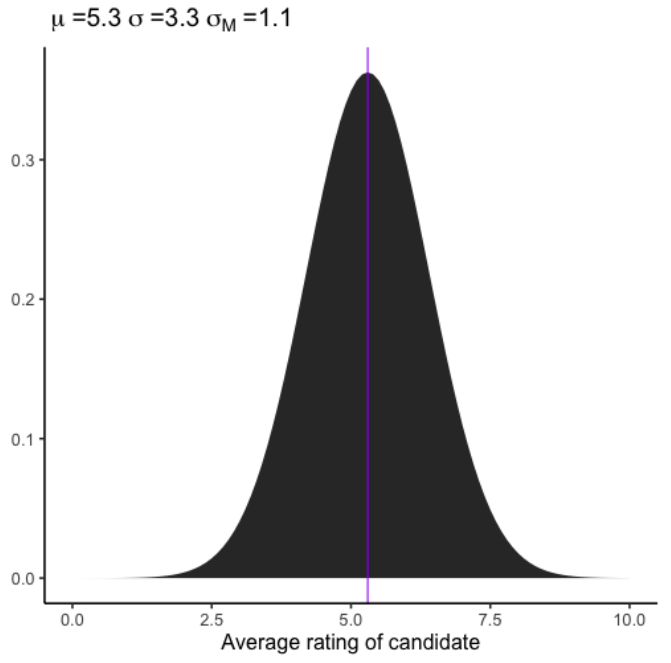


The mean of the sampling distribution = the mean of the null hypothesis

The standard deviation of the sampling distribution:

$$SEM = \frac{\sigma}{\sqrt{N}}$$





We have a normal distribution for which we have a score of interest (new applicant, $\bar{X}$), we know the mean of the population (the before times), the standard deviation of the before times, so we can get the standard deviation (SEM).

$$Z = \frac{\bar{X} - M}{SEM} = \frac{2.9 - 5.3}{1.1}$$

$$-2.18$$

We can use this information to calculate a Z-score; in the context of comparing one mean to a sampling distribution of means, we call this a **Z-statistic**.

$$Z = \frac{\bar{X} - M}{SEM} = \frac{2.9 - 5.3}{1.1} = -2.18$$

Now we use the properties of the Standard Normal Distribution to calculate probabilities, specifically the probability of getting a score this far away from μ or more extreme:

```
pnorm(-2.18) + pnorm(2.18, lower.tail = F)
```

```
## [1] 0.02925746
```

```
pnorm(-2.18)*2
```

```
## [1] 0.02925746
```

The probability that the average current applicant's score would be at least 2.18 units away from the average score of the population of applicants is 0.029.

0.029 is our p-value!

Congrats! 🎉

You just ran your first z-test.

It's useless.

We use it when we want to know if our mean is the same or different from a population mean.

The probability that the average post-pandemic applicant's score would be at least 2.18 units away from the average score is 0.029.

0.029 is our p-value.

A p-value DOES NOT:

- Tell you that the probability that the null hypothesis is true.
- Prove that the alternative hypothesis is true.
- Tell you anything about the size or magnitude of any observed difference in your data.

Again, but with a twist...

All well and good.

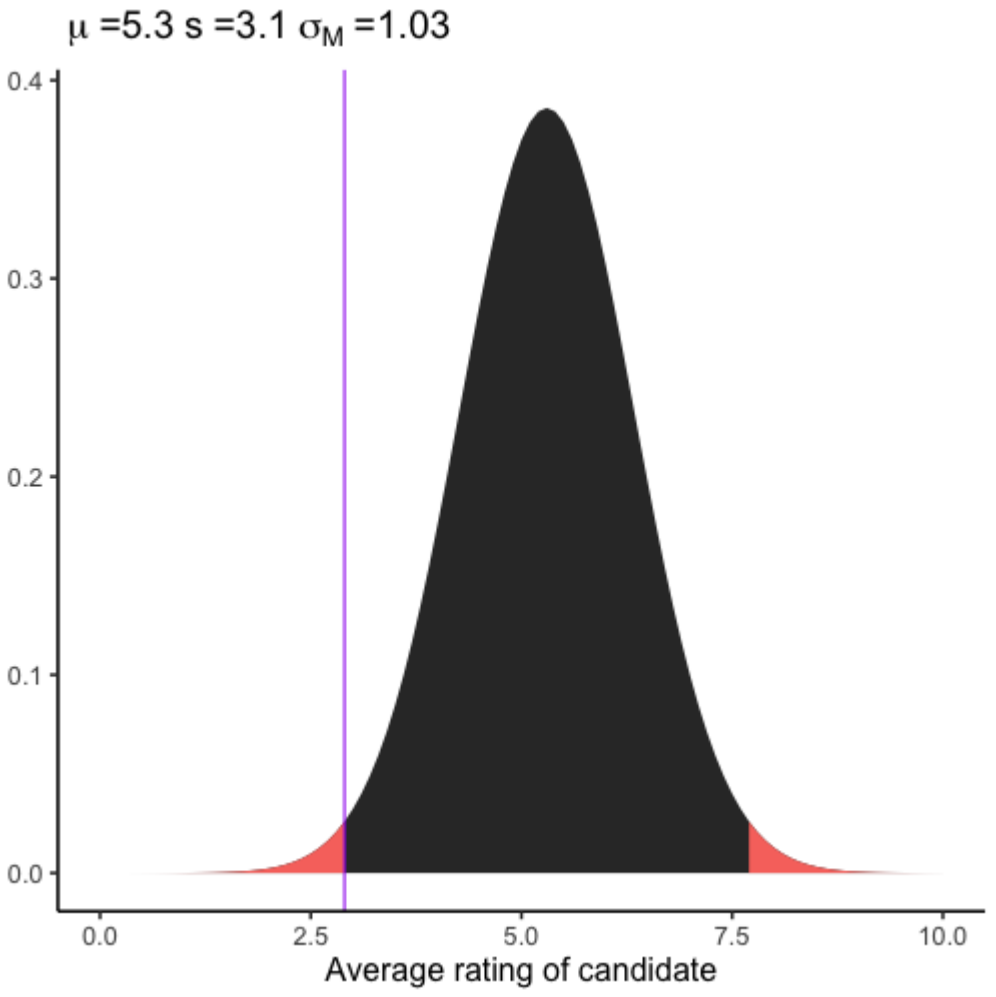
But rarely will you have access to all the data in your population, so you won't be able to calculate the population standard deviation. What will you do?

We still have our population mean of 5.3, but now let's say we *only* have the **sample** standard deviation...

$$SEM = \frac{\hat{\sigma}}{\sqrt{N}} = \frac{s}{\sqrt{N}}$$

If you didn't know the population standard deviation, you would use the sample of post-pandemic applicants to estimate the population standard deviation.

$$SEM = \frac{\hat{\sigma}}{\sqrt{N}}$$



Again, but with a twist...

$$t = \frac{\bar{X} - M}{SEM} = \frac{2.9 - 5.3}{1.03} = -2.32$$

```
pnorm(-2.32)*2
```

```
## [1] 0.02034088
```

```
pt(-2.32, df = N-1)*2
```

```
## [1] 0.04891943
```

Congrats! 🎉

You just ran your first t-test.

It's extremely useful

We use it when we want to know if our mean is the same or different from a population mean, but we don't have the population variance. Most of the time (not always, but most), we use this in the context of "is some value different from 0".

Note, this is specifically called a "one sample t-test" but we'll talk more about that in a bit

$.029, .020, .049$

Is that a really low probability?

Before you test your hypotheses -- ideally, even before you collect the data -- you have to determine how low is too low.

Researchers set an alpha (α) level that is the probability at which you declare your result to be "statistically significant." How do we determine this?

Consider what the p-value means. In a world where the null (H_0) is true, then by chance, we'll get statistics in the extreme. Specifically, we'll get them α proportion of the time. So α is our tolerance for False Positives or incorrectly rejecting the null.

Historically, psychologists have chosen to set their α level at .05, meaning any p-value less than .05 is considered "statistically significant" or the null is rejected.

This means that, among the times we examine a relationship that is truly null, we will reject the null 1 in 20 times.

Some have argued that this is not conservative enough and we should use $\alpha < .005$ (Benjamin et al., 2018).

1. Define H_0 and H_1.

2. Choose your α level.

3. Collect data.

4. Define your sampling distribution using your null hypothesis and either the knowns about the population or estimates of the population from your sample.

5. Calculate the probability of your data or more extreme under the null. (To get the probability, you'll need to calculate some kind of standardized score, like a z-statistic.)

6. Compare your probability (p-value) to your α level and decide whether your data are "statistically significant" (reject the null) or not (fail to reject the null).

Next time...

more NHST