STA130H1F

Class #4

Prof. Nathalie Moon

2018-01-10

Today's Class

Answering the question:
is something we observe in data meaningful, or could it simply be due to chance?

Examples for:

- a single proportion

Next week:

- extend to more situations

Recommended reading:
Sections 2.3.1, 2.3.2, 2.3.7 and 2.4 of *Introductory Statistics with Randomization and Simulation* from OpenIntro
(a free open-source textbook)

Statistical Inference

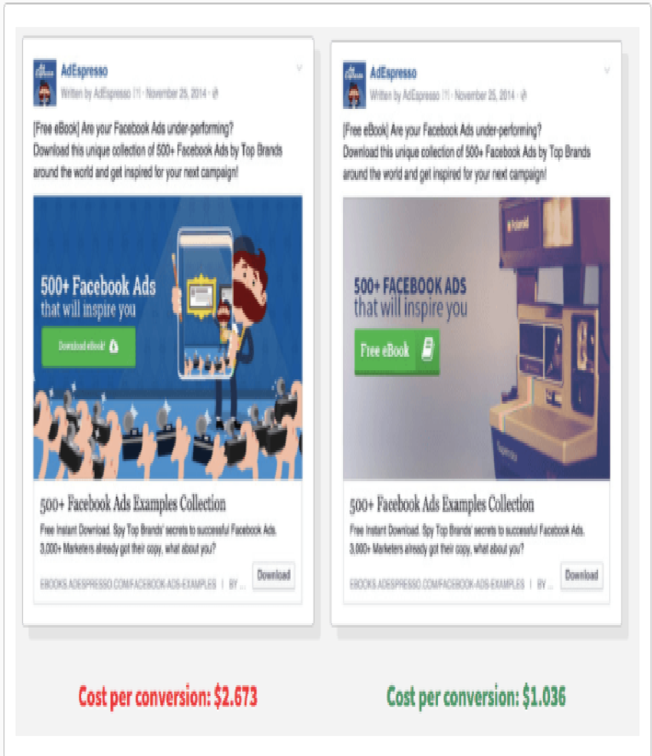
- A **statistical inference** helps to make conclusions or decisions using data.

Statistical Inference

- A **statistical inference** helps to make conclusions or decisions using data.
- An *inference* in general is an uncertain conclusion.

Statistical Inference

- A **statistical inference** helps to make conclusions or decisions using data.
- An *inference* in general is an uncertain conclusion.
- Two things mark out **statistical inference**:
  - the information on which they are based is statistical (i.e., subject to randomness);
  - our conclusion is uncertain, and attempt to measure the uncertainty involved.



Experimental TB vaccine shows promi in clinical trials

By HELEN BRANSWELL @HelenBranswell / SEPTEMBER 28, 2018

A child is given a new TB vaccine as part of a clinical trial in South Africa in 2011.
RODGER BOSCH/AFP/GETTY IMAGES

As world leaders pledged support for the fight against tuberculosis at the United Nations this week, some good news in the effort to develop weapons to combat the bacterium nearly slipped under the radar.

An experimental TB vaccine showed solid protection in a clinical trial reported Tuesday in the New England Journal of Medicine. The vaccine is being developed by GSK and Aeras, a nonprofit organization working on affordable tuberculosis vaccines.

The vaccine was tested in volunteers with latent tuberculosis — in other words, people who had been infected, but who did not at the time of vaccination have active TB disease. People who received placebo vaccine progressed from latent to active disease at roughly twice the rate of people in the trial who received the active vaccine.

Statistical Inference

Statistical inference can answer questions such as:

1. If more people buy the product after seeing two versions of the same web page (A and B) is the difference due to chance or due to different versions of the web page?

Statistical Inference

Statistical inference can answer questions such as:

1. If more people buy the product after seeing two versions of the same web page (A and B) is the difference due to chance or due to different versions of the web page?

2. If less people who received the experimental TB vaccine are infected with TB compared to people that didn't receive the vaccine is the difference due to chance or due to receiving the vaccine?

Statistical Inference

Statistical inference can answer questions such as:

1. If more people buy the product after seeing two versions of the same web page (A and B) is the difference due to chance or due to different versions of the web page?

2. If less people who received the experimental TB vaccine are infected with TB compared to people that didn't receive the vaccine is the difference due to chance or due to receiving the vaccine?

Sometimes inference isn't appropriate. For example, if we have data for all possible observations, there may be nothing to infer.

Statistical Inference

Significance Testing

"If I calculate *something* in my data, say a difference between two groups or a relationship between two variables or a value that is different than what I'd expect then, could this be simply due to chance, or is it an actual real difference or relationship?"

Significance Testing for a Single Proportion

Kissing the Right Way

Rodin's sculpture *The Kiss*

- Güntürkün (2003) recorded the direction kissing couples tilted their heads.
- Of the 124 couples he observed, 80 turned their heads to the right.
- 64.5% of couples tilted their heads to the right.
- Is this evidence of a right-side preference?

What would you expect to see if couples had no preference?

What would you expect to see if couples had no preference?

- In order to explore what we might expect to see if couples had no preference for tilting their heads to the left or right when kissing, we'll use **simulation**.

What would you expect to see if couples had no preference?

- In order to explore what we might expect to see if couples had no preference for tilting their heads to the left or right when kissing, we'll use **simulation**.
- Randomly generate data that under the assumption that couples have no preference (i.e. they are equally likely to tilt their heads to the left or right.)

What would you expect to see if couples had no preference?

- In order to explore what we might expect to see if couples had no preference for tilting their heads to the left or right when kissing, we'll use **simulation**.

- Randomly generate data that under the assumption that couples have no preference (i.e. they are equally likely to tilt their heads to the left or right.)

- We'll do this many times to see what values are possible under the assumption of no preference.

What simple activity simulates an event that can occur one way or another with equal probability?

Flip a coin once

```
sample(c("heads","tails"),
       size = 1,
       prob = c(0.5, 0.5))
```

```
## [1] "tails"
```

the choices from which to sample

← # of trials (how many times to flip the coin)

↳ or one couple

↳ probability of each of the choices (H or T here).

Flip a coin 124 times

The R code below simulates 124 flips of a coin or *simulating* 124 flips of a coin.

```
# randomly generate 124 flips of a coin -- a "simulation"
# probability is c(0.5, 0.5) by default
n_flips <- 124
coin_flips <- sample(c("heads", "tails"),
                     size = n_flips,
                     replace = TRUE)
```

# of times to flip the coin
(# of couples in the experiment)

```
data.frame(coin_flips) %>% head() #result of first 6 flips
```

```
##   coin_flips
## 1      heads
## 2      tails
## 3      heads
## 4      tails
## 5      tails
## 6      tails
```

← data frame

head() → returns first 6 rows pf data frame

```
table(coin_flips) #counts number of heads and tails
```

```
## coin_flips
## heads tails
##    53    71
```

Calculate the proportion of heads

Which of the 124 flips are heads?

```
coin_flips == "heads"
```

```
##   [1]  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
##  [12]  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE  TRUE  TRUE  TRUE FALSE
##  [23] FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE
##  [34] FALSE  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE FALSE  TRUE FALSE
##  [45] FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE
##  [56]  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE  TRUE
##  [67] FALSE FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE
##  [78]  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
##  [89]  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE  TRUE  TRUE
## [100]  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE  TRUE
## [111] FALSE FALSE FALSE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE
## [122]  TRUE  TRUE FALSE
```

to test equality while "=" is used in functs and for assignment

x <- 5 ✶ recommended

or x = 5

Calculate the proportion of heads

Count the number of heads (count how often `flips == "heads"` is TRUE).

```
sum(coin_flips == "heads")
```

from prev slide

```
## [1] 53
```

→ 53 heads in our simulation.

Calculate the proportion of heads in the simulation.

```
p_heads <- sum(coin_flips == "heads") / n_flips
p_heads
```

53/124

```
## [1] 0.4274194
```

Recall: how to reproduce 'randomness' in R?

- Simulations use functions in `R` that produce (apparently) random outcomes (for example, `sample`).

Recall: how to reproduce 'randomness' in R?

- Simulations use functions in `R` that produce (apparently) random outcomes (for example, `sample`).
- We can force such a function to produce the same outcome every time by setting a parameter called the "seed".

Recall: how to reproduce 'randomness' in R?

- Simulations use functions in `R` that produce (apparently) random outcomes (for example, `sample`).
- We can force such a function to produce the same outcome every time by setting a parameter called the "seed".
- The seed can be any integer.

Recall: how to reproduce 'randomness' in R?

- Simulations use functions in `R` that produce (apparently) random outcomes (for example, `sample`).
- We can force such a function to produce the same outcome every time by setting a parameter called the "seed".
- The seed can be any integer.
- I'll do that now, so that you can reproduce my results exactly with the following command:

```
set.seed(130)
```

(any integer

Simulate 124 head tilts when kissing, assuming that left or right is equally likely

Set the random seed to get the same answer every time

```
set.seed(130)
n_observations <- 124
```

Create an empty vector to store the results to store 1000 results, initially it's filled with missing values (NAs).

this is where we'll put our simulated values

```
simulated_stats <- rep(NA, 1000)
sim <- sample(c("right", "left"),
              size = n_observations,
              replace = TRUE)
sim_p <- sum(sim == "right") / n_observations
sim_p
```

→ vector of 124 "right" and "left" values

↳ proportion of 'right" in our vector

```
## [1] 0.4435484
```

Add the new simulated value to the first entry in the vector of results.

```
simulated_stats[1] <- sim_p
```

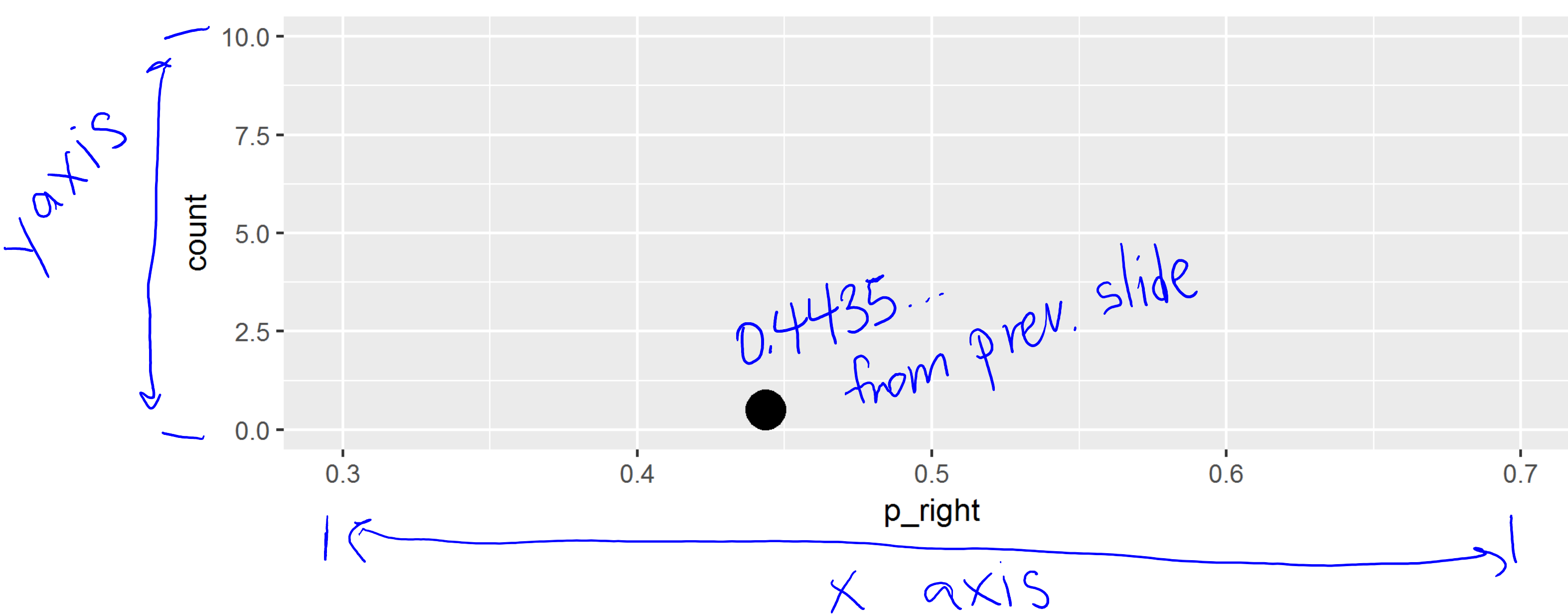
↳ vector for the proportion from each simulated dataset

Turn results into a data frame.

```
sim1 <- data_frame(p_right = simulated_stats)
```

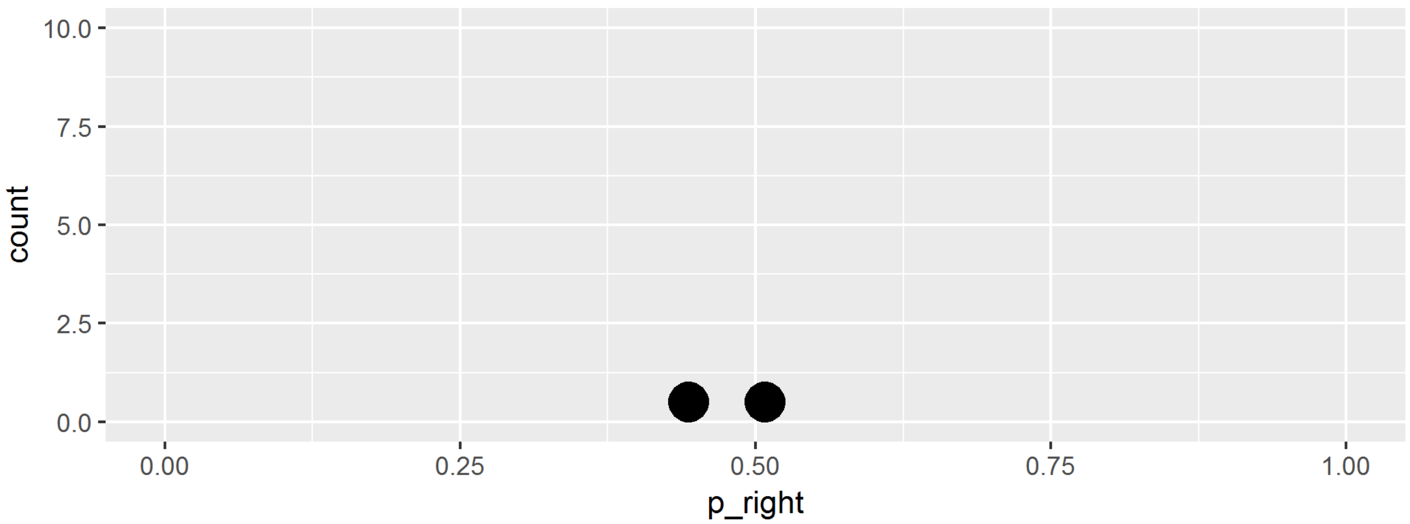
Plot using `ggplot`

```
sim1 %>% ggplot(aes(x=p_right)) +
  geom_dotplot() +
  xlim(0.3, 0.7) +
  ylim(0, 10)
```



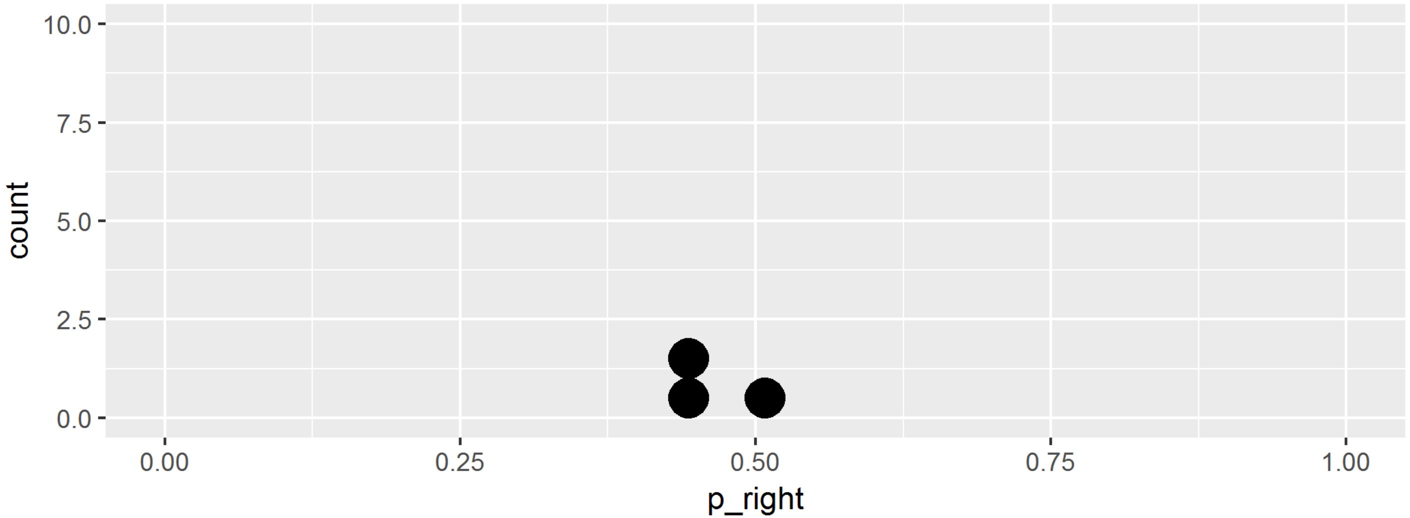
Add another simulation

```
## [1] 0.5080645
```



And another simulation

```
## [1] 0.4435484
```



`for` loops

- Automate the process of generating many simulations

`for` loops

- Automate the process of generating many simulations
- Evaluate a block of code for each value of a sequence

`for` loops

- Automate the process of generating many simulations
- Evaluate a block of code for each value of a sequence
- The following `for` loop will evaluate SOME CODE 1000 times, for `i=1` and `i=2` and ... and `i=1000`

`for` loops

- Automate the process of generating many simulations
- Evaluate a block of code for each value of a sequence
- The following `for` loop will evaluate SOME CODE 1000 times, for `i=1` and `i=2` and ... and `i=1000`
- Note that SOME CODE is within curly brackets

# of times you want to repeat your code

```
for (i in 1:1000)
{
  SOME CODE
}
```

Set values for simulation.

```
n_observations <- 124 # number of obeservations
repetitions <- 1000 # 1000 simulations
simulated_stats <- rep(NA, repetitions) # 1000 missing values
set.seed(101)
```

Automate simulation with a `for` loop.

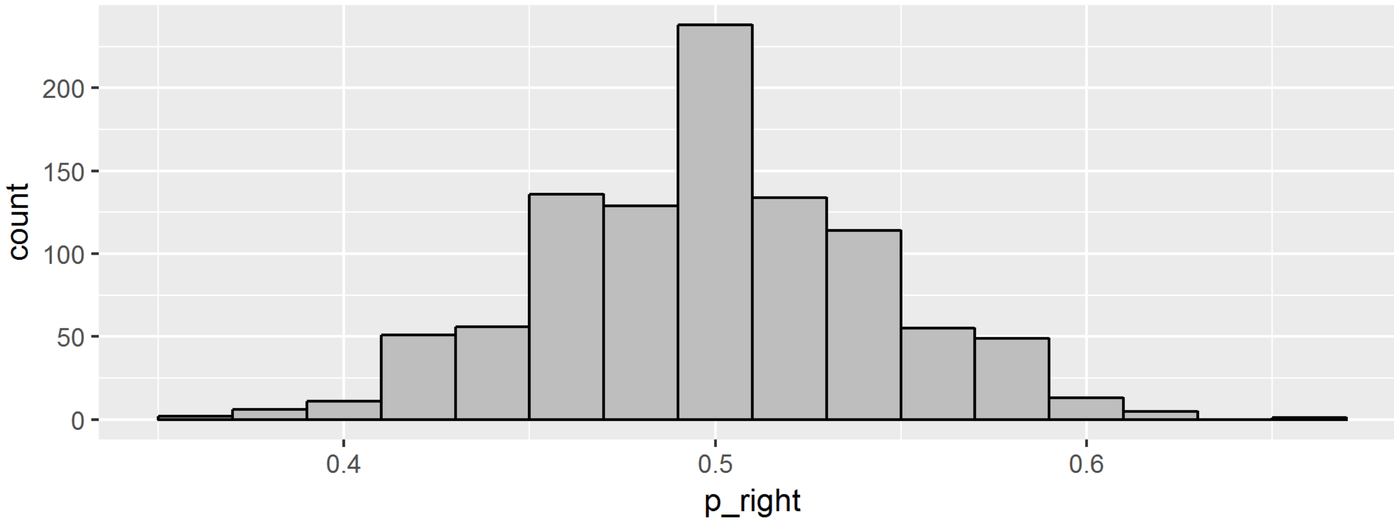
```
for (i in 1:repetitions)
{
  new_sim <- sample(c("right", "left"),
                    size = n_observations,
                    replace = TRUE)
  sim_p <- sum(new_sim == "right") / n_observations
  # add the new simulated value to the ith entry
  # in the vector of results
  simulated_stats[i] <- sim_p
}
```

Turn results into a data frame.

```
sim <- data_frame(p_right = simulated_stats)
```

Plot results

```
sim %>% ggplot(aes(x = p_right)) +
  geom_histogram(binwidth = 0.02, colour = "black", fill = "grey")
```



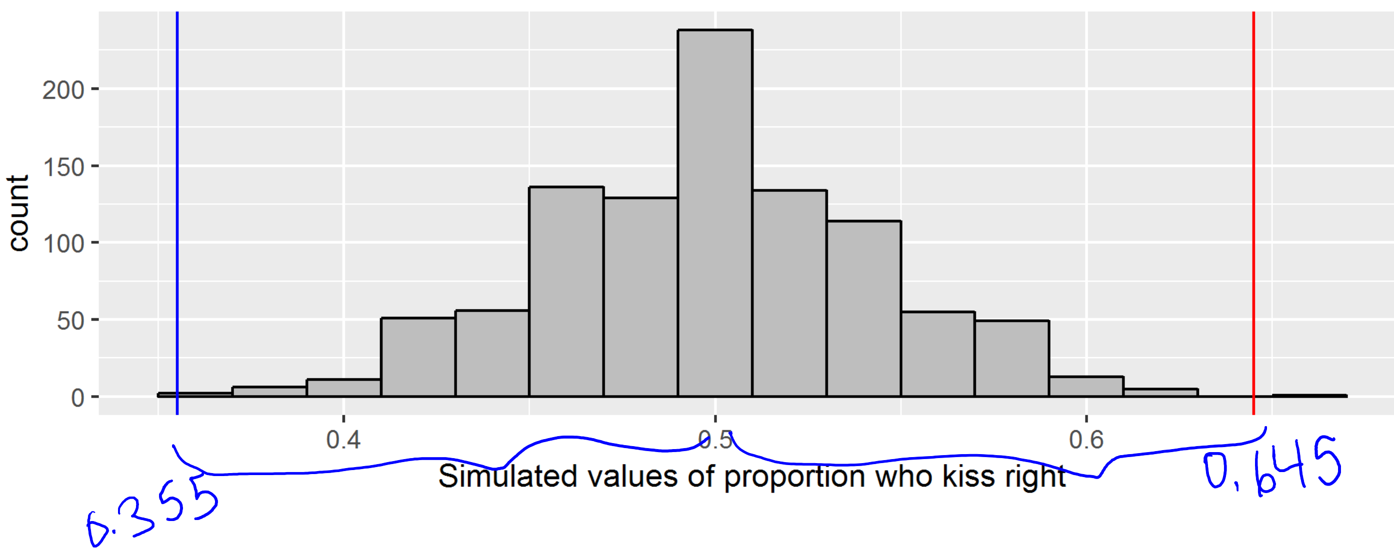
This is a histogram of the proportions of "right" obtained in each of the 1000 simulations

How unusual is a value of 0.645, if tilting to the right or left is equally likely?

⇒ How many simulated proportions are further away from 0.5 than our true observation

```
sim %>% ggplot(aes(p_right)) +
  geom_histogram(binwidth = 0.02, colour = "black", fill = "grey") +
  geom_vline(xintercept = 0.645, color = "red") +
  geom_vline(xintercept = 0.355, color = "blue") +
  labs(x = "Simulated values of proportion who kiss right")
```

← drawing the red line



How unusual is a value of 0.645, if tilting to the right or left is equally likely?

This includes values that are ≥ 0.645 as well as values that are ≤ 0.355 since 0.355 is as far from 0.5 as 0.645.

Calculate the proportion of our simulated observations that are as unusual or more unusual than 0.645:

How unusual is a value of 0.645, if tilting to the right or left is equally likely?

This includes values that are ≥ 0.645 as well as values that are ≤ 0.355 since 0.355 is as far from 0.5 as 0.645.

Calculate the proportion of our simulated observations that are as unusual or more unusual than 0.645:

In R, the vertical bar `|` means **or**.

```
sim %>%
filter(p_right >= 0.645 | p_right <= 0.355) %>%
summarise(p_value = n() / repetitions)
```

OR

keeping only the rows satisfying condition

counting # of rows

here this is 1000

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.001
```

The Logic of Hypothesis Testing

1. The hypotheses

Two claims:

1. Couples are equally likely to tilt to the right or left. This is the **null hypothesis**, written H_0. The proportion who kiss to the right is one-half.

$$H_0: p = 0.5$$

1. The hypotheses

Two claims:

1. Couples are equally likely to tilt to the right or left. This is the **null hypothesis**, written H_0. The proportion who kiss to the right is one-half.

$$H_0 : p = 0.5$$

p= proportion who tilt to the right

2. Couples are more likely to prefer one side. This is the **alternative hypothesis**, written H_A (or H_a or H_1).
   For the kissing example, if there is something going on, the proportion who kiss to the right should be something other than one-half.

$$H_A : p \neq 0.5$$

2. Parameters, Statistics, Test Statistics

A **parameter**: "true" value of what we're interested in, typically, because it's what holds for the population.

A **statistic** is a number that describes the sample. The value of a statistic will changes from sample to sample.

A **test statistic** measures the compatiability between null hypothesis and the data.

Kissing example

p : proportion who tilt to the right (or left...)

$\hat{p}$: sample proportion

"hat" denotes that it comes from a sample

2. Parameters, Statistics, Test Statistics

A **parameter**: "true" value of what we're interested in, typically, because it's what holds for the population.

A **statistic** is a number that describes the sample. The value of a statistic will changes from sample to sample.

A **test statistic** measures the compatiability between null hypothesis and the data.

In the kissing example:

Parameter: p: the true proportion of people who kiss to the right

Statistic: $\hat{p}$: the proportion of people who kiss to the right. The value of a statistic can be different from sample to sample.

The **test statistic** is a number, calculated from the data. For the kissing example, the test statistic we'll use is $\hat{p} = 80/124 = 0.645$

3. Simulate what the null hypothesis predicts will happen

The **distribution** of the statistic is the pattern of values it could be, including an indication of how likely those values are to occur.

3. Simulate what the null hypothesis predicts will happen

The **distribution** of the statistic is the pattern of values it could be, including an indication of how likely those values are to occur.

A simulation is a way to explore random events, such as what some data or a test statistic could look like under certain assumptions. By observing many simulated outcomes, we can see what values are possible and the distribution of these possible values.

3. Simulate what the null hypothesis predicts will happen

The **distribution** of the statistic is the pattern of values it could be, including an indication of how likely those values are to occur.

A simulation is a way to explore random events, such as what some data or a test statistic could look like under certain assumptions. By observing many simulated outcomes, we can see what values are possible and the distribution of these possible values.

We want to know the distribution of what the test statistic could be if the null hypothesis were true.

3. Simulate what the null hypothesis predicts will happen

The **distribution** of the statistic is the pattern of values it could be, including an indication of how likely those values are to occur.

A simulation is a way to explore random events, such as what some data or a test statistic could look like under certain assumptions. By observing many simulated outcomes, we can see what values are possible and the distribution of these possible values.

We want to know the distribution of what the test statistic could be if the null hypothesis were true.

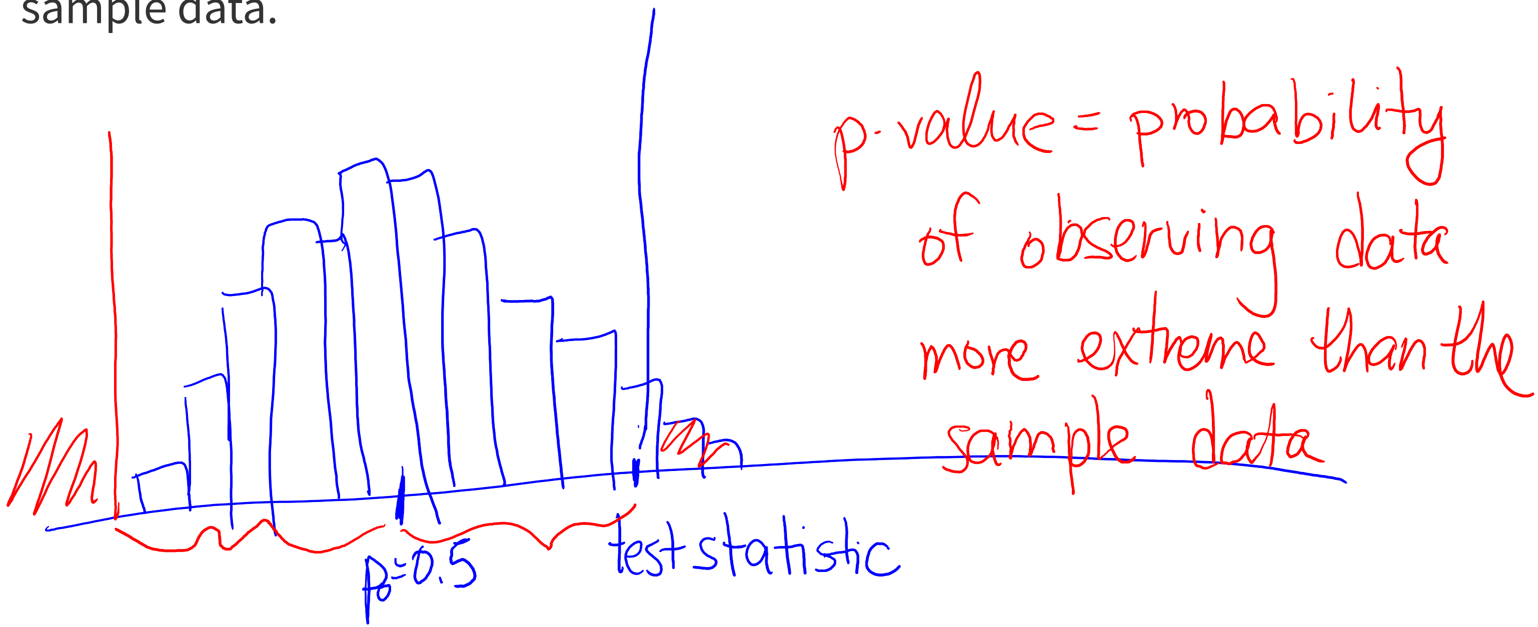
To get an estimate of this, simulate many possible values of the statistic under the assumption that the null hypothesis is true.

This is the **empirical distribution** of the test statistic under the null hypothesis.

↳ ex: histogram from a few slides ago

4. The P-value

- Assuming that the null hypothesis is true, the **P-value** is the probability of observing data that are **at least as unusual** as the sample data.



4. The P-value

- Assuming that the null hypothesis is true, the **P-value** is the probability of observing data that are **at least as unusual** as the sample data.

- We estimate the P-value as the proportion of observations in the empirical distribution that yield a statistic as extreme or more extreme than the test statistic calculated from our data.

4. The P-value

- What does "as extreme or more extreme" mean?

4. The P-value

- What does "as extreme or more extreme" mean?
  - Values that are as far away or even farther from the null hypothesis value.

4. The P-value

- What does "as extreme or more extreme" mean?
  - Values that are as far away or even farther from the null hypothesis value. → in either direction

For the kissing example:

- the null hypothesis value: $p = 0.5$

4. The P-value

- What does "as extreme or more extreme" mean?
  - Values that are as far away or even farther from the null hypothesis value.

For the kissing example:

- the null hypothesis value: $p = 0.5$
- the observed estimate from the data: $\hat{p} = 0.645$

4. The P-value

- What does "as extreme or more extreme" mean?
  - Values that are as far away or even farther from the null hypothesis value.

For the kissing example:

- the null hypothesis value: $p = 0.5$
- the observed estimate from the data: $\hat{p} = 0.645$
- values at least as unusual as the data values: all values **greater than or equal to 0.645** and all values **less than or equal to 0.355**

4. The P-value

- What does "as extreme or more extreme" mean?
  - Values that are as far away or even farther from the null hypothesis value.

For the kissing example:

- the null hypothesis value: $p = 0.5$
- the observed estimate from the data: $\hat{p} = 0.645$
- values at least as unusual as the data values: all values **greater than or equal to 0.645** and all values **less than or equal to 0.355**
- This is a **two-sided test** because it considers differences from the null hypothesis that are both larger and smaller than what you observed.
  (It is also possible to carry out one-sided tests. They are useful in some specific applications.)

5. Make a conclusion

- P-values are probabilities so are between 0 and 1. Small probabilities correspond to events that are unlikely to happen and large values correspond to events that are likely to happen.

5. Make a conclusion

- P-values are probabilities so are between 0 and 1. Small probabilities correspond to events that are unlikely to happen and large values correspond to events that are likely to happen.

- A small P-value means the data are inconsistent with the null hypothesis. A **statistically significant** result is associated with a small P-value.

If pvalue is small:
↳ either we observed something very unusual
or ↳ the null hypothesis is "wrong" (i.e. the data is inconsistent with the null)

5. Make a conclusion

- P-values are probabilities so are between 0 and 1. Small probabilities correspond to events that are unlikely to happen and large values correspond to events that are likely to happen.

- A small P-value means the data are inconsistent with the null hypothesis. A **statistically significant** result is associated with a small P-value. ⇒ reject the null hypothesis

- A large P-value means the data are consistent with the null hypothesis.
  ↳ does not prove that H_0 is true
  ↳ fail to reject the null hypothisis

You can never prove that the null hypothesis is true

5. Make a conclusion

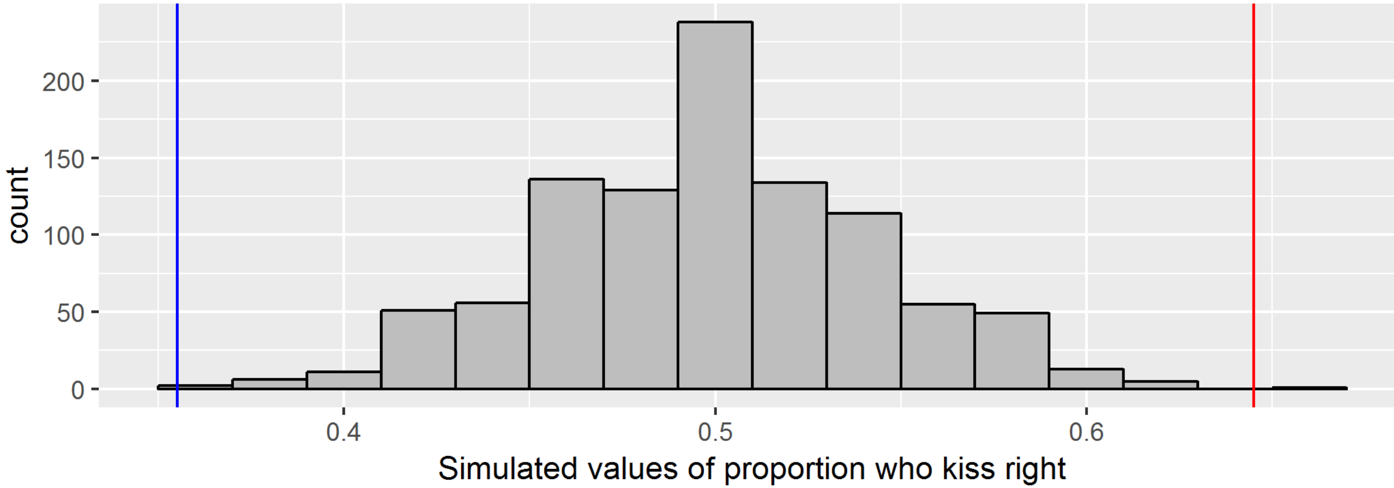
Some guidelines for how small is small:

P-value	Evidence
p-value > 0.10	no evidence against H_0
0.05 < p-value < 0.10	weak evidence against H_0
0.01 < p-value < 0.05	moderate evidence against H_0
0.001 < p-value < 0.01	strong evidence against H_0
p-value < 0.001	very strong evidence against H_0

generally statistical significant

Simulation results and P-value for kissing ex.

from before




```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.001
```

Conclusion for the Kissing Example

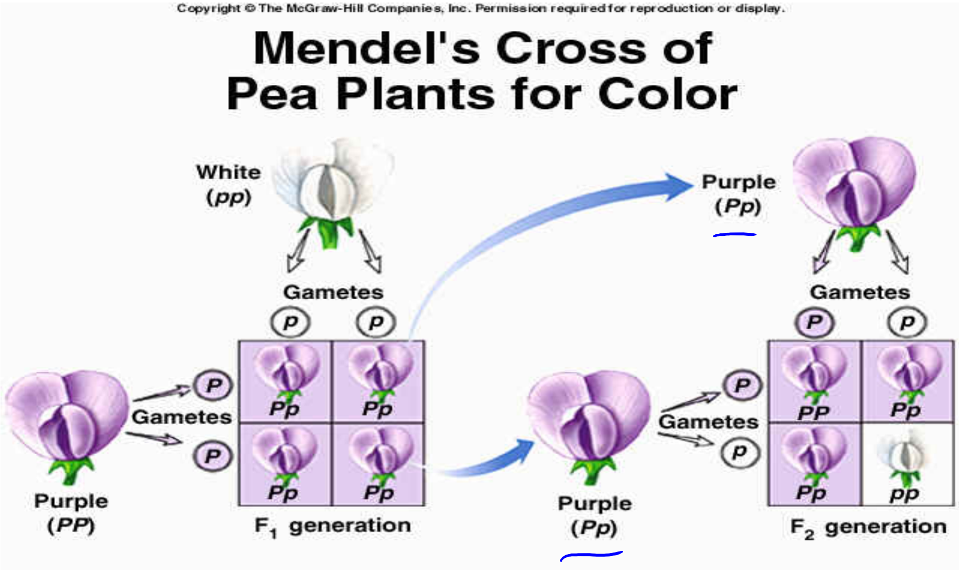
Since the P-value is 0.001 we conclude that we have we have strong evidence against the null hypothesis. The data provide convincing evidence that people are more likely to tilt their heads to one direction when they kiss.

Another example: Mendel's Pea Flowers

(This example is adapted from Computational and Inferential Thinking by Adhikari and DeNero.)

- Mendel (1822-1884) conducted experiments that resulted in the development of some fundamental laws of genetics.
- He formulated assumptions which gave theoretical models for how genetics work in pea plants and collected data to test the validity of his models.

Mendel's Pea Flowers



- In one variety of pea plant, his model predicted that the plants should have purple or white flowers, determined randomly, occurring in the ratio: 3 plants with purple flowers for every 1 plant with white flowers.

data

- He grew 929 plants. 705 had purple flowers and 224 had white flowers.

Steps to testing whether the data are consistent with Mendel's model

1. Formulate null and alternative hypotheses.
2. Calculate a test statistic from the data.
3. Simulate many values of what the test statistic could possibly have been if the null hypothesis were true.
4. Calculate the P-value.
5. Make a conclusion.

What would be appropriate null and alternative hypotheses to test Mendel's theory?

H_0: $p = 0.75$ p: the proportion of purple flowers

H_A: $p \neq 0.75$

What would be an appropriate test statistics?

The data: He grew 929 plants. 705 had purple flowers and 224 had white flowers.

$$\hat{p} = \frac{705}{929} = \frac{\text{\# of purple}}{\text{\# of flowers}}$$

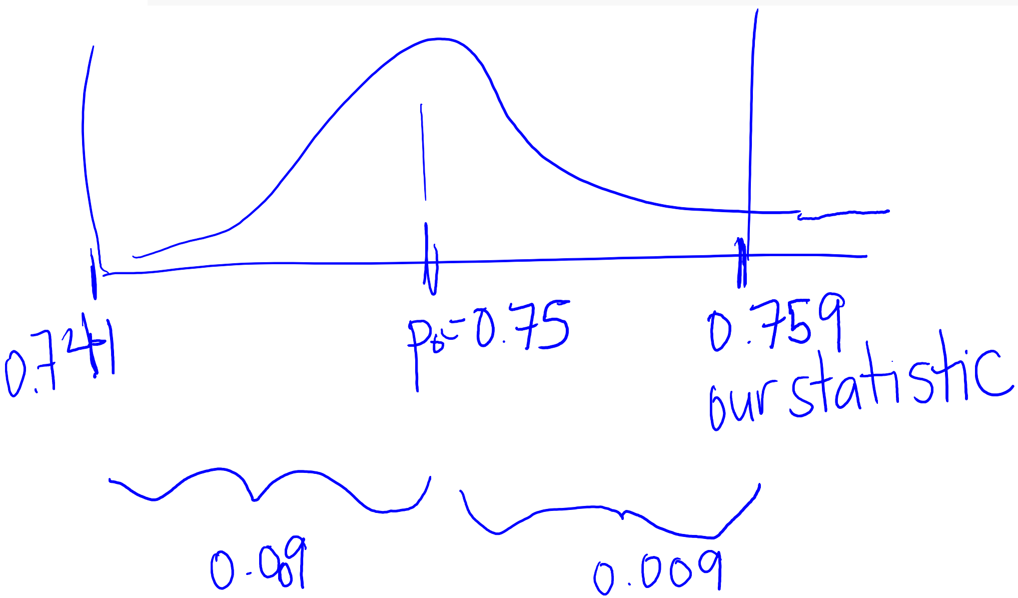
Simulate many values of what we'd observe if the null hypothesis were true

Here is the code for the kissing example. What values do we need to change?

```
repetitions <- 1000
simulated_stats <- rep(NA, repetitions) # 1000 missing values

n_observations <- 124

test_stat <- 80/124

set.seed(101)
for (i in 1:repetitions)
{
  new_sim <- sample(c("right", "left"),
                    size=n_observations,
                    replace=TRUE)
  sim_p <- sum(new_sim == "right") / n_observations
  simulated_stats[i] <- sim_p
}
```

Here is more code for the kissing example. What values do we need to change?

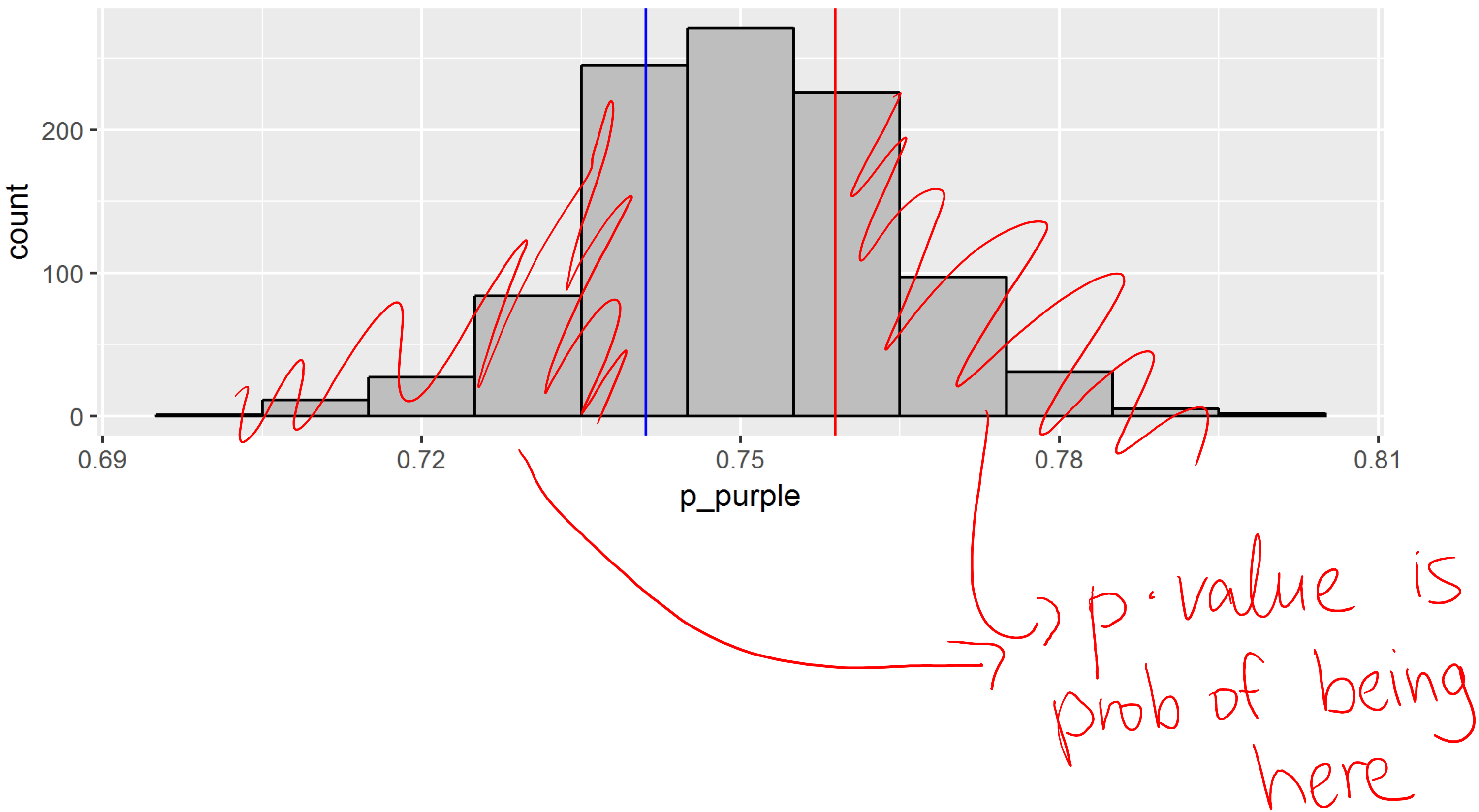
```
sim <- data_frame(p_right = simulated_stats)

ggplot(sim, aes(p_right)) +
geom_histogram(binwidth=0.02, colour = "black", fill = "grey") +
geom_vline(xintercept = 0.645, color="red") +
  geom_vline(xintercept = 0.355, color="blue")

sim %>%
filter(p_right >= 0.645 | p_right <= 0.355) %>%
summarise(p_value = n() / repetitions)
```



Results for Mendel's pea plant example

```r
set.seed(130)
repetitions <- 1000
simulated_stats <- rep(NA, repetitions) # 1000 missing values

n_observations <- 929

test_stat <- 705/929
other_extreme <- 0.75 - (705/929 - 0.75)

set.seed(101)
for (i in 1:repetitions)
{
  new_sim <- sample(c("purple", "white"),
                    size = n_observations,
                    prob = c(0.75,0.25),
                    replace = TRUE)
  sim_p <- sum(new_sim == "purple") / n_observations
  simulated_stats[i] <- sim_p
}
```

```
sim <- data_frame(p_purple = simulated_stats)

ggplot(sim, aes(p_purple)) +
geom_histogram(binwidth = 0.01, colour = "black", fill = "grey") +
geom_vline(xintercept = test_stat, color = "red") +
  geom_vline(xintercept = other_extreme, color = "blue")
```



```
sim %>%
filter(p_purple >= test_stat | p_purple <= other_extreme) %>%
summarise(p_value = n() / repetitions)
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.547
```

Conclusion

Mendel observed 705 purple-flowered plants in his 929 plants, a proportion of 0.7589.

Assuming that the probability that a pea plant will produce purple flowers is 0.75, the probability of observing a proportion that differs from 0.75 as much or more than 0.7589 is 0.55. Therefore we have no evidence against the null hypothesis that the probability that a pea plant will produce a purple flower is 0.75.

How many simulations is enough?

- In our examples, we've looked at 1000 simulated values assuming the null hypothesis is true, to compare to the value of our test statistic.
- In practice, the number of simulations is more typically on the order of 10,000.
- But that can take a long time to run.

A mathematical note

[Not responsible for on test and exam.]

- You could determine the P-value exactly using a *binomial* probability model.
- A binomial probability model is used to count the number of "successes" in n independent trials, where each trial has two possible outcomes: "success" with probability p or "failure" with probability $(1 - p)$.
- The probability of k successes in n trials is

$$\binom{n}{k} p^k (1 - p)^{n-k}$$

You'll study binomial probability models in second year statistics courses.