STA130H1S - Class #4

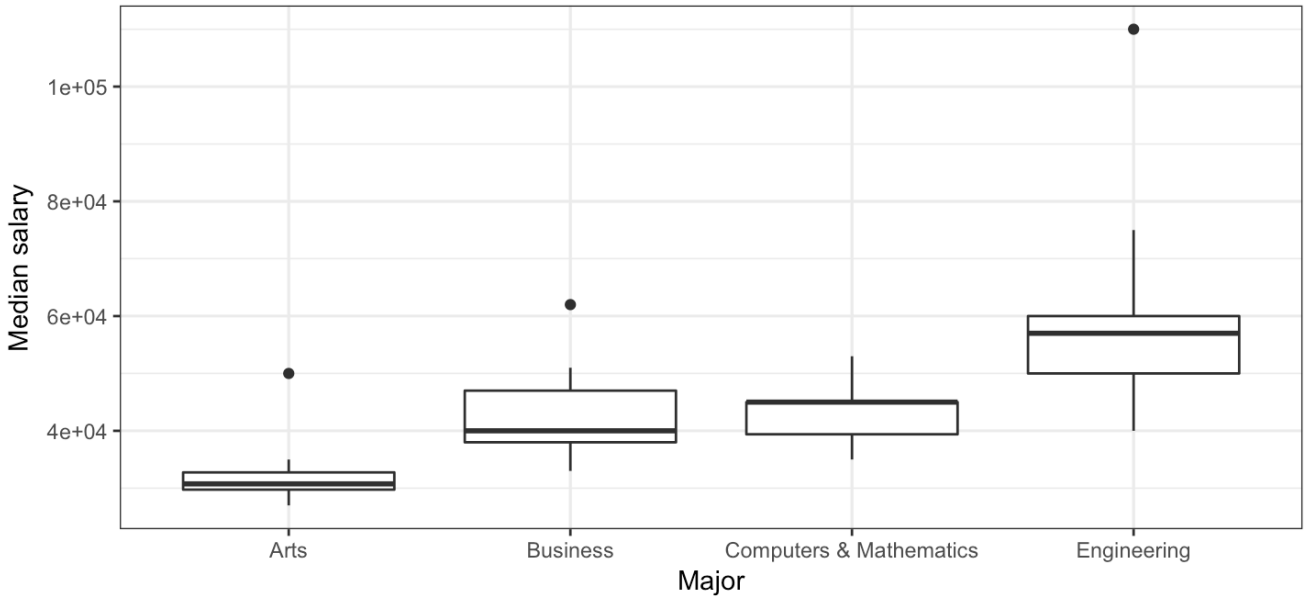
Inferential Thinking Part 1: Testing hypotheses

Prof. A. Gibbs

January 29, 2018

Another look at the "Economic Guide to Picking a Major" from last week:

- Is there evidence that salaries are, on average, higher for Engineering graduates than Arts graduates?
- Is there evidence that salaries are, on average, higher for Computers & Mathematics graduates than Business graduates?



Today

Answering the question:
is something we observe meaningful?
or could it just 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

- Sometimes the goal of statistical work is to make conclusions or decisions based on the incomplete information we have in our data – this procedure is known as *statistical inference*.
- Want to answer the question: "If I see 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, could this be simply due to chance or is it an actual real difference or relationship?"
- If we obtain results that we think are not due to chance,
  - Can we generalize them to a larger group or even to the whole world?
  - Do our results support a novel theoretical model?
  - When we do see a relationship between two variables, can we say one variable causes the other to change?

Statistical Inference

- Imagine we have a "real world" where we observe data, and a "theoretical world" where we have scientific models. Inference connects what we have observed in the real world to what we can say about the theoretical world.
- Sometimes the "theoretical world" is based on a scientific or mathematical theory and the "real world" observations are data that may confirm or contradict that theory.
- Sometimes the "real world" is a *sample* and the "theoretical world" is a *population*.
- Sometimes inference isn't appropriate. For example, if we have data for all possible observations, there may be nothing to make inference about.

- Today: **Hypothesis testing**
  "If I see 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, could this be simply due to chance or is it an actual real difference or relationship?"

STATISTICAL INTERENCE

"Real world"

"Theoretical world"

Data

Scientific model

What I observe

CONCLUSIONS

Sample

Population

Inference

Hypothesis Testing for a Single Proportion

Kissing the Right Way

Rodin's sculpture *The Kiss*.
Photo from http://www.musee-rodin.fr/en/collections/sculptures/kiss

Kissing the Right Way

- 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?

proportion: $p_{right} = 0.645$

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

50-50

left - 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**.
- We'll randomly generate data that follows the property that couples have no preference, that is 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?

coin flip, equally likely heads, tails

Flip a coin once

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

```
## [1] "heads"
```

Flip a coin 124 times

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

# view the sample
head(flips) # the first 6 values
```

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

```r
table(flips)
```

```
## flips
## heads tails
##    62    62
```

Calculate the proportion of heads

```
# check which of the 124 flips are heads
flips == "heads"
```

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

Calculate the proportion of heads

```
# count the number of heads (count how often flips == "heads" is TRUE)
sum(flips == "heads") # the sum function treats TRUE as 1 and FALSE as 0
```

sum treats TRUE as 1 and FALSE as 0 and adds them up.

```
## [1] 62
```

```
# calculate the proportion of heads in the simulation
p_heads <- sum(flips == "heads") / n_flips
p_heads
```

```
## [1] 0.5
```

A side note on generating random quantities in `R`

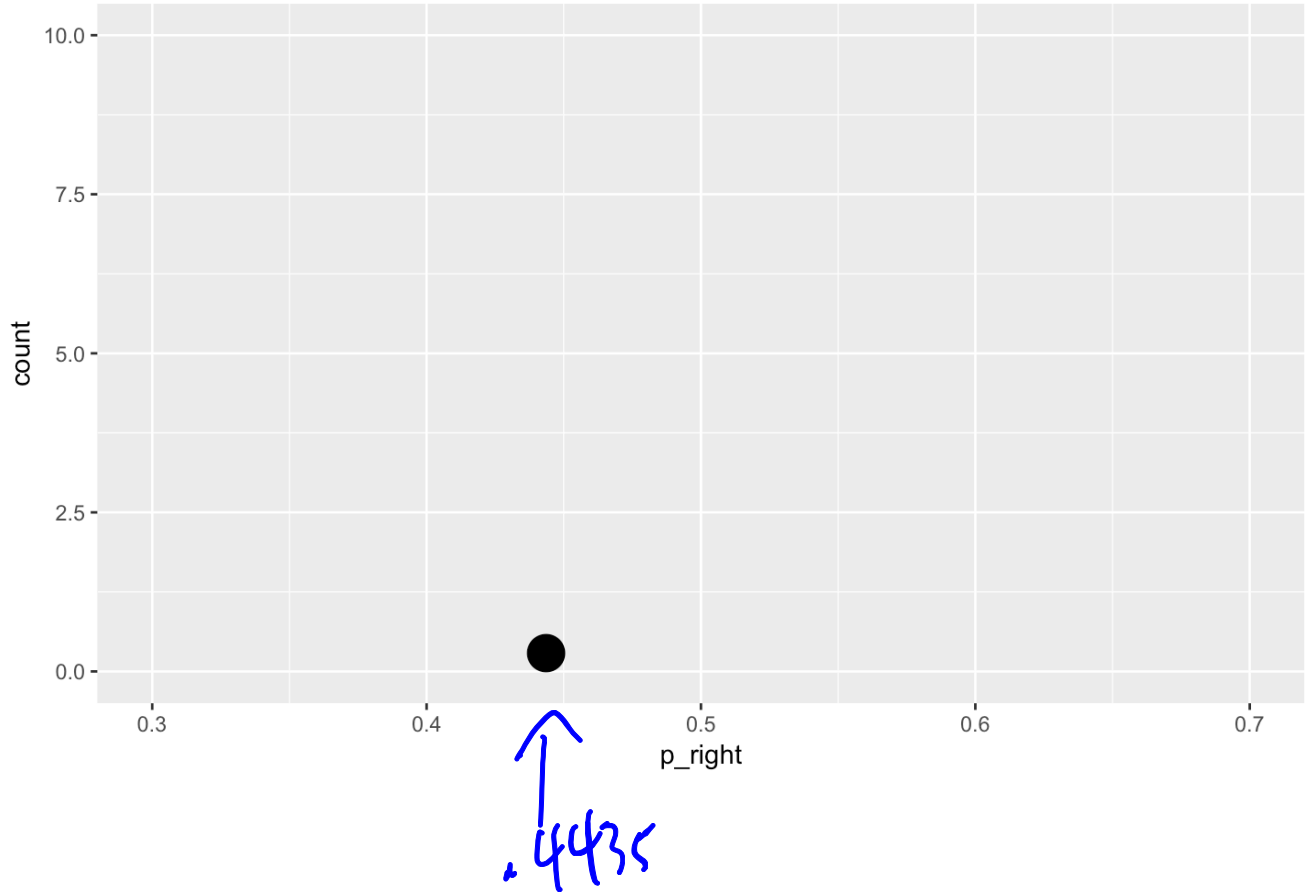
- Simulations use functions in `R` that produce (apparently) random outcomes (for example, `sample`).
- We can force such a function to produce the same outcome every time by setting a parameter called the "seed".
- The seed can be any integer.
- I'll do that now, so that you can reproduce my results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

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

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

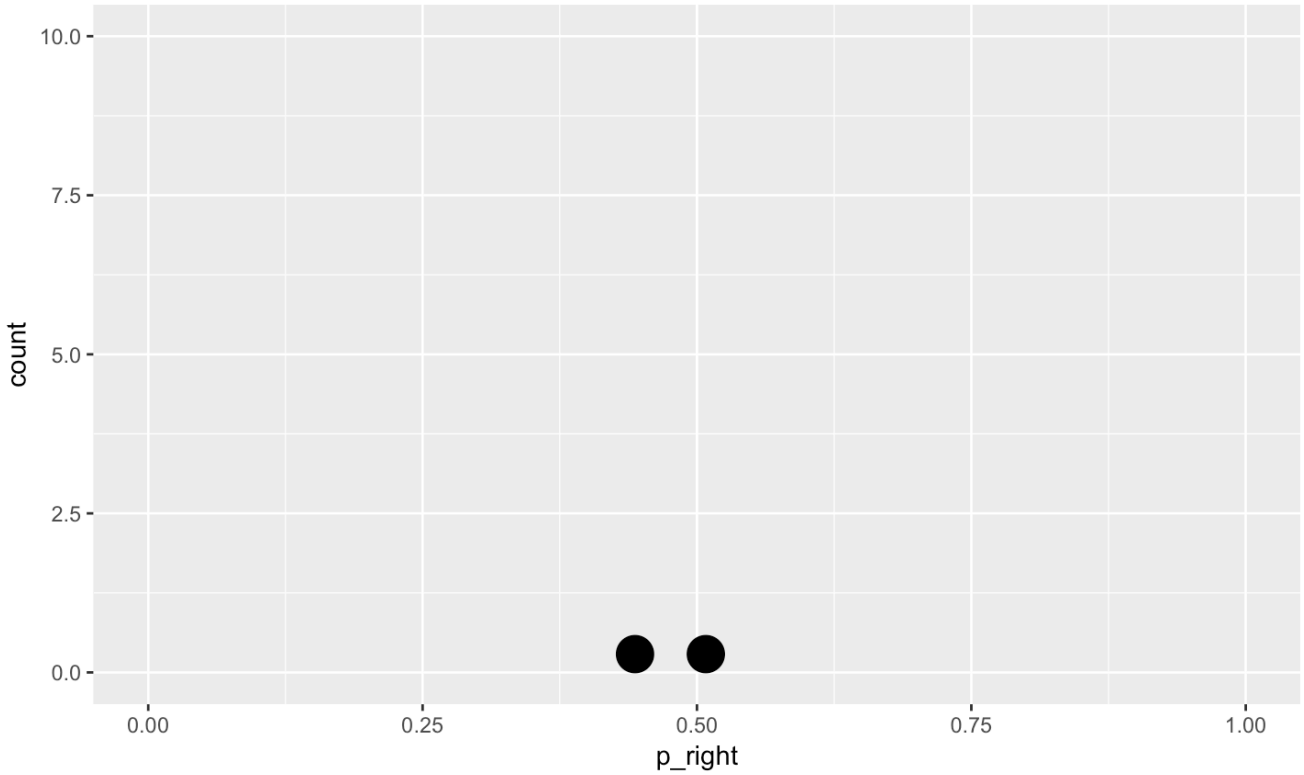
```
set.seed(130) # set the random number seed if you want to get the same answer every time
n_observations <- 124

# create an empty vector to store the results
# this vector can store 1000 results, and is filled with missing values (NA's)
simulated_stats <- rep(NA, 1000)

sim <- sample(c("right", "left"), size=n_observations, replace=TRUE)
sim_p <- sum(sim == "right") / n_observations
sim_p

## [1] 0.4435484

# add the new simulated value to the first entry in the vector of results
simulated_stats[1] <- sim_p
```

```
# turn results into a data frame
sim1 <- data_frame(p_right = simulated_stats)

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



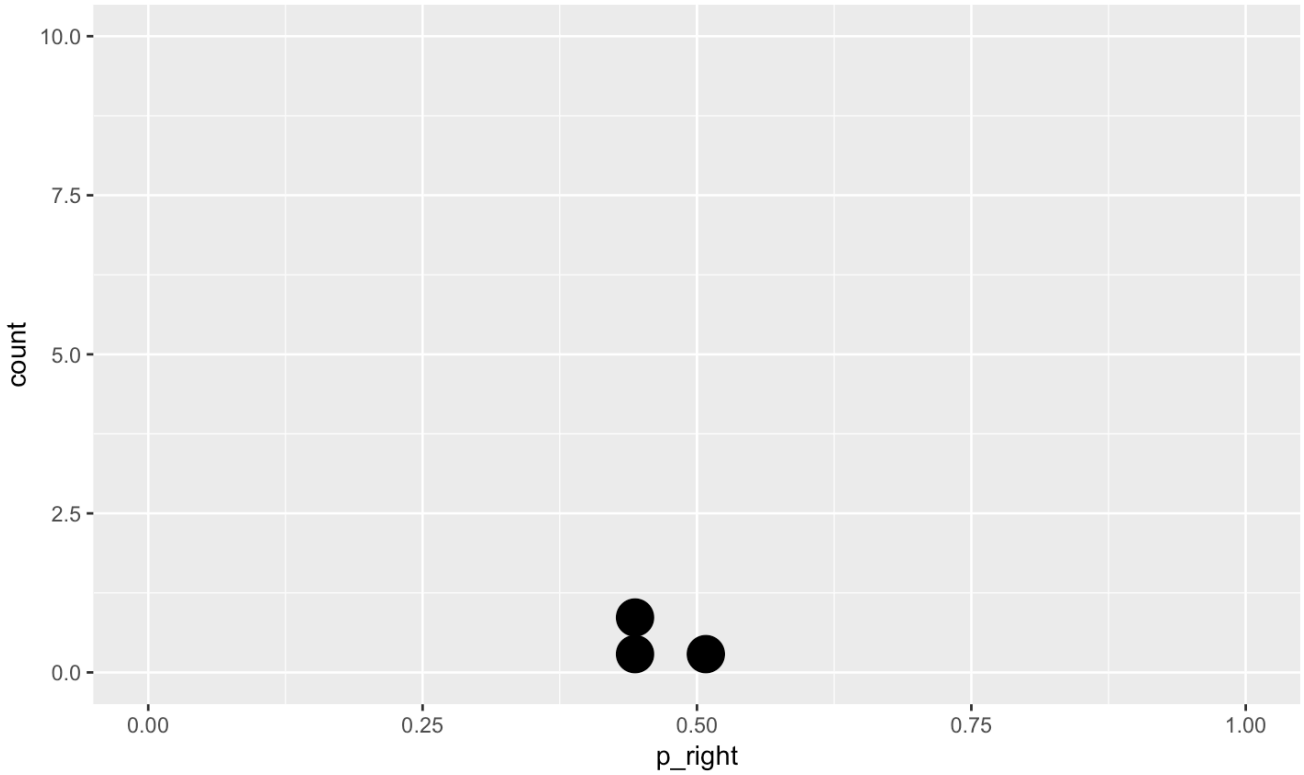
Add another simulation

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



And another simulation

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



`for` loops

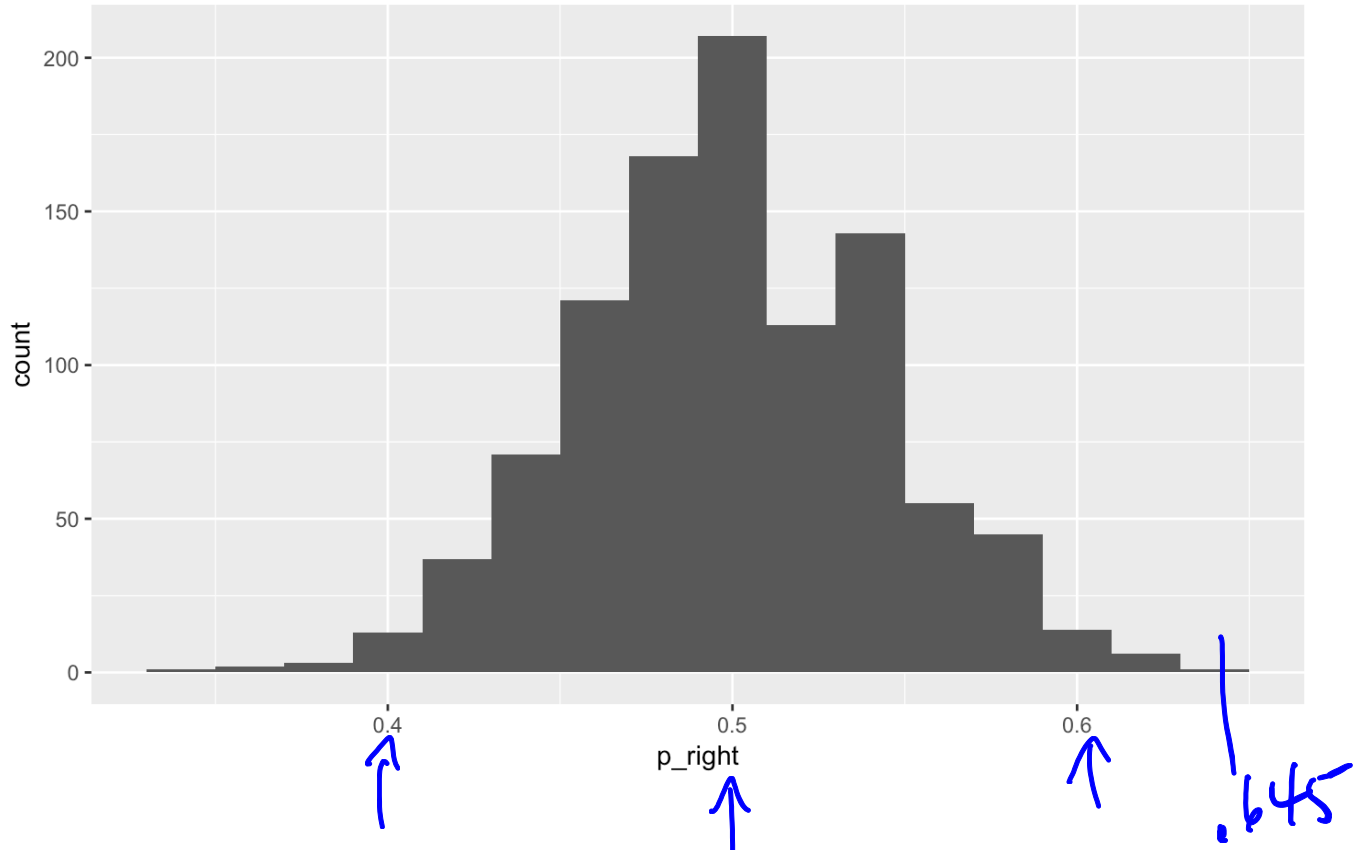
- Automate the process of generating many simulations
- Evaluate a block of code for each value of a sequence
- The following `for` loop will evaluate SOME CODE 1000 times, for i=1 and i=2 and … and i=1000
- Note that SOME CODE is within curly brackets

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

i is an index

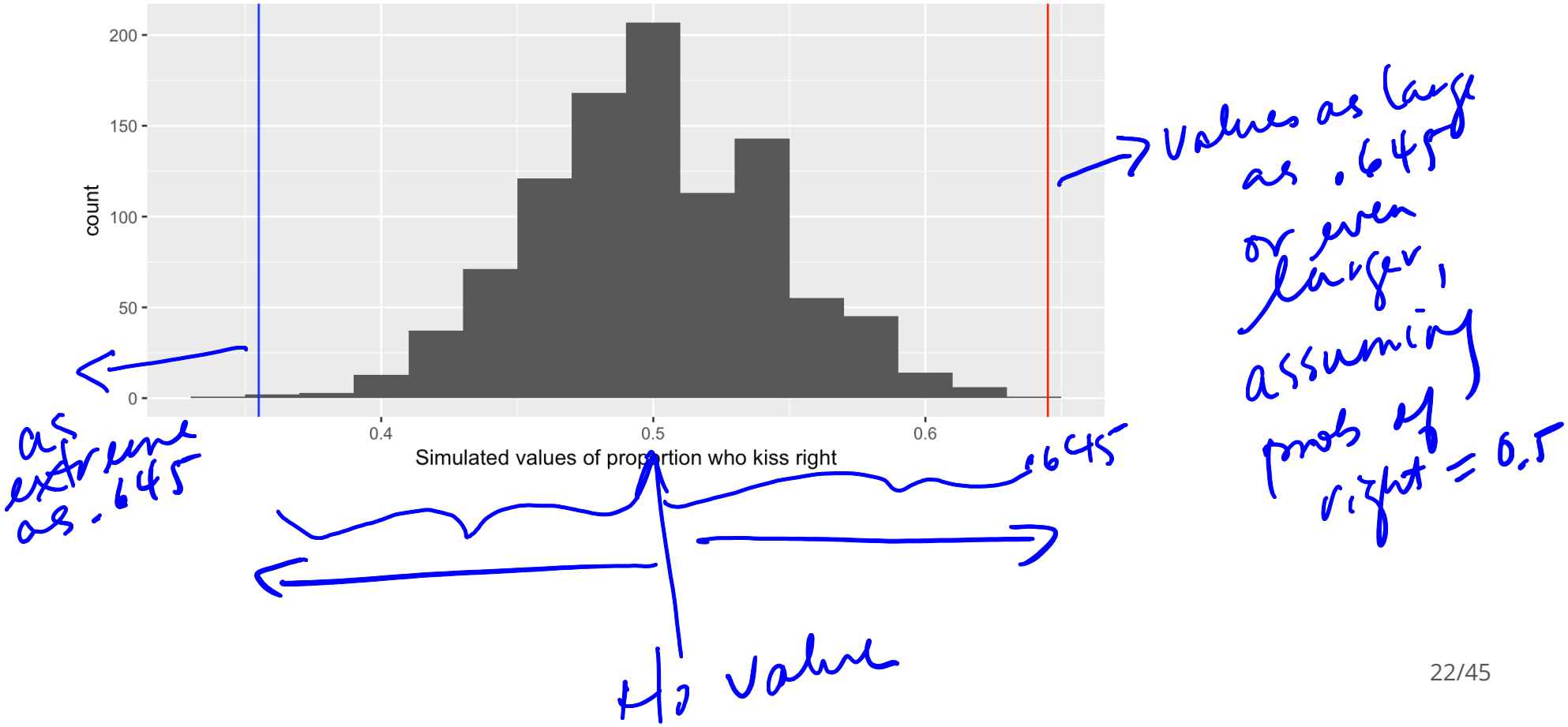
```r
# create a vector of missing values to store results
repetitions <- 1000
simulated_stats <- rep(NA, repetitions) # 1000 missing values

n_observations <- 124

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
ggplot(sim, aes(x=p_right)) + geom_histogram(binwidth=0.02)
```



How unusual is a value as unusual as 0.645?

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



How unusual is a value as unusual as 0.645?

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)
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.003
```

count ÷ by repetitions (1000) to get the proportion of observations in the extremes

The Logic of Hypothesis Testing

1. The hypotheses

Two claims:

1. There is nothing going on. This is the **null hypothesis**, written H_0. For the kissing example, if there is nothing going on, the proportion who kiss to the right should be one-half.

$$H_0 : p = 0.5$$

2. There is something going on. 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.

$$H_A : p \neq 0.5$$

   The alternative is almost always what the research wants to find evidence for.

2. The test statistic

Parameter vs statistic

A **parameter** is the "true" value of what we're interested in, typically, because it's what holds for the population.
A **statistic** is a quantity calculated from the data.
We estimate the **parameter** using a **statistic** calculated from our sample data to estimate the parameter. Often, estimates are specified by putting a "hat" ^ over the symbol for its corresponding parameter.

For 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 in our sample
$\hat{p} = 80/124 = 0.645$

The **test statistic** is a number, calculated from the data, that captures what we're interested in. For the kissing example, the test statistic we'll use is $\hat{p}$.

3. Simulate what the null hypothesis predicts will happen

The **distribution** of the test 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 test statistic under the assumption that the null hypothesis is true.
This is the **empirical distribution** of the test statistic under the null hypothesis.

4. The P-value

- Assuming that the null hypothesis is true, the **P-value** gives a measure of the probability of getting 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 gave a statistic as extreme or more extreme than the test statistic calculated from our data.

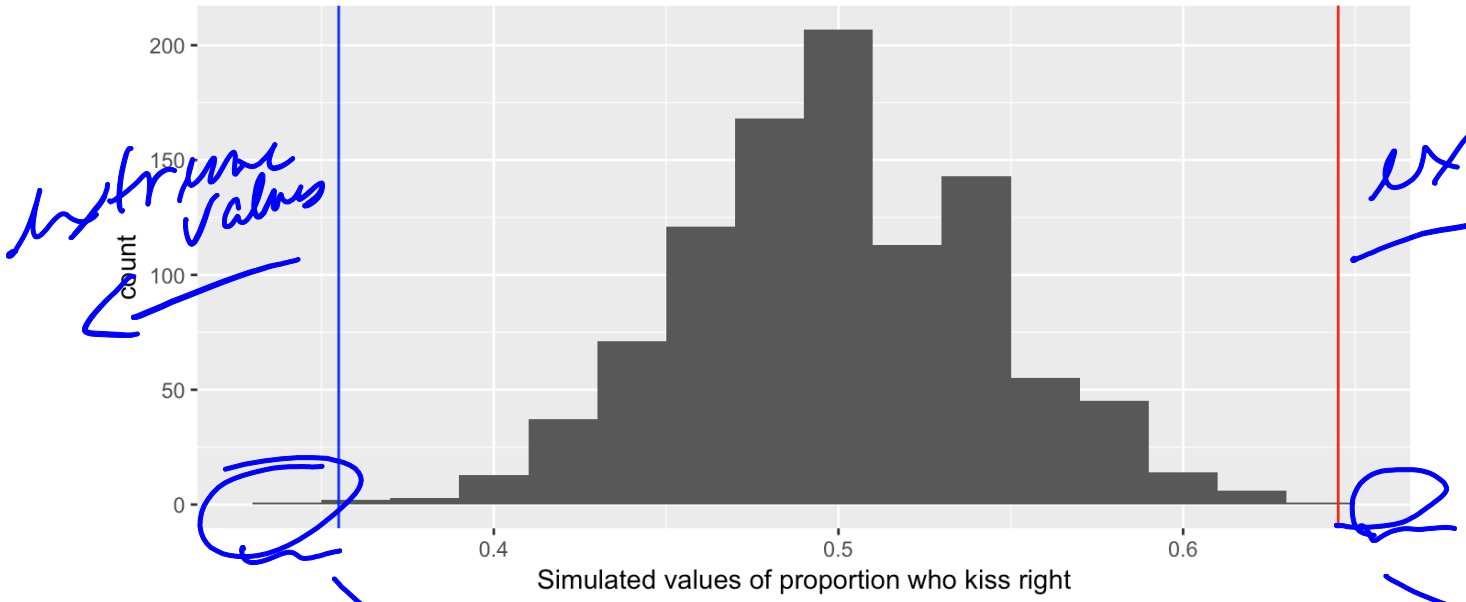
4. The P-value

- What does "as extreme or more extreme" mean?
  Values that are as far away or even farther from the null hypothesis value.
  For the kissing example:
  - the null hypothesis value is $p = 0.5$
  - the observed estimate from the data is $\hat{p} = 0.645$
  - values at least as unusual as the data values inclues 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. We'll only use two-sided tests, which are more objective.)

4. 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 large P-value means the data are consistent with the null hypothesis.
  - A small P-value means the data are inconsistent with the null hypothesis. A **statistically significant** result is associated with a small P-value.
- Some guidelines for how small is small:
  - A P-value above 0.10 means we have **no evidence** against H_0
  - A P-value less than 0.10 but greater than 0.05 means we have **weak evidence** against H_0
  - A P-value less than 0.05 but greater than 0.01 means we have **moderate evidence** against H_0
  - A P-value less than 0.01 but greater than 0.001 means we have **strong evidence** against H_0
  - A P-value less than 0.001 means we have **very strong evidence** against H_0

Simulation results and P-value for kissing ex.




```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.003
```

Conclusion for the Kissing Example

Since the P-value is 0.003 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

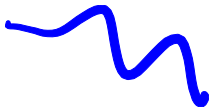
Mendel's Pea Flowers

(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.
- 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.
- 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: $p_{purple} = 0.75$

H_A: $p_{purple} \neq 0.75$

- p_{purple} be the proportion of plants with purple flower

What would be an appropriate test statistic?

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

$$\hat{p}_{purple} = \frac{705}{929} = 0.75888$$

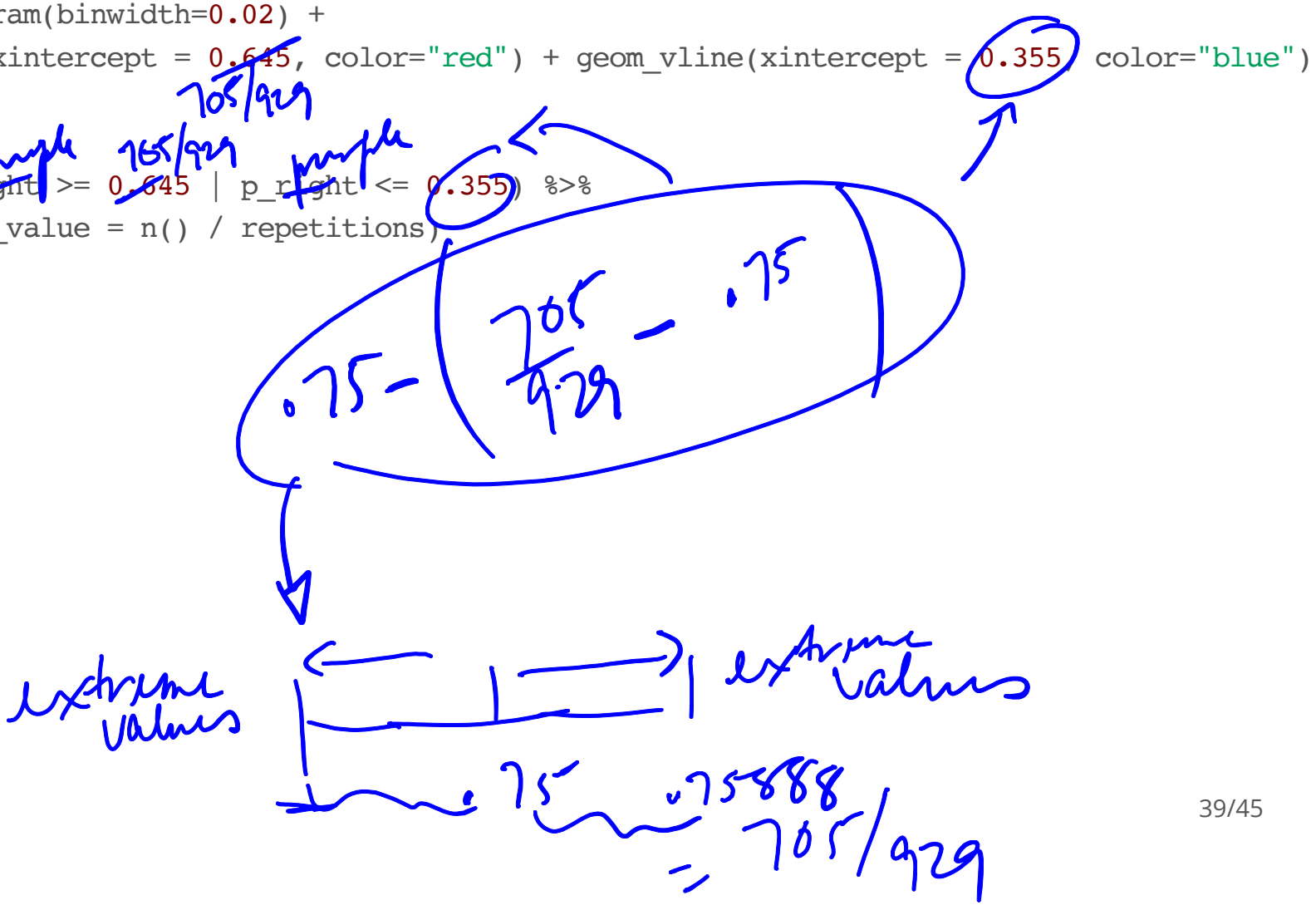
Simulate many values of what we'd observe if the null hypothesis were true

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

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?

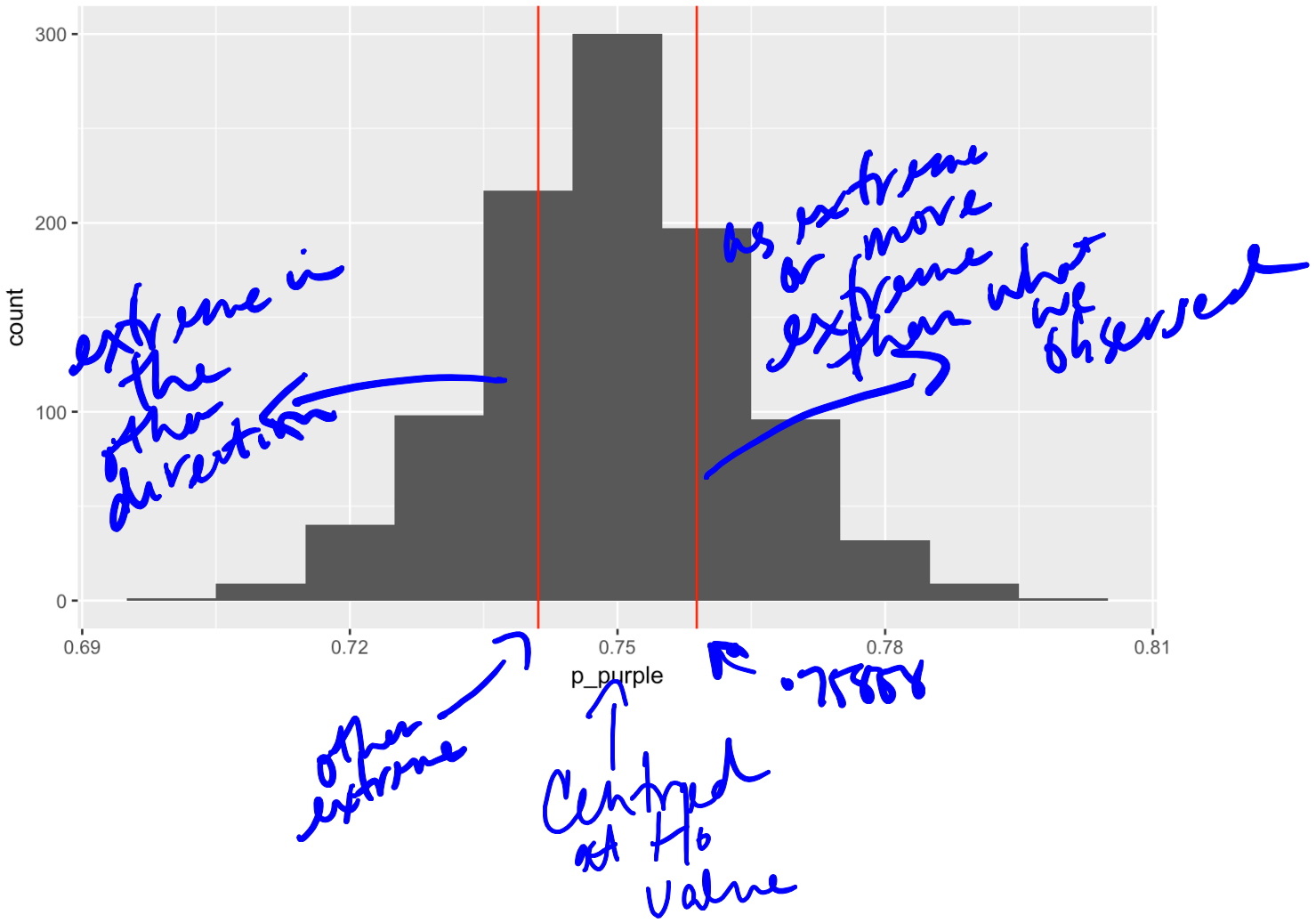
```
sim <- data_frame(p_right = simulated_stats)

ggplot(sim, aes(p_right)) +
  geom_histogram(binwidth=0.02) +
  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

```
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)

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) +
  geom_vline(xintercept = test_stat, color="red") + geom_vline(xintercept = other_extreme, colo
```



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

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.524
```

P-value

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 0.52. Therefore we have no evidence against the null hypothesis that the probability that a pea plant will produce a purple flower is 0.75. The observed data are consistent with Mendel's theory.

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

For your next statistics course

[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.