Hypothesis testing

Data Science in a Box

datasciencebox.org

Hypothesis testing for a single proportion

Packages

```
library(tidyverse)
library(tidymodels)
```

Organ donors

People providing an organ for donation sometimes seek the help of a special "medical consultant". These consultants assist the patient in all aspects of the surgery, with the goal of reducing the possibility of complications during the medical procedure and recovery. Patients might choose a consultant based in part on the historical complication rate of the consultant's clients.

One consultant tried to attract patients by noting that the average complication rate for liver donor surgeries in the US is about 10%, but her clients have only had 3 complications in the 62 liver donor surgeries she has facilitated. She claims this is strong evidence that her work meaningfully contributes to reducing complications (and therefore she should be hired!).

Data

```
organ_donor %>%
  count(outcome)
```

```
## # A tibble: 2 × 2
##   outcome            n
##   <chr>          <int>
## 1 complication       3
## 2 no complication   59
```

Parameter vs. statistic

A **parameter** for a hypothesis test is the "true" value of interest. We typically estimate the parameter using a **sample statistic** as a **point estimate**.

p : true rate of complication

$\hat{p}$: rate of complication in the sample = $\frac{3}{62}$ = 0.048

Correlation vs. causation

> Is it possible to assess the consultant's claim using the data?

No. The claim is that there is a causal connection, but the data are observational. For example, maybe patients who can afford a medical consultant can afford better medical care, which can also lead to a lower complication rate.

While it is not possible to assess the causal claim, it is still possible to test for an association using these data. For this question we ask, could the low complication rate of $\hat{p} = 0.048$ be due to chance?

Two claims

- **Null hypothesis:** "There is nothing going on"

Complication rate for this consultant is no different than the US average of 10%

- **Alternative hypothesis:** "There is something going on"

Complication rate for this consultant is **lower** than the US average of 10%

Hypothesis testing as a court trial

- **Null hypothesis**, H_0: Defendant is innocent
- **Alternative hypothesis**, H_A: Defendant is guilty
- **Present the evidence:** Collect data
- **Judge the evidence:** "Could these data plausibly have happened by chance if the null hypothesis were true?"
  - Yes: Fail to reject H_0
  - No: Reject H_0

Hypothesis testing framework

- Start with a null hypothesis, H_0, that represents the status quo
- Set an alternative hypothesis, H_A, that represents the research question, i.e. what we're testing for
- Conduct a hypothesis test under the assumption that the null hypothesis is true and calculate a **p-value** (probability of observed or more extreme outcome given that the null hypothesis is true)
    - if the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, stick with the null hypothesis
    - if they do, then reject the null hypothesis in favor of the alternative

Setting the hypotheses

Which of the following is the correct set of hypotheses?

(a) $H_0 : p = 0.10; H_A : p \neq 0.10$

(b) $H_0 : p = 0.10; H_A : p > 0.10$

(c) $H_0 : p = 0.10; H_A : p < 0.10$

(d) $H_0 : \hat{p} = 0.10; H_A : \hat{p} \neq 0.10$

(e) $H_0 : \hat{p} = 0.10; H_A : \hat{p} > 0.10$

(f) $H_0 : \hat{p} = 0.10; H_A : \hat{p} < 0.10$

Simulating the null distribution

Since $H_0 : p = 0.10$, we need to simulate a null distribution where the probability of success (complication) for each trial (patient) is 0.10.

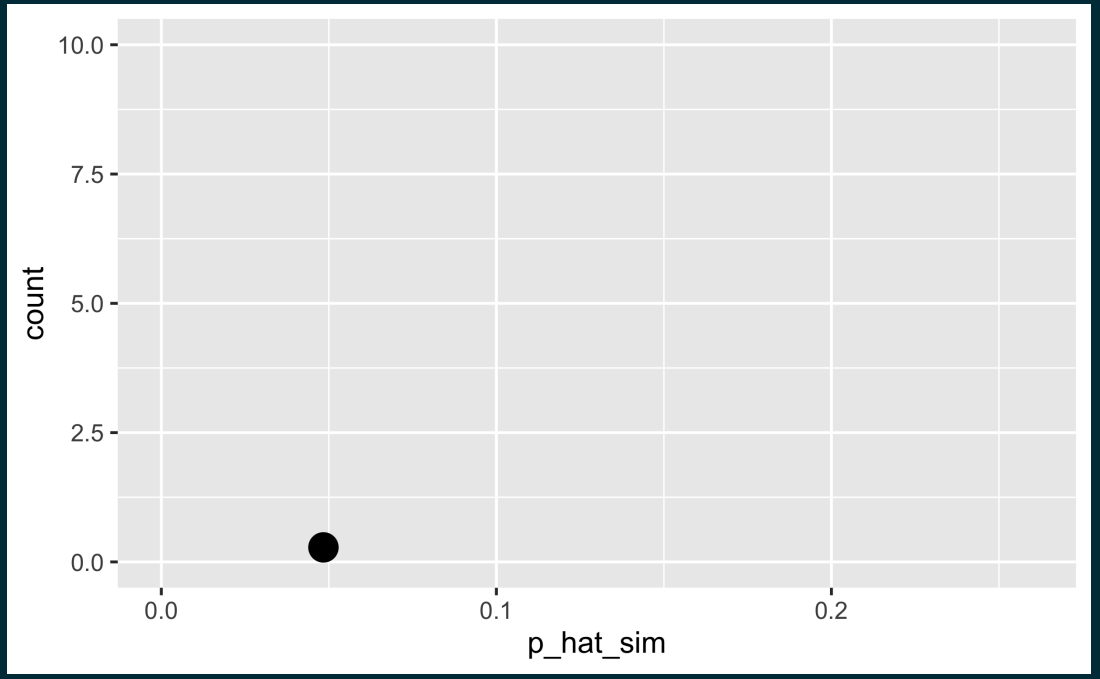
> Describe how you would simulate the null distribution for this study using a bag of chips. How many chips? What colors? What do the colors indicate? How many draws? **With replacement** or **without replacement**?

What do we expect?

> When sampling from the null distribution, what is the expected proportion of success (complications)?

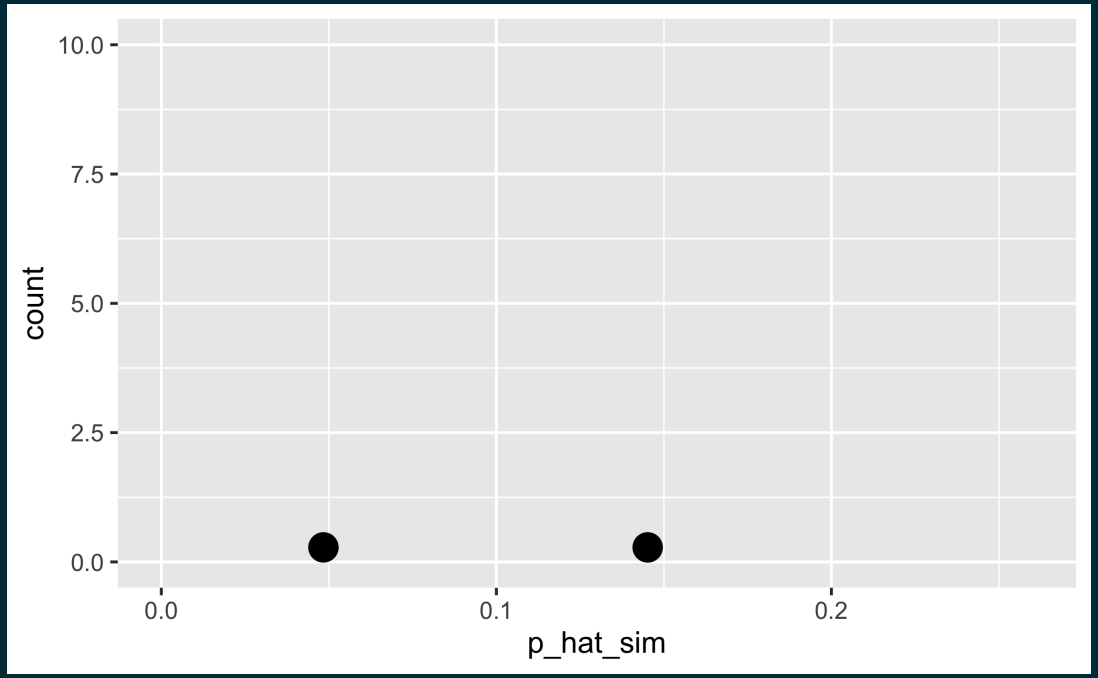
Simulation #1

```
## sim1
##    complication no complication
##               3              59

## [1] 0.0483871
```



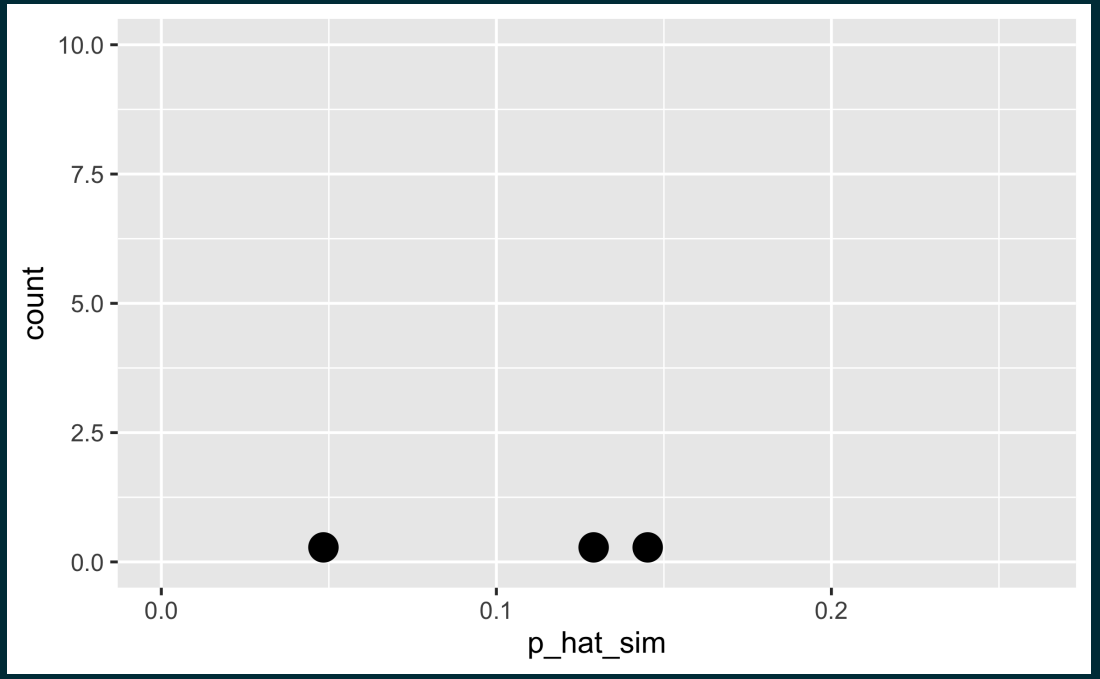
Simulation #2

```
## sim2
##    complication no complication
##               9              53

## [1] 0.1451613
```



Simulation #3

```
## sim3
##    complication no complication
##               8              54

## [1] 0.1290323
```



This is getting boring...

We need a way to automate this process!

Using tidymodels to generate the null distribution

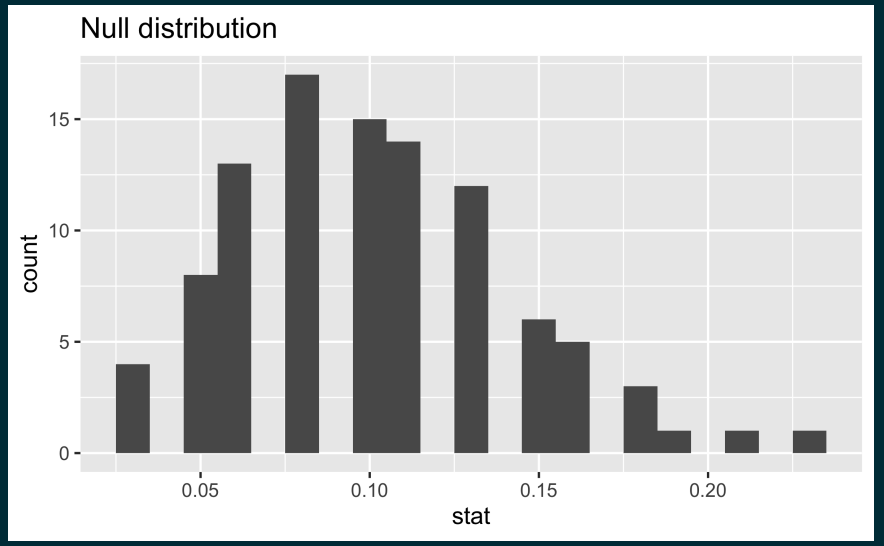
```
null_dist <- organ_donor %>%
  specify(response = outcome, success = "complication") %>%
  hypothesize(null = "point", p = c("complication" = 0.10, "no complication" = 0.90)) %>%
  generate(reps = 100, type = "simulate") %>%
  calculate(stat = "prop")
```

```
## Response: outcome (factor)
## Null Hypothesis: point
## # A tibble: 100 × 2
##   replicate  stat
##       <dbl> <dbl>
## 1         1 0.161
## 2         2 0.081
## 3         3 0.161
## 4         4 0.145
## 5         5 0.097
## 6         6 0.145
## # … with 94 more rows
```

Visualizing the null distribution

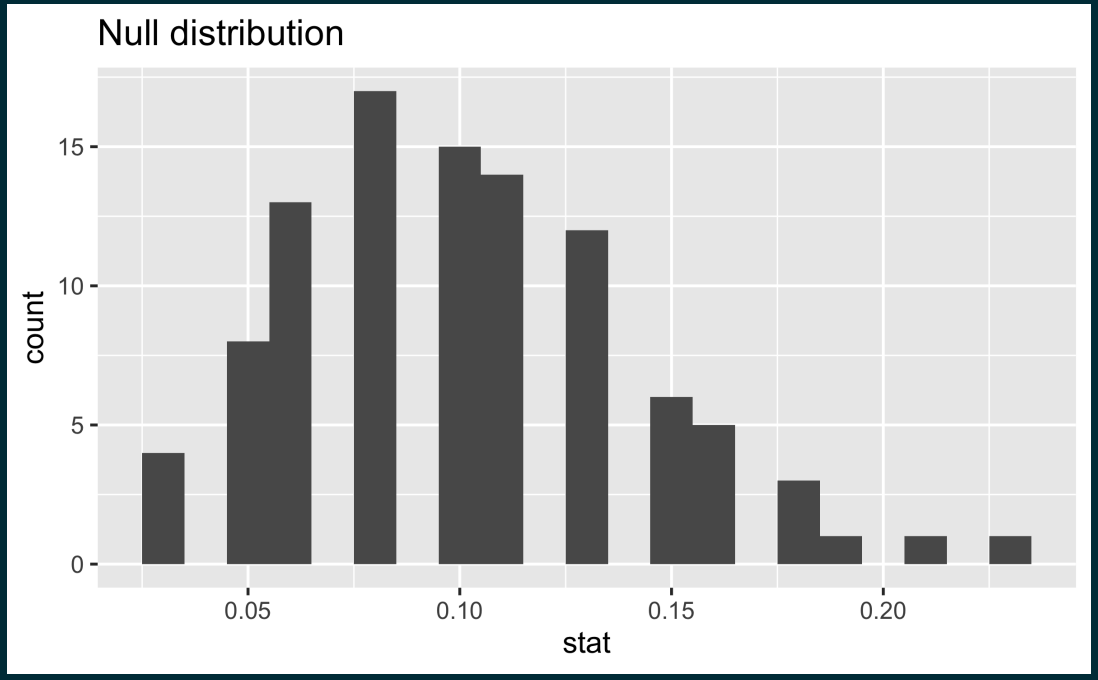
What would you expect the center of the null distribution to be?

```
ggplot(data = null_dist, mapping = aes(x = stat)) +
  geom_histogram(binwidth = 0.01) +
  labs(title = "Null distribution")
```



Calculating the p-value, visually

What is the p-value, i.e. in what % of the simulations was the simulated sample proportion at least as extreme as the observed sample proportion?



Calculating the p-value, directly

```
null_dist %>%
  filter(stat <= (3/62)) %>%
  summarise(p_value = n()/nrow(null_dist))
```

```
## # A tibble: 1 × 1
##   p_value
##     <dbl>
## 1    0.12
```

Significance level

We often use 5% as the cutoff for whether the p-value is low enough that the data are unlikely to have come from the null model. This cutoff value is called the **significance level**, α.

- If p-value < α, reject H_0 in favor of H_A: The data provide convincing evidence for the alternative hypothesis.
- If p-value > α, fail to reject H_0 in favor of H_A: The data do not provide convincing evidence for the alternative hypothesis.

Conclusion

> What is the conclusion of the hypothesis test?

Since the p-value is greater than the significance level, we fail to reject the null hypothesis. These data do not provide convincing evidence that this consultant incurs a lower complication rate than 10% (overall US complication rate).

Let's get real

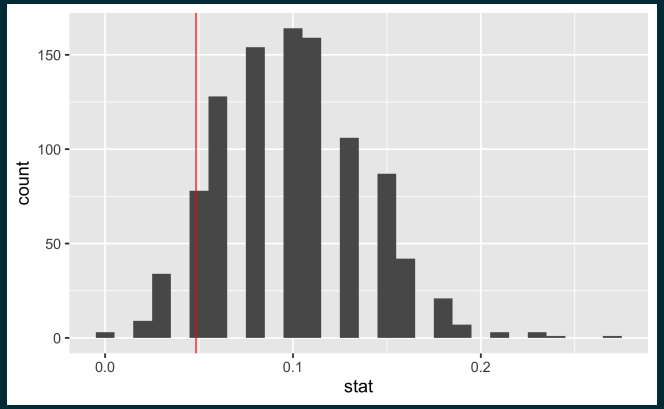
- 100 simulations is not sufficient
- We usually simulate around 15,000 times to get an accurate distribution, but we'll do 1,000 here for efficiency.

Run the test

```
null_dist <- organ_donor %>%
  specify(response = outcome, success = "complication") %>%
  hypothesize(null = "point", p = c("complication" = 0.10, "no complication" = 0.90)) %>%
  generate(reps = 1000, type = "simulate") %>%
  calculate(stat = "prop")
```

Visualize and calculate

```
ggplot(data = null_dist, mapping = aes(x = stat)) +
  geom_histogram(binwidth = 0.01) +
  geom_vline(xintercept = 3/62, color = "red")
```




```
null_dist %>%
  filter(stat <= 3/62) %>%
  summarise(p_value = n()/nrow(null_dist))
```

```
## # A tibble: 1 × 1
##   p_value
##     <dbl>
## 1   0.124
```

One vs. two sided hypothesis tests

Types of alternative hypotheses

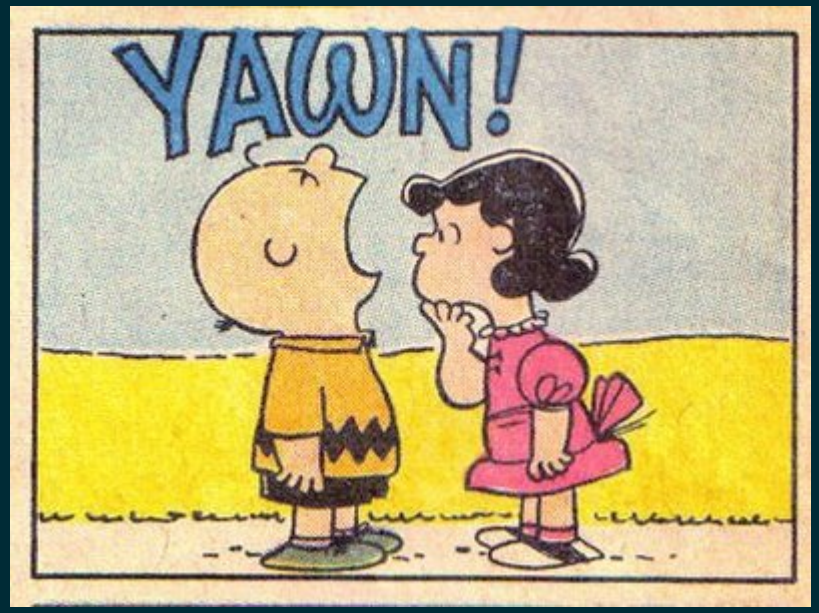
- One sided (one tailed) alternatives: The parameter is hypothesized to be less than or greater than the null value, < or >
- Two sided (two tailed) alternatives: The parameter is hypothesized to be not equal to the null value, $\neq$
  - Calculated as two times the tail area beyond the observed sample statistic
  - More objective, and hence more widely preferred

> Average systolic blood pressure of people with Stage 1 Hypertension is 150 mm Hg. Suppose we want to use a hypothesis test to evaluate whether a new blood pressure medication has **an effect** on the average blood pressure of heart patients. What are the hypotheses?

Testing for independence

Is yawning contagious?

Do you think yawning is contagious?



Is yawning contagious?

An experiment conducted by the MythBusters tested if a person can be subconsciously influenced into yawning if another person near them yawns.

https://www.discovery.com/tv-shows/mythbusters/videos/is-yawning-contagious-2

Study description

In this study 50 people were randomly assigned to two groups: 34 to a group where a person near them yawned (treatment) and 16 to a control group where they didn't see someone yawn (control).

The data are in the **openintro** package: `yawn`

```
yawn %>%
  count(group, result)
```

```
## # A tibble: 4 × 3
##   group result       n
##   <fct> <fct>    <int>
## 1 ctrl  not yawn    12
## 2 ctrl  yawn         4
## 3 trmt  not yawn    24
## 4 trmt  yawn        10
```

Proportion of yawners

```
yawn %>%
  count(group, result) %>%
  group_by(group) %>%
  mutate(p_hat = n / sum(n))
```

```
## # A tibble: 4 × 4
## # Groups:   group [2]
##   group result       n p_hat
##   <fct> <fct>    <int> <dbl>
## 1 ctrl  not yawn    12 0.75
## 2 ctrl  yawn         4 0.25
## 3 trmt  not yawn    24 0.706
## 4 trmt  yawn        10 0.294
```

- Proportion of yawners in the treatment group: $\frac{10}{34} = 0.2941$
- Proportion of yawners in the control group: $\frac{4}{16} = 0.25$
- Difference: $0.2941 - 0.25 = 0.0441$
- Our results match the ones calculated on the MythBusters episode.

Independence?

> Based on the proportions we calculated, do you think yawning is really contagious, i.e. are seeing someone yawn and yawning dependent?

```
## # A tibble: 4 × 4
## # Groups:   group [2]
##   group result       n p_hat
##   <fct> <fct>    <int> <dbl>
## 1 ctrl  not yawn    12 0.75
## 2 ctrl  yawn         4 0.25
## 3 trmt  not yawn    24 0.706
## 4 trmt  yawn        10 0.294
```

Dependence, or another possible explanation?

- The observed differences might suggest that yawning is contagious, i.e. seeing someone yawn and yawning are dependent.
- But the differences are small enough that we might wonder if they might simple be **due to chance**.
- Perhaps if we were to repeat the experiment, we would see slightly different results.
- So we will do just that - well, somewhat - and see what happens.
- Instead of actually conducting the experiment many times, we will **simulate** our results.

Two competing claims

- "There is nothing going on." Yawning and seeing someone yawn are **independent**, yawning is not contagious, observed difference in proportions is simply due to chance. $\rightarrow$ Null hypothesis

- "There is something going on." Yawning and seeing someone yawn are **dependent**, yawning is contagious, observed difference in proportions is not due to chance. $\rightarrow$ Alternative hypothesis

Simulation setup

1. A regular deck of cards is comprised of 52 cards: 4 aces, 4 of numbers 2-10, 4 jacks, 4 queens, and 4 kings.

2. Take out two aces from the deck of cards and set them aside.

3. The remaining 50 playing cards to represent each participant in the study:

    - 14 face cards (including the 2 aces) represent the people who yawn.
    - 36 non-face cards represent the people who don't yawn.

Running the simulation

1. Shuffle the 50 cards at least 7 times[1] to ensure that the cards counted out are from a random process.

2. Count out the top 16 cards and set them aside. These cards represent the people in the control group.

3. Out of the remaining 34 cards (treatment group) count the \red{number of face cards} (the number of people who yawned in the treatment group).

4. Calculate the difference in proportions of yawners (treatment - control), and plot it on the board.

5. Mark the difference you find on the dot plot on the board.

[1] http://www.dartmouth.edu/~chance/course/topics/winning_number.html

Simulation by hand

> Do the simulation results suggest that yawning is contagious, i.e. does seeing someone yawn and yawning appear to be dependent?

Simulation by computation

```
null_dist <- yawn %>%
  specify(response = result, explanatory = group,
          success = "yawn") %>%
  hypothesize(null = "independence") %>%
  generate(100, type = "permute") %>%
  calculate(stat = "diff in props",
            order = c("trmt", "ctrl"))
```

Simulation by computation - 1

- Start with the data frame
- **Specify the variables**
  - **Since the response variable is categorical, specify the level which should be considered as "success"**

```
yawn %>%
{{  specify(response = result, explanatory = group,
          success = "yawn") }}
```

Simulation by computation - 2

- Start with the data frame
- Specify the variables
  - Since the response variable is categorical, specify the level which should be considered as "success"
- **State the null hypothesis (yawning and whether or not you see someone yawn are independent)**

```
yawn %>%
  specify(response = result, explanatory = group,
          success = "yawn") %>%
  hypothesize(null = "independence")
```

Simulation by computation - 3

- Start with the data frame
- Specify the variables
  - Since the response variable is categorical, specify the level which should be considered as "success"
- State the null hypothesis (yawning and whether or not you see someone yawn are independent)
- **Generate simulated differences via permutation**

```
yawn %>%
  specify(response = result, explanatory = group,
          success = "yawn") %>%
  hypothesize(null = "independence") %>%
  generate(100, type = "permute")
```

Simulation by computation - 4

- Start with the data frame
- Specify the variables
  - Since the response variable is categorical, specify the level which should be considered as "success"
- State the null hypothesis (yawning and whether or not you see someone yawn are independent)
- Generate simulated differences via permutation
- **Calculate the sample statistic of interest (difference in propotions)**
  - **Since the explanatory variable is categorical, specify the order in which the subtraction should occur for the calculation of the sample statistic, $(\hat{p}_{treatment} - \hat{p}_{control})$.**

```
yawn %>%
  specify(response = result, explanatory = group,
          success = "yawn") %>%
  hypothesize(null = "independence") %>%
  generate(100, type = "permute") %>%
{{ calculate(stat = "diff in props",
           order = c("trmt", "ctrl")) }}
```

Simulation by computation - 0

- **Save the result**
- Start with the data frame
- Specify the variables
  - Since the response variable is categorical, specify the level which should be considered as "success"
- State the null hypothesis (yawning and whether or not you see someone yawn are independent)
- Generate simulated differences via permutation
- Calculate the sample statistic of interest (difference in propotions)
  - Since the explanatory variable is categorical, specify the order in which the subtraction should occur for the calculation of the sample statistic, $(\hat{p}_{treatment} - \hat{p}_{control})$.

```
null_dist <- yawn %>%
  specify(response = outcome, explanatory = group,
          success = "yawn") %>%
  hypothesize(null = "independence") %>%
  generate(100, type = "permute") %>%
  calculate(stat = "diff in props",
            order = c("treatment", "control"))
```

Visualizing the null distribution

> What would you expect the center of the null distribution to be?

```
ggplot(data = null_dist, mapping = aes(x = stat)) +
  geom_histogram(binwidth = 0.05) +
  labs(title = "Null distribution")
```

Calculating the p-value, visually

What is the p-value, i.e. in what % of the simulations was the simulated difference in sample proportion at least as extreme as the observed difference in sample proportions?

Calculating the p-value, directly

```
null_dist %>%
  filter(stat >= 0.0441) %>%
  summarise(p_value = n()/nrow(null_dist))
```

```
## # A tibble: 1 × 1
##   p_value
##     <dbl>
## 1    0.53
```

Conclusion

What is the conclusion of the hypothesis test?

Do you "buy" this conclusion?