SciencesPo

DEPARTMENT OF ECONOMICS

Applied Data Analysis for Public Policy Studies

Hypothesis Testing

Michele Fioretti
SciencesPo Paris
2020-08-25

Packages used in this set of slides

```r
library(tidyverse)
library(infer)
library(moderndive)
```

Is There Gender Discrimination In Promotions?

- An article published in the *Journal of Applied Psychology* in 1974 investigates whether female employees at Banks are discriminated against.

- 48 supervisors were given *identical* candidate CVs - identical up to the first name, which was male or female.

- Many similar experiments have been conducted with other groups. Arabic Names, Black names, Jewish names or other groups that can be identified from typical name choice. [1], [2], [3], …

Is There Gender Discrimination In Promotions?

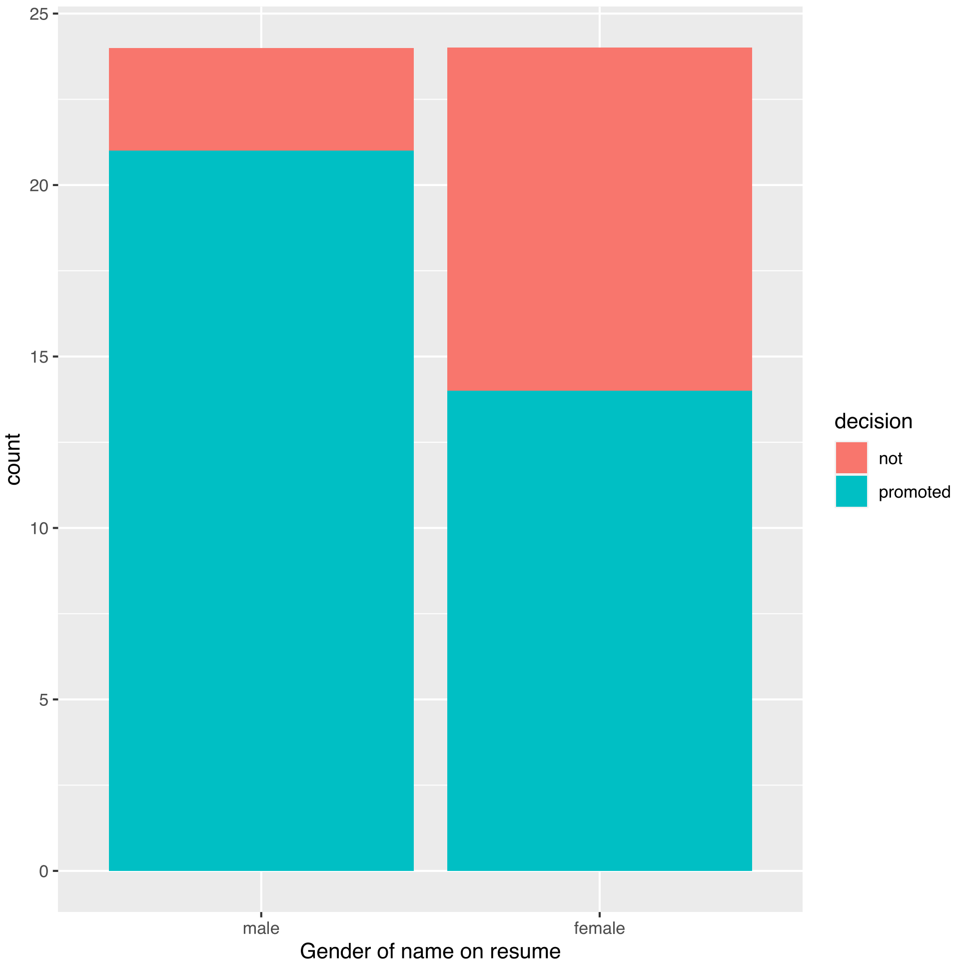
- An article published in the *Journal of Applied Psychology* in 1974 investigates whether female employees at Banks are discriminated against.
- 48 supervisors were given *identical* candidate CVs - identical up to the first name, which was male or female.
- Many similar experiments have been conducted with other groups. Arabic Names, Black names, Jewish names or other groups that can be identified from typical name choice. [1], [2], [3], ...

```r
library(moderndive)
promotions
```

```
## # A tibble: 48 x 3
##       id decision gender
##    <int> <fct>    <fct>
##  1     1 promoted male
##  2     2 promoted male
##  3     3 promoted male
##  4     4 promoted male
##  5     5 promoted male
##  6     6 promoted male
##  7     7 promoted male
##  8     8 promoted male
##  9     9 promoted male
## 10    10 promoted male
## # … with 38 more rows
```

```r
# for info on the `dataset`
?promotions
# Data from a 1970's study on whether gender
# influences hiring recommendations [...]
```

Looking At Promotions




```
promotions %>%
  group_by(gender, decision) %>%
  summarize(n = n()) %>%
  mutate(proportion = n / sum(n))
```

```
## # A tibble: 4 x 4
## # Groups:   gender [2]
##   gender decision     n proportion
##   <fct>  <fct>    <int>      <dbl>
## 1 male   not          3      0.125
## 2 male   promoted    21      0.875
## 3 female not         10      0.417
## 4 female promoted    14      0.583
```

- 87.5% of "men" were promoted.
- 58.3% of "women" were promoted.
- That's a difference of 87.55 - 58.3% = 29.2%.
- Is the 29% advantage for men in this sample **conclusive evidence**?
- In a *hyoptethical world* **without gender discrimination**, could we have observed a 29% difference *by chance*?

Imposing A Hypothetical World: No Gender Discriminiation

- Suppose we lived in a world without gender discrimination.
- The label `gender` in our dataframe would be meaningless.
- Let's randomly reassign `gender` to each row and see how this affects the result.
- Suppose we have 48 playing cards: 24 red (female) and 24 (black)
- Shuffle the cards, and lay down the cards in a row, record `f` if **red**.

Imposing A Hypothetical World: No Gender Discriminiation

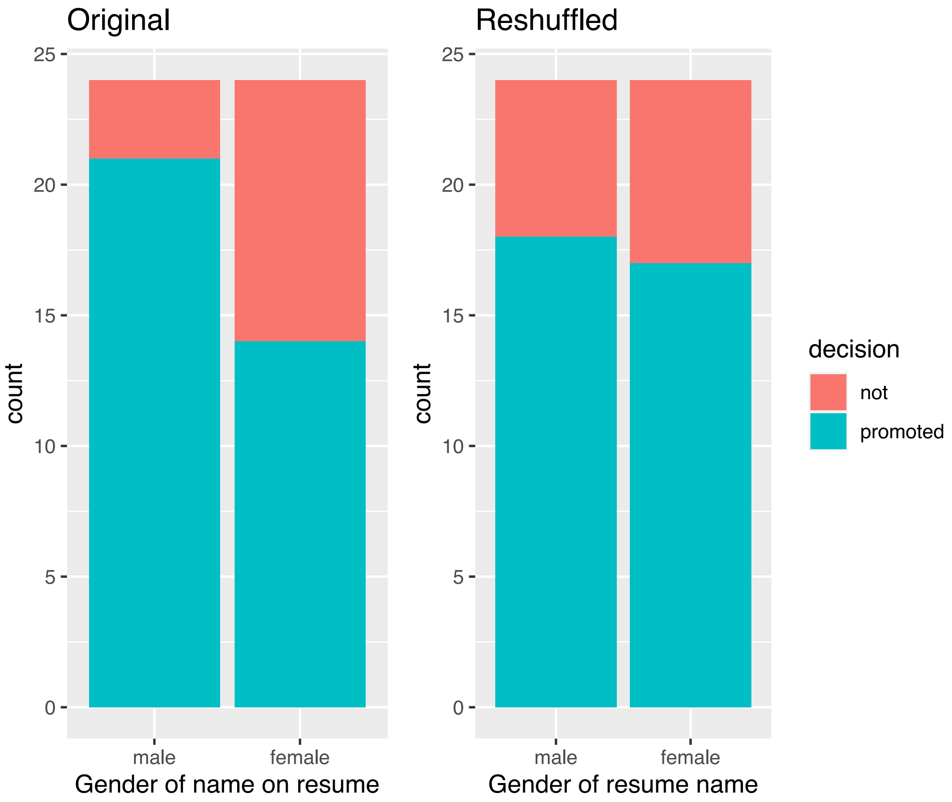
- Suppose we lived in a world without gender discrimination.
- The label `gender` in our dataframe would be meaningless.
- Let's randomly reassign `gender` to each row and see how this affects the result.
- Suppose we have 48 playing cards: 24 red (female) and 24 (black)
- Shuffle the cards, and lay down the cards in a row, record `f` if **red**.

```
bind_cols(promotions, promotions_shuffled) #?promotio
```

```
## # A tibble: 48 x 6
##    id...1 decision...2 gender...3 id...4 decision...5 gender..
##     <int> <fct>        <fct>       <int> <fct>        <fct>
##  1      1 promoted     male            1 promoted     female
##  2      2 promoted     male            2 promoted     female
##  3      3 promoted     male            3 promoted     male
##  4      4 promoted     male            4 promoted     female
##  5      5 promoted     male            5 promoted     male
##  6      6 promoted     male            6 promoted     male
##  7      7 promoted     male            7 promoted     male
##  8      8 promoted     male            8 promoted     female
##  9      9 promoted     male            9 promoted     male
## 10     10 promoted     male           10 promoted     female
## # … with 38 more rows
```

- Observe how in `promotions_shuffled` we randomly assigned `gender1`.
- The `decision` column is the same!
- What does this now look like?

Reshuffled Promotions




```
promotions %>%
  group_by(gender, decision) %>%
  summarize(n = n()) %>%
  mutate(proportion = n / sum(n))
```

```
## # A tibble: 4 x 4
## # Groups:   gender [2]
##   gender decision     n proportion
##   <fct>  <fct>    <int>      <dbl>
## 1 male   not          3      0.125
## 2 male   promoted    21      0.875
## 3 female not         10      0.417
## 4 female promoted    14      0.583
```

```
promotions_shuffled %>%
  group_by(gender, decision) %>%
  summarize(n = n()) %>%
  mutate(proportion = n / sum(n))
```

```
## # A tibble: 4 x 4
## # Groups:   gender [2]
##   gender decision     n proportion
##   <fct>  <fct>    <int>      <dbl>
## 1 male   not          6      0.25
## 2 male   promoted    18      0.75
## 3 female not          7      0.292
## 4 female promoted    17      0.708
```

Sampling Variation?

- In the hypothetical world, the difference was only 4.2%.
- But what's the role of ***sampling variation***? How representative of that hypothetical world is 4.2%?
- Let's construct the sampling distribution ourselves!

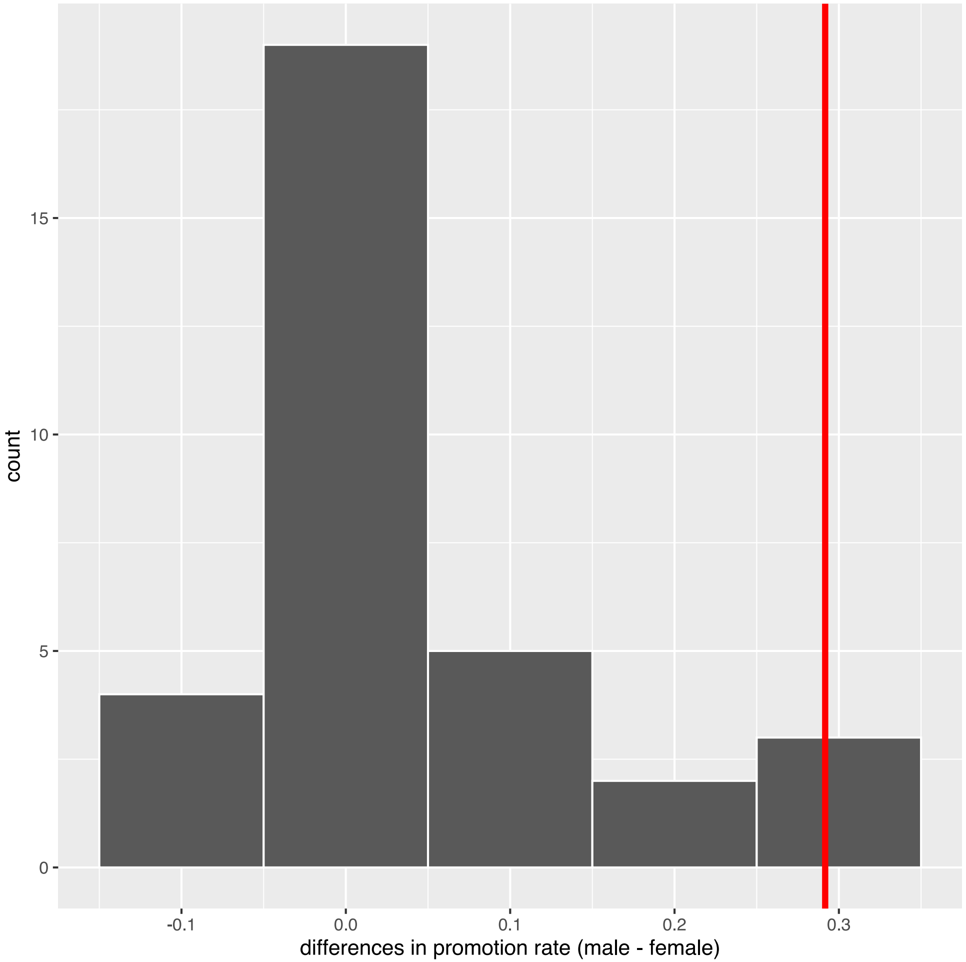
Sampling Variation?

- In the hypothetical world, the difference was only 4.2%.
- But what's the role of ***sampling variation***? How representative of that hypothetical world is 4.2%?
- Let's construct the sampling distribution ourselves!

1. You need to shuffle a deck of 48 cards, 24 red, 24 black, and lay out card after card in front of you.
2. You do **not** put the cards back into the deck!
3. You could use the function `sample` for example. Look at `?sample` to find out more.
4. fill in your results into this shared spreadsheet!

```
sample(promotions$gender, replace = FALSE) # Note: create a new google csv file
```

Sampling Variation in Reshuffling



- This distribution was created in our **hypothetical** scenario: no discrimination.
- We see how sampling variation affects the difference in promotion rates.
- The red line denotes the *observed difference* in the **real world** (29.2%).
- Now: How *likely* is it that the red line is part of this **hypothetical** distribution?

Recap: Permutation vs Bootstrap

- We just did a **permutation test**. We randomly reshuffled and checked if it makes a difference.
- Again Resampling: boostrapping is **with** replacment, permutation is **without**.
- Bootstrapping: we put the paper slips **back** after recording them.
- Permutation: we took card after card from our deck (*without* putting it back!)

Recap: Permutation vs Bootstrap

- We just did a **permutation test**. We randomly reshuffled and checked if it makes a difference.
- Again Resampling: boostrapping is **with** replacment, permutation is **without**.
- Bootstrapping: we put the paper slips **back** after recording them.
- Permutation: we took card after card from our deck (*without* putting it back!)

- We observed the estimate $\hat{p}_m - \hat{p}_f = 29.2\%$ in the real world.
- We *tested* whether in a hypothetical universe with no discrimination, 29.2% *likely* to occur.
- We concluded *rather not*. We tended to **reject** that hypothesis.
- The real question was: is 29.2% **really** different from zero? What is the role of sampling variation?

Hypothesis Testing Setup

Hypothesis Test Notation and Definitions

- In Hypothesis testing we compare two **competing hypothesis**.
  - In our example:

$$
\begin{aligned}
H_0 : & p_m - p_f = 0 \\
H_A : & p_m - p_f > 0
\end{aligned}
$$

  - H_0 stands for the **null hypothesis**, where *no effect* is observed. That's our hypothetical world from above.
- H_A or H_1 is the **alternative** hypothesis. Here, we have a *one-sided* alternative, saying that $p_m > p_f$, ie women are discriminated against. The *two-sided* formulation is just $H_A : p_m - p_f \neq 0$

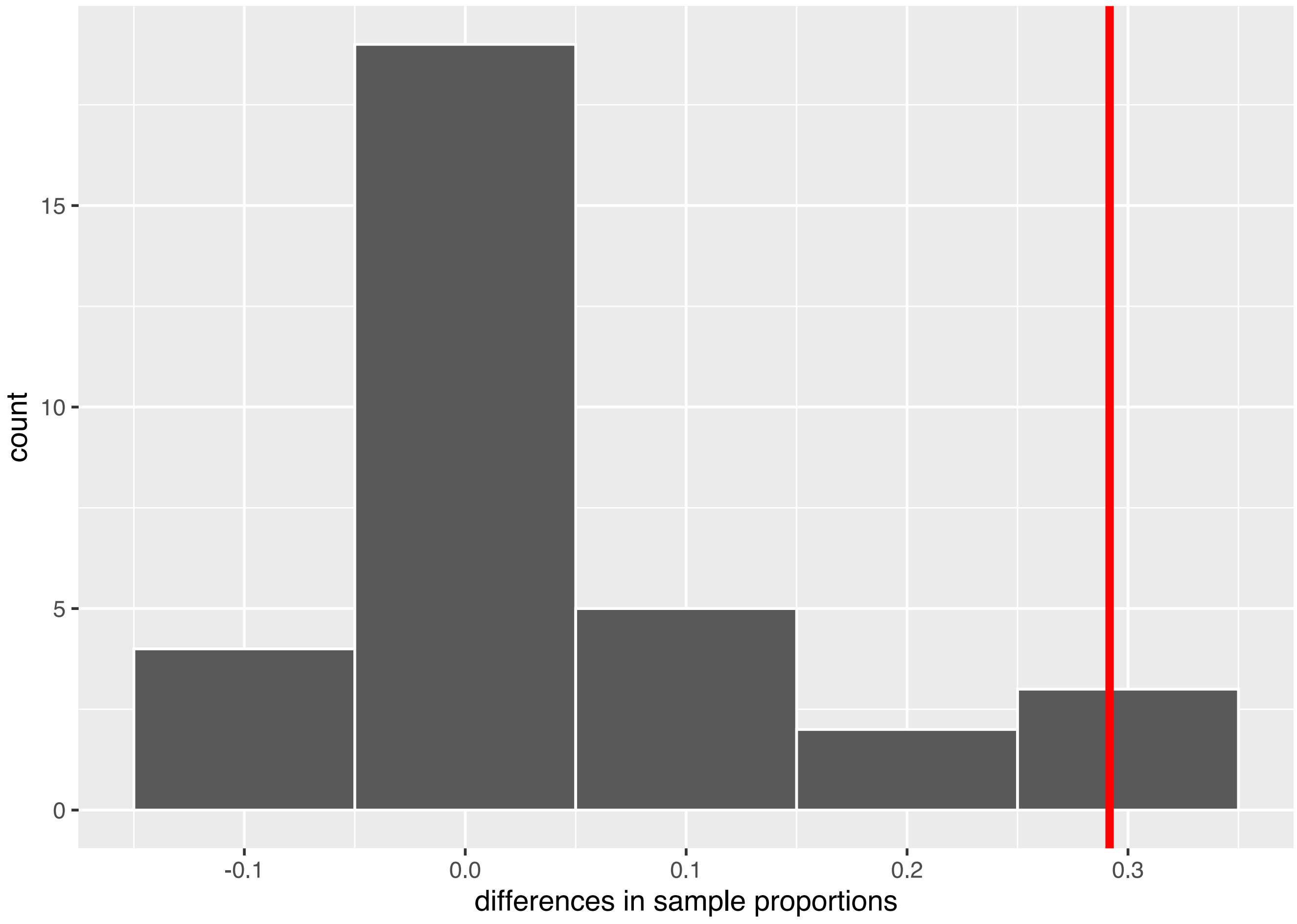
Hypothesis Test Notation and Definitions

- In Hypothesis testing we compare two **competing hypothesis**.
    - In our example:

$$
\begin{aligned}
H_0 &: p_m - p_f = 0 \\
H_A &: p_m - p_f > 0
\end{aligned}
$$

    - H_0 stands for the **null hypothesis**, where *no effect* is observed. That's our hypothetical world from above.
- H_A or H_1 is the **alternative** hypothesis. Here, we have a *one-sided* alternative, saying that $p_m > p_f$, ie women are discriminated against. The *two-sided* formulation is just $H_A : p_m - p_f \neq 0$
- A **test statistic** is a summary statistic which we use to summarise a certain aspect of our sample. Here: $\hat{p}_m - \hat{p}_f$
- The *observed test statistic* is the number we get from our real world sample: $\hat{p}_m - \hat{p}_f = 29\%$
- The **null distribution** is the sampling distribution of our test statistic, assuming the Null hypothesis is **true**. That's our hypothetical world without discrimination.
- We have seen such a null distribution just above:

Null Distribution



- This **is** the sampling distribution of $\hat{p}_m - \hat{p}_f$, assuming H_0 is true.
- The red line is the *observed* test statistic.

P-Value and Significance Level α

- The **p-value** is the probability of observing a test statistic *more extreme* than the one we obtained, assuming H_0 is true. 🤔

- How *strong* a piece of evidence is it to observe $\hat{p}_m - \hat{p}_f = 29\%$ in a world where $p_m - p_f = 0$ is assumed true? Very strong? Not so strong?

- How many samples did we obtain that had a difference *greater* than 29%? Many, or not so many?

- The p-value quantifies this by measuring the probability to the right of the red line in the previous plot.

P-Value and Significance Level α

- The **p-value** is the probability of observing a test statistic *more extreme* than the one we obtained, assuming H_0 is true. 🤔

- How *strong* a piece of evidence is it to observe $\hat{p}_m - \hat{p}_f = 29\%$ in a world where $p_m - p_f = 0$ is assumed true? Very strong? Not so strong?

- How many samples did we obtain that had a difference *greater* than 29%? Many, or not so many?

- The p-value quantifies this by measuring the probability to the right of the red line in the previous plot.

- The **significance level** α is a *cutoff* on the p-value.

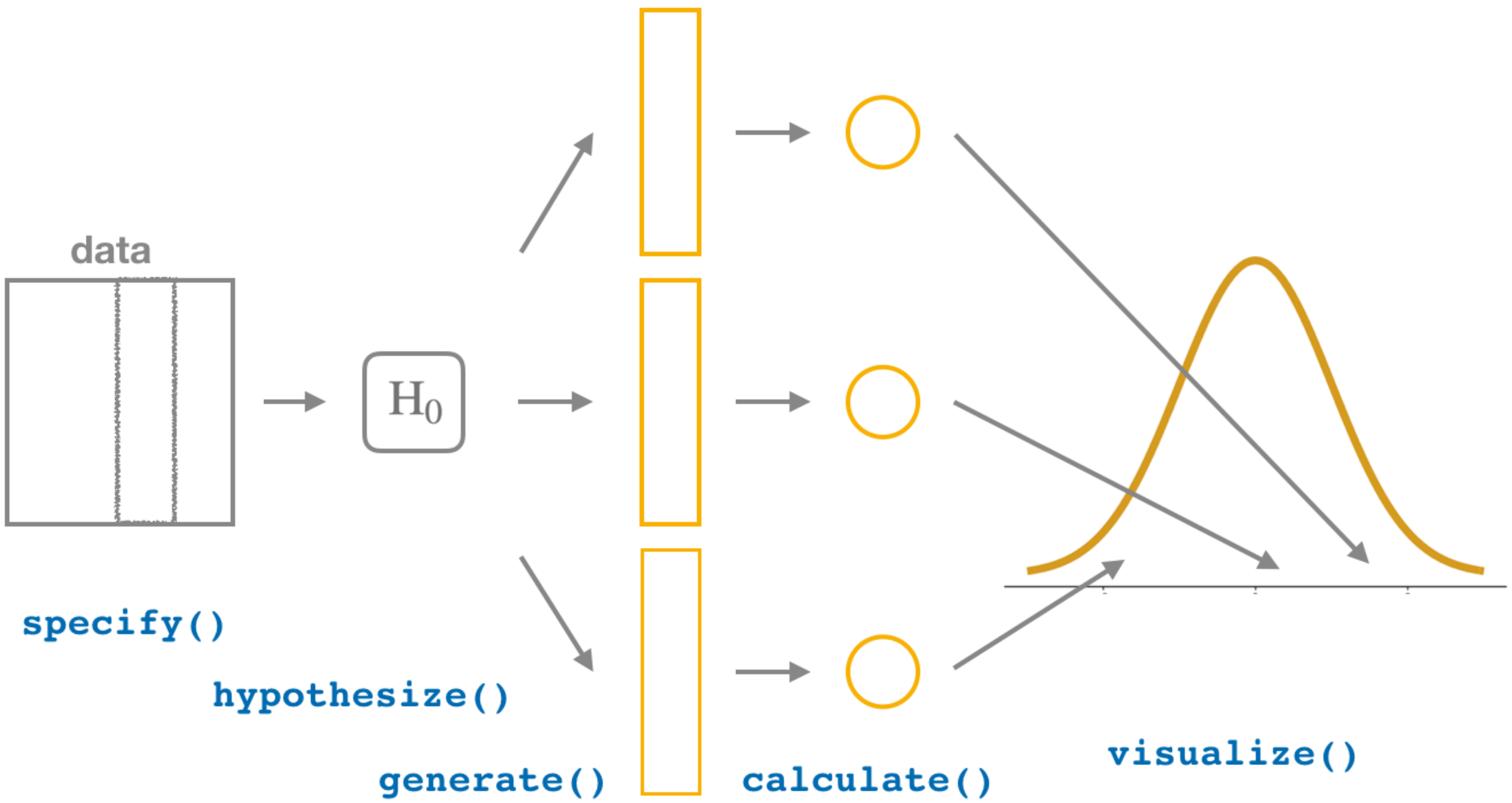
- We choose it *before* conducting our hypothesis test. It's common to assume $\alpha = 5\%$.

- If the p-value falls below the cutoff α, we **reject** the null hypothesis on the grounds that *what we observe is too unlikely to happen* under the Null.

- Small p-value: The red line is *too far* from the center of the Null distribution. Observing the red line would have happened with very small probability only.

Conducting Hypothesis Tests

Testing with `infer`



`infer` Testing Pipeline

- Here we follow closely the infer workflow given in moderndive.
- We augment our previous pipeline with the `hypothesize` function, defining the type of null hypothesis.
- Also, we give a `formula` to `specify()` this time, instead of only a variable name as before.
- We create the Null Distribution by *reshuffling* (deck of cards), and *not* by *resampling* (pennies).

`infer` Testing Pipeline

- Here we follow closely the infer workflow given in moderndive.
- We augment our previous pipeline with the `hypothesize` function, defining the type of null hypothesis.
- Also, we give a `formula` to `specify()` this time, instead of only a variable name as before.
- We create the Null Distribution by *reshuffling* (deck of cards), and *not* by *resampling* (pennies).

```
null_distribution <- promotions %>%
  # takes formula, defines success
  specify(formula = decision ~ gender,
          success = "promoted") %>%
  # decisions are independent of gender
  hypothesize(null = "independence") %>%
  # generate 1000 reshufflings of data
  generate(reps = 1000, type = "permute") %>%
  # compute p_m - p_f from each reshuffle
  calculate(stat = "diff in props",
            order = c("male", "female"))
null_distribution
```

```
## # A tibble: 1,000 x 2
##    replicate    stat
##        <int>   <dbl>
##  1         1 -0.0417
##  2         2 -0.0417
##  3         3  0.0417
##  4         4 -0.208
##  5         5 -0.208
##  6         6  0.0417
##  7         7 -0.125
##  8         8  0.208
##  9         9  0.125
## 10        10 -0.0417
## # … with 990 more rows
```

Back to Reality: What did we *Observe*?

- We computed $\hat{p}_m - \hat{p}_f$ from our *real-world* sample before.

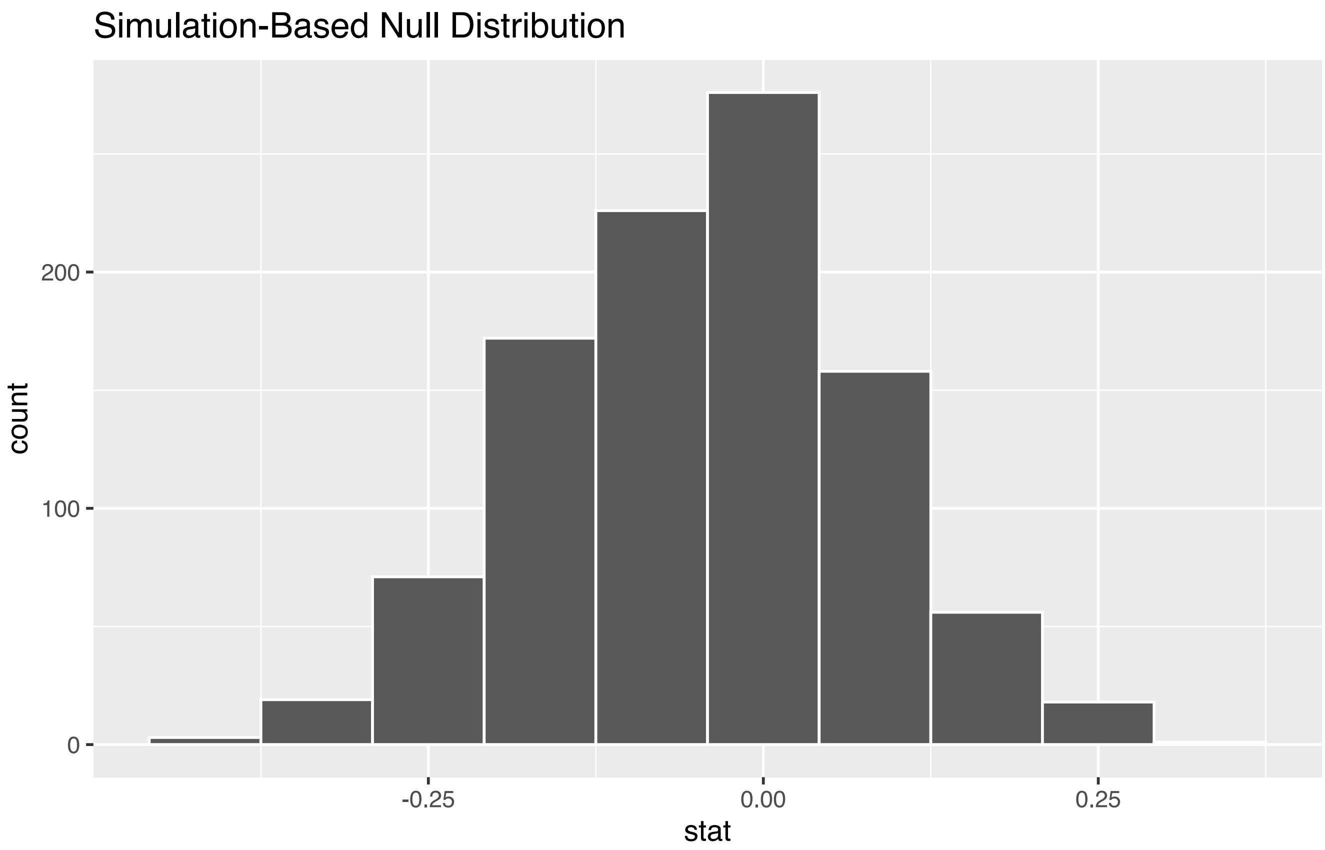
```
obs_diff_prop <- promotions %>%
  specify(decision ~ gender, success = "promoted") %>%
  calculate(stat = "diff in props", order = c("male",
obs_diff_prop
```

```
## # A tibble: 1 x 1
##    stat
##   <dbl>
## 1 0.292
```

- How does that observed statistic compare the distribution of **this** test statistic, assuming that H_0 is true?
- We **created** that distribution on the previous slide: `null_distribution`.
- Let's confront `null_distribution` with `obs_diff_prop`, and let's compute the p-value!

Visualize the Null

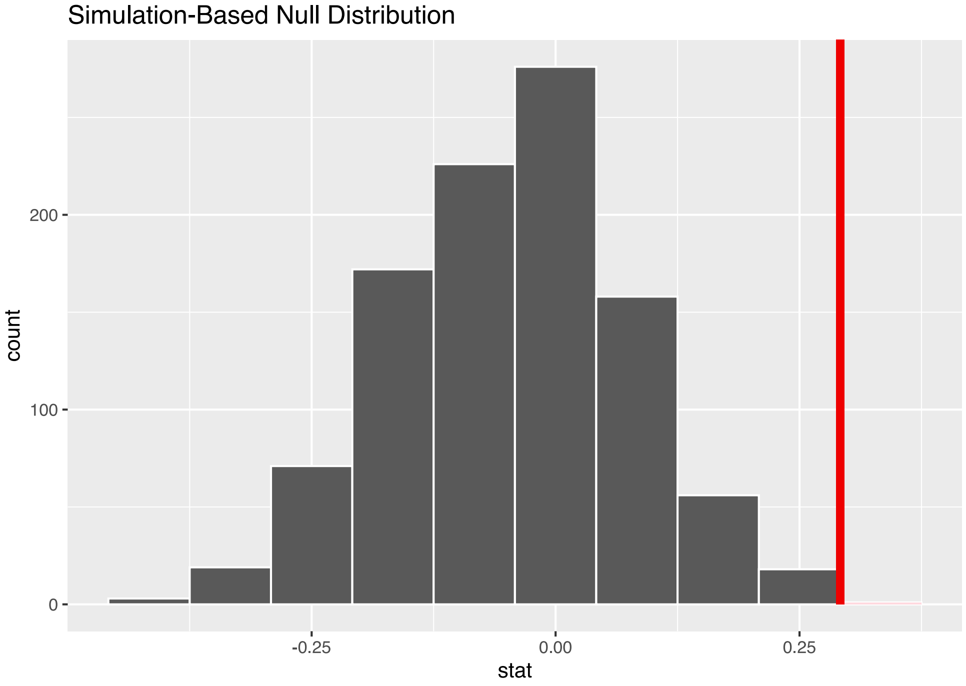
```
visualize(null_distribution, bins = 10)
```



- This is the distribution of $\hat{p}_m - \hat{p}_f$ under H_0 .
- No Discrimination in that world.

Visualize the P-value

```
visualize(null_distribution, bins = 10) +
  shade_p_value(obs_stat = obs_diff_prop,
                direction = "right")
```



- `shade_p_value` adds the p-value based on `obs_diff_prop`, i.e 0.29.
- `direction = "right"` represents our one-sided alternative $H_A: p_m - p_f > 0$
- *more extreme* means *bigger difference* here, hence *more to the right*.
- If $H_A: p_m - p_f < 0$, we'd set `direction = "left"`
- The red area **is the p-value**!
- Is that a *big* or a *small* area?

Obtaining the p-value and Deciding to Reject

- Obtain the precise p-value with

```
p_value <- null_distribution %>%
  get_p_value(obs_stat = obs_diff_prop, direction
p_value
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.019
```

- So, the probability of observing a 29% difference in a world with no discrimination is only 1.9%. That probability is due to sampling variation.

Obtaining the p-value and Deciding to Reject

- Obtain the precise p-value with

```
p_value <- null_distribution %>%
  get_p_value(obs_stat = obs_diff_prop, direction
p_value
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.019
```

- So, the probability of observing a 29% difference in a world with no discrimination is only 1.9%. That probability is due to sampling variation.

- Suppose we had set $\alpha = 0.001 = 0.1\%$

- Given that the p-value is *greater* than α,

  - i.e. 1.9% > 0.1%,
  - we would **fail to reject** the null $H_0 : p_m - p_f = 0$.

- The p-value was not sufficiently small to convince us in this case.

- What would have happened, had we set cutoff $\alpha = 0.05 = 5\%$ instead?

Testing Errors

- Working with probabilities implies that sometimes, we make an error.
- 29% may be *unlikely* under H_0, but that doesn't mean it's *impossible* to occur.
- So, it may happen that we sometimes reject H_0, when in fact it was true.

Testing Errors

- Working with probabilities implies that sometimes, we make an error.
- 29% may be *unlikely* under H_0, but that doesn't mean it's *impossible* to occur.
- So, it may happen that we sometimes reject H_0, when in fact it was true.
- This is similar to a verdict reach in a court trial:

	Truly not guilty	Truly guilty
Verdict		
Not guilty verdict	Correct	Type II error
Guilty verdict	Type I error	Correct

- In fact, in hypothesis testing:

	H0 true	HA true
Verdict		
Fail to reject H0	Correct	Type II error
Reject H0	Type I error	Correct

Type I and Type II Errors

- So, there are even two types of errors to make! 😲

- Type I: We convict an innocent person. We Reject a *true* Null.

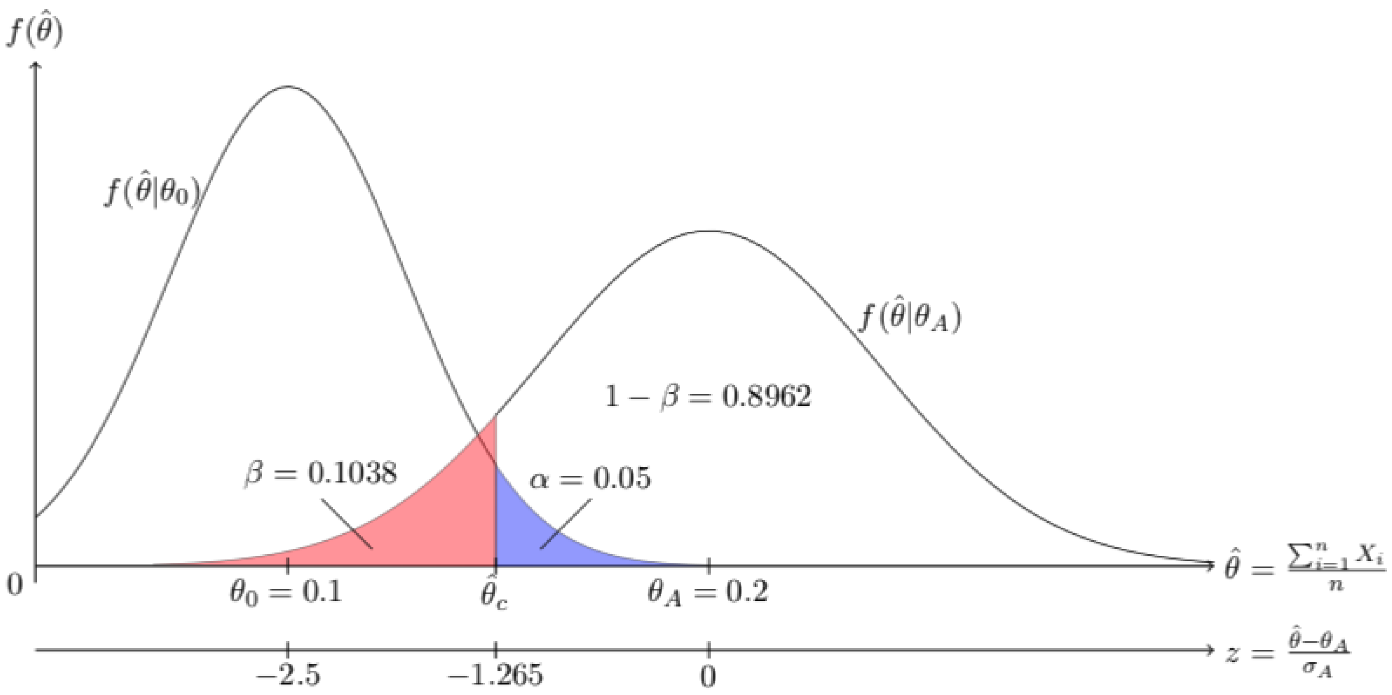
- Type II: We *fail* to convict a criminal. We *fail* to reject a *wrong* Null.

- We **choose** the frequency of a Type I error by setting α, called the **significance level**.

Type I and Type II Errors

- So, there are even two types of errors to make! 😲
- Type I: We convict an innocent person. We Reject a *true* Null.
- Type II: We *fail* to convict a criminal. We *fail* to reject a *wrong* Null.
- We **choose** the frequency of a Type I error by setting α, called the **significance level**.
- The probability of committing a type II error is called β. The value $1-\beta$, i.e. the prob. of *not* making such an error, is called the **power** of a hypothesis test.
- Ideally, $\alpha = \beta = 0$. However, with random sampling this is impossible. Also, both errors are inversely related. (see next slide)
- So, typically we fix α and try to maximize the power of the test.
- Given a certain frequency of convicting an innocent person, we try to make sure we convict as many true criminals as possible.

Type I and II Errors are Inversely related



- $\hat{\theta}$ is *some* test statistic.
- $f(\hat{\theta}|\theta_0)$ and $f(\hat{\theta}|\theta_A)$ are Null and Alternative distributions.
- Changing α moves critical value $\hat{\theta}_c$.
- This example is fully worked out here by Florian Oswald

THANKS

To the amazing moderndive team!

SciencesPo
DEPARTMENT OF ECONOMICS

END

	michele.fioretti@sciencespo.fr
<a href="https://gmichelefioretti.github.io/ScPoEconometrics-Slides/>	Slides
	Book
	@ScPoEcon
	@ScPoEcon