

# ScPoEconometrics

## Confidence Intervals and Hypothesis Testing

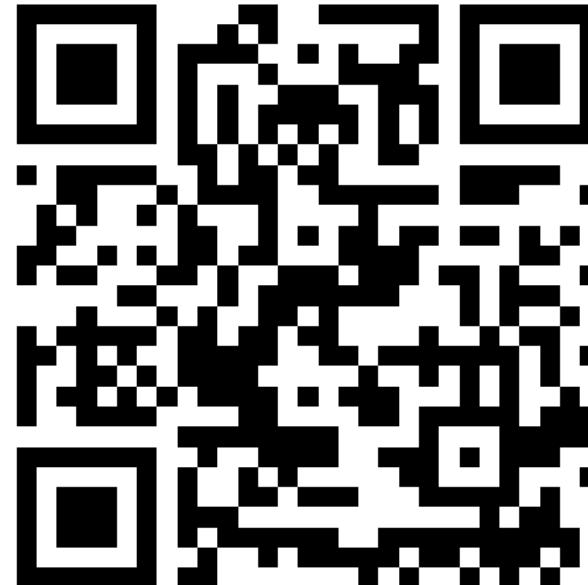
Florian Oswald, Gustave Kenedi and Pierre Villedieu  
SciencesPo Paris  
2021-04-07

# Quick "Quiz" on Last Week's Material

1. From your *computer* ↗ connect to [www.wooclap.com/SCPOSAMP](http://www.wooclap.com/SCPOSAMP)

*OR*

2. From your *phone* ↗ flash QR code below



# Today - Deeper dive into *statistical inference*<sup>1</sup>

- *Confidence intervals*: providing plausible *range* of values
- *Hypothesis testing*: comparing statistics between groups



[1]: This lecture is based on the wonderful [confidence interval](#) and [hypothesis testing](#) chapters of [ModernDive](#)

# Back to reality (there goes gravity 😊)



- In real life we only get to take **one** sample from the population (not **1000**!).
- Also, we obviously don't know the true population parameter, that's what we are interested in!
- So what on earth was all of this good for? Fun only?! 😕



# Back to reality (there goes gravity 😊)



- In real life we only get to take **one** sample from the population (not **1000**!).
- Also, we obviously don't know the true population parameter, that's what we are interested in!
- So what on earth was all of this good for? Fun only?! 😕

- Even unobserved, we **know** that the sampling distribution does exist, and even better, we know how it behaves!
- Let's see what we can do with this...



# Confidence Intervals

# From Point Estimates to Confidence Intervals

- Until now, we have only estimated ***point estimates*** from our samples: *sample means*, *sample proportions*, *regression coefficients*, etc.



# From Point Estimates to Confidence Intervals

- Until now, we have only estimated ***point estimates*** from our samples: *sample means*, *sample proportions*, *regression coefficients*, etc.
- We know that this ***sample statistic*** differs from the ***true population parameter*** due to ***sampling variation***.



# From Point Estimates to Confidence Intervals

- Until now, we have only estimated ***point estimates*** from our samples: *sample means*, *sample proportions*, *regression coefficients*, etc.
- We know that this ***sample statistic*** differs from the ***true population parameter*** due to ***sampling variation***.
- Rather than a point estimate, we could give a ***range of plausible values*** for the population parameter.



# From Point Estimates to Confidence Intervals

- Until now, we have only estimated ***point estimates*** from our samples: *sample means*, *sample proportions*, *regression coefficients*, etc.
- We know that this ***sample statistic*** differs from the ***true population parameter*** due to ***sampling variation***.
- Rather than a point estimate, we could give a ***range of plausible values*** for the population parameter.
- This is precisely what a ***confidence interval*** (CI) provides.



# Constructing Confidence Intervals

There are several approaches to constructing confidence intervals:

1. *Theory*: use mathematical formulas (*Central Limit Theorem*) to derive the sampling distribution of our point estimate under certain conditions → *what R does under the hood!*



# Constructing Confidence Intervals

There are several approaches to constructing confidence intervals:

1. *Theory*: use mathematical formulas (*Central Limit Theorem*) to derive the sampling distribution of our point estimate under certain conditions → *what R does under the hood!*
2. *Simulation*: use the *bootstrapping* method to *reconstruct* the sampling distribution of our point estimate



# Constructing Confidence Intervals

There are several approaches to constructing confidence intervals:

1. *Theory*: use mathematical formulas (*Central Limit Theorem*) to derive the sampling distribution of our point estimate under certain conditions → *what R does under the hood!*
2. *Simulation*: use the *bootstrapping* method to *reconstruct* the sampling distribution of our point estimate

We'll focus on simulation to give you the intuition and come back to the maths approach next week.



# Constructing Confidence Intervals

There are several approaches to constructing confidence intervals:

1. *Theory*: use mathematical formulas (*Central Limit Theorem*) to derive the sampling distribution of our point estimate under certain conditions → *what R does under the hood!*
2. *Simulation*: use the *bootstrapping* method to *reconstruct* the sampling distribution of our point estimate

We'll focus on simulation to give you the intuition and come back to the maths approach next week.

In practice, you *don't* need to compute your confidence intervals using *bootstrap*, R uses statistical theory to do it for you.



# Back to Pasta

- As in real life, imagine we had access to *only one random sample* from our bowl of past.



# Back to Pasta

- As in real life, imagine we had access to *only one random sample* from our bowl of pasta.
- How could we study the effect of sampling variation with a single sample? ↗ **bootstrap resampling with replacement!**



# Back to Pasta

- As in real life, imagine we had access to *only one random sample* from our bowl of pasta.
- How could we study the effect of sampling variation with a single sample? ↗ **bootstrap resampling with replacement!**
- Let's start by drawing one random sample of size  $n = 50$  from our bowl.

```
library(tidyverse)
bowl <- read.csv("https://www.dropbox.com/s/qpjsk0rfgc0gx80/pasta.csv?dl=1")

my_sample = bowl %>%
  mutate(color = ifelse(color == "green", "green", "non-green")) %>%
  rep_sample_n(size = 50) %>%
  ungroup() %>%
  select(pasta_ID, color)

head(my_sample, 3)

## # A tibble: 3 x 2
##   pasta_ID color
##       <int> <fct>
## 1        4 non-green
## 2       41 non-green
## 3       79 non-green
```



# Back to Pasta

- As in real life, imagine we had access to *only one random sample* from our bowl of pasta.
- How could we study the effect of sampling variation with a single sample? ↗ **bootstrap resampling with replacement!**
- Let's start by drawing one random sample of size  $n = 50$  from our bowl.

```
library(tidyverse)
bowl <- read.csv("https://www.dropbox.com/s/qpjsk0rfgc0gx80/pasta.csv?dl=1")
```

```
my_sample = bowl %>%
  mutate(color = ifelse(color == "green", "green", "non-green")) %>%
  rep_sample_n(size = 50) %>%
  ungroup() %>%
  select(pasta_ID, color)
```

```
head(my_sample, 3)
## # A tibble: 3 x 2
##   pasta_ID color
##       <int> <fct>
## 1        4 non-green
## 2       41 non-green
## 3       79 non-green
```

```
p_hat = mean(my_sample$color == "green")
p_hat
## [1] 0.46
```

The proportion of green pasta in this sample is:  $\hat{p} = 0.46$ .



# Resampling our Pasta Sample

How do we obtain a *bootstrap sample*?



# Resampling our Pasta Sample

How do we obtain a *bootstrap sample*?

1. Randomly pick *one* pasta from the sample and record the associated color.



# Resampling our Pasta Sample

How do we obtain a *bootstrap sample*?

1. Randomly pick *one* pasta from the sample and record the associated color.
2. Put this pasta back in the sample.



# Resampling our Pasta Sample

How do we obtain a *bootstrap sample*?

1. Randomly pick *one* pasta from the sample and record the associated color.
2. Put this pasta back in the sample.
3. Repeat steps 1 and 2 49 times, i.e. *until the new sample is of the same size as the original sample*.



# Resampling our Pasta Sample

How do we obtain a *bootstrap sample*?

1. Randomly pick *one* pasta from the sample and record the associated color.
2. Put this pasta back in the sample.
3. Repeat steps 1 and 2 49 times, i.e. *until the new sample is of the same size as the original sample*.
4. Compute the proportion of green pasta in the bootstrap sample.



# Resampling our Pasta Sample

How do we obtain a *bootstrap sample*?

1. Randomly pick *one* pasta from the sample and record the associated color.
2. Put this pasta back in the sample.
3. Repeat steps 1 and 2 49 times, i.e. *until the new sample is of the same size as the original sample*.
4. Compute the proportion of green pasta in the bootstrap sample.

This procedure is called *resampling with replacement*:

- *resampling*: drawing repeated samples from a sample.
- *with replacement*: each time the drawn pasta is put back in the sample.



# Resampling our Pasta Sample

Here is one bootstrap sample:

```
one_bootstrap = my_sample %>%
  rep_sample_n(size = 50, replace = TRUE) %>%
  arrange(pasta_ID)

head(one_bootstrap, 8)
```

```
## # A tibble: 8 x 3
## # Groups:   replicate [1]
##   replicate pasta_ID color
##       <int>     <int> <fct>
## 1         1         4 non-green
## 2         1        41 non-green
## 3         1        41 non-green
## 4         1        79 non-green
## 5         1        79 non-green
## 6         1       103 non-green
## 7         1       103 non-green
## 8         1       103 non-green
```

```
nrow(one_bootstrap)
```

```
## [1] 50
```



# Resampling our Pasta Sample

Here is one bootstrap sample:

```
one_bootstrap = my_sample %>%
  rep_sample_n(size = 50, replace = TRUE) %>%
  arrange(pasta_ID)

head(one_bootstrap, 8)

## # A tibble: 8 x 3
## # Groups:   replicate [1]
##   replicate pasta_ID color
##       <int>    <int> <fct>
## 1         1        4 non-green
## 2         1       41 non-green
## 3         1       41 non-green
## 4         1       79 non-green
## 5         1       79 non-green
## 6         1      103 non-green
## 7         1      103 non-green
## 8         1      103 non-green

nrow(one_bootstrap)

## [1] 50
```

Several pasta have been drawn multiple times. How come?

What's the proportion of green pasta in this bootstrap sample?

```
mean(one_bootstrap$color == "green")

## [1] 0.4
```

The proportion is different than that in our sample! This is due to resampling **with replacement**.

What if we repeated this resampling procedure many times? Would the proportion be the same each time?



# Obtaining the Bootstrap Distribution

- Let's repeat the resampling procedure 1,000 times: there will be 1,000 bootstrap samples and 1,000 bootstrap estimates!



# Obtaining the Bootstrap Distribution

- Let's repeat the resampling procedure 1,000 times: there will be 1,000 bootstrap samples and 1,000 bootstrap estimates!

We use the `infer` package to ease the bootstrapping procedure.

```
library(infer)

bootstrap_distrib = my_sample %>%
  # specify the variable and level of interest
  specify(response = color, success = "green") %>%
  # generate 1000 bootstrap samples
  generate(reps = 1000, type = "bootstrap") %>%
  # calculate the proportion of green pasta for each
  calculate(stat = "prop")
```



# Obtaining the Bootstrap Distribution

- Let's repeat the resampling procedure 1,000 times: there will be 1,000 bootstrap samples and 1,000 bootstrap estimates!

We use the `infer` package to ease the bootstrapping procedure.

```
library(infer)

bootstrap_distrib = my_sample %>%
  # specify the variable and level of interest
  specify(response = color, success = "green") %>%
  # generate 1000 bootstrap samples
  generate(reps = 1000, type = "bootstrap") %>%
  # calculate the proportion of green pasta for each
  calculate(stat = "prop")
```

Here are the first 6 rows:

```
head(bootstrap_distrib)

## # A tibble: 6 x 2
##   replicate stat
##       <int> <dbl>
## 1         1  0.44
## 2         2  0.36
## 3         3  0.46
## 4         4  0.46
## 5         5  0.52
## 6         6  0.52
```

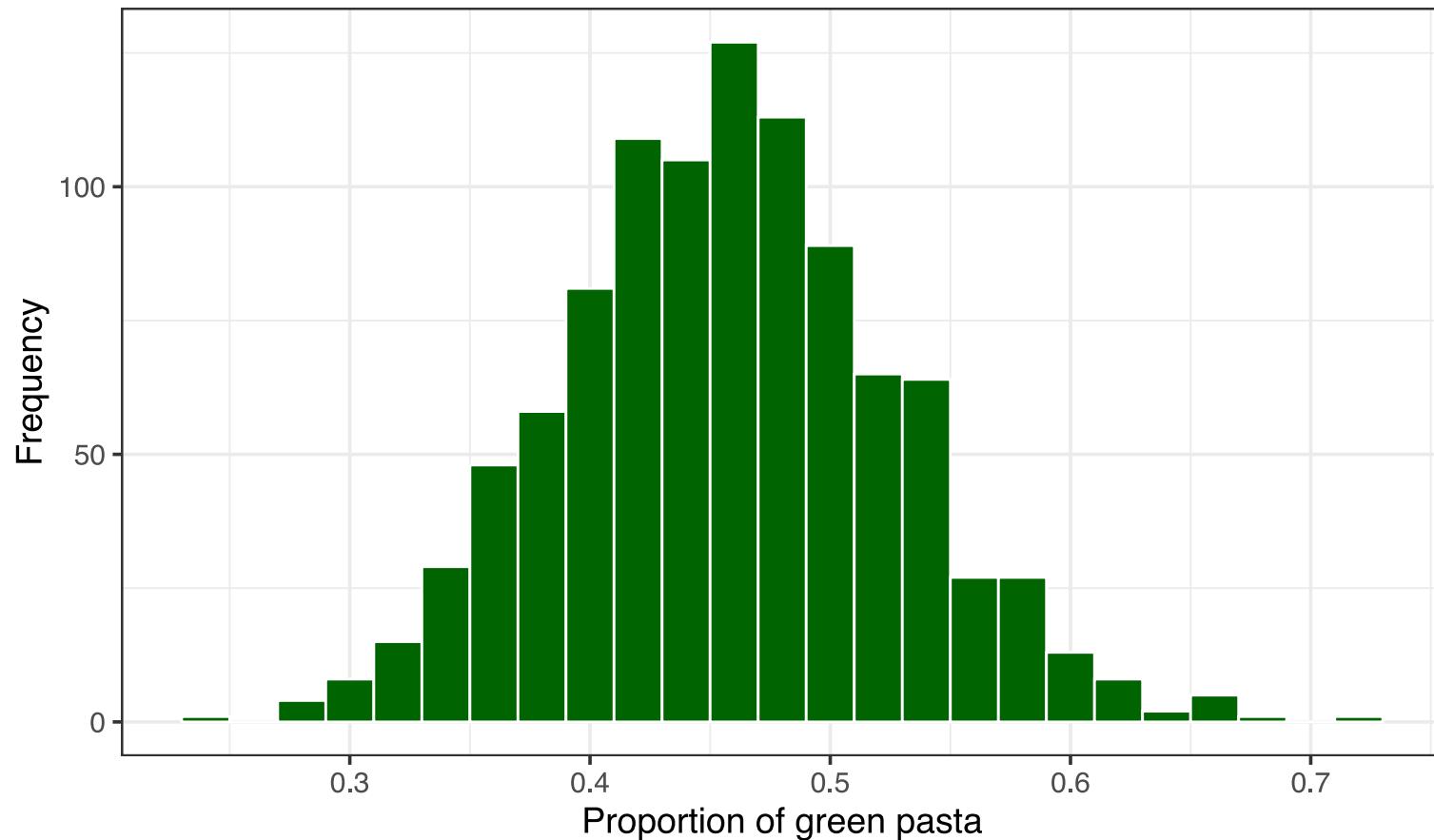
```
nrow(bootstrap_distrib)

## [1] 1000
```

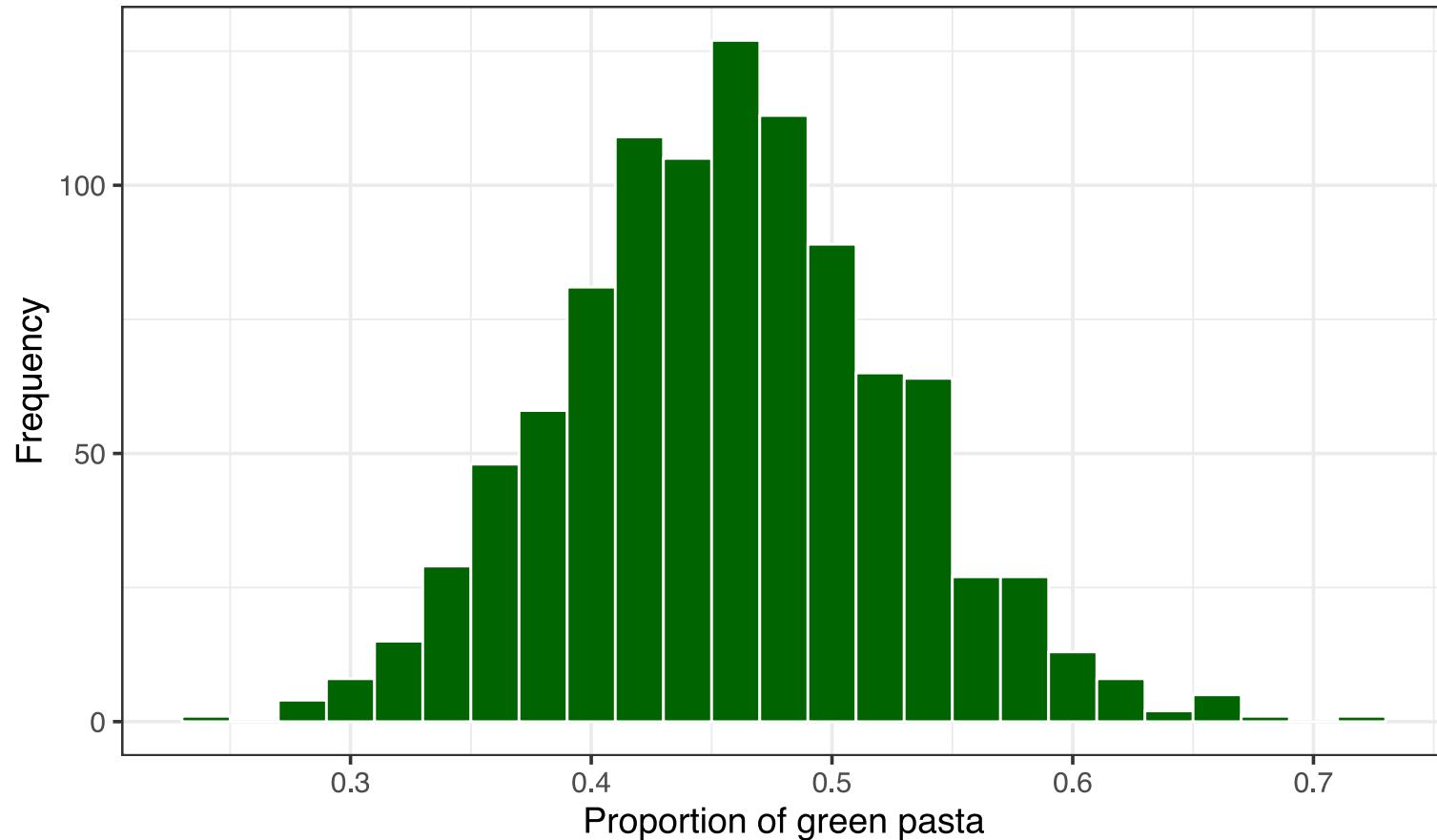
Let's visualize this sampling variation!



# Bootstrap Distribution



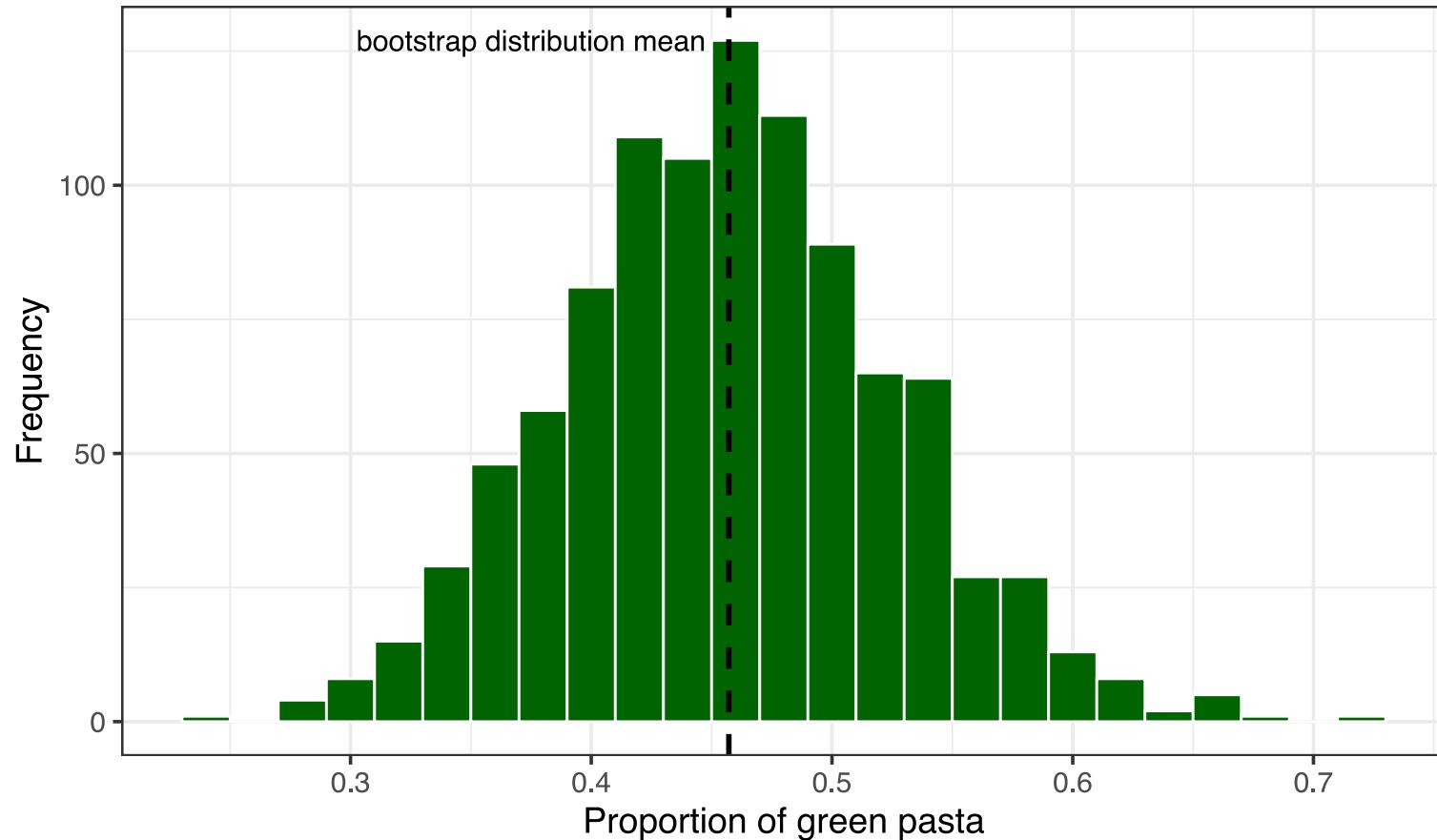
# Bootstrap Distribution



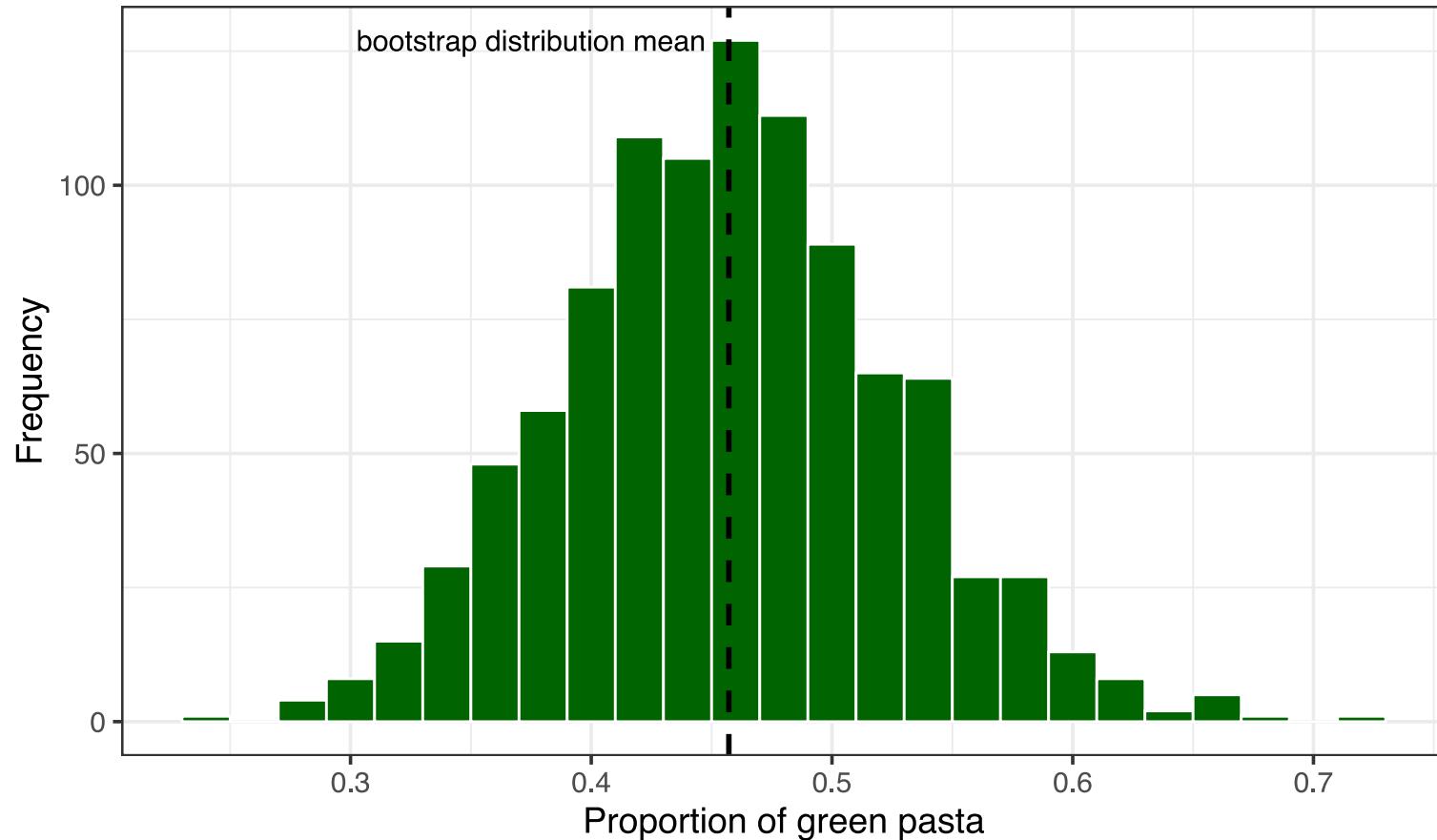
The *bootstrap distribution* is an approximation of the *sampling distribution*.



# Bootstrap Distribution with Mean



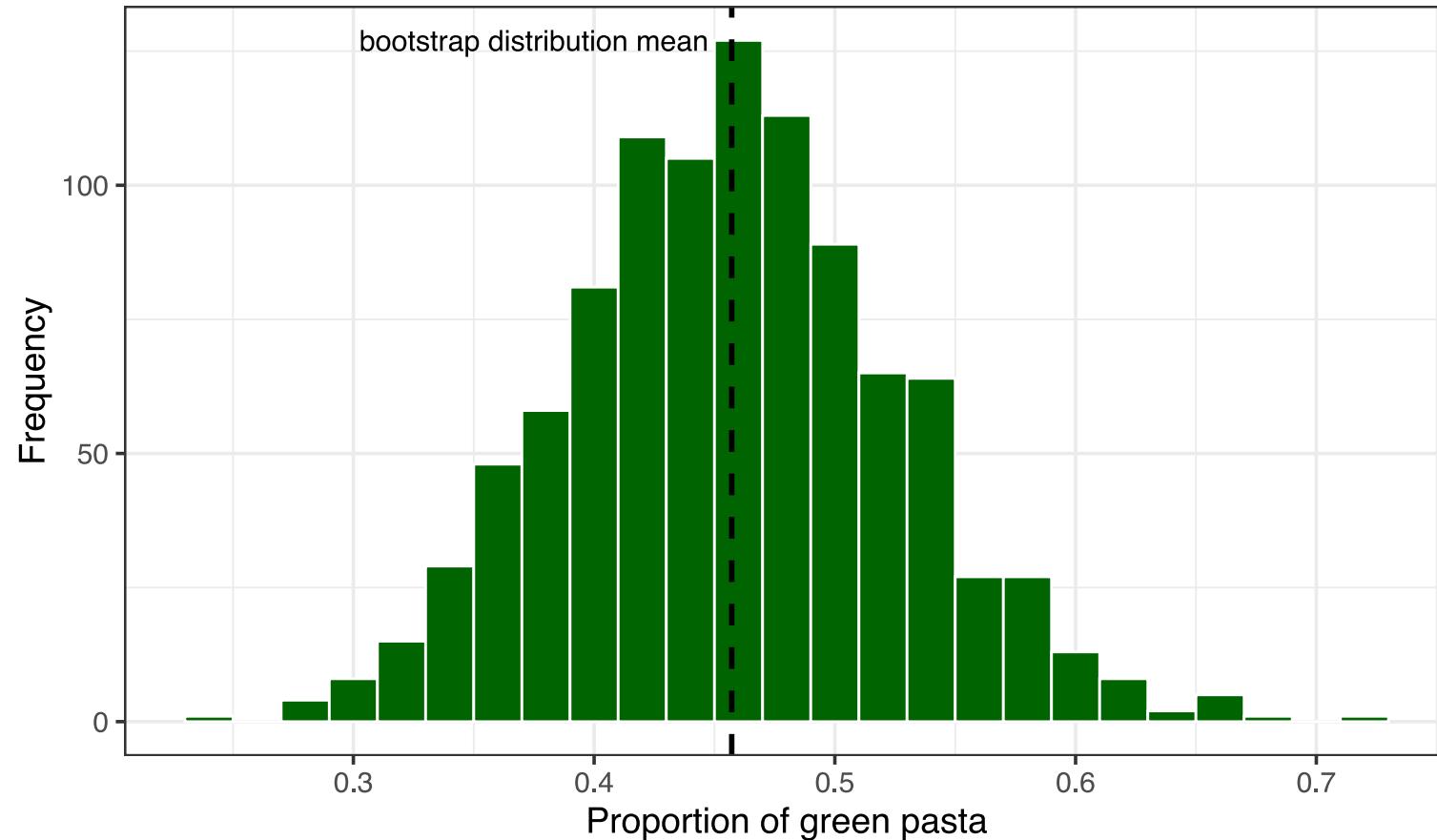
# Bootstrap Distribution with Mean



The **bootstrap distribution** mean is very close to the original sample proportion.



# Bootstrap Distribution with Mean



Let's use this **bootstrap distribution** to construct confidence intervals!



# Understanding Confidence Intervals

- Analogy with fishing:
  - *point estimate*: fishing with a spear.



# Understanding Confidence Intervals

- Analogy with fishing:
  - *point estimate*: fishing with a spear.
  - *confidence interval*: fishing with a net.



# Understanding Confidence Intervals

- Analogy with fishing:
  - *point estimate*: fishing with a spear.
  - *confidence interval*: fishing with a net.
- In our case the fish is the true proportion of pasta in the bowl that are green ( $p$ ).



# Understanding Confidence Intervals

- Analogy with fishing:
  - *point estimate*: fishing with a spear.
  - *confidence interval*: fishing with a net.
- In our case the fish is the true proportion of pasta in the bowl that are green ( $p$ ).
- The *point estimate* would be the proportion of green pasta obtained from a random sample ( $\hat{p}$ ).



# Understanding Confidence Intervals

- Analogy with fishing:
  - *point estimate*: fishing with a spear.
  - *confidence interval*: fishing with a net.
- In our case the fish is the true proportion of pasta in the bowl that are green ( $p$ ).
- The *point estimate* would be the proportion of green pasta obtained from a random sample ( $\hat{p}$ ).
- The *confidence interval*: from the previous bootstrap distribution, ***where do most proportions lie?***



# Understanding Confidence Intervals

- Analogy with fishing:
  - *point estimate*: fishing with a spear.
  - *confidence interval*: fishing with a net.
- In our case the fish is the true proportion of pasta in the bowl that are green ( $p$ ).
- The *point estimate* would be the proportion of green pasta obtained from a random sample ( $\hat{p}$ ).
- The *confidence interval*: from the previous bootstrap distribution, ***where do most proportions lie?***
- Method for confidence interval construction: ***percentile method***.



# Understanding Confidence Intervals

- Analogy with fishing:
  - *point estimate*: fishing with a spear.
  - *confidence interval*: fishing with a net.
- In our case the fish is the true proportion of pasta in the bowl that are green ( $p$ ).
- The *point estimate* would be the proportion of green pasta obtained from a random sample ( $\hat{p}$ ).
- The *confidence interval*: from the previous bootstrap distribution, ***where do most proportions lie?***
- Method for confidence interval construction: ***percentile method***.
- Requires specifying a ***confidence level***: 90%, 95%, and 99% are the most common.



# Percentile Method: 95% Confidence Interval

- Construct a confidence interval as the middle 95% of values of the bootstrap distribution.



# Percentile Method: 95% Confidence Interval

- Construct a confidence interval as the middle 95% of values of the bootstrap distribution.
- For that, we compute the 2.5% and 97.5% percentile:

```
quantile(bootstrap_distrib$stat, 0.025)
```

```
## 2.5%
## 0.32
```

```
quantile(bootstrap_distrib$stat, 0.975)
```

```
## 97.5%
## 0.6
```

- Therefore the 95% confidence interval is [0.32; 0.6].
- It is a *range* of values.



# Percentile Method: 95% Confidence Interval

- Construct a confidence interval as the middle 95% of values of the bootstrap distribution.
- For that, we compute the 2.5% and 97.5% percentile:

```
quantile(bootstrap_distrib$stat, 0.025)
```

```
## 2.5%
## 0.32
```

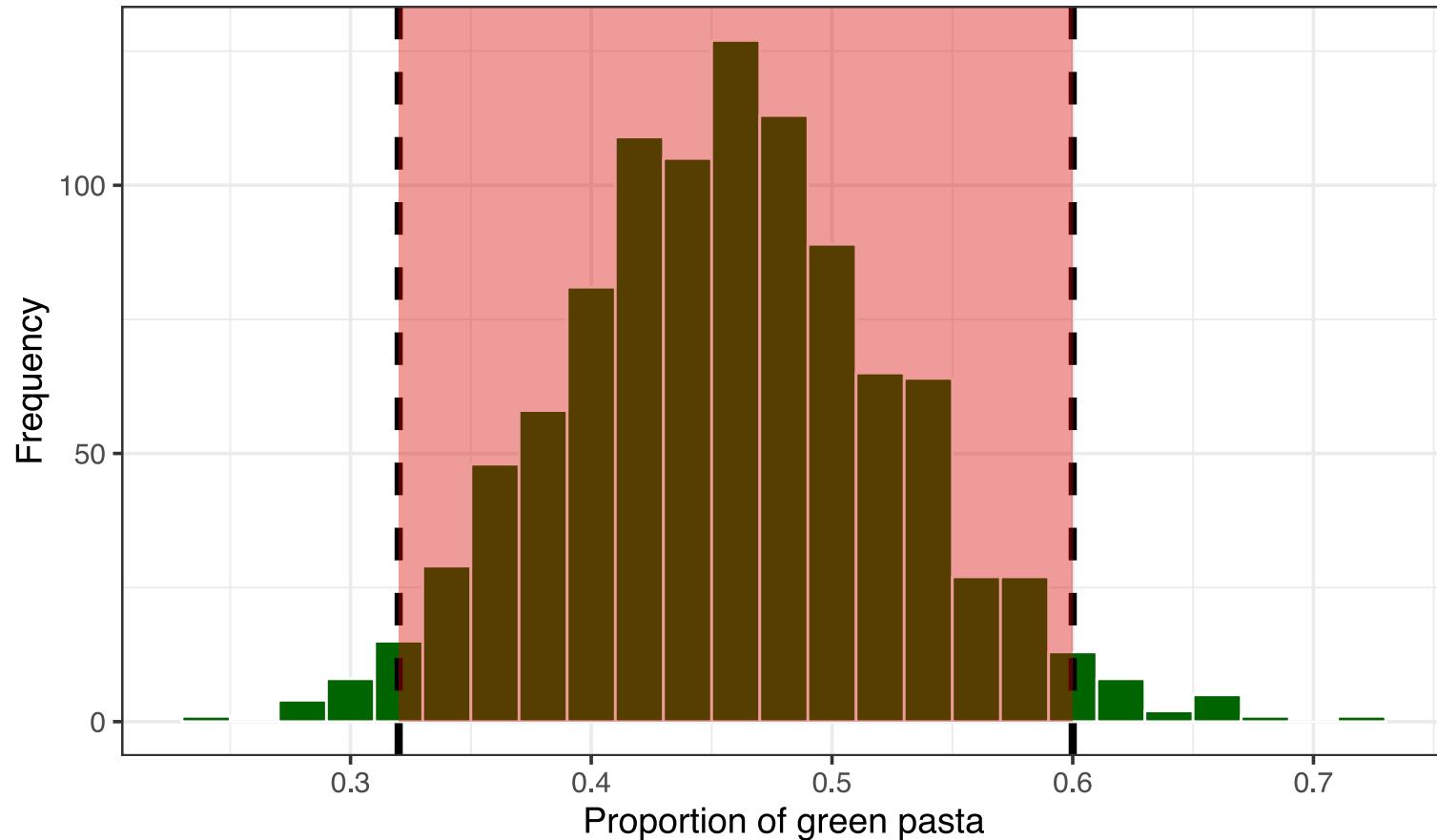
```
quantile(bootstrap_distrib$stat, 0.975)
```

```
## 97.5%
## 0.6
```

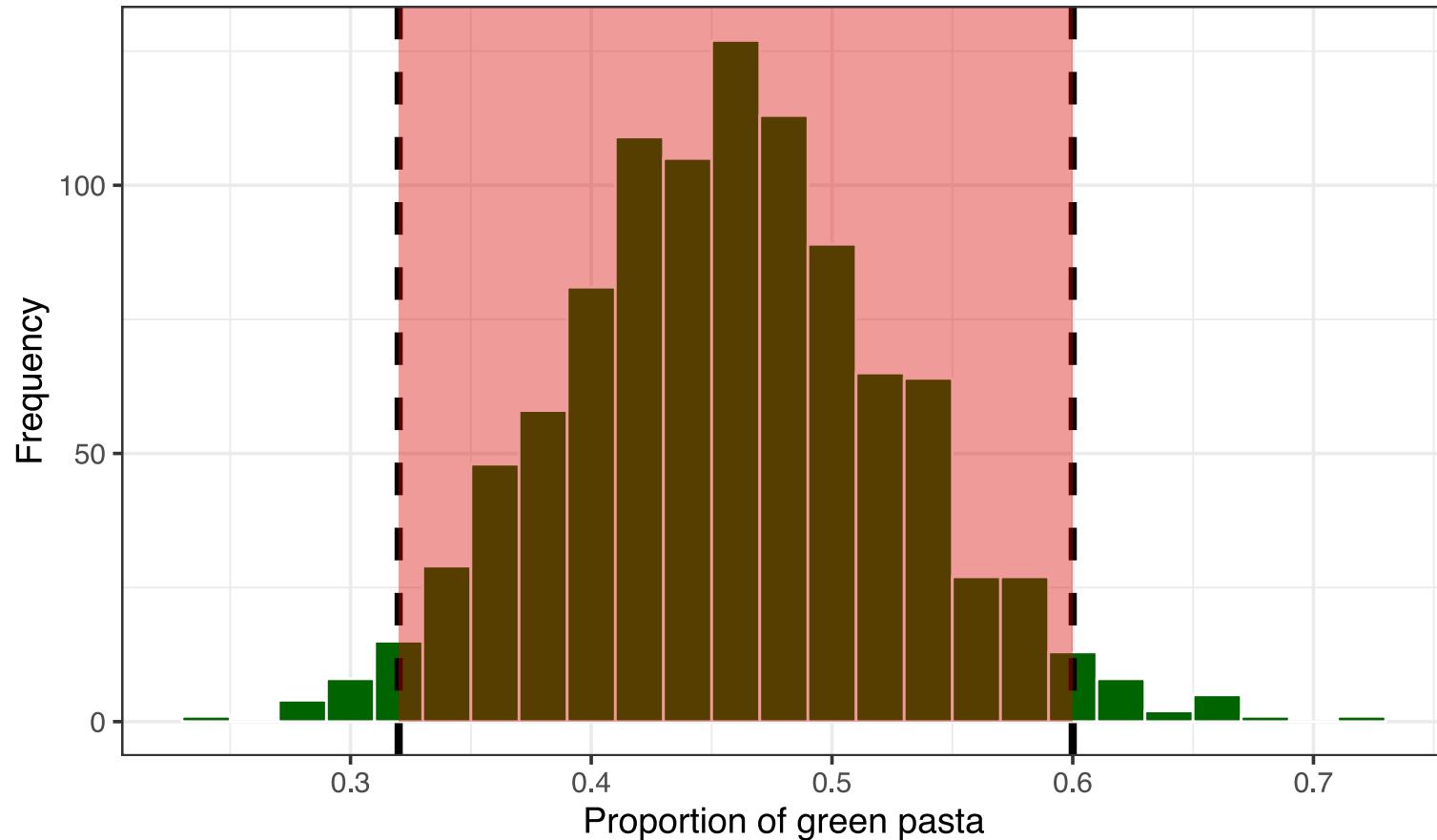
- Therefore the 95% confidence interval is [0.32; 0.6].
- It is a *range* of values.
- Let's see this confidence interval on the sampling distribution.



# Percentile Method: 95% Confidence Interval Visually



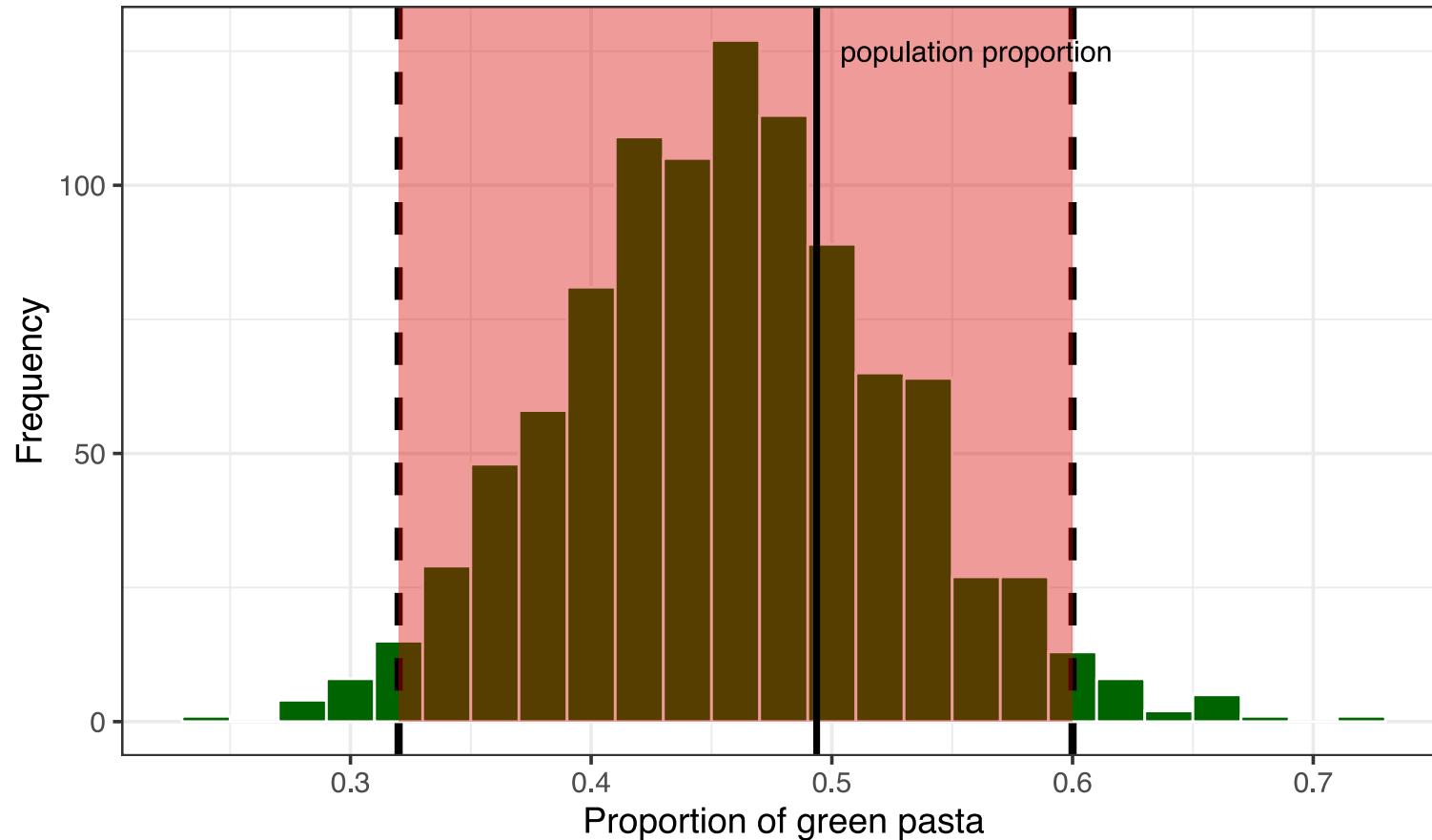
# Percentile Method: 95% Confidence Interval Visually



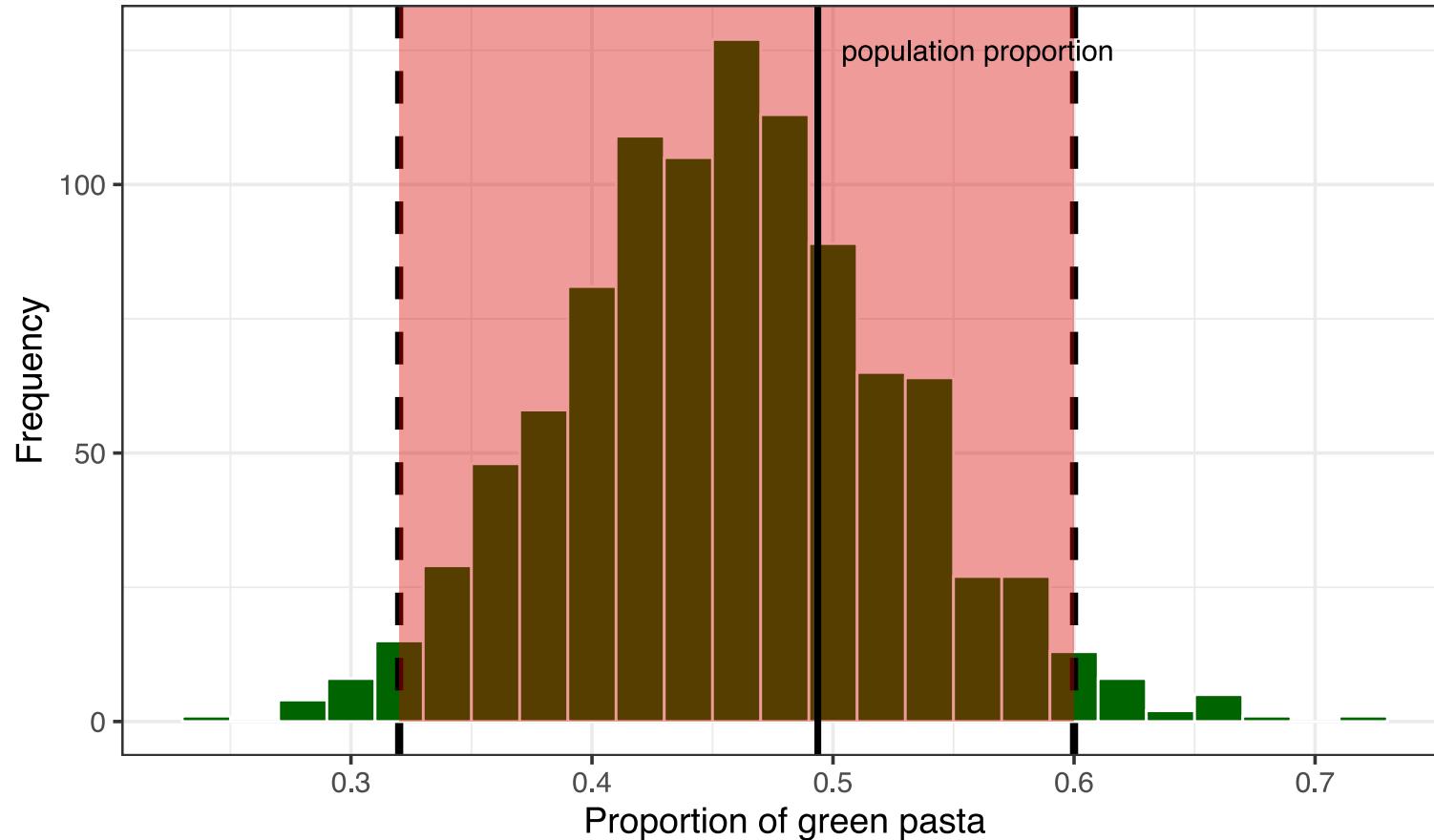
Does the interval contain the true population proportion?



# Percentile Method: 95% Confidence Interval Visually



# Percentile Method: 95% Confidence Interval Visually



True population parameter is indeed in our 95% interval! Will it always be?

# Interpreting a 95% Confidence Interval

Let's repeatedly draw 100 different samples from our `bowl` and for each sample compute the associated 95% CI using the percentile method.



# Interpreting a 95% Confidence Interval

Let's repeatedly draw 100 different samples from our `bowl` and for each sample compute the associated 95% CI using the percentile method.



How many confidence intervals contain the true parameter? Why?

# Interpreting a 95% Confidence Interval

*Precise interpretation:* If we repeated our sampling procedure ***a large number of times***, we ***expect about 95%*** of the resulting confidence intervals to capture the value of the population parameter.

In other words, 95% of the time, the 95% confidence interval will contain the true population parameter.



# Interpreting a 95% Confidence Interval

*Precise interpretation:* If we repeated our sampling procedure ***a large number of times***, we ***expect about 95%*** of the resulting confidence intervals to capture the value of the population parameter.

In other words, 95% of the time, the 95% confidence interval will contain the true population parameter.

*Short-hand interpretation:* We are ***95% “confident”*** that a 95% confidence interval captures the value of the population parameter.



# Interpreting a 95% Confidence Interval

*Precise interpretation:* If we repeated our sampling procedure ***a large number of times***, we ***expect about 95%*** of the resulting confidence intervals to capture the value of the population parameter.

In other words, 95% of the time, the 95% confidence interval will contain the true population parameter.

*Short-hand interpretation:* We are ***95% “confident”*** that a 95% confidence interval captures the value of the population parameter.

## ***Questions:***

- How does the width of the confidence interval change as the ***confidence level*** increases?
- How does the width of the confidence interval change as the ***sample size*** increases?



# Interpreting a 95% Confidence Interval

*Precise interpretation:* If we repeated our sampling procedure ***a large number of times***, we ***expect about 95%*** of the resulting confidence intervals to capture the value of the population parameter.

In other words, 95% of the time, the 95% confidence interval will contain the true population parameter.

*Short-hand interpretation:* We are ***95% “confident”*** that a 95% confidence interval captures the value of the population parameter.

***Impact of confidence level:*** the greater the confidence level, the wider the confidence intervals.

- *Intuition:* a greater confidence level means the confidence interval needs to contain the true population parameter more often, and thus needs to be wider to ensure this.



# Interpreting a 95% Confidence Interval

*Precise interpretation:* If we repeated our sampling procedure **a large number of times**, we **expect about 95%** of the resulting confidence intervals to capture the value of the population parameter.

In other words, 95% of the time, the 95% confidence interval will contain the true population parameter.

*Short-hand interpretation:* We are **95% “confident”** that a 95% confidence interval captures the value of the population parameter.

**Impact of confidence level:** the greater the confidence level, the wider the confidence intervals.

- *Intuition:* a greater confidence level means the confidence interval needs to contain the true population parameter more often, and thus needs to be wider to ensure this.

**Impact of sample size:** the greater the confidence level, the wider the confidence intervals.

- *Inuition:* a larger sample size leads to less sampling variation and therefore a narrower bootstrap distribution, which in turn leads to thinner confidence intervals.



# From Confidence Intervals to Hypothesis Testing

- *Confidence intervals* can be thought of as an extension of *point estimation*.



# From Confidence Intervals to Hypothesis Testing

- *Confidence intervals* can be thought of as an extension of *point estimation*.
- What if we want to **compare** a sample statistic for two groups?
  - *Example:* differences in average wages between men and women. Are the observed differences **significant?**



# From Confidence Intervals to Hypothesis Testing

- *Confidence intervals* can be thought of as an extension of *point estimation*.
- What if we want to **compare** a sample statistic for two groups?
  - *Example:* differences in average wages between men and women. Are the observed differences **significant**?
- These comparisons are the realm of ***hypothesis testing***.



# From Confidence Intervals to Hypothesis Testing

- *Confidence intervals* can be thought of as an extension of *point estimation*.
- What if we want to **compare** a sample statistic for two groups?
  - *Example:* differences in average wages between men and women. Are the observed differences **significant**?
- These comparisons are the realm of ***hypothesis testing***.
- Just like confidence intervals, hypothesis tests are used to make claims about a population based on information from a sample.
- However, we'll see that the framework for making such inferences is slightly different.



# Hypothesis Testing

# Is There Gender Discrimination In Promotions?

- We will use data from an [article](#) published in the *Journal of Applied Psychology* in 1974 which investigated whether female employees at banks were discriminated against.
- 48 (male) supervisors were given *identical* candidate CVs, differing only with respect to the first name, which was male or female.
  - Each CV was "*in the form of a memorandum requesting a decision on the promotion of an employee to the position of branch manager.*"



# Is There Gender Discrimination In Promotions?

- We will use data from an **article** published in the *Journal of Applied Psychology* in 1974 which investigated whether female employees at banks were discriminated against.
- 48 (male) supervisors were given *identical* candidate CVs, differing only with respect to the first name, which was male or female.
  - Each CV was "*in the form of a memorandum requesting a decision on the promotion of an employee to the position of branch manager.*"
- **Hypothesis** we want to test: *Is there gender discrimination?*



# Is There Gender Discrimination In Promotions?

- We will use data from an **article** published in the *Journal of Applied Psychology* in 1974 which investigated whether female employees at banks were discriminated against.
- 48 (male) supervisors were given *identical* candidate CVs, differing only with respect to the first name, which was male or female.
  - Each CV was "*in the form of a memorandum requesting a decision on the promotion of an employee to the position of branch manager.*"
- **Hypothesis** we want to test: *Is there gender discrimination?*
- The data from the experiment are provided in the **promotions** dataset from the **moderndive** package.



# Is There Gender Discrimination In Promotions?

- We will use data from an **article** published in the *Journal of Applied Psychology* in 1974 which investigated whether female employees at banks were discriminated against.
- 48 (male) supervisors were given *identical* candidate CVs, differing only with respect to the first name, which was male or female.
  - Each CV was "*in the form of a memorandum requesting a decision on the promotion of an employee to the position of branch manager.*"
- **Hypothesis** we want to test: *Is there gender discrimination?*

- The data from the experiment are provided in the **promotions** dataset from the **moderndive** package.

```
library(moderndive)
head(promotions)

## # A tibble: 6 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
```



# Evidence of Discrimination?

How many men and women were offered a promotion (and not)?

```
promotions %>%
  group_by(gender, decision) %>%
  tally() %>%
  mutate(percentage = 100 * n / sum(n))

## # A tibble: 4 x 4
## # Groups:   gender [2]
##   gender decision     n percentage
##   <fct>   <fct>   <int>      <dbl>
## 1 male    not         3        12.5
## 2 male    promoted   21       87.5
## 3 female  not        10       41.7
## 4 female  promoted  14       58.3
```

There is a **29.2 percentage points difference** in promotions between men and women!



# Evidence of Discrimination?

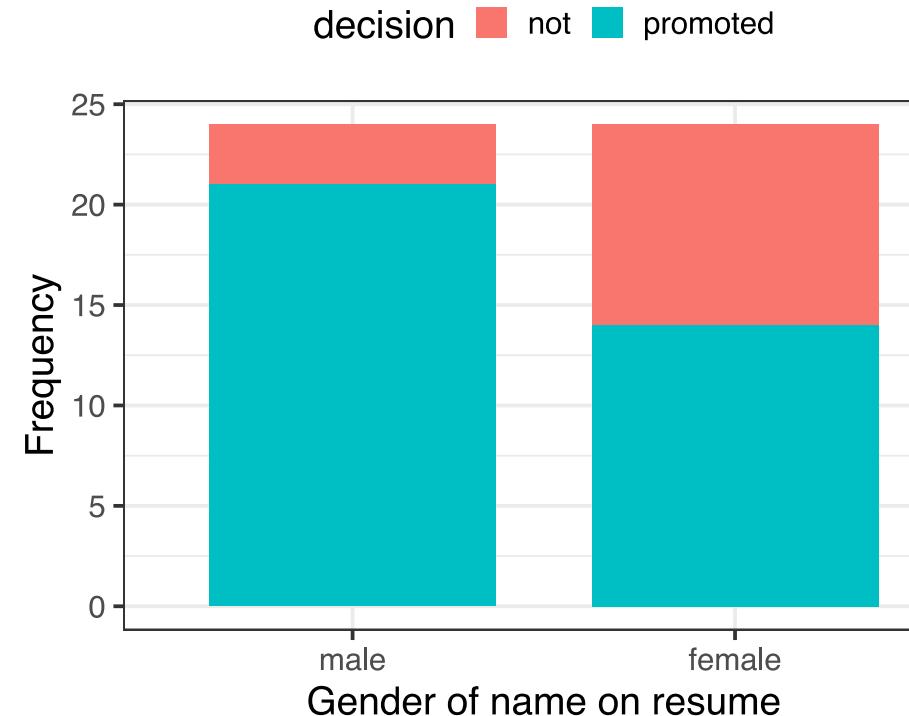
How many men and women were offered a promotion (and not)?

```
promotions %>%
  group_by(gender, decision) %>%
  tally() %>%
  mutate(percentage = 100 * n / sum(n))

## # A tibble: 4 x 4
## # Groups:   gender [2]
##   gender decision     n percentage
##   <fct>   <fct>   <int>      <dbl>
## 1 male    not        3       12.5
## 2 male    promoted  21      87.5
## 3 female  not       10      41.7
## 4 female  promoted 14      58.3
```

There is a **29.2 percentage points difference** in promotions between men and women!

Promotion decision



# Evidence of Discrimination?

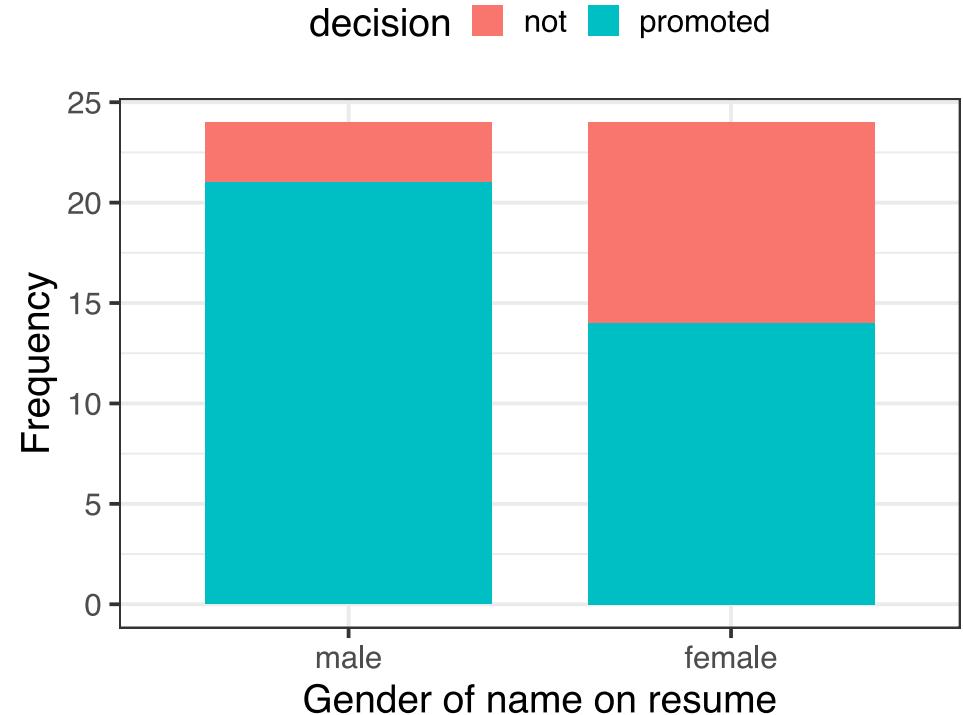
How many men and women were offered a promotion (and not)?

```
promotions %>%
  group_by(gender, decision) %>%
  tally() %>%
  mutate(percentage = 100 * n / sum(n))

## # A tibble: 4 x 4
## # Groups:   gender [2]
##   gender decision     n percentage
##   <fct>   <fct>   <int>      <dbl>
## 1 male    not        3       12.5
## 2 male    promoted  21      87.5
## 3 female  not       10      41.7
## 4 female  promoted 14      58.3
```

There is a **29.2 percentage points difference** in promotions between men and women!

Promotion decision



**Question:** Is this difference **conclusive evidence** of differences in promotion rates between men and women? Could such a difference have been observed **by chance**?



# Imposing A Hypothetical World: No Gender Discrimination

- Suppose we lived in a world without gender discrimination: the promotion decision would be completely *independent* from gender.
- Let's randomly reassign `gender` to each row and see how this affects the result.



# Imposing A Hypothetical World: No Gender Discrimination

- Suppose we lived in a world without gender discrimination: the promotion decision would be completely *independent* from gender.
- Let's randomly reassign `gender` to each row and see how this affects the result.

```
promotions %>%
  left_join(promotions_shuffled %>%
             rename(shuffled_gender = gender)) %
  head()

## # A tibble: 6 x 4
##       id decision gender shuffled_gender
##   <int> <fct>    <fct>    <fct>
## 1     1 promoted male    female
## 2     2 promoted male    female
## 3     3 promoted male     male
## 4     4 promoted male    female
## 5     5 promoted male     male
## 6     6 promoted male     male
```

How do the promotion rates look like in our reshuffled sample?



# Imposing A Hypothetical World: No Gender Discrimination

- Suppose we lived in a world without gender discrimination: the promotion decision would be completely *independent* from gender.
- Let's randomly reassign `gender` to each row and see how this affects the result.

```
promotions %>%
  left_join(promotions_shuffled %>%
             rename(shuffled_gender = gender)) %
  head()

## # A tibble: 6 x 4
##       id decision gender shuffled_gender
##   <int> <fct>    <fct>    <fct>
## 1     1 promoted male    female
## 2     2 promoted male    female
## 3     3 promoted male     male
## 4     4 promoted male    female
## 5     5 promoted male     male
## 6     6 promoted male     male
```

How do the promotion rates look like in our reshuffled sample?

```
promotions_shuffled %>%
  group_by(gender, decision) %>%
  tally() %>%
  mutate(percentage = 100 * n / sum(n))

## # A tibble: 4 x 4
## # Groups:   gender [2]
##   gender decision      n percentage
##   <fct>    <fct> <int>     <dbl>
## 1 male      not        6        25
## 2 male      promoted   18       75
## 3 female    not        7       29.2
## 4 female    promoted  17      70.8
```

The difference is much lower: **4.2 percentage points!**



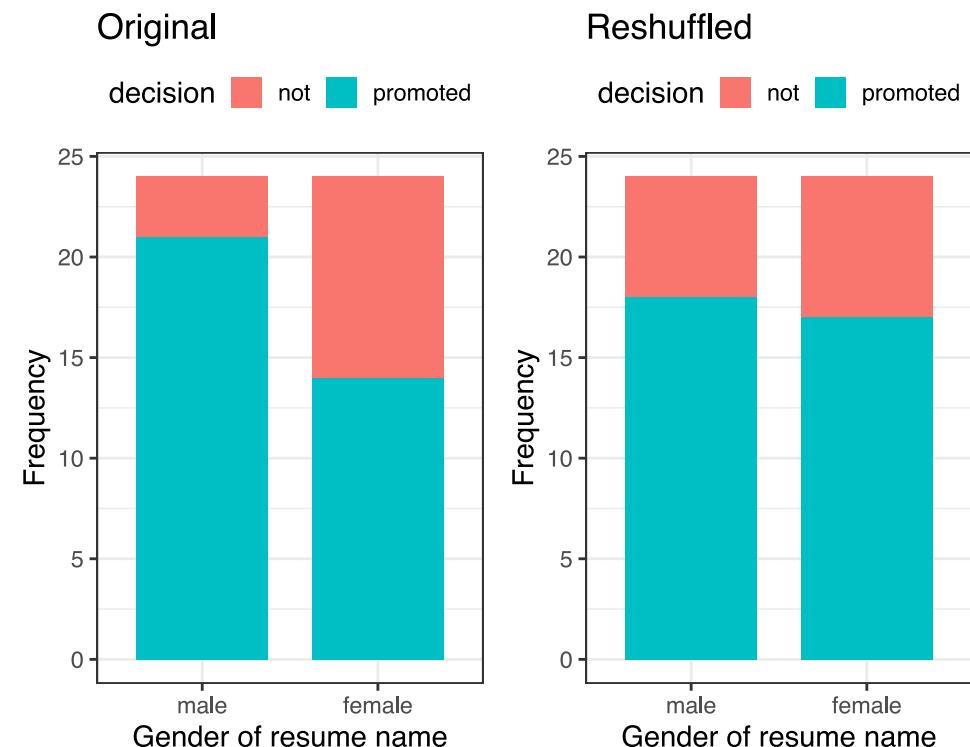
# Imposing A Hypothetical World: No Gender Discrimination

- Suppose we lived in a world without gender discrimination: the promotion decision would be completely *independent* from gender.
- Let's randomly reassign `gender` to each row and see how this affects the result.

```
promotions %>%
  left_join(promotions_shuffled %>%
             rename(shuffled_gender = gender)) %
  head()

## # A tibble: 6 x 4
##       id decision gender shuffled_gender
##   <int> <fct>    <fct>    <fct>
## 1     1 promoted male     female
## 2     2 promoted male     female
## 3     3 promoted male      male
## 4     4 promoted male     female
## 5     5 promoted male      male
## 6     6 promoted male      male
```

How do the promotion rates look like in our reshuffled sample?



# Sampling Variation

- In our hypothetical world, the difference in promotion rates was only 4.2 percentage points.
- Can we answer our initial question about the existence of gender discrimination now?



# Sampling Variation

- In our hypothetical world, the difference in promotion rates was only 4.2 percentage points.
- Can we answer our initial question about the existence of gender discrimination now?
- No, we must investigate the role of *sampling variation!*
  - What if we reshuffle once again, how different from 4.2% $p$  (*percentage points*) would the difference be?



# Sampling Variation

- In our hypothetical world, the difference in promotion rates was only 4.2 percentage points.
- Can we answer our initial question about the existence of gender discrimination now?
- No, we must investigate the role of *sampling variation!*
  - What if we reshuffle once again, how different from 4.2%p (*percentage points*) would the difference be?
  - In other words, how representative of that hypothetical world is 4.2%p?



# Sampling Variation

- In our hypothetical world, the difference in promotion rates was only 4.2 percentage points.
- Can we answer our initial question about the existence of gender discrimination now?
- No, we must investigate the role of *sampling variation!*
  - What if we reshuffle once again, how different from 4.2%p (*percentage points*) would the difference be?
  - In other words, how representative of that hypothetical world is 4.2%p?
  - How likely is a 29%p difference to occur in such a world?



# Sampling Variation

- In our hypothetical world, the difference in promotion rates was only 4.2 percentage points.
- Can we answer our initial question about the existence of gender discrimination now?
- No, we must investigate the role of *sampling variation!*
  - What if we reshuffle once again, how different from 4.2%p (*percentage points*) would the difference be?
  - In other words, how representative of that hypothetical world is 4.2%p?
  - How likely is a 29%p difference to occur in such a world?
- We need to know about the whole sampling distribution under the *no discrimination* hypothesis.

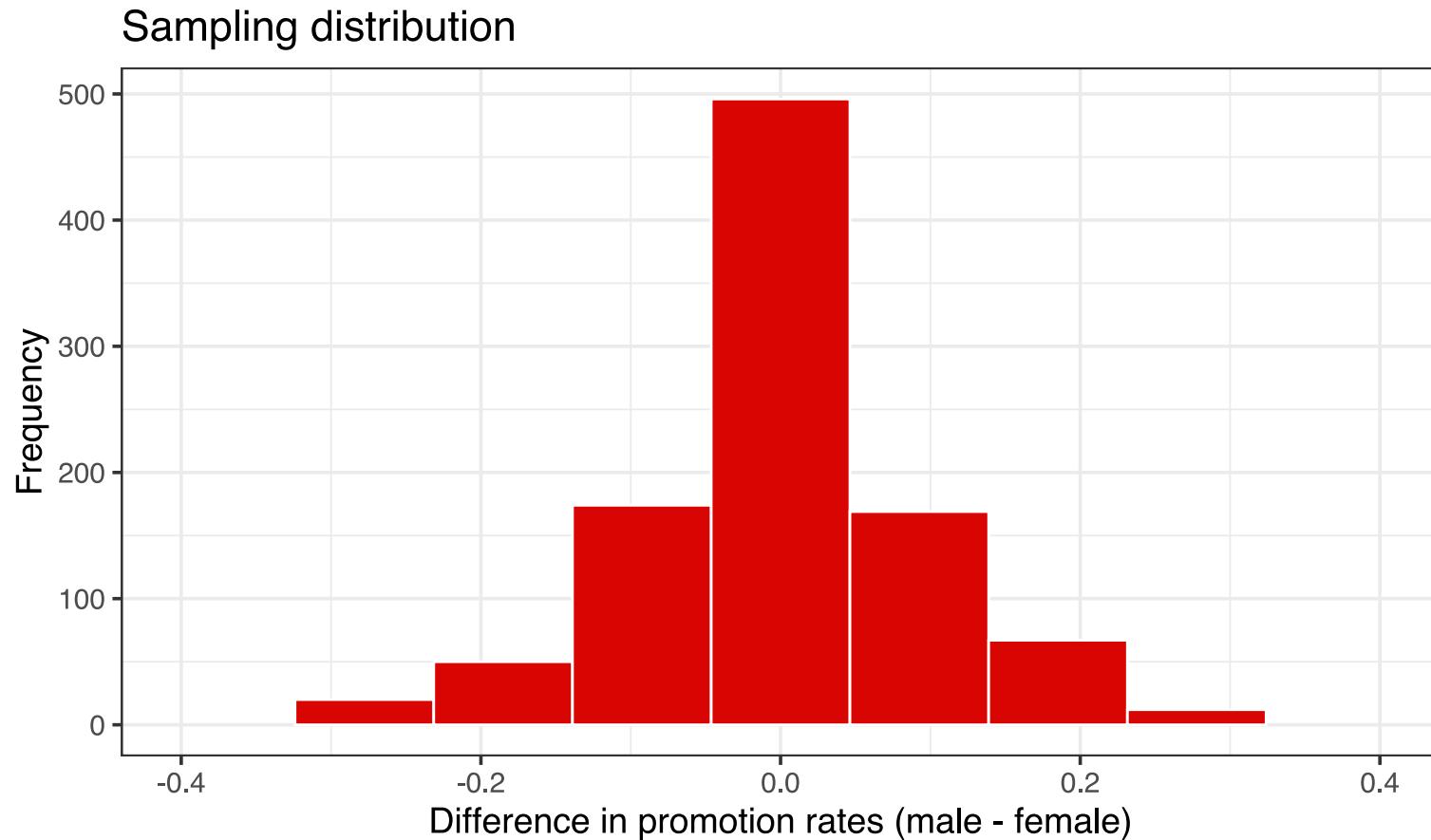


# Sampling Variation

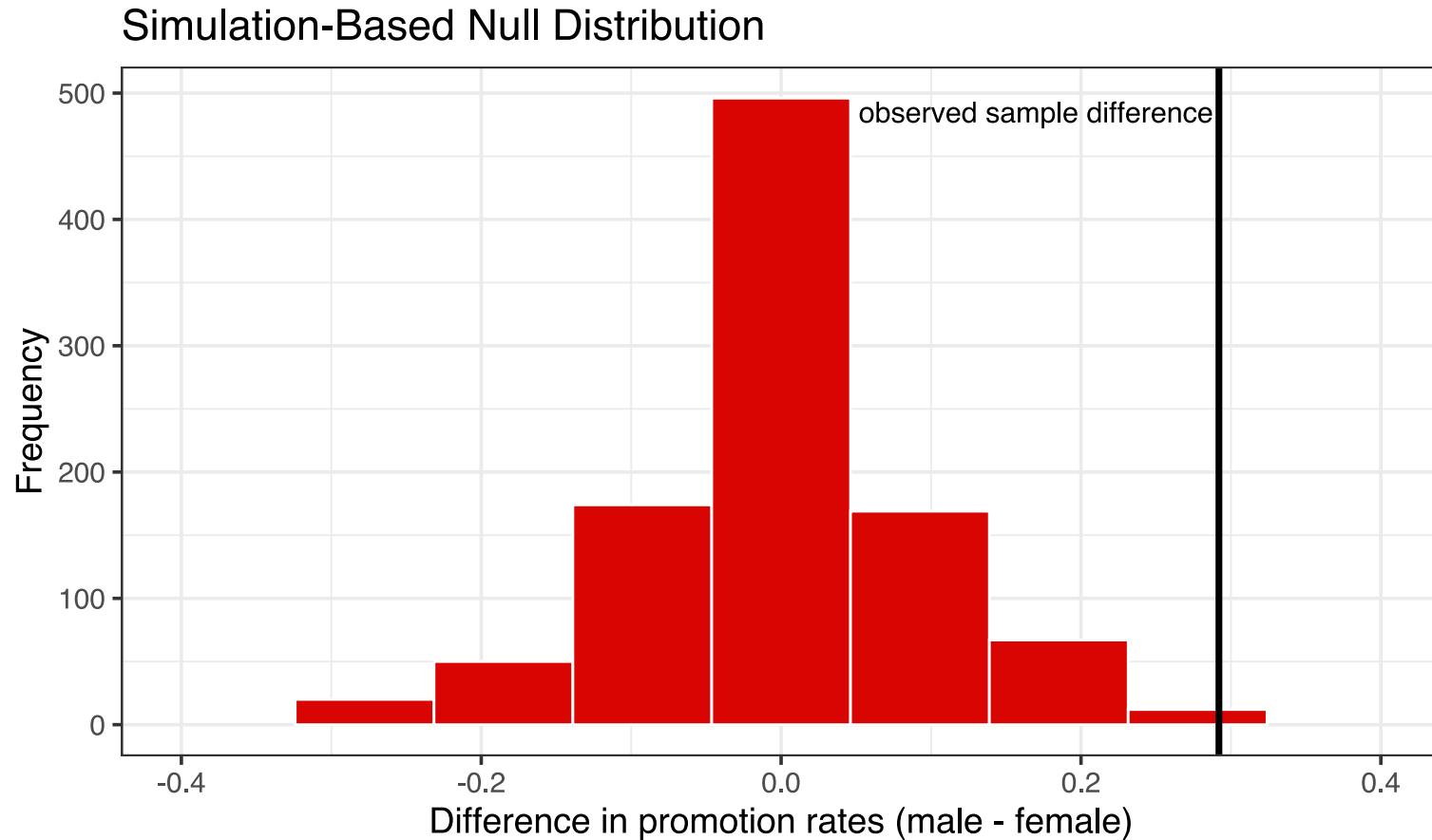
- In our hypothetical world, the difference in promotion rates was only 4.2 percentage points.
- Can we answer our initial question about the existence of gender discrimination now?
- No, we must investigate the role of *sampling variation!*
  - What if we reshuffle once again, how different from 4.2%p (*percentage points*) would the difference be?
  - In other words, how representative of that hypothetical world is 4.2%p?
  - How likely is a 29%p difference to occur in such a world?
- We need to know about the whole sampling distribution under the *no discrimination* hypothesis.
- How? Just by redoing the reshuffling a large number of times, and computing the difference each time.



# Sampling Distribution with 1000 Reshufflings



# Sampling Distribution with 1000 Reshufflings



How *likely* is it to observe a 0.292 difference in a world with no discrimination?



# What did we just do?

- We just demonstrated the statistical procedure known as *hypothesis testing* using a *permutation test*.



# What did we just do?

- We just demonstrated the statistical procedure known as *hypothesis testing* using a *permutation test*.
- The question is how likely the observed difference in promotion rates is to occur in a hypothetical universe with no discrimination.



# What did we just do?

- We just demonstrated the statistical procedure known as *hypothesis testing* using a *permutation test*.
- The question is how likely the observed difference in promotion rates is to occur in a hypothetical universe with no discrimination.
- We concluded *rather not*, i.e. we tended to *reject* the no discrimination hypothesis.



# What did we just do?

- We just demonstrated the statistical procedure known as *hypothesis testing* using a *permutation test*.
- The question is how likely the observed difference in promotion rates is to occur in a hypothetical universe with no discrimination.
- We concluded *rather not*, i.e. we tended to *reject* the no discrimination hypothesis.
- Let's introduce the formal framework of hypothesis testing now.



# Hypothesis Test Notation and Definitions

- A ***hypothesis test*** consists of a test between ***two competing hypotheses*** about the population parameter:



# Hypothesis Test Notation and Definitions

- A ***hypothesis test*** consists of a test between ***two competing hypotheses*** about the population parameter:
  - The ***null hypothesis*** ( $H_0$ ): generally hypothesis of no difference;



# Hypothesis Test Notation and Definitions

- A ***hypothesis test*** consists of a test between ***two competing hypotheses*** about the population parameter:
  - The ***null hypothesis*** ( $H_0$ ): generally hypothesis of no difference;
  - The ***alternative hypothesis*** ( $H_A$  or  $H_1$ ): the research hypothesis.



# Hypothesis Test Notation and Definitions

- A ***hypothesis test*** consists of a test between ***two competing hypotheses*** about the population parameter:
  - The ***null hypothesis*** ( $H_0$ ): generally hypothesis of no difference;
  - The ***alternative hypothesis*** ( $H_A$  or  $H_1$ ): the research hypothesis.
- In the previous example:

$$H_0 : p_m - p_f = 0$$

$$H_A : p_m - p_f > 0,$$

where  $p_m$  = promotion rate of men, and  $p_f$  = promotion rate of women.



# Hypothesis Test Notation and Definitions

- A ***hypothesis test*** consists of a test between ***two competing hypotheses*** about the population parameter:
  - The ***null hypothesis*** ( $H_0$ ): generally hypothesis of no difference;
  - The ***alternative hypothesis*** ( $H_A$  or  $H_1$ ): the research hypothesis.
- In the previous example:

$$H_0 : p_m - p_f = 0$$

$$H_A : p_m - p_f > 0,$$

where  $p_m$  = promotion rate of men, and  $p_f$  = promotion rate of women.

- Here, we considered a ***one-sided*** alternative, stating that  $p_m > p_f$ , i.e. women are discriminated against.
- The ***two-sided*** formulation is just  $H_A : p_m - p_f \neq 0$ .



# Hypothesis Test Notation and Definitions

- **Test statistic:** *point estimate/sample statistic formula used for hypothesis testing.*



# Hypothesis Test Notation and Definitions

- **Test statistic:** point estimate/sample statistic formula used for hypothesis testing.
  - In our previous case: difference in sample proportions  $\hat{p}_m - \hat{p}_f$ .



# Hypothesis Test Notation and Definitions

- **Test statistic:** point estimate/sample statistic formula used for hypothesis testing.
  - In our previous case: difference in sample proportions  $\hat{p}_m - \hat{p}_f$ .
- **Observed test statistic:** value of the test statistic that we observed in real life.



# Hypothesis Test Notation and Definitions

- **Test statistic:** point estimate/sample statistic formula used for hypothesis testing.
  - *In our previous case:* difference in sample proportions  $\hat{p}_m - \hat{p}_f$ .
- **Observed test statistic:** value of the test statistic that we observed in real life.
  - *In our previous case:* observed difference  $\hat{p}_m - \hat{p}_f = 0.292 = 29.2$  percentage points.



# Hypothesis Test Notation and Definitions

- **Test statistic:** point estimate/sample statistic formula used for hypothesis testing.
  - *In our previous case:* difference in sample proportions  $\hat{p}_m - \hat{p}_f$ .
- **Observed test statistic:** value of the test statistic that we observed in real life.
  - *In our previous case:* observed difference  $\hat{p}_m - \hat{p}_f = 0.292 = 29.2$  percentage points.
- **Null distribution:** sampling distribution of the test statistic *assuming the null hypothesis  $H_0$  is true.*

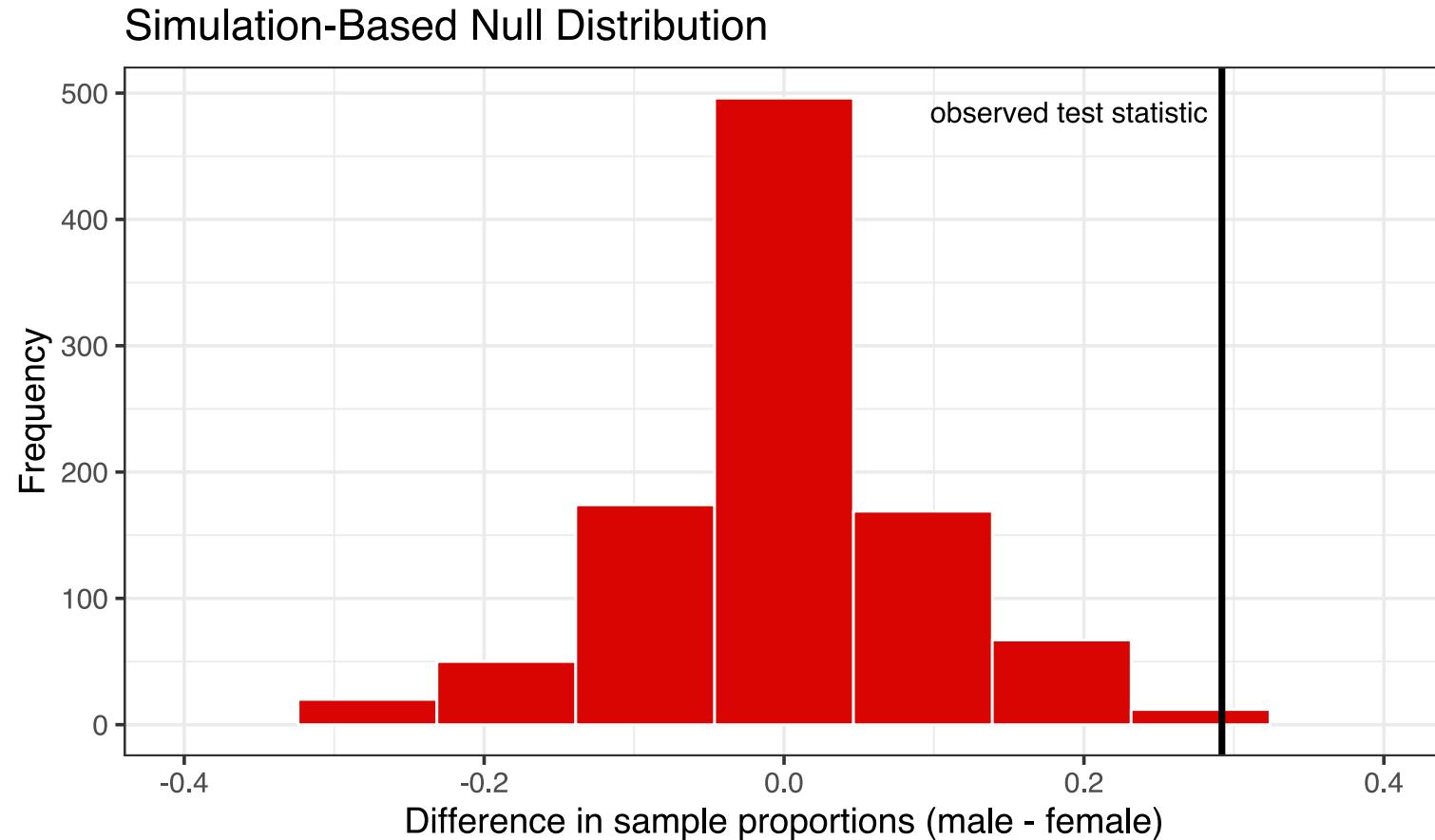


# Hypothesis Test Notation and Definitions

- **Test statistic:** point estimate/sample statistic formula used for hypothesis testing.
  - *In our previous case:* difference in sample proportions  $\hat{p}_m - \hat{p}_f$ .
- **Observed test statistic:** value of the test statistic that we observed in real life.
  - *In our previous case:* observed difference  $\hat{p}_m - \hat{p}_f = 0.292 = 29.2$  percentage points.
- **Null distribution:** sampling distribution of the test statistic *assuming the null hypothesis  $H_0$  is true.*
  - *In our previous case:* All the possible values that  $\hat{p}_m - \hat{p}_f$  can take assuming there is no discrimination.
  - That's the distribution we have seen just before.



# Null Distribution



# Hypothesis Test Notation and Definitions

**p-value:** probability of observing a test statistic *just as or more extreme* than the one we obtained, assuming the null hypothesis  $H_0$  is true. 🤔



# Hypothesis Test Notation and Definitions

**p-value:** probability of observing a test statistic *just as or more extreme* than the one we obtained, assuming the null hypothesis  $H_0$  is true. 🤔

- How surprised are we that we observed a difference in promotions rates of 0.292 in our sample assuming  $H_0$  is true, that is a world without discrimination? Very surprised? Kind of surprised?



# Hypothesis Test Notation and Definitions

**p-value:** probability of observing a test statistic *just as or more extreme* than the one we obtained, assuming the null hypothesis  $H_0$  is true. 🤔

- How surprised are we that we observed a difference in promotions rates of 0.292 in our sample assuming  $H_0$  is true, that is a world without discrimination? Very surprised? Kind of surprised?
- What do we mean by **more extreme**?
  - Defined in terms of the alternative hypothesis: in this case, men are **more likely** to be promoted than women. Therefore, **more extreme** in our case means observing a difference in promotion rates **greater than 0.292**.



# Hypothesis Test Notation and Definitions

**p-value:** probability of observing a test statistic *just as or more extreme* than the one we obtained, assuming the null hypothesis  $H_0$  is true. 🤔

- How surprised are we that we observed a difference in promotions rates of 0.292 in our sample assuming  $H_0$  is true, that is a world without discrimination? Very surprised? Kind of surprised?
- What do we mean by **more extreme**?
  - Defined in terms of the alternative hypothesis: in this case, men are **more likely** to be promoted than women. Therefore, **more extreme** in our case means observing a difference in promotion rates **greater than 0.292**.
- **Interpretation:** the lower the p-value, the *less consistent our null hypothesis is with the observed statistic.*



# Hypothesis Test Notation and Definitions

**p-value:** probability of observing a test statistic *just as or more extreme* than the one we obtained, assuming the null hypothesis  $H_0$  is true. 🤔

- How surprised are we that we observed a difference in promotions rates of 0.292 in our sample assuming  $H_0$  is true, that is a world without discrimination? Very surprised? Kind of surprised?
- What do we mean by **more extreme**?
  - Defined in terms of the alternative hypothesis: in this case, men are **more likely** to be promoted than women. Therefore, **more extreme** in our case means observing a difference in promotion rates **greater than 0.292**.
- **Interpretation:** the lower the p-value, the *less consistent our null hypothesis is with the observed statistic.*
- When do we decide to **reject**  $H_0$  or not?



# Hypothesis Test Notation and Definitions

- To decide whether we reject  $H_0$  or not, we set a *significance level* for the test.



# Hypothesis Test Notation and Definitions

- To decide whether we reject  $H_0$  or not, we set a **significance level** for the test.
- **Significance level ( $\alpha$ )**: acts as a *cutoff* on the p-value.
  - Common values are  $\alpha = 0.01, 0.05$ , or  $0.1$ .



# Hypothesis Test Notation and Definitions

- To decide whether we reject  $H_0$  or not, we set a **significance level** for the test.
- **Significance level ( $\alpha$ ):** acts as a *cutoff* on the p-value.
  - Common values are  $\alpha = 0.01, 0.05$ , or  $0.1$ .
- **Decision:**
  - If the p-value falls ***below the cutoff  $\alpha$*** , we "***reject the null hypothesis at the significance level  $\alpha$*** ."



# Hypothesis Test Notation and Definitions

- To decide whether we reject  $H_0$  or not, we set a **significance level** for the test.
- **Significance level ( $\alpha$ ):** acts as a *cutoff* on the p-value.
  - Common values are  $\alpha = 0.01, 0.05$ , or  $0.1$ .
- **Decision:**
  - If the p-value falls *below the cutoff  $\alpha$* , we "*reject the null hypothesis at the significance level  $\alpha$* ."
  - Alternatively, if the p-value is *greater than  $\alpha$* , we say that we "*fail to reject the null hypothesis  $H_0$  at the significance level  $\alpha$* ."



# Hypothesis Test Notation and Definitions

- To decide whether we reject  $H_0$  or not, we set a **significance level** for the test.
- **Significance level ( $\alpha$ ):** acts as a *cutoff* on the p-value.
  - Common values are  $\alpha = 0.01, 0.05$ , or  $0.1$ .
- **Decision:**
  - If the p-value falls *below the cutoff  $\alpha$* , we "**reject the null hypothesis at the significance level  $\alpha$ .**"
  - Alternatively, if the p-value is *greater than  $\alpha$* , we say that we "**fail to reject the null hypothesis  $H_0$  at the significance level  $\alpha$ .**"
- **Interpretation:** If what we observe is *too unlikely to happen* under the null hypothesis, it means that this hypothesis is **likely to be false**.

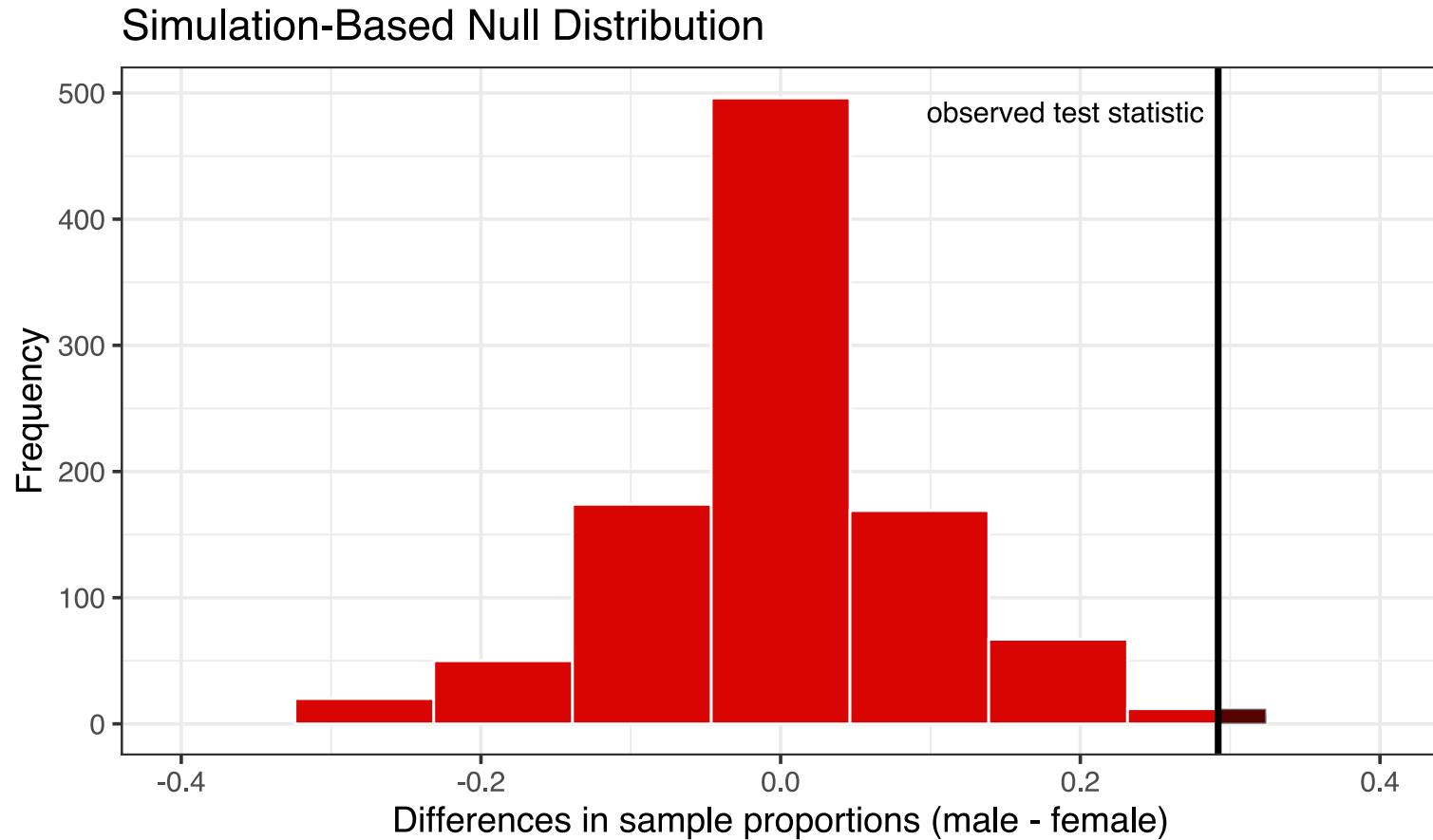


# Hypothesis Test Notation and Definitions

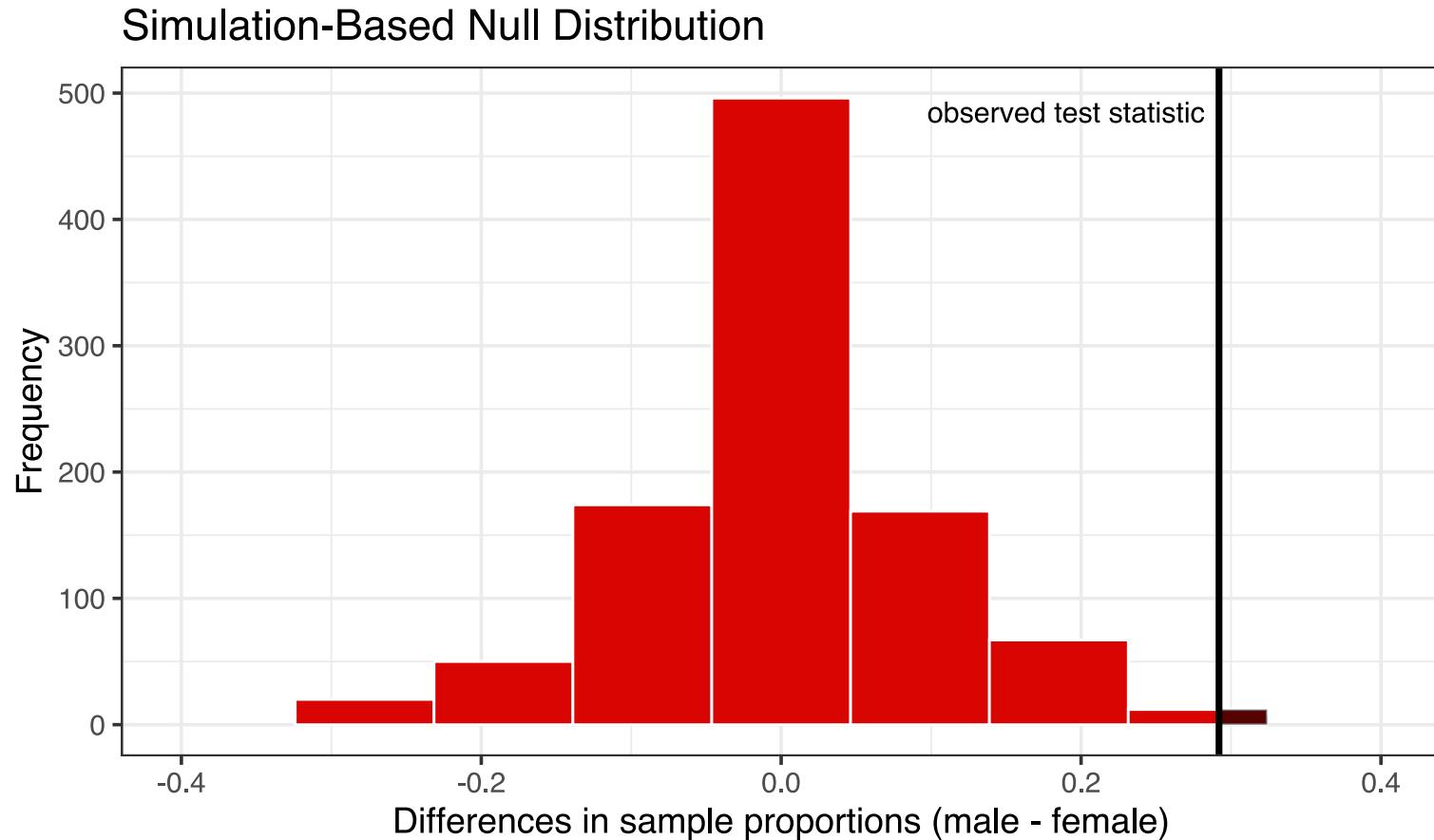
- To decide whether we reject  $H_0$  or not, we set a **significance level** for the test.
- **Significance level ( $\alpha$ ):** acts as a *cutoff* on the p-value.
  - Common values are  $\alpha = 0.01, 0.05$ , or  $0.1$ .
- **Decision:**
  - If the p-value falls *below the cutoff  $\alpha$* , we "**reject the null hypothesis at the significance level  $\alpha$ .**"
  - Alternatively, if the p-value is *greater than  $\alpha$* , we say that we "**fail to reject the null hypothesis  $H_0$  at the significance level  $\alpha$ .**"
- **Interpretation:** If what we observe is *too unlikely to happen* under the null hypothesis, it means that this hypothesis is **likely to be false**.
- Let's illustrate how it works in our example.



# Visualizing the P-value



# Visualizing the P-value



The shaded area corresponds to the p-value!



# Obtaining the p-value and Deciding

- Recall the definition of the p-value: *probability of observing a test statistic just as or more extreme than the one we obtained, assuming the null hypothesis  $H_0$  is true.*



# Obtaining the p-value and Deciding

- Recall the definition of the p-value: *probability of observing a test statistic just as or more extreme than the one we obtained, assuming the null hypothesis  $H_0$  is true.*

```
p_value <- mean(null_distribution$stat >= 0.292)
p_value
## [1] 0.007
```

- In a world without discrimination, we would get  $\hat{p}_m - \hat{p}_f$  superior (or equal) to 0.292 only 0.7% of the time.



# Obtaining the p-value and Deciding

- Recall the definition of the p-value: *probability of observing a test statistic just as or more extreme than the one we obtained, assuming the null hypothesis  $H_0$  is true.*

```
p_value <- mean(null_distribution$stat >= 0.292)
p_value
## [1] 0.007
```

- In a world without discrimination, we would get  $\hat{p}_m - \hat{p}_f$  superior (or equal) to 0.292 only 0.7% of the time.
- So, we can reject  $H_0$ , i.e. the absence of discrimination, at the 5% significance level.
  - We also say that  $\hat{p}_m - \hat{p}_f = 0.292$  is *statistically significantly different from 0* at the 5% level.



# Obtaining the p-value and Deciding

- Recall the definition of the p-value: *probability of observing a test statistic just as or more extreme than the one we obtained, assuming the null hypothesis  $H_0$  is true.*

```
p_value <- mean(null_distribution$stat >= 0.292)
p_value
## [1] 0.007
```

- In a world without discrimination, we would get  $\hat{p}_m - \hat{p}_f$  superior (or equal) to 0.292 only 0.7% of the time.
- So, we can reject  $H_0$ , i.e. the absence of discrimination, at the 5% significance level.
  - We also say that  $\hat{p}_m - \hat{p}_f = 0.292$  is *statistically significantly different from 0* at the 5% level.
- Question:** Suppose we had set  $\alpha = 0.01 = 1\%$ , would we have rejected the absence of discrimination at this level?



# Testing Errors

Working with probabilities implies that sometimes, we make **errors**.



# Testing Errors

Working with probabilities implies that sometimes, we make **errors**.

- A 29% $p$  difference may be *unlikely* under  $H_0$ , but that **doesn't mean it's impossible to occur**.
  - In fact, such a difference (or higher) would occur (approximately) in 0.007% of cases.



# Testing Errors

Working with probabilities implies that sometimes, we make **errors**.

- A 29% p difference may be *unlikely* under  $H_0$ , but that **doesn't mean it's impossible to occur**.
  - In fact, such a difference (or higher) would occur (approximately) in 0.007% of cases.
- So, it may happen that we sometimes reject  $H_0$ , when in fact it was true.
  - Setting 5% significance level, you make sure it won't happen more than 5% of the time.



# Testing Errors

In hypothesis testing, there are **two types of errors**:

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

**Type I error:** reject the null hypothesis when in fact it was true. **false positive**

**Type II error:** don't reject the null hypothesis when in fact it was false. **false negative**

- In practice, we choose the frequency of a Type I error by setting  $\alpha$  and try to minimize the type II error.



# How does all of this relate to regression analysis?

- Now you have all the tools to make *statistical inference* for real!



# How does all of this relate to regression analysis?

- Now you have all the tools to make *statistical inference* for real!
- Regression analysis is based on a *sample* of data.



# How does all of this relate to regression analysis?

- Now you have all the tools to make *statistical inference* for real!
- Regression analysis is based on a *sample* of data.
- So your *regression coefficient* is subject to *sampling variation*, it's not the true population coefficient.



# How does all of this relate to regression analysis?

- Now you have all the tools to make *statistical inference* for real!
- Regression analysis is based on a *sample* of data.
- So your *regression coefficient* is subject to *sampling variation*, it's not the true population coefficient.
- *Question:* Is the estimated effect statistically significantly different from some value  $z$ ?



# How does all of this relate to regression analysis?

- Now you have all the tools to make **statistical inference** for real!
- Regression analysis is based on a **sample** of data.
- So your **regression coefficient** is subject to **sampling variation**, it's not the true population coefficient.
- **Question:** Is the estimated effect statistically significantly different from some value  $z$ ?
- The answer in the next episode of *Introduction to Econometrics with R*! 🧑



# THANKS

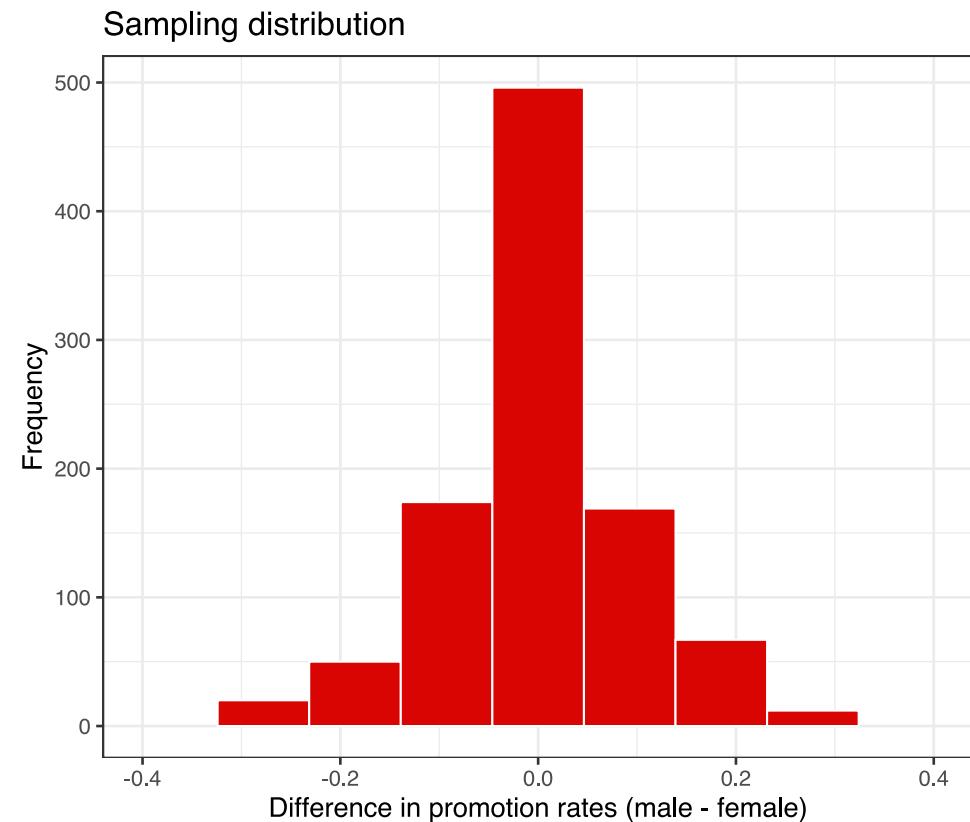
To the amazing **moderndive** team!



# Appendix: code to generate the null distribution

```
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"))

visualize(null_distribution,
          bins = 10,
          fill = "#d90502") +
  labs(title = "Sampling distribution",
       x = "Difference in promotion rates (male - female)",
       y = "Frequency") +
  xlim(-0.4, 0.4) +
  theme_bw(base_size = 14)
```



END

---

✉ florian.oswald@sciencespo.fr

🔗 Slides

🔗 Book

🐦 @ScPoEcon

Ⓜ️ @ScPoEcon

---

