Hypotheses testing

09.06.2022, Data Science (SpSe 2022): T16

Prof. Dr. Claudius Gräbner-Radkowitsch
Europa-University Flensburg, Department of Pluralist Economics
www.claudius-graebner.com | @ClaudiusGraebner | claudius@claudius-graebner.com

Europa-Universität Flensburg

Europa-Universität Flensburg
International Institute of Management and Economic Education
Department of Pluralist Economics

Prologue:

Prologue
Feedback and exercises

- XX of you filled out the feedback survey. Main take-aways:
  - TBA
- What were the main problems with the exercises?

Learning Goals

- Understand the idea behind hypothesis testing
- Learn how to implement the hypothesis testing in R using `infer`
- Understand the relation between p-values, statistical significance, and confidence intervals

Motivation

Motivation

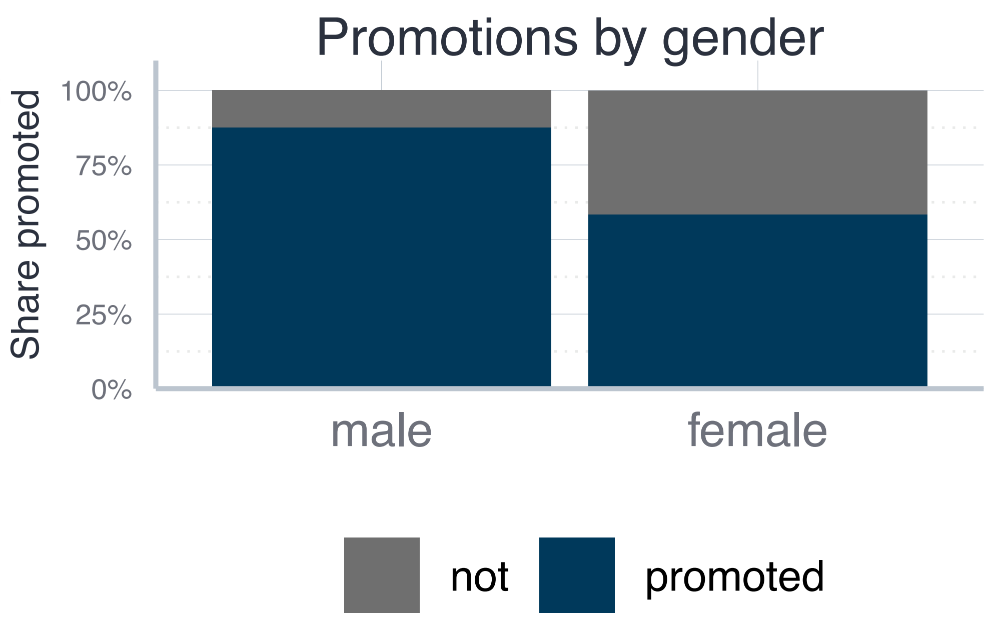
- We learned earlier that hypotheses and their test is an essential part of **scientific progress**
- We now learn how to test hypotheses quantitatively and how this relates to the idea of confidence intervals
- To this end, we will build directly on our knowledge about sampling
  - Hypothesis tests are meant to assess hypothesis using random samples
- To illustrate the idea we will start with an introductory example...
- ...then learn about the different steps of the hypothesis testing workflow....
- ...and then conclude with some remarks about interpreting hypothesis tests
- We will see numerous similarities to the computation of CI from last session🎉

Introductory example

Introductory example

- Does gender affect promotion? A study from the 1970s...
- Bank directors were given resumes of either men and women and needed to decide whether the quality for a promotion
  - Catch: the resumes were completely identical except the name
- First step: descriptive analysis of the data

```
> prom_data %>%
+   group_by(gender, decision) %>%
+   tally()
# A tibble: 4 x 3
# Groups:   gender [2]
  gender decision     n
  <fct>  <fct>    <int>
1 male   not          3
2 male   promoted    21
3 female not         10
4 female promoted    14
```



- It seems that women are less likely to get a promotion $(\Delta = 29.2\,\%)$
- But might this effect only appear due to sampling variation?

Introductory example

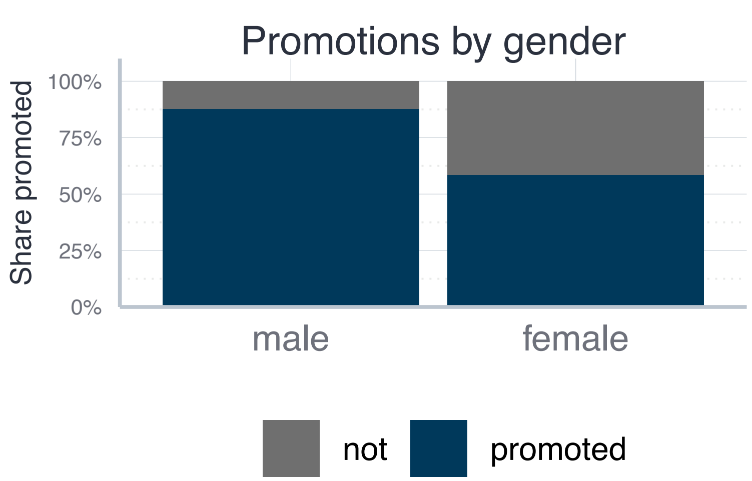
- When we ask whether the effect is due to sampling variation we are effectively asking the following:
  - Could it be that in reality there is no association between gender and promotion likelihood, but that we drew a sample in which this association exists?

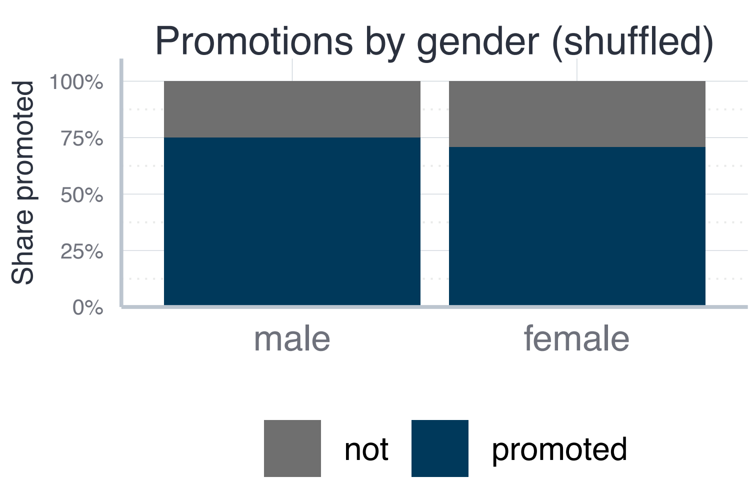
- In other words: is it likely to draw as sample such as ours in a world without gender discrimination?

- Unfortunately, we can explore this possibility only via a computer experiment 🤦🏽‍♀️

Introductory example

- What would be the association between gender and promotion likelihood in a world without gender discrimination?
  - Assume that not only gender has no association on promotions...
  - ...but also that no determinant of promotions is associated with gender
- Then there should be no association between gender and promotion!
- We could simulate this world by taking the promotion decision and re-shuffling the gender variable across observations → permutation





Much less of a difference, but its still just one sample!

Do many permutations and check how likely the original result would be → **permutation test**

Exercise 1: an MCS or a fair promotion world

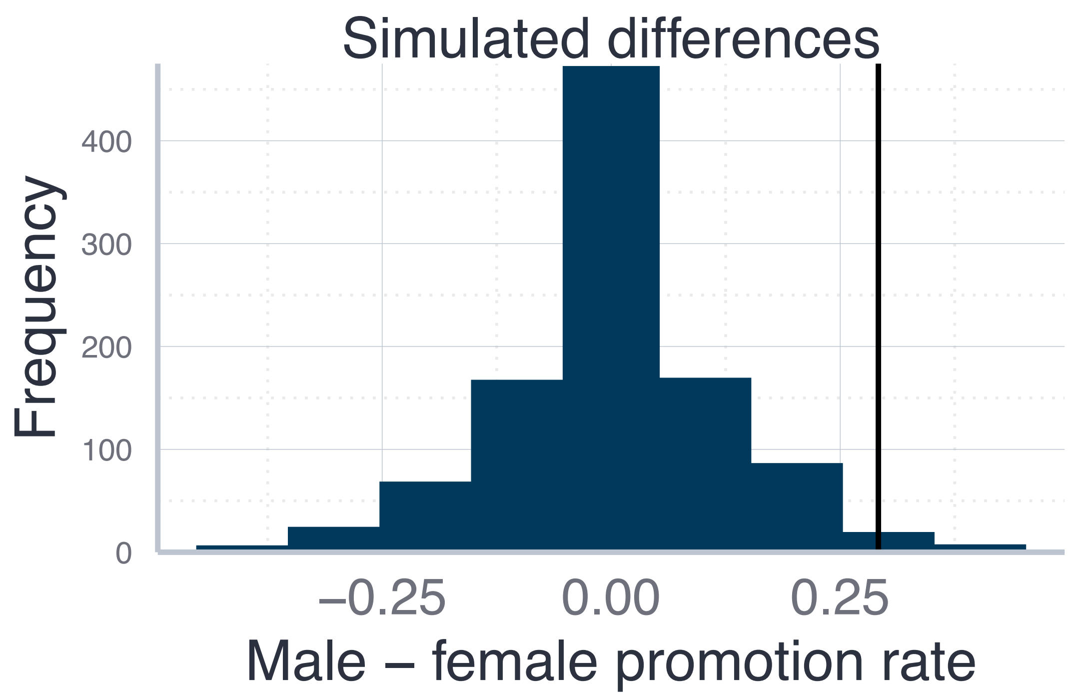
- For a rigorous permutation test we need to do the following:
  - Reshuffle the gender category
  - Compute the difference between promotion rates for men and women
  - Repeat the process for 1000 times and visualise differences
- Then we can check how likely our original difference of 29.2 % would be in a world without gender discrimination

- **Your turn:**
  - Take the code snippet on the right as a starting point and conduct an MCS as described above!

```
prom_data_shuffled <- prom_data %>%
  dplyr::mutate(
    gender_shuffled = sample(gender)
    ) %>%
    dplyr::group_by(
        gender_shuffled, decision) %>%
    dplyr::tally() %>%
    dplyr::mutate(prop=prop.table(n))
```

Exercise 1: an MCS or a fair promotion world

- This experiments indicates that in a world without gender discrimination as defined above...
  - ...it would be *very unlikely* to get a sample as we did



- Given this, it seems rather **implausible** that we are really living in a world without gender discrimination

- In other world: the hypothesis that there is no gender discrimination enjoys little evidence

- This is the fundamental idea behind hypothesis tests

Summary and take-away from the example

- We collected a sample on the promotions received by men and women
- We want to assess the hypothesis that men are more likely to get a promotion than women
- While more men than women received promotion in our sample, this is no conclusive evidence → difference might be due to sampling variation
- We conducted a permutation test by computing the probability to draw our sample in a world in which men and women are equally likely to get promoted
- By simulating this process, we found out that it would have been extremely unlikely to draw a sample such as ours if there were no gender discrimination
- We concluded that the sample provides evidence for the existence of discrimination

Workflow for hypothesis testing

The workflow using `infer`

- A more comprehensive workflow would make use of the package infer and is very similar to the one for computing confidence intervals:
  1. Specify the relevant variables using `infer::specify()`
  2. Explicate the underlying hypothesis using infer::`hypothesize()`
  3. Generate hypothetical data sets using `infer::generate()`
  4. Analyse the data sets and compute a null distribution using `infer::calculate()`
  5. Visualize and/or quantify the results
- We now go through all steps using the same example as above
  - Then we compare hypothesis testing and computing confidence intervals

The workflow using `infer`

1. Specify the relevant variables

```
> head(prom_data, 2)
# A tibble: 2 × 3
     id decision gender
  <int> <fct>    <fct>
1     1 promoted male
2     2 promoted male
```

- At this stage we need to specify the variables of interest
  - In contrast to last session, we now have an explanatory variable:

```
prom_data %>%
    infer::specify(
        formula = decision ~ gender,
        success = "promoted"
  )
```

The response variable

The explanatory variable

Value that counts as success

Same notation as in `lm()`

  - The result is identical to the initial data set, except some meta data:

```
Response: decision (factor)
Explanatory: gender (factor)
# A tibble: 48 × 2
  decision gender
  <fct>    <fct>
1 promoted male
2 promoted male
```

The workflow using `infer`

2. Explicate the underlying hypothesis

- At this stage we explicate the hypothesis that we want to test
  - This hypothesis is called **Null hypothesis** and denoted H_0
- This hypothesis determines the imagined world against which we compare our actually obtained sample → a world without gender discrimination
  - Its standard to have the Null hypothesis referring to a situation where there is no effect, or a relationship is absent
  - $H_0: p_m - p_f = 0$

```
...
    ) %>%
infer::hypothesize(
    null = "independence")
```

"independence" if we want to test whether one variable is independent of another one that determines groups

"point" if we want to test whether one variable corresponds to a particular point value

The workflow using `infer`

3. Generate hypothetical data

- We now generate hypothetical data as if H_0 were true

```
...
    ) %>%
infer::generate(
    reps = 1000,
    type = "permute")
```

The number of permutations to be computed

We want to compute permutations (without replacement), not conducting a bootstrap (with replacement)

- This results in a new tibble with `reps` times n rows:

```
Response: decision (factor)
Explanatory: gender (factor)
Null Hypothesis: independence
# A tibble: 48,000 × 3
# Groups:   replicate [1,000]
   decision gender replicate
   <fct>    <fct>      <int>
 1 promoted male           1
 2 promoted male           1
```

Now we need to analyse the 1000 samples!

The workflow using `infer`

4. Analyse the data and compute null distribution

- For the 1000 iterations we need to compute the adequate summary statistic
  - This means we compute our **sample statistic**, which in the context of hypothesis testing is called a **test statistic**
  - The relevant test statistic is determined by the population statistic of interest
  - Here the latter is $p_m - p_f$, so we need to compute $\hat{p}_m - \hat{p}_f$:

```
null_distribution <- ...
    ) %>%
infer::calculate(
    stat = "diff in props",
    order = c("male", "female"))
)
```

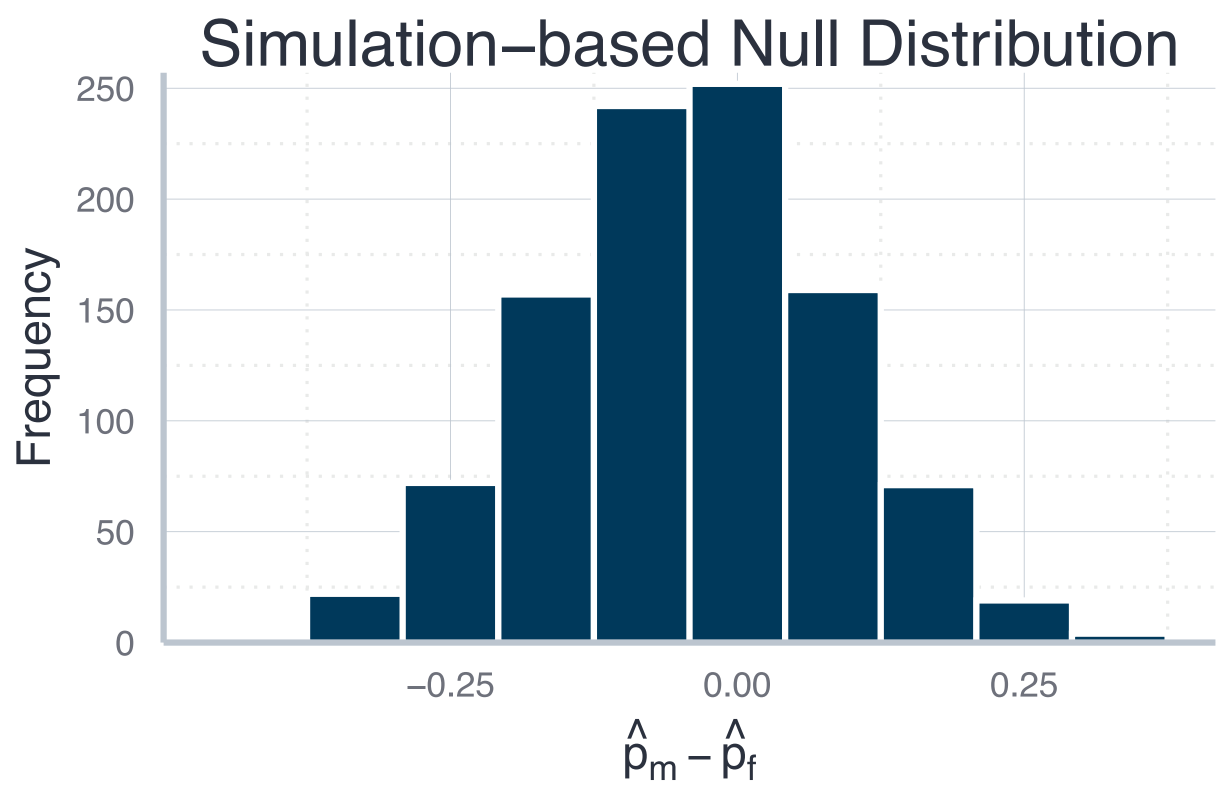
The test statistic of interest (alternatives would be `mean`, `median`, `prop`, etc.)

The order for subtraction (or comparable) operations

  - This creates a distribution as if H_0 were true → **Null distribution**

The workflow using `infer`

5. Visualize and quantify the results




```
Response: decision (factor)
Explanatory: gender (factor)
Null Hypothesis: independence
# A tibble: 1,000 × 2
  replicate    stat
      <int>   <dbl>
1         1 -0.0417
2         2 -0.125
3         3  0.125
4         4  0.208
```

- Given this distribution, what is the probability to observe $\hat{p}_m - \hat{p}_f = 0.292$ as in our actual sample? → This probability is called the **p-value**

> Probability to obtain a test statistic just as or more extreme than the actually observed test statistic, assuming H_0 is true

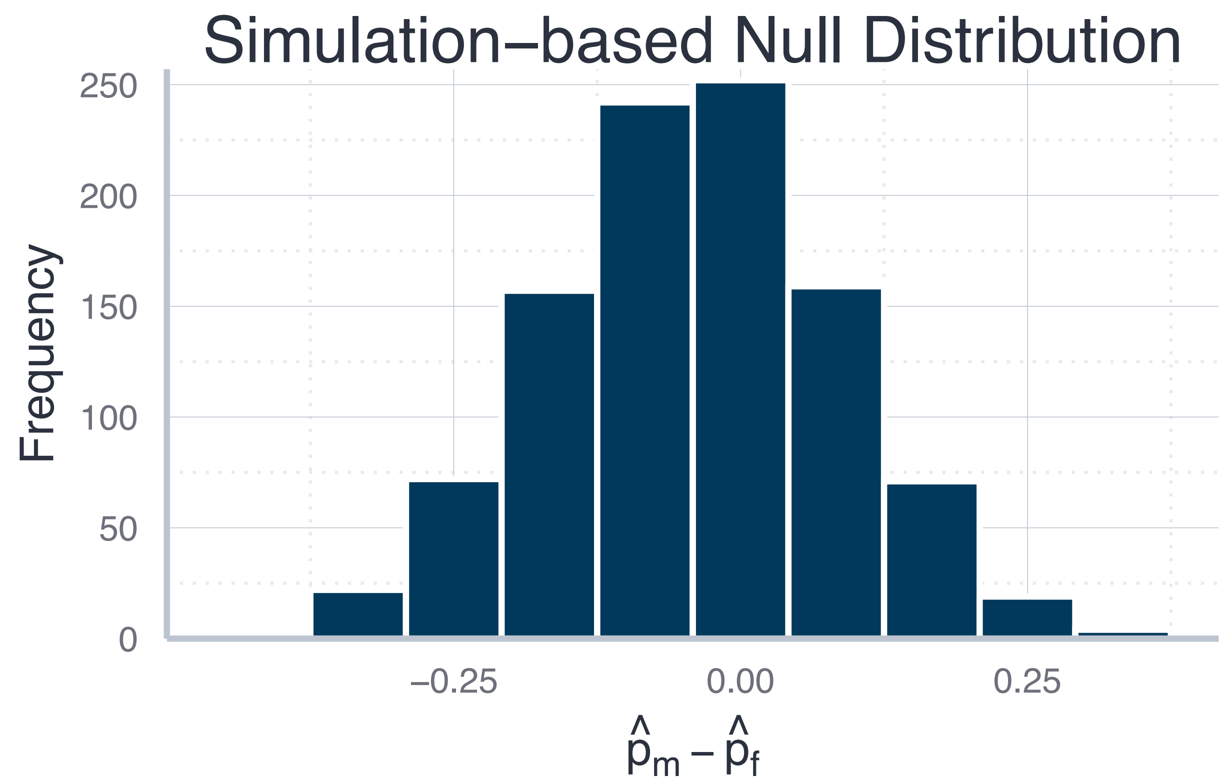
The workflow using `infer`

5. Visualize and quantify the results

p-value

Probability to obtain a test statistic just as or more extreme than the actually observed test statistic, assuming H_0 is true

```
null_distribution %>%
    infer::get_p_value(
        obs_stat = 0.292,
        direction = "greater"
)
```



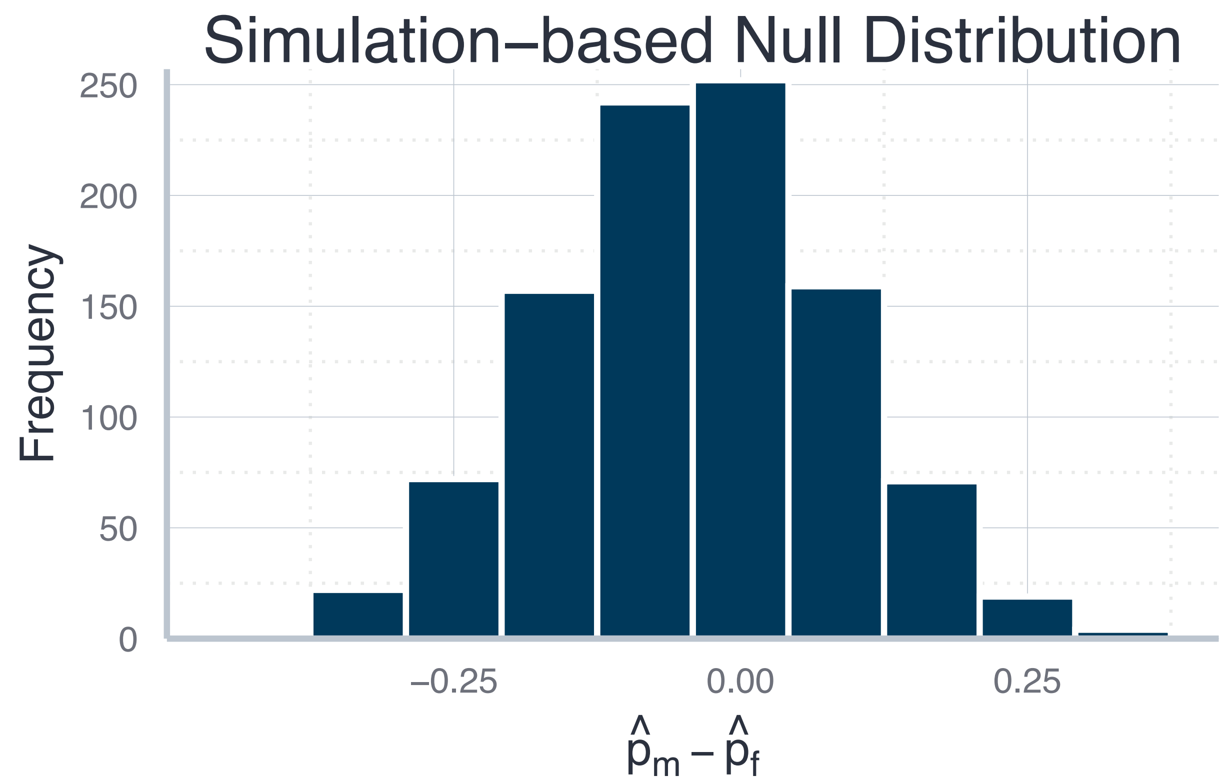
Depends on what we want to test H_0 against; here whether p_m is greater than p_f (alternative: `"smaller"` or `"two-sided"`)

The workflow using `infer`

5. Visualize and quantify the results

p-value

Probability to obtain a test statistic just as or more extreme than the actually observed test statistic, assuming H_0 is true




```
null_distribution %>%
    infer::get_p_value(
        obs_stat = 0.292,
        direction = "greater"
)
```

$$p = 0.023 = 2.3\,\%$$

If H_0 were true, the probability to draw a sample with a test statistic of 0.292 or higher is 2.3%.

- Very small p-values suggest that H_0 is quite implausible
    - We reject H_0 when p is below a pre-specified threshold α (the **significance level**)

P-Values and confidence intervals

- You might have recognised that the syntax to compute p-values and confidence intervals is very similar:

```
p_val <- prom_data %>%
  infer::specify(
    formula = decision ~ gender,
    success = "promoted"
  ) %>%
  infer::hypothesize(null = "independence") %>%
  infer::generate(
    reps = 1000,
    type = "permute") %>%
  infer::calculate(
    stat = "diff in props", order = c("male", "female")) %>%
  infer::get_p_value(
    obs_stat = obs_diff_prop,  direction = "right")
```

```
conf_intervals <- prom_data %>%
  infer::specify(
    formula = decision ~ gender,
    success = "promoted"
  ) %>%
  # infer::hypothesize(null = "independence") %>%
  infer::generate(
    reps = 1000,
    type = "bootstrap") %>% # <- changed from "permute"
  infer::calculate(
    stat = "diff in props", order = c("male", "female")) %>%
  get_confidence_interval(
    level = 0.95, type = "percentile")
```

- In our case $CI_{95\%} = \left[0.042; 0.525\right]$, so we are 95% confident that the true value is contained in this interval
  - Then we would reject $H_0 : \hat{p}_m - \hat{p}_f = 0$ since $CI_{95\%}$ does not contain 0

P-Values and confidence intervals

- In our case $CI_{95\%} = [0.042; 0.525]$, so we are 95% confident that the true value is contained in this interval
    - Then we would reject $H_0 : \hat{p}_m - \hat{p}_f = 0$ since $CI_{95\%}$ does not contain 0
- This bridge between hypothesis testing with p-values and CI is the **significance level**:
    - If we set $\alpha = 5\ \%$, then we reject H_0 when $p < 0.05$
    - This corresponds to the situation in which $0 \notin CI_{95\%}$
- We set our significance level based on theoretical considerations and according to conventions → usually 0.1%, 1%, 5% or 10%

Terminology

Hypothesis tests: terminology

A **hypothesis** is s statement about an unknown population parameter.

A **hypothesis test** is a test that aims to distinguish between two hypotheses. A **null hypothesis** H_0 of "no effect" or "no difference" and an **alternative hypothesis** H_1.
H_1 can refer to a **one-sided** or a **two-sided alternative**.

- **Example: Studying bank promotions**
  - Hypothesis: men get promoted more frequently than women
  - $H_0 : p_m - p_f = 0$ and the one-sided alternative $H_1 : p_m - p_f > 0$
  - If we chose a two-sided alternative we had $H_1 : p_m - p_f \neq 0$

Hypothesis tests: terminology

A **test statistic** is sample statistic used for hypothesis testing.

The **observed test statistic** is the sample statistic computed from our actually obtained sample.

- **Example: Studying bank promotions**
  - Observed test statistic: the difference $\hat{p}_m - \hat{p}_f = 29.2\,\%$ as computed from our sample with $n = 48$
  - Test statistic: The difference $\hat{p}_m - \hat{p}_f$ (but from the actual sample, or one of the re-samples)

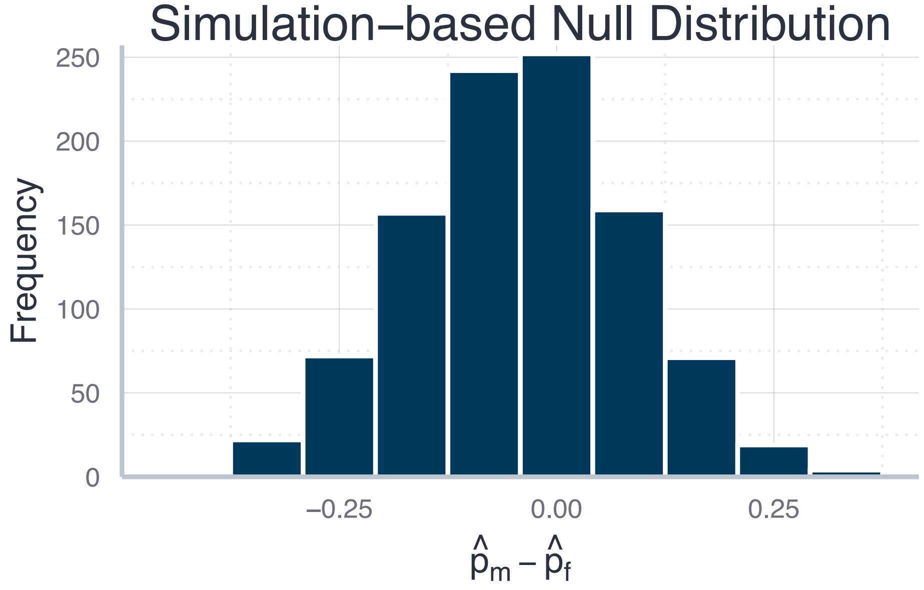
Hypothesis tests: terminology

A **Null distribution** is the sampling distribution of the test statistic under H_0.

This means it is a **hypothetical distribution** that is not informed by empirical observation.

It gives information about how the test statistic would vary due to sampling variation **if H_0 was true**.

- **Example: Studying bank promotions**
  - The Null distribution was obtained by generating 1000 permutations from the original sample...
  - ...and computing the test statistic for each sample



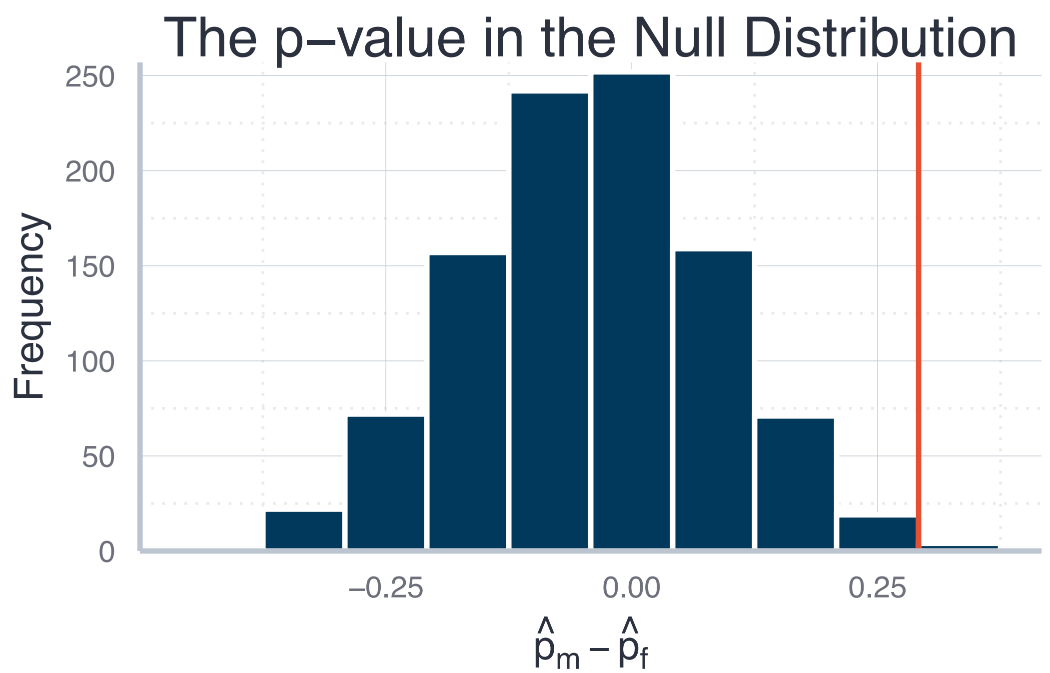
Hypothesis tests: terminology

The **p-value** is the probability to obtain a test statistic just as or more extreme than the actually observed test statistic, if H_0 was true.

The size of the p-value depends on the formulation of H_1 as one-sided or two-sided.

It could be interpreted as a **measure of surprise**: the smaller p, the more surprised we were to observe a test statistic.

- **Example: Studying bank promotions**
  - The probability to observe the difference $\hat{p}_m - \hat{p}_f = 29.2\,\%$ as computed from our actual sample was $p = 1.5\,\%$



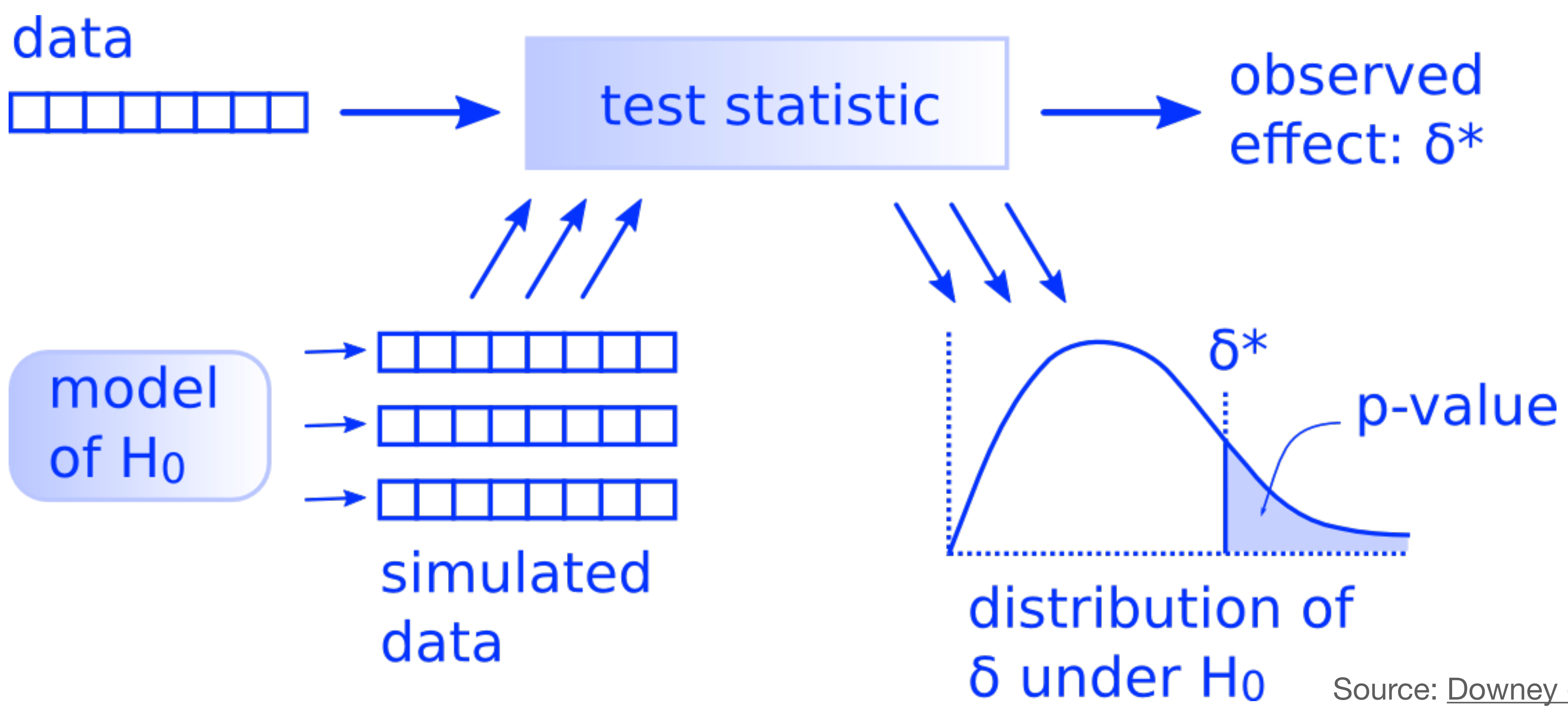
Hypothesis tests: terminology

The **p-value** is the probability to obtain a test statistic just as or more extreme than the actually observed test statistic, if H_0 was true.

The **significance level** α is a threshold that should be set before conducting the test. If the p-value falls below the level α one should reject H_0, if not, one speaks about "failing to reject H_0" (not: accept H_0).

- **Example: Studying bank promotions**
  - The p-value was 0.015. Thus, we would reject H_0 for the commonly used significance level 0.1 and 0.05, but not 0.01 and 0.001

A birds eye view on hypothesis tests



Source: Downey (2016)

Exercise 2:

- In the previous session, you computed a confidence intervals for the average height of EUF students using the data set

- Now we want to test the hypothesis that men and women differ in their height
  - $H_0 : \mu_m - \mu_f = 0, H_1 : \mu_m - \mu_f \neq 0$

- Compute the p-value using the workflow described above. For which confidence level can you reject H_0?

- Also compute the p-value when $H_1 : \mu_m - \mu_f > 0$; how do the two p-values differ?

Interpreting p-values

On the interpretation of p-values

- Before starting to implement a hypothesis test we should set α
  - But based on what should this decision be made?
- To answer this question, consider the two possible outcomes:

$p < \alpha$	$p \geq \alpha$
Reject H_0 in favour of H_1	Fail to reject H_0
Not: verify H_1!	Not: verify H_0!

On the interpretation of p-values

- Based on these considerations, a number of things could go wrong 🫣

	In reality...	H_0 is true	H_0 is false
Based on our sample we...	Fail to reject H_0	Correct 🍾	Type II error (or: false negative)
	Reject H_0	Type I error (or: false positive)	Correct 🍾

- We choose α by deciding on the **acceptable risk for a Type-I-error**

On the interpretation of p-values

- With α we set the probability for a Type-I-error explicitly
  - α is the significance level of the test
- The probability for a Type-II-error is denoted by β
  - $1-\beta$ is the power of the test
- When α goes down, so does $1-\beta$ → Trade-off between errors
- The conservative scientific culture tends to prioritise avoiding Type-I-errors

---

- A final word of caution: p-values are often misused in scientific and public communication
  - As Ismay & Kim (2022) I tend to agree that confidence intervals are usually a better way for communicating your results

Summary & outlook

Summary

- Vantage point: hypotheses and their test is an essential part of **scientific progress**
- We learned how to conduct a hypothesis test in R using infer
  - Formulate your Null hypothesis H_0 and alternative hypothesis H_1
  - Obtain a random sample from which you compute a test statistic
  - Generate a null distribution, which corresponds to the sampling distribution of the test statistic *if* H_0 *was true*
  - Assess the likelihood of the actual sample occurring under this setting: p-value
  - If the p-value is below the significance level of our test, reject it
- The process was syntactically similar to the computation of confidence intervals 🎉

Outlook

- Next session we will return to the method of regression analysis
    - Using the concepts of sampling theory and hypothesis testing we can qualify our regression results more precisely
    - We learn how to assess the adequateness of the regression assumptions
- We use the linear regression models for the purpose of prediction and explanation

Tasks until next week:

1. Fill in the **quick feedback survey** on Moodle
2. Read the **tutorials** and **lecture notes** posted on the course page
3. Do the **exercises** provided on the course page and **discuss problems** and difficulties via the Moodle forum