Hypothesis testing

with randomization

Prof. Dr. Jan Kirenz
HdM Stuttgart

Statistical inference is primarily concerned with understanding and quantifying the **uncertainty of parameter estimates**.

Sample vs population

- Our data is usually a (hopefully) **representative subset (sample)** of a larger population

- Sometimes the dataset at hand represents the **population** for the entire research question.

How different is one dataset from another?

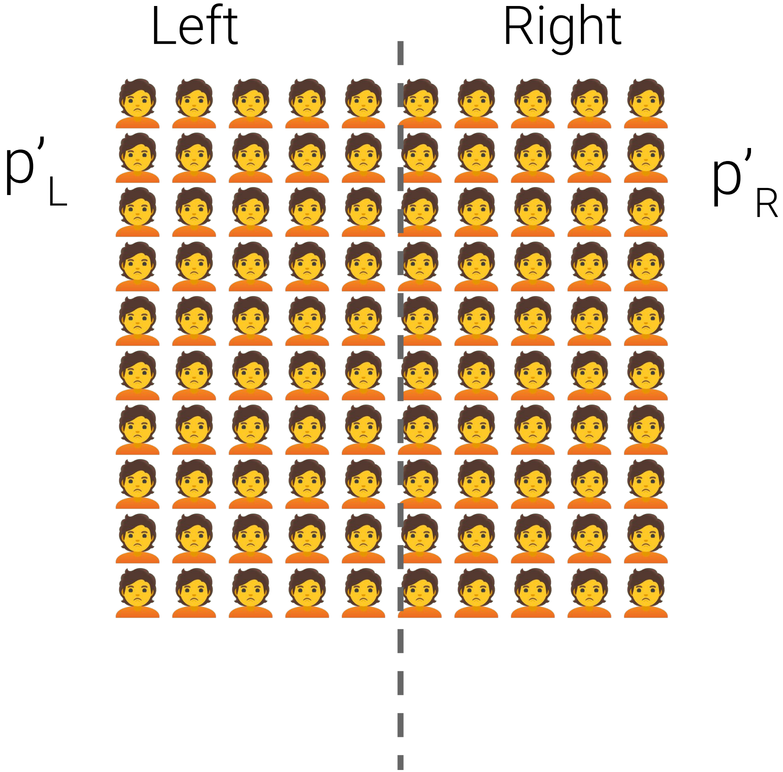
Notation for samples and population

- $\hat{p}$ and p’ sample proportion
- $\bar{x}$ and x’ sample mean

- p population proportion
- μ population mean

Proportion of students who prefer reading books on screen

- What are the variables?
- What data type?
- Levels of the variables?
- What is the explanatory variable?



- Variables and levels:
    - **side of the room**: left, right
    - **prefer to read books on screen**: yes; no
- Assumption about the relationship
    - A: independent
    - B: dependent

Studying **randomness** is a key focus of statistics.

Three different approaches for quantifying the **variability inherent in data**:

1. Randomization
2. bootstrapping
3. mathematical models

Randomization in experiments

Randomized experiments

- Explanatory variable (or treatment) is **randomly assigned** to the observational units.

- Used to assess whether or not one variable *(the explanatory variable)* **causes changes** in a second variable *(the response variable)*

Types of variability

1. **Causal** mechanism
    a. the randomized explanatory variable in the experiment

2. **Natural** variability
    a. inherent to the data

Sex discrimination case study

48 male bank supervisors

👴 👴 👴 👴 👴 👴 👴 👴 👴 👴
👴 👴 👴 👴 👴 👴 👴 👴 👴 👴
👴 👴 👴 👴 👴 👴 👴 👴 👴 👴
👴 👴 👴 👴 👴 👴 👴 👴 👴 👴
👴 👴 👴 👴 👴 👴 👴 👴

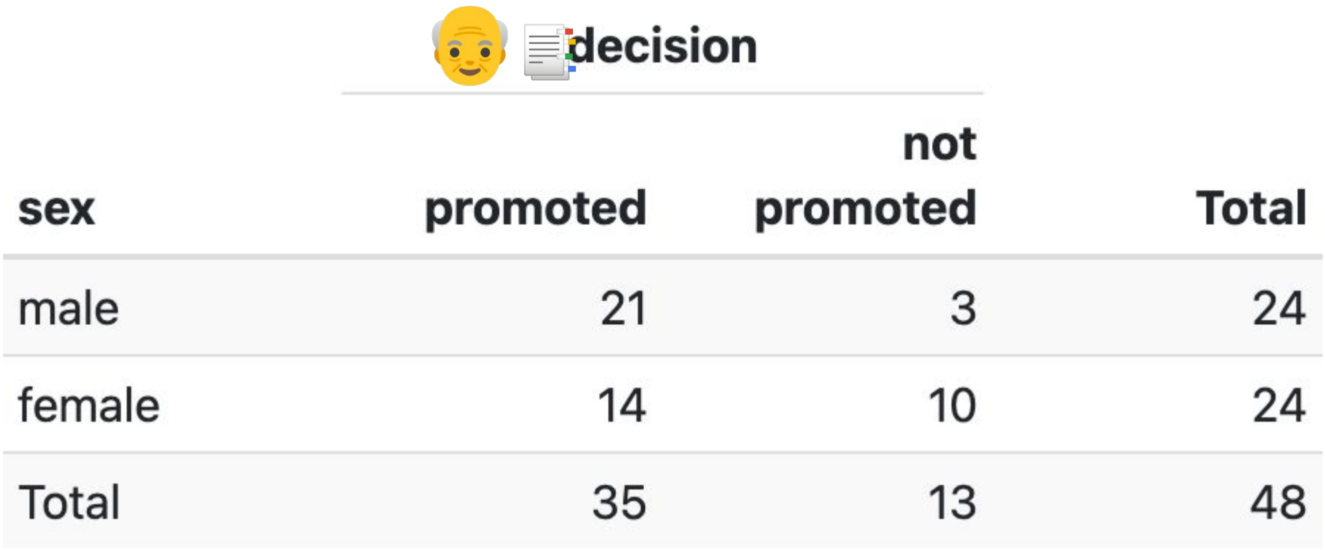
- Judge whether a person should be promoted to a branch manager position
- 📑 Personnel file
  - 👨 or 👩

📑 given to the participants were identical except ...

👩👩👩👩👩👩👩👩👩👩
👩👩👩👩👩👩👩👩👩👩
👩👩👩👩

👨👨👨👨👨👨👨👨👨👨
👨👨👨👨👨👨👨👨👨👨
👨👨👨👨

Summary results for the sex discrimination study

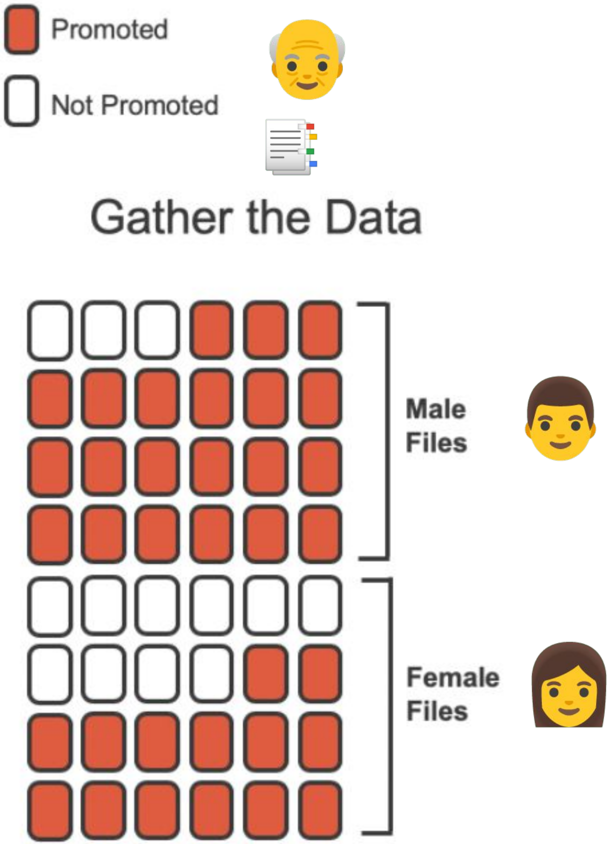


	👴 📑decision		
sex	promoted	not promoted	Total
male	21	3	24
female	14	10	24
Total	35	13	48

$p'_{MP} = 21/24 = 87{,}5\%$

$p'_{FP} = 14/24 = 58{,}3\%$

Source: Rosen and Jerdee (1974);



- **Difference in promotion rates**
  - $p'_{MP} - p'_{FP}$ =
  - 21/24 - 14/24 =
  - 29.2%
- We call the summary value (here: difference in promotion rates) the **statistic** of interest (or often the **test statistic**).

Null hypothesis: the difference is due to chance

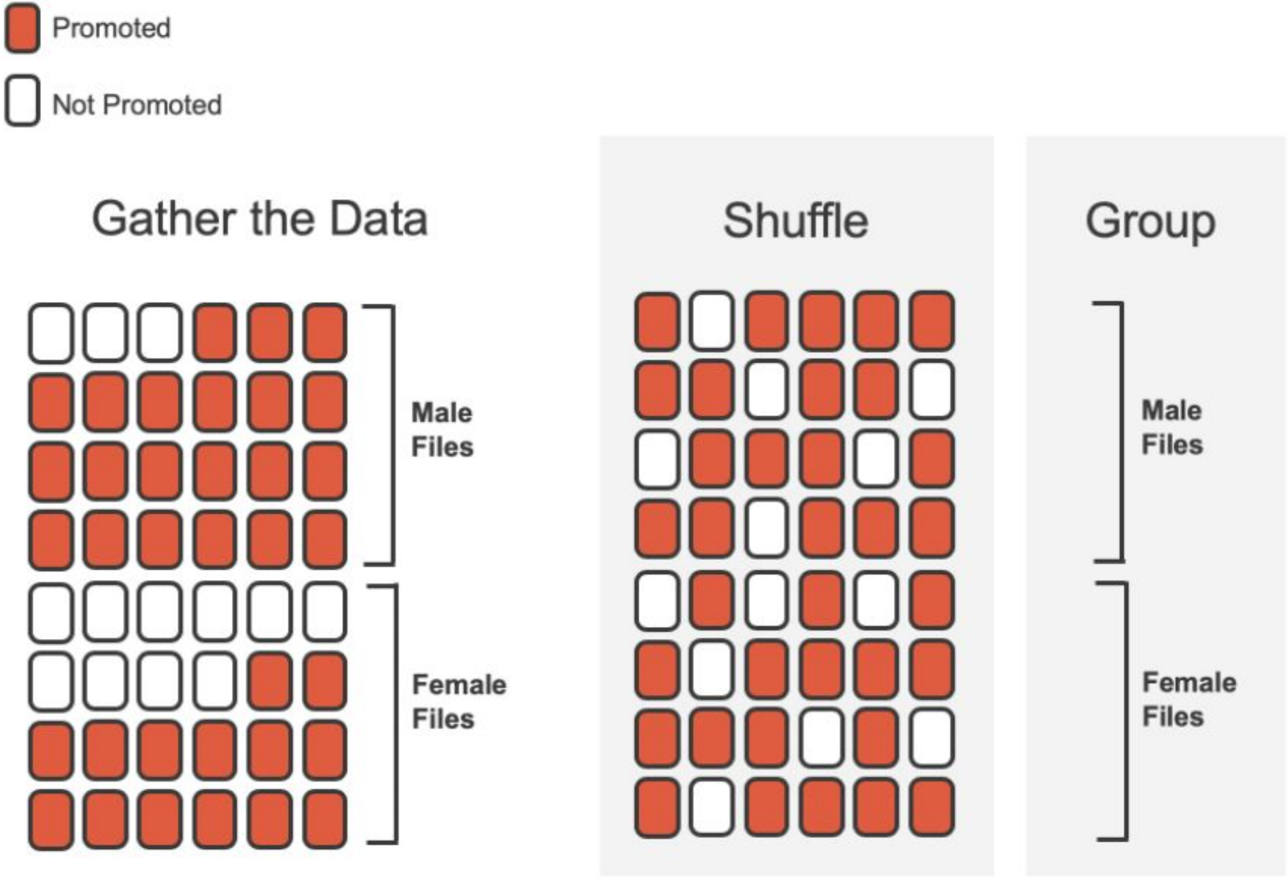
- **Null hypothesis** (H_0)
- The variables sex and decision are **independent**.
- They have **no relationship**
- The observed difference between the proportion of males and females who were promoted, 29.2%, was due to the **natural variability** inherent in the population.

Alternative hypothesis: the difference is <u>not</u> due to chance

- **Alternative hypothesis** (H_1)
- The variables sex and decision are **not independent**.
- There is a **relationship** between the variables
- The difference in promotion rates of 29.2% was **not due** to **natural variability**
  - equally qualified female personnel are less likely to be promoted than male personnel.

Randomization

- We **simulate** what would have happened if the bankers' decisions had been independent of sex but we had distributed the file sexes differently.
- In the simulation, we shuffle the 48 personnel files,
  - **35** labelled **promoted** and
  - **13** labelled **not promoted**, together
  - and we deal files into two new stacks.



🟥 Promoted

□ Not Promoted

Gather the Data

Male Files

Female Files

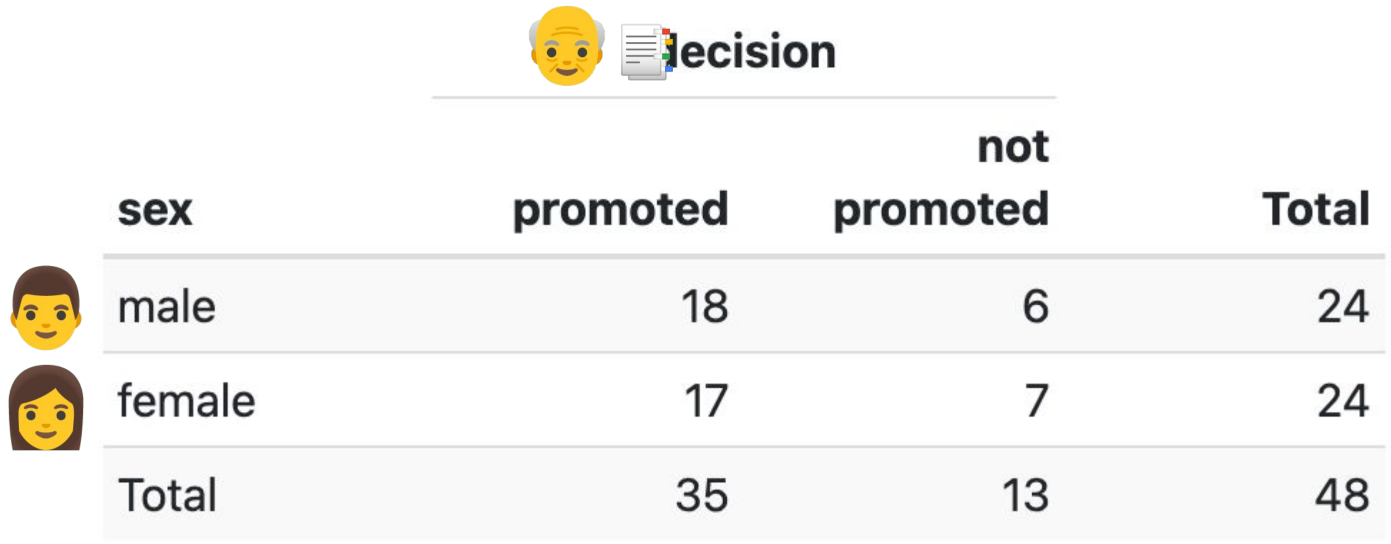
Shuffle

Group

Male Files

Female Files

Simulation results: difference in promotion rates between male and female is purely due to random chance



	👴 📑decision		
		not	
sex	promoted	promoted	Total
👨 male	18	6	24
👩 female	17	7	24
Total	35	13	48

$p'_{MP} = 18/24 = 75\%$

$p'_{FP} = 17/24 = 71\%$

What is the difference in promotion rates between the two simulated groups ?

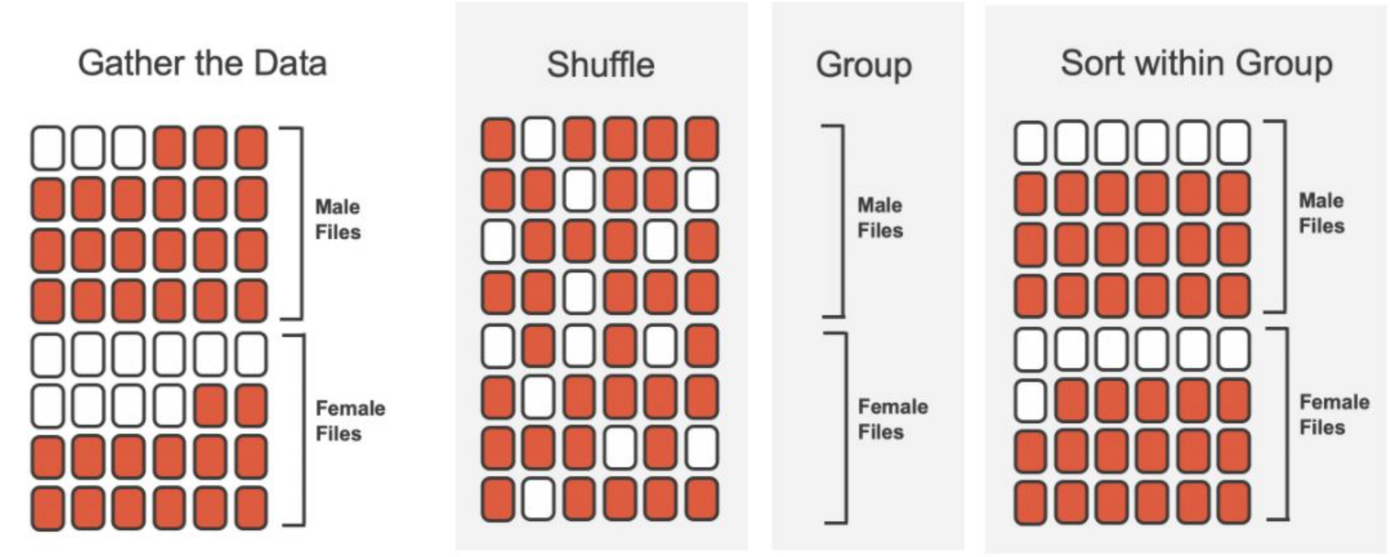
Difference between experiment and simulation

Experiment

$\hat{p}_M - \hat{p}_F = 0.292$

Randomization Shuffle #1

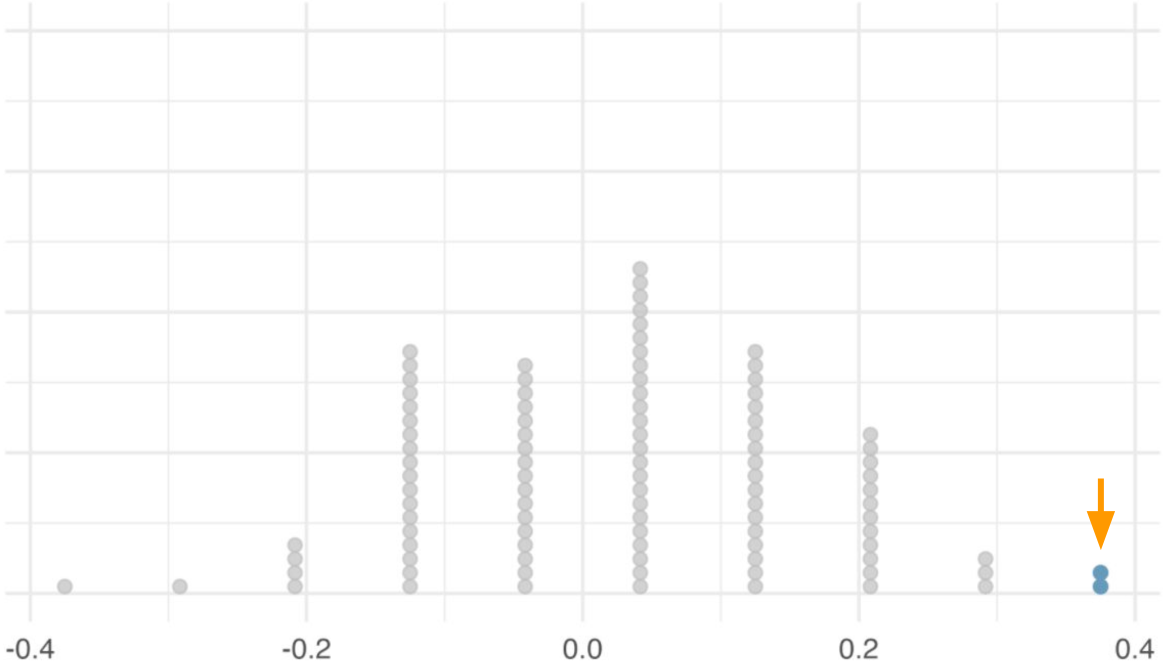
$\hat{p}_{M,shfl1} - \hat{p}_{F,shfl1} = 0.042$



Doing repeated simulations

- We use the computer to perform multiple simulations
- We repeat the simulation enough times that we have a good idea of the **shape of** the **distribution** of differences under the **null hypothesis**.

Differences in promotion rates (male-female) across 100 simulations



The difference of 29.2% is a rare event

Prof. Dr. Jan Kirenz

We reject a null position if the data strongly conflict with that null position

Null hypothesis is true?

Sex has no effect on promotion decision

Alternative hypothesis is true?

Sex has an effect on promotion decision

- Simulation result:
    - **2% probability** of obtaining a sample where ≥ 29.2% more male candidates than female candidates get promoted under the null hypothesis,
- We conclude
    - The data provide s**trong evidence of sex discrimination** against female candidates
    - We **reject the null hypothesis** in favor of the alternative

Statistical inference

- **Statistical inference:**
  - making decisions and
  - conclusions from data
  - in the context of **uncertainty**

- A given dataset (sample) may not always lead us to a correct conclusion
- Statistical inference gives us tools to control and evaluate **how often these errors occur.**

Opportunity cost case study

money not spent now
can be spent later

One-hundred and fifty students were recruited for the study, and each was given the following statement:

Imagine that you have been saving some extra money on the side to make some purchases, and on your most recent visit to the video store you come across a special sale on a new video. This video is one with your favorite actor or actress, and your favorite type of movie (such as a comedy, drama, thriller, etc.).

This particular video that you are considering is one you have been thinking about buying for a long time.

It is available for a special sale price of $14.99. What would you do in this situation? Please circle one of the options below

Control group and **treatment** group

A. Buy this entertaining video

B. Not buy this entertaining video

A. Buy this entertaining video.

B. Not buy this entertaining video. Keep the $14.99 for other purchases.

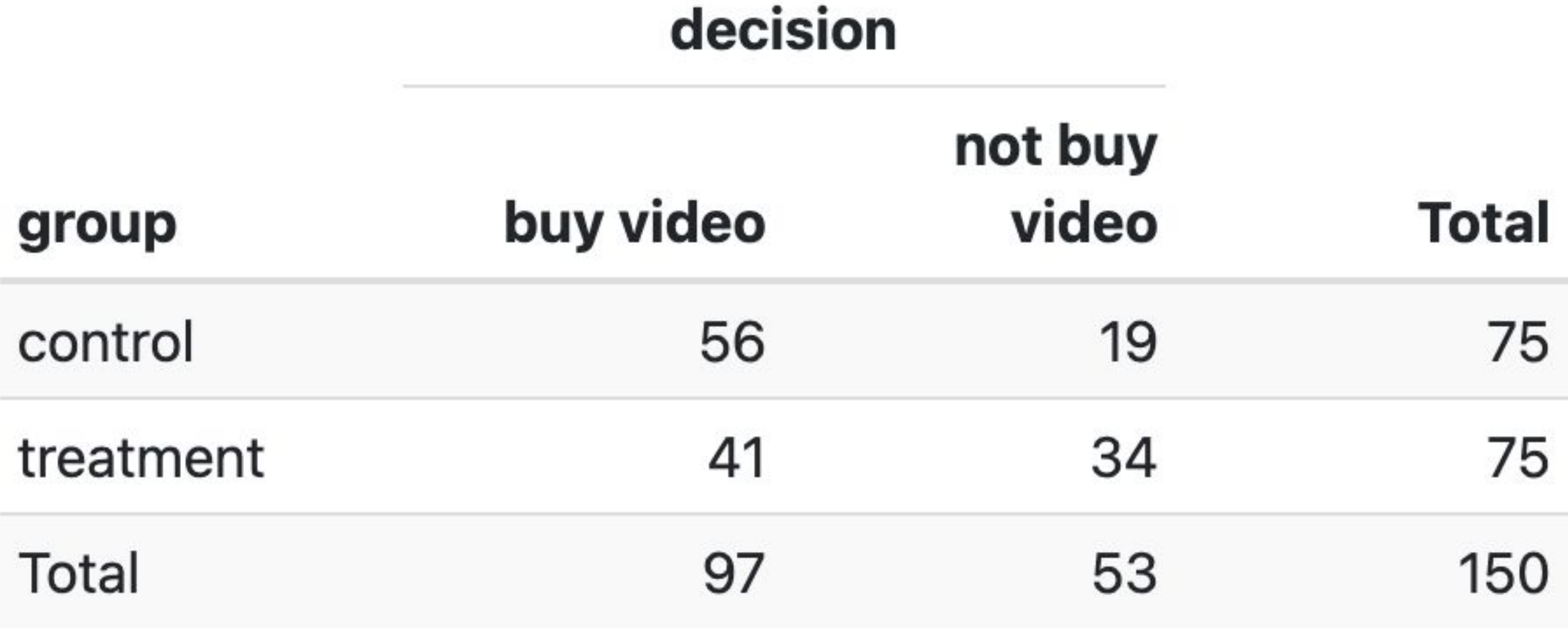
Null hypothesis.

Reminding students that they can save money for later purchases will not have any impact on students' spending decisions.

Alternative hypothesis.

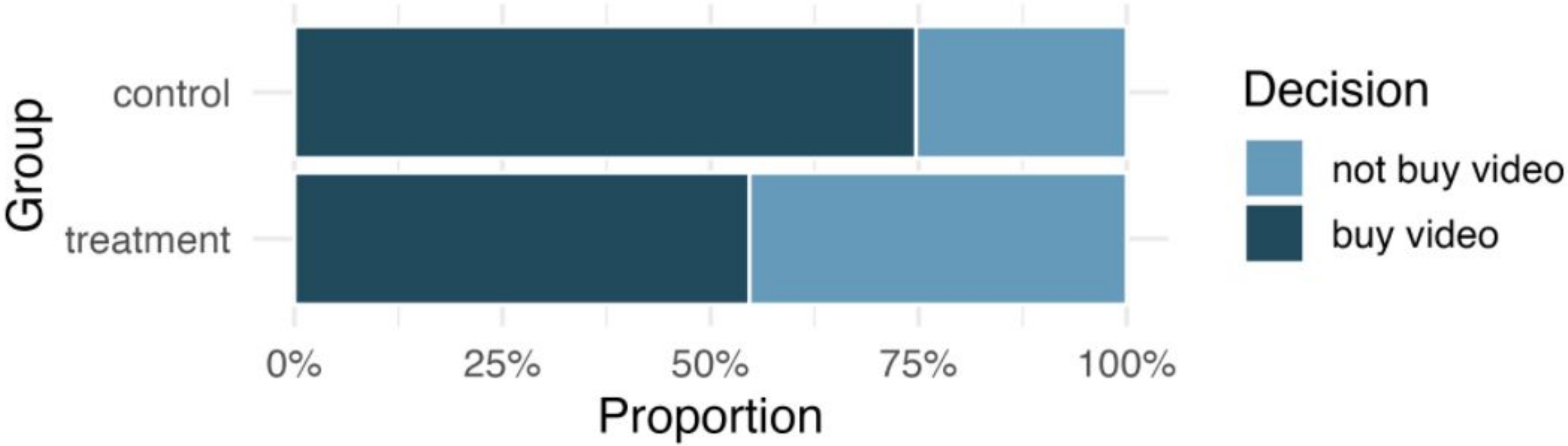
Reminding students that they can save money for later purchases will reduce the chance they will continue with a purchase.

Summary results of the opportunity cost study

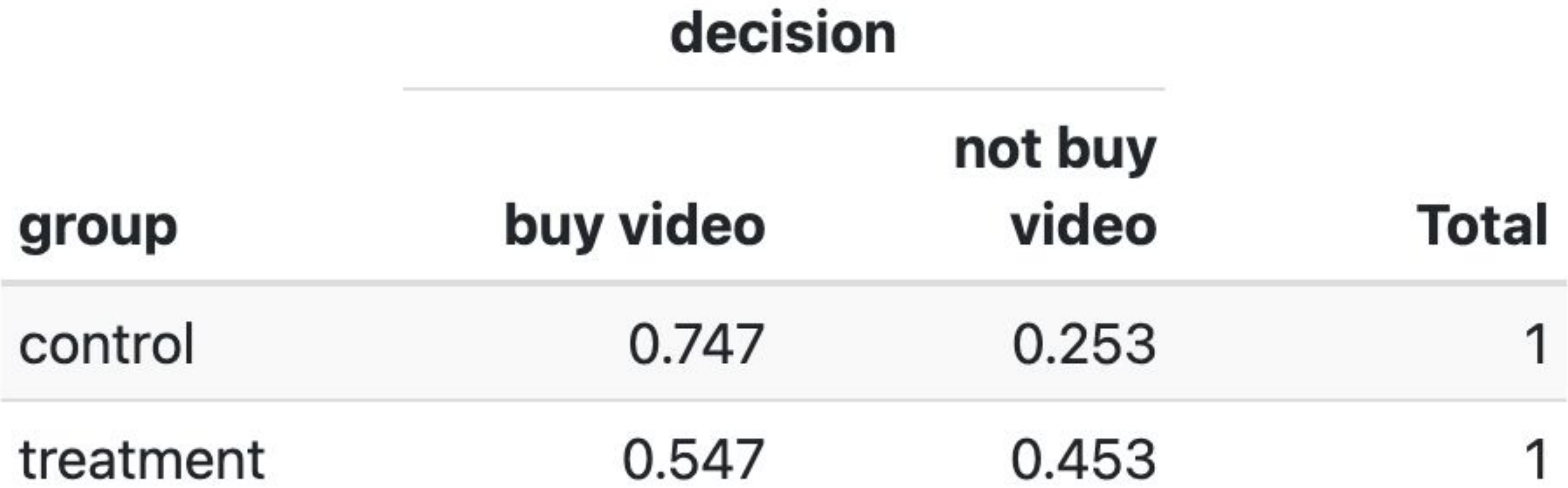


	decision		
group	buy video	not buy video	Total
control	56	19	75
treatment	41	34	75
Total	97	53	150

Stacked bar plot of results of the opportunity cost study.



Row proportions are particularly useful here since we can view the proportion of buy and not buy decisions



	decision		
group	buy video	not buy video	Total
control	0.747	0.253	1
treatment	0.547	0.453	1

Difference between groups

- **Success:** a student who chooses not to buy the video
- **Value of interest**: change in video purchase rates of non-buyers

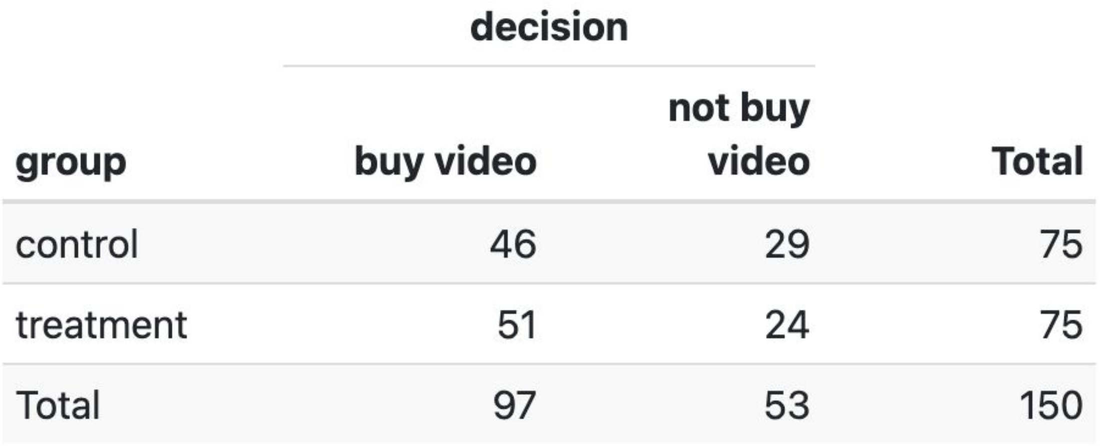
$$\hat{p}_T - \hat{p}_C = \frac{34}{75} - \frac{19}{75} = 0.453 - 0.253 = 0.200$$

- Is the difference due to **chance**?

- **Null hypothesis:**
- Treatment has no impact on student decisions
- Observed difference between the two groups of 20% could be attributed entirely to random chance.

- **Alternative hypothesis**
- Difference indicates that reminding students about saving for later purchases actually impacts their buying decisions.

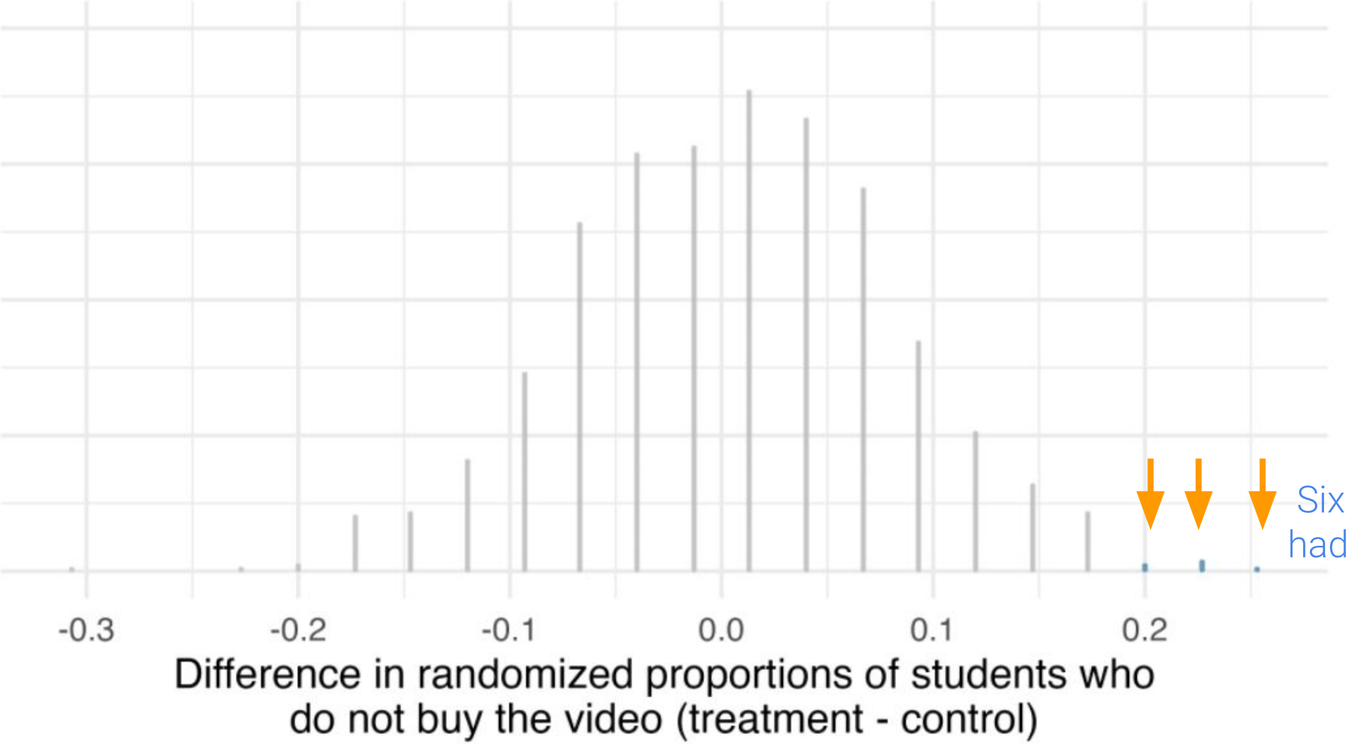
The results of a single randomization



	decision		
group	buy video	not buy video	Total
control	46	29	75
treatment	51	24	75
Total	97	53	150

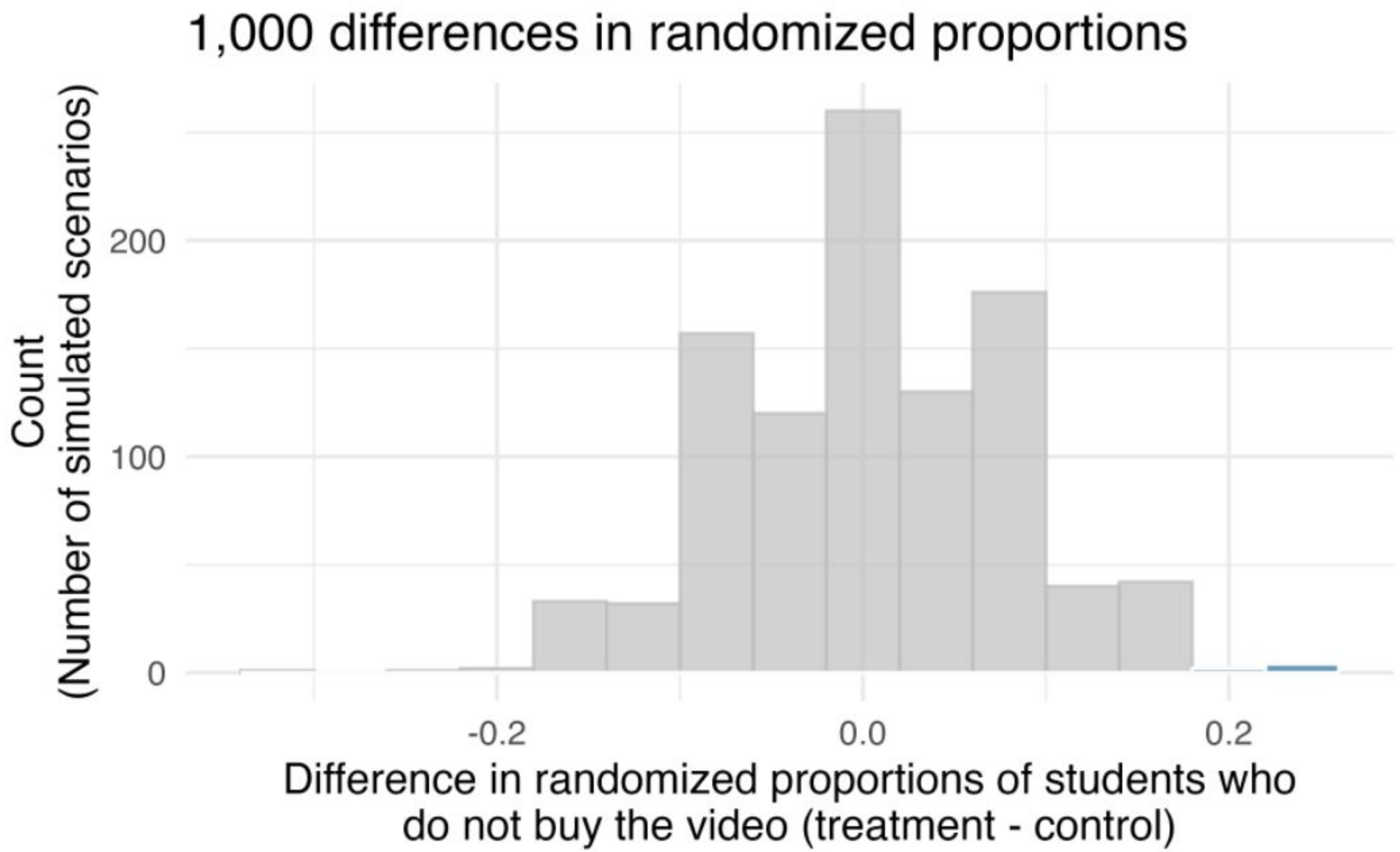
$$\hat{p}_{T,shfl1} - \hat{p}_{C,shfl1} = \frac{24}{75} - \frac{29}{75} = 0.32 - 0.387 = -0.067$$

1,000 differences in randomized proportions



Six of the 1,000 simulations had a difference of at least 20%

Stacked dot plot


1,000 differences in randomized proportions
Count
(Number of simulated scenarios)
200
100
0
-0.2
0.0
0.2
Difference in randomized proportions of students who
do not buy the video (treatment - control)

- Under the null hypothesis (no treatment effect), we'd observe
  - a difference of at least +20% about **0.6%** of the time.

- Data provide strong evidence there is a **treatment effect**:
- *"Reminding students before a purchase that they could instead spend the money later on something else lowers the chance that they will continue with the purchase"*

Hypothesis testing

- **Null hypothesis (H_0)**
- Often represents a
  - skeptical perspective or
  - a claim of “no difference” to be tested

- **Alternative hypothesis (H_A)**
- Represents an
  - alternative claim under consideration
  - is often represented by a range of possible values for the value of interest.

If a person makes a somewhat unbelievable claim, we are i**nitially skeptical.**

However, if there is **sufficient evidence** that supports the claim, we set aside our skepticism.

The US court system

EQUAL JUSTICE UNDER LAW

The US court considers two possible claims about a defendant:

- not guilty
- guilty

- The jury considers whether the evidence is so convincing (strong) that there is no reasonable doubt regarding the person’s guilt.
- That is, the skeptical perspective (**null hypothesis**) is that the person is innocent until evidence is presented that convinces the jury that the person is guilty (**alternative hypothesis**).

- If a jury finds a defendant **not guilty**, this does not necessarily mean the jury is confident in the person's **innocence**.
- They are simply not convinced of the alternative, that the person is **guilty**.

This is also the case with hypothesis testing:

Even if we fail to reject the null hypothesis, we do not accept the null hypothesis as truth.

Remember

Failing to find evidence in favor of the alternative hypothesis

is not equivalent to finding evidence that the null hypothesis is true.

p-value and statistical significance

Discrimination experiment

- H_0: Gender has **no effect** on promotion decisions.

- H_A:Female candidates are **discriminated** against in promotion decisions.

Discrimination experiment

Simulation result:

- Difference from chance alone
  - Assuming the null hypothesis was true
  - Would only happen about **2 in 100 times**.

- The 2-in-100 chance is what we call a **p-value**
- Is a probability
  - quantifying the strength of the evidence against the null hypothesis
  - given the observed data.

p-value

1. Probability of observing data at least as favorable to the alternative hypothesis as our current dataset
2. if the null hypothesis were true.

- We typically use a **summary statistic** of the data to compute the p-value
    - such as a difference in proportions
- This summary value is often called the **test statistic**.

Statistical significance

- When the p-value is “small” we say the results are **statistically significant**.
- This means the data provide such **strong evidence against H_0**
- We reject the null hypothesis in favor of the alternative hypothesis.

significance level α

- The threshold, called the **significance level** and often represented by α (the Greek letter alpha).

- The value of α represents how rare an event needs to be in order for the null hypothesis to be rejected.

p-value

- Historically, many fields have set α = 0.05
- Meaning that the results need to occur less than 5% of the time, if the null hypothesis is to be rejected.
- The value of α can vary depending on the the field or the application.

statistically significant vs significant

Although in everyday language “significant” would indicate that a difference is large or meaningful, that is not necessarily the case here.

The term “statistically significant” only indicates that the p-value from a study fell below the chosen significance level.

Summary

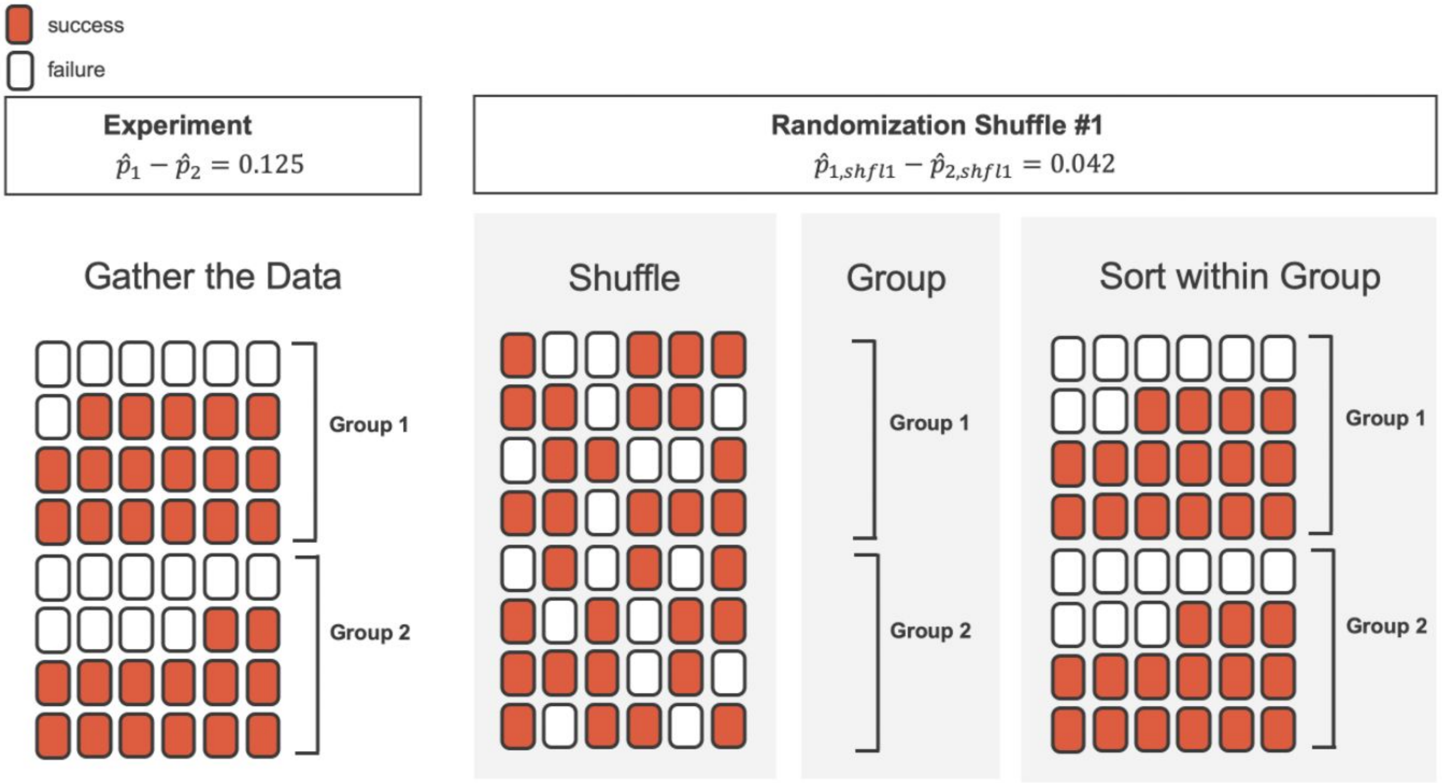


Figure 11.8: An example of one simulation of the full randomization procedure from a hypothetical dataset as visualized in the first panel. We repeat the steps hundreds or thousands of times.

Randomization test procedure

1. Frame the research question in terms of hypotheses.
2. Collect data with an observational study or experiment.
3. Model the randomness that would occur if the null hypothesis was true
4. Analyze the data
5. Form a conclusion

Terms you should know

- alternative hypothesis
- confidence interval
- hypothesis test
- independent
- null hypothesis
- p-value
- permutation test
- point estimate
- randomization test
- significance level
- simulation
- statistic
- statistical inference
- statistically significant
- success
- test statistic