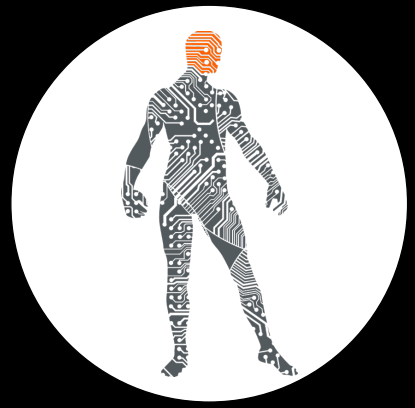
RECAP

Aalto-yliopisto
Aalto-universitetet
Aalto University



Research Methods in Engineering Psychology – Lecture 2

B.Sc. Engineering Psychology

Prof. Dr. Robin Welsch

Modules today

Advancing theory

Study designs and variables

Advancing Theory

Inference

- Hypothetico-deductive method
- Falsification
- Bayesianism
- Statistical test workflow

Science “breakthroughs” in 2022

A ‘Reversible’ Form of Death? Scientists Revive Cells in Dead Pigs’ Organs.

Researchers who previously revived some brain cells in dead pigs succeeded in repeating the process in more organs.

AI gets creative

BY MATTHEW HUTSON

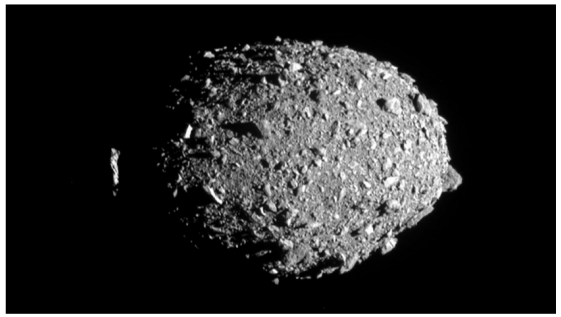
g work, *Théâtre D'opéra Spatial*, was created using text-to-image artificial intelligence. © 2022 JASON M

l) is making inroads in areas once considered uniquely human, including artistic discovery. The machines’ encroachment was slow at first, but this year it

ng evidence—inescapable on social media—came from so-called text-to-image e learning to analyze pairings of text and images online, finding patterns that images based on new text. Last year, the research lab OpenAI presented a ALL-E that when asked for “an armchair in the shape of an avocado” could spit mples. This spring, OpenAI released a large upgrade, DALL-E 2. It implemented

Asteroid deflected

BY ZACK SAVITSKY



The asteroid Dimorphos was the target for this year’s test run of a potential planetary defense system—an impact from a fridge-size spacecraft that successfully altered its orbit. NASA/JOHNS HOPKINS APL

For thousands—if not millions—of years, a little moon named Dimorphos made laps around a larger asteroid, millions of kilometers from Earth. On 26 September, NASA smacked into it with a spacecraft, forever altering its orbit—and demonstrating a strategy that might one day save humanity.

When the fridge-size Double Asteroid Redirection Test (DART) satellite barreled into the 160-meter-wide Dimorphos at 6 kilometers per second, scientists celebrated the first-ever mock trial of a planetary defense mission. NASA’s goal was to knock Dimorphos slightly closer to its partner, shortening its orbital period and demonstrating a strategy for thwarting real threats, should future Earth-bound asteroids be detected.

PUBLIC HEALTH

We Need to Improve Indoor Air Quality: Here’s How and Why

Upgrading buildings’ ventilation, filtration and other factors would not only decrease COVID transmission but also improve health and cognitive performance in general

By Tanya Lewis on June 8, 2022 اعرض هذا بالثقافة العربية

Air purification device stands in a classroom at an elementary school in Berlin. Credit: Annette Riedl/dpa/Alamy Live News

We spend 90 percent of our lives indoors, yet most of us seldom spare a thought for the quality of the air we breathe there.

READ THIS NEXT

PUBLIC HEALTH
Why We Need to Upgrade Our Face Masks--and Where to Get Them
Tanya Lewis

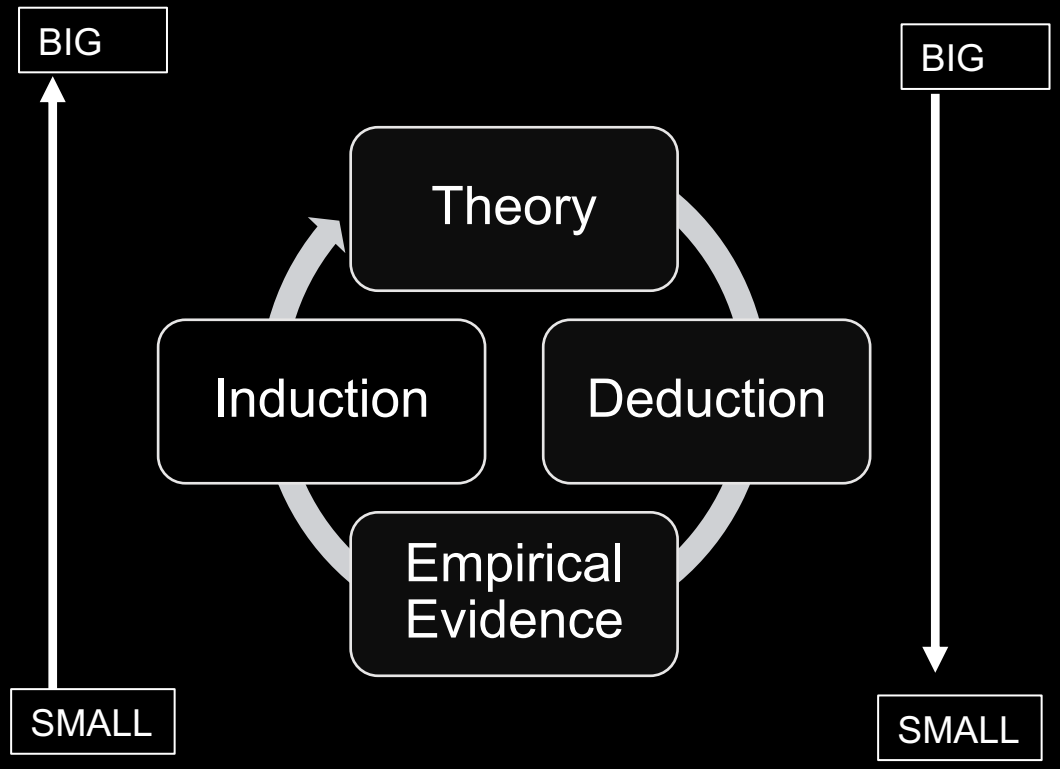
EPIDEMIOLOGY
How to Make Smart COVID Risk-Benefit Decisions
Devabhaktuni Srikrishna

PUBLIC HEALTH
Is Your Office Safe from COVID? What to Know Now That Your Boss Wants You Back
Christie Aschwanden

PUBLIC HEALTH
Safer Indoor Air, and People Want Masks on Planes and Trains: COVID Quickly,

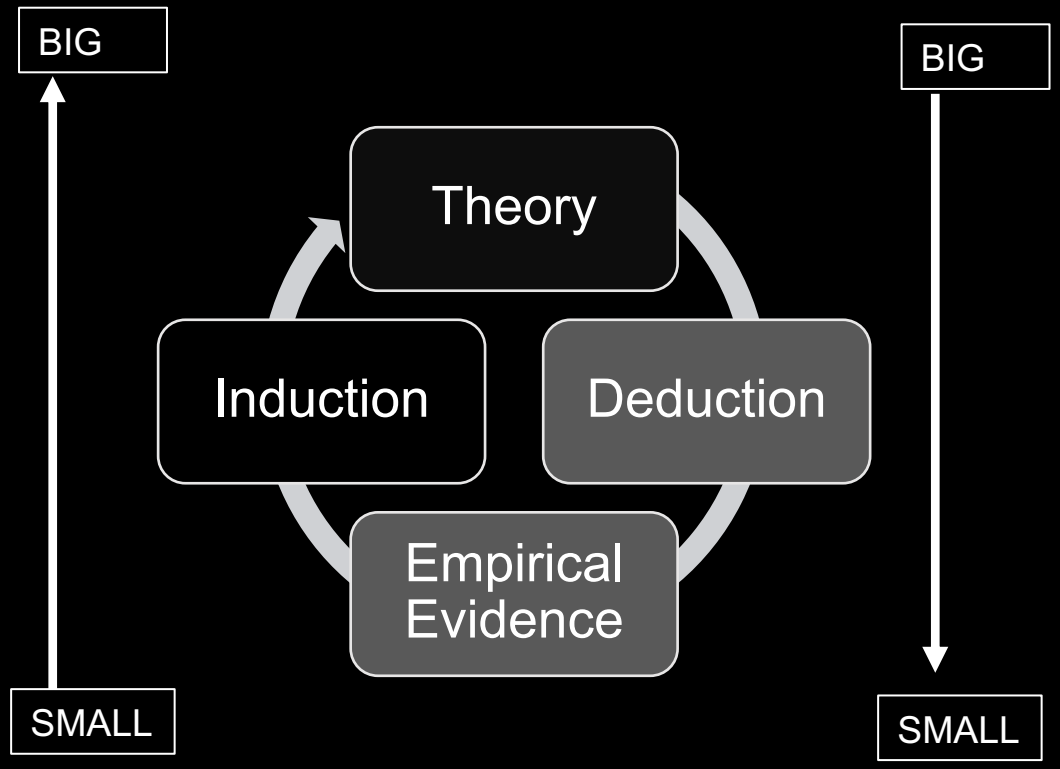
https://www.science.org/content/article/breakthrough-2022

Hypothetico-deductive method



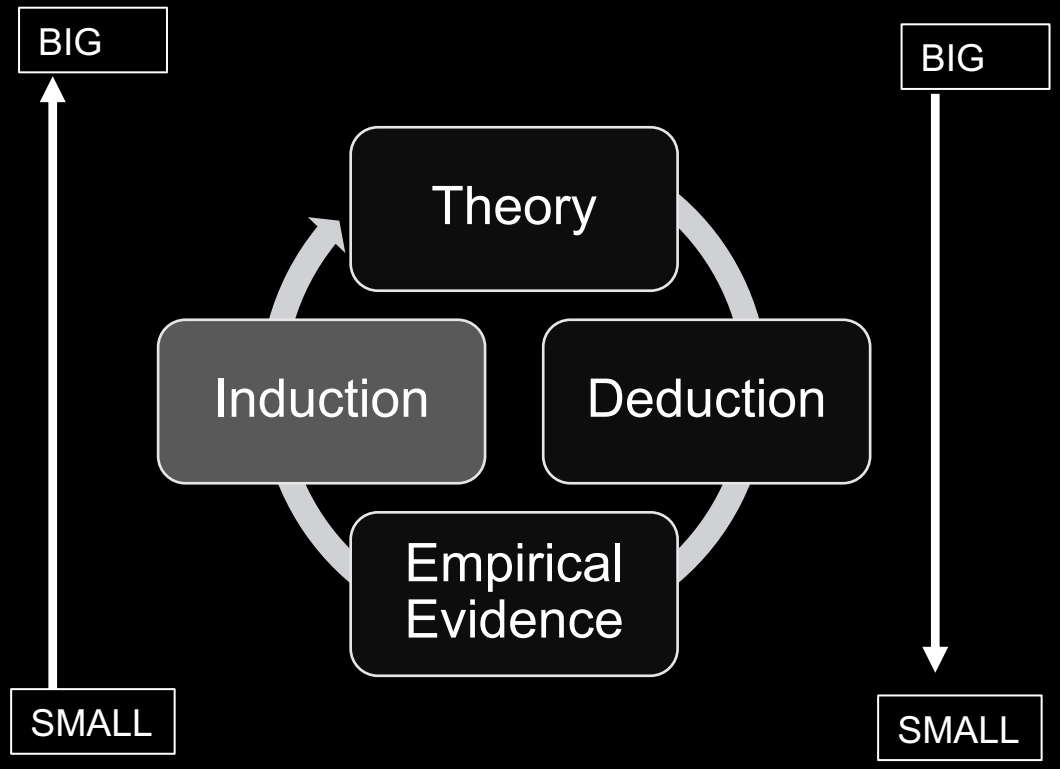
1. Established Theory
2. Hypothesis
3. Study design
4. Data collection
5. Analysis
6. Report results
7. Revise Theory

Hypothetico-deductive method

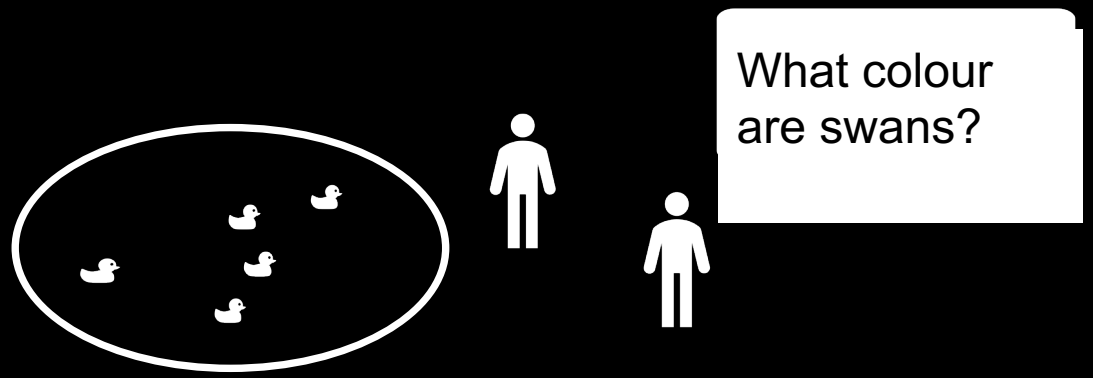


1. Established Theory
2. Hypothesis
3. Study design
4. Data collection
5. Analysis
6. Report results
7. Revise Theory

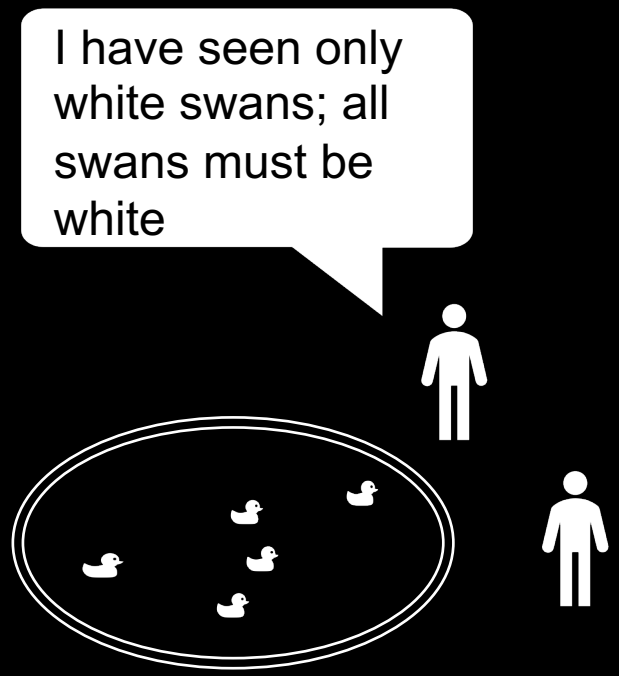
Induction



1. Established Theory
2. Hypothesis
3. Study design
4. Data collection
5. Analysis
6. Report results
7. Revise Theory


What colour are swans?

Folk explanation



Deduction

Theory

- **All swans are white**

Hypothesis/Model

- **All swans in New Zeeland are white**

Gather Evidence

- **Record collour of all swans in New Zeeland**

BIG

SMALL

Ways of Inference from data

All swans in New Zeeland are white

Falsification

- Observing one black swan falsifies the hypothesis

BIG

SMALL

Bayesianism

- Observing one black swan lowers the probability of the hypothesis being true

Ways of Inference from data

All swans in New Zeeland are white

Falsification

- Observing one black swan falsifies the hypothesis -->theory has to be updated

Bayesianism (model-based reasoning)

- Observing one black swan lowers the probability of the hypothesis being true

Abductive reasoning

- Swan color varies more than previously thought or it is an entire new species

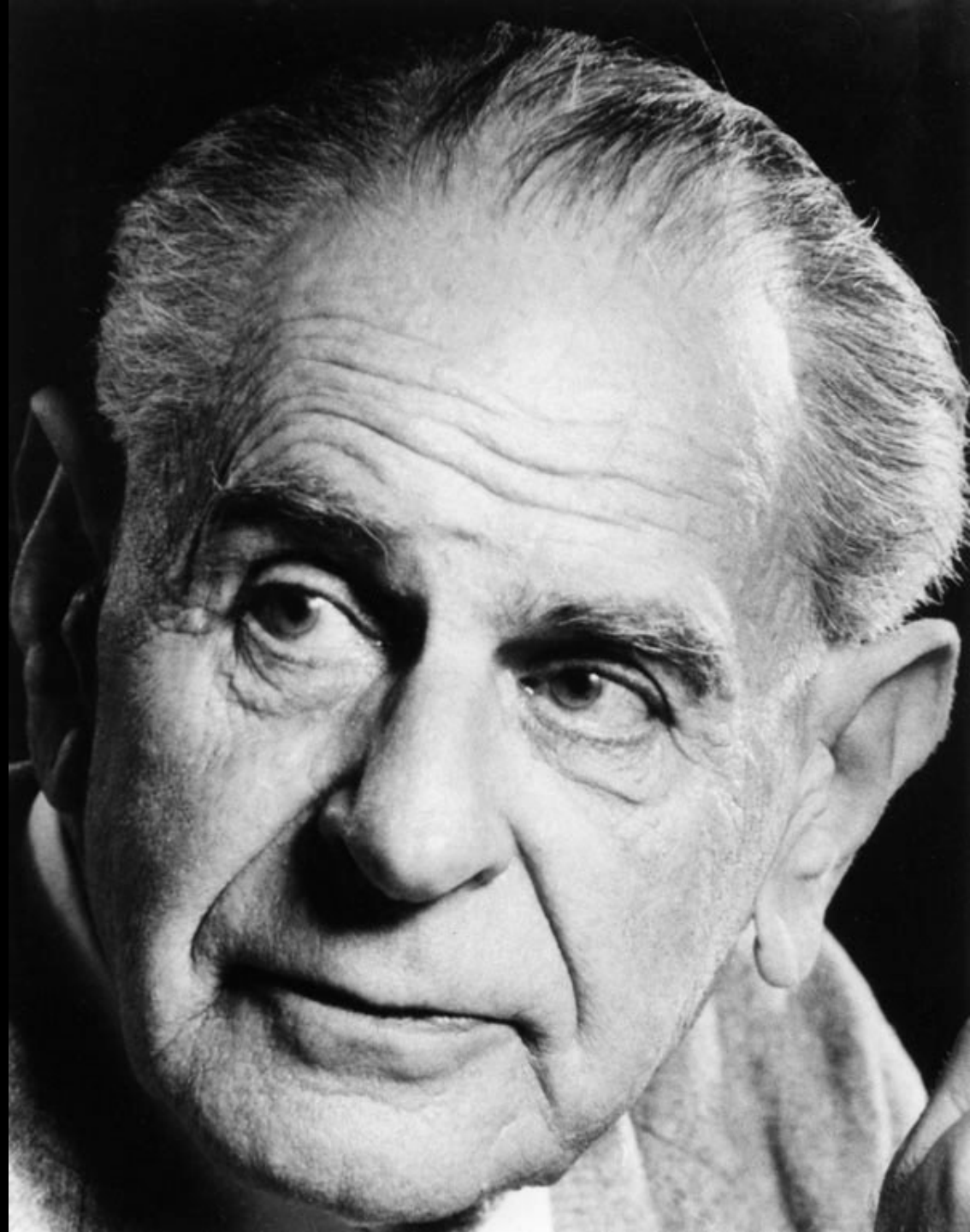
BIG

SMALL

The next few slides might be more advanced and will be covered more closely in your stats course

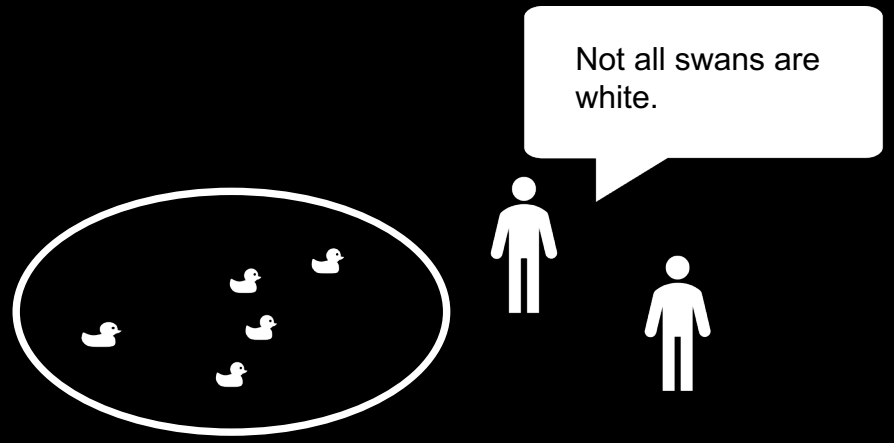
Falsification

- Hypothesis makes a clear prediction
  - Null-Hypothesis: ”All swans are white”
  - Alternative-hypothesis: “Not all swans are white”
- Observation or measurement that contradicts this prediction falsifies the hypothesis
  - one black swan falsifies this hypothesis



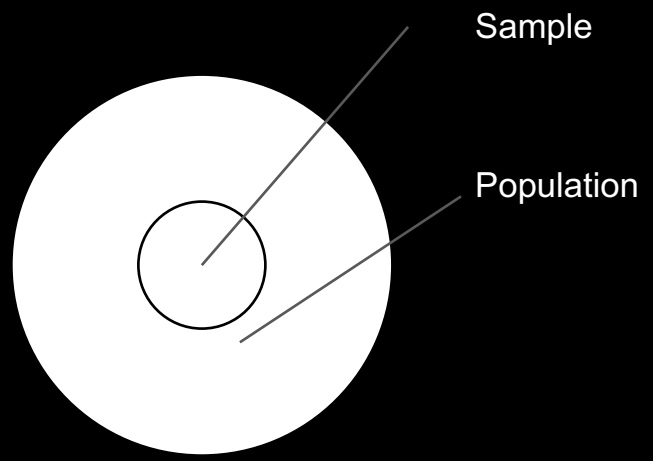
Falsification is dominant in Engineering Psychology

- Null Hypothesis Testing (NHST) assigns a probability to data assuming the NULL-hypothesis under the assumption that the procedure could be replicated
- In practice:
  - The probability of the data under the NULL-hypothesis is low (conventionally $p < .05$) and it thus needs to be rejected or
  - The probability of the data under the NULL-hypothesis is high and therefore, one cannot reject the NULL-hypothesis



Sampling

- refers to the process of selecting a subset of individuals from a larger group, or population, to study or represent the population as a whole
- make inferences about the population based on the characteristics of the sample
- sample characteristics may affect the validity of the results → does the sample represent the population we talk about



Applied NHST workflow

Website A is different from Website B

1. Derive statistical hypothesis
2. Define alpha-level (5%)
3. Define statistical model
4. Gather data
5. Compute test-statistics
6. Compute *p*-value
    1. The probability of NULL-hypothesis given the the test-statistic
7. Compare *p*-value to alpha-level
8. Report results concerning data given the hypothesis

Time Website A != Time Website B

5%

t-test

M_1 = 100 minutes
SD_1 = 20
n_1 = 60

M_2 = 80 minutes
SD_2 = 10
n_2 = 65

$$t = \frac{M_1 - M_2}{\sqrt{\frac{SD1^2}{n1} + \frac{SD2^2}{n2}}}$$ →6.98

p = **5.967e-10**

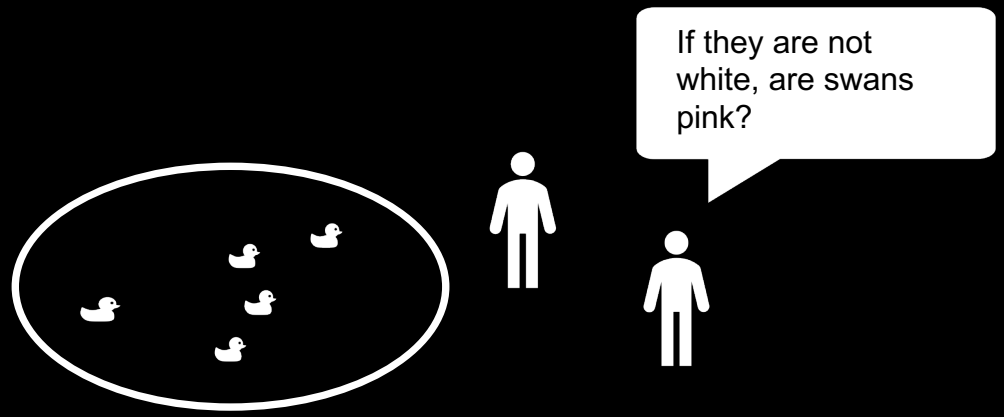
p < 5%

There was a significant difference between websites, $t(85.19) = 6.98$, $p < .001$, d = 1.28. Participants spent more time on Website A ($M = 100$, $SD = 20$) as compared to Website B ($M = 80$, $SD = 10$)

Problems of Falsification in practice

Testing models is not black and white

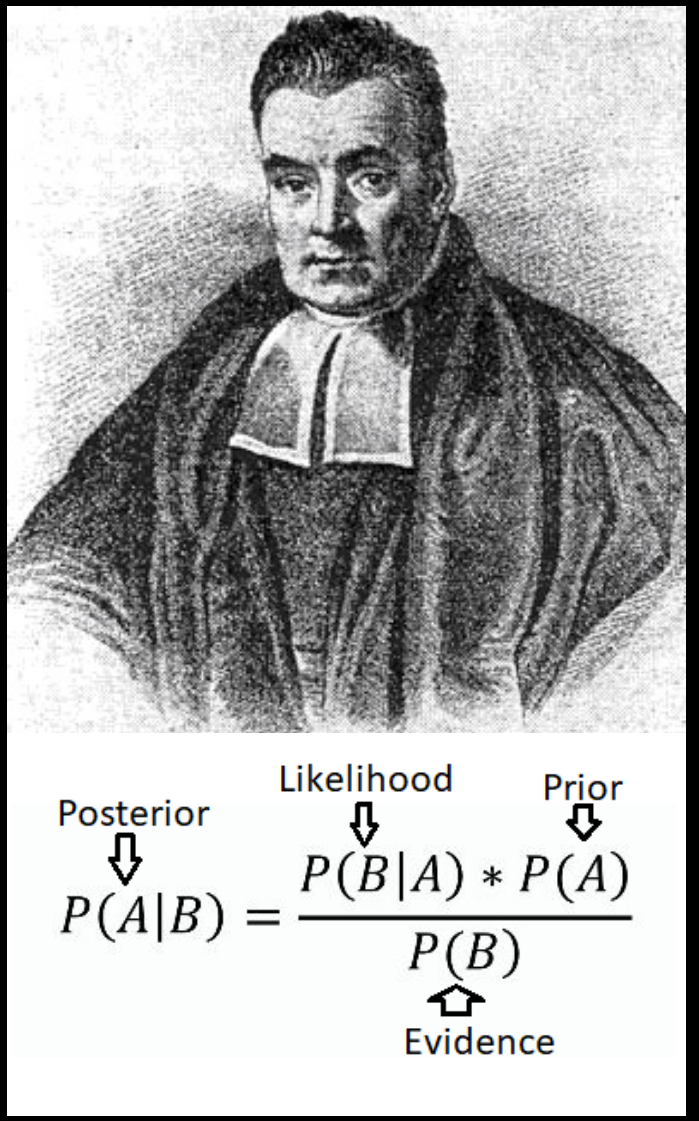
1. Direct observation is rarely the case in psychology
2. Hypotheses cannot be condensed to simple predictions with binary outcomes
   - What do we conclude if we observe a grey swan or a bird that looks much like a swan but has small genetic differences?
3. Models/Hypotheses are only an approximation and are always false in terms of falsification



Bayesianism

To the rescue

- Degrees of belief in response to a varying body of evidence
- Models/hyotheses cannot be rejected or „verified“, they can only be compared and updated given the evidence
- Problem
  - Scientistits need to formulate an initial probability of the hypothesis -- a prior



By Unknown author - [2][3], Public Domain, https://commons.wikimedia.org/w/index.php?curid=14532025

Applied Bayesian workflow

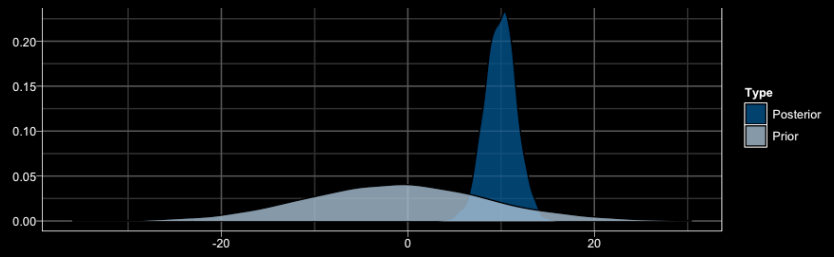
1. Gather prior knowledge
2. Define models
3. Collect data
4. Evaluate the updated model
5. Report results concerning hypotheses given the data

Differences between websites should be substantial but i do not know which one is better

Bayesian *t*-test prior on group parameter normal(010)

M_1 = 100 minutes
SD_1 = 20
n_1 = 60

M_2 = 80 minutes
SD_2 = 10
n_2 = 65



The effect of group (coded as Web A –1, Web B = +1, M = 10.93, $HDI_{95\%}$[7.56, 14.14]) has a 100.00% probability of being positive (> 0). Participants spent more time on Website A (M = 100, SD = 20) as compared to Website B (M = 80, SD = 10)

Schad, D. J., Betancourt, M., & Vasishth, S. (2021). Toward a principled Bayesian workflow in cognitive science. *Psychological methods*, *26*(1), 103.

Advanced part is over!

Statistical inference comparison

NHST

- Asssigns probability to data given the NULL-hypothesis
- Needs a threshold (e.g. 5%)
- Results in binary decisions
- Empirical Science progresses by rejecting the NULL-hypothesis

Bayesian inference

- Assigns credibility to hypotheses given the data
- Needs prior
- Results in probability statements
- Empirical Science progresses by developing models

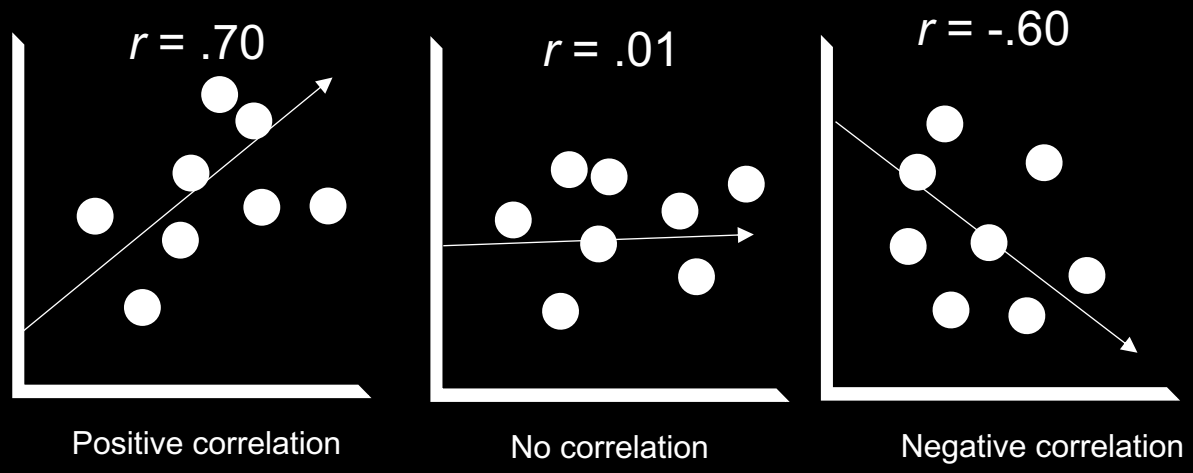
Most basic statistical models

Regression model

Correlation represents the strength and direction (+1 to -1) of the relationship between two variables

$y = x * b + e$

$$r_{xy} = \frac{s_{xy}}{SxSy}$$



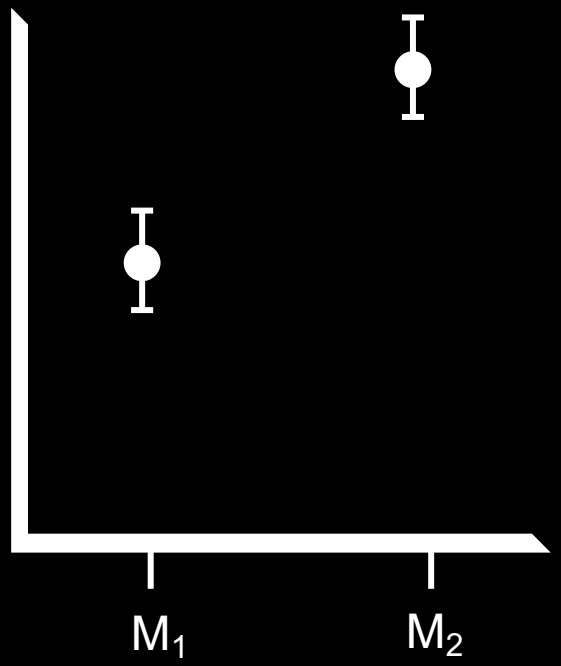
More on this in your stats course!

Mean comparison

Cohen‘s d represents the strength (0; --) of the mean difference for two samples

$$t = \frac{M_1 - M_2}{\sqrt{\frac{SD1^2}{n1} + \frac{SD2^2}{n2}}}$$

$$d = \frac{M_1 - M_2}{SD\ (pooled)}$$



Vocabulary statistical inference

- **Significance** refers to the likelihood that test statustic occurs by chance. In psychological research, we deem sth. significant if the probability of obtaining the test-statistic by chance is less than 5%.
- **Null-result** is a result that does not show a significant relationship between variables
- **Test-statistic** is a statistical measure that is used to evaluate the significance of a research result. Typically based on sample statistics (e.g. standardised difference of means in a *t*-test)
- **Mean** represents the centre of a distribution as an average of measurements
- **Standard deviation** is the amount of variation in measurments. It is calculated by taking the square root of the variance, which is the average of the squared differences between each measurement and the mean
- **Central-limit theorem** is a statistical principle that states that the distribution of sample means will be approximately normal (bell-shaped)

Checklist inference

- ❑ Are assumptions of statistical models met?
- ❑ Does your hypothesis translate into a statistical model?
- ❑ Do not write “verify” or your hypothesis “is true” in NHST.
- ❑ Do you have proper priors? (Inference in Bayesian statistics needs prior knowledge and thus can be sensitive to specifying this prior knowledge)
- ❑ Have you considered alternative data analyses? Do they result in the same inferences?

Summary

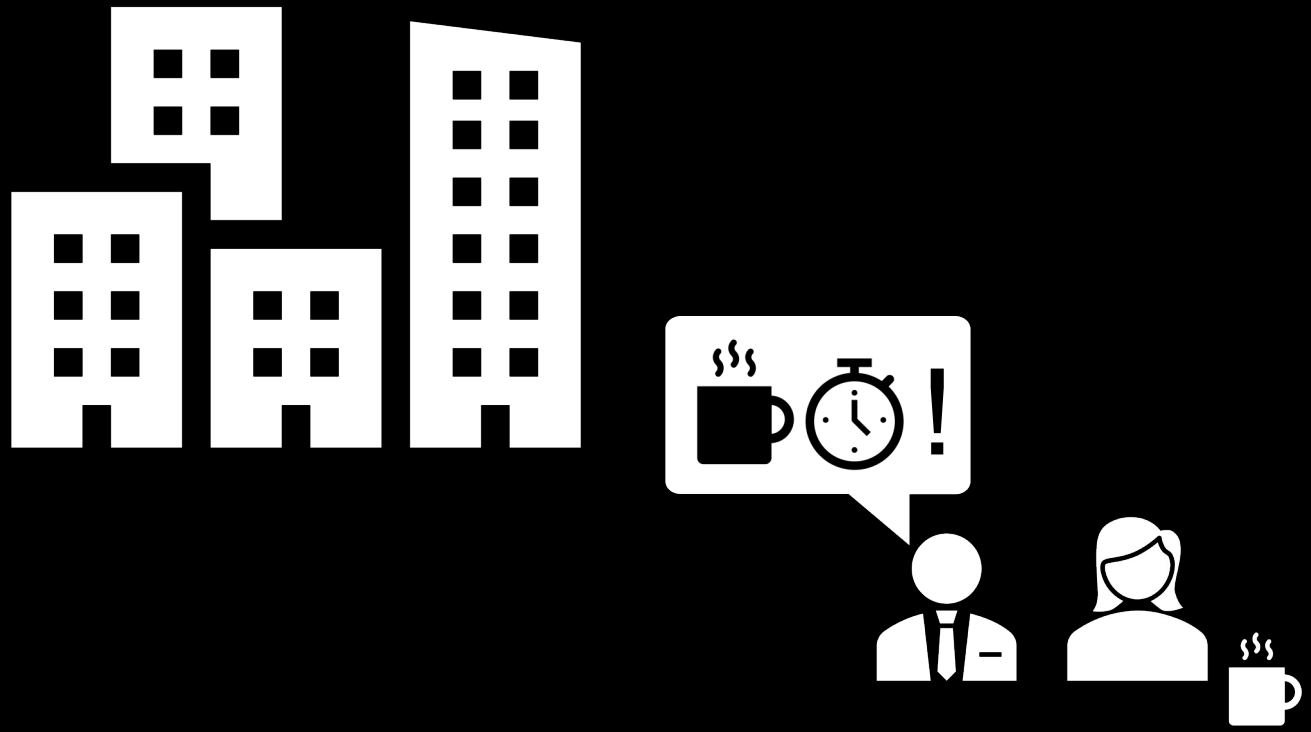
- Scientists use the hypothetico-deductive method to ensure that the knowledge they produce is robust
- Statistical inference in response to empirical data can be made using Falsification or Bayesianism
- Falsification is dominant in engineering psychology, but Bayesianism is accepted as a viable counter-movement

Questions?

Study design and Variables

- Experimental designs
  - Survey study
  - Within/Between-experiments
  - RCT
- Reading graphs
- Effects to control for in study design

Case study



How does coffee consumption affect productivity?

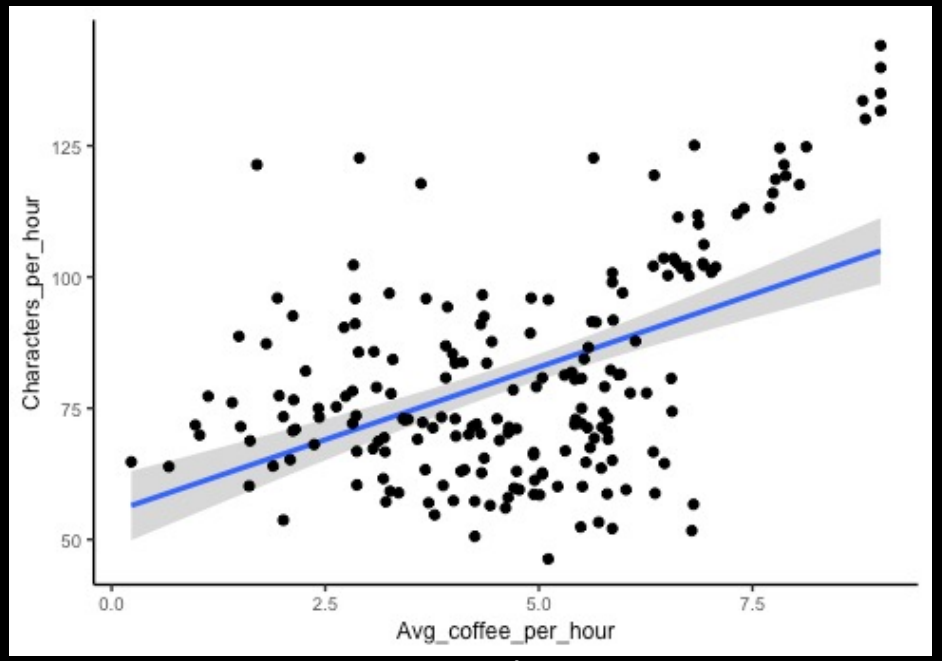
How does coffee consumption affect productivity?

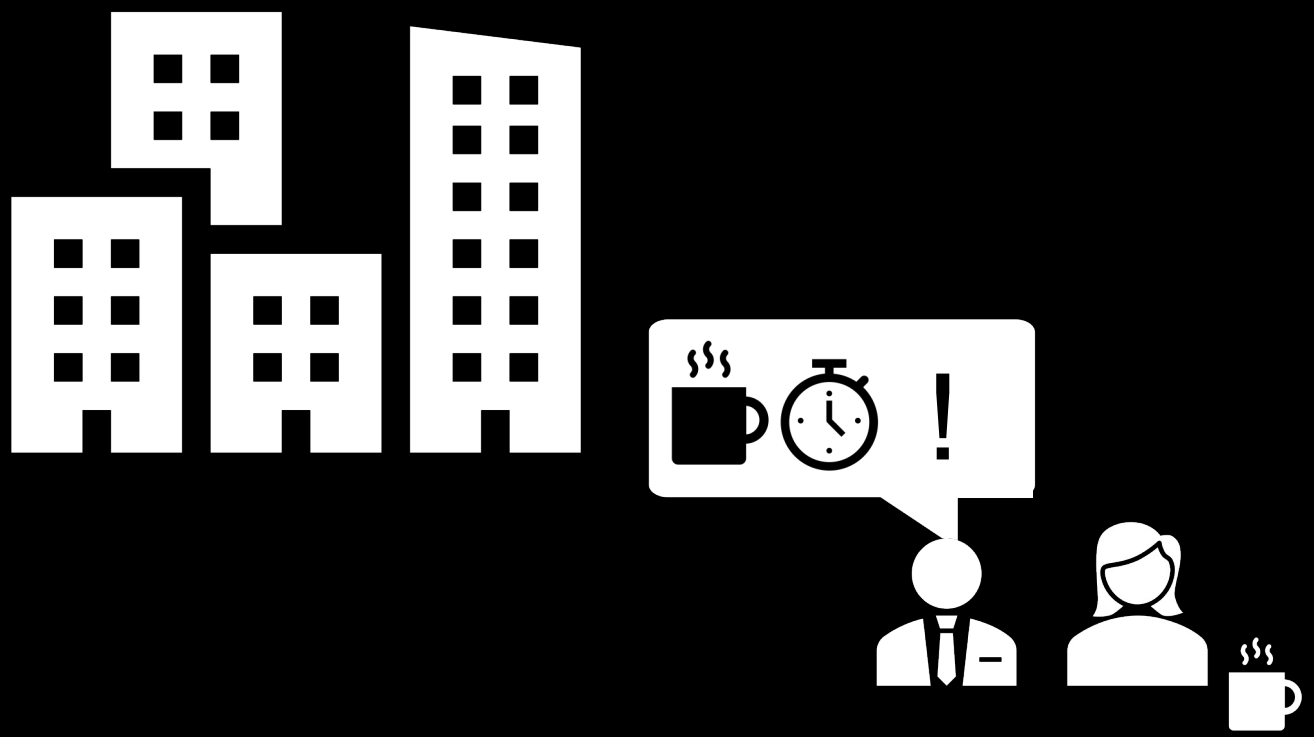
Hypothesis: The more coffee people drink the more characters they‘ll type during work

Survey study

Cross-sectional study

1. You ask your coworkers to log the characters they typed and also about their coffee intake in a given day
2. You calculate the average consumption of coffee hour and average number of characters typed per hour
3. You collect the data and relate the user-ids of the maschine to your coworkers counts and plot them





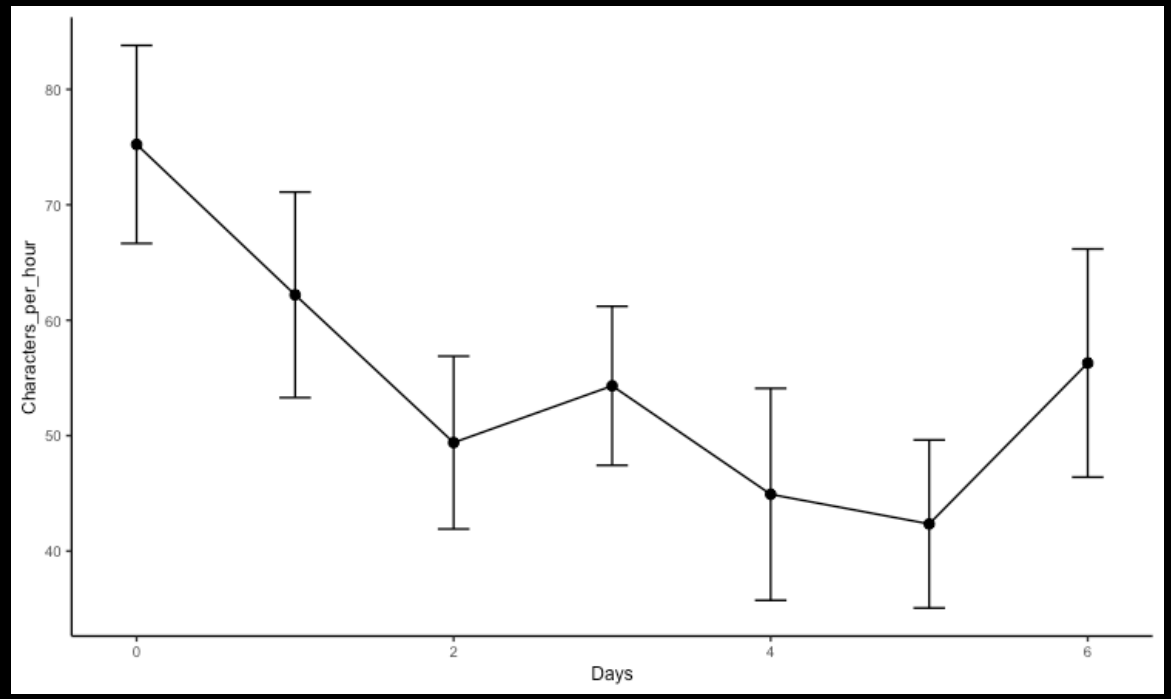
- Cause and effect relationship
  - Does coffee cause faster-typing speed or do people that type a lot like coffee more?
- Control of confounding variables
  - Did they all have strong black coffee?
  - Did people drink decaffeinated coffee?
- Experimental control
  - When did people drink the coffee?
- Sampling bias
  - Did only people who drink coffee answer the survey?

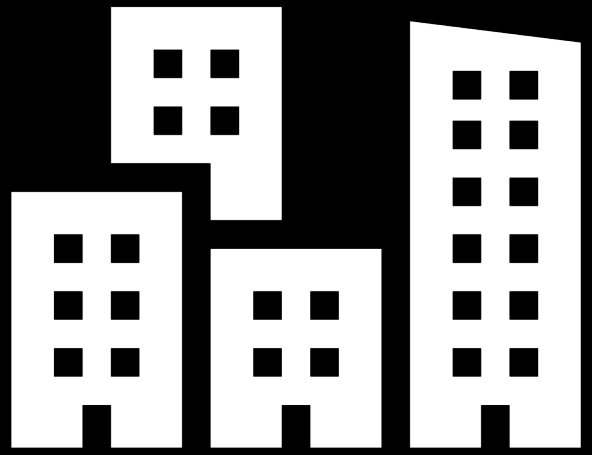
How does coffee consumption affect productivity?

Hypothesis: If coffee intake is prohibited productivity will decrease

Longitidunal study

1. You ask your coworkers to log the characters they typed every day
2. You ask them to stop drinking coffee for a week
3. You collect data on 6 days
4. You calculate the average number of characters typed per hour





- Did people really abstain from drinking coffee?
- Is their tolerance relevant?
- Were they less productive because they didn’t participate in coffee breaks?
- Did the expectation of a performance decline produce a nocebo effect?
- Can any other factor explain the decline, e.g., workshops that distracted from the typical schedule?

How does coffee consumption affect productivity?

Hypothesis: Coffee consumption causes an increase of words typed

Study designs for experimental research

- In a experiment, we want to **measure the effect of the independent variables on the dependent variables**.
  - In other words: We measure the **dependent variables** while varying the **levels** of the **independent variables**.
- How do we do that?
  - We expose our participants to the different **experimental conditions** and measure the DVs.
- Factorial designs allow us to investigate two types of effects
  - **Main effects** (the individual influence of a the respective independend variables)
  - **Interaction effects** (the effect of differing levels of one IV on another IV)

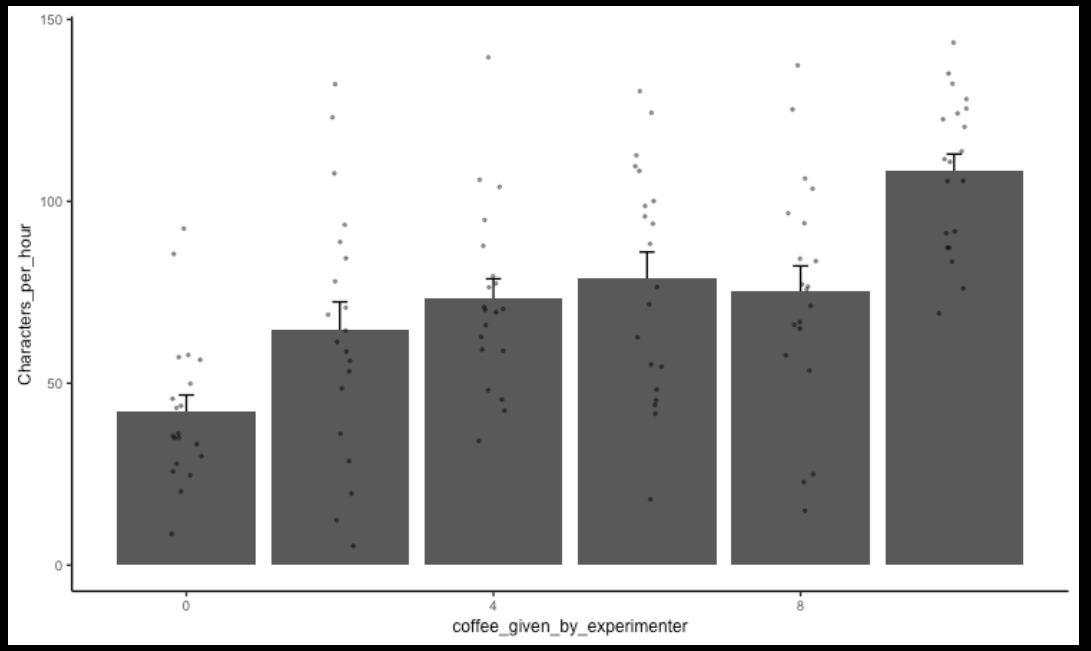
Vocabulary for experiments

- **Experimental blocks**: a sequence of trials or tasks in an experiment that are grouped together to minimize the influence of confounding variables
- **Counterbalancing**: Matching of order in simple experiments (e.g. Group 1 tests Prototype A → B; Group 2 tests Prototype B → A)
- **Trials**: Measurements that are potentially repeated within an experimental condition to increase the reliability of the results
- **Randomization**: Randomly assigning participants to different treatments, conditions, trials etc. in a study to control for confounding variables
- **Quasi-randomization**: Having a fixed-randomized order where full-randomization is not feasible
- **Latin-Square design**: Administration of conditions in a sequence that systematically varies the occurance of conditions so that a condition occurs equally often at a position in a given sequence.

Experimental Study

Between-subjects

1. You invites 120 people into the lab
2. Manipulate the dose of coffee between-subjects
   - 20 people are given each 0,2,4,6,8,10 cups
   - Each cup has 240mg Caffeein
3. You randomly allocate participants to one of the groups
4. After drinking they have to type a text about cats for one hour
5. You calculate the average characters typed per group

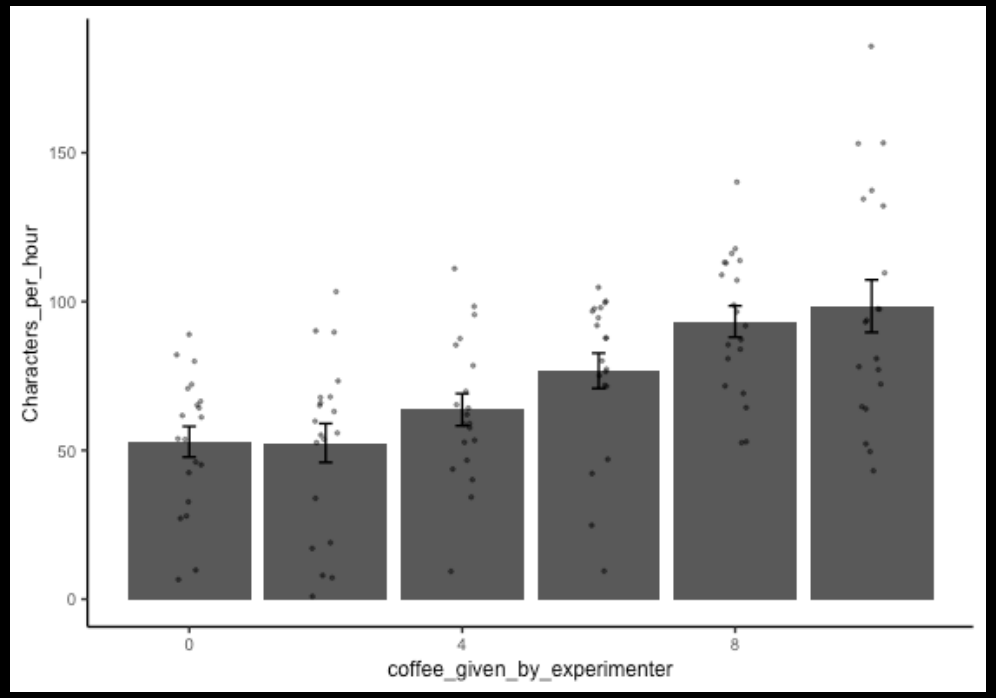


- Independent variable is **Cups of coffee given by the experimenter**
- The dependen variable is **Characters per hour**

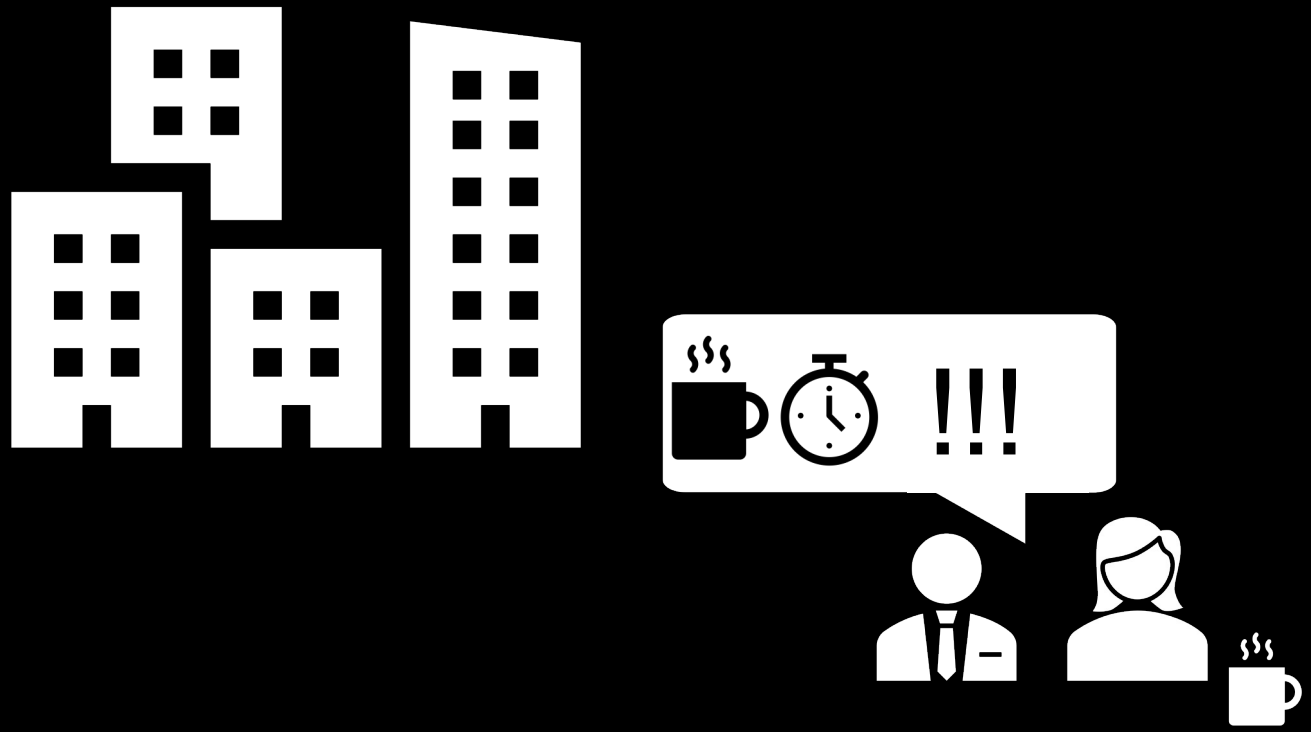
Experimental Study

Within-subjects

1. You invite 20 people on 6 days to the lab
2. Manipulate the dose of coffee wihtin subjects on each day
   - 20 people are given all conditions 0,2,4,6,8,10 different days
   - Each cup has 240mg Caffeein
3. You randomly allocate participants to start with one condition
4. After drinking they have to type a text about animals for one hour
5. You calculate the average characters typed per group



- Independent variable is **Cups of coffee given by the experimenter**
- The dependent variable is **Characters per hour**

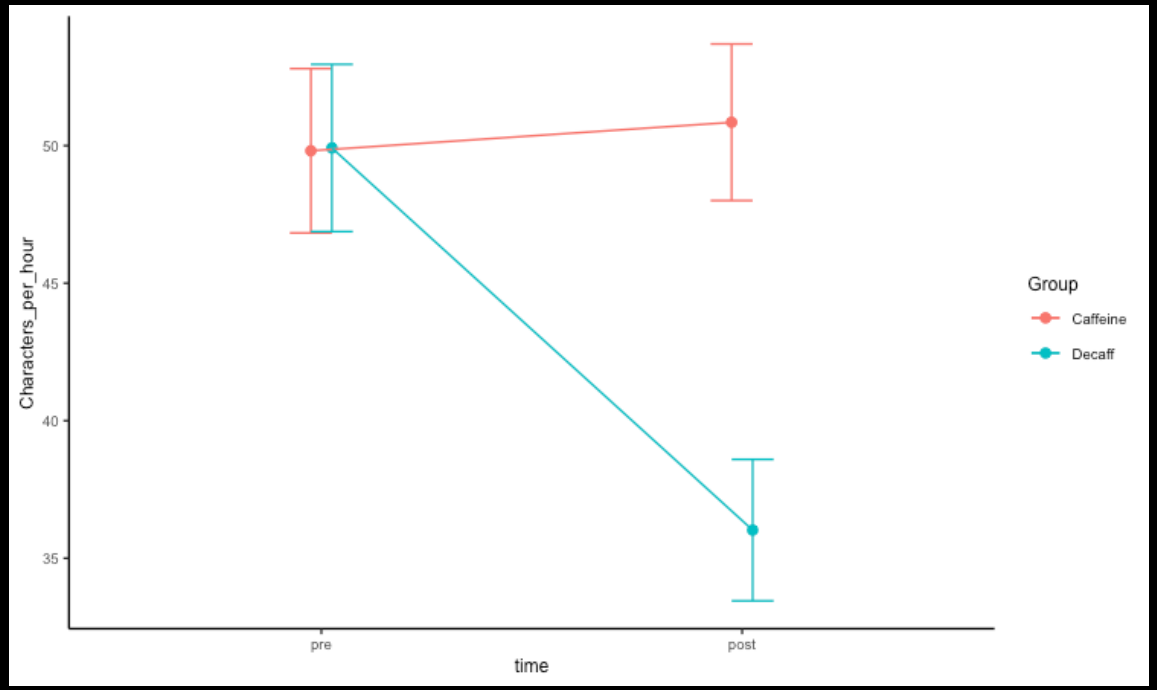


- Participants are forced to drink coffee
- Manipulation is not adequate for situation or the person
- Context is artificial and doesnt resemble coffee drinking (e.g. socializing)
- Experimenter effects: Participants in the high-coffee condition may write more words to conform with experimenter‘s hypothesis

RCT – randomized controlled trial

Interaction effect of Time x Group

1. You invite 20 people for each experimental group
2. Manipulate the dose of Caffeine between subjects (Caffeinated coffee vs. Decaff)
3. You randomly allocate participants to the groups
4. They write an essay at “pre”
5. After three days of drinking only the study-coffee they are invited to the lab-test again
6. They write an essay at “post”
7. You calculate the average characters typed per group

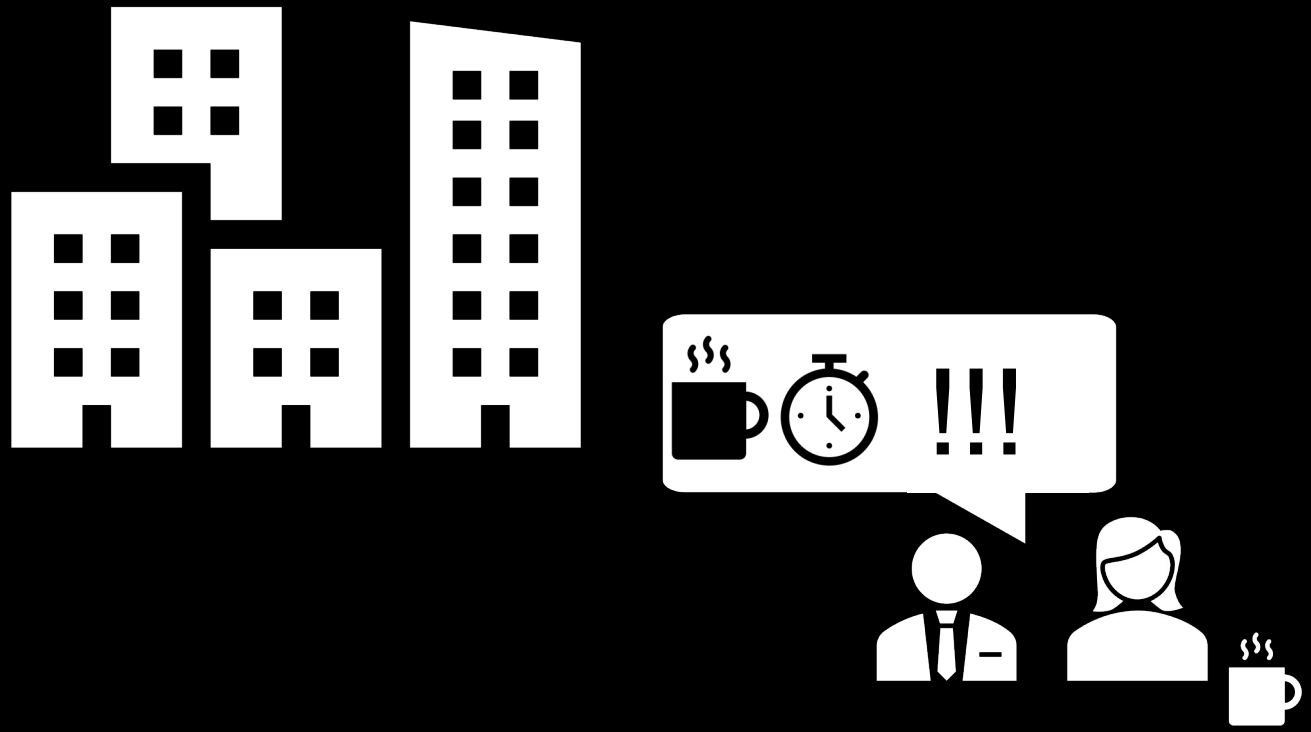


- Independent variable is **Group (between; blinded)**
- Independent variable is **time (within)**
- The dependent variable is **Characters per hour**

→ **Idealy groups are equivalent on DV at the pre-test and differ only in the post-test**

Other experimental designs

- Pre-post design: One experimental group (no control) is measured before and after intervention
- Ex-post Facto design: Groups are formed based on existing characteristics (e.g., people with schizophrenia or not)
- Double-blind study: Participant and experimenter do not know in which condition the participant is(often in RCT)



Effects to control for

- **Randomizing order**
  - **Learning effects**: changes in a participant's performance on a task or measure as a result of practicing or becoming more familiar with the task
  - **Carry-over effects**: influence of one treatment or condition on a subsequent treatment or condition in a study
- **Consider the experiment as an experience**
  - **Response bias**: participants do not respond to a question or task in a neutral or unbiased way
  - **Fatigue**: physical or mental exhaustion that can affect a participant's performance on a task or measure

Effects to control for

- **Blinding**
  - **Demand characteristics**: participant's perception of what the experimenter expects or wants them to do in a study
  - **Experimenter bias**: influence of the experimenter's expectations or beliefs on the results of a study
- **Placebo-control**
  - **Placebo effects**: changes in a participant's behavior or symptoms that occur as a result of receiving a "placebo" treatment, which is a treatment that has no active ingredients and is not expected to have any therapeutic effect

Checklist for a study design

- ❑ Are confounding variables controlled?
- ❑ Is the statistical model for analysis known and applicable?
- ❑ Are dependent and independent variables properly defined?
- ❑ Is randomization applied where possible?
- ❑ Does the design allow for the evaluation of the hypothesis?
- ❑ Is the study feasible?
- ❑ Are participants accessible?
- ❑ Does the design violate ethical principles?
- ❑ Do you have appropriate statistical power?

Summary

Study designs and variables

- Studies need to be carefully designed to identify cause-effect relationships
- Intervention studies can be tested with RCT designs
- Studies must be carefully designed in order to control for influencing variables/effects that could render results invalid (e.g., placebo effect)

Classroom exercise

Work with the person next to you (5 minutes)

What is the dependent and independent (within vs. between) variable in these studies?

1. “Participants in the exercise group lost more weight as compared to participants in the non-exercise waiting group”
2. “We measured skin conductance in response to 10 happy facial expressions and 10 angry facial expressions. Participants saw all stimuli 10 times leading to overall 200 trials within the experiment.”
3. “We found that Aalto students with better grades reported to spend more time reading the news (in h/day).

Questions?

Self-assesment

Lecture 2

1. What is the difference between Falsification and Bayesianism in hypothesis testing?
2. You are conducting a usability study for a novel transportation app. The investors want to see in an empirical study that users it's more fun for participants to use the novel app and that the booking works more smoothly.
   - Operationalize the variables and propose a study design.