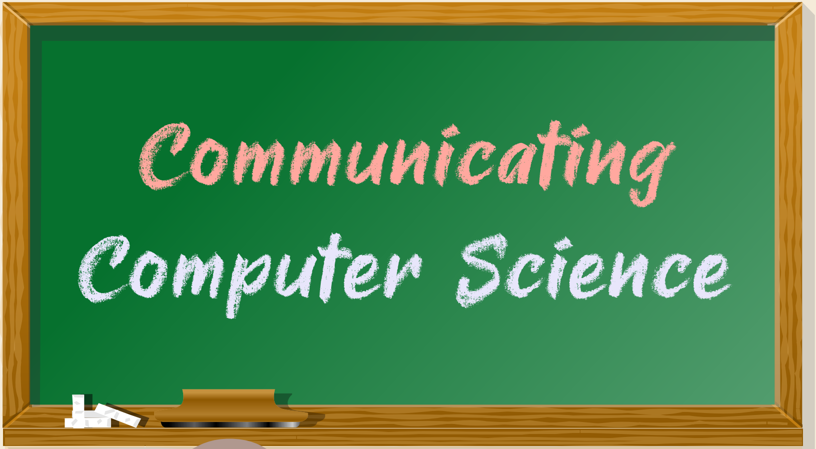



# 2 Empirical Science & Statistics

*14 October 2022*

Sebastian Wild

COMP335 (2022-23)
University of Liverpool
*version* 2022-10-14 07:50 H

# Outline

# Goals for today

1. Recognize key vocabulary for empirical research
2. Gain intuition behind statistical tools
3. Be aware of common pitfalls

Background:

- CS is not a *natural* science
  - one half is a structural science like mathematics, with theorems and proofs
  - the other half is an engineering discipline, with toolboxes and best practices

$\rightsquigarrow$ scientific method not part of standard CS curriculum

- but: many parts of CS use empirical science
  (and might profit from a more structured methodology . . . )

# Terminology

⚠ *We use many terms in day-to-day lingo with a different meaning.*

## Colloquial use

- **theory** = unproven suspicion/opinion

  *"Detectives are working on a theory that he knew his murderer."*

- **experiment** = trying something out

  *"Artists now experiment with many media, from canvas to computers."*

- **statistics** = any systematic collection or presentation of (numerical) facts

  *"The statistics show that computer science is not a popular GCSE subject."*

- **significant** = important, big, impactful

## Scientific meaning

- **theory** = set of ideas intended to explain something about life or the world

  *Darwin's theory of evolution*

- **experiment** = carefully designed scientific test, must be reproducible

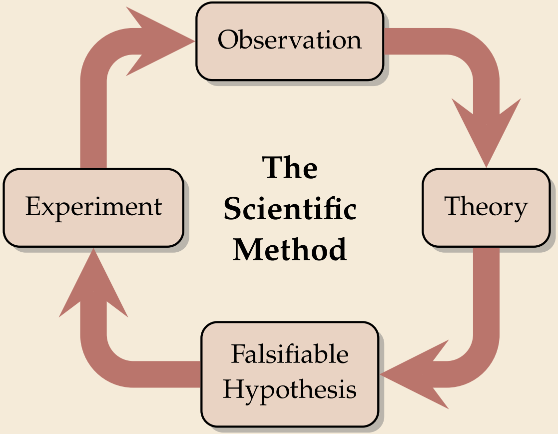
  *Gregor Mendel's inheritance experiment using pea plants*

- **statistics** = science of collecting and analyzing numerical data, hypothesis testing and inference, sampling

- **statistically significant** = $p$-value, i. e., probability to see the data assuming the null hypothesis, is below a threshold (often 5% or 1% or 0.5%)

## 2.1 The Scientific Method

# Cycle of Science




1. From an *observation* about the real world, we formulate a *theory* (= how things work)
2. To test whether a theory is true, we derive a *hypothesis* that follows from the theory. It must be possible prove this hypothesis *wrong* by factual observations (*falsifiable*).
3. We design an *experiment* that will either prove the hypothesis wrong, or will *fail* to prove the hypothesis wrong. In the latter case, it *supports* the theory.
4. The experiment might lead to new observations and refined theories.

*Note: The scientific method cannot prove a theory correct!*
*We can only collect supporting evidence (or refute it).*

# Falsifiable Hypotheses

- Not all hypotheses can be tested scientifically.
  - Example: *There is an almighty God.*
  - ⇝ We may never find any evidence of such a God.
    But it can always be claimed to have been God's will not be discovered.
    After all, God is almighty.
  - ⇝ There is no possible evidence that would irrevocable disprove this hypothesis.
- ⇝ Science is restricted to **falsifiable hypotheses**.

- The world is not usually black and white.
  - Evidence usually contains imperfections
    - theories are idealized (simplified) models of reality
    - measurements include inaccuracies/noise
  - ⇝ Need a *fuzzy/quantitative* way of falsification:
    **To what extent** does given evidence support or disprove a hypothesis?
  - ⇝ That's why we need *statistics*

# Running example

Running example for this lesson:

- **Theory:** Expectancy-value theory of motivation;
  being in control improves self-efficacy, hence expectancy,
  and hence (hopefully) performance
- **Hypothesis:** Giving students a **choice** for their assessment
  improves their learning outcomes.

  (in the specific context of teaching X in environment Y to age group Z . . . )

⇝ *How can we test this hypothesis?*

# Controlled experiments

**Hypothesis:** *Giving students a **choice** for their assessment improves marks.*

Gold standard of empirical science: ✨ *The Controlled Experiment* ✨

1. Randomly assign half of students to *experimental group* and *control group*
2. Experimental group gets the *choice* which of two essay topics $A$ or $B$ they write about.
3. Control group gets essay topics $A$ and $B$ *randomly assigned* by teacher.
4. Both groups' essays are marked according to same marking scheme.
5. **Analyze data:** If marks in experimental group are *"better"* than in control,
   we *support* our hypothesis.
   Other we *refute* it.

need a reliable way to determine that in presence of random variance!

## 2.2 Concepts of Statistics

# Null Hypotheses

**Hypothesis:** *Giving students a **choice** for their assessment ~~improves~~ affects marks.*

- Apparent problem: The scientific method can convincingly *refute* hypotheses, but not *prove* them correct.

Apply it to the *negation* of our hypothesis!

- ***Null hypothesis $H_0$***: $\mu_1 = \mu_0$
  $\mu_1$: Average mark for students with choice
  $\mu_0$: Average mark for students without choice

⇝ If experiment refutes $H_0$, we have evidence for the *alternative hypothesis* $H_1$: $\mu_1 \neq \mu_0$.

# Hypothesis testing

How to refute $H_0$?

- Idealistic way to refute $H_0$: test all humans . . . ✓
- unless that is possible (census!), must resort to (small) sample of population
  - How can we know whether our sample is large enough?
  - What if, by chance, we assigned all strong students to one group?

$\rightsquigarrow$ Inherently have to deal with randomness

$\rightsquigarrow$ statistical tests

# Statistical significance

- Statistical hypothesis tests are given
  - hypotheses $H_0$ and $H_1$
  - Observations $X_c$ and $X_e$ from control group and experimental group
  - Computes ***p-value*** = likelihood of data $(X_c, X_e)$ **assuming $H_0$ is true**
  - Many different tests depending on form of hypotheses and data   $Z$-test, $t$-test, ANOVA, $\chi^2$ test, ...

    But: Basic principle always computes

$$\text{test statistic} = \frac{\text{observed data} - \text{expected data assuming } H_0}{\text{average variation}}$$

    large test statistic $\iff$ small $p$-value

- In our example (using a $t$-test)
  - observed data = difference in average marks between control and experimental group
  - expected data = 0   (no difference under $H_0$)
  - average variation = $\sqrt{\frac{\sigma_c^2}{n_c} + \frac{\sigma_e^2}{n_e}}$ standard error of the two groups
  - $p$-value can be computed from Student $t$-distribution

## 2.3 Pitfalls & Dangers

# Correlation vs. Causation

- Controlled experiments are not always possible.
  - can be unethical
    *"The Forbidden Experiment" (language deprivation experiments with children)*
  - or impractically expensive

⇝ resort to ***observational study***

- record data just by observing
  - ⇝ cheap, as it can be done after the fact, using available data!
  - lots of examples, e. g., POLAR data on higher education participation

***Can only ever observe <u>correlations</u>,***
***but cannot infer causal relationships.***

- For example, might find correlation between post code average income and HE participation.
- But cannot infer from this data whether higher income *brings* children into unis or
  whether uni degrees *generate* higher income,
  or neither!

- Not all correlations are meaningful! `https://tylervigen.com/spurious-correlations`

# My taxonomy of (CS) education articles

Articles in educational research venues have a huge range

- Personal experience report
- Experience report with feedback analysis
- Observational study
- Controlled experiment    A/B testing

The analysis (was it "good"?) can likewise differ

- qualitative experiences (e. g., testimonials)
- quantitative performance metrics (e. g., test scores)

# The "Replication Crisis"

Is Most Published Research Wrong?

- Some classic scientific results could not be *reproduced*
  - some false positives inherently expected
  - misaligned incentives in "publish or perish"
  - clickbait mentalities
  - publication bias (only positive studies published)
  - selection bias (unrepresentative survey participants)
  - conflict of interest through funding

- Some concerns less relevant for education, but be skeptical of
  - binary "statistical significance"
  - studies with unclear setup

# Misunderstanding $p$-values

- Recall: ***p-value*** = likelihood of data $(X_c, X_e)$ **assuming $H_0$ is true**
- It is **NOT** the likelihood that $H_0$ is true!

- Also, a statistically significant rejection of $H_0$ might just say: There is *some* difference.
- But: The difference might not be "significant" (large) in value!