Hypothesis testing

Basic concepts and examples

Ivano Malavolta

VU VRIJE UNIVERSITEIT AMSTERDAM

LOOKING FURTHER

Quick Recap: Hypotheses

Hypothesis: a **formal** statement about a phenomenon

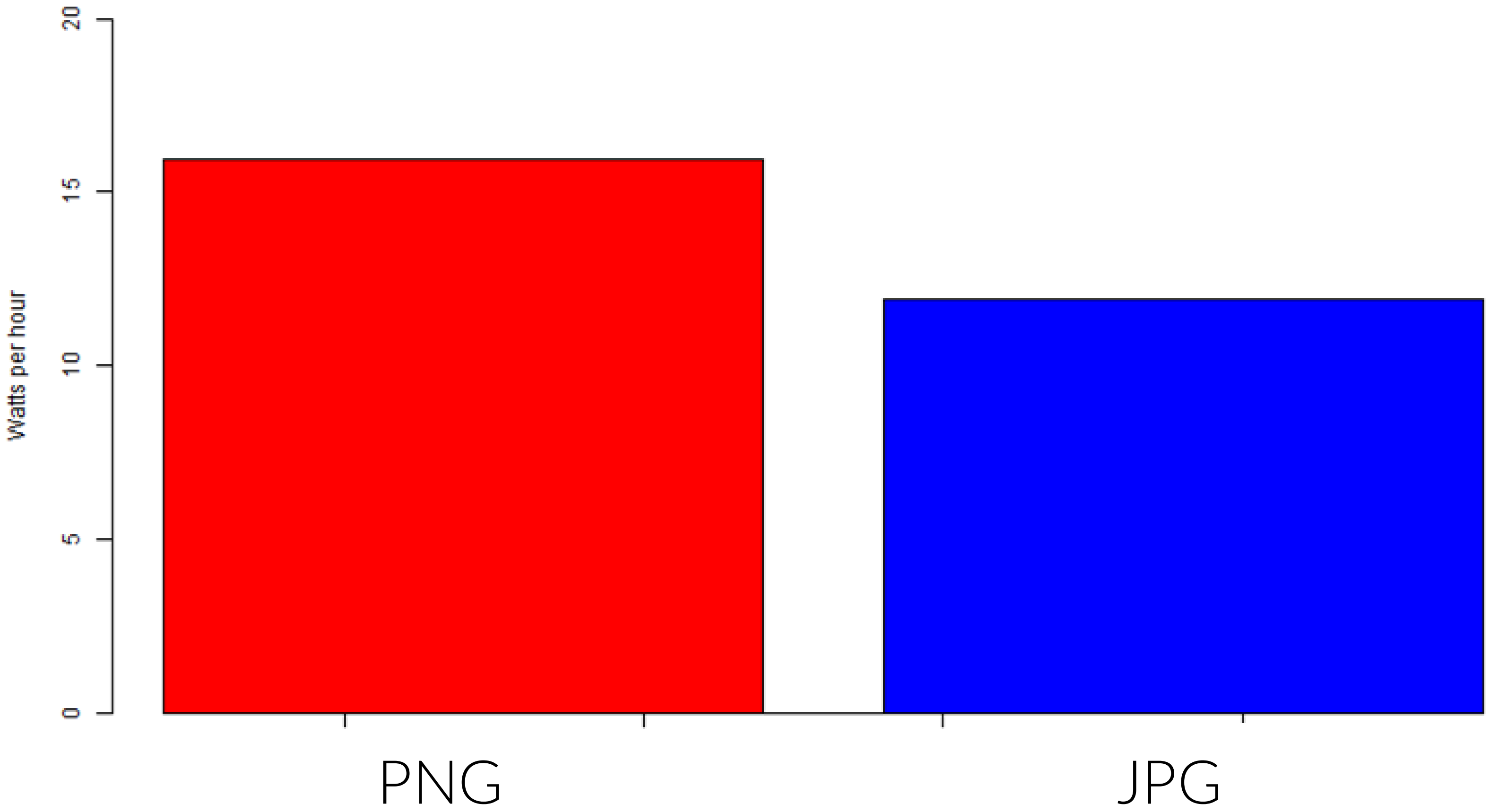
- **Null** hypothesis H_0: no real trends or patterns in the experiment setting
- **Alternative** hypothesis H_a: there are real trends or patterns in the experiment setting

Type of Hypotheses

- **Research Hypothesis:** a statement of what we believe will be the outcome (from GQM questions)
  e.g. "Using different image encoding algorithms implies different energy consumption".
- **Statistical hypothesis:** the formalization of our research hypothesis.
  e.g. $H_1: avg(P_{PNG}) < avg(P_{JPG})$

- Hypotheses may be **only** rejected, **never** confirmed!

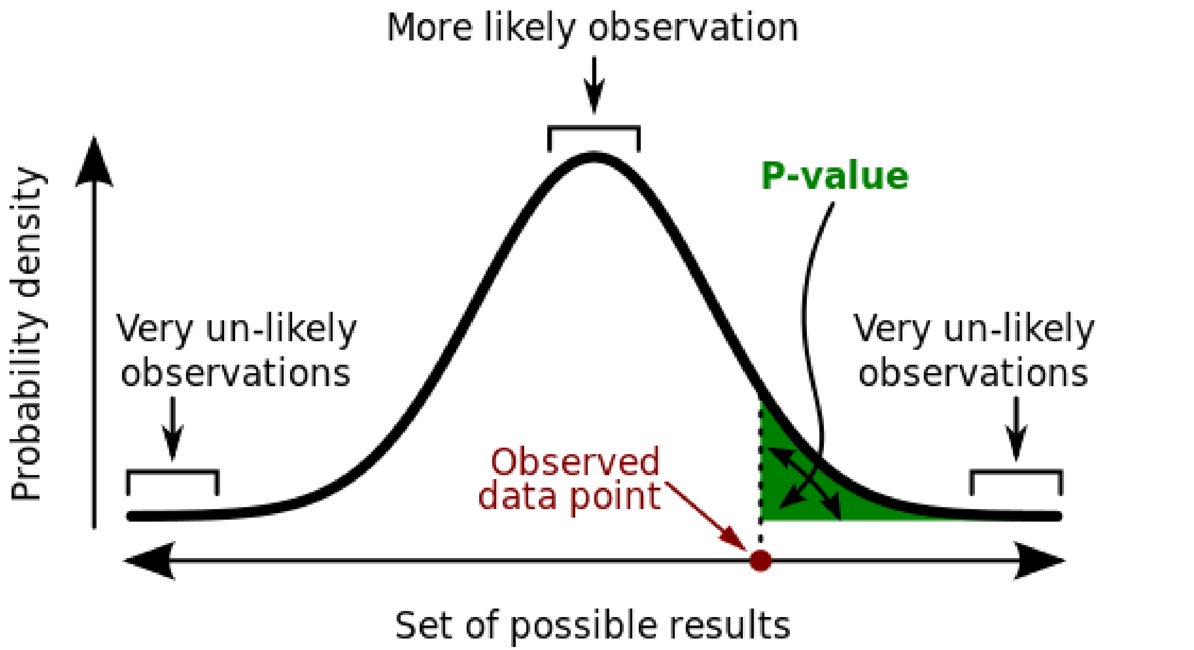
Observation and Significance



...how do we know we were not just lucky?

Observation and significance: p-value

- **p-value:** the probability of obtaining an effect at least as extreme as the one in our sample data
  - if the null hypothesis is true



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

Observation and significance: p-value

- p = Pr(observation | H_0)

- If the P-value is “low enough”, we can **reject** the null hypothesis (i.e., consider it extremely unlikely)

- If the P-value is close to 1, there is no difference between groups other than that due to random variation

  → the null hypothesis is confirmed

Observation and significance: p-value

		H_0	
		True	False
Test decision	Reject	Type I Error	OK
	Fail to reject	OK	Type II Error

Observation and significance: p-value

- Type I error (false positive)
  - we conclude the existence of a trend\pattern when there actually is not
  - $\alpha = Pr(\text{reject } H_0 \mid H_0 \text{ is true})$

		H_0	
		True	False
Test decision	Reject	Type I Error	OK
	Fail to reject	OK	Type II Error

- Type II error (false negative)
  - we neglect the existence of a trend\pattern when there actually is one
  - $\beta = Pr(\text{confirm } H_0 \mid H_0 \text{ is false})$

Observation and Significance: power

- Power: the opposite of type II errors
  - $1 - \beta$

- Power is the probability of actually observing a true effect

Observation and significance: cut-offs

- α = 0.05 (5%)
  - Confidence = 1 - α
  - If $p < 0.05$ we are 95% confident of rejecting H_0
- β = 0.20 (20%)
  - Power = 80%
  - We allow a 20% rate of false negatives
- Those cut-offs values are **empirically defined**

p-value: example

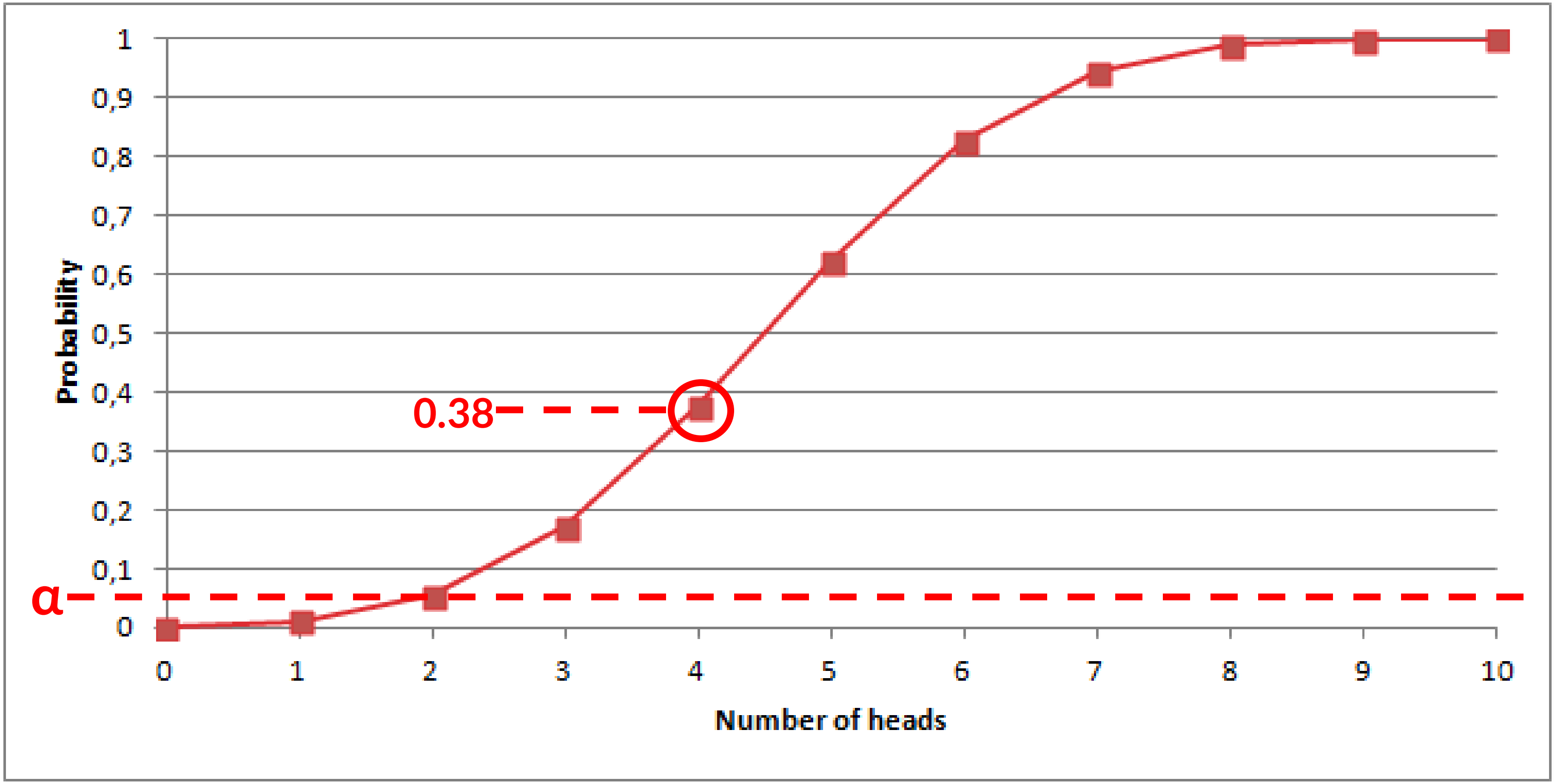
- Conjecture: the coin is tricky and disfavours heads
- Consequence: after a series of tosses, number of heads is smaller than number of tails
- Hypotheses
  - H_0: Heads = Tails = #Tosses/2
    - assuming that the initial position of the coin is balanced [1]
  - H_1: Heads < Tails
    - $\alpha = 5\%$

p-value: example

- Experiment Result: 4 heads in 10 trials

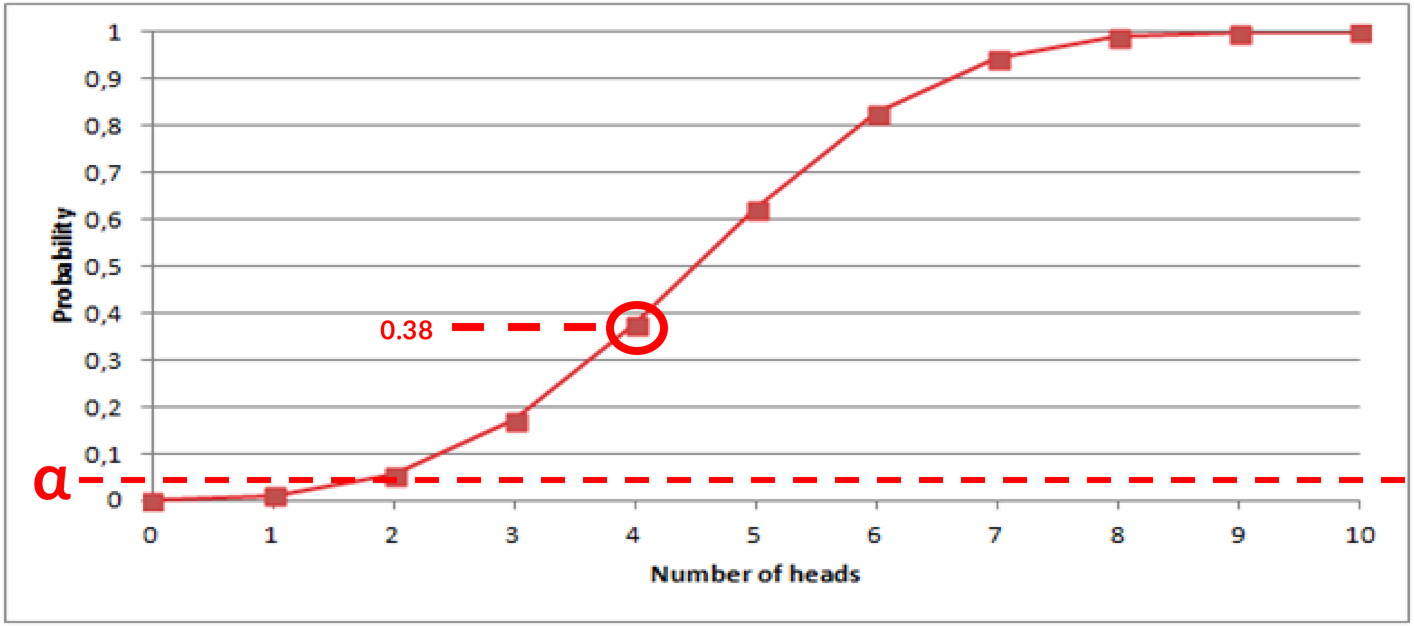
- *Hypothesis testing*: **Assuming H_0 is true,** what is the probability of having 4 or less heads in 10 trials?
  - p-value
- Binomial distribution
  - probability of heads/tails : 0.5
  - number of trials: 10

p-value: example



p-value: example

- p-value > α
  - we cannot reject the null hypothesis
- If we had 2 or less heads in 10 trials, we could have



Hypothesis testing: pitfalls

- Statistical significance **does not prove** causation
  - context analysis and study design are crucial

- Check sample size and power of your test
  - "evidence of no effect" is rather "**no evidence of effect**"
  - "Low statistical power" validity threat

- Over-emphasis over p-value
  - a significant p-value does not mean effect is relevant

Readings

Chapter 10
(section 3)

Chapter 6

Acknowledgements

- Coin toss example and other content from *Empirical Methods in Software Engineering*, Marco Torchiano, Politecnico di Torino - http://softeng.polito.it/EMSE/
- Giuseppe Procaccianti's lectures at VU