Introduction to Bayesian thinking and modelling

Gianluca Baio

Department of Statistical Science | University College London

g.baio@ucl.ac.uk
http://www.statistica.it/gianluca/
https://egon.stats.ucl.ac.uk/research/statistics-health-economics/
https://github.com/giabaio
https://github.com/StatisticsHealthEconomics
@gianlubaio

Know your Professor, UCL Statistical Society

28 October 2021

Disclaimer...

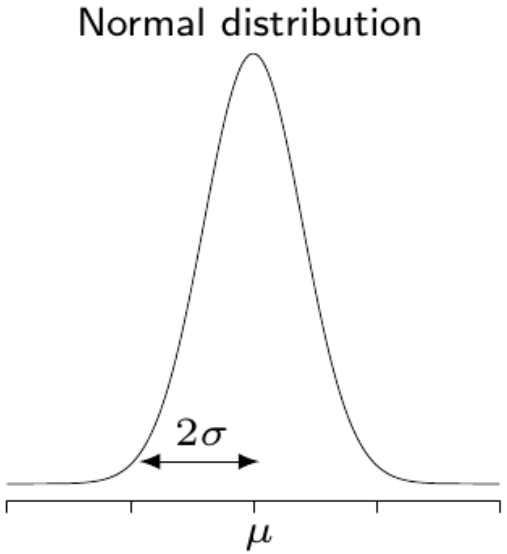
...Just so you know what you're about to get into... 😉

Summary

- Sampling variability
  - Probability calculus vs Statistics
- Deductive inference
  - "Standard" statistical methods
  - Confidence intervals & testing
- Inductive inference
  - Bayesian reasoning
  - Basic ideas
  - Forming "priors"
- Some examples

What is statistics all about?

- Typically, we observe some data and we want to use them to learn about some unobservable feature of the general population in which we are interested
- To do this, we use statistical models to describe the probabilistic mechanism by which (**we assume!**) that the data have arisen



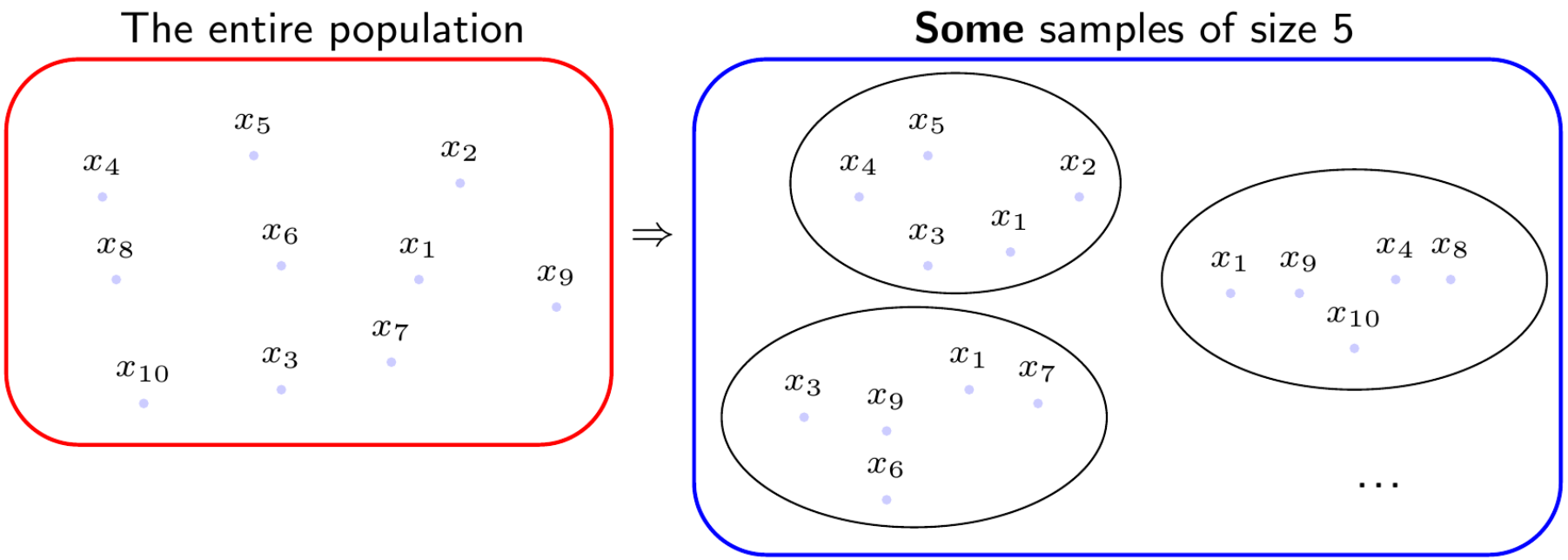
θ Parameters

$\downarrow$

$p(y \mid \theta)$ y Data

NB: Roman letters $(y$ or $x)$ typically indicate **observable data**, while Greek letters $(\theta, \mu, \sigma, \ldots)$ indicate **population parameters**

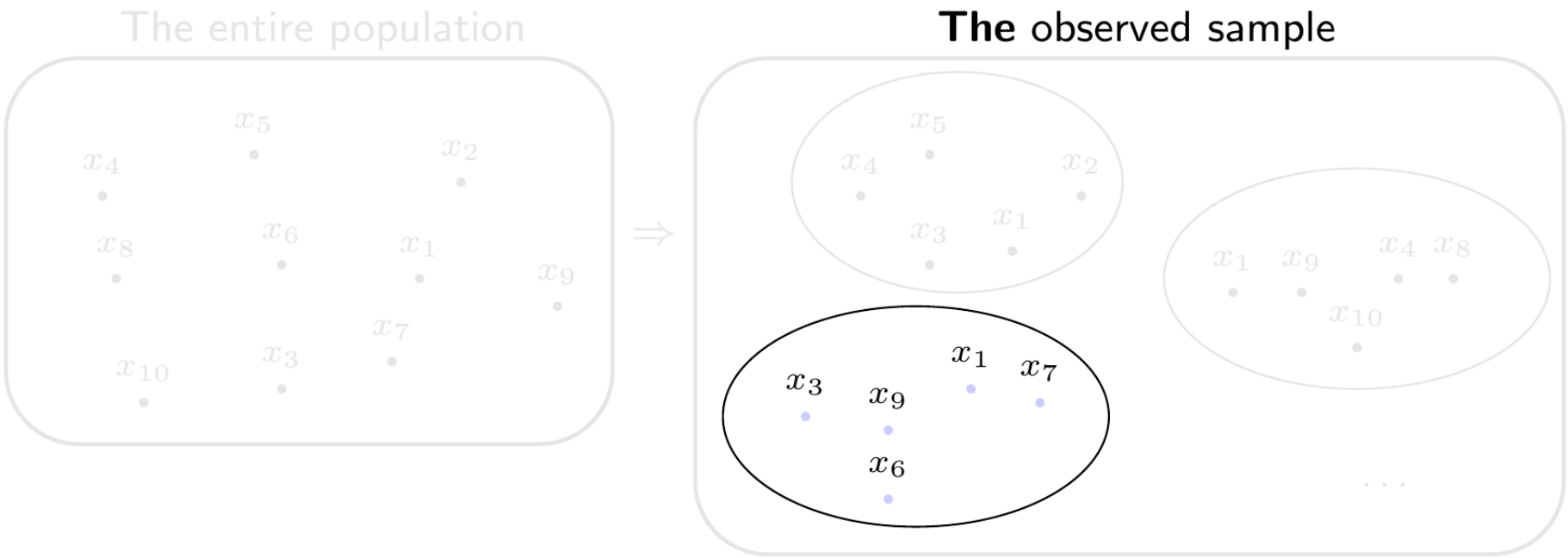
Sampling variability



- Population size $N = 10$
- "True" population Mean μ
- "True" Standard deviation σ

- Sample size $n = 5$
- Sample Mean $\bar{x}$
- Sample Standard deviation s_x

Sampling variability



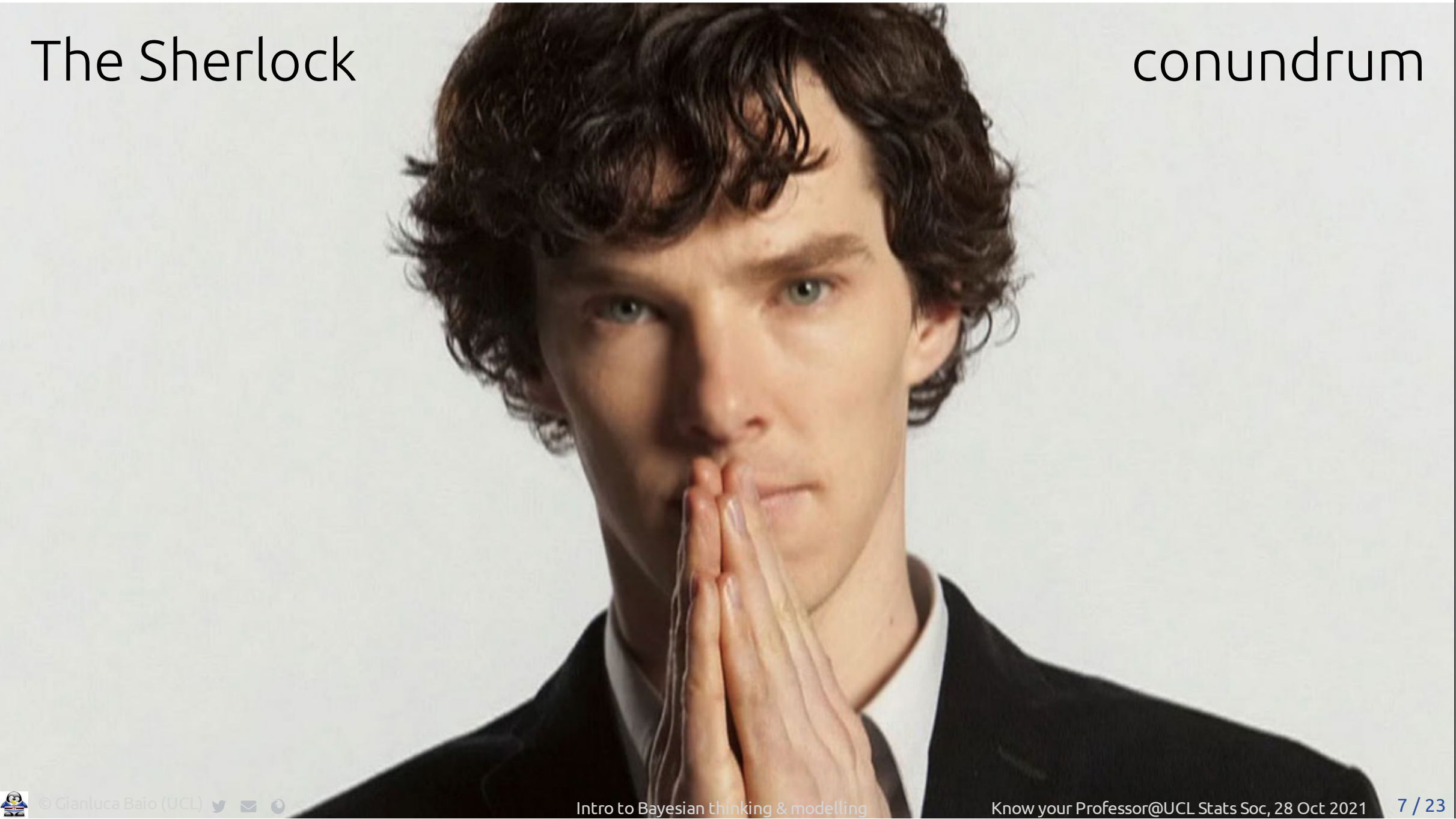
- Population size $N = 10$
- "True" population Mean μ
- "True" Standard deviation σ

$\Leftarrow$

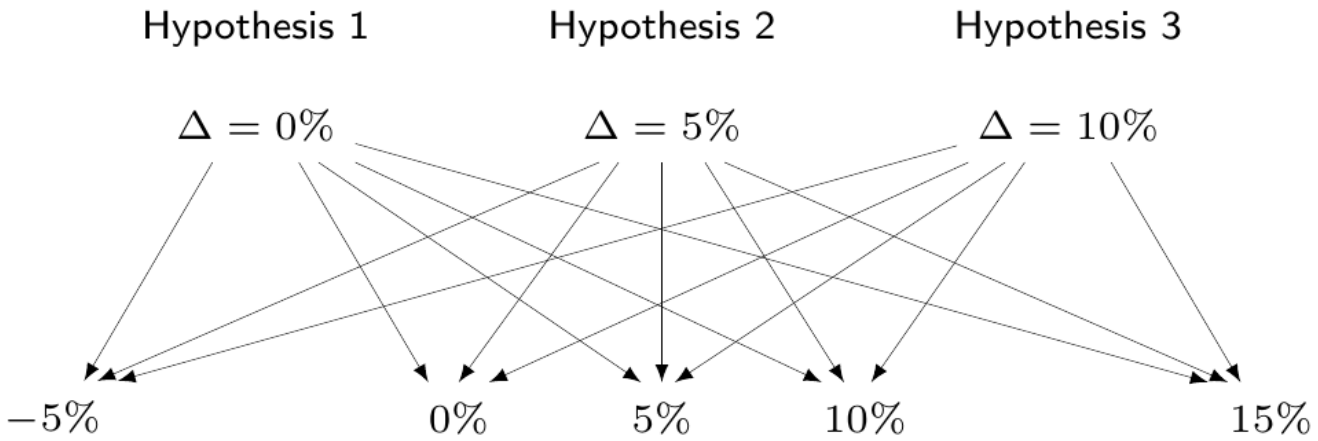
- Sample size $n = 5$
- Sample Mean $\bar{x}$
- Sample Standard deviation s_x

In reality we observe **only one** such sample (out of the many possible – in fact there are **252** different ways of picking **at random** 5 units out of a population of size 10!) and we want to use the information contained in **that** sample to **infer** about the population parameters (e.g. the true mean and standard deviation)

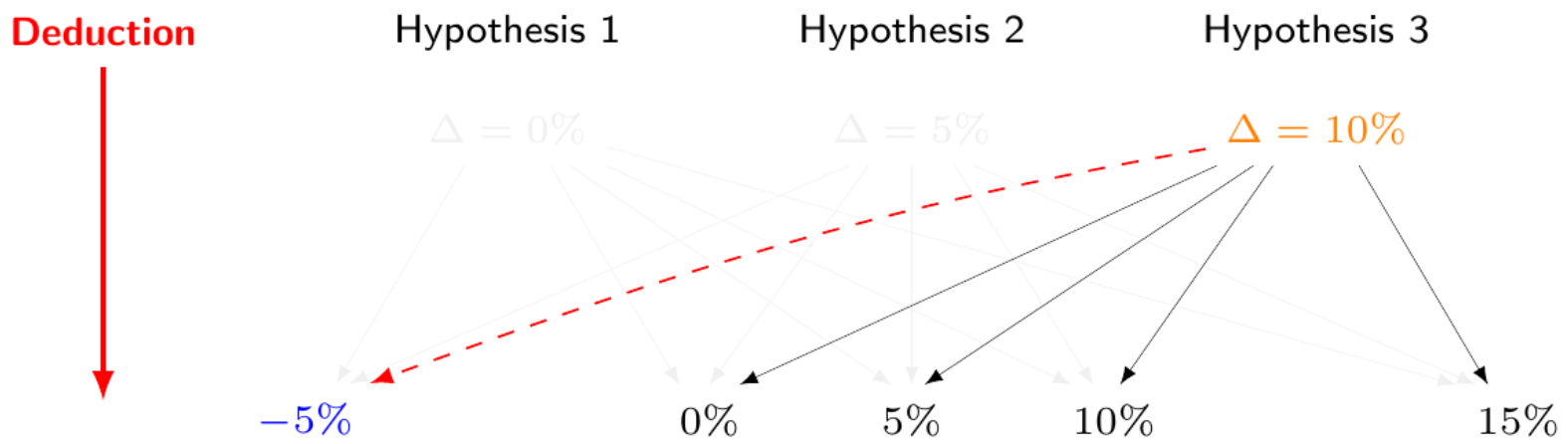
The Sherlock conundrum



Deductive vs inductive inference

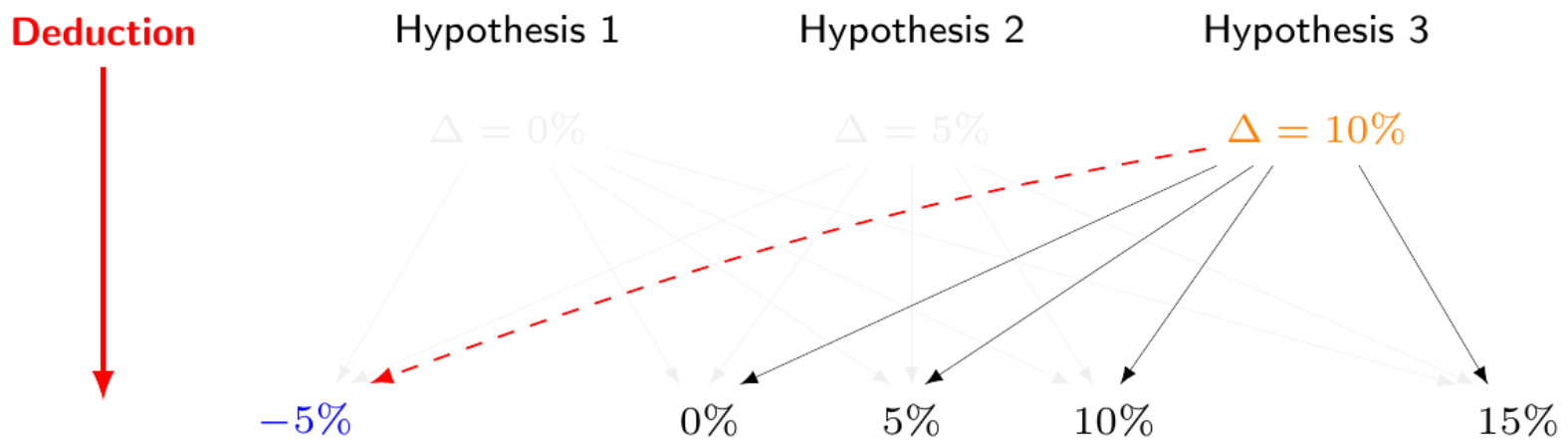


Deductive vs inductive inference



- Standard (frequentist) procedures fix the working hypotheses and, **by deduction**, make inference on the observed data:
  – If my hypothesis is true, what is the probability of randomly selecting the data that I actually observed? If small, then *deduce* weak support of the evidence to the hypothesis

Deductive vs inductive inference



- Standard (frequentist) procedures fix the working hypotheses and, **by deduction**, make inference on the observed data:
  - If my hypothesis is true, what is the probability of randomly selecting the data that I actually observed? If small, then *deduce* weak support of the evidence to the hypothesis
  - Assess $\Pr(\text{Observed data} \mid \text{Hypothesis})$
  - Directly relevant for standard frequentist summaries, eg p-values, Confidence Intervals, etc
  - **NB**: Comparison with data that could have been observed, but haven't!

Adapted from Goodman (1999)

Confidence intervals

⚠ See http://www.statistica.it/gianluca/teaching/intro-stats/interval-estimation.html!

Drug to cure headaches - "true" probability of success: $\pi = 40/73 \approx 0.55$

Confidence intervals

⚠ See http://www.statistica.it/gianluca/teaching/intro-stats/interval-estimation.html!

Drug to cure headaches - "true" probability of success: $\pi = 40/73 \approx 0.55$

- You get to see data for, say, $n = 10$ individuals, under the "true" **data generating process** (DGP): $\boldsymbol{y} = (y_1, \ldots, y_{10}) = (0, 0, 1, 1, 0, 1, 0, 1, 0, 1)$
- Can make estimates to infer from sample to population
  - Sample mean: $\bar{y} = \hat{\pi} = \sum_{i=1}^{n} \frac{y_i}{n} = \frac{5}{10} = 0.5$ Standard error: $\text{se}(\hat{\pi}) = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} = 0.16$

Confidence intervals

⚠ See http://www.statistica.it/gianluca/teaching/intro-stats/interval-estimation.html!

Drug to cure headaches - "true" probability of success: $\pi = 40/73 \approx 0.55$

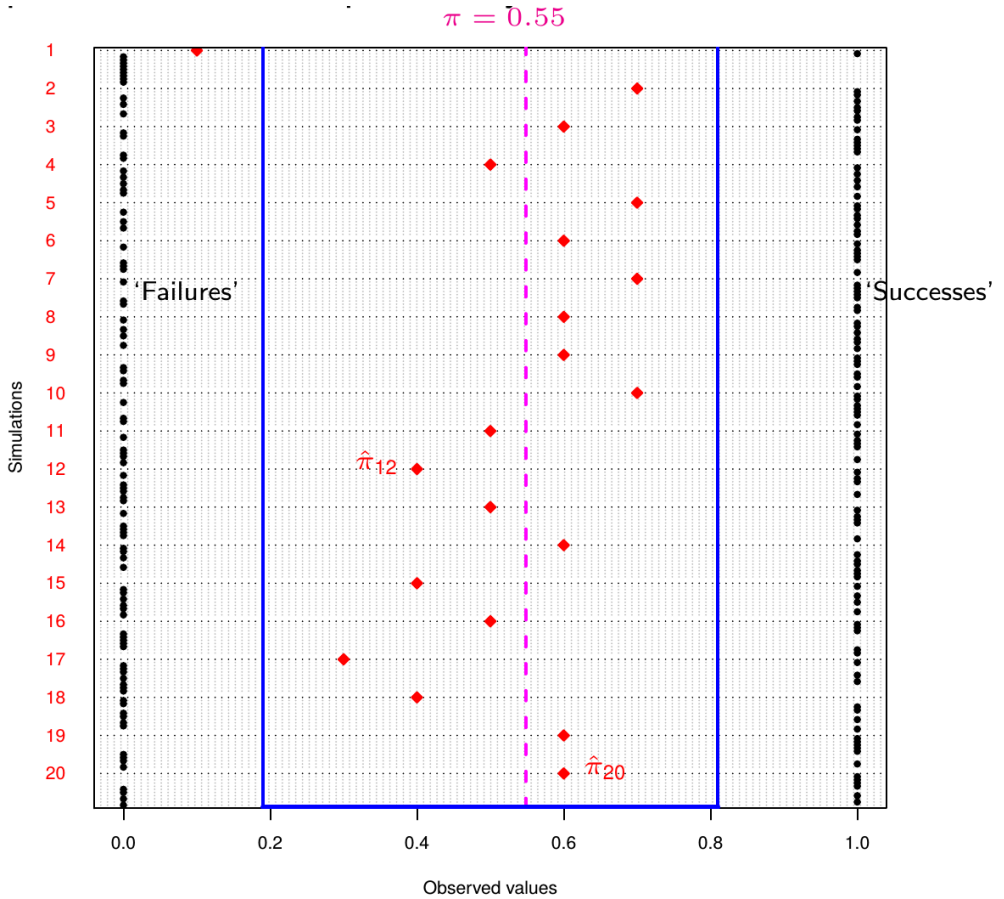
- You get to see data for, say, $n = 10$ individuals, under the "true" **data generating process** (DGP): $\boldsymbol{y} = (y_1, \dots, y_{10}) = (0, 0, 1, 1, 0, 1, 0, 1, 0, 1)$
- Can make estimates to infer from sample to population
  - Sample mean: $\bar{y} = \hat{\pi} = \sum_{i=1}^{n} \frac{y_i}{n} = \frac{5}{10} = 0.5$ Standard error: $\text{se}(\hat{\pi}) = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} = 0.16$
- Can compute the interval estimate (using some approximation/theoretical results...)

$$95\%\ \text{CI} \approx [\hat{\pi} - 2\text{se}(\hat{\pi});\ \hat{\pi} + 2\text{se}(\hat{\pi})] = [0.5 - 0.32; 0.5 + 0.32] = [0.19; 0.81]$$

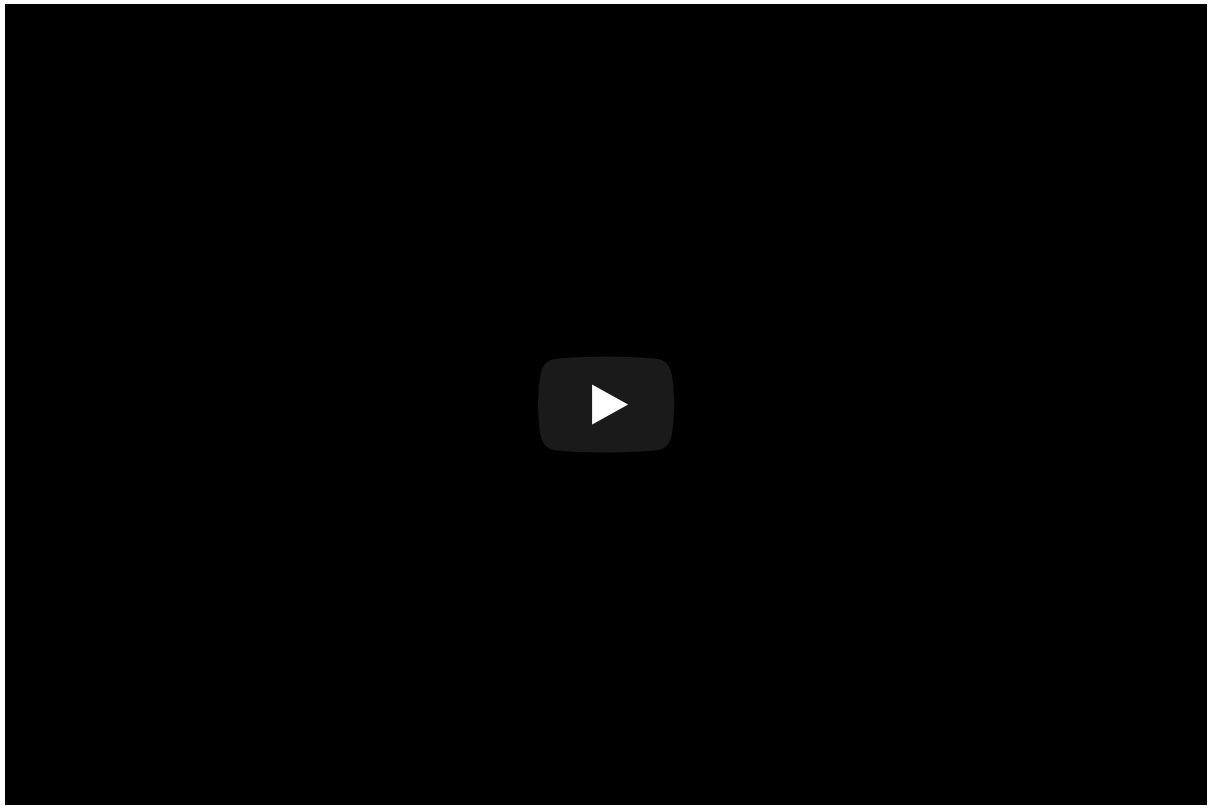
- Assuming the observed sample is representative of the DGP and using the sample estimates, **if** we were able to replicate the experiment over and over again under the same conditions, 95% of the times, the estimate for the "true" probability of success will be included in the interval $[0.19; 0.81]$
- **That** is how you interpret a 95% Confidence Interval!

Confidence intervals

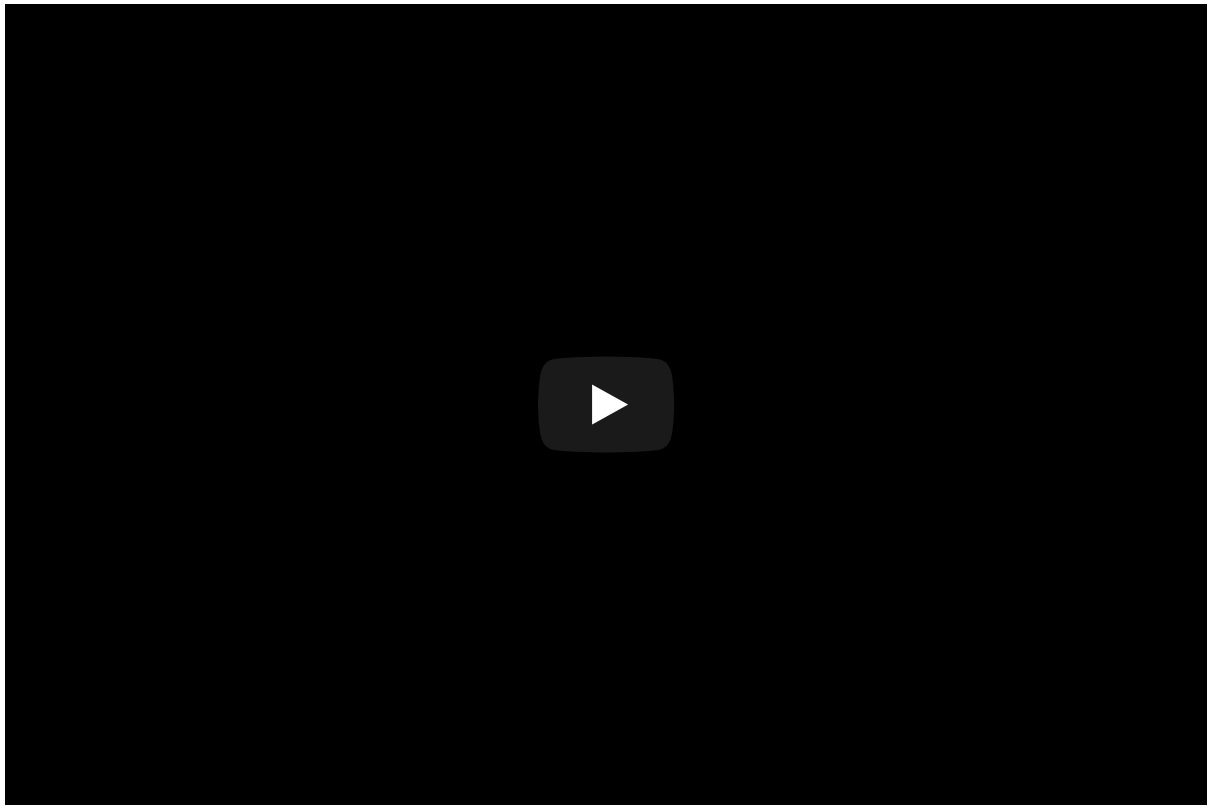
- Simulate n_{sim} (e.g. $= 20$) studies sampling data from a DGP assuming a "true" $\pi = 0.5$ (although in fact, $\pi = 0.55$!) and $n = 10$
- For each, estimate the probability $\hat{\pi}$



Confidence intervals



Sample size calculations



Sample size calculations

Designing a study

- Designing a study is just as important as analysing it
  – If we don't have "enough" information in the data, we won't be able to detect an underlying signal
- Related to "hypothesis" testing

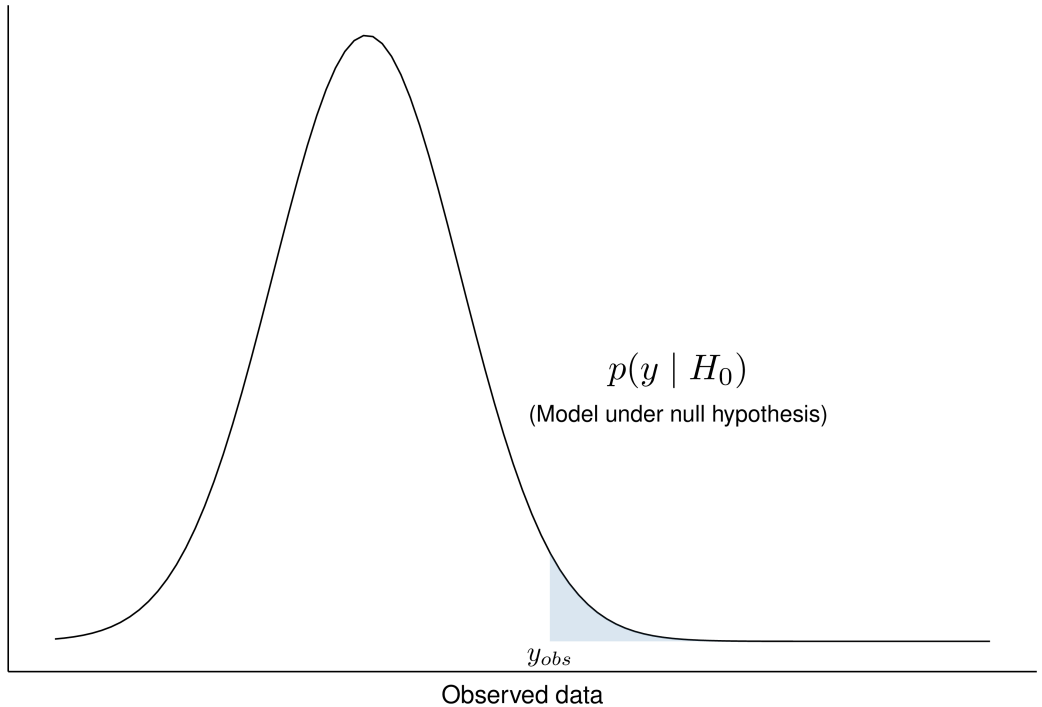
		"Null" hypothesis H_0	
		True	*False*
Decision on "Null"	*Reject*	Type I error α (False positive)	Correct inference (True positive)
	Fail to reject	Correct inference (True negative)	Type II error β (False negative)

1. Set the Type I error to some low level (**typically:** $\alpha = 0.05$)
2. Set the Type II error to some set level (**typically:** $\beta = 0.10$ or $\beta = 0.20$)
3. Define the "clinically relevant outcome" (eg difference in treatment effects), δ
4. Set an estimate of variability in the underlying population
5. Use assumptions about sampling variability and determine minimum number of observations to be able to detect δ

I need a pee(-value)...

Analysing a study

- Interpretation: Under the null hypothesis (ie **IF** it is true), what is the probability of observing something as extreme or even more extreme as the observed data?



- If $p < 0.01$ then **strong** evidence against the null hypothesis
- If $0.01 < p < 0.05$ then **fairly strong** evidence against the null hypothesis
- If $p > 0.05$ then **little or no evidence** against the null hypothesis

Two sides of the same coin?

- Often, hypothesis testing and p-values are seen as the same thing. **They are not!**

Two sides of the same coin?

- Often, hypothesis testing and p-values are seen as the same thing. **They are not!**
- **Hypothesis testing** (HT)
  - Considers formally two competing hypotheses – a "null" and an "alternative" (**NB**: that determines the treatment effect)
  - **Sets** the probabilities of error α and β
  - Aims at "rejecting" the null – so it has a binary outcome (yes/no)
- **Significance testing** (ST, p-values)
  - Concerned with the sampling distribution of the data under the null hypothesis
  - Measures the strength of the evidence for/against the null, but has no formal involvement of alternative explanations for the observed data

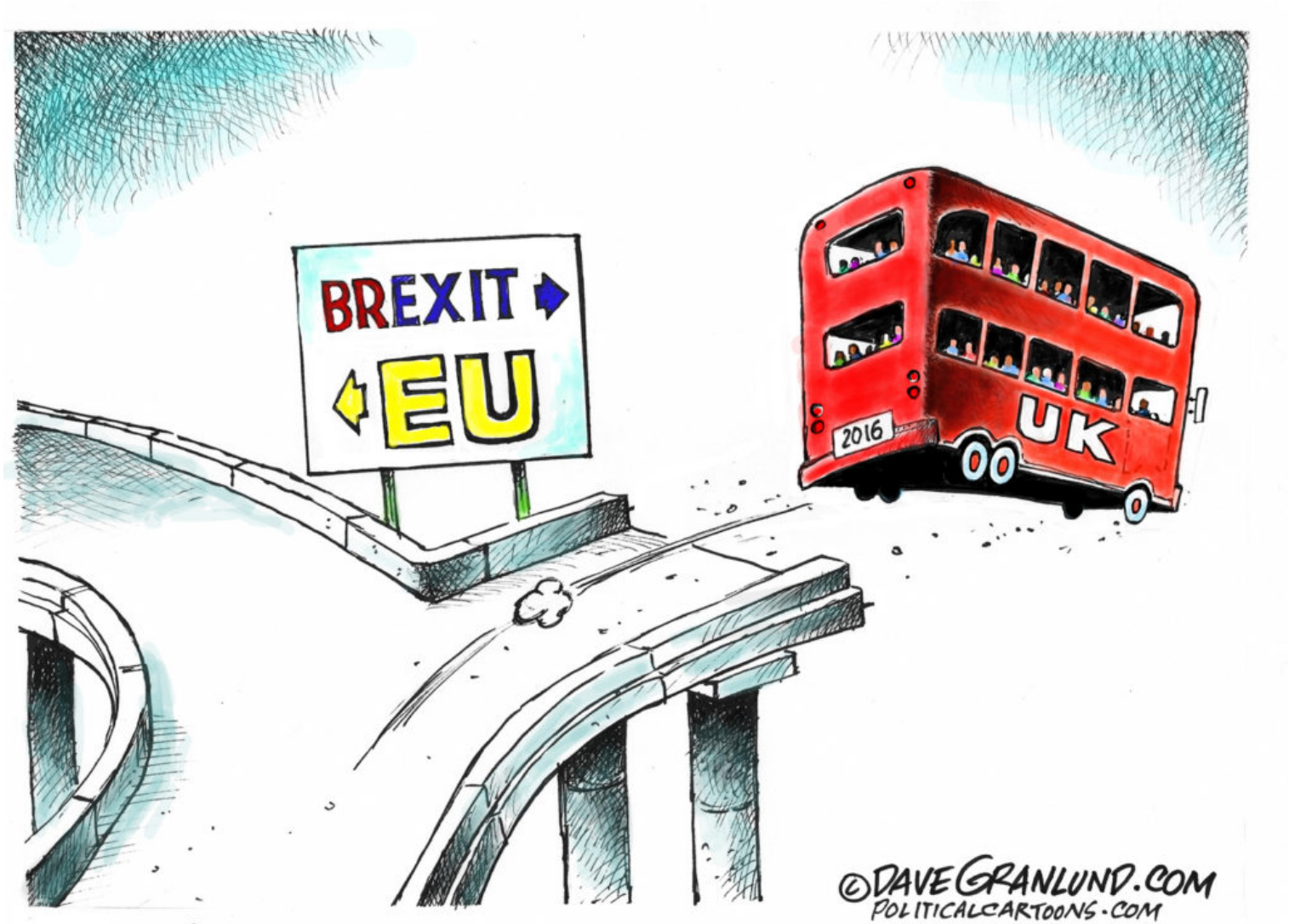
Two sides of the same coin?

- Often, hypothesis testing and p-values are seen as the same thing. **They are not!**
- **Hypothesis testing** (HT)
  - Considers formally two competing hypotheses – a "null" and an "alternative" (**NB**: that determines the treatment effect)
  - **Sets** the probabilities of error α and β
  - Aims at "rejecting" the null – so it has a binary outcome (yes/no)
- **Significance testing** (ST, p-values)
  - Concerned with the sampling distribution of the data under the null hypothesis
  - Measures the strength of the evidence for/against the null, but has no formal involvement of alternative explanations for the observed data
- $p \neq \alpha$ even if often the **threshold** is set at 0.05 for both!
  - α is **set** by the researcher
  - p is **computed** from the data (as extreme or more extreme than those observed)

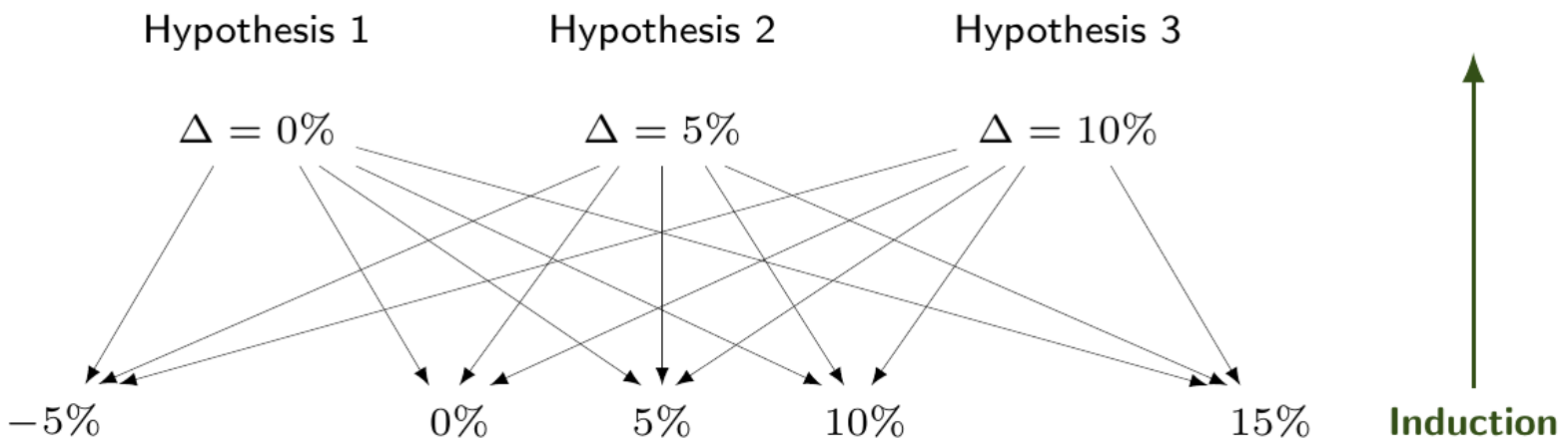
Two sides of the same coin?

- Often, hypothesis testing and p-values are seen as the same thing. **They are not!**
- **Hypothesis testing** (HT)
  - Considers formally two competing hypotheses – a "null" and an "alternative" (**NB**: that determines the treatment effect)
  - **Sets** the probabilities of error α and β
  - Aims at "rejecting" the null – so it has a binary outcome (yes/no)
- **Significance testing** (ST, p-values)
  - Concerned with the sampling distribution of the data under the null hypothesis
  - Measures the strength of the evidence for/against the null, but has no formal involvement of alternative explanations for the observed data
- $p \neq \alpha$ even if often the **threshold** is set at 0.05 for both!
  - α is **set** by the researcher
  - p is **computed** from the data (as extreme or more extreme than those observed)
- **NB**: Confusingly, experimental studies are designed under a HT setting, but analysed under a ST setting!
- Increasing recognition of pitfalls in science (here and here)

Is there another way?...

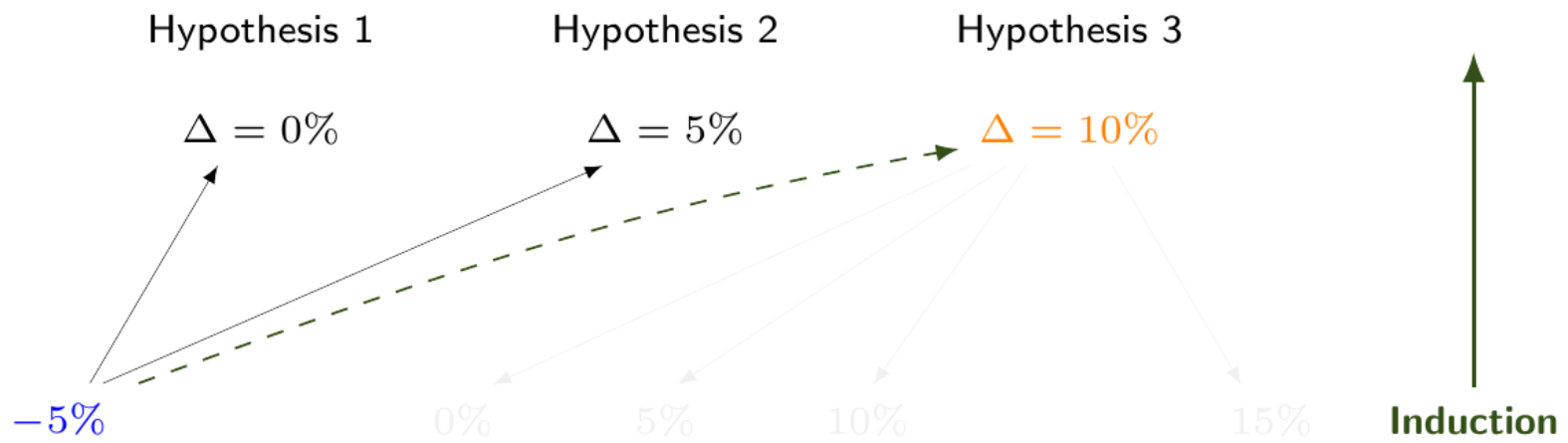


Deductive vs **inductive** inference



- The **Bayesian** philosophy proceeds fixing the value of the observed data and, **by induction**, makes inference on unobservable hypotheses
  – What is the probability of my hypothesis, given the data I observed? If less than the probability of other competing hypotheses, then weak support of the evidence to the hypothesis

Deductive vs **inductive** inference



- The **Bayesian** philosophy proceeds fixing the value of the observed data and, **by induction**, makes inference on unobservable hypotheses
  - What is the probability of my hypothesis, given the data I observed? If less than the probability of other competing hypotheses, then weak support of the evidence to the hypothesis
  - Assess $\Pr(\text{Hypothesis} \mid \text{Observed data})$
  - Can express in terms of an **interval** estimate: $\Pr(a \leq \text{parameter} \leq b \mid \text{Data})$
  - **NB**: Unobserved data have no role in the inference!

Bayesian inference

How did it all start?

In 1763, Reverend Thomas Bayes of Tunbridge Wells wrote

PROBLEM.

Given the number of times in which an unknown event has happened and failed: *Required* the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named.

In modern language, given $r \sim \text{Binomial}(\theta, n)$, what is $\Pr(\theta_1 < \theta < \theta_2 \mid r, n)$?

Some historical references

- http://www.bayesian.org/resources/bayes.html
- S. Bertsch McGrayne (2011). *The Theory That Would Not Die*
- S. Fienberg (2006). When did Bayesian inference become Bayesian?

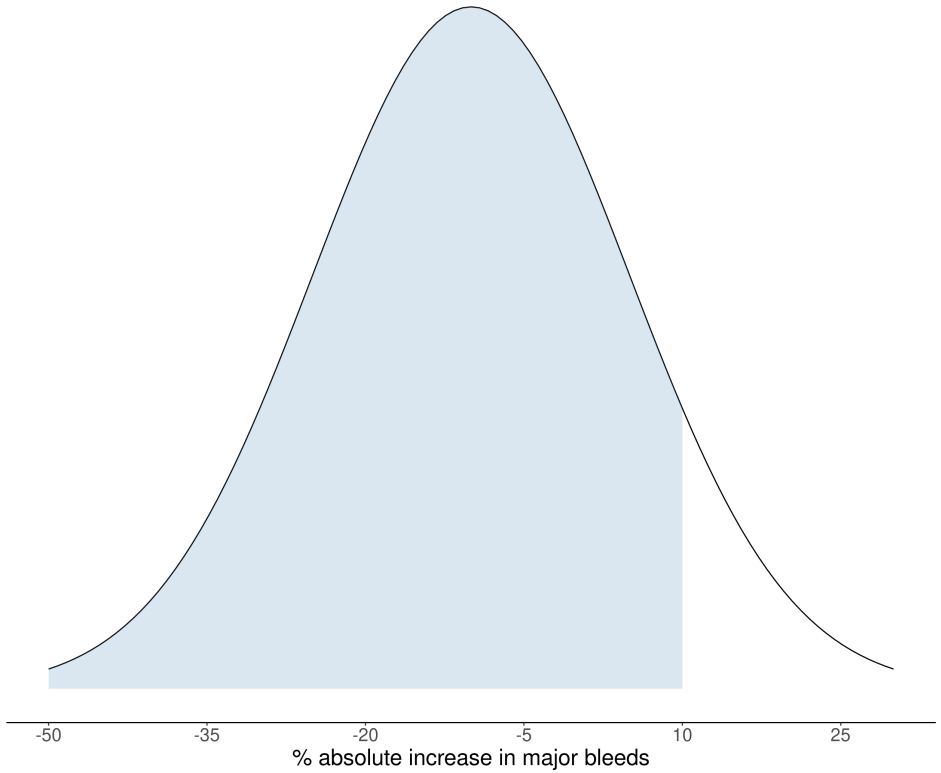
Bayesian inference

Basic ideas

Direct expression of uncertainty about unknown parameters

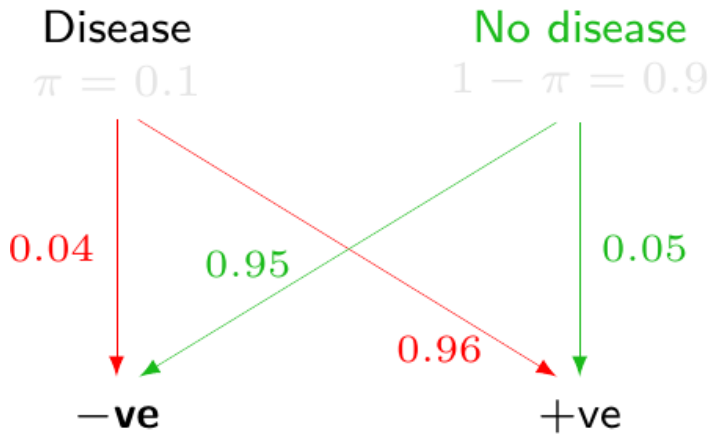
"There is an 89% probability that the absolute increase in major bleeds is less than 10 percent with low-dose PLT transfusions"

(Tinmouth et al, *Transfusion*, 2004)



Bayesian inference

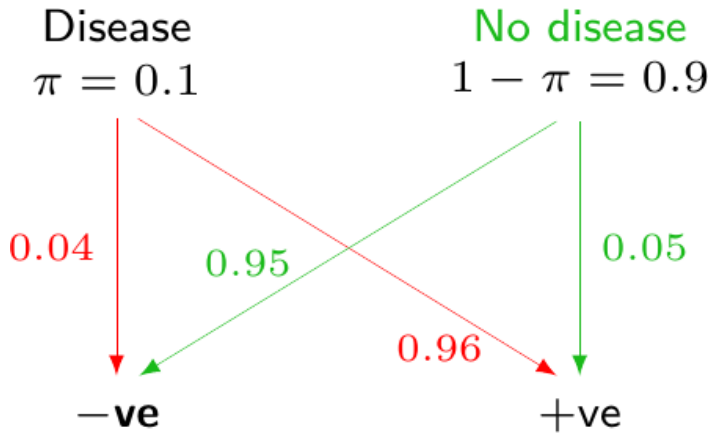
Basic ideas



- Suppose a patient is tested for HIV. The test comes up negative (–ve)
- Given the assumptions/model, this indicates **fairly strong** evidence against the hypothesis that the true status is "Disease", so basically $p = 0.04$

Bayesian inference

Basic ideas



- Suppose a patient is tested for HIV. The test comes up negative (–ve)
- Given the assumptions/model, this indicates **fairly strong** evidence against the hypothesis that the true status is "Disease", so basically $p = 0.04$
- But: how **prevalent** is the disease in the population?
  - We can model our prior knowledge about this and combine this information with the evidence from the data (using **Bayes theorem**)

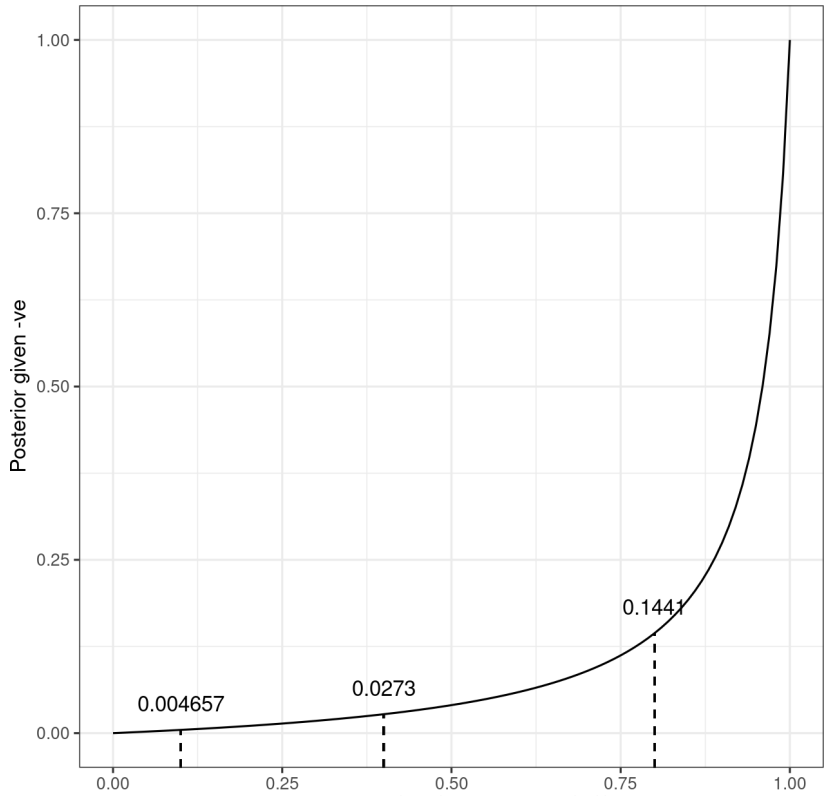
$$\Pr(\text{Disease} \mid \text{-ve}) = \frac{\Pr(\text{Disease})\Pr(\text{-ve} \mid \text{Disease})}{\Pr(\text{-ve})}$$

- Update uncertainty given the evidence provided by the data

Bayesian inference

Prior vs posterior

- The evidence **from the data alone** tells us that the observed result is extremely unlikely under the hypothesis of "Disease"
- This is strongly dependent on the **context**, as provided by the prior knowledge/epistemic uncertainty, though!



Bayesian inference

Basic ideas

- A Bayesian model specifies a **full probability distribution** to describe uncertainty
- This applies to
  - **Data**, which are subject to **sampling variability**
  - **Parameters** (or hypotheses), typically unobservable and thus subject to **epistemic uncertainty**
  - And even future, yet unobserved realisations of the observable variables (data)

Bayesian inference

Basic ideas

- A Bayesian model specifies a **full probability distribution** to describe uncertainty
- This applies to
  - **Data**, which are subject to **sampling variability**
  - **Parameters** (or hypotheses), typically unobservable and thus subject to **epistemic uncertainty**
  - And even future, yet unobserved realisations of the observable variables (data)
- Probability is the only languange in the Bayesian framework to assess any form of imperfect information or knowledge
  - No need to distinguish between probability and confidence
  - Before even seeing the data, we need to identify a suitable probability distribution to describe the overall uncertainty about the data $\boldsymbol{y}$ and the parameters $\boldsymbol{\theta}$

Bayesian inference

Basic ideas

- A Bayesian model specifies a **full probability distribution** to describe uncertainty
- This applies to
  - **Data**, which are subject to **sampling variability**
  - **Parameters** (or hypotheses), typically unobservable and thus subject to **epistemic uncertainty**
  - And even future, yet unobserved realisations of the observable variables (data)
- Probability is the only languange in the Bayesian framework to assess any form of imperfect information or knowledge
  - No need to distinguish between probability and confidence
  - Before even seeing the data, we need to identify a suitable probability distribution to describe the overall uncertainty about the data $\boldsymbol{y}$ and the parameters $\boldsymbol{\theta}$

$$p(\boldsymbol{y}, \boldsymbol{\theta}) = p(\boldsymbol{\theta})p(\boldsymbol{y} \mid \boldsymbol{\theta}) = p(\boldsymbol{y})p(\boldsymbol{\theta} \mid \boldsymbol{y})$$

from which we derive Bayes Theorem

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{\theta})p(\boldsymbol{y} \mid \boldsymbol{\theta})}{p(\boldsymbol{y})}$$

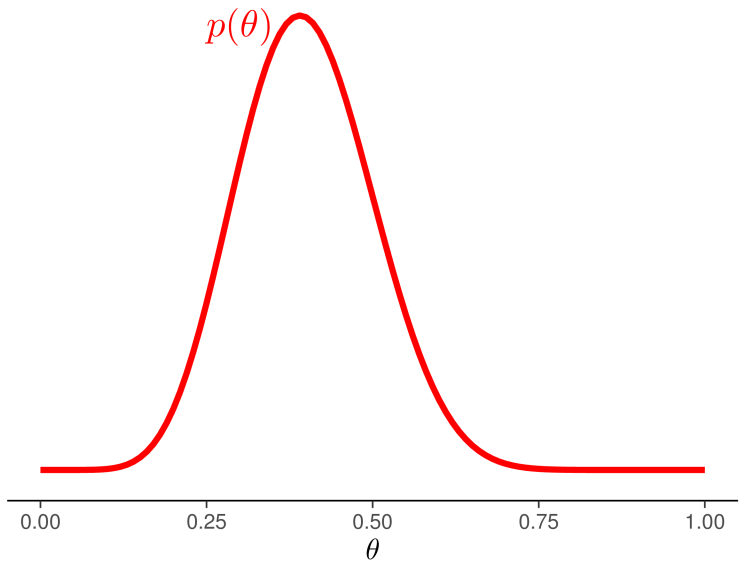
- **Express beliefs in form of a probability distribution**

Bayesian modelling

(Super) silly example: drug

Existing knowledge

- Population registries
- Observational studies
- Small/pilot RCTs
- Expert opinion



Encode the assumption that a drug has a response rate between 20% and 60%

Bayesian modelling

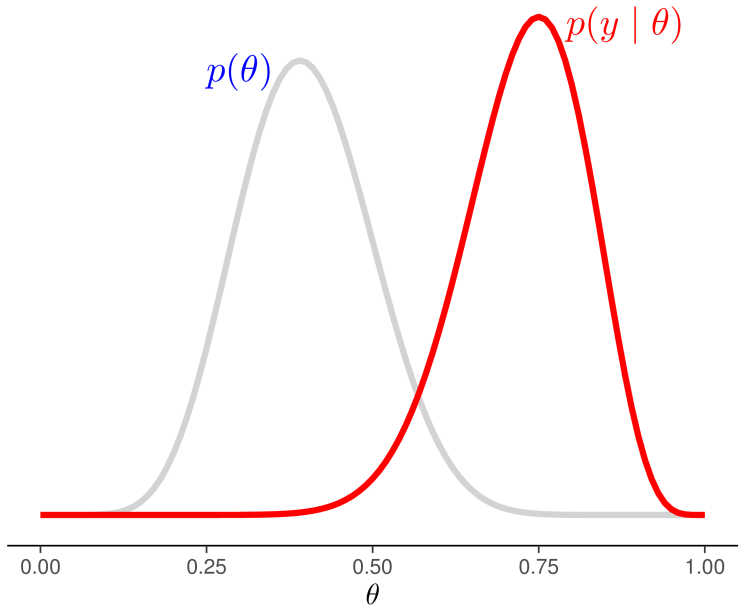
(Super) silly example: drug

Existing knowledge

- Population registries
- Observational studies
- Small/pilot RCTs
- Expert opinion

Current data

- Large(r) scale RCT
- Observational study
- Relevant summaries



Observe a study with 150 responders out of 200 patients given the drug

Bayesian modelling

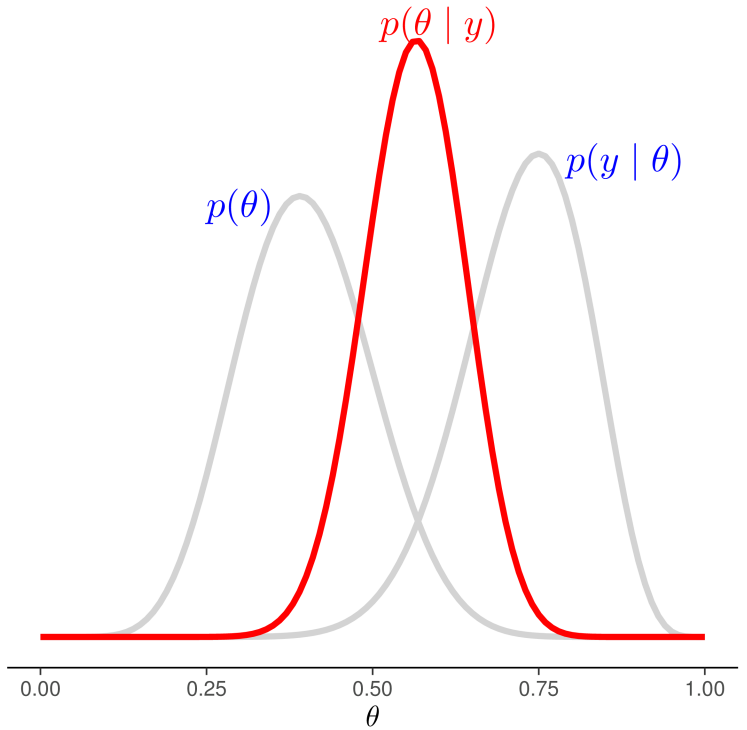
(Super) silly example: drug

Existing knowledge

- Population registries
- Observational studies
- Small/pilot RCTs
- Expert opinion

Current data

- Large(r) scale RCT
- Observational study
- Relevant summaries



- Update knowledge to describe revised "state of science"

*But how can I form a prior? I know **nothing** about this parameter!...*

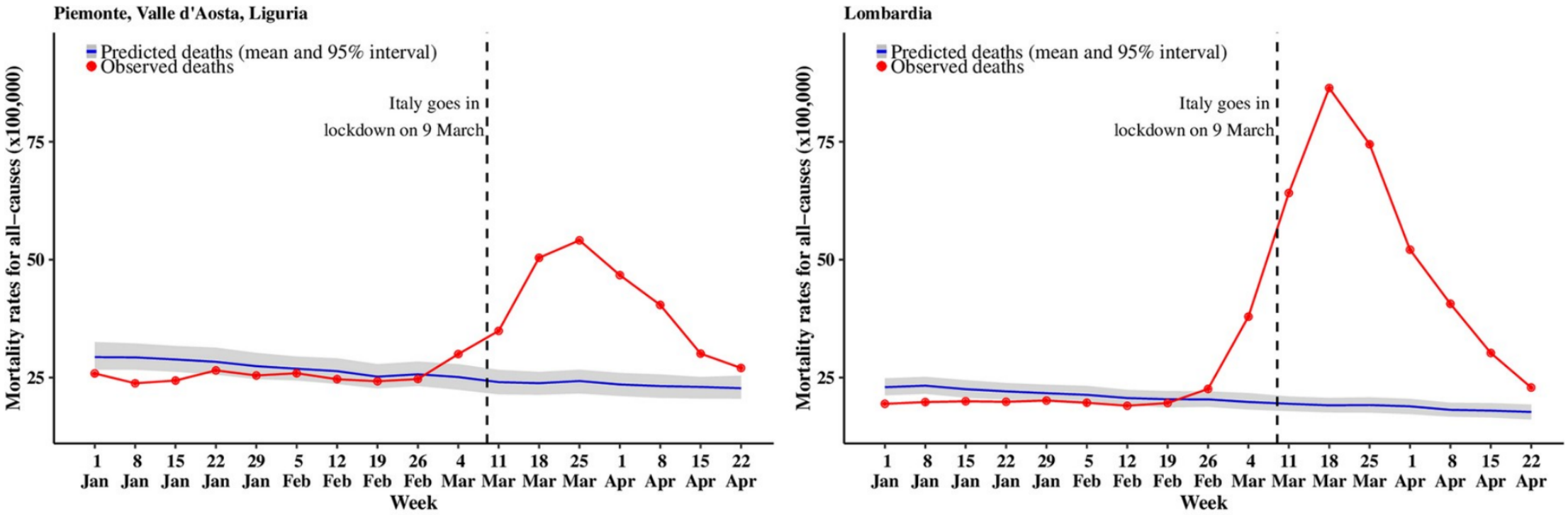
*But how can I form a prior? I know **nothing** about this parameter!...*

*But how can I form a prior? I know **nothing** about this parameter!...*

*But how can I form a prior? I know **nothing** about this parameter!...*

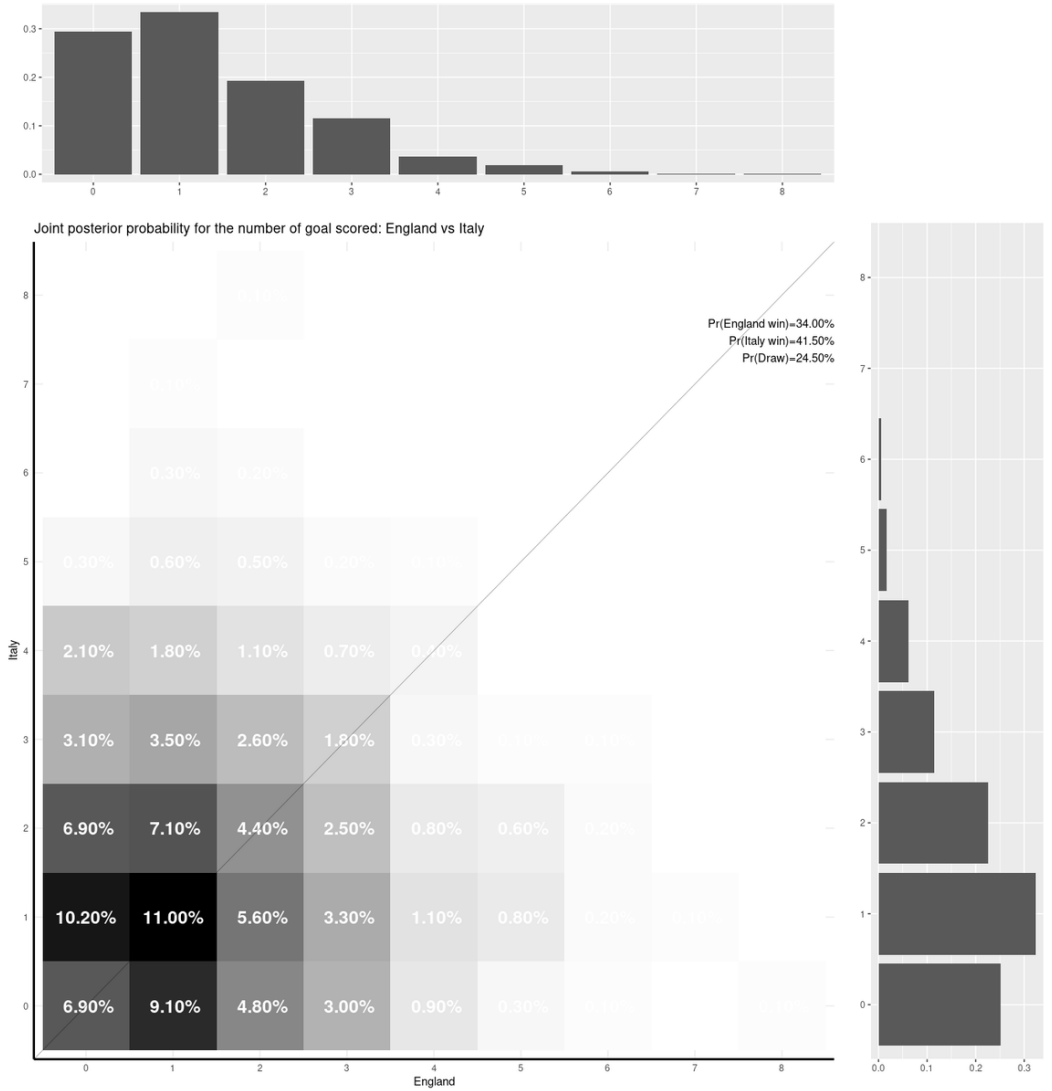
- Predicting the output of the 2017 UK General Election using poll data (see here and subsequent posts)
  – Data: number of people out of the N_i respondents in poll i intending to vote for party p (multinomial counts)
  – **Objective of estimation**: $(\pi_1, \ldots, \pi_P) =$ population vote share for each party
  – Can model $\pi_p = (\phi_p / \sum \phi_p)$ and $\log(\phi_p) = \alpha_p + \beta_p X_p$

Covid-19 excess mortality in Italy



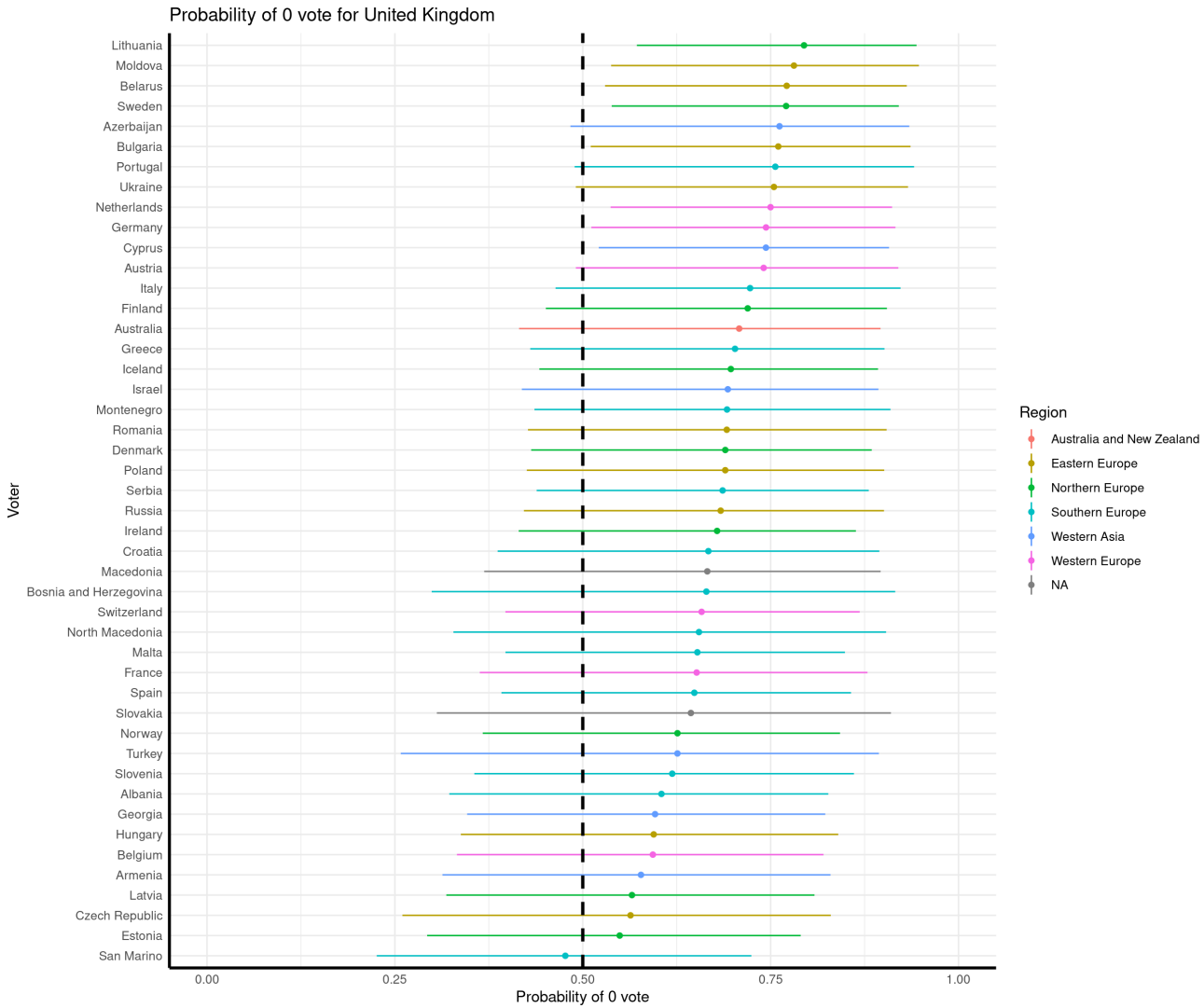
doi Blangiardo et al 2020

Predicting outcome of football games



Blog

Voting bias at the Eurovision




PBS NEWS HOUR
Thank you, all!