Statistical inference

Mabel Carabali

EBOH, McGill University

Updated: 2025-09-02

Objectives

- Review the concept of statistical inference.
- Appreciate the value, limitations & misconceptions of frequentist paradigm.
- Understand the general philosophy, basic mechanism, advantages and limitations of Bayesian inference

References

1. The ASA's Statement on p-Values: Context, Process, and Purpose. T The American Statistician, 2016,70, (2), 129–133
2. Greenland, S et. al.“Statistical Tests, P-values, Confidence Intervals, and Power: A Guide to Misinterpretations.” The American Statistician, Online Supplement 2016.

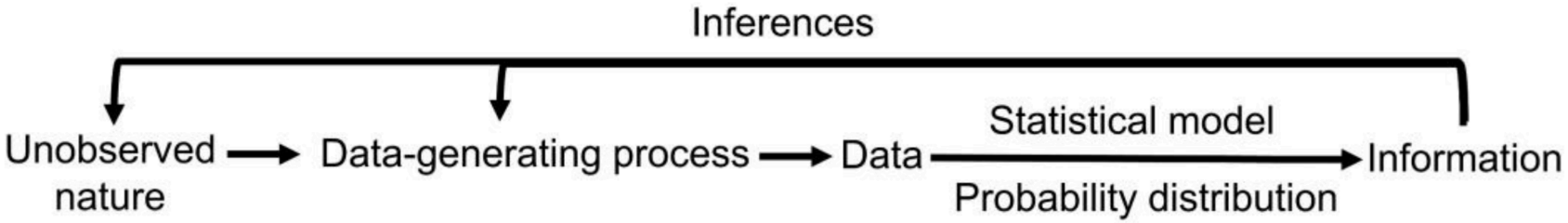
- Some notes from J. Brophy

Expected competencies

- Basic knowledge about probabilities
- Basic knowledge about distributions
- Understand the difference between population and sampling
- Understand the rationale for statistical analysis
- Knows how to correctly interpret confidence interval and p-values

Statistical Inference

Inference



Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med. 1999 Jun 15;130(12):995-1004. doi: 10.7326/0003-4819-130-12-199906150-00008. PMID: 10383371.

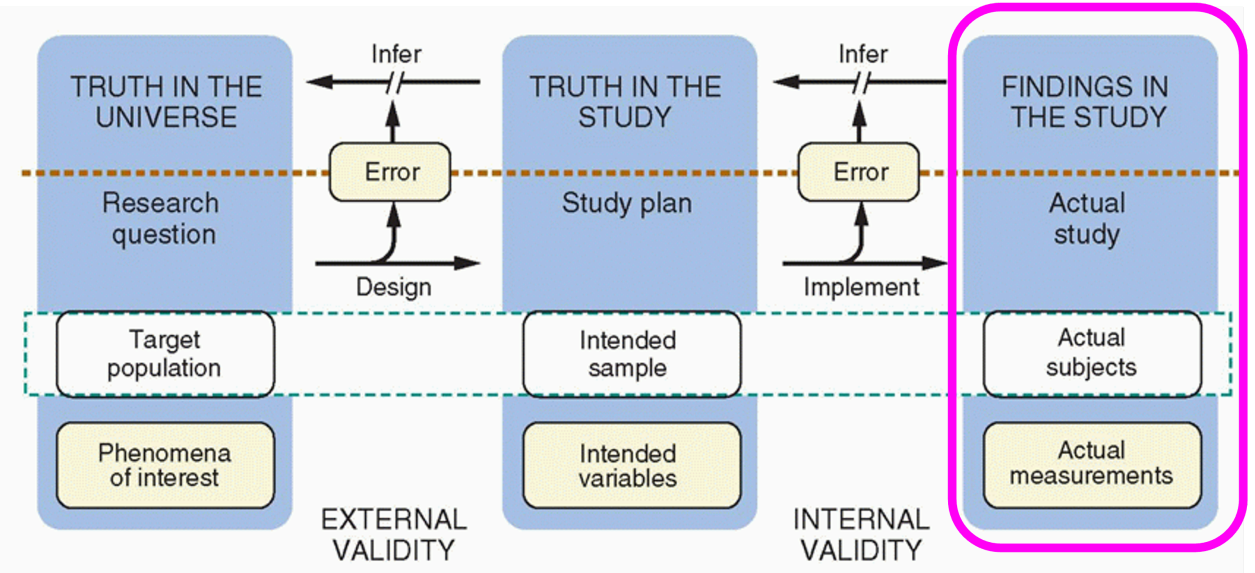
Statistical inference

- Statistical inference is the process of generating associations about a population from a **sample**, without it we're left simply with our data.

- **Probability** models connect noisy sample data and populations and represent the most effective way to obtain inference.
    - Statistical models are insufficient for causality.
    - Paradox: models that are causally incorrect can make better predictions than those that are causally correct.

- Inference is about belief **revision**:
    - Frequentist by *deduction* and Bayesians by *induction*

Sampling



INFERENCE IN THE REAL WORLD... Actual Measurements !!!

Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB.Designing Clinical Research. Lippincott Williams & Wilkins; 2013

Implications of *Sampling* ?

Enough for what? Finding a **null** effect? a **small** effect? a clinically **meaningful** effect?

Sakpal, Tushar Vijay. “Sample size estimation in clinical trial.” Perspectives in clinical research. vol. 1,2 (2010): 67-9.

Probabilities

> - *"A probability describes the possibility of an event to occur given a series of circumstances (or under a series of pre-event factors)."*
> - *"It is a form of **inference**, a way to predict what may happen, based on what happened before under the same (never exactly the same) circumstances."*
> - *"Probability can vary from 0 to 1."*
> - *"Probability could be described by a formula, a graph, in which each event is linked to its probability."*

See: Viti A., et al. A practical overview on probability distributions

Inference depends on the assumed statistical model

- The *probability* of sudden infant death syndrome (SIDS) =

$$\frac{1}{8500}$$

- A UK mother, a lawyer, was on trial for infanticide as she had 2 children die of SIDS

- An expert testified that the probability of 2 deaths in 1 family was

$$(\frac{1}{8500})^2$$

  or 1 in 72 million

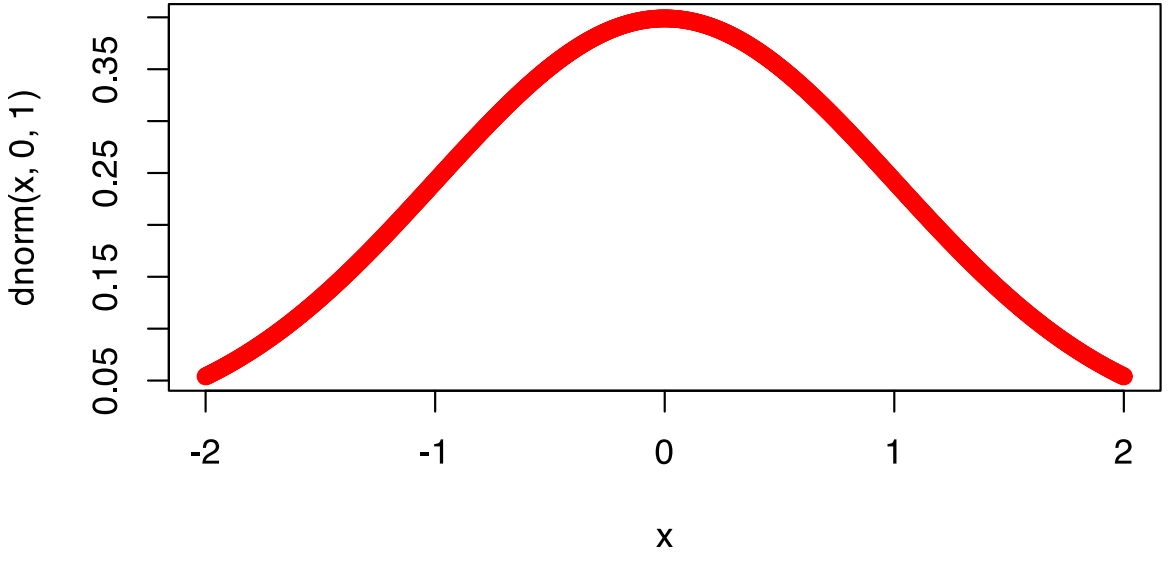
- **The mother was convicted. Do you agree with the conviction?**

Inference depends on the assumed statistical model

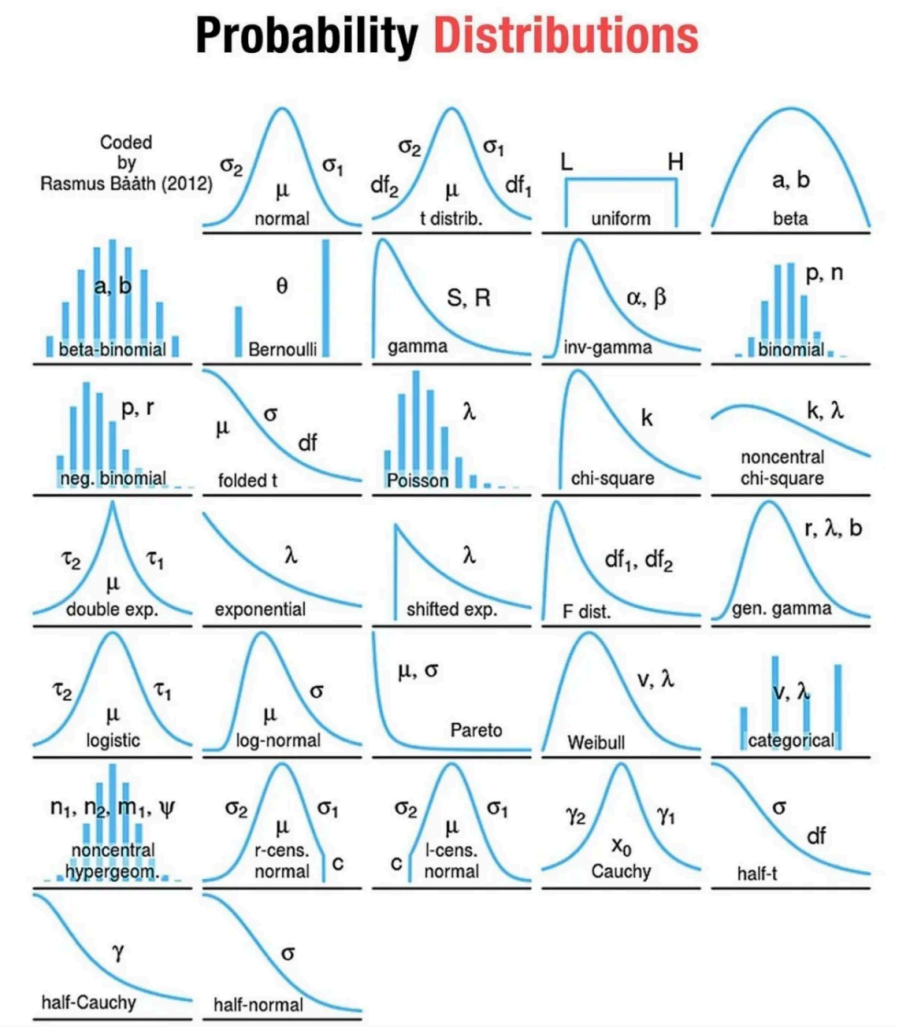
- There are 700,000 annual UK births and therefore about 82 first SIDS deaths
- SIDS deaths are **not** independent as assumed -> strong family occurrence & the risk of a 2nd death is $\neq$ 1 in 8500 but = 1 in 300
- If SIDS families have a 2nd child, E(2nd death) $\approx$ 4 years, >> 1 in 72 million
- Don't know about her guilt but **statistical model and hence inference** was wrong!

Distributions

To make inferences we need to either refer to some common statistical distributions (e.g., normal) or do simulations.



But... normal is boring!?



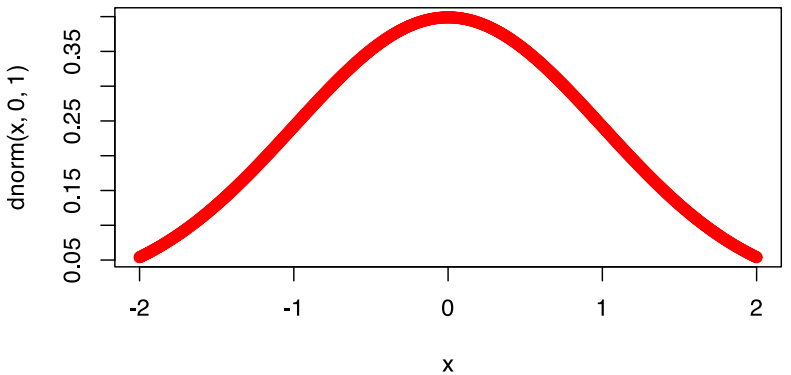
Distributions

- A probability density function (pdf), is a function associated with a continuous random variable
- This leads us to the central dogma of pdfs, namely the areas under the curve corresponds to probabilities for that random variable. To be a valid pdf, a function must:

1. be larger than or equal to zero everywhere
2. the total area under it must be one

Some R code

```
x <- seq(-2, 2, length.out =1000);
plot(x, dnorm(x, 0, 1), col ="red")
```



Is it borning?

Some probability distributions in R:

"A probability distribution displays on a graph the statistical ***likelihood*** of every possible outcome that could occur within a specific time period."

- dnorm: density function of the normal distribution
- pnorm: cumulative density function of the normal distribution
- qnorm: quantile function of the normal distribution
- rnorm: random sampling from the normal distribution

```
dnorm(0); dnorm(2); pnorm(0); qnorm(.975);
```

```
## [1] 0.3989423
```

```
## [1] 0.05399097
```

```
## [1] 0.5
```

```
## [1] 1.959964
```

```
mean(rnorm(10000,0,1))
```

```
## [1] 0.006115893
```

Check this resources

- Statistical Inference for Everyone
- Distribution functions in R
- A Guide to dnorm, pnorm, rnorm, and qnorm in R

Likelihood Principle

The likelihood of a set of data is the probability of obtaining that particular set of data given the chosen probability model.

- The likelihood function is not a probability density function.
- This expression contains the unknown parameters.
- It's constructed by taking the joint probability distribution of your data and treating it as a function of the model's unknown parameters.
- Those values of the parameter that maximize the sample likelihood are known as the maximum likelihood estimates.
- It measures the support provided by the data for each possible value of the parameter.

Likelihood Principle

- The likelihood principle implies that likelihood function can be used to compare the plausibility of various parameter values. For a random sample where $x = (x_1, x_2, \ldots x_n)$ the likelihood function is defined as:

$$L(\theta|x) = \prod_{n=1}^{N} f(x_i; \theta)$$

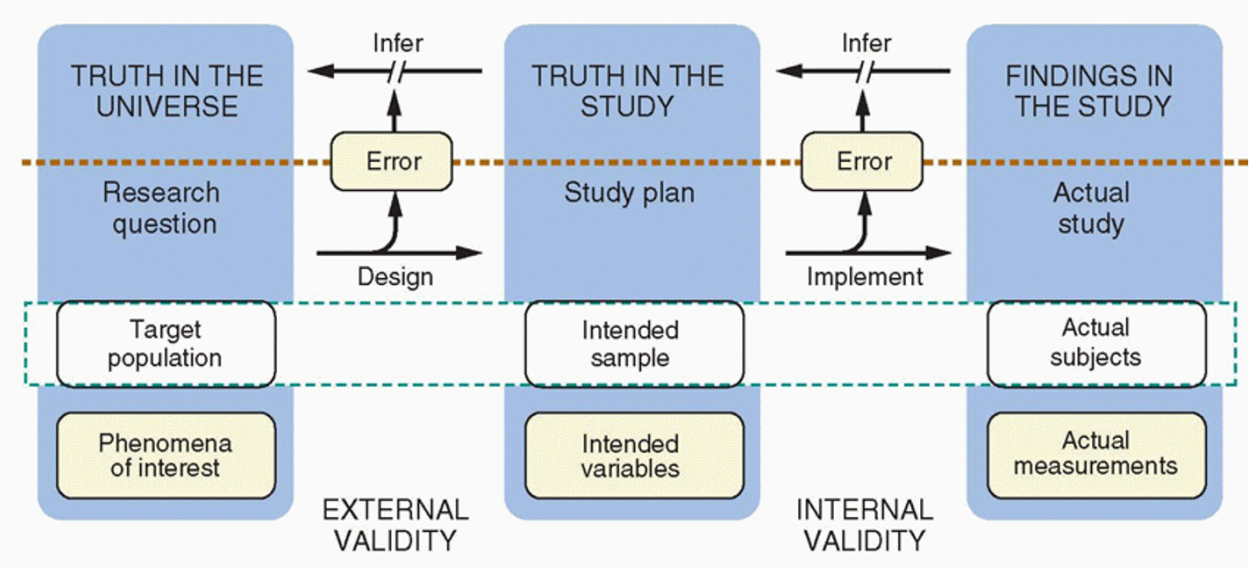
Properties of the Likelihood Function

1. Invariance under Reparameterization:
2. Not a Probability: Likelihoods do not integrate to 1. It's a function of parameters, not a probability distribution.
3. Relative Scale Matters: The absolute value of the likelihood is often less important than relative likelihoods across different parameter values.

Before Inference

Before statistical inference, there **should be** a proper **study design** and **data collection**

- Plenty of places to go wrong before statistical inference



Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB.Designing Clinical Research. Lippincott Williams & Wilkins; 2013

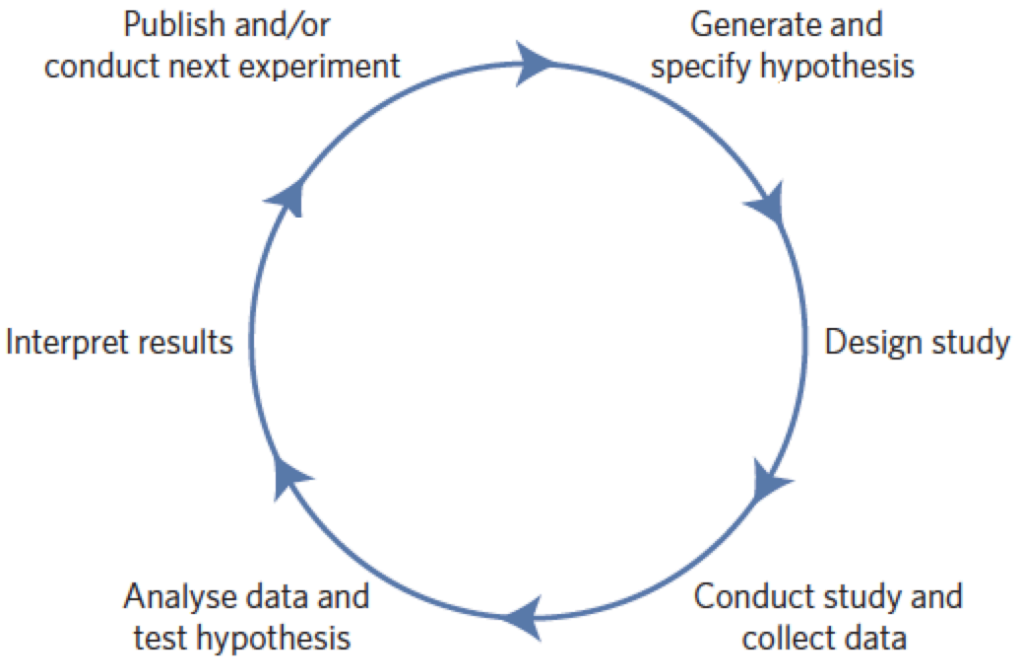
Before statistical inference

Questions to be asked:

- Is the sample representative of the population that we'd like to draw inferences about?
- Are there systematic bias created by selection, misclassification or missing data at the design or during conduct of the study?
- Are there known and observed, known and unobserved or unknown and unobserved variables that contaminate our conclusions?
- What are the criteria for choosing a model (statistical vs causal)?
- What analytical choices are made for the chosen model?

What else can go wrong?

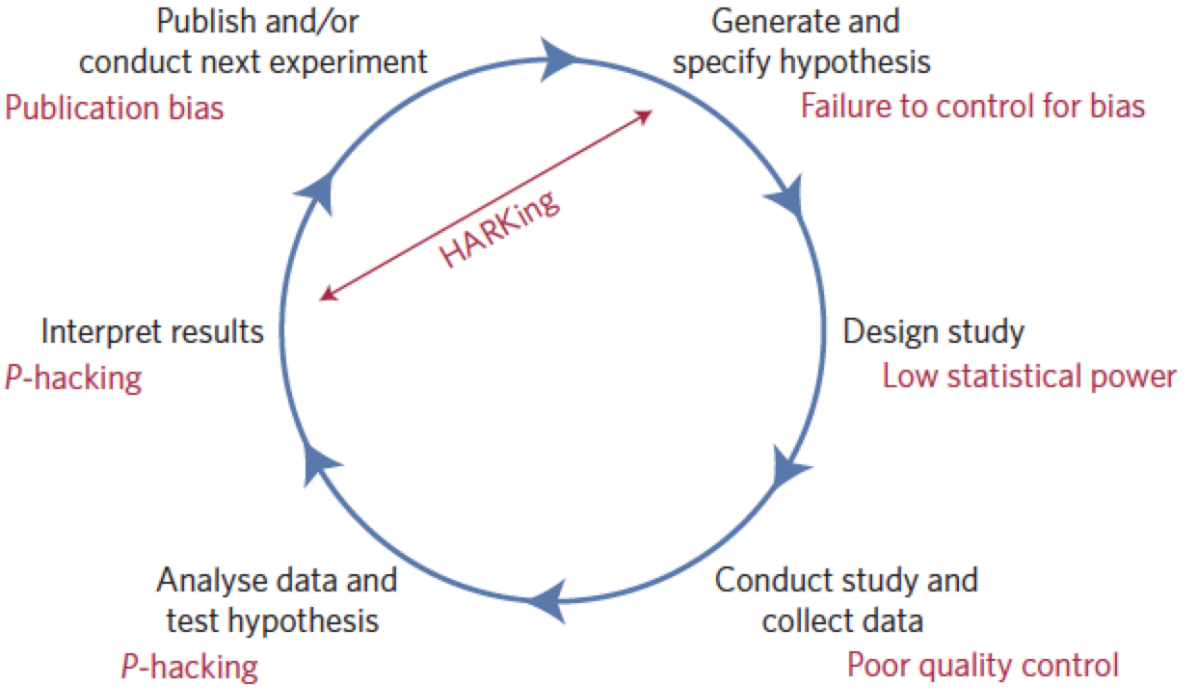
Metascience: The scientific study of science itself: **Hypothetico-deductive model of the scientific method**



Munafò, M., Nosek, B., Bishop, D. et al. A manifesto for reproducible science. Nat Hum Behav 1, 0021 (2017).

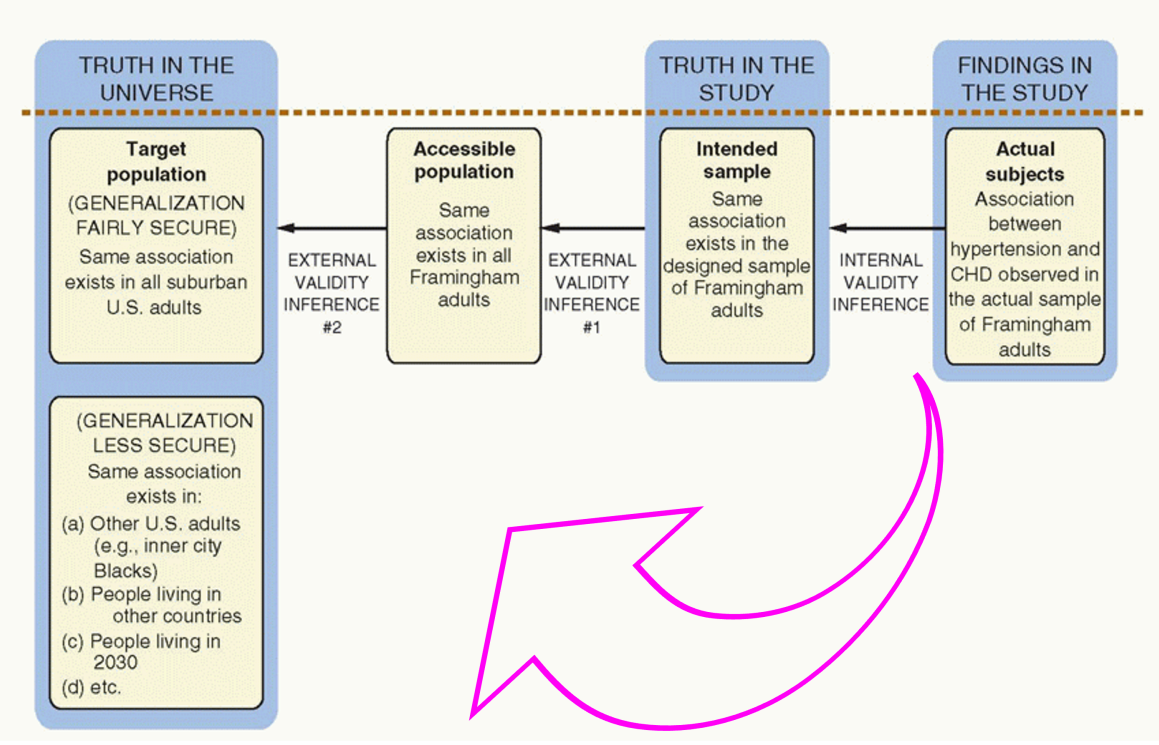
Metascience

Plenty of places to go wrong



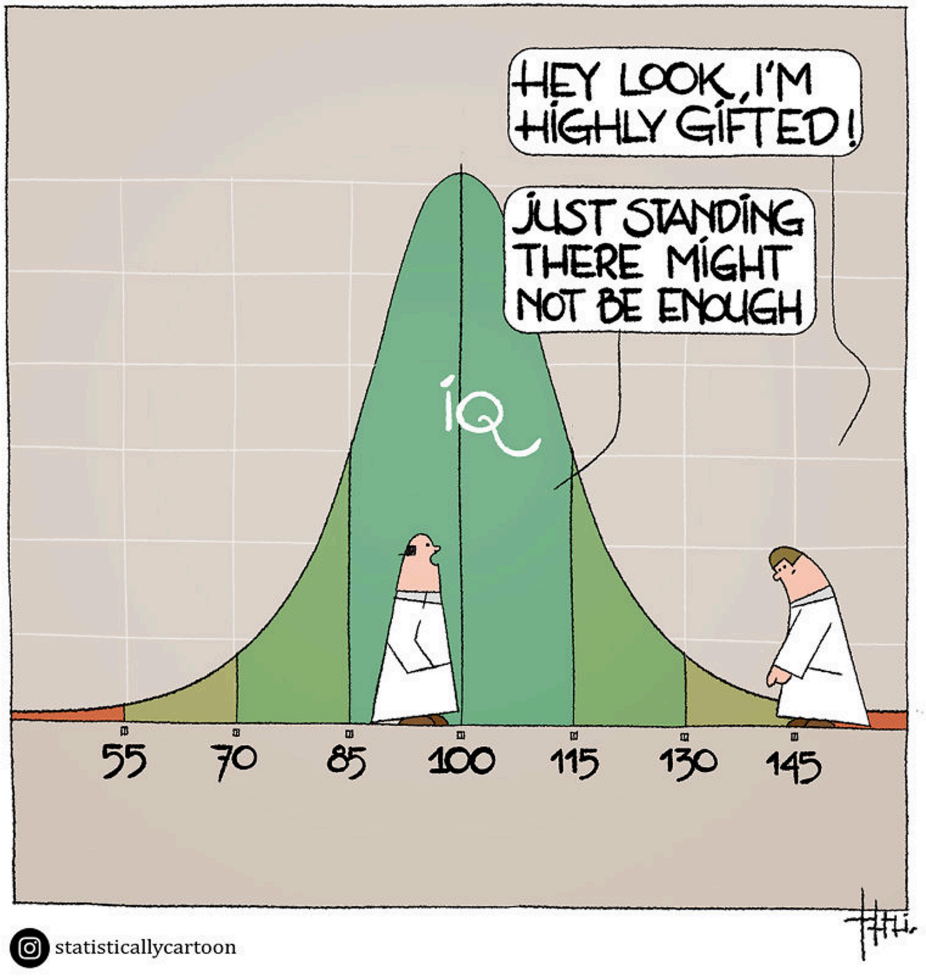
Rubin M. The cost of HARKing and Munafò, M., Nosek, B., Bishop, D. et al. A manifesto for reproducible science. Nat Hum Behav 1, 0021 (2017).

Inference and Validity



Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB.Designing Clinical Research. Lippincott Williams & Wilkins; 2013

Significance?



Significance?

- First, how can I know if the observation (a.k.a) "effect" is "likely" or possible?
- How can I express certainty about an effect or its size?
  - *Using probabilities, which are quantities that are hypothetical frequencies of data patterns under an assumed statistical model.*
  - Ideally, "*the probability of the observed data given the parameter value; it does not refer to a probability of the parameter taking on the given value*"
- How can I measure the distance between the data and the model prediction?
  - **Using a test statistic (e.g., location, t-statistic or a Chi squared statistic)**

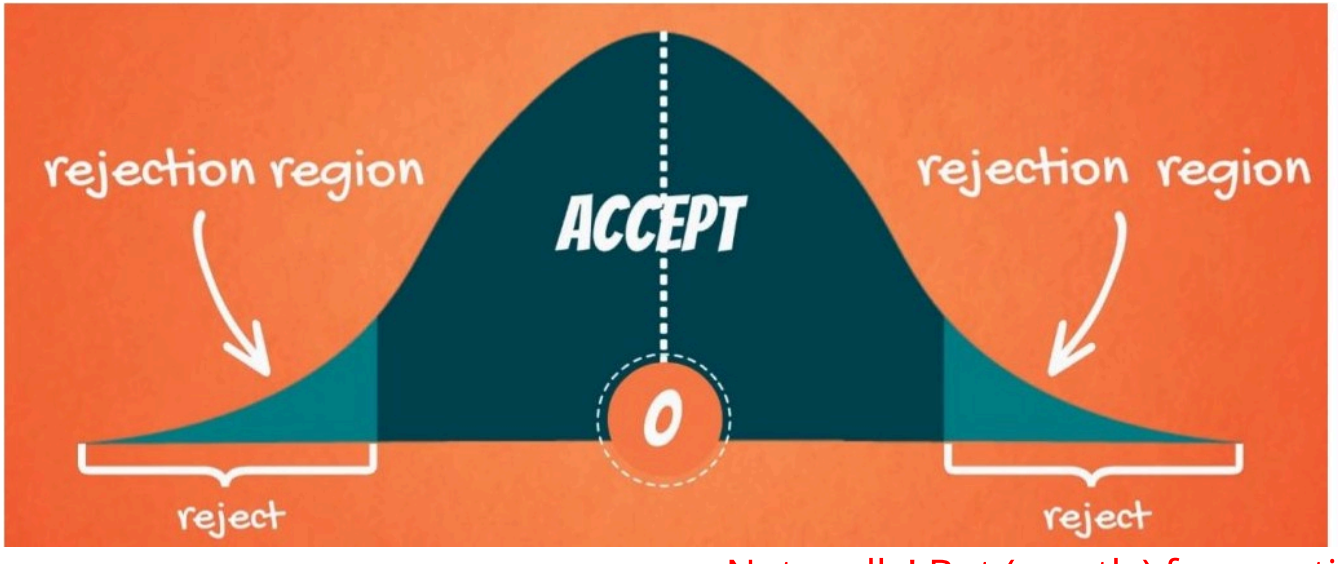
Do you know what are those statistical tests?

Statistical Tests

Distribution	Statistical Test
Standard normal (z distribution)	One-sample location test
Student's *t* distribution	*t* test (all forms); Linear Regression; Pearson correlation
F distribution	ANOVA; Comparison of nested linear models; Equality of two variances
Chi-square	Chi-square goodness of fit test; Chi-square test of independence; McNemar's test; Test of a single variance

For you to review! including critical values and degrees of freedom

... OK, this means that with a statistical test and a probability (or probabilities) I can decide on the **"significance"** of my result?



Source: significant level

Not really! But (mostly) frequentists call it significance level

Level of Significance & *p-values* ...

- Probability of Type I error is called the level of significance and denoted as α (typically 5% or less); but the p-value is not the α error
- Type I error = (Reject H_0 WHEN H_0 is true).
- Type I error opportunity is inversely proportional to **sample size**.

Sample Size and Precision

- Small effect $\rightarrow$ LARGE SAMPLE
- LARGE EFFECT $\rightarrow$ small (decent) sample

The p-value is not the α error

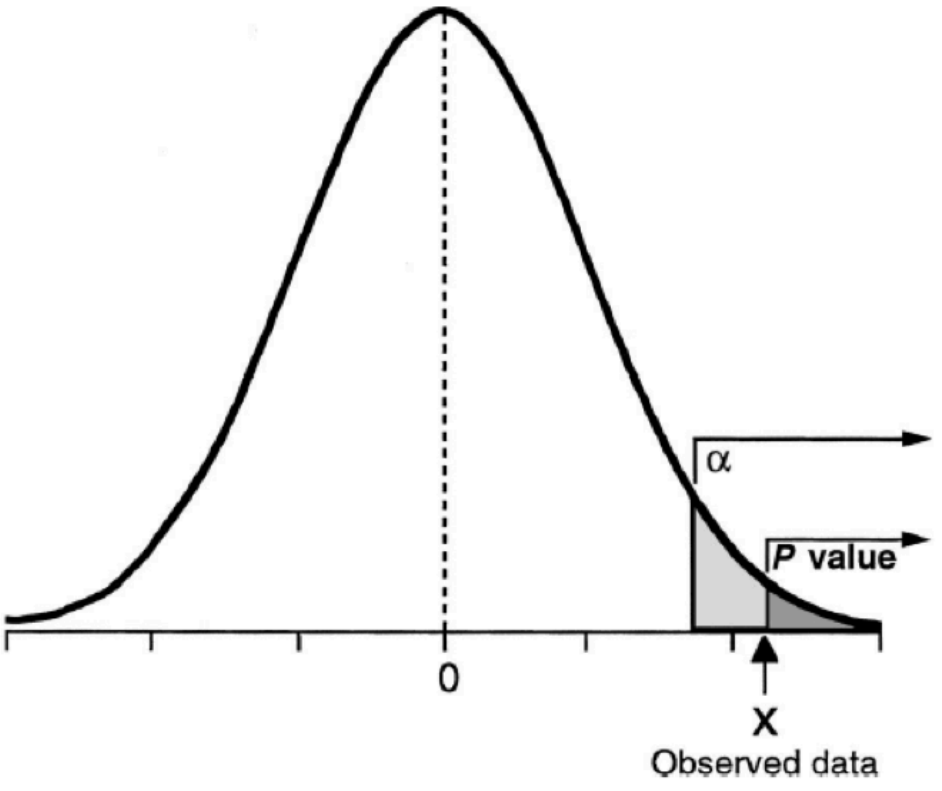


Figure 3. The bell-shaped curve represents the probability of every possible outcome under the null hypothesis. Both α (the type I error rate) and the *P* value are "tail areas" under this curve. The tail area for α is set before the experiment, and a result can fall anywhere within it. The *P* value tail area is known only after a result is observed, and, by definition, the result will always lie on the border of that area.

Level of Significance & *p-values*

So, if *p-value* is the probability under a specified statistical model that a statistical summary of the data would be equal to or more extreme than its observed value... what else can I say about this "metric"?

> - P-values can indicate how incompatible the data are with a specified statistical model.
> - P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
> - Scientific conclusions or policy decisions should not be based only on whether a p-value passes a specific threshold
> - A p-value, or statistical significance, does not measure the size of an effect or the importance of a result
> - Proper inference requires full reporting and transparency

p-values ...

> *"Not only does a P value not tell us whether the hypothesis targeted for testing is true or not; it says nothing specifically related to that hypothesis unless we can be completely assured that every other assumption used for its computation is correct—an assurance that is lacking in far too many studies."*

Greenland S., et al. (2026). Eur J Epidemiol (2016) 31:337–350

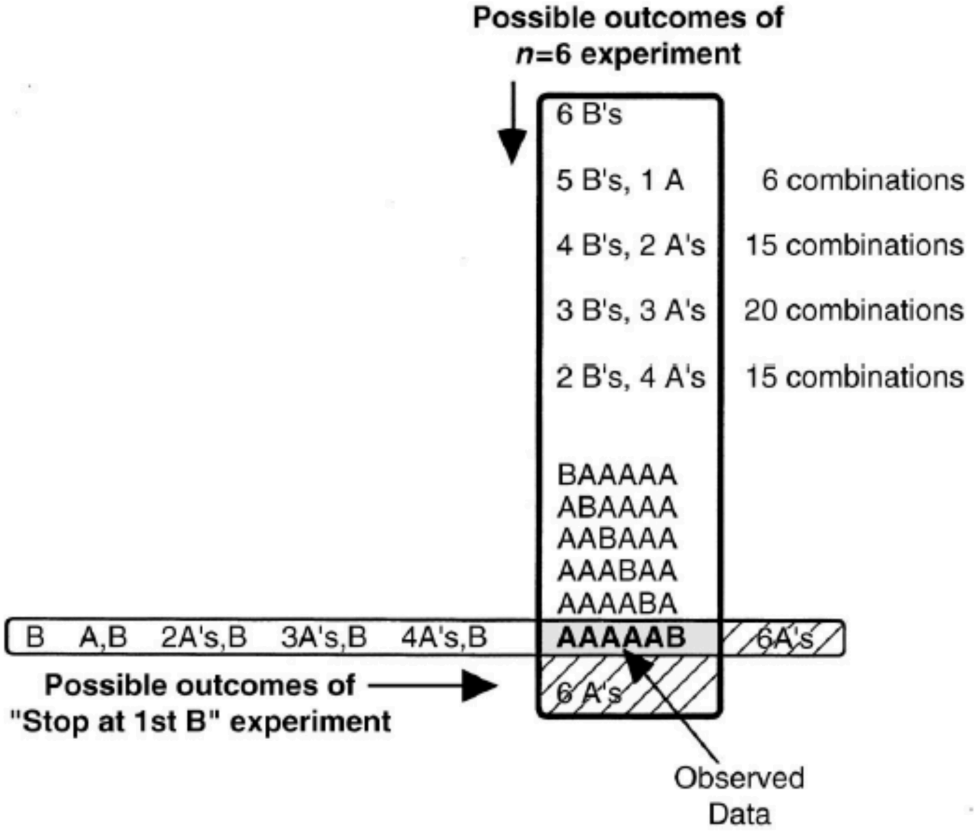
p-values ...

> *"P value is the [conditional] probability that the **chosen test statistic** would have been at least as large as its observed value if **every model assumption were correct**, including the test hypothesis."*

Greenland S., et al. (2026). Eur J Epidemiol (2016) 31:337–350

Recall the Likelihood principle?

Imagine an experiment where you are testing 2 drugs in 6 patients; 5 prefer A and one prefers B. What is the p value?
Well it depends...



Likelihood principle

The n = *6 design:* The probability of the observed result (one treatment B success and five treatment A successes) is $6 \times (1/2) \times (1/2)^5$. The factor "6" appears because the success of treatment B could have occurred in any of the six patients. The more extreme result would be the one in which treatment A was superior in all six patients, with a probability (under the null hypothesis) of $(1/2)^6$. The one-sided *P* value is the sum of those two probabilities:

$$\underbrace{6\,\frac{1^5}{2}\,\frac{1^1}{2}}_{\substack{\text{Probability}\\ \text{of}\\ \text{observed data}}} + \underbrace{\frac{1^6}{2}}_{\substack{\text{Probability of}\\ \text{"more extreme"}\\ \text{data}}} = 0.11$$

"Stop at first treatment B preference" design: The possible results of such an experiment would be either a single instance of preference for treatment B or successively more preferences for treatment A, followed by a case of preference for treatment B, up to a total of six instances. With the same data as before, the probability of the observed result of 5 treatment A preferences − 1 treatment B preference would be $(1/2)^5 \times (1/2)$ (without the factor of "6" because the preference for treatment B must always fall at the end) and the more extreme result would be six preferences for treatment As, as in the other design. The one-sided *P* value is:

$$\underbrace{\frac{1^5}{2}\,\frac{1^1}{2}}_{\substack{\text{Probability}\\ \text{of}\\ \text{observed data}}} + \underbrace{\frac{1^6}{2}}_{\substack{\text{Probability of}\\ \text{"more extreme"}\\ \text{data}}} = 0.03$$

Confidence Intervals

What does it refers to?

> *"refers only to how often 95% confidence intervals computed from very many studies would contain the true size if all the assumptions used to compute the intervals were correct."*

... This speaking Frequentist :)

- What would it be the interpretation of Bayesian "Credible Intrervals"?
- How does precision and sampling relate to the Confidence intervals?

Example

data on years of PhD training, scientific manuscripts, Sex assigned at birth, marital status, whether the individual has children, and a variable about mentoring.

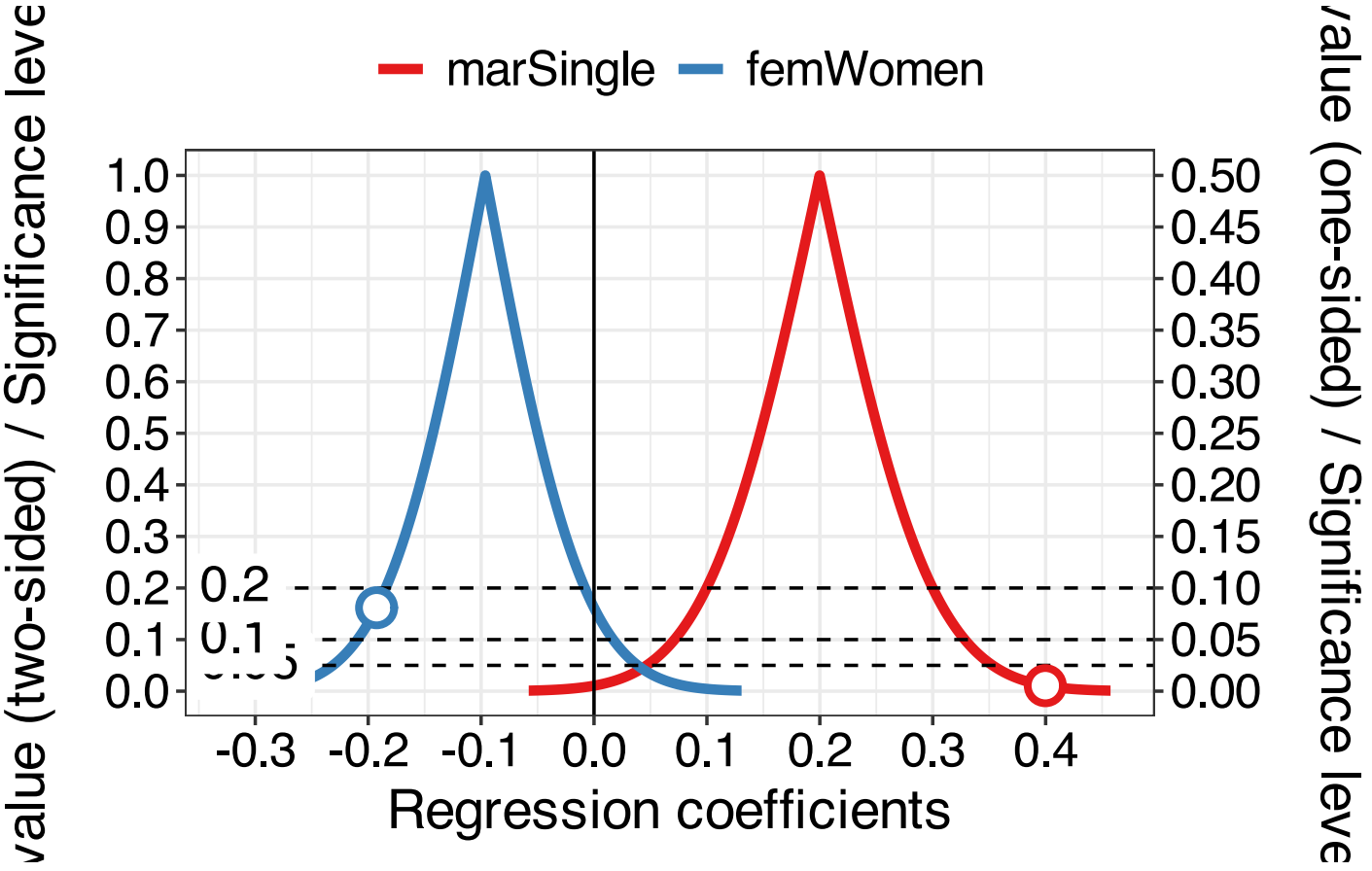
```
##       art             fem                mar                kid5       
##  Min.   : 0.000   Length:915         Length:915         Min.   :0.0000  
##  1st Qu.: 0.000   Class :character   Class :character   1st Qu.:0.0000  
##  Median : 1.000   Mode  :character   Mode  :character   Median :0.0000  
##  Mean   : 1.693                                         Mean   :0.4951  
##  3rd Qu.: 2.000                                         3rd Qu.:1.0000  
##  Max.   :19.000                                         Max.   :3.0000  
##       phd             ment       
##  Min.   :0.755   Min.   : 0.000  
##  1st Qu.:2.260   1st Qu.: 3.000  
##  Median :3.150   Median : 6.000  
##  Mean   :3.103   Mean   : 8.767  
##  3rd Qu.:3.920   3rd Qu.:12.000  
##  Max.   :4.620   Max.   :77.000
```

How should I interpret this?

```
L1mod<- lm(phd ~ mar + fem +kid5, data=articlesdata)
jtools::summ(L1mod, digits=2, confint = TRUE)
```

Observations					915
Dependent variable					phd
Type					OLS linear regression
	F(3,911)	2.78			
	R^2	0.01			
	Adj. R^2	0.01			
	Est.	**2.5%**	**97.5%**	**t val.**	**p**
(Intercept)	3.08	2.95	3.20	48.81	0.00
marSingle	0.20	0.05	0.35	2.56	0.01
femWomen	-0.10	-0.23	0.04	-1.40	0.16
kid5	0.00	-0.09	0.10	0.09	0.93
Standard errors: OLS					

p - value functions (confidence curves)



Mean test; *t* test

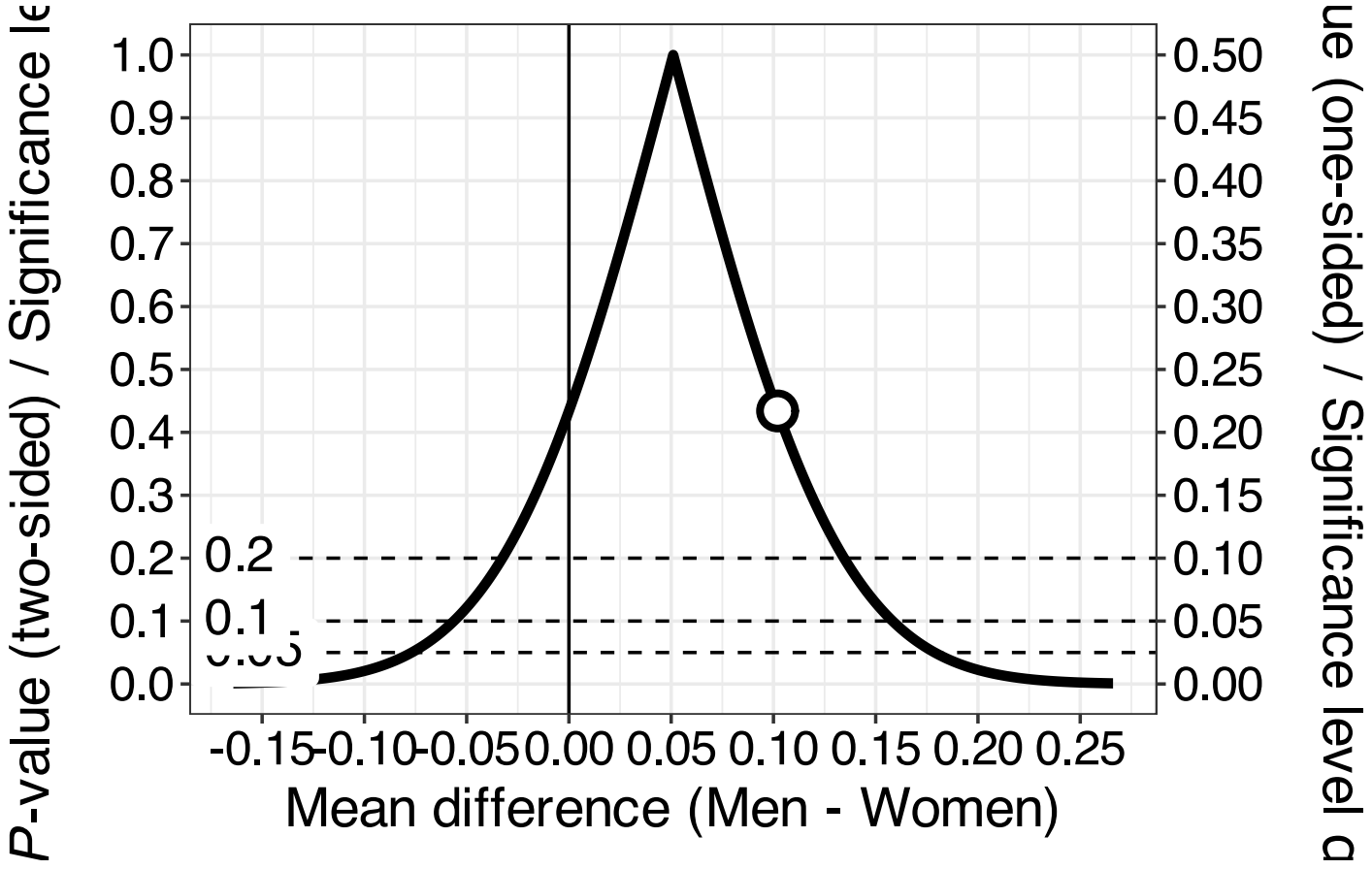
```
with(articlesdata, mean(phd[fem == "Men"])) - with(articlesdata, mean(phd[fem == "Women"]))
```

```
## [1] 0.05098688
```

```
# 0.05098688 #<
t.test(phd ~ fem, data = articlesdata, var.equal = FALSE)
```

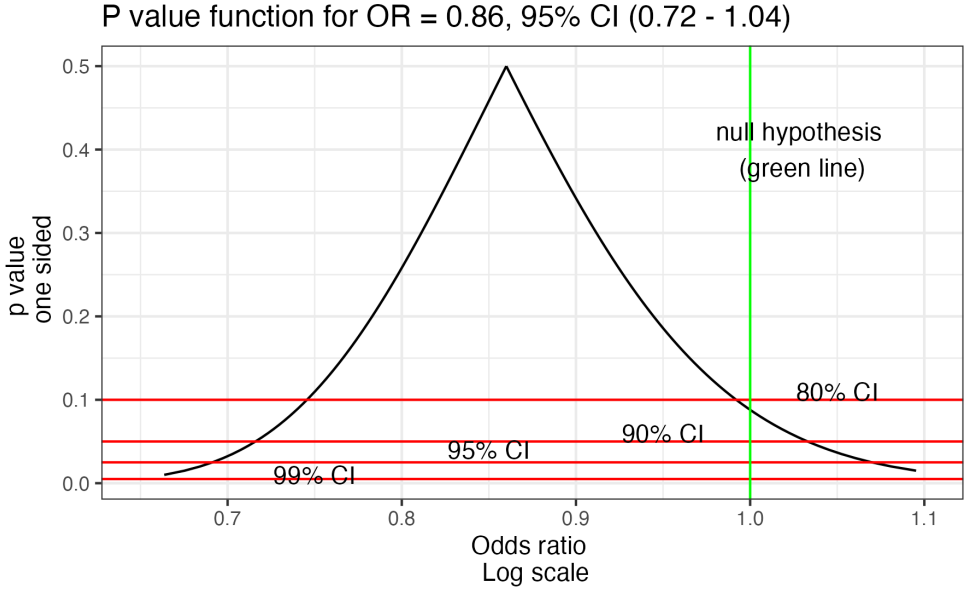
```
##
##     Welch Two Sample t-test
##
## data:  phd by fem
## t = 0.78265, df = 897.58, p-value = 0.434
## alternative hypothesis: true difference in means between group Men and group Women is not equal t
## 95 percent confidence interval:
##  -0.07687023  0.17884399
## sample estimates:
##   mean in group Men mean in group Women
##            3.126569            3.075582
```

Mean test; *t* test; p-value function



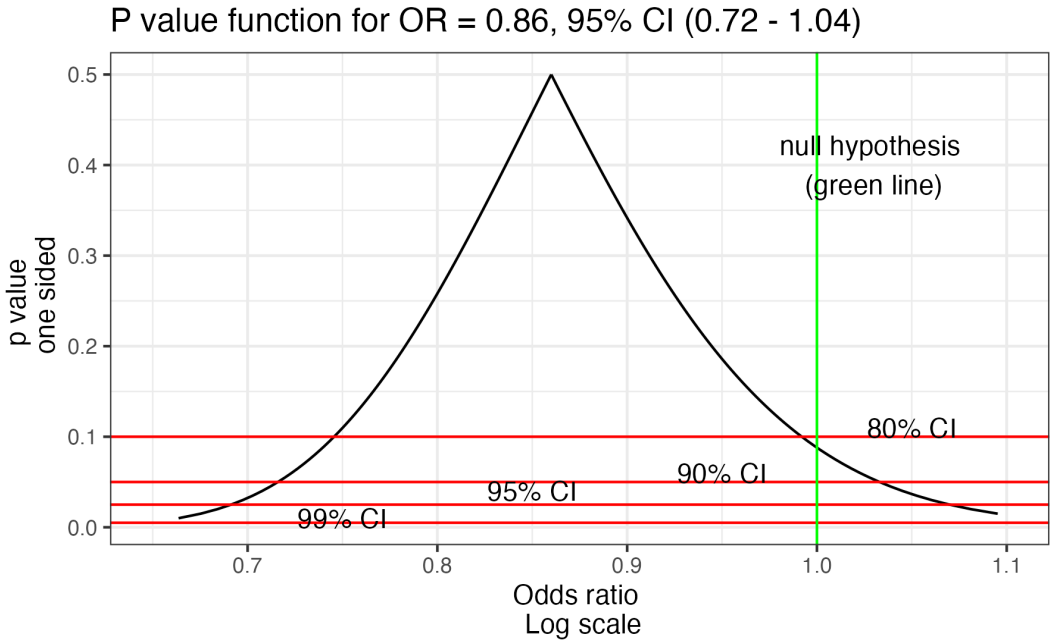
Avoid dichotomania

- Selection of the level of significance or confidence is arbitrary
- Better to interpret the totality of the **p-value function graph**
- NEJM study "Coronary-Artery Bypass Surgery in Patients with Left Ventricular Dysfunction"
  - Reported: HR with CABG, 0.86; 95% CI, 0.72-1.04; P = 0.12) $\rightarrow$ "*no significant difference between treatments*".



Avoid dichotomania

- CIs interpreted dichotomized if HR = 1 $\longrightarrow$ *Not Significant* **BUT** Results support opposite conclusion
- Δ exist between the 2 treatments, and it favors CABG!



P-value function graph (R-code)

```
library(tidyverse)
se <- (log(1.04)-log(0.72))/(2*1.65); x <- seq(0.01, 0.50,by = .005)
p1 <- log(0.86) - (qnorm(x) * se); p2 <- log(0.86) + (qnorm(x) * se)
p1 <- exp(p1); p2 <- exp(p2); p <- data.frame(x, p2, p1)
gg <- ggplot(p, aes( p2, x)) +
  geom_line() +
  geom_line(aes(p1, x)) +
  xlim(0.65,1.1) +
  ylab("p value \n one sided") +
  xlab("Odds ratio \n Log scale") +
  ggtitle("P value function for OR = 0.86, 95% CI (0.72 - 1.04)" ) +
  geom_hline(yintercept=c(.005,.025,0.05,0.10), color = "red") +
  annotate("text", x=0.75,y=.01, label="99% CI") +
  annotate("text", x=0.85,y=.04, label="95% CI") +
  annotate("text", x=0.95,y=.06, label="90% CI") +
  annotate("text", x=1.05,y=.11, label="80% CI") +
  geom_vline(xintercept=1.0, color = "green") +
  annotate("text", x=1.03,y=.4, label="null hypothesis \n(green line)") + theme_bw()
gg <- ggsave("images/01_gg2.png") #To save the figute
```

Reference: Infanger D, Schmidt-Trucksäss A. P value functions: An underused method to present research results and to promote quantitative reasoning. Statistics in Medicine. 2019;38:4189–4197.Original paper here and Tutorial here

Avoiding nullism

Evidence **against** not only H_o but against any specific H_a better appreciated by considering the binary Shannon information, surprisal or **S value**.

- $s = log_2(\frac{1}{P})$ or $P = (1/2)^s$, i.e = P(successive tosses of an unbiased coin showing only heads)

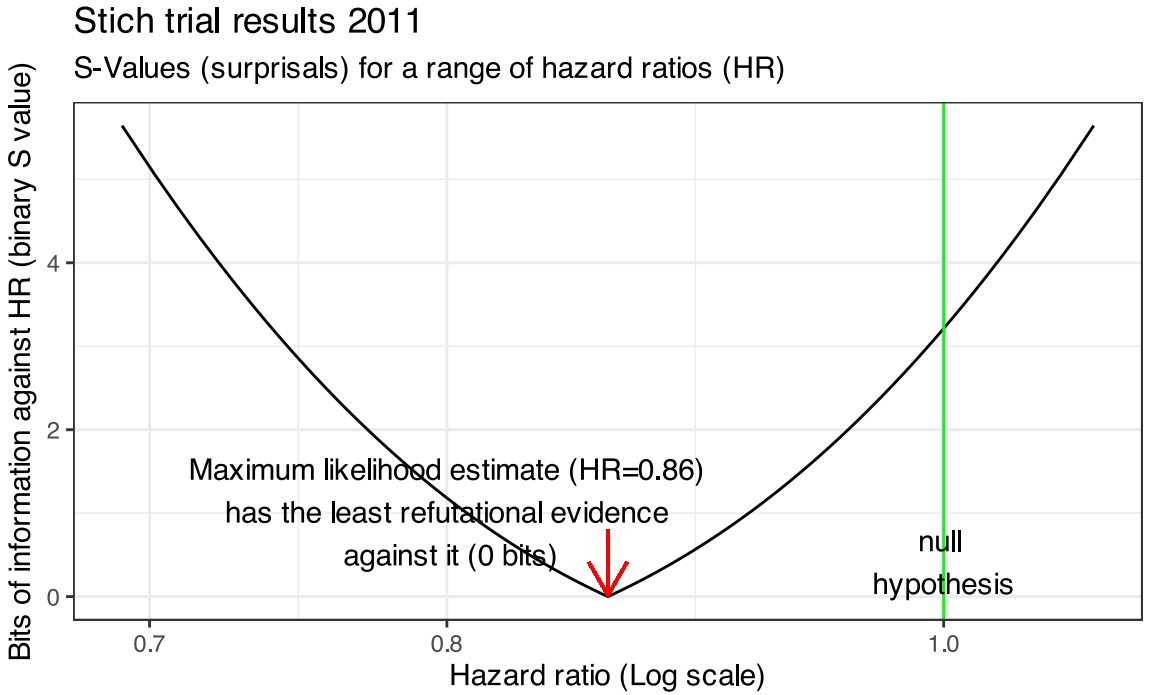
- S *"as measuring our evidence against acceptability"*

- *"The S-value is designed to reduce incorrect probabilistic interpretations of statistics by providing a nonprobability measure of information supplied by the test statistic against the test hypothesis H"*

Rafi, Z., Greenland, S. Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise. BMC Med Res Methodol 20, 244 (2020).

Avoiding nullism

- Evidence against it's minimized at point estimate
- $\downarrow$ evidence against H_a of a 25% $\downarrow$, decrease with CABG than there is against H_o, which we have been told to accept!



Simple R function

```
inter_test <- function(rr1, rr1LL, rr1UL, rr2, rr2LL, rr2UL, sig=0.975) {
  #se of log(rr1), default 95%CI, sig = 1 sided value
  logSE1 <- abs(log(rr1UL) - log(rr1LL))/(2 * qnorm(sig))
  logSE2 <- abs(log(rr2UL) - log(rr2LL))/(2 * qnorm(sig)) #se of log(rr1)
  diffLogRR <- log(rr1) - log(rr2) #diff of log rr
  logRR_SE <- sqrt(logSE1^2 + logSE2^2) #log (se) of differences
  logRR_UCI <- diffLogRR + qnorm(sig) * logRR_SE
  logRR_LCI <- diffLogRR - qnorm(sig) * logRR_SE
  RR <- exp(diffLogRR) # RR point estimate
  RR_UCI <- exp(logRR_UCI) # RR upper CI
  RR_LCI <- exp(logRR_LCI) # RR lower CI
  RR_SE <- (RR_UCI - RR_LCI) / (2*1.96)
  pvalue <- round(2*(1 - pnorm(sig,RR,RR_SE)),2) #p value for the interaction term
  state1 <- cat("The relative risk for the interaction is ",
                round(RR, 2),", 95% CI ", round(RR_LCI, 2), "-",
                round(RR_UCI,2), " and p value =" , round(pvalue, 3))
}

inter_test(0.65,0.46,0.92,0.91,0.74,1.11)
```

```
## The relative risk for the interaction is  0.71 , 95% CI  0.48 - 1.07  and p value = 0.08
```

How different are these two results?

```
inter_test(0.65,0.46,0.92,0.91,0.74,1.11)
```

```
## The relative risk for the interaction is  0.71 , 95% CI  0.48 - 1.07  and p value = 0.08
```

Author's Conclusion:

Annual screening was associated with reduced risk of PCSM among Black men but not among White men, suggesting that annual screening may be particularly important for Black men.

More than 20 years on, and still making the same errors and drawing incorrect conclusions!

If inference is about belief revision... How do I *Revise* my beliefs?

- Frequentist Paradigm (deductive)
- Bayesian Paradigm by (inductive)

Consider two claims

1. John claims that they can predict dice rolls/throws. To test John's claim, you roll a fair dice 10 times and John correctly predicts all 10.

2. Jane claims that they can distinguish between natural and artificial sweeteners. To test Jane's claim, you give her 10 sweetener samples and Jane correctly identifies all 10
   Given this evidence, which of the 2 statements below do you most agree with?

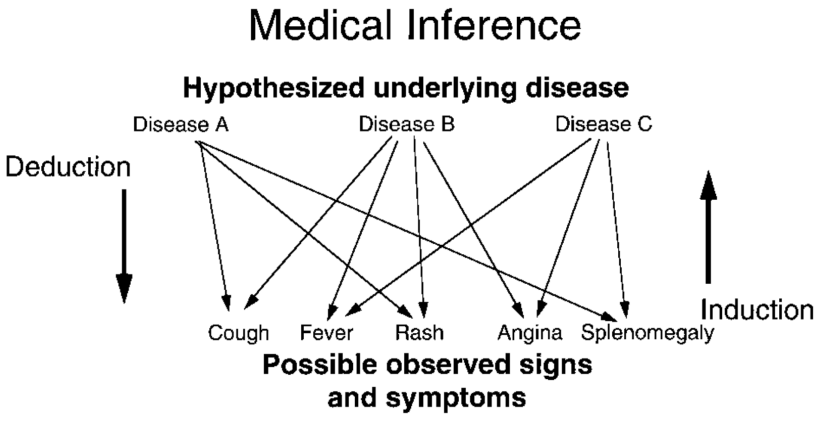
- **A.** John’s claim is just as strong as Jane’s claim

- **B.** Jane’s claim is stronger than John’s claim

Choose A: - Hardcore frequentist

Choose B: - Latent Bayesian

Deductive vs Inductive Inference (I)



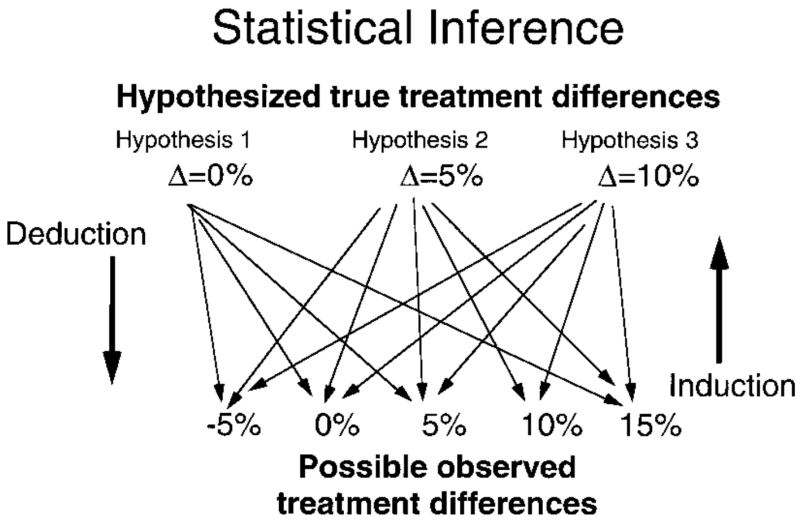
Deduction appears objective; predictions true **only if** H are true

- Can't expand knowledge beyond H
- Analogous to "frequentist" with Fisherian p values, & Neyman-Pearson hypothesis testing, long term errors rates
- 2 schools presented as unified theory, but actually separate (?irreconcilable)
- Pr(Observed data | Hypothesis) (p value definition)

Deductive vs Inductive Inference (II)

Induction is harder but provides a broader, more useful, view of nature

- Drawback can't be sure that what we conclude about nature is actually true - problem of induction
- Analogous to "Bayesian" approach to statistical inference
- Pr(Hypothesis | Observed data)



Frequentist and Bayesian views of probability

Frequency viewpoint: probability parameters considered as **fixed** but unknown quantities, can't make probability statements about them. Probability limited to sampling variability, i..e. in the long run proportion of times an event occurs in independent, identically distributed (iid) repetitions.

Frequency style inference: uses frequency interpretations of probabilities to control error rates. Answers questions like *"What should I decide given my data controlling the long run proportion of mistakes I make at a tolerable level."*

Bayesian viewpoint: probability is the calculus of beliefs, with parameters that are considered **random** variables with probability distributions that follow the rules of probability

Bayesian style inference: uses of probability representation of beliefs to perform inference. Answers questions like *"Given my subjective beliefs and the objective information from the data, what should I believe now?"*

Frequentist vs Bayesian (summary)

Frequentist	Bayesian
Probability is "long-run frequency"	Probability is "degree of certainty"
$Pr(X \mid \theta)$ is a sampling distribution (function of X with θ fixed)	$Pr(X \mid \theta)$ is a likelihood (function of θ with X fixed)
No prior	Prior
P-values (NHST)	Full probability model available for summary/decisions
Confidence intervals	Credible intervals
Violates the "likelihood principle": Sampling intention matters Corrections for multiple testing Adjustment for planned/post hoc testing	Respects the "likelihood principle": Sampling intention is irrelevant No corrections for multiple testing No adjustment for planned/post hoc testing
Objective?	Subjective?

Frequentists Inference

Frequentist statistical inference (known falsehoods)

- Statistical methods alone can provide a number that by itself reflects a probability of reaching true / erroneous conclusions
- Biological understanding and previous research have little formal role in the interpretation of quantitative results
- Standard statistical approach implies that conclusions can be produced with certain “random error rates,” without consideration of internal biases and external information
- p values and hypothesis tests, are a mathematically coherent approach to inference

Bayesian Inference

Bayesian Inference - What is it?

- "Bayesian inference is **reallocation** of **credibility** across **possibilities**." (Kruschke, p. 15)
- "Bayesian data analysis takes a **question** in the form of a **model** and uses **logic** to produce an **answer** in the form of **probability distributions**." (McElreath, p. 10)
- "Bayesian inference is the **process** of **fitting** a **probability model** to a set of **data** and summarizing the result by a **probability distribution on the parameters** of the model and on **unobserved quantities** such as predictions for new observations." (Gelman, p. 1)

References

- Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2013. Bayesian Data Analysis, Third Edition. Boca Raton: Chapman; Hall/CRC.
- Kruschke, John K. 2014. Doing Bayesian Data Analysis: A Tutorial Introduction with R. 2nd Edition. Burlington, MA: Academic Press.
- McElreath, Richard. 2020. Statistical Rethinking: A Bayesian Course with Examples in R and Stan. 2nd ed. CRC Texts in Statistical Science. Boca Raton: Taylor; Francis, CRC Press.

Bayesian Inference

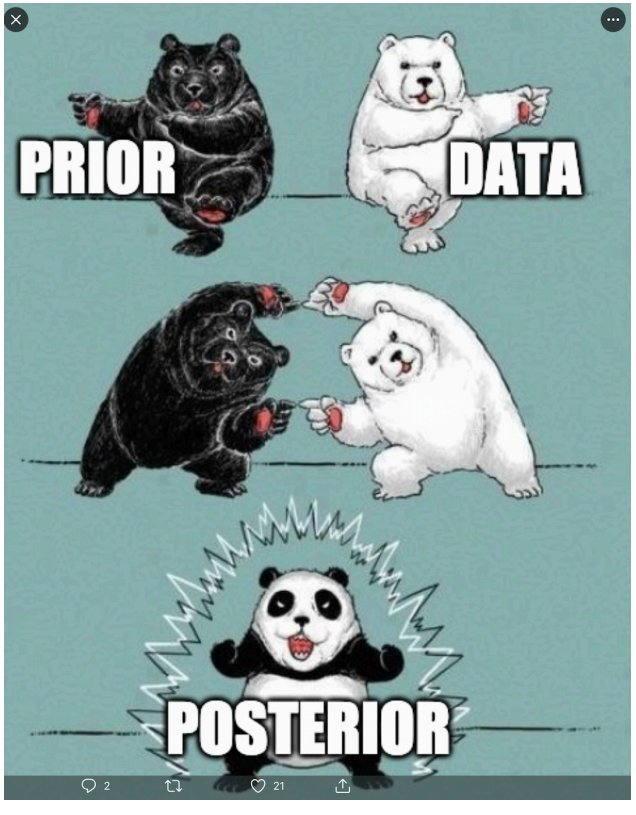
Bayes' Theorem $\rightarrow$ probability statements about hypotheses, model parameters or anything else that has associated uncertainty

Advantages
Treats unknown parameters as random variables -> direct and meaningful answers (estimates)

- Allows integration of all available information -> mirrors sequential human learning with constant updating
- Allows consideration of complex questions / models where all sources of uncertainty can be simultaneously and coherently considered

Disadvantages
Subjectivity (?) Problem of induction (Hume / Popper - difficulty generalizing about future)

Bayes rule (conceptual)



$$posterior = \frac{likelihood * prior}{normalizing\,constant}$$

Bayes rule

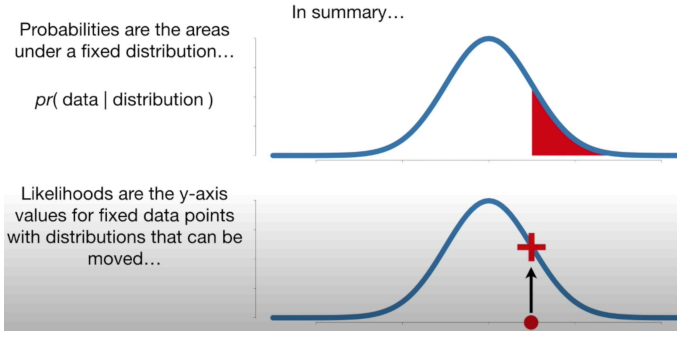
Likelihood - propensity for observing the data given a certain value of θ

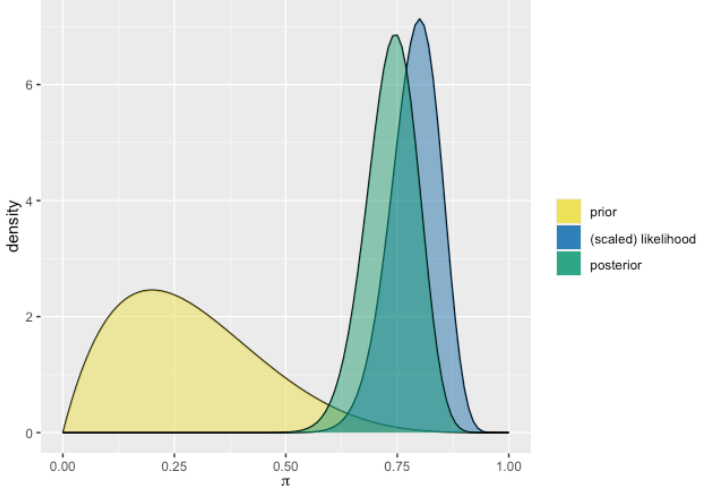
Prior - what we know of θ **before** seeing the data

$$\Pr(\theta \mid \text{data}) = \frac{\Pr(\text{data} \mid \theta) \times \Pr(\theta)}{\Pr(\text{data})}$$

Posterior - what we know of θ **after** seeing the data

Pr(data) - called the average likelihood because it is obtained by integrating the likelihood WRT the prior





Calculations

$$p(\theta|Y) \propto p(Y|\theta)p(\theta)$$

How the likelihood of each data point contributes

$$p(\theta|Y) \propto p(\theta) \prod_{n=1}^{N} p(y_n|\theta)$$

For programming, add individual log probabilities

$$\log p(\theta|Y) \propto \log p(\theta) + \sum_{n=1}^{N} \log p(y_n|\theta)$$

Calculations

- Stan and other Markov Chain Monte Carlo (MCMC) techniques approximate high dimensional probability distributions
- Stan uses Hamiltonian MCMC to approximate $p(\theta|Y)$
- We can write out (almost) any probabilistic model and get full probability distributions to express our uncertainty about model parameters
- Higher-level interfaces allow us to avoid writing raw Stan code

```
library(rstan)
library(brms)
library(rstanarm)
```

- Converts R modelling syntax to Stan language *and extends it in interesting ways*

Bayesian workflow

To get started with Bayesian data analysis (BDA), it is useful to first informally define what a "Bayesian workflow" might look like.

Five key data analysis steps follow;

1. Identify data relevant to the research question
2. Define a descriptive model, whose parameters capture the research question
3. Specify prior probability distributions on parameters in the model
4. Update the prior to a posterior distribution using Bayesian inference
5. Check your model against data, and identify possible problems

Defining the model

Usually model written as

$$y_n = \mu + \epsilon_n$$

where

$$\epsilon_n \sim N(0, \sigma^2)$$

Bayesian usually prefer the following equivalent form

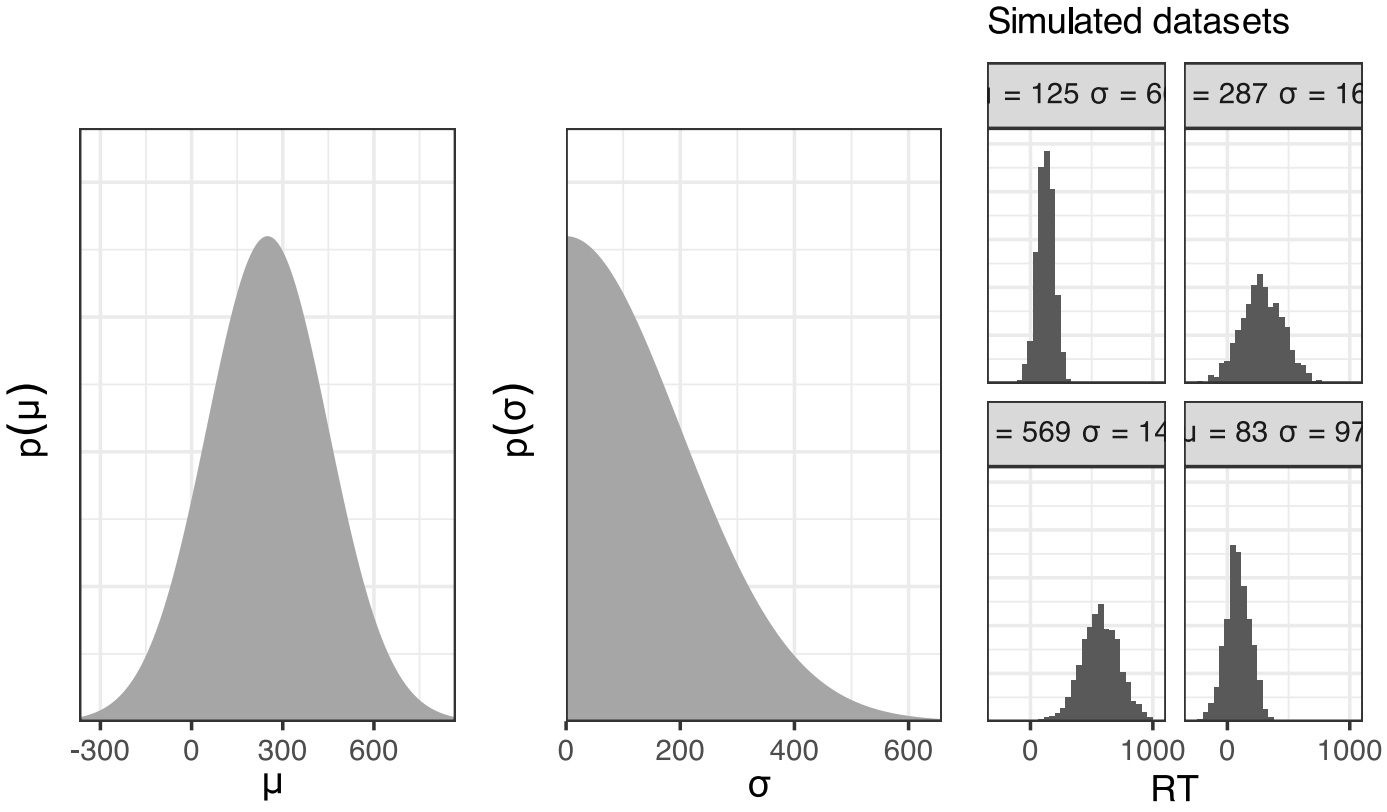
$$y_n \sim N(\mu, \sigma^2)$$

Need to define prior beliefs, before the data are observed. Requires care, and often a vague or non-informative priors are useful starting points.

$$\mu \sim N(250, 200)$$
$$\sigma \sim N^+(0, 200)$$

Defining the priors

$$\mu \sim N(250, 200)$$
$$\sigma \sim N^{+}(0, 200)$$



Bayesian example - non-informative prior

The NEJM 2011 Coronary-Artery Bypass Surgery in Patients with Left Ventricular Dysfunction study, cited > 1200 times, concluded no significant difference between medical therapy alone and medical therapy plus CABG.

Table 2. Study Outcomes.*

Outcome	Medical Therapy (N=602)	CABG (N=610)	Hazard Ratio with CABG (95% CI)	P Value†
	no. (%)			
Primary outcome: rate of death from any cause	244 (41)	218 (36)	0.86 (0.72–1.04)	0.12

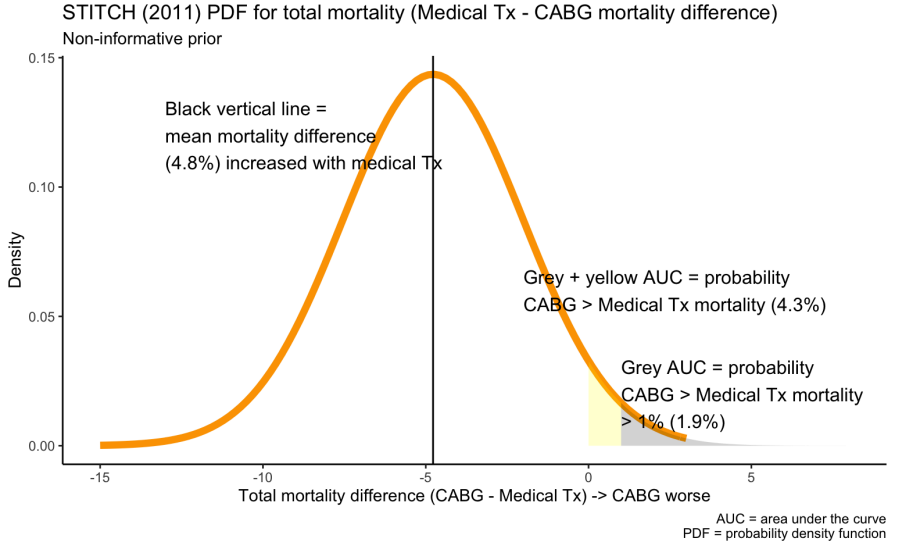
Likelihood - propensity for observing the data given a certain value of θ

Prior - what we know of θ **before** seeing the data

$$\Pr(\theta \mid \text{data}) = \frac{\Pr(\text{data} \mid \theta) \times \Pr(\theta)}{\Pr(\text{data})}$$

Posterior - what we know of θ **after** seeing the data

Pr(data) - called the average likelihood because it is obtained by integrating the likelihood WRT the prior



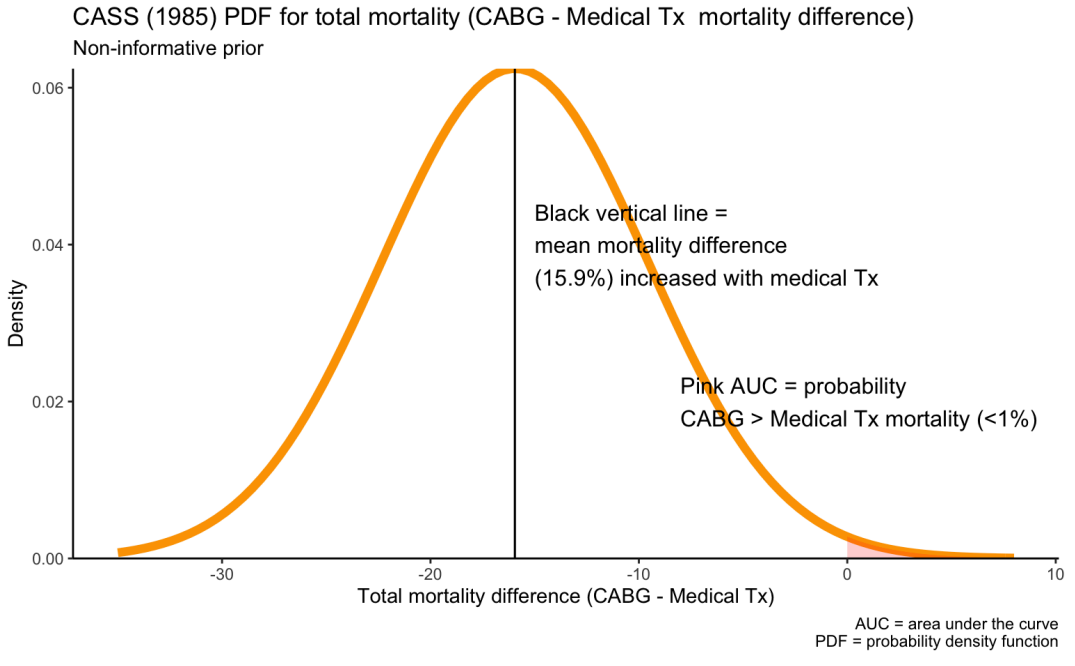
Bayesian example - informative prior

THE NEW ENGLAND JOURNAL OF MEDICINE June 27, 1985

A RANDOMIZED TRIAL OF CORONARY ARTERY BYPASS SURGERY

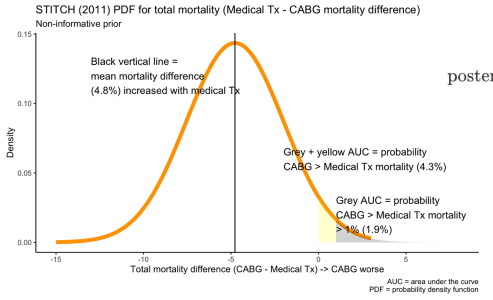
Survival of Patients with a Low Ejection Fraction

7 year mortality - 25 / 82 (medical 30%) versus 11 / 78 (CABG 14%)



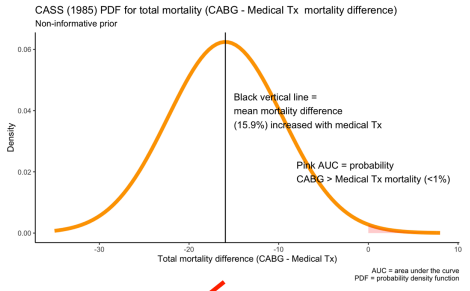
Bayesian example - updated

STICH data

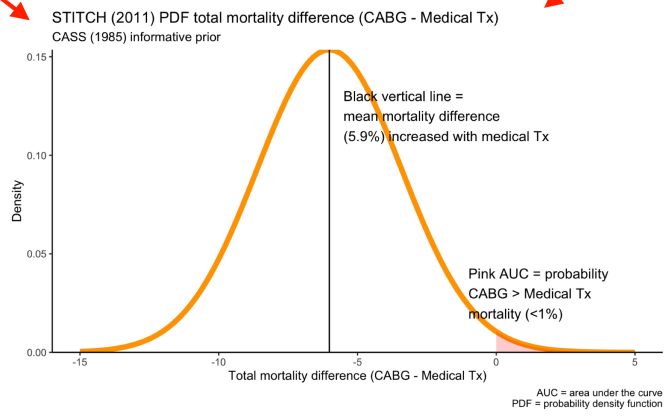


$$\text{posterior} = \frac{\text{prior} \cdot \text{likelihood}}{\text{normalizing constant}} \propto \text{prior} \cdot \text{likelihood}$$

Informative prior - CASS (1985)



STICH updated belief

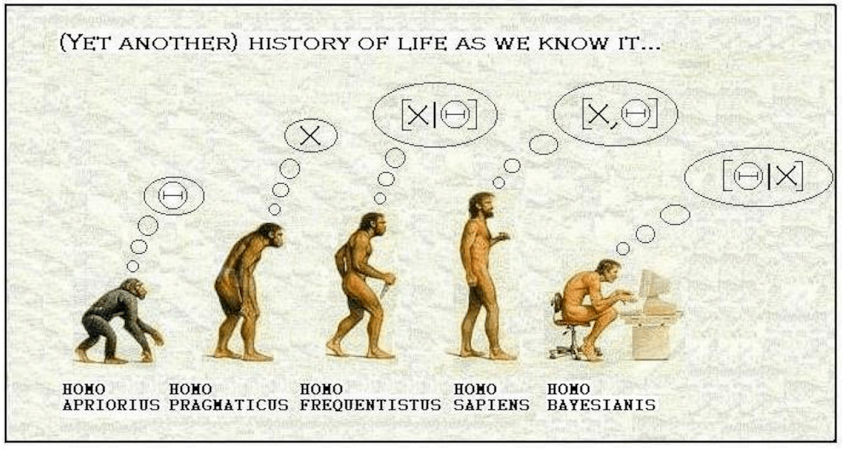


NEJM (2011) conclusion - **no significant changes in mortality**

Bayesian conclusion - **99% probability of decreased mortality with CABG**

NEJM (2016) conclusion - **mortality significantly lower with CABG**

Fighting for truth, justice and subjective probability



- Possibilities consistent with the data $\rightarrow$ more credibility,
- Possibilities not consistent $\rightarrow$ lose credibility.
- Bayesian analysis $\rightarrow$ mathematics of re-allocating credibility in a logically coherent and precise way.

- Street cred (https://twitter.com/d_spiegel/status/550677361205977088)

Ready to make "Inferences"?

QUESTIONS?

COMMENTS?

RECOMMENDATIONS?

Other resources

- Goodman S. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med. 1999;130:995-1004.
- Goodman S. Toward Evidence-Based Medical Statistics. 2: The Bayes Factor. Annals Int Med 1999;130:1005-13.

Statistical inference - Example 3

A case-control study of statins and risk of glioma, reported OR = 0.75; 95 % CI 0.48–1.17 when comparing users (>90 Rx) to non-users.

The authors then made the following statements
1) "As compared with non-use of statins, use of statins was not associated with risk of glioma"
2) "This matched case-control study revealed a null association between statin use and risk of glioma"
Do you agree?

Both statements are flat-out wrong

- Misinterpreting that their CI included the null as meaning no association
- Tests of significance, by comparing p to α or by looking for null values within CI, are worse than useless, they are misleading and inhibit critical discussion
- Values just beyond the CI are only slightly less likely to have given rise to the observed data than are some of the values included in the CI

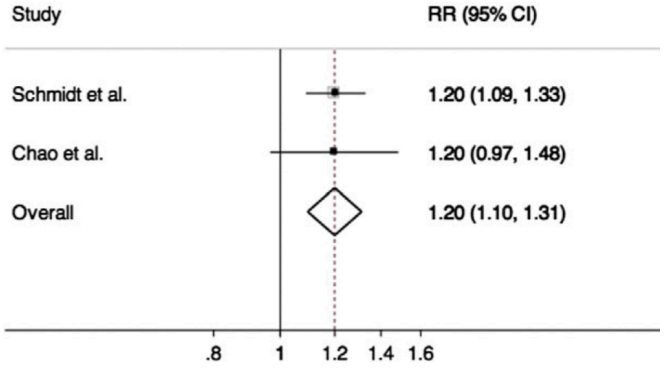
S-value graph (R - code)

```
s_graph <- function(hr, uci, lci){
  se <- (log(uci)-log(lci))/(2*1.96); x <- seq(0.01, 0.50,by = .005)
  lci <- exp(log(hr) - (qnorm(x) * se));uci <- exp(log(hr) + (qnorm(x) * se))
  lci <- rev(lci); hr <- rev(c(uci, lci))
  yy <- 2*x; yy <- c(yy,rev(yy)); ss <- -log(yy, base=2); df1 <- data.frame(hr,ss);
  df1 <- df1[-297,]
  s <- ggplot(df1, aes( hr,ss)) +  geom_line() + xlim(0.01,1.2) +
    scale_x_continuous(trans='log10') +
    ylab("Bits of information against HR (binary S value)") +
    xlab("Hazard ratio (Log scale)") +
    labs (subtitle = "S-Values (surprisals) for a range of hazard ratios (HR)") +
    geom_vline(xintercept=1.0, color = "green") +
    annotate("text", x=1,y=.4, label="null \nhypothesis") + theme_bw()
  return(s) }
gg <- s_graph(0.86, 1.04, 0.72) + labs(title="Stich trial results 2011") +
  annotate("text", x=.8,y=1, label="Maximum likelihood estimate (HR=0.86)\n
           has the least refutational evidence \n against it (0 bits)") +
  geom_segment(aes(x = .86, y = 0.8, xend = .86, yend = 0.015),
               arrow = arrow(length = unit(0.5, "cm")),color="red")
```


- Rafi, Z., Greenland, S. Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise. BMC Med Res Methodol 20, 244 (2020).
- Greenland, S. (2019). Valid P-Values Behave Exactly as They Should: Some Misleading Criticisms of P-Values and Their Resolution With S-Values.

Statistical inference - Example 1

- A study reported that selective COX-2 inhibitors (NSAIDs) **were associated** with atrial fibrillation (RR 1.20, 95% CI 1.09 - 1.33, p<0.01)
- A 2nd study concluded “use of selective COX-2 inhibitors **was not significantly related** to atrial fibrillation occurrence” (RR 1.20, 95% CI 0.97 - 1.47, p=.23)
- Authors elaborated why the results were different - different populations, etc
  Are the 2 results are really different?



Only difference is better precision in 1st study, the 2nd study actually supports the 1st
Data visualization helps again!

Message: Don’t rely on statistical significance testing for inferences

Statistical inference - Example 2

A recent 2022 study reported *"annual screening (vs some screening) was associated with a* ***significant reduction in risk of prostate cancer–specific mortality (PCSM) among Black men (sHR, 0.65; 95% CI, 0.46-0.92; P = .02)***

- *but not among White men (sHR, 0.91; 95%CI, 0.74-1.11; P = .35)"* and then concluded:
- *Annual screening was associated with reduced risk of PCSM among Black men but not among White men, suggesting that annual screening may be particularly important for Black men.*

Are the 2 results are really different?

Probably NOT!

- **Reference #1** The Difference Between “Significant” and “Not Significant” is not Itself Statistically Significant
- **Reference #2** Interaction revisited: the difference between two estimates