Statistical inference

Mabel Carabali

EBOH, McGill University

2024/08/29 (updated: 2024-09-03)

Housekeeping:

- **Rapid Reviews**
  - Presentations Schedule
  - **Perussall** course can be accessed here
  - Readings and *Perusall* links will be posted on mycourses
- **Slides, coding and data**
  - EPIB 704 GitHub Repository
  - Data for assignments can be found in the `/EPIB-704/tree/main/data` folder.
  - Slides on html format can be accessed using the README.md Table of content

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

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

Before statistical inference

Before statistical inference, there is proper **study design** and **data collection**

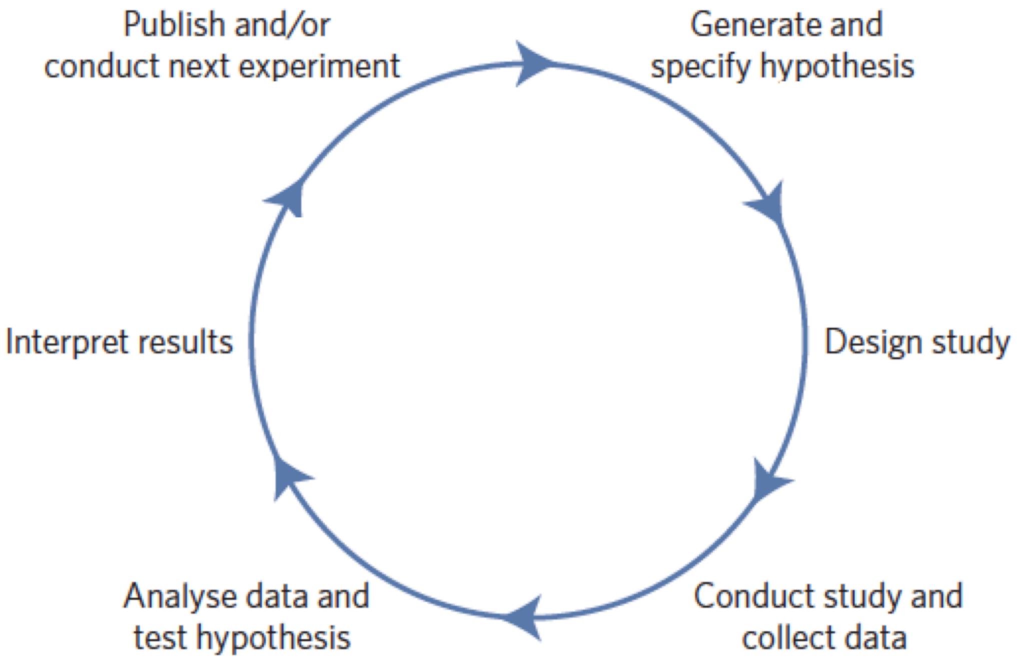
- Plenty of places to go wrong before statistical inference

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?

Metascience

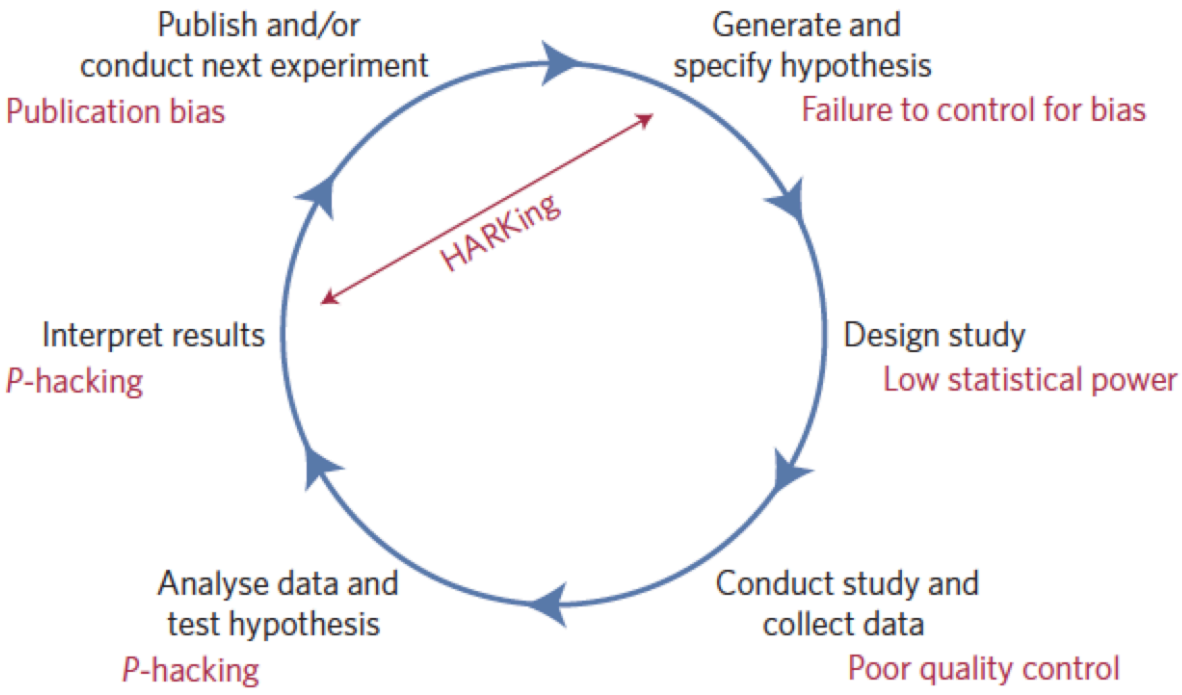
The scientific study of science itself: **Hypothetico-deductive model of the scientific method**


Generate and specify hypothesis
Design study
Conduct study and collect data
Analyse data and test hypothesis
Interpret results
Publish and/or conduct next experiment

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


Publish and/or conduct next experiment
Publication bias
Generate and specify hypothesis
Failure to control for bias
HARKing
Interpret results
P-hacking
Design study
Low statistical power
Analyse data and test hypothesis
P-hacking
Conduct study and collect data
Poor quality control

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).

Researcher degrees of freedom

Most often done in good faith → vibration of effects

Consider the question: "Does "skin color" influence red cards in football (soccer)?"

Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results

- Crowd source research project used **1 dataset** and provided to 29 experienced analytic teams
- Teams initially worked independently
- But before final submission, each team's methods (without results) were circulated to the other teams and experts for review comments
- Teams could then revise their methods or even change them before their final submission

Silberzahn R, et al.2018;1(3):337-356. here

Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results

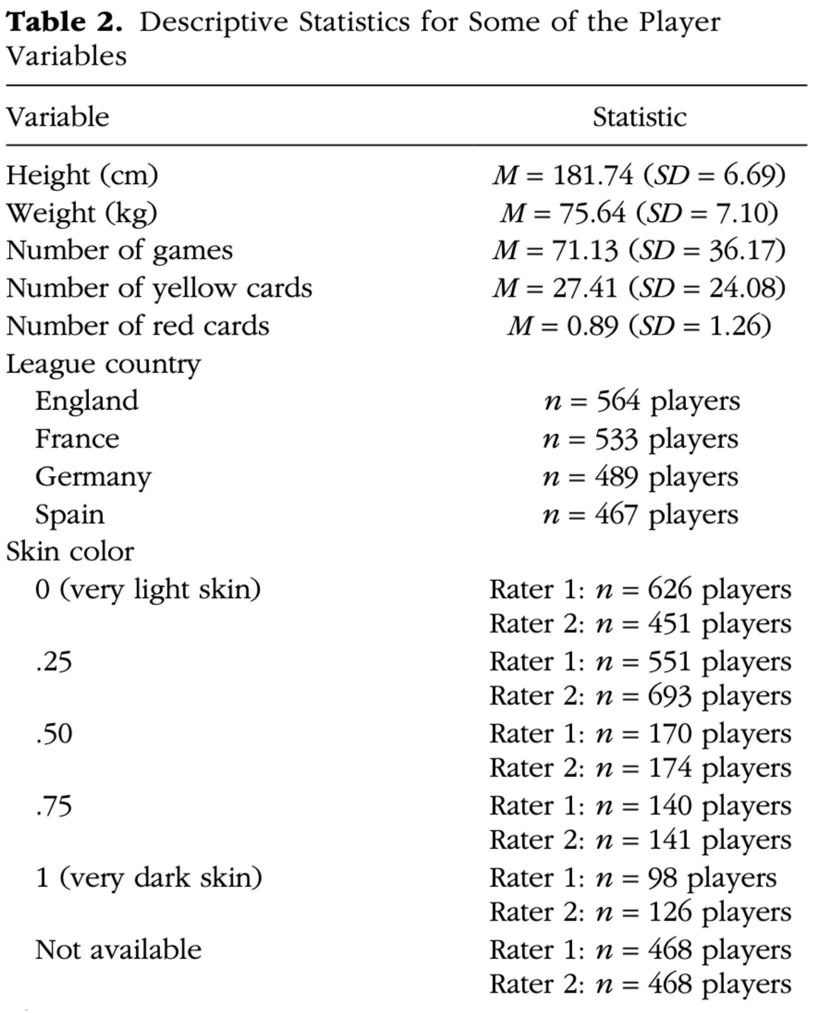


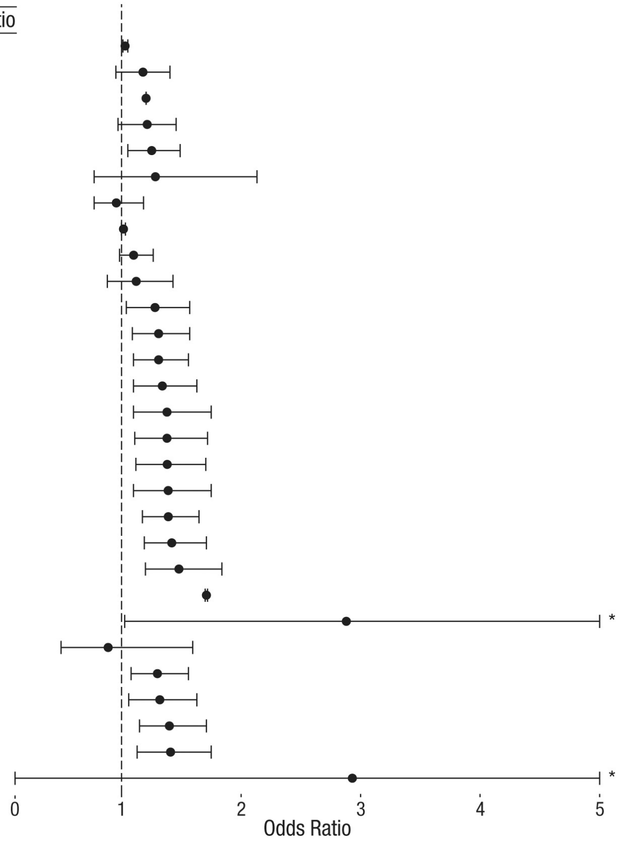
Table 2. Descriptive Statistics for Some of the Player Variables

Variable	Statistic
Height (cm)	*M* = 181.74 (*SD* = 6.69)
Weight (kg)	*M* = 75.64 (*SD* = 7.10)
Number of games	*M* = 71.13 (*SD* = 36.17)
Number of yellow cards	*M* = 27.41 (*SD* = 24.08)
Number of red cards	*M* = 0.89 (*SD* = 1.26)
League country	
England	*n* = 564 players
France	*n* = 533 players
Germany	*n* = 489 players
Spain	*n* = 467 players
Skin color	
0 (very light skin)	Rater 1: *n* = 626 players Rater 2: *n* = 451 players
.25	Rater 1: *n* = 551 players Rater 2: *n* = 693 players
.50	Rater 1: *n* = 170 players Rater 2: *n* = 174 players
.75	Rater 1: *n* = 140 players Rater 2: *n* = 141 players
1 (very dark skin)	Rater 1: *n* = 98 players Rater 2: *n* = 126 players
Not available	Rater 1: *n* = 468 players Rater 2: *n* = 468 players

"Photos for 1,586 of the 2,053 players were available from our source….The variable player's skin tone was coded by two independent raters blind to the research question. On the basis of the photos, the raters categorized the players on a 5-point scale ranging from 1 (very light skin) to 3 (nei- ther dark nor light skin) to 5 (very dark skin)."

Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results

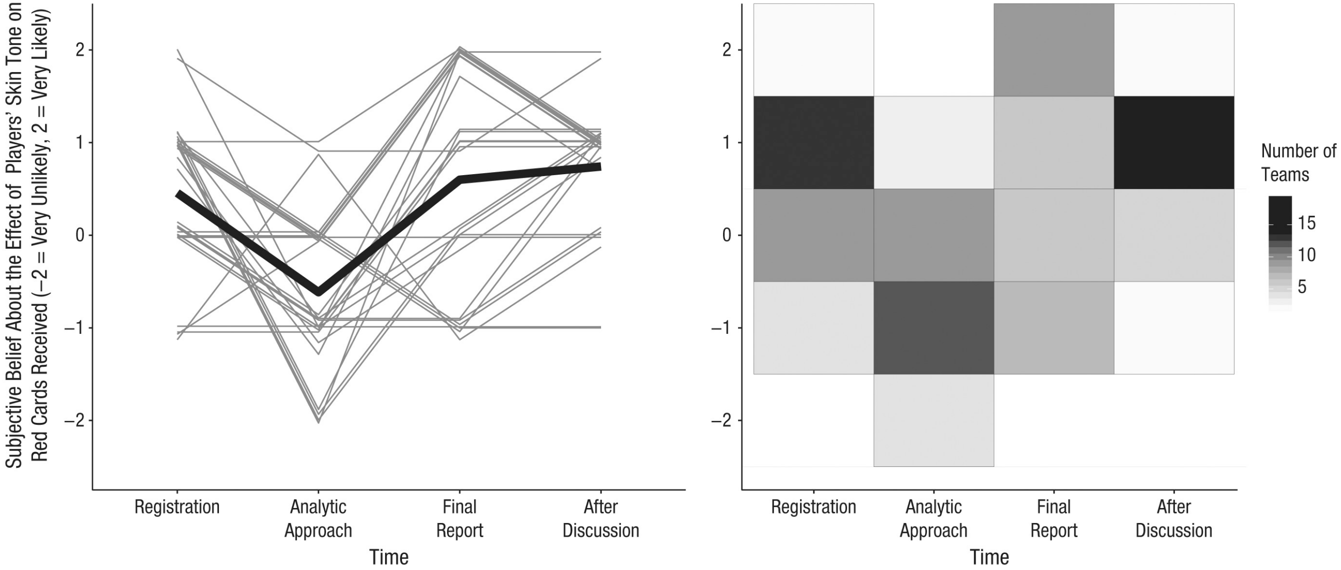
Team	Analytic Approach	Distribution	Odds Ratio
10	Multilevel Regression and Logistic Regression	Linear	1.03
1	OLS Regression With Robust Standard Errors, Logistic Regression	Linear	1.18
4	Spearman Correlation	Linear	1.21
14	WLS Regression With Clustered Standard Errors	Linear	1.21
11	Multiple Linear Regression	Linear	1.25
6	Linear Probability Model	Linear	1.28
17	Bayesian Logistic Regression	Logistic	0.96
15	Hierarchical Log-Linear Modeling	Logistic	1.02
18	Hierarchical Bayes Model	Logistic	1.10
31	Logistic Regression	Logistic	1.12
30	Clustered Robust Binomial Logistic Regression	Logistic	1.28
3	Multilevel Logistic Regression Using Bayesian Inference	Logistic	1.31
23	Mixed-Model Logistic Regression	Logistic	1.31
2	Linear Probability Model, Logistic Regression	Logistic	1.34
5	Generalized Linear Mixed Models	Logistic	1.38
24	Multilevel Logistic Regression	Logistic	1.38
28	Mixed-Effects Logistic Regression	Logistic	1.38
32	Generalized Linear Models for Binary Data	Logistic	1.39
8	Negative Binomial Regression With a Log Link	Logistic	1.39
25	Multilevel Logistic Binomial Regression	Logistic	1.42
9	Generalized Linear Mixed-Effects Models With a Logit Link	Logistic	1.48
7	Dirichlet-Process Bayesian Clustering	Misc	1.71
21	Tobit Regression	Misc	2.88
12	Zero-Inflated Poisson Regression	Poisson	0.89
26	Hierarchical Generalized Linear Modeling With Poisson Sampling	Poisson	1.30
16	Hierarchical Poisson Regression	Poisson	1.32
20	Cross-Classified Multilevel Negative Binomial Model	Poisson	1.40
13	Poisson Multilevel Modeling	Poisson	1.41
27	Poisson Regression	Poisson	2.93



Note: Each team's presented different effect sizes, here converted to ORs 95% confidence intervals (CIs). **OR ranged from 0.89 to 2.93 (median = 1.31)**; 21 unique covariate combinations; 69% p-values < 0.05; variability **not** explained by quality of analyses.

Researcher degrees of freedom

Teams’ subjective beliefs about the primary research question across time.



Analysts’ subjective beliefs about the research hypothesis were ***assessed four times during the project:*** *at registration, after accessing the data and submitting their analytic approach, when submitting final analyses, and after a group discussion of all the teams’ approaches and results. Responses were centered at 0, the range was from -2, for very unlikely, to +2, for very likely."* Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results

Author's Conclusion

"*The observed results from analyzing a complex data set can be highly contingent on justifiable, but subjective, analytic decisions. Uncertainty in interpreting research results is therefore not just a function of statistical power or the use of questionable research practices; it is also* ***a function of the many reasonable decisions that researchers must make in order to conduct the research****.*

This does not mean that analyzing data and drawing research conclusions is a subjective enterprise with no connection to reality. It does mean that ***many subjective decisions are part of the research process and can affect the outcomes. The best defense against subjectivity in science is to expose it****. Transparency in data, methods,and process gives the rest of the community opportunity to see the decisions, question them, offer alternatives, and test these alternatives in further research.*"

Another take: Subjective in data analysis is not restricted to Bayesian analyses which indeed make their subjectivities fully transparent (priors)

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
- Statistical models insufficient for causality
- Paradox - models that are causally incorrect can make better predictions than those that are causally correct
- Probability models connect noisy sample data and populations and represent the most effective way to obtain inference
- Inference is about belief revision, so Bayesian perspective seems logical and may provide additional insights (my personal, but not universally shared, belief)

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

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!

Statistical inference

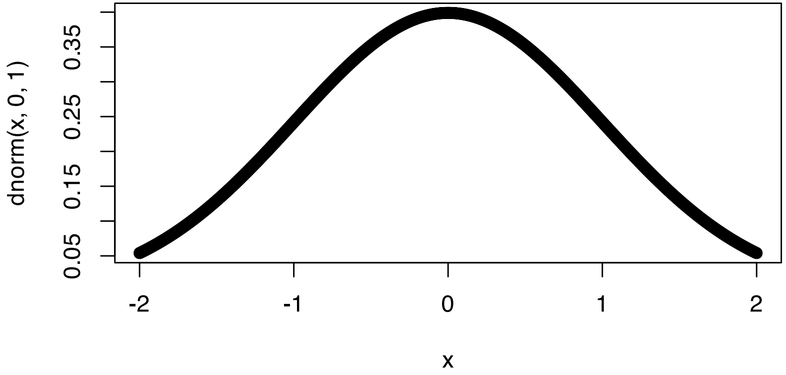
To make inferences we need to either refer to some common statistical distributions (normal, binomial, etc) or do simulations.

- 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))
```



Some probability distributions in R:

- 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 distributio

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

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

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

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

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

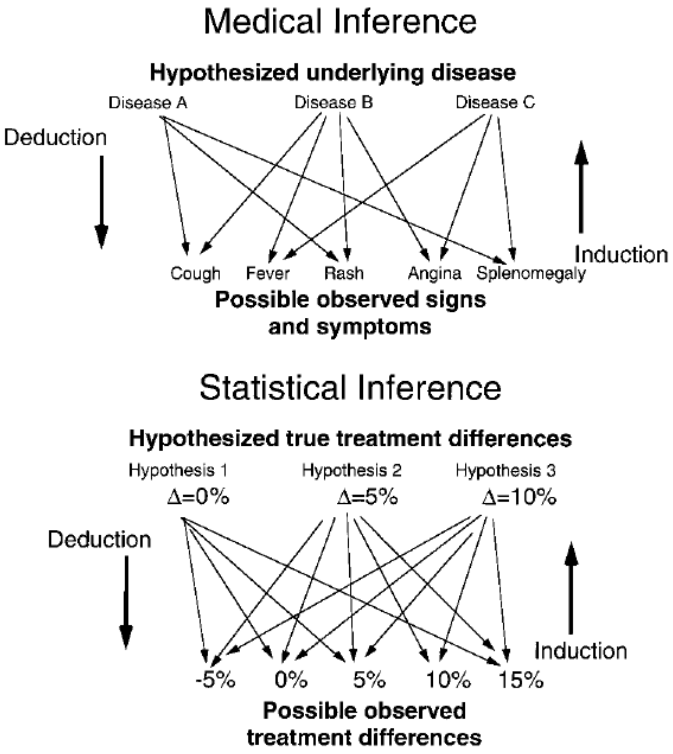
```
mean(rnorm(10000,0,1))
```

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

Check this resources

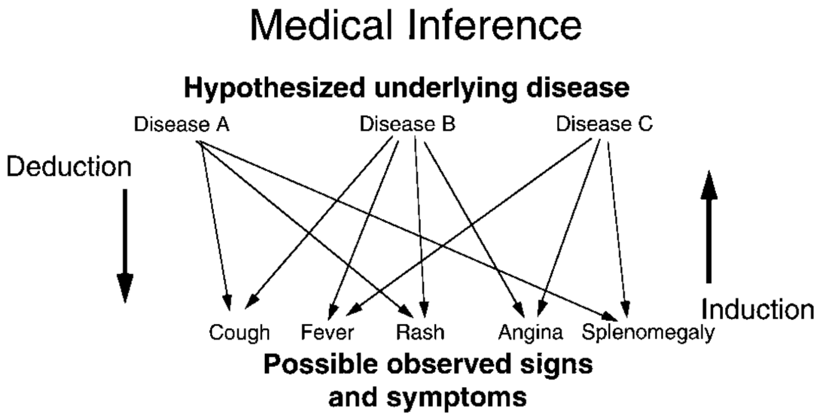
- Statistical Inference for Everyone
- Distribution functions in R
- A Guide to dnorm, pnorm, rnorm, and qnorm in R

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.

Deductive vs Inductive Inference (I)



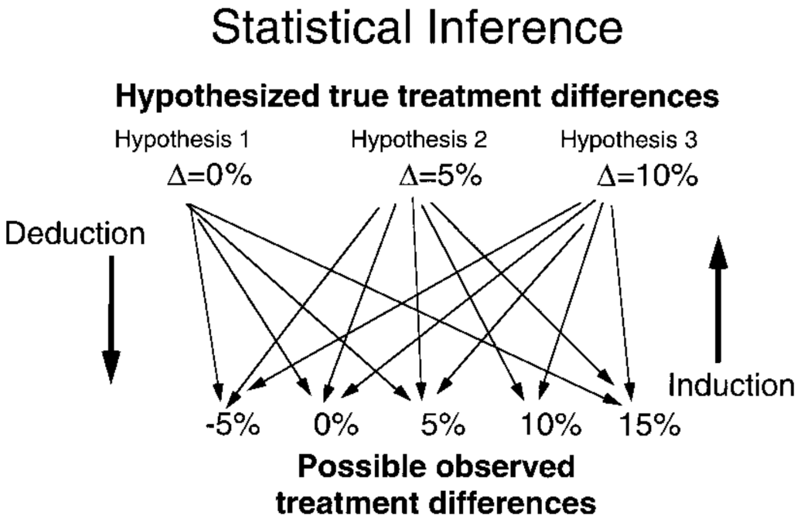
Deduction appears objective; predictions true **only if** H are true

- 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)



Contrasting 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?"*

Null hypothesis significance testing (NHST)

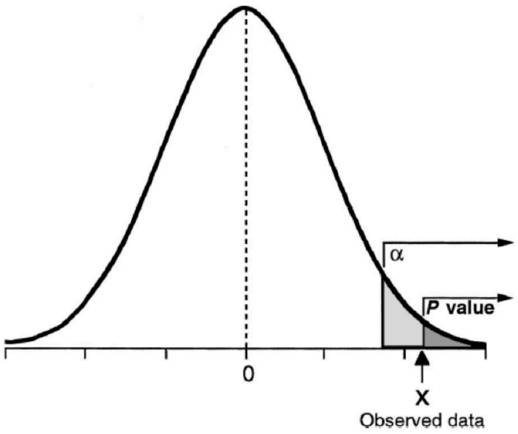


Figure 3. The bell-shaped curve represents the probability of every possible outcome under the null hypothesis. Both α (the type I error rate) and the *P* value are "tail areas" under this curve. The tail area for α is set before the experiment, and a result can fall anywhere within it. The *P* value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

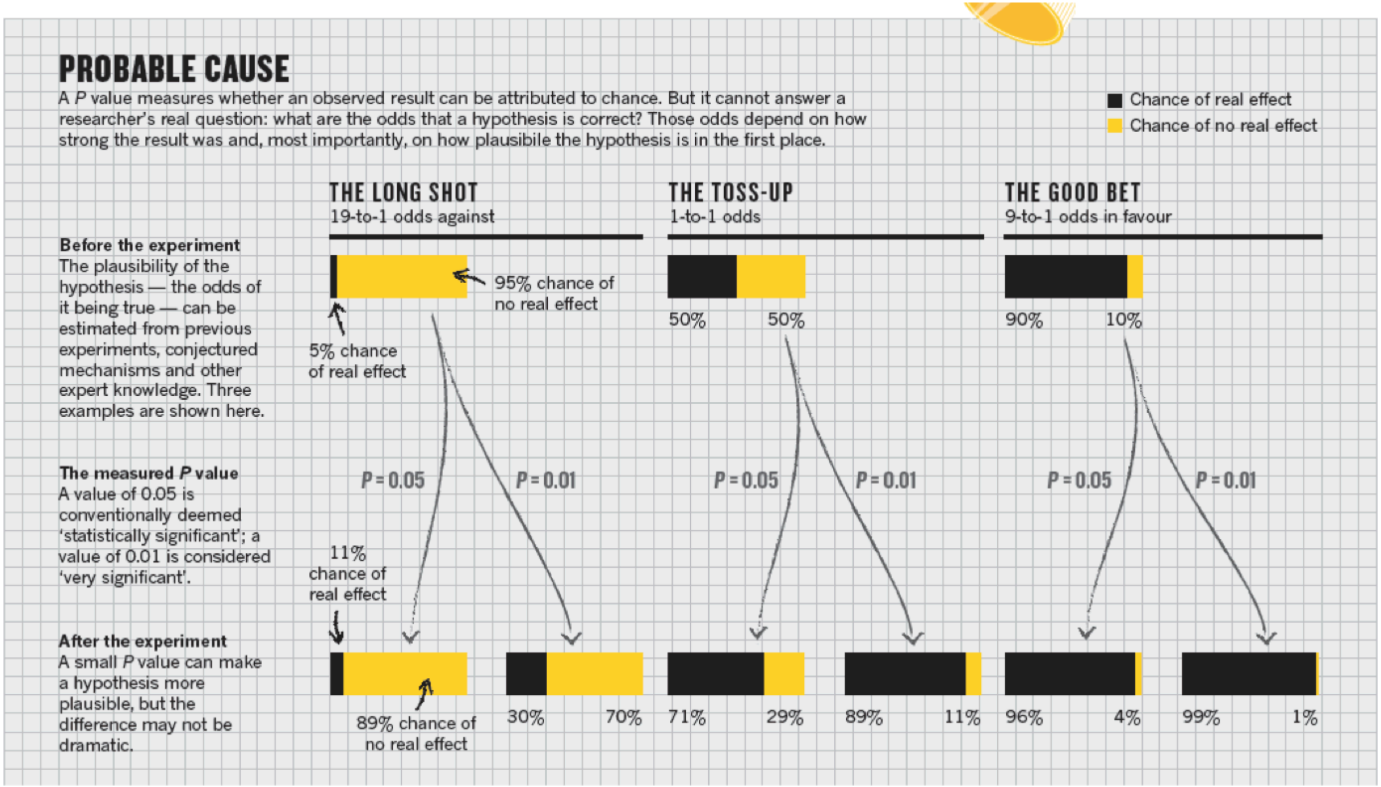
- State H_o, H_a, α error → rejection area
- Check if data falls into the rejection area
- If yes, reject the null and accept the alternative, if no, can only say you don't have enough evidence to reject

Concerns with p values

- misinterpret as the "probability that the studied hypothesis is true"
- poor measure of strength of evidence; same value with small effect & large study as with large effect in small study
- often confused with α error
- can't provide both "short run" evidential perspective which is inductive & the long-run perspective, which is error-based and deductive experiment
- often used to make "scientific conclusions & policy decisions" when it provides no measure of effect size

P value fallacy

The mistaken idea that a single number can capture both the long-run outcomes of an experiment and the evidential meaning of a single result



PROBABLE CAUSE

A *P* value measures whether an observed result can be attributed to chance. But it cannot answer a researcher's real question: what are the odds that a hypothesis is correct? Those odds depend on how strong the result was and, most importantly, on how plausibile the hypothesis is in the first place.

■ Chance of real effect

🟨 Chance of no real effect

Before the experiment
The plausibility of the hypothesis — the odds of it being true — can be estimated from previous experiments, conjectured mechanisms and other expert knowledge. Three examples are shown here.

The measured *P* value
A value of 0.05 is conventionally deemed 'statistically significant'; a value of 0.01 is considered 'very significant'.

After the experiment
A small *P* value can make a hypothesis more plausible, but the difference may not be dramatic.

Other problems with statistical significance

- Statistical significance ≠ practical significance
- Non-significance $\neq$ zero effect
- Δ between statistically significant and not statistically significant is not itself statistically significant
- Research degrees of freedom, p hacking & forking paths
- Statistical significance filter
- Doesn't respect the likelihood principle (all the evidence in a sample relevant to model parameters is contained in the likelihood function)

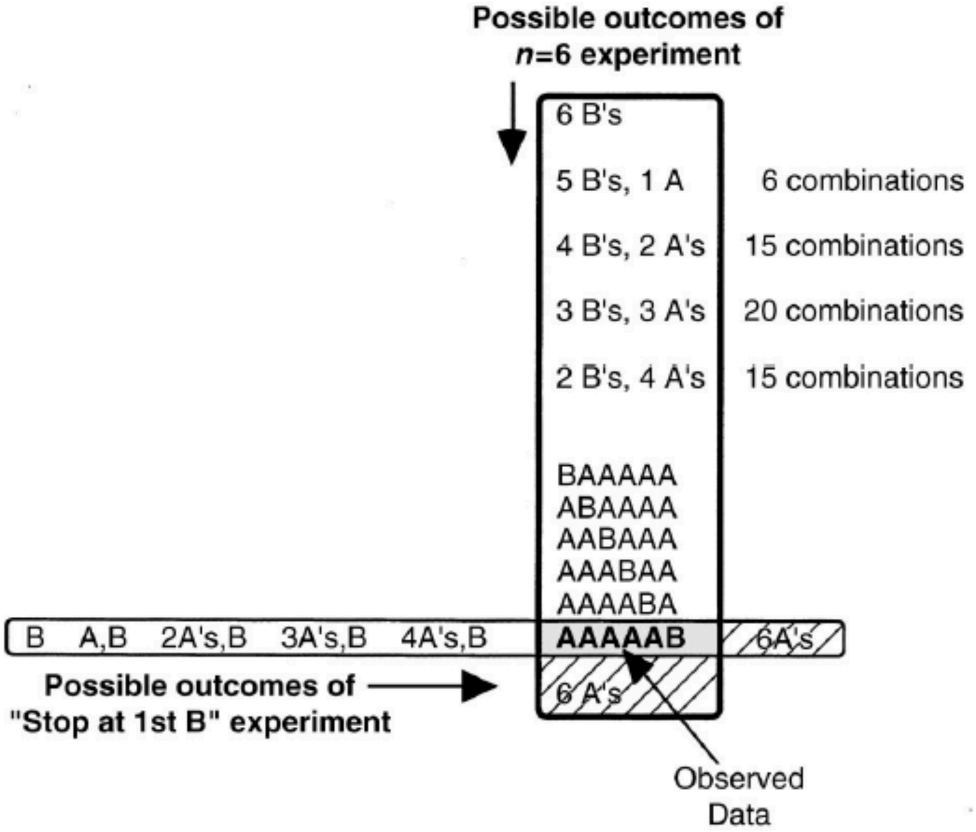
Reference Gelman, Andrew; Hill, Jennifer; Vehtari, Aki. Regression and Other Stories

> A final concern is that statistically significant estimates tend to be overestimates.
>
> This is the type M, or magnitude, error problem discussed in Section 4.4. Any estimate with $p < 0.05$ is by necessity at least two standard errors from zero. If a study has a high noise level, standard errors will be high, and so statistically significant estimates will automatically be large, no matter how small the underlying effect. Thus, **routine reliance on published, statistically significant results will lead to systematic overestimation of effect sizes and a distorted view of the world.** All the problems discussed above have led to what has been called a replication crisis, in which studies published in leading scientific journals and conducted by researchers at respected universities have failed to replicate. Many different problems in statistics and the culture of science have led to the replication crisis; for our purposes here, what is relevant is to understand how to avoid some statistical misconceptions associated with overcertainty.

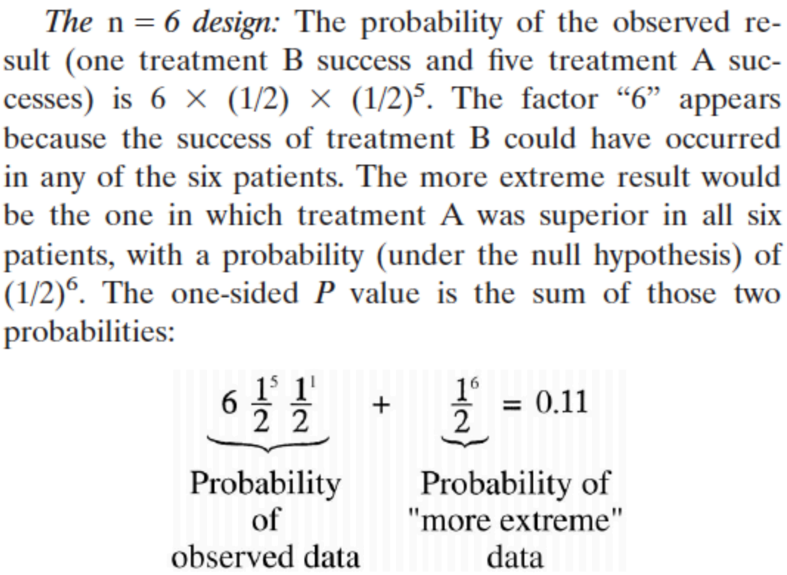
- Gelman, Andrew; Hill, Jennifer; Vehtari, Aki. Regression and Other Stories
- Why Most Published Research Findings Are False

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...

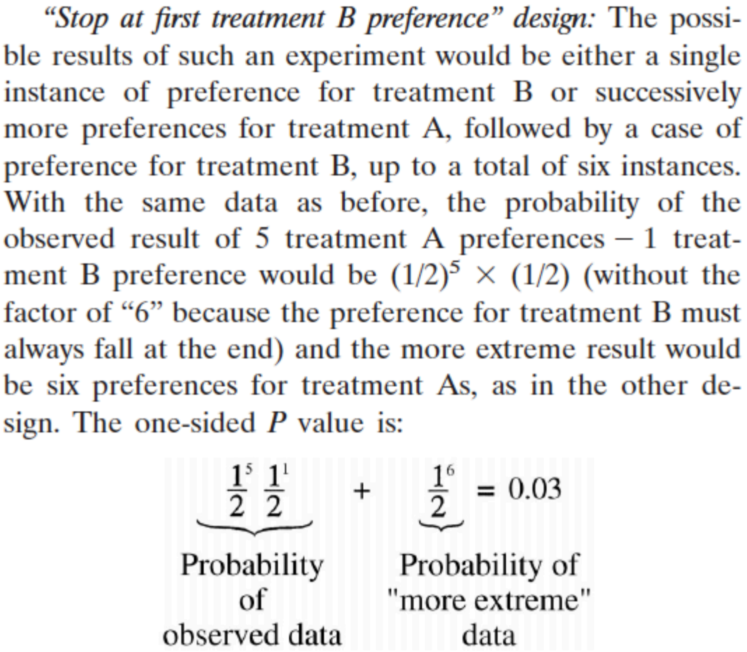


Likelihood principle



The n = *6 design:* The probability of the observed result (one treatment B success and five treatment A successes) is 6 × (1/2) × $(1/2)^5$. 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:

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

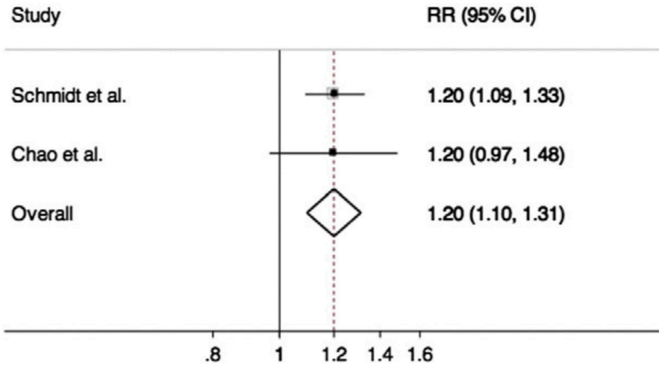


"Stop at first treatment B preference" design: The possible results of such an experiment would be either a single instance of preference for treatment B or successively more preferences for treatment A, followed by a case of preference for treatment B, up to a total of six instances. With the same data as before, the probability of the observed result of 5 treatment A preferences – 1 treatment B preference would be $(1/2)^5$ × (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:

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

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

Simple R function

```r
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
```

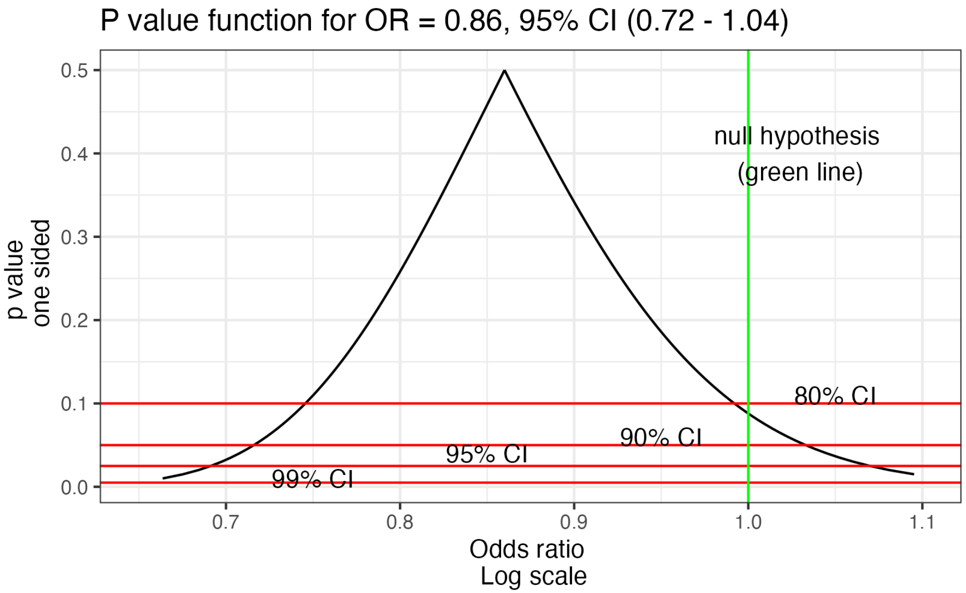
Author's Conclusion:

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!

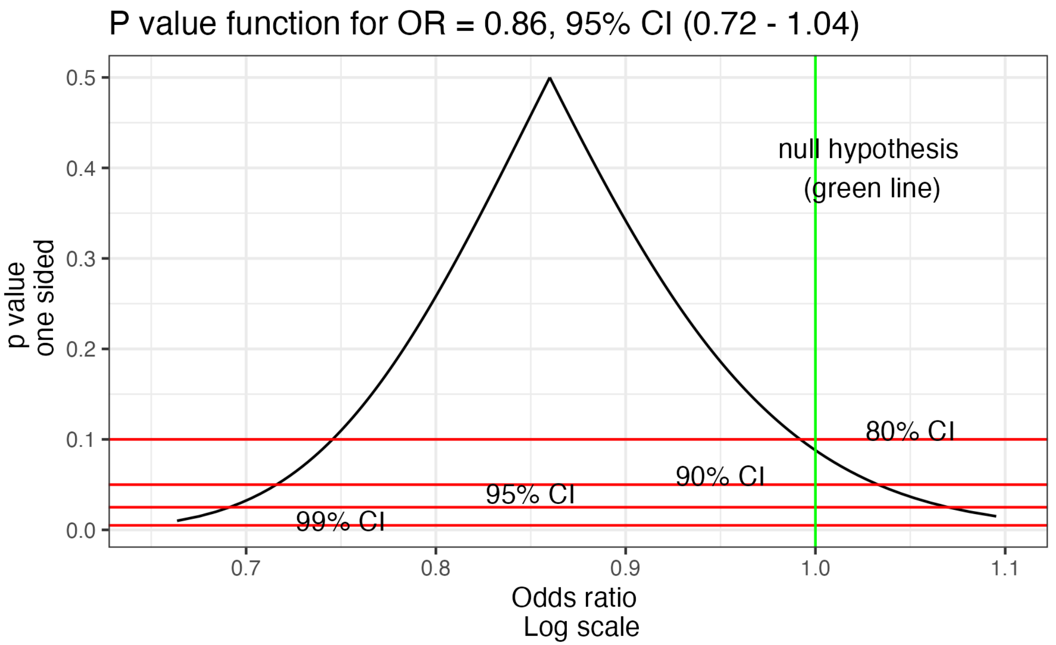
Avoid dichotomania

- 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) → “*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)

```r
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

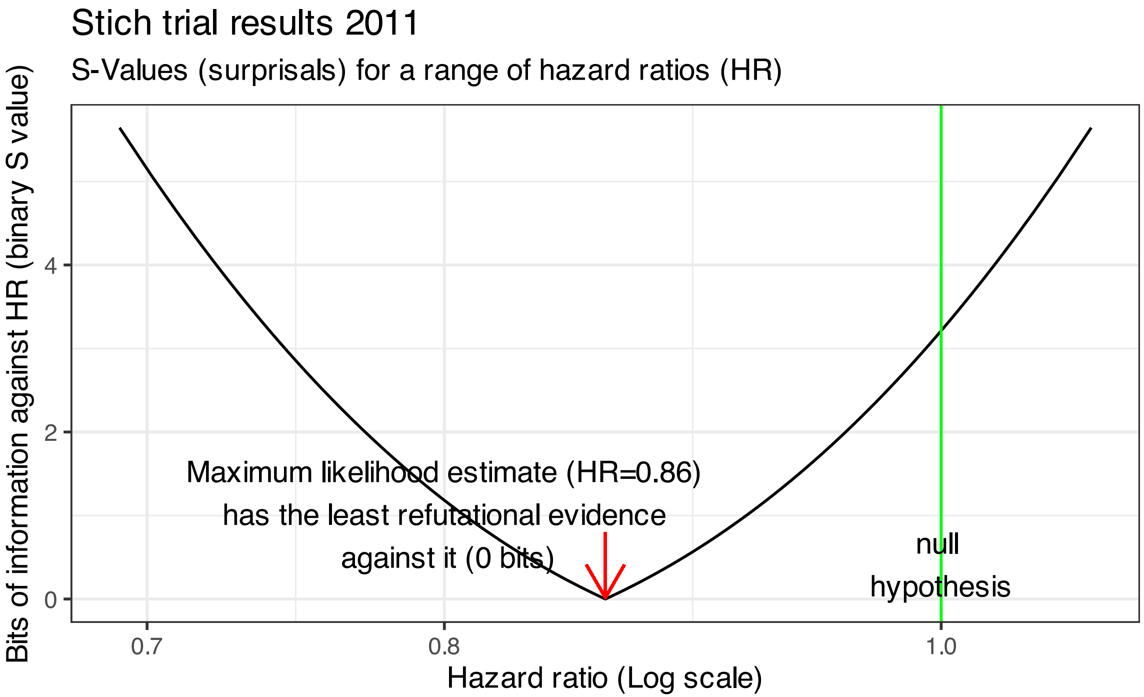
Evidence **against** not only H_o but against any specific H_a better appreciated by considering the binary Shannon information, surprisal or **S value**.

- $s = log_2(\frac{1}{P})$ 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!



S-value graph (R - code)

```r
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.

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 $\longrightarrow$ 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

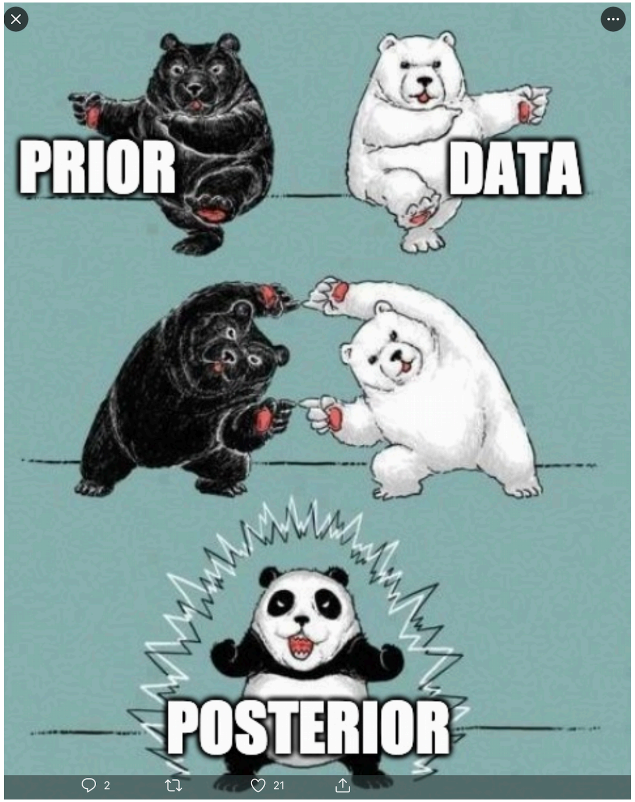
Disadvantages

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

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?

Bayes rule (conceptual)



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

Bayes rule

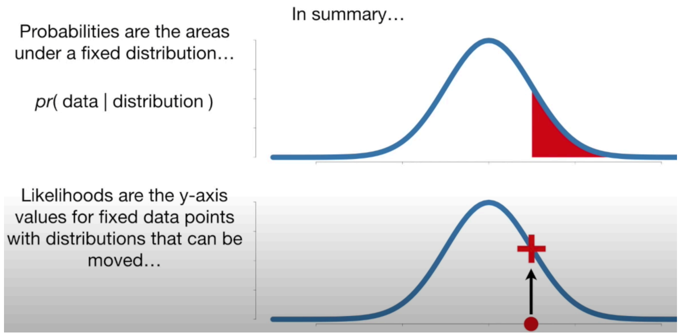
Likelihood - propensity for observing the data given a certain value of θ

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

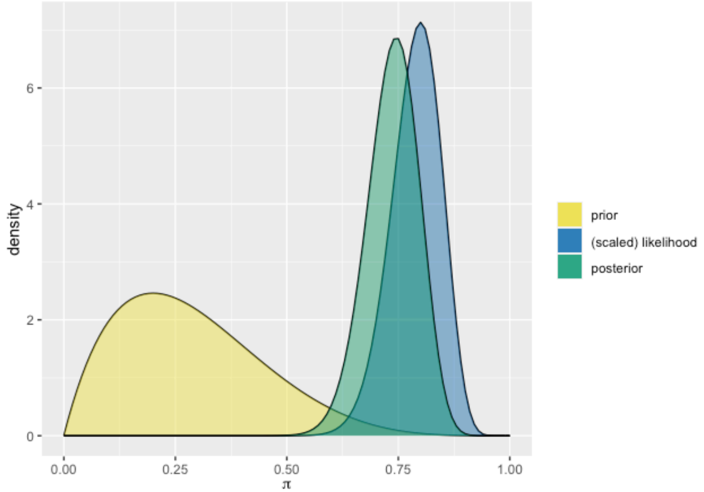


In summary…

Probabilities are the areas under a fixed distribution…

pr(data | distribution)

Likelihoods are the y-axis values for fixed data points with distributions that can be moved…



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

```r
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

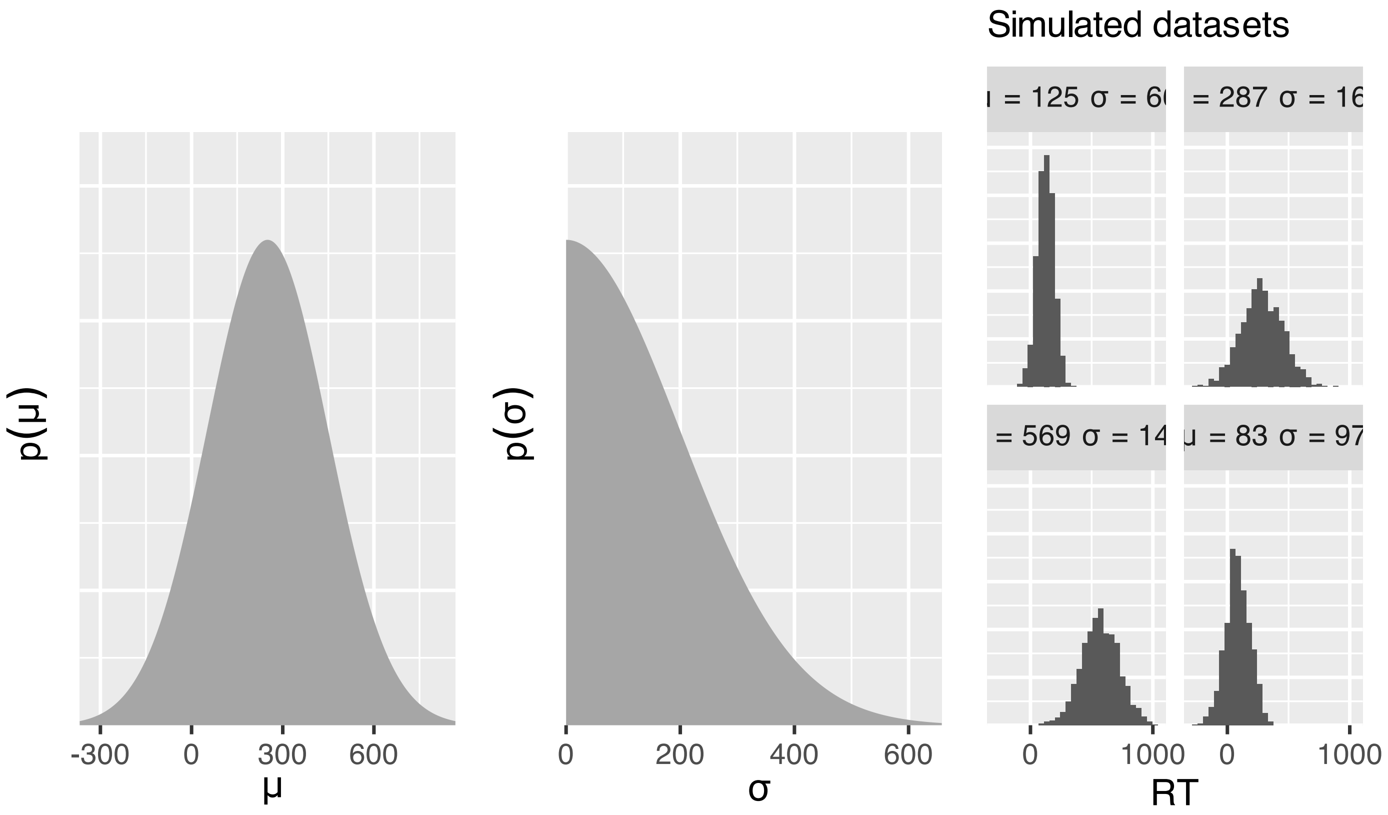
$$y_n \sim N(\mu, \sigma^2)$$

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.

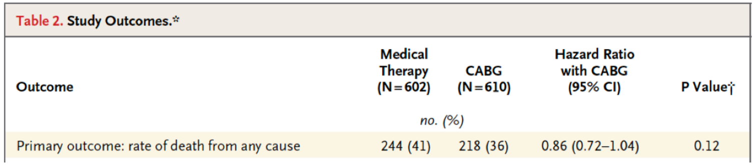
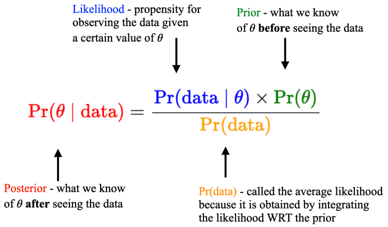


Table 2. Study Outcomes.*

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



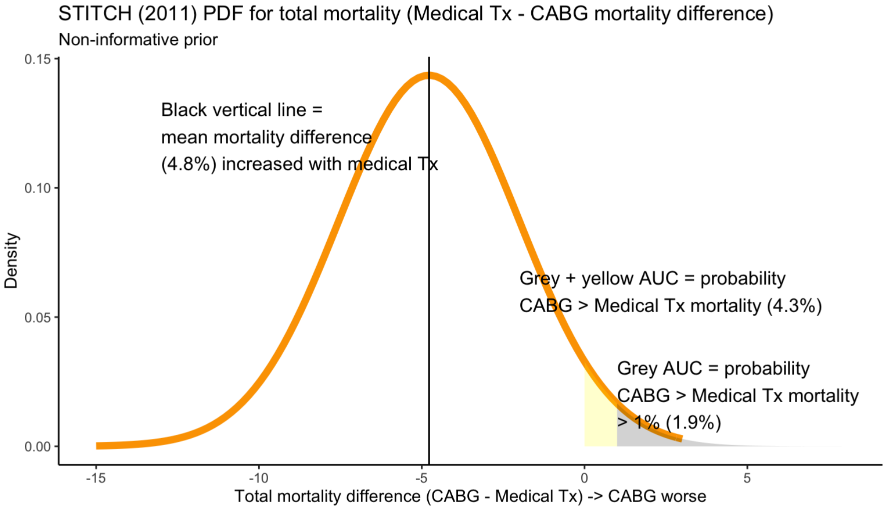
Likelihood - propensity for observing the data given a certain value of θ

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



AUC = area under the curve
PDF = probability density function

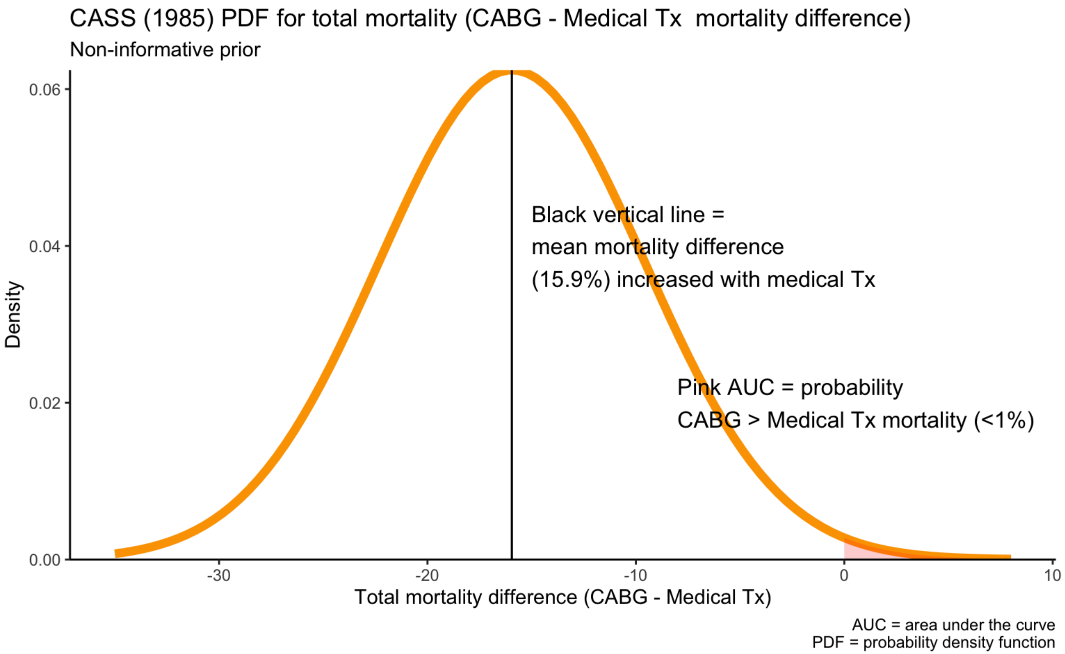
Bayesian example - informative prior

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%)



CASS (1985) PDF for total mortality (CABG - Medical Tx mortality difference)

Non-informative prior

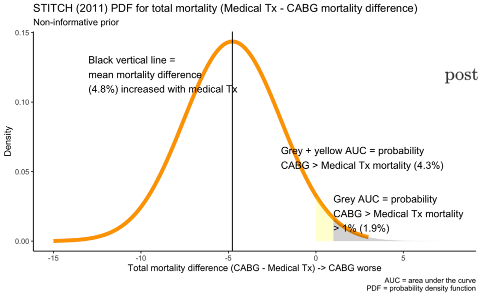
Black vertical line = mean mortality difference (15.9%) increased with medical Tx

Pink AUC = probability CABG > Medical Tx mortality (<1%)

AUC = area under the curve
PDF = probability density function

Bayesian example - updated

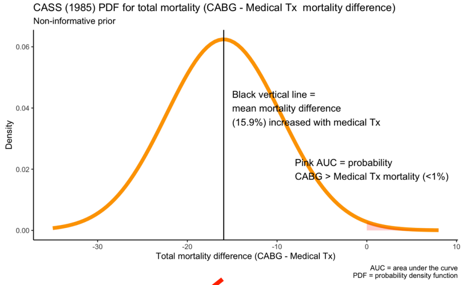
STICH data


STITCH (2011) PDF for total mortality (Medical Tx - CABG mortality difference)

Non-informative prior

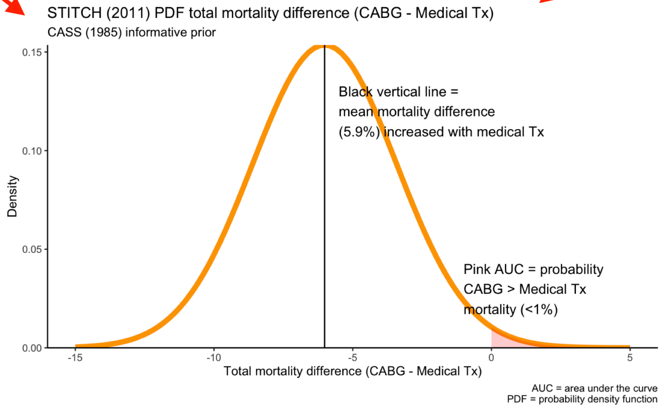
$$\text{posterior} = \frac{\text{prior} \cdot \text{likelihood}}{\text{normalizing constant}} \propto \text{prior} \cdot \text{likelihood}$$

Informative prior - CASS (1985)


CASS (1985) PDF for total mortality (CABG - Medical Tx mortality difference)

Non-informative prior

STICH updated belief


STITCH (2011) PDF total mortality difference (CABG - Medical Tx)

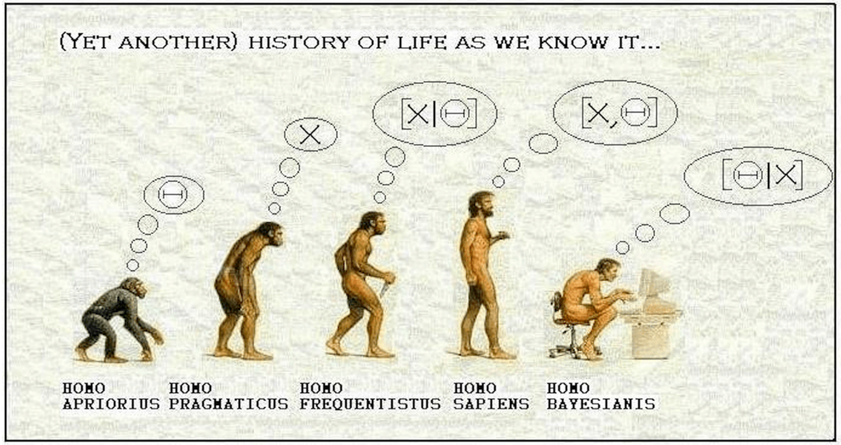
CASS (1985) informative prior

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 → more credibility,
- Possibilities not consistent → lose credibility.
- Bayesian analysis → mathematics of re-allocating credibility in a logically coherent and precise way.

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

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