PHARMALEX
CONFIDENCE BEYOND COMPLIANCE

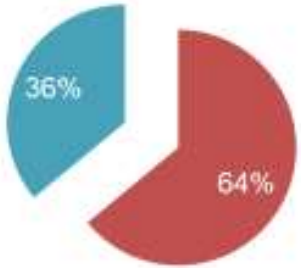
Reproducibility from Discovery to Clinical Research
What could be the way out?

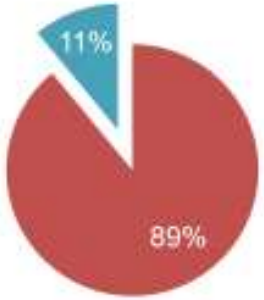
Bruno Boulanger
November 2019

Agenda

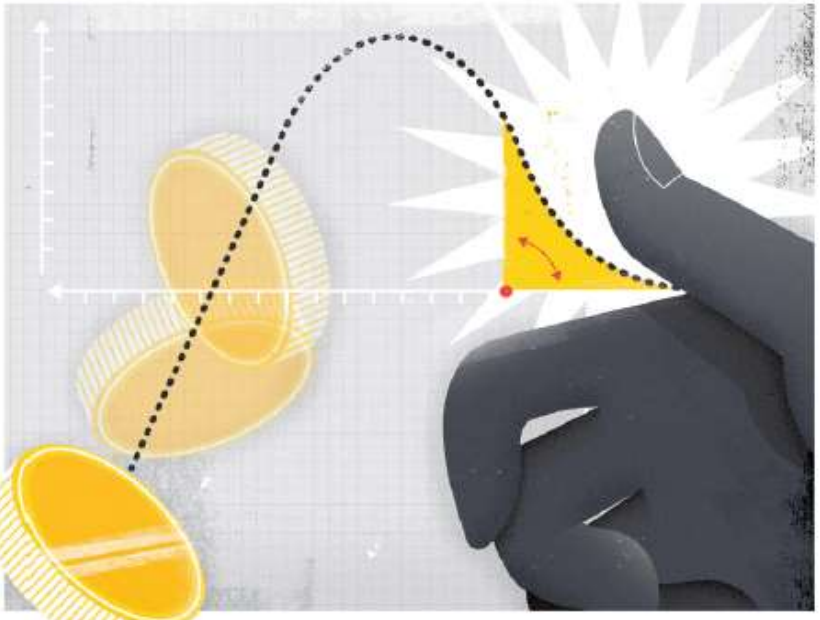
- The reproducibility crisis, i.e. the replicability crisis
- The p-value crisis
- How to make a decision? What is the question ?
- The Bayesian learning process in Pharmaceutical R&D
- The posterior predictive distribution
- The power, the Bayesian power and the Assurance
- The missing component: the elephant in the room ?
- Take away message

The Replicability crisis: the beginning





Nature, 2014



STATISTICAL ERRORS

P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume.

BY REGINA NUZZO

As the American Statistical Association officially reminded in March 2016….

REPRODUCIBILITY

Statisticians issue warning on *P* values

Statement aims to halt missteps in the quest for certainty.

BY MONYA BAKER

Misuse of the *P* value — a common test for judging the strength of scientific evidence — is contributing to the number of research findings that cannot be reproduced, the American Statistical Association (ASA) warned on 8 March. The group has taken the unusual step of issuing principles to guide use of the *P* value, which it says cannot determine whether a hypothesis is true or whether results are important.

This is the first time that the 177-year-old ASA has made explicit recommendations on such a foundational matter, says executive director Ron Wasserstein. The society's members had become increasingly concerned that the *P* value was being misapplied, in ways that cast doubt on statistics generally, he adds.

cannot indicate the importance of a finding; for instance, a drug can have a statistically significant effect on patients' blood glucose levels without having a therapeutic effect.

Giovanni Parmigiani, a biostatistician at the Dana Farber Cancer Institute in Boston, Massachusetts, says that misunderstandings about what information a *P* value provides often crop up in textbooks and practice manuals. A course correction is long overdue, he adds. "Surely if this happened twenty years ago, biomedical research could be in a better place now."

FRUSTRATION ABOUNDS

Criticism of the *P* value is nothing new. In 2011, researchers trying to raise awareness about false positives gamed an analysis to reach a statistically significant finding: that listening to music by the Beatles makes undergraduates younger

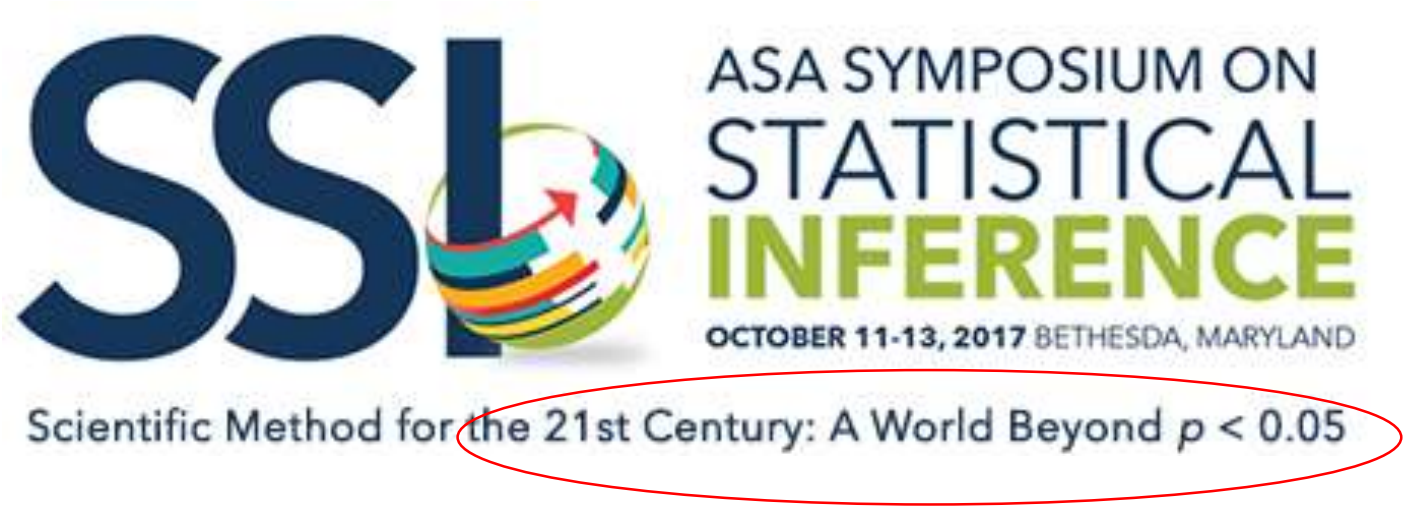
Misuse of the *P* value — a common test for judging the strength of scientific evidence — is contributing to the number of research findings that cannot be reproduced, the American Statistical Association (ASA) warned on 8 March. The group

The American Statistical Association reminded in a press release some key points:

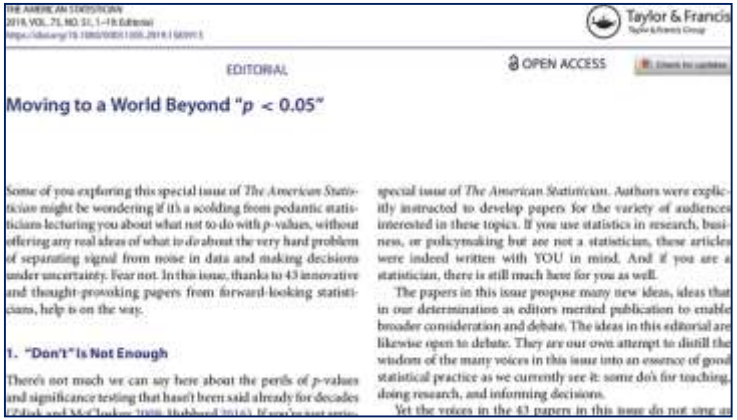
1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. **Scientific conclusions and business or policy decisions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

Reproducibility now a public concern

Emergency meeting: ASA Symposium in October 2017.



The American Statistician - 2019



Taylor & Francis

EDITORIAL

OPEN ACCESS

Moving to a World Beyond "$p < 0.05$"

Some of you exploring this special issue of *The American Statistician* might be wondering if it's a scolding from pedantic statisticians lecturing you about what not to do with *p*-values, without offering any real ideas of what to *do* about the very hard problem of separating signal from noise in data and making decisions under uncertainty. Fear not. In this issue, thanks to 43 innovative and thought-provoking papers from forward-looking statisticians, help is on the way.

1. "Don't" Is Not Enough

There's not much we can say here about the perils of *p*-values and significance testing that hasn't been said already for decades

special issue of *The American Statistician*. Authors were explicitly instructed to develop papers for the variety of audiences interested in these topics. If you use statistics in research, business, or policymaking but are not a statistician, these articles were indeed written with YOU in mind. And if you are a statistician, there is still much here for you as well.

The papers in this issue propose many new ideas, ideas that in our determination as editors merited publication to enable broader consideration and debate. The ideas in this editorial are likewise open to debate. They are our own attempt to distill the wisdom of the many voices in this issue into an essence of good statistical practice as we currently see it: some do's for teaching, doing research, and informing decisions.

Yet the voices in the 43 papers in this issue do not sing as

THE AMERICAN STATISTICIAN
2019, VOL. 73, NO. S1, 26–30: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2018.1558111

Taylor & Francis

OPEN ACCESS

Why is Getting Rid of *P*-Values So Hard? Musings on Science and Statistics

Steven N. Goodman

Departments of Medicine and Health Research and Policy, Stanford University, Stanford, CA

ABSTRACT

The current concerns about reproducibility have focused attention on proper use of statistics across the sciences. This gives statisticians an extraordinary opportunity to change what are widely regarded as statistical practices detrimental to the cause of good science. However, how that should be done is enormously complex, made more difficult by the balkanization of research methods and statistical traditions across scientific subdisciplines. Working within those sciences while also allying with science reform movements—operating simultaneously on the micro and macro levels—are the key to making lasting change in applied science.

ARTICLE HISTORY
Received April 2018
Revised November 2018

KEYWORDS
P-values; Reproducible research; Statistical inference; Scientific inference

THE AMERICAN STATISTICIAN
2019, VOL. 73, NO. S1, 20–25: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2018.1447512

Taylor & Francis

OPEN ACCESS

What Have We (Not) Learnt from Millions of Scientific Papers with *P* Values?

John P. A. Ioannidis

Departments of Medicine, of Health Research and Policy, of Biomedical Data Science, and of Statistics, Stanford University and Meta-Research Innovation Center at Stanford (METRICS), Stanford, CA

ABSTRACT

P values linked to null hypothesis significance testing (NHST) is the most widely (mis)used method of statistical inference. Empirical data suggest that across the biomedical literature (1990–2015), when abstracts use *P* values 96% of them have *P* values of 0.05 or less. The same percentage (96%) applies for full-text articles. Among 100 articles in PubMed, 55 report *P* values, while only 4 present confidence intervals for all the reported effect sizes, none use Bayesian methods and none use false-discovery rate. Over 25 years (1990–2015), use of *P* values in abstracts has doubled for all PubMed, and tripled for meta-analyses, while for some

ARTICLE HISTORY
Received November 2017
Revised February 2018

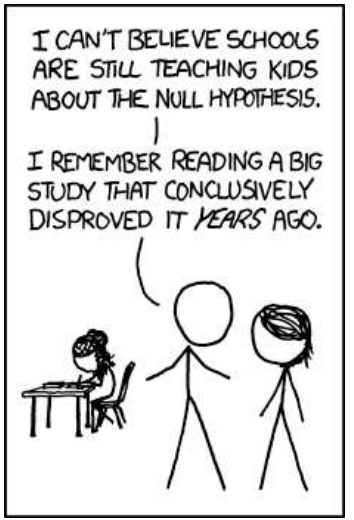
KEYWORDS
Bias; *P*-value; Statistical significance

Maybe the key issue is the training in statistics

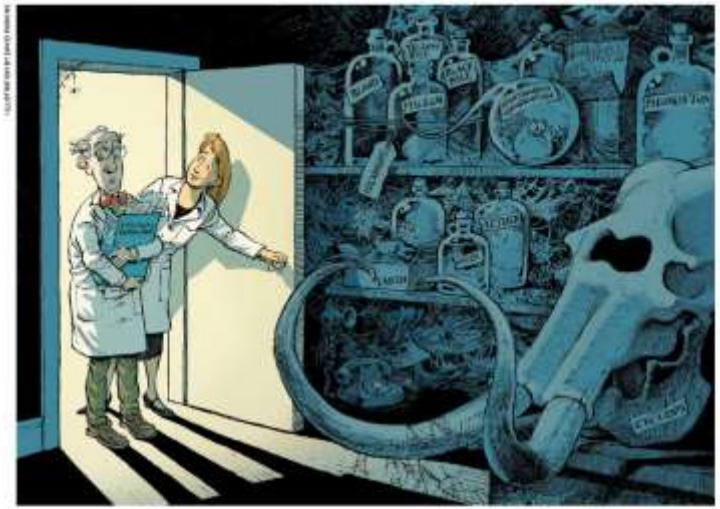
Why It Is Hard to Eliminate *P*-Values?

This brings us to the question of why eliminating P-value is so hard. The basic explanation is neither philosophical nor scientific, but sociologic; everyone uses them. It is the same reason we can use money. When everyone believes in something's value, we can use it for real things; money for food, and *P*-values for knowledge claims, publication, funding, and promotion. It does not matter if the P-value does not mean what people think it means; it becomes valuable because of what it buys.

S. Goodman: The American Statistician - 2019



Nature March 2019



Retire statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.

When was the last time you heard a seminar speaker claim there was 'no difference' between two groups because the difference was 'statistically non-significant'?

If your experience matches ours, there's a good chance that this happened at the last talk you attended. We hope that at least someone in the audience was perplexed if, as frequently happens, a plot or table showed that there actually was a difference.

How do statistics so often lead scientists to deny differences that those not educated in statistics can plainly see? For several generations, researchers have been warned that a statistically non-significant result does not 'prove' the null hypothesis (the hypothesis that there is no difference between groups or no effect of a treatment on some measured outcome)[1]. Nor do statistically significant results 'prove' some other hypothesis. Such misconceptions have famously warped the literature with overstated claims and, less famously, led to claims of conflicts between studies where none exists.

We have some proposals to keep scientists from falling prey to these misconceptions.

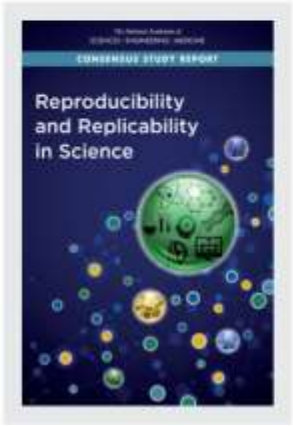
PERVASIVE PROBLEM

Let's be clear about what must stop: we should never conclude there is 'no difference' or 'no association' just because a *P* value is larger than a threshold such as 0.05 ▶

National Academies May 2019

The National Academies of SCIENCES ENGINEERING MEDICINE

THE NATIONAL ACADEMIES PRESS



Reproducibility and Replicability in Science (2019)

DETAILS

218 pages | 6 x 9 | PAPERBACK
ISBN 978-0-309-48616-3 | DOI 10.17226/25303

To « p » or not to « p »: what is the question?

The objective: is my product effective ?

How to make a decision ?

A

What is the probability of obtaining the observed data, if the product is not effective?

B

What is the probability that the product is effective, given the observed data?

Two different ways to make a decision based on

A

Pr(**observed data** | **product is not effective**)

- Better known as the **p-value** concept
- Used in the **null hypothesis** test (or decision)
- This is the likelihood of the data assuming an hypothetical explanation (eg the “null hypothesis”)
- **Classical statistics** perspective (Frequentist)

B

Pr(**product effective** | **observed data**)

- **Bayesian** perspective
- It is the probability of efficacy given the data

The Bayesian perspective allows to directly address the question of interest.

Nature 2017

POINTS OF SIGNIFICANCE

Interpreting *P* values

A *P* value measures a sample's compatibility with a hypothesis, not the truth of the hypothesis.

Although *P* values are convenient and popular summaries of experimental results, we can be led astray if we consider them as our only metric[1]. Even in the ideal case of a rigorously designed randomized study fit to a predetermined model, *P* values still need to be supplemented with other information to avoid misinterpretation.

A *P* value is a probability statement about the observed sample in the context of a hypothesis, not about the hypotheses being tested.

Nothing has changed in 20 years

1999

Toward Evidence-Based Medical Statistics. 1: The *P* Value Fallacy

Steven N. Goodman, MD, PhD

What is the question ?

New patient : **Cancer?**

© Can Stock Photo

Diagnostic test

Test Result

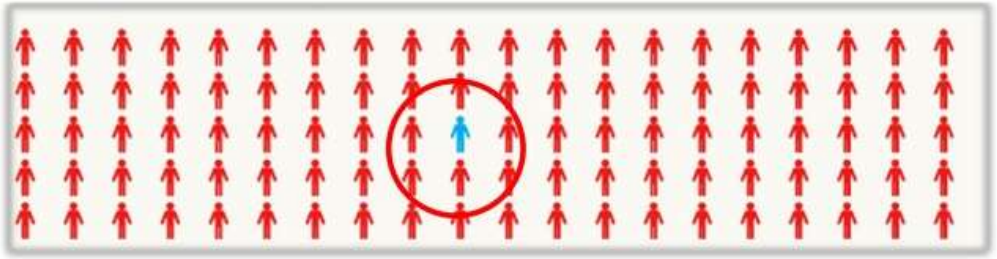
Interpretation ?

What is the probability that the patient has Cancer given the observed positive results ?

Context

A disease **D** with a low prevalence

1 % of the population is diseased = **D+**



Major consequences if the disease is not detected

A problem of decision making

The accuracy of a diagnostic test is assessed as follows:

- Sensitivity: $\Pr(\textbf{positive result} \mid \textbf{cancer})$
- Specificity: $\Pr(\textbf{negative result} \mid \textbf{no cancer})$

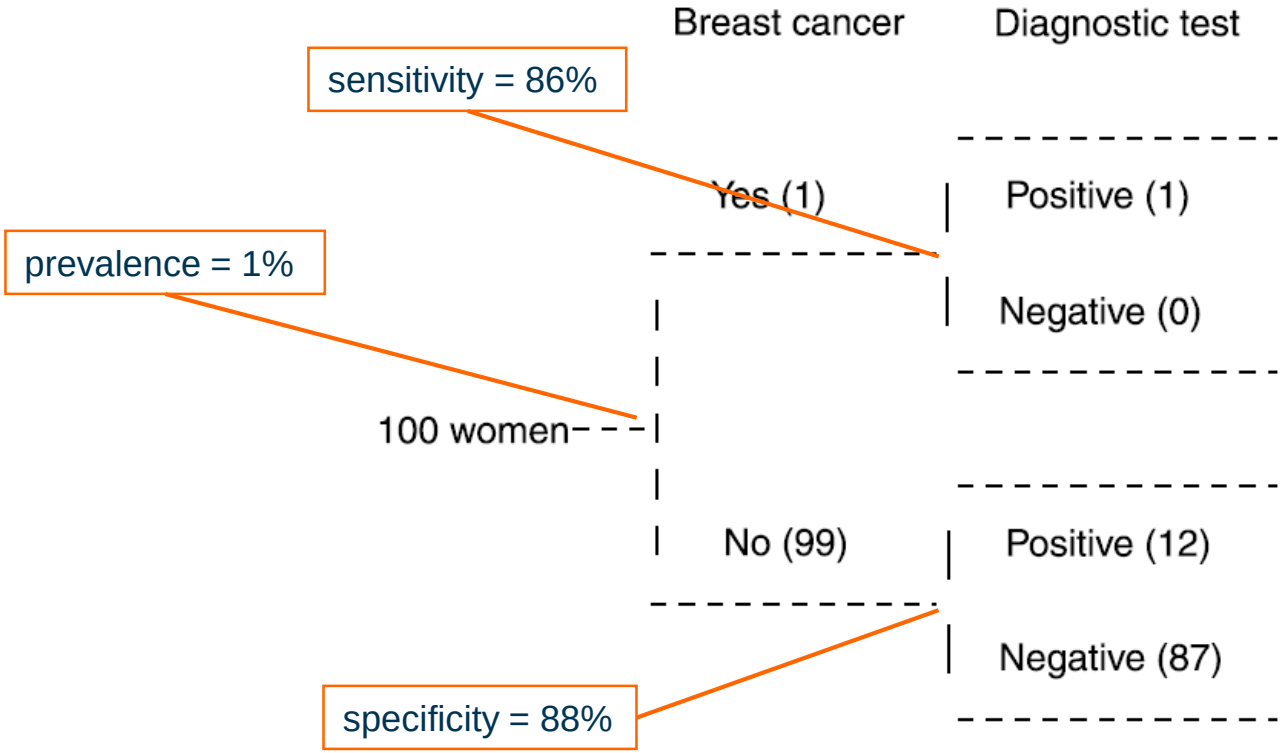
In practice:

Given that the diagnostic test result is positive, what is the probability you truly have cancer?

$$\Pr(\textbf{cancer} \mid \textbf{positive result}) = ?$$

Example

$$\Pr(\text{cancer} \mid \text{positive result}) = \frac{1}{12+1} = 0.077$$



How can that be so low?
The small proportion of errors for the large majority of women who do not have breast cancer swamps the large proportion of correct diagnoses for the few women who have it.

The probability of interest dependents on the underlying prevalence of the disease.

Agresti, A. (2007). *An Introduction to Categorical Data Analysis.* Wiley, 2nd ed.
Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of *p*-values. *R. Soc. Open sci.* **1**(3): 140216

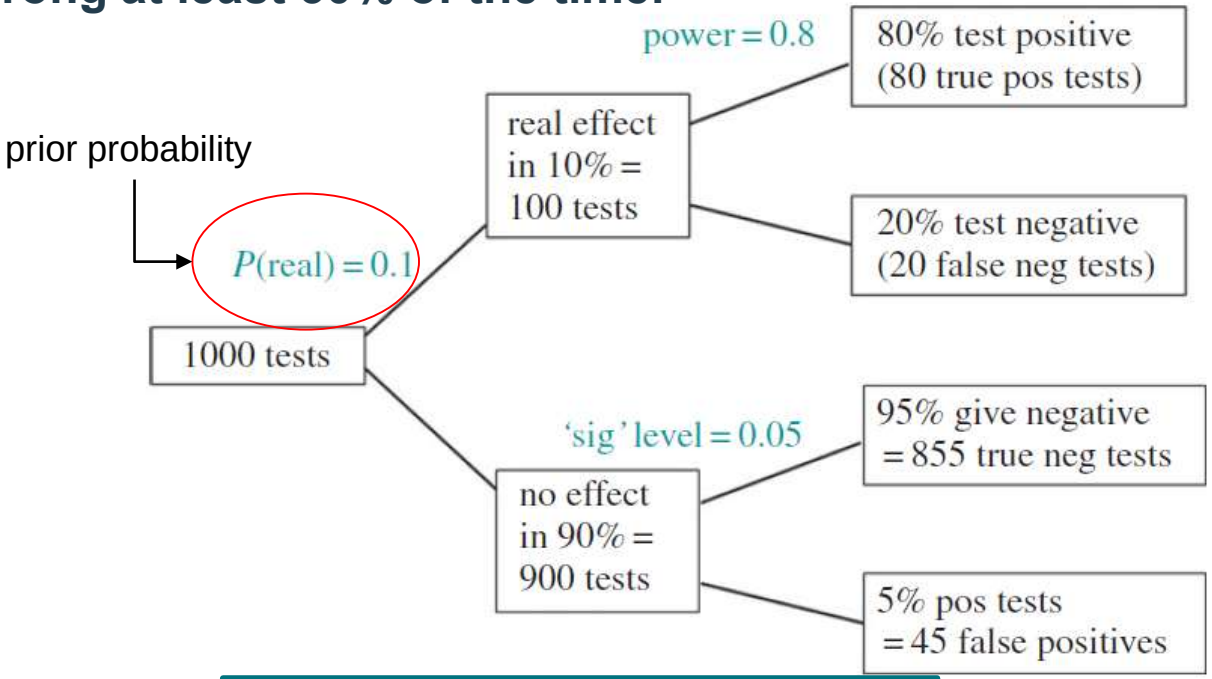
The clinical trial analogy

effective? → clinical trial → data

$$\Pr(\textbf{drug effective} \mid \textbf{data}) = ?$$

depends largely on **prior probability** that there is a real effect

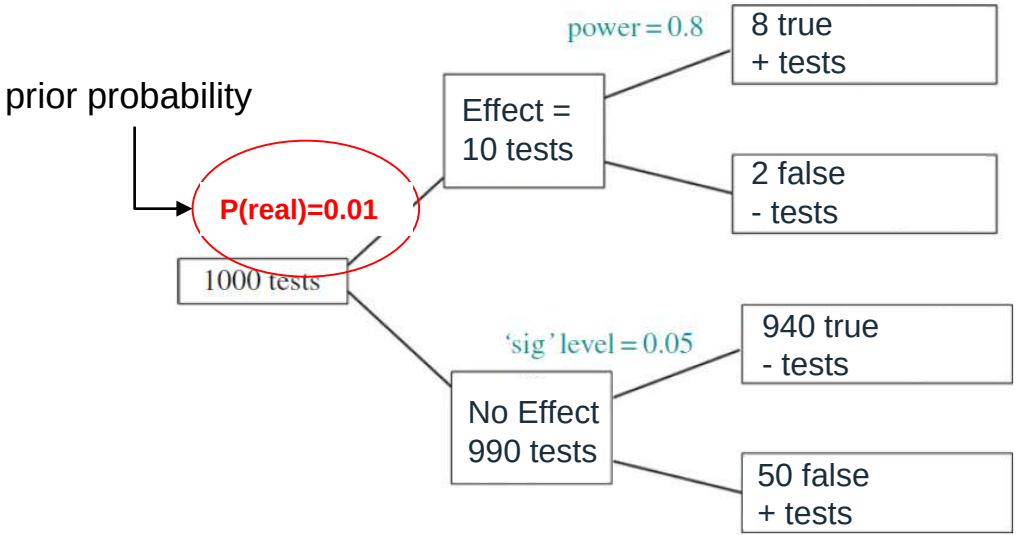
“If you use p = 0.05 to suggest that you have made a discovery, you will be wrong at least 30% of the time.”



$$\Pr(\textbf{real effect} \mid \mathbf{p} < 0.05) = \frac{80}{80 + 45} = 0.64$$

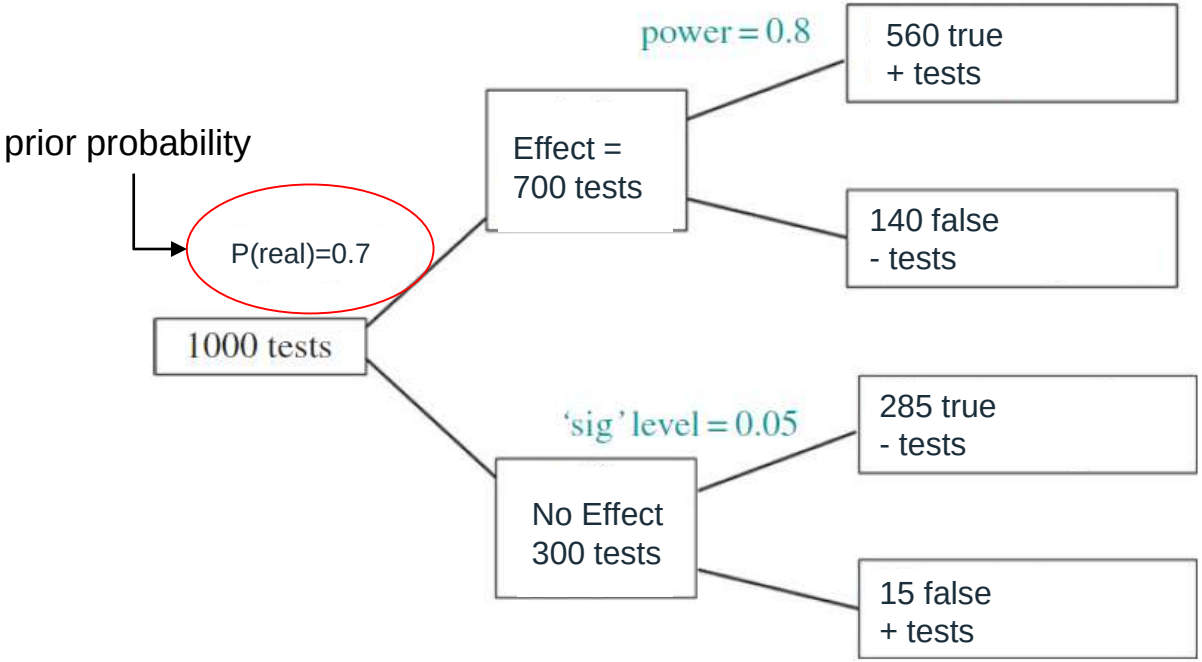
Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of *p*-values. *R. Soc. Open sci.* **1**(3): 140216.

"If you use p = 0.05 "….when you are in early discovery



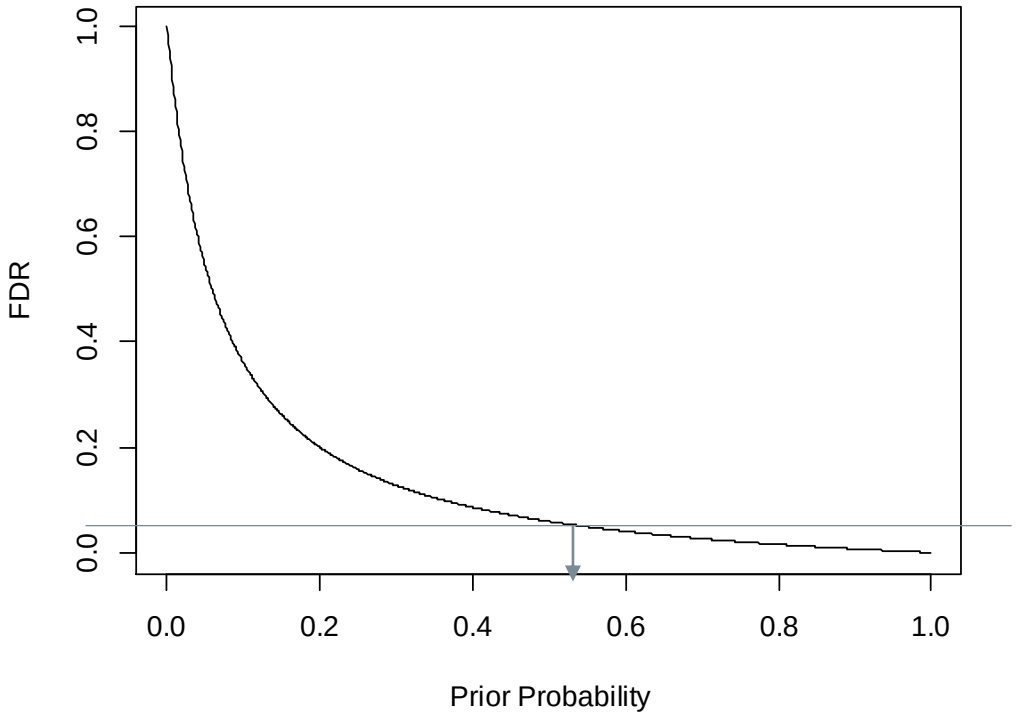
$$\Pr(\textbf{real effect} \mid \mathbf{p} < 0.05) = \frac{8}{8+50} = \mathbf{0.14}\ !!!!$$

…. if the prior is good

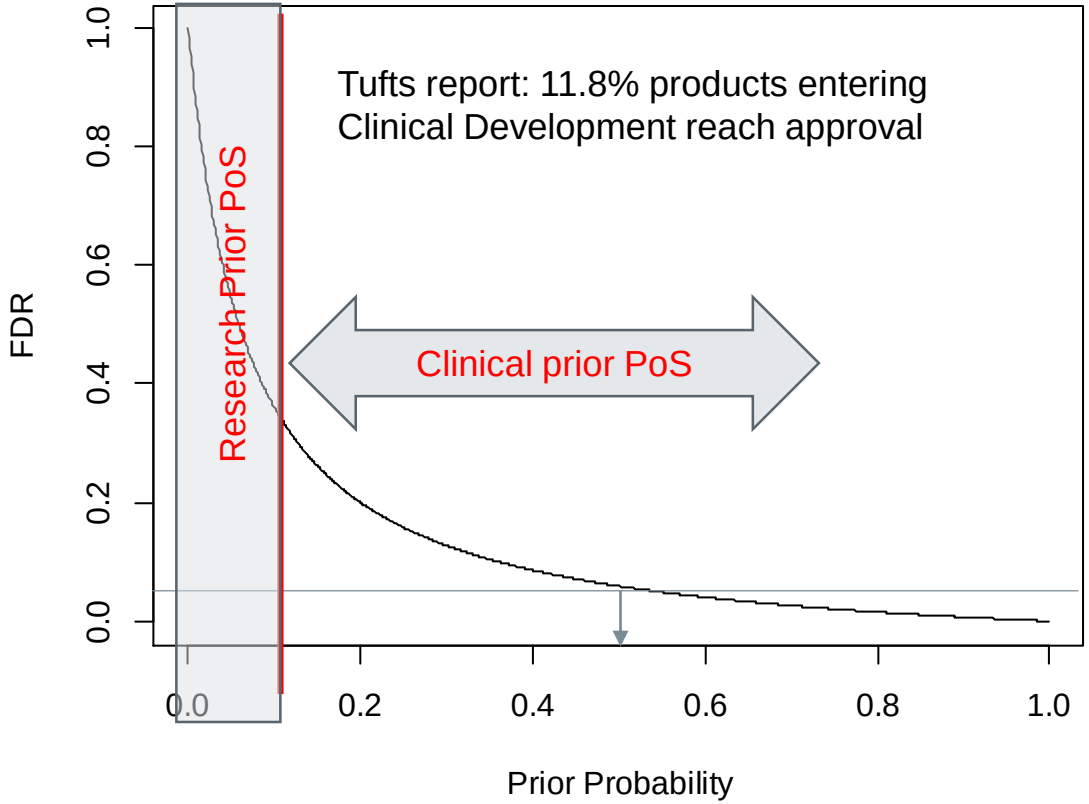


$$\Pr(\textbf{real effect} \mid \mathbf{p} < 0.05) = \frac{560}{560 + 15} = \mathbf{0.97}$$

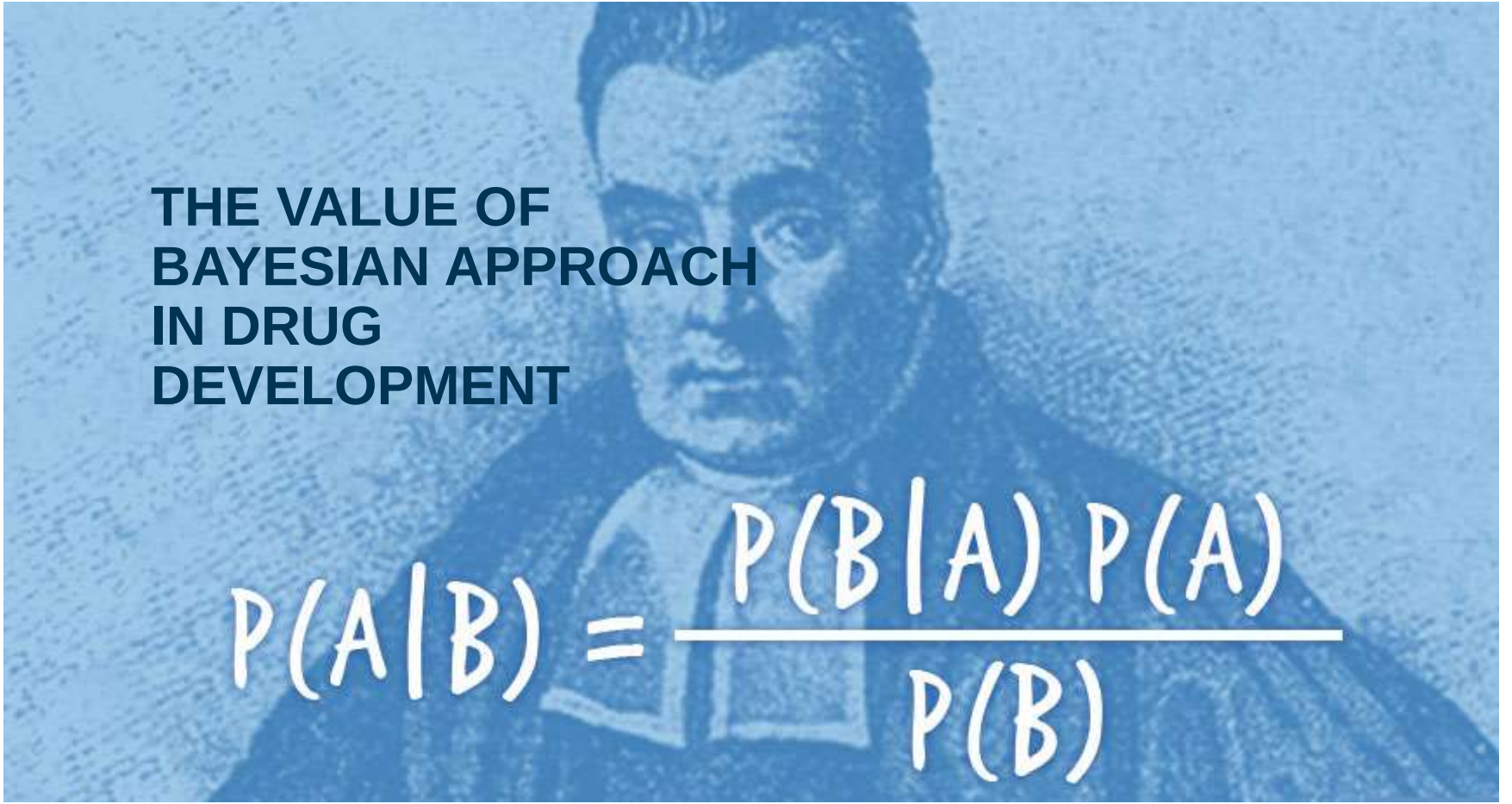
False “Discovery" Rate for p<0.05, power=0.8 as function of Prior Probability



False “Discovery” Rate for p<0.05, power=0.8 as function of Prior Probability

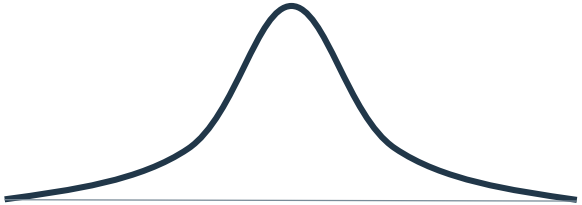


THE VALUE OF BAYESIAN APPROACH IN DRUG DEVELOPMENT



Bayesian inference is the mechanism used to update the state of knowledge

prior information
$p(\theta)$

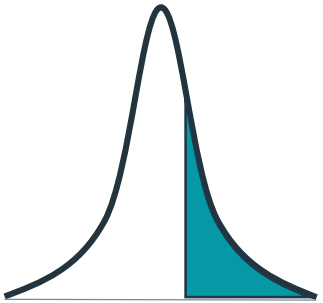


X

data information
$p(\text{data}|\theta)$

$\propto$

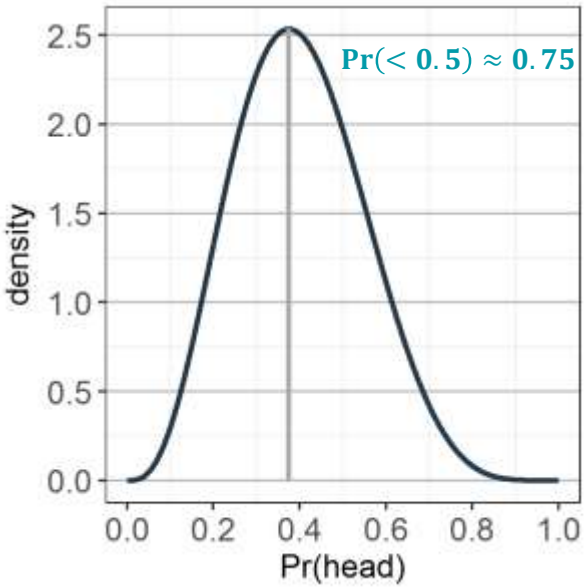
posterior information
$p(\theta|\text{data})$



The process to arrive at a posterior distribution makes use of Bayes' formula.

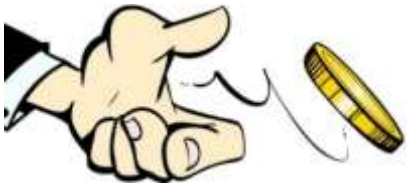
The coin flipping experiment

prior information



X

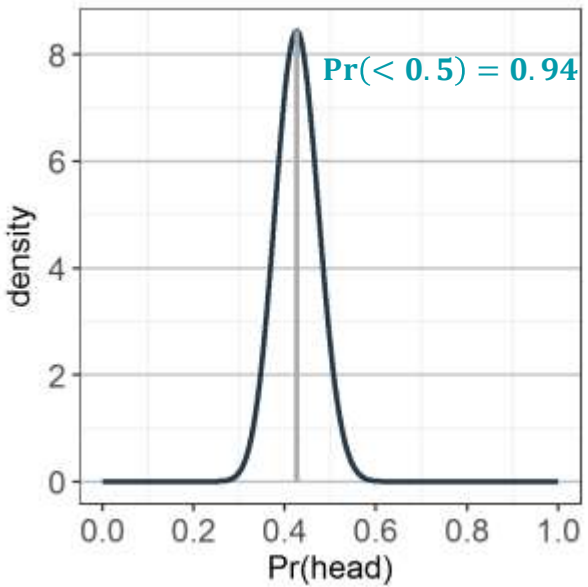
data information



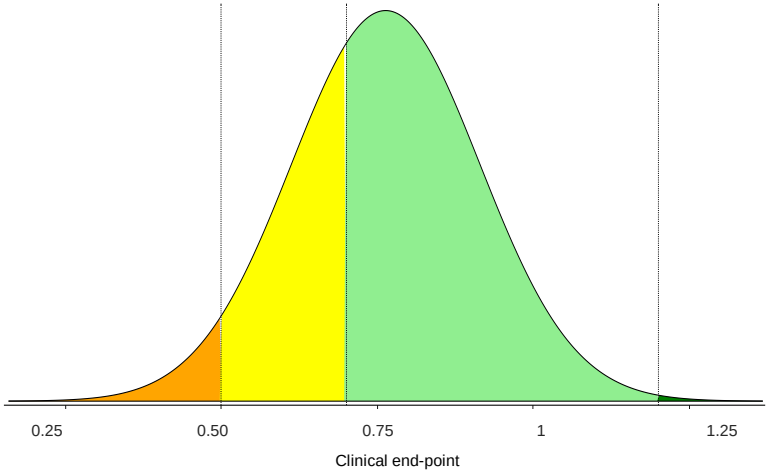
43 heads in 100 flips

∝

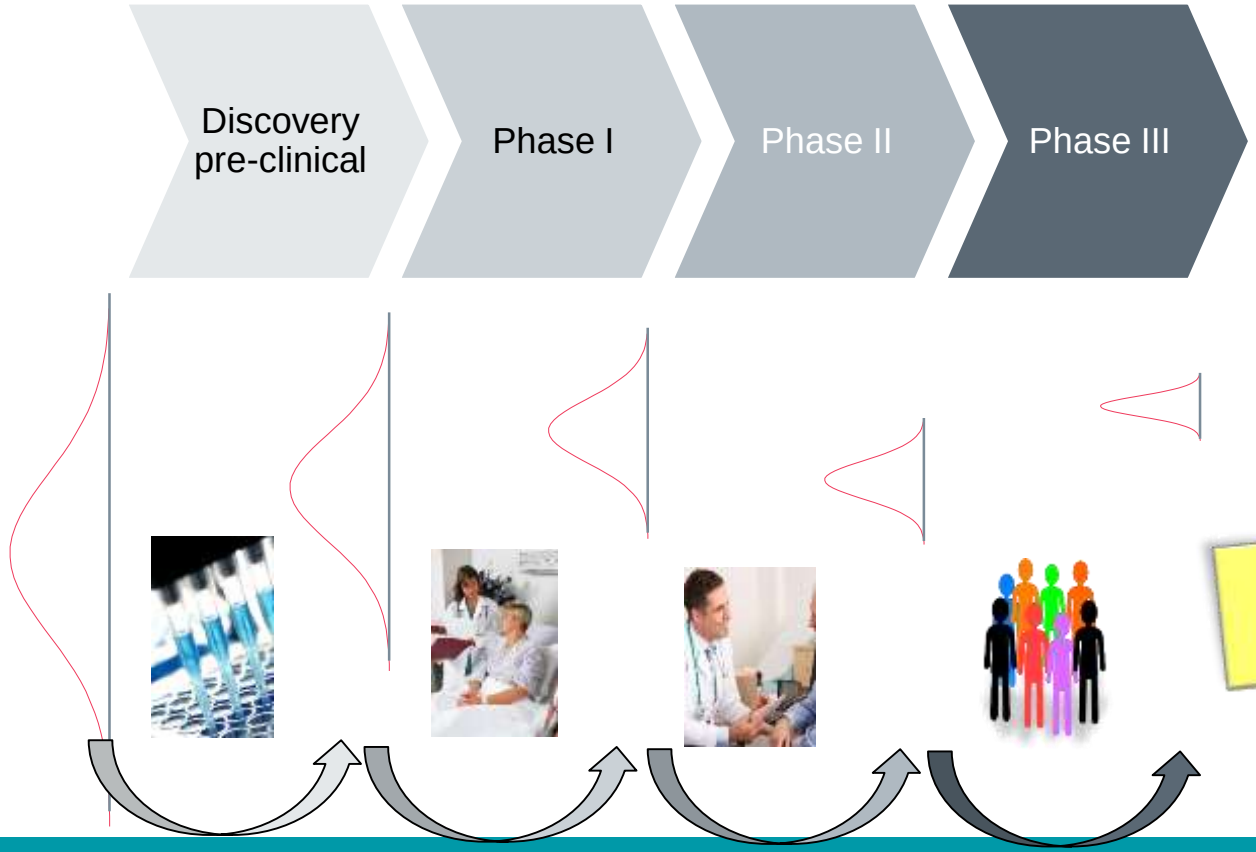
posterior information



Decision rules based on Posterior Probability



Drug development is a learning process



We now have computing power to apply Bayesian statistics

Regulatory point of view

- 2010 & 2016 Guidance for medical device clinical trials

Guidance

for Industry and FDA Staff

Guidance for the Use of

Bayesian Statistics in

Medical Device Clinical Trials

Document issued on: February 5, 2010

Leveraging Existing Clinical Data for Extrapolation to Pediatric Uses of Medical Devices

Guidance for Industry and Food and Drug Administration Staff

Document issued on June 21, 2016.
This document will be in effect as of September 19, 2016.

The draft of this document was issued on May 6, 2015.

For questions regarding this document, contact Jacqueline Francis (CDRH) at (301) 796-6405 (Jacqueline.Francis@fda.hhs.gov), CDRHPediatricExtrapolation@fda.hhs.gov, or the Office of Communication, Outreach, and Development (CBER) at 800-835-4709 or 240-402-8010.

FDA CID initiative

U.S. FOOD & DRUG ADMINISTRATION

Bayesian Applications

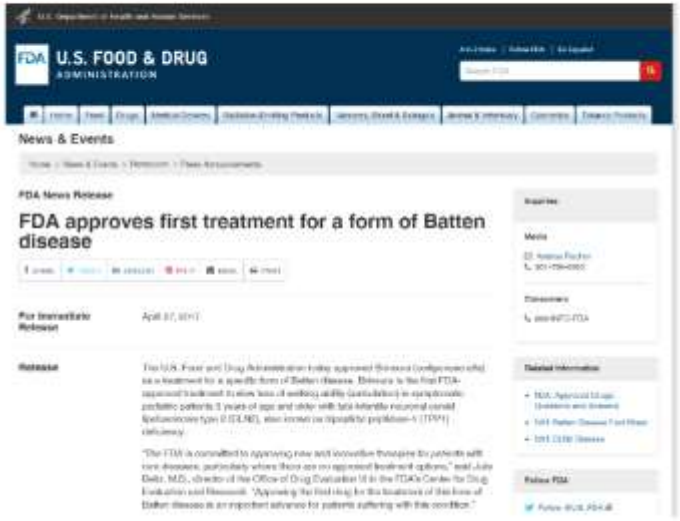
2018

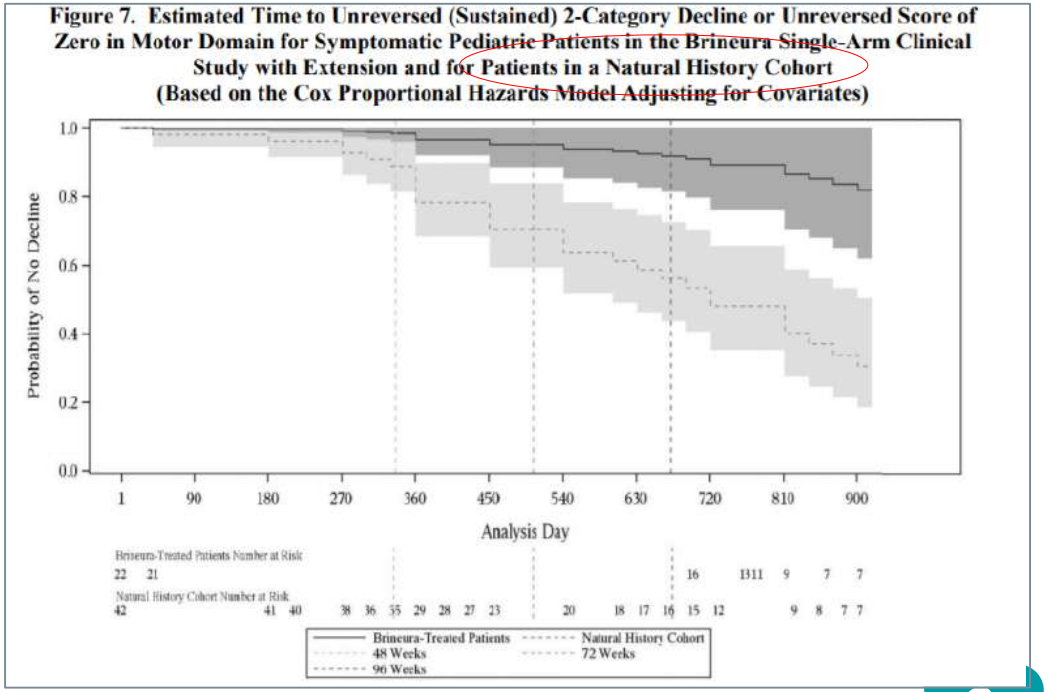
- Safety monitoring
  - Large CV risk studies that leverage control patient data from other sources via Bayesian adaptive designs
- Oncology
  - Early phase dose-finding trial designs, e.g., CRM
  - Bayesian adaptive trials that use intermediate or accelerated approval endpoints for decision-making
- Rare diseases
  - Incorporate prior information from early phase trials
  - Use information about disease progression in analytical model
  - Compute shrinkage estimators of effects in rare subsets of disease
  - Incorporate prior information from adult trials to improve efficiency of pediatric trials

Historical control

2018

Rare Disease Example





Part II: Power, Bayesian power and Assurance

Power vs assurance

independent samples t-test ($H_0: \mu_1 = \mu_2$ vs $H_1: \mu_1 \neq \mu_2$)

frequentist approach (power)

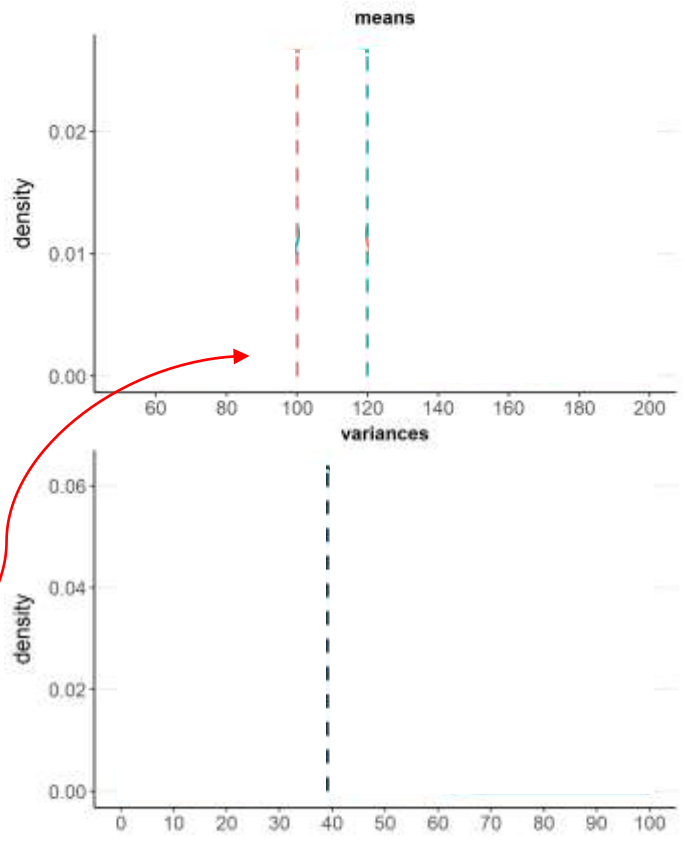
- A power calculation takes a particular value of the effect within the range of possible values given by H_1 and poses the question: if this particular value happens to obtain, what is the probability of coming to the correct conclusion that there is a difference?

assumptions:

- $\mu_1 = 100;$
- $\mu_2 = 120;$
- $\sigma_1^2 = \sigma_2^2 = 39$

very strong priors!

assumptions:



Power vs assurance

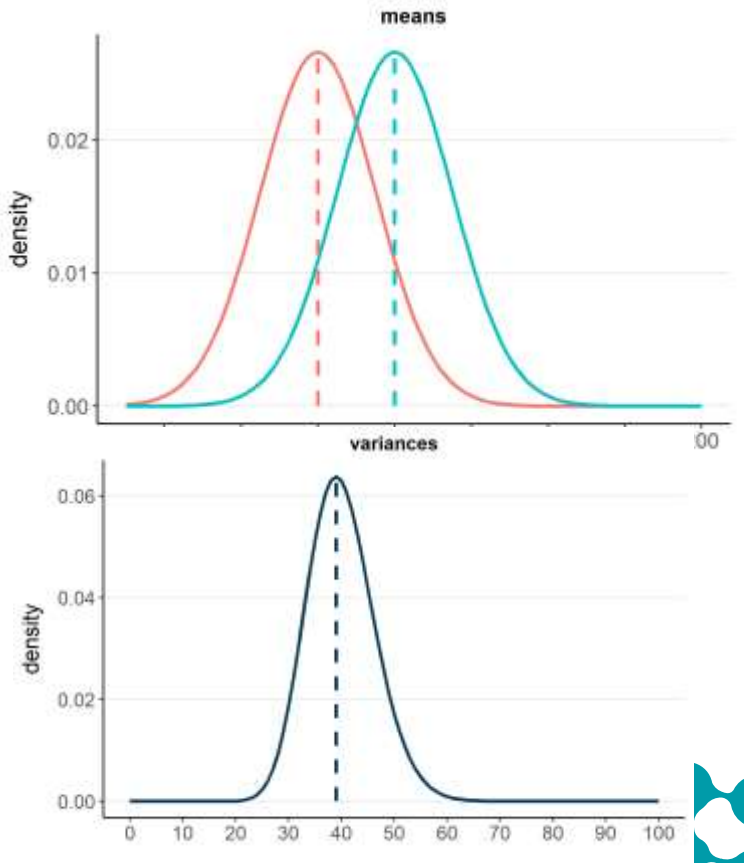
independent samples t-test ($H_0: \mu_1 = \mu_2$ vs $H_1: \mu_1 \neq \mu_2$)

bayesian approach (assurance)

- In order to reflect the uncertainty, a large number of effect sizes, i.e. $(\mu_1 - \mu_2)/\sigma_{\text{pooled}}$, are generated using the prior distributions.
- A power curve is obtained for each effect size
- the expected (weighted by prior beliefs) power curve is calculated

- *Note1: Given those priors, using the Frequentist power approach, the Probability of Success of the trial is 50%*
- *Note2: about 50% of Phase III trials are failing because of lack of efficacy (S. Wang, FDA, 2008)*

assumptions:



(Frequentist) Power

- Let R denote the rejection of the null hypothesis, the power is, assuming parameter values of $\boldsymbol{\theta} = \boldsymbol{\theta}^*$

$$\pi(\boldsymbol{\theta}^*, n) \coloneqq \Pr(R|\boldsymbol{\theta}^*, \boldsymbol{n})$$

- It is a conditional probability. It is conditional on the parameters of the model, e.g. the “true effect size” in a frequentist test and the sample size.

Assurance

- *“Assurance is the **unconditional** probability that a trial will lead to a specific outcome”*

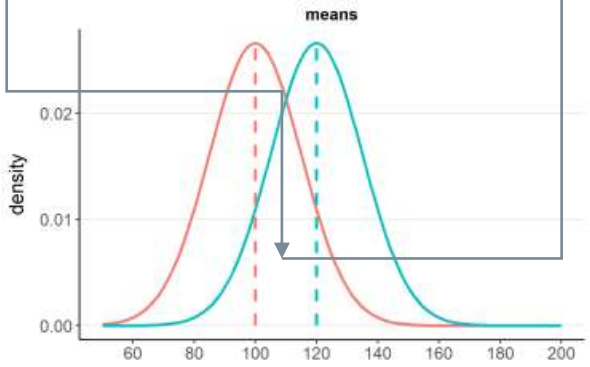
$$\gamma(n) \coloneqq \int \pi(\theta, n) f(\theta) d\theta$$
$$\gamma(n) \coloneqq \Pr(R) = E_\theta[\pi(\boldsymbol{\theta})]$$

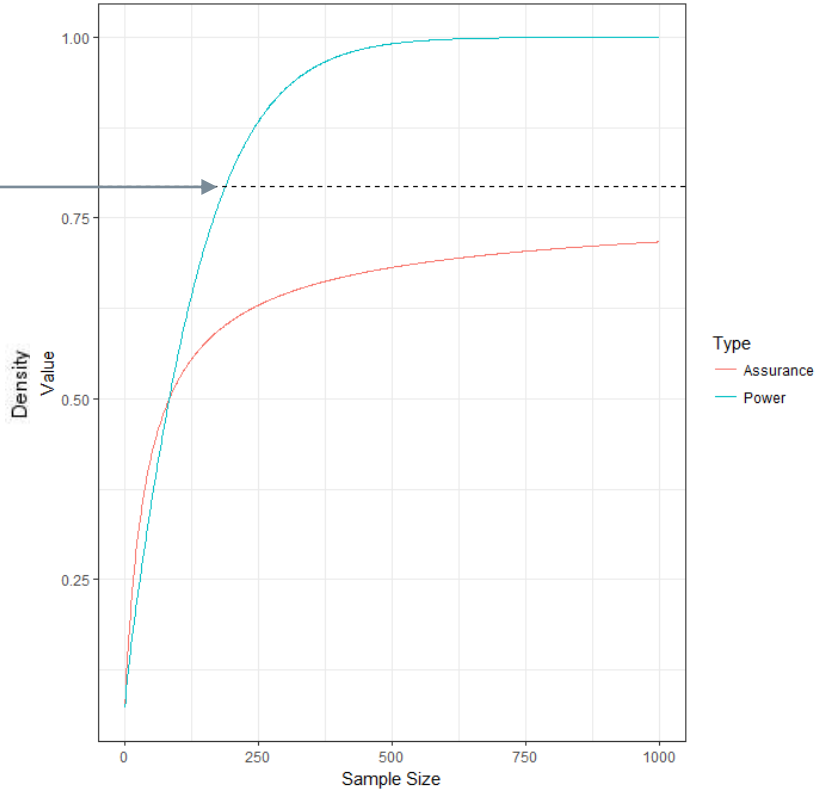
It is thus also a function of n (and eventually other nuisance parameters)

The assurance is the expected power over all possible values of theta (-> over its prior distribution…)

An example: Power vs Assurance

In this example the **assurance** converges to 0.793, that is the prior probability that the new drug is indeed superior

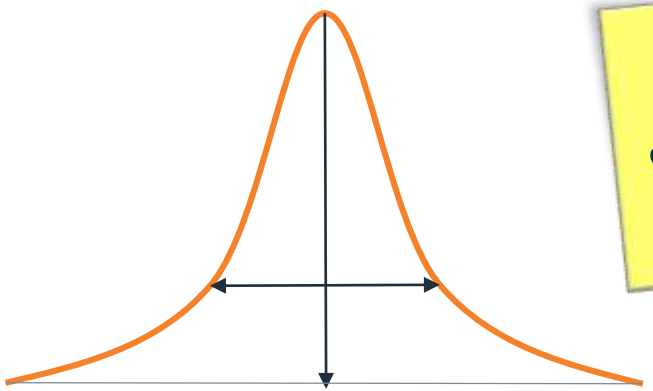




Difference Simulations/Predictions

Simulations

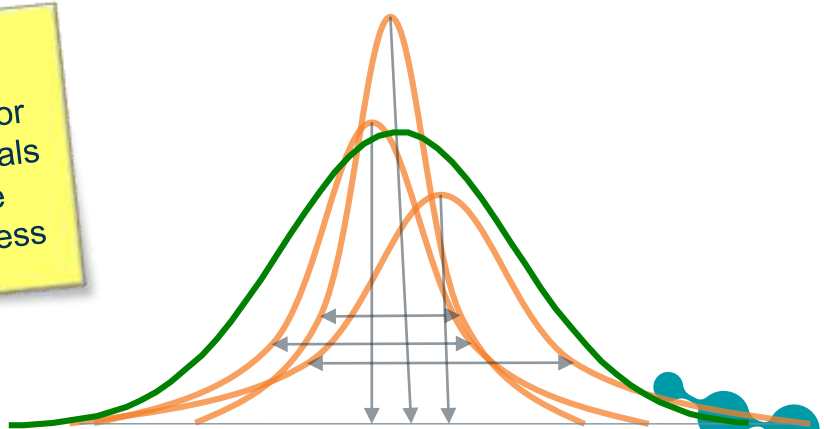
the “new observations” are drawn from distribution “centered” on estimated location and dispersion parameters (treated as “true values”).



The predictive distribution is key for optimizing future trials and evaluate the probability of success

Predictions

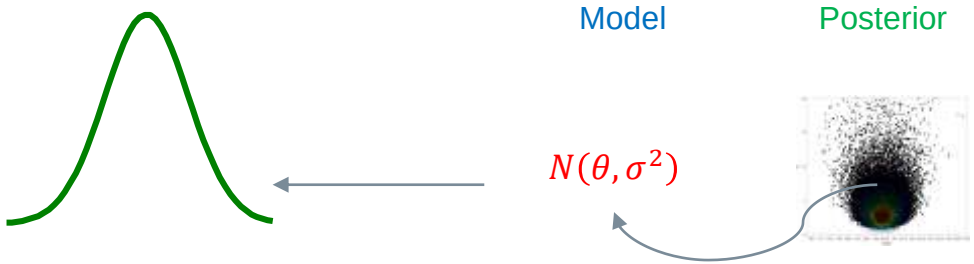
the uncertainty of parameter estimates (location and dispersion) is taken into account before drawing “new observations” from relevant distribution



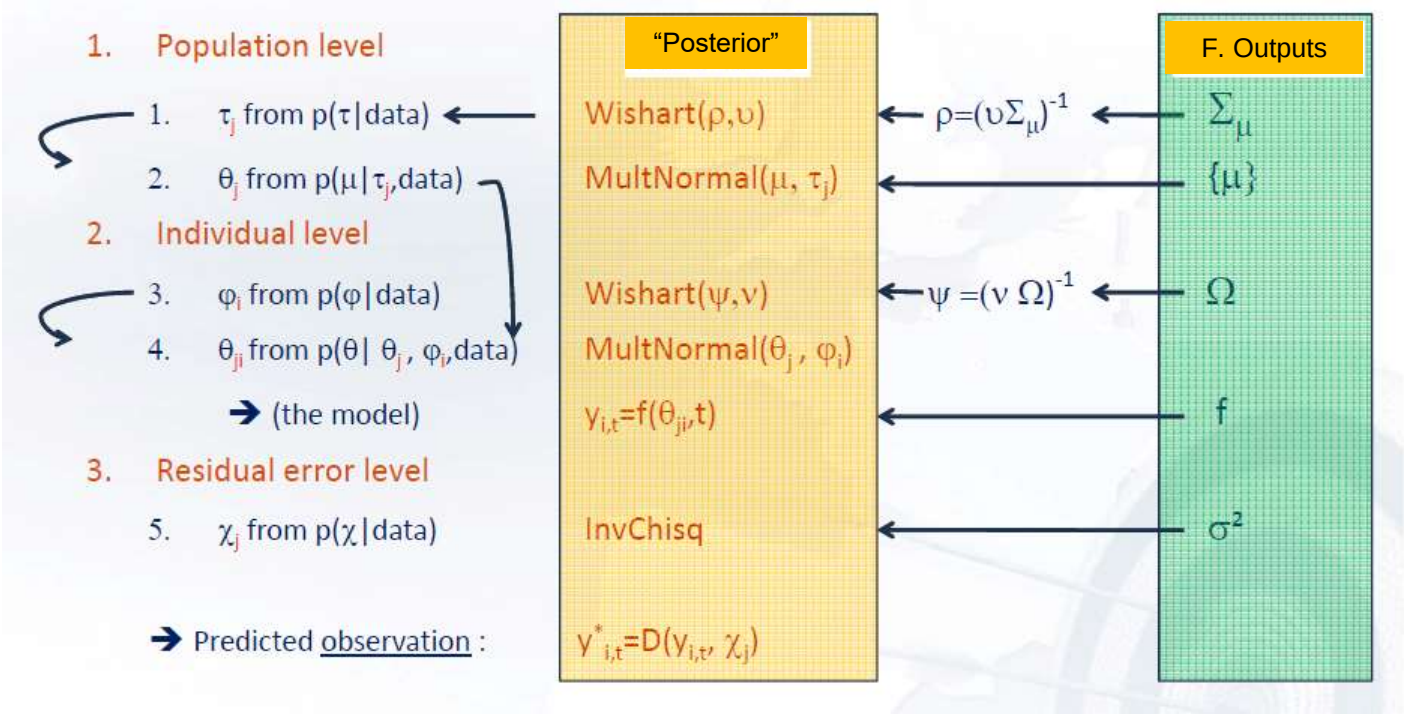
Predictions

- Given the model and the posterior distribution of its parameters, what are the plausible values for a future observation $\tilde{y}$?

$$p(\tilde{y}|data) = \int p(\tilde{y}|\theta)\, p(\theta|data)\, d\theta$$

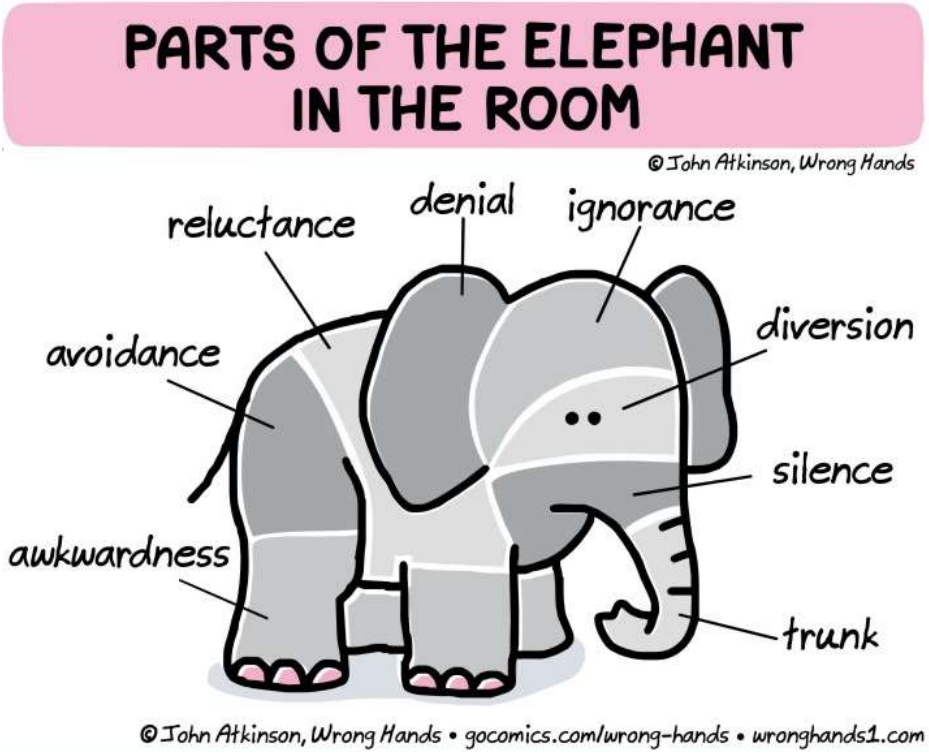


Note: It's easy to approximate the predictive distribution from Frequentist outputs

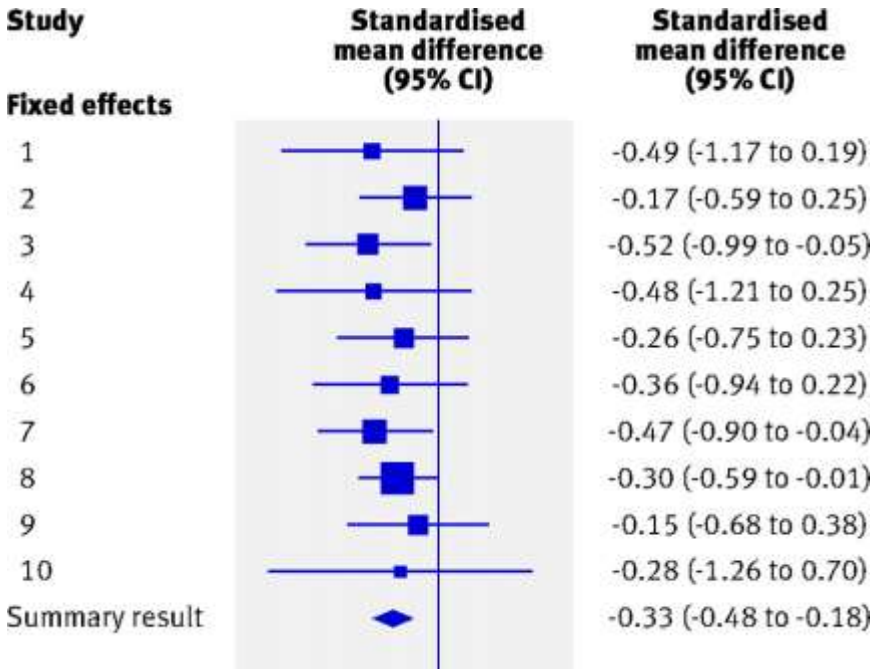


Part III: The missing component

The elephant in the room? The study-to-study variability

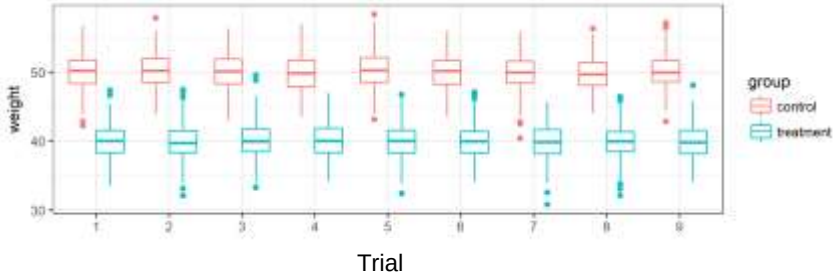


You know this: Meta-analysis showing study-to-study differences

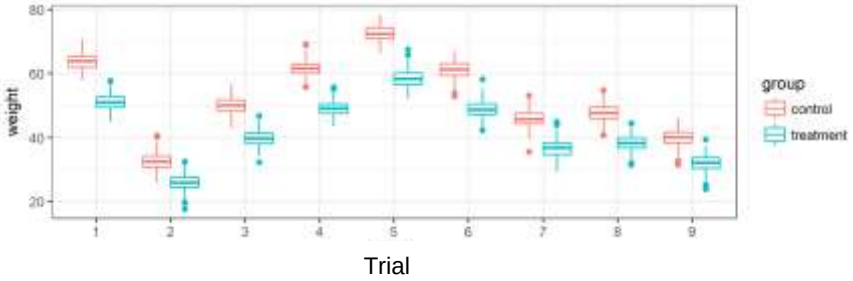


Different scenarios may happen

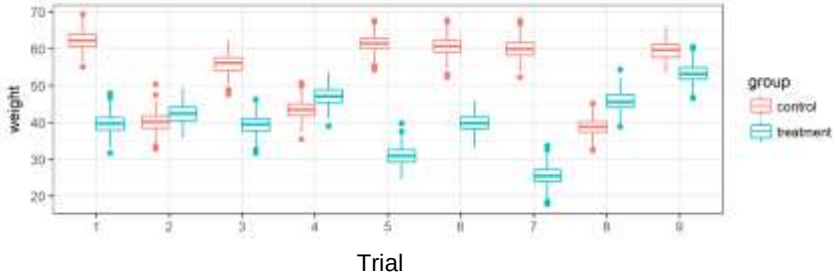
No variability



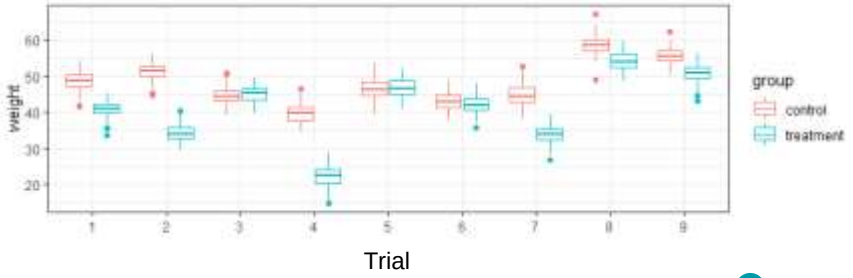
Groups vary together (ρ=1)



Groups vary independently (ρ=0)

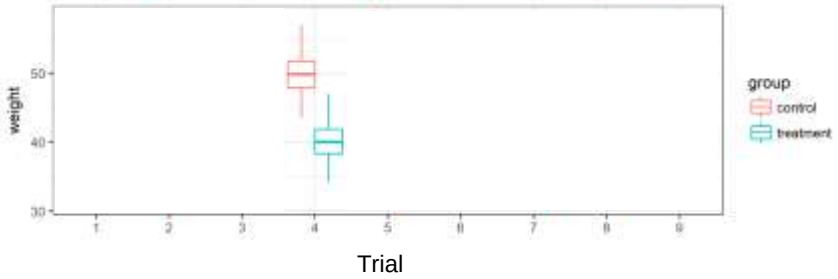


Groups vary with some dependencies (ρ~0.5)

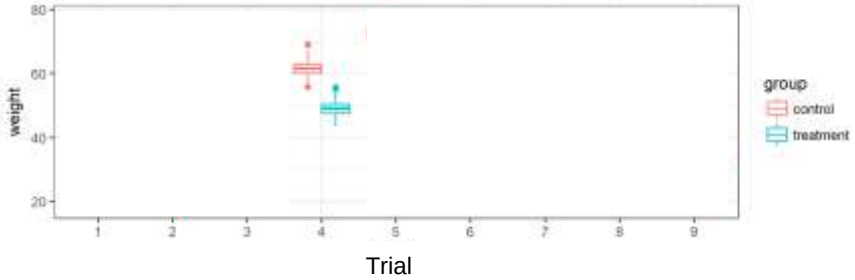


If you do one trial you may get one of those outcomes….

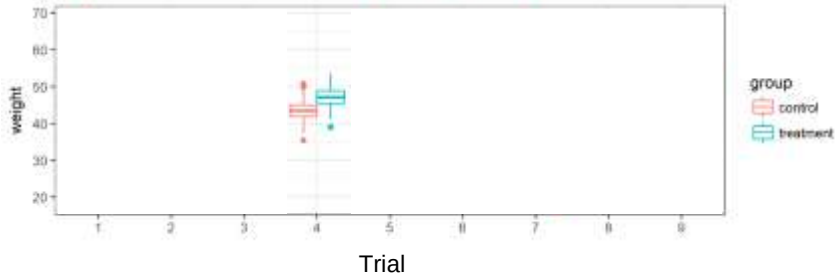
No variability



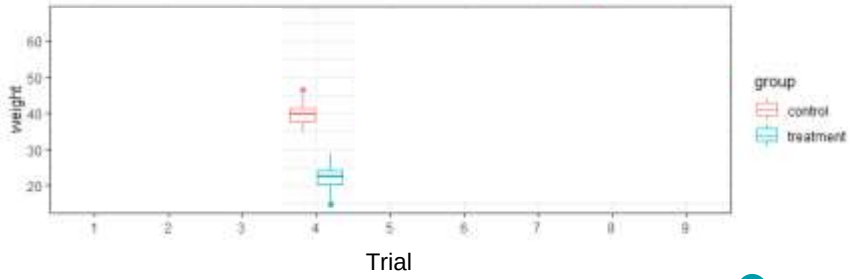
Groups vary together



Groups vary independently

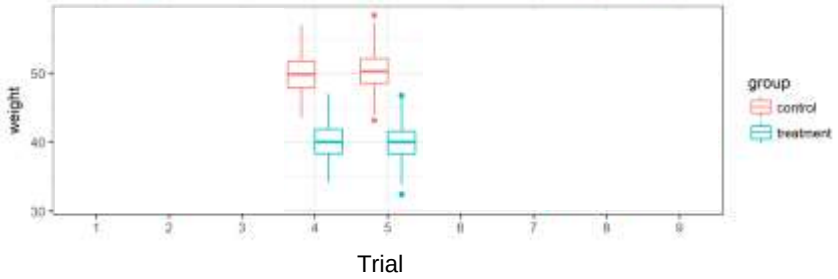


Groups vary with some dependencies

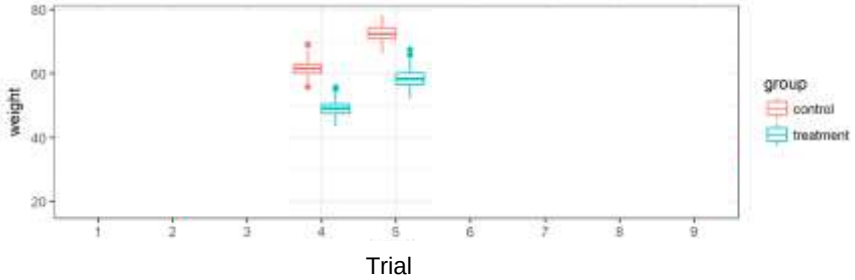


If you do two trials you may get one of those outcomes….

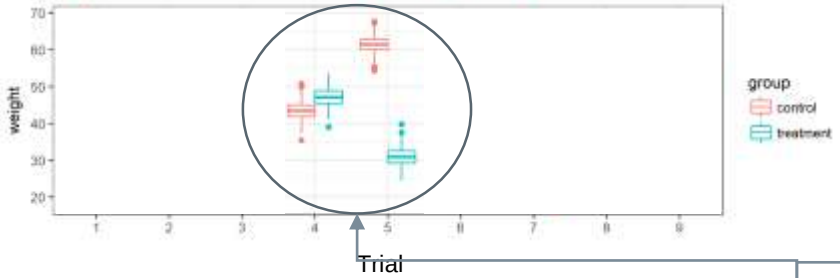
No variability



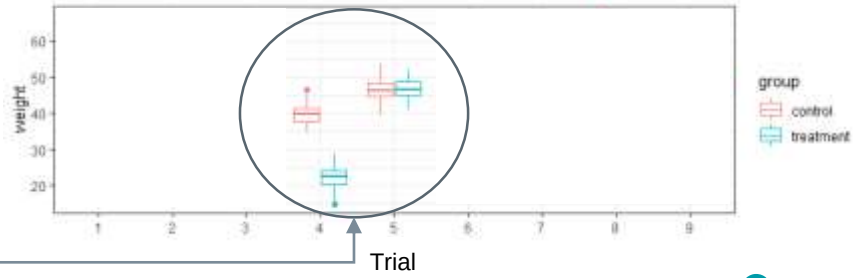
Groups vary together



Groups vary independently



Groups vary with some dependencies

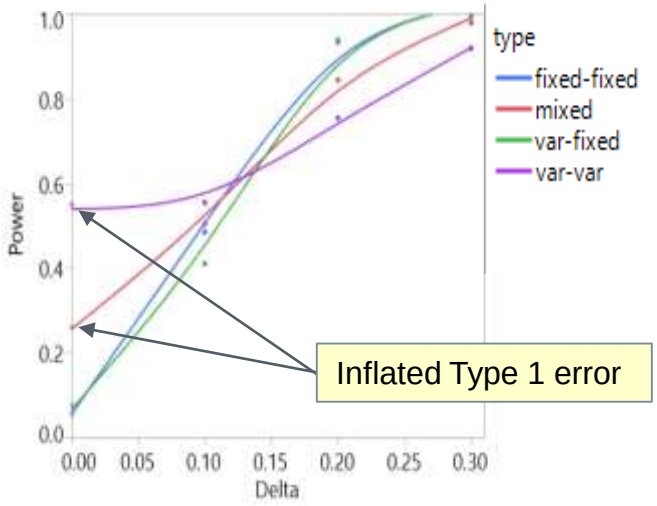


Does this new treatment works ?

Impact of study-to-study variability

- Assumptions are made that there is no study-to-study variability.
- Everyone know there is such variability but this is ignored in design, power calculation, evaluation, …..
- This variance component is fundamental
- It is related to the “replicability” issue, achieving a conclusion regardless of the study

- If ignored and existing:
  - then there is a major risk of type I error-inflation!
  - the estimates are biased since confounded with study effect
  - It violates fundamental DoE practices: maximize D-optimality, ie sources of variability in studies



Study “formats”: example in pre-clinical pharmacology

Classic (Common practice):

Model: $Y_{ij} = \mu + t_i + \varepsilon_{ij}$

Optimal designs:

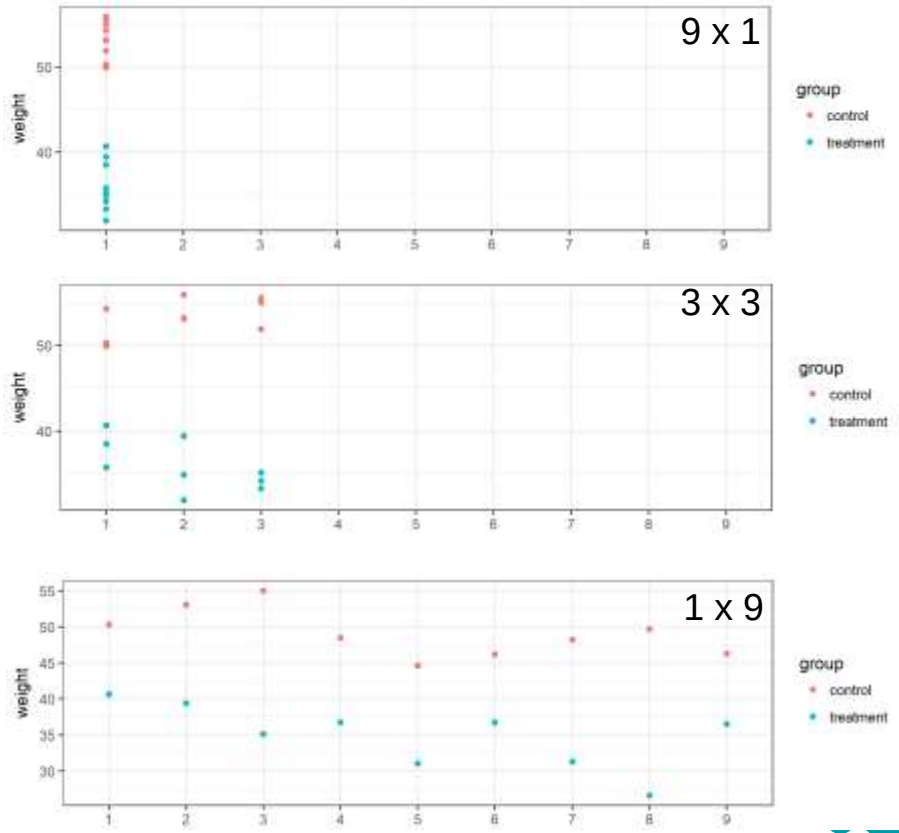
- Intermediate design

Model: $Y_{ijk} = \mu + t_i + r_j + \varepsilon_{ijk}$

  - $r_j \sim N(0, {\sigma_{study}}^2)$, random effect due to the jth study, Same for both groups

- Extreme design

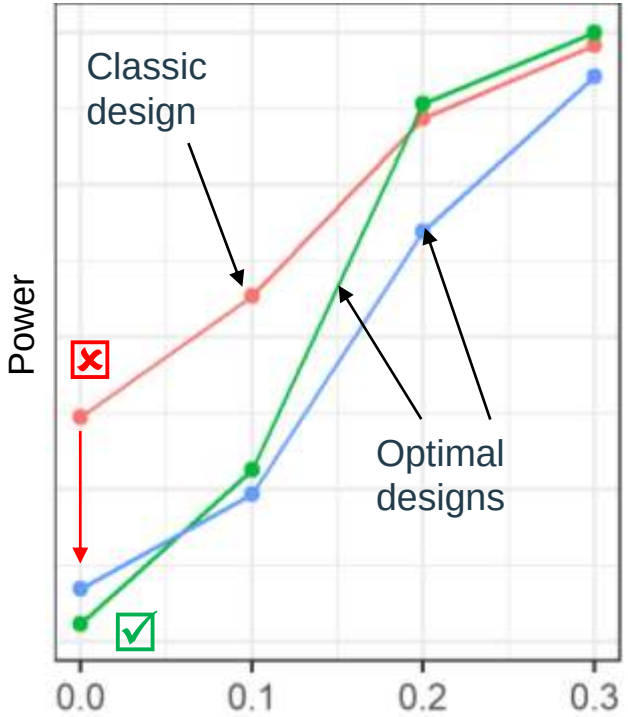
Model: $Y_{ij} = \mu + t_i + \varepsilon_{ij}$



Performance comparison of the tested designs (“formats”)

- Current approach -> ‘’classic” design (all in one study)
- Convention (USP <1032>)
  - “Convert **bias** into lack of precision of the estimate”
  - Control precision with sample size
- Concept of study “format”:

  N(total) = R (runs) x r (replicates)



Optimal designs allow to control for Type I error in all cases

Improving precision of measurements by adding noise sources

- Assume that:
  - θ is the parameter of interest
  - you can perform R studies of r patients
- The variance of θ is: $V(\theta) = \frac{\sigma^2_{Study}}{R} + \frac{\sigma^2_r}{R \times r}$
- Currently most consider that:

$$V(\theta) = \frac{\sigma_r^2}{1 \times r}$$

But in reality it is:

$$V(\theta) = \frac{\sigma^2_{Study}}{1} + \frac{\sigma_r^2}{1 \times r}$$

- How to design trials / allocate patients to have best precision of θ ?

$$\sigma^2_{\text{Study}} \ll \sigma^2_{\text{r}}$$

1 trial, 10 patients

$$\frac{\sigma^2_{Study}}{R} + \frac{\sigma^2_r}{R \times r} = \frac{3}{1} + \frac{10}{1 \times 10} = 4$$

2 trials, 5 patients/trial

$$\frac{\sigma^2_{Study}}{R} + \frac{\sigma^2_r}{R \times r} = \frac{3}{2} + \frac{10}{2 \times 5} = 2.5$$

$$\sigma^2_{\text{Study}} \gg \sigma^2_{\text{r}}$$

1 trial, 10 patients

$$\frac{\sigma^2_{Study}}{R} + \frac{\sigma^2_r}{R \times r} = \frac{10}{1} + \frac{3}{1 \times 10} = 10.3$$

2 trials, 5 patients/trial

$$\frac{\sigma^2_{Study}}{R} + \frac{\sigma^2_r}{R \times r} = \frac{10}{2} + \frac{3}{2 \times 5} = 5.3$$

Conclusions: To reduce partially the issue of Replicability:

1. **What is the question ?**
2. Consider the study-to-study variance component in designing and sizing trials
   - Available via literature and control groups used in many trials
3. Consider the uncertainty of parameters estimates
4. Use prior distributions to compute the Assurance instead of the Power
   - Focus on probability of success of the trial beyond the power
5. Use Bayesian thinking and practices all the way through
   - This is a easy way to carry on the uncertainty
   - This this available now
   - This is the answer to most of your **questions: Pr(drug is effective | data)**

20-22 SEPTEMBER 2020

BETHESDA NORTH MARRIOTT HOTEL & CONFERENCE CENTER, ROCKVILLE, MARYLAND, USA

September 20 Short course on

Bayesian Complex Innovative Designs: a path for regulatory acceptance

Telba Irony	FDA	Scott Berry	Berry Consultants
John Scott	FDA	Roger Lewis	UCLA and Berry Consultants
Dionne Price	FDA	Karen Price	Lilly

The Adolphe Quetelet Society

ISBA
INTERNATIONAL SOCIETY FOR BAYESIAN ANALYSIS

Contacts

Bruno Boulanger
Chief Scientific Officer
bruno.boulanger@pharmalex.com

Phone +32 10 461010
Mobile +32 476 813936

5 Rue Edouard Belin
Mont-saint-Guibert
Belgium

Thank you !

Follow us on social media

/pharmalex-gmbh

@PharmalexGlobal