PHARMALEX
CONFIDENCE BEYOND COMPLIANCE

Using Bayesian methods to improve reproducibility in preclinical research

Bruno Boulanger
Clément Laloux
October 2022

NCS 2022

Agenda

- The reproducibility crisis (the replicability crisis ?)
- What is the question (in discovery & preclinical research) ?
- The design
  - The power and the Assurance
  - The ignored components
    - Lessons learned from bioassay development and validation
- Conclusions

Nature, 2014

STATISTICAL ERRORS

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

BY REGINA NUZZO

ASA, 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

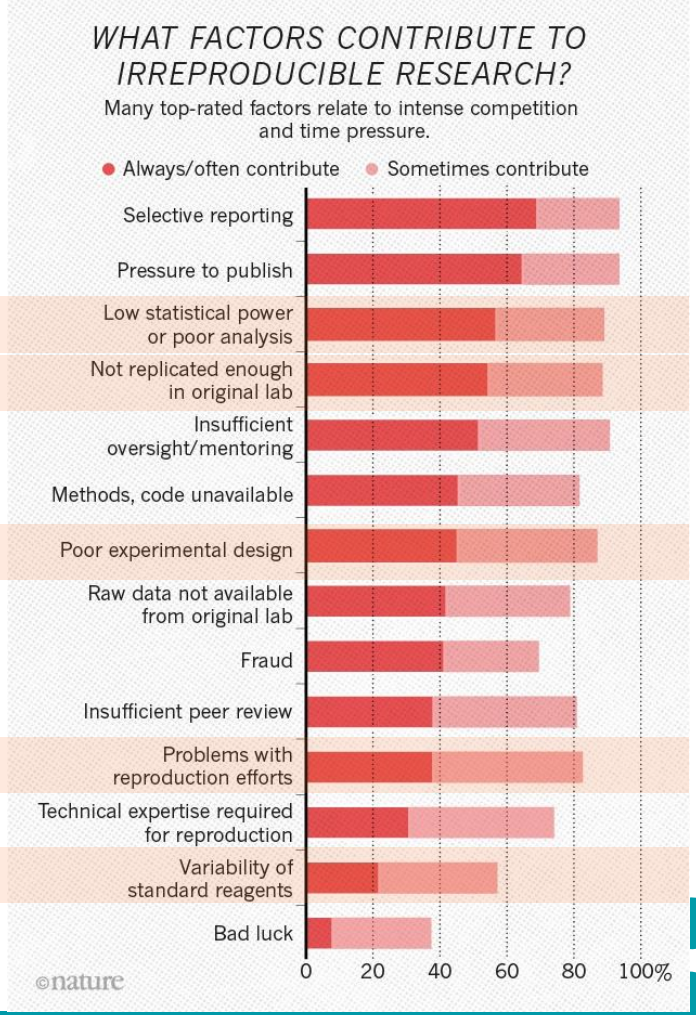
Nature, 2016

Statistics & DoE

Statistics & DoE

Statistics & DoE

Statistics & DoE



Nature, 2016

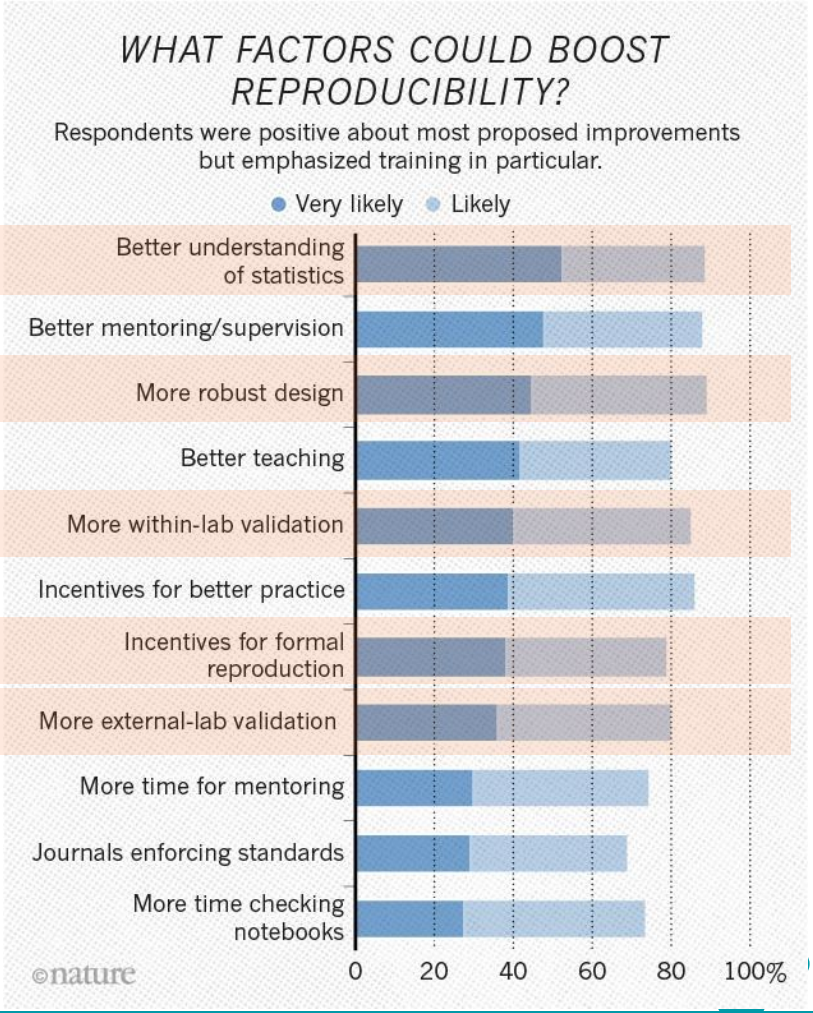
Statistics & DoE

Statistics & DoE

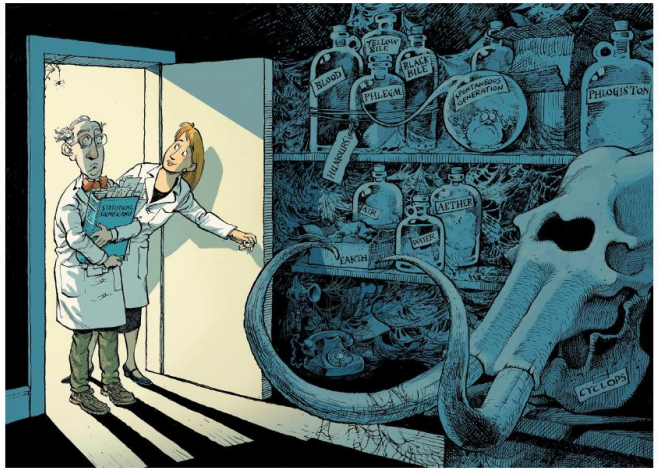
Statistics & DoE

Statistics & DoE

Statistics & DoE



Nature, 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.

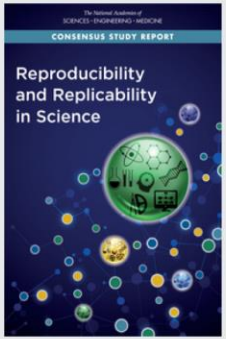
We have some proposals to keep scientists from falling prey to these misconceptions.

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 ▶

The National Academies Press, 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

The question: 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?

Currently 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”)

This essentially depends on the very question of interest

B

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

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

A problem of decision making

The accuracy of a diagnostic test is assessed as follows:

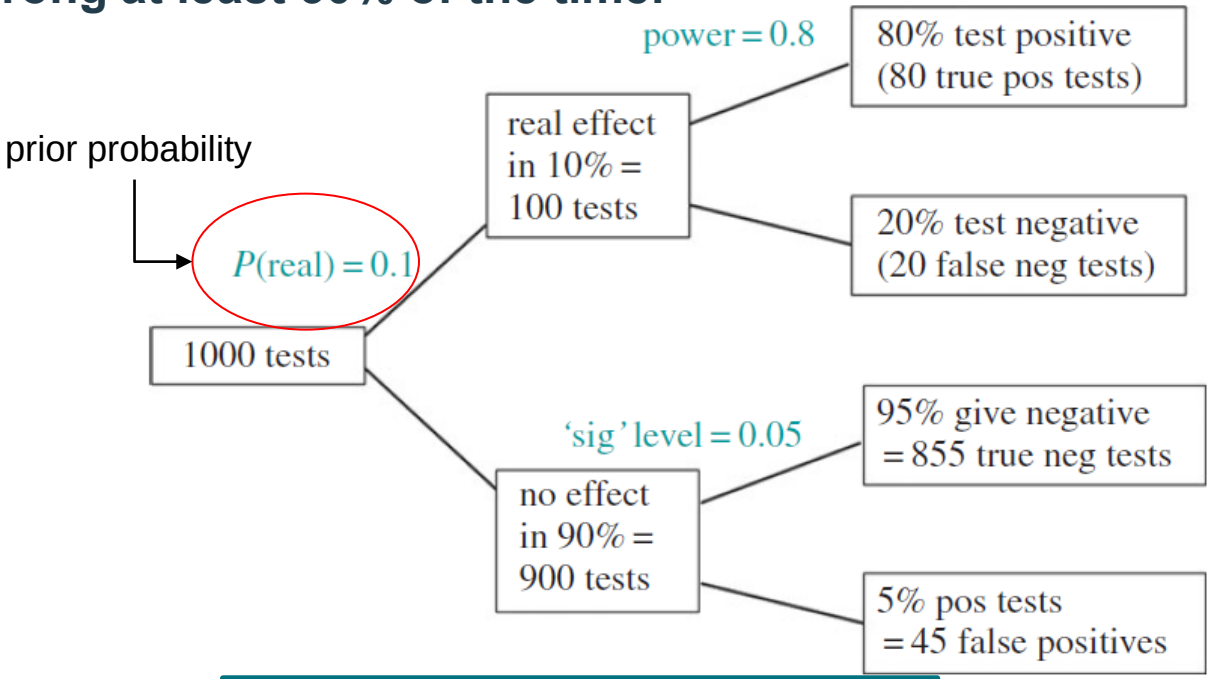
- Sensitivity: $\Pr(\textbf{positive result} \mid \textbf{cancer})$
- Specificity: $\Pr(\textbf{negative result} \mid \textbf{no cancer})$

In practice:

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

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

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

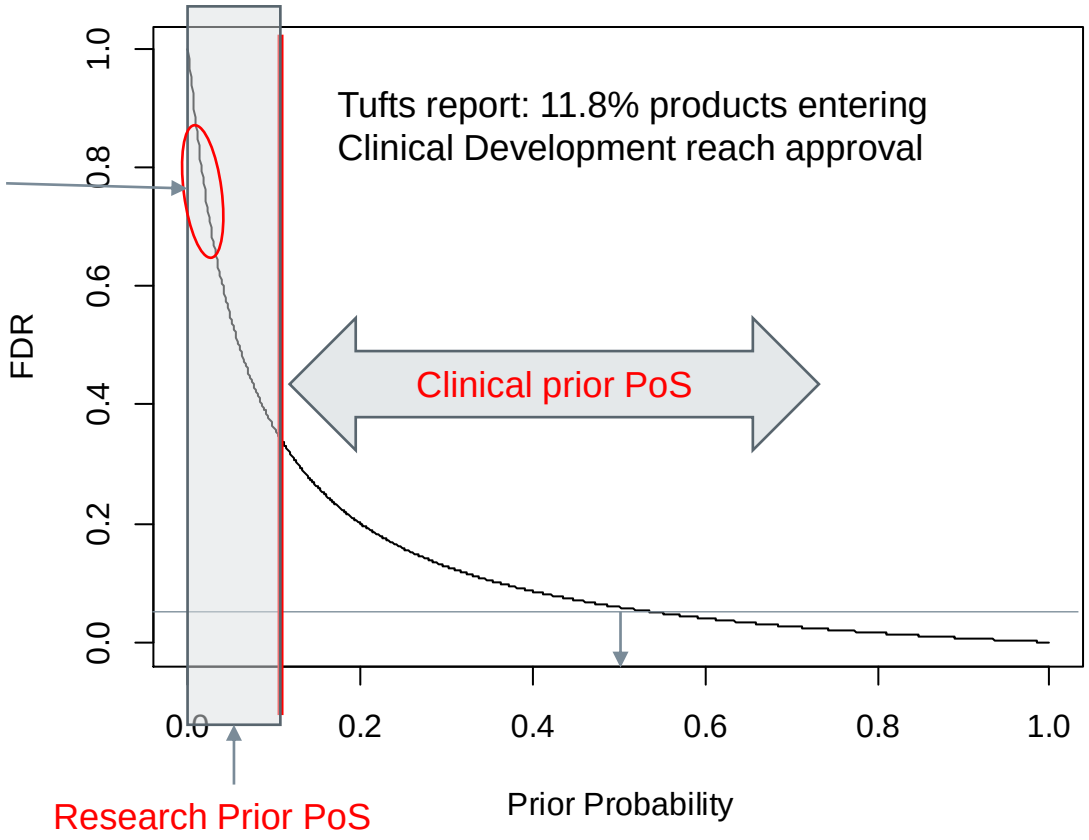


$$\Pr(\textbf{real effect} \mid \mathbf{p} < 0.05) = \frac{80}{80 + 45} = 0.64$$

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

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

Published rate of irreproducible studies



A lesson from bioassay and diagnostic world

- Pr(**Good results**| **Good "Known" product**) = Sensitivity
- Pr(**Bad results** |**Bad "Known" product**) = Specificity

Validation assay

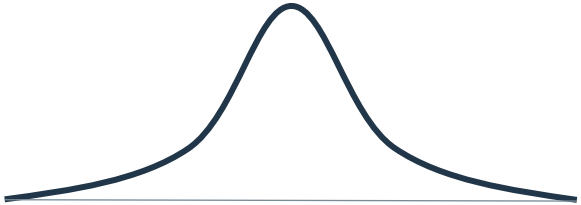
Use of assay

- Pr(**Good Unknown product**| **Good results**) = Positive Predictive Value (PPV)
- Pr(**Bad Unknown product** |**Bad Results**) = Negative Predictive Value (NPV)

$$PPV = P(D^+ \mid T^+)$$
$$= \frac{P(T^+|D^+) \times P(D^+)}{P(T^+)}$$
$$= \frac{SE \times PR}{SE \times PR + (1 - SP) \times (1 - PR)}$$

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

prior information
$p(\theta)$

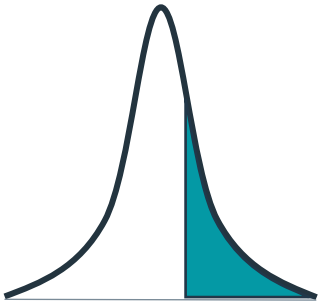


X

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

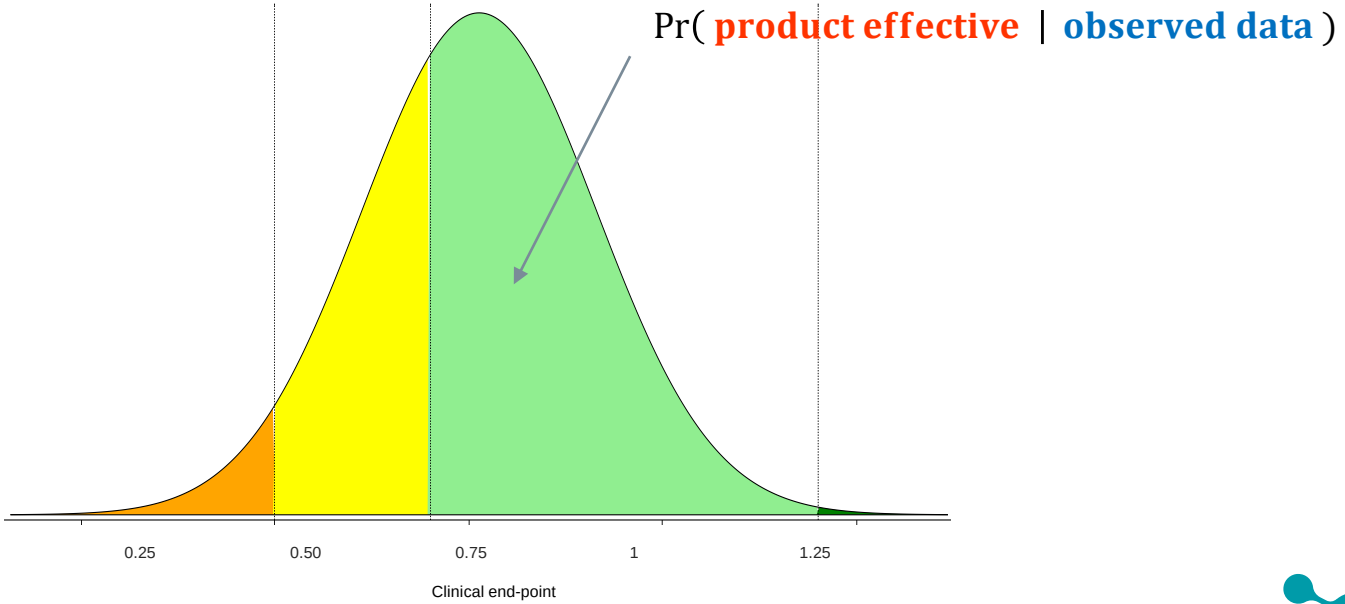
$\propto$

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



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

Decision rules based on Posterior Probability



Direct answer to the question

Design I: 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)

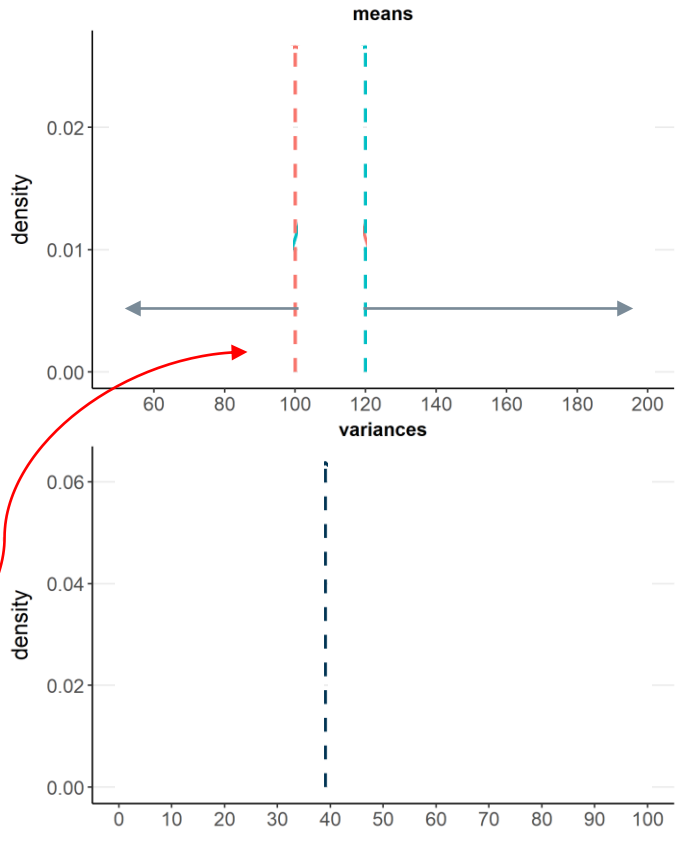
- A power calculation takes a particular value of the effect within the range of possible values given by H_1 and poses the question: if this particular value happens to obtain, what is the probability of coming to the correct conclusion that there is a difference?

assumptions:

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

very strong priors!

assumptions:



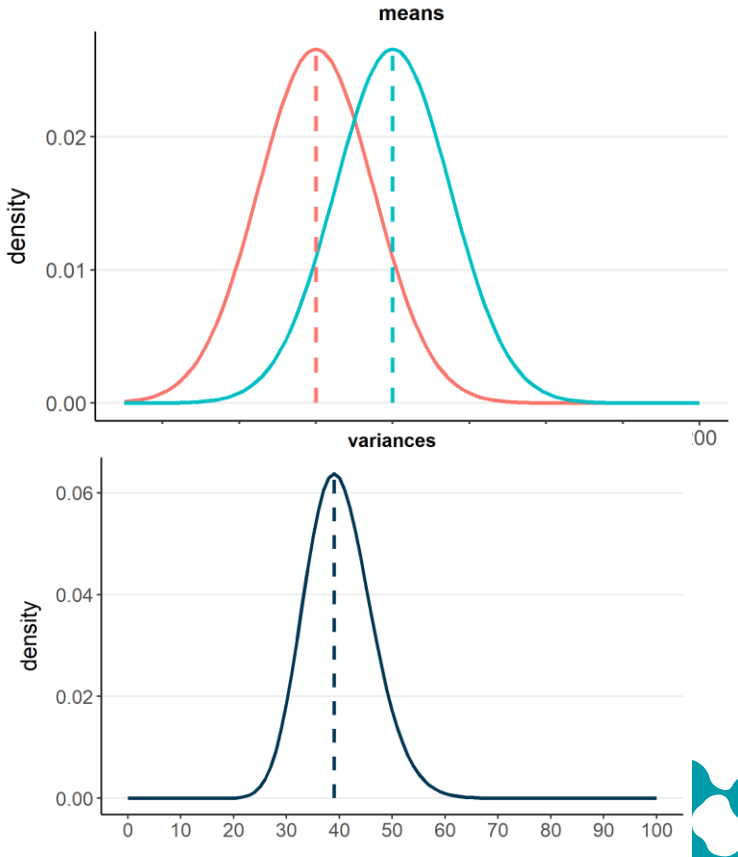
Power vs assurance

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

bayesian approach (assurance)

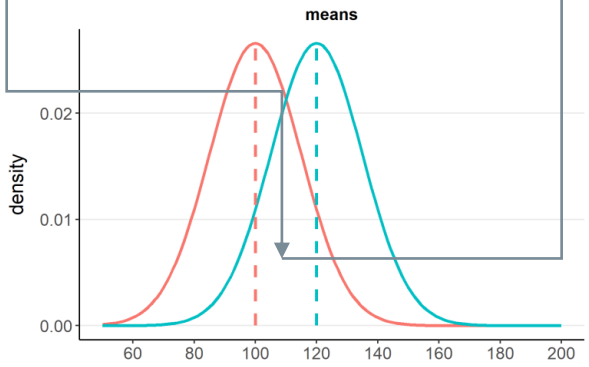
- In order to reflect the uncertainty, a large number of effect sizes, i.e. $(\mu_1 - \mu_2)/\sigma_{\text{pooled}}$, are generated using the prior distributions.
- A power curve is obtained for each effect size
- the expected (weighted by prior beliefs) power curve is calculated

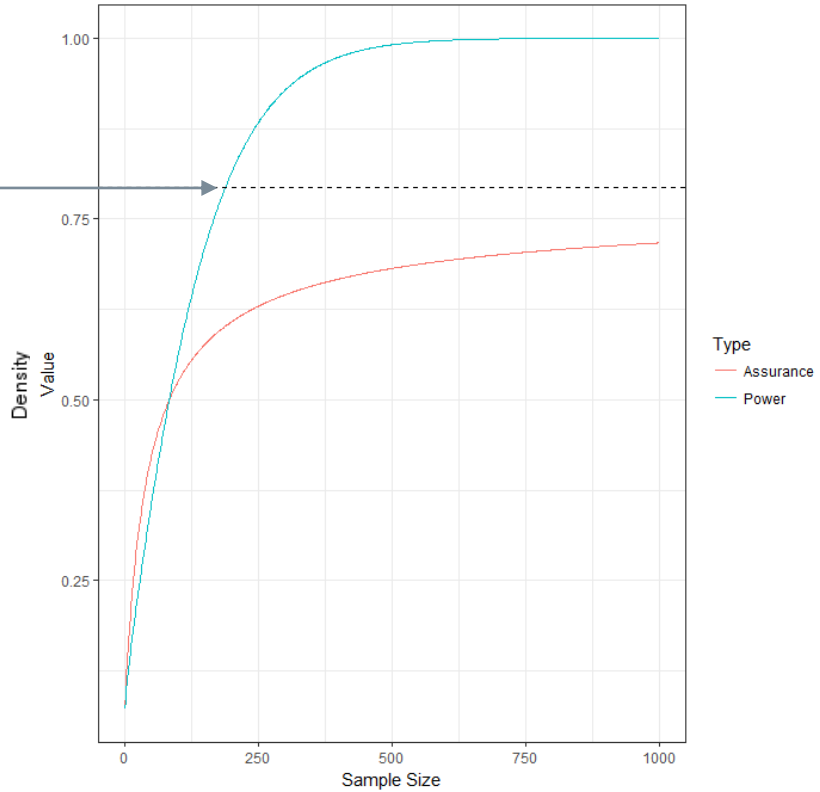
assumptions:



An example: Power vs Assurance

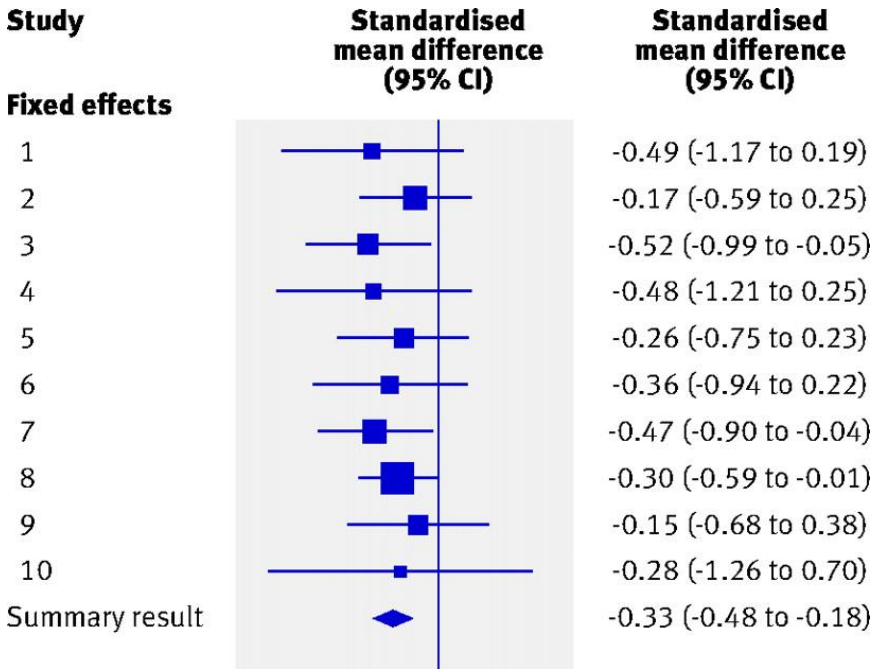
In this example the **assurance** converges to 0.793, that is the prior probability that the new drug is indeed superior





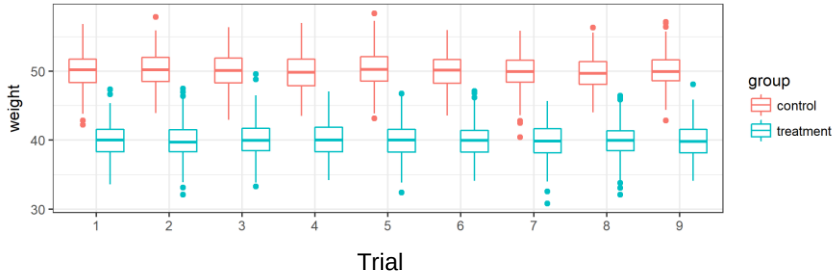
Design II: The missing components

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

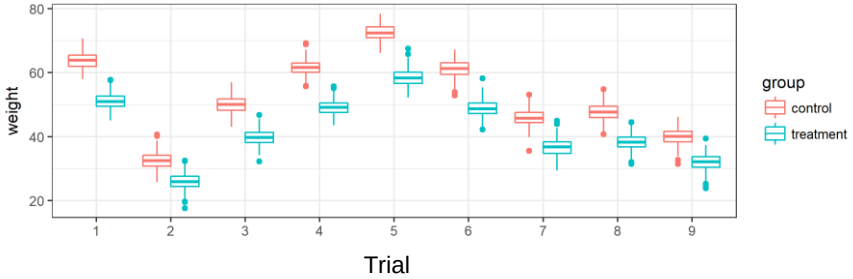


Different scenarios may happen

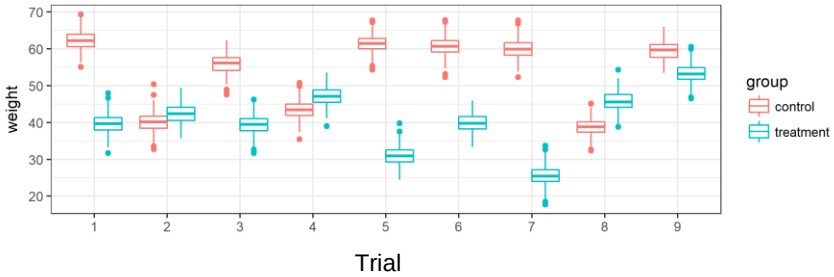
No variability



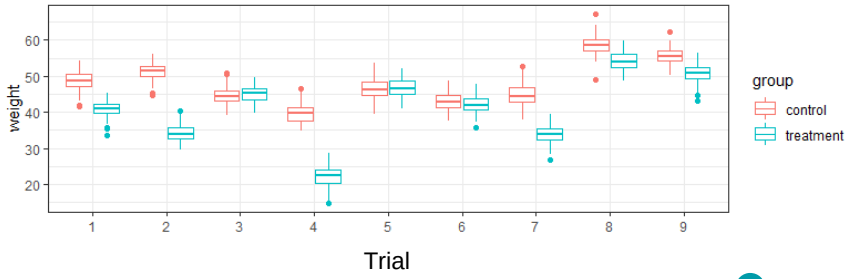
Groups vary together (ρ=1)



Groups vary independently (ρ=0)

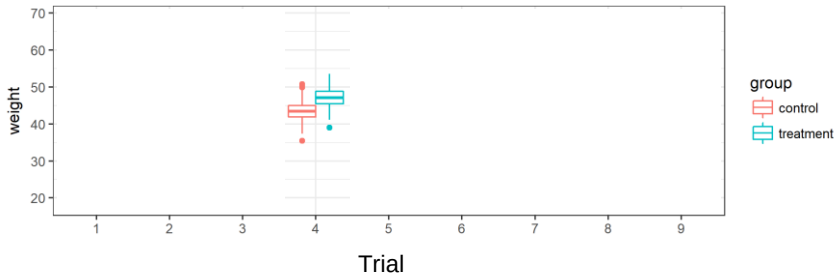


Groups vary with some dependencies (ρ~0.5)

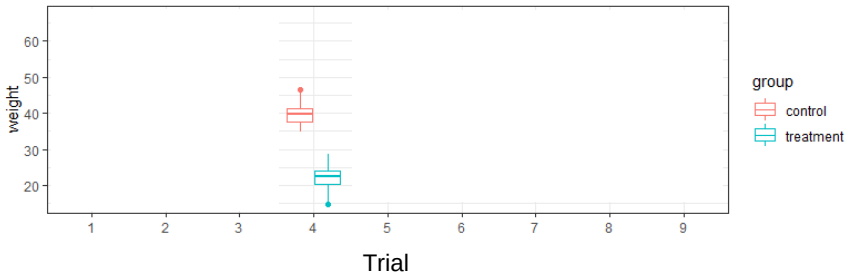


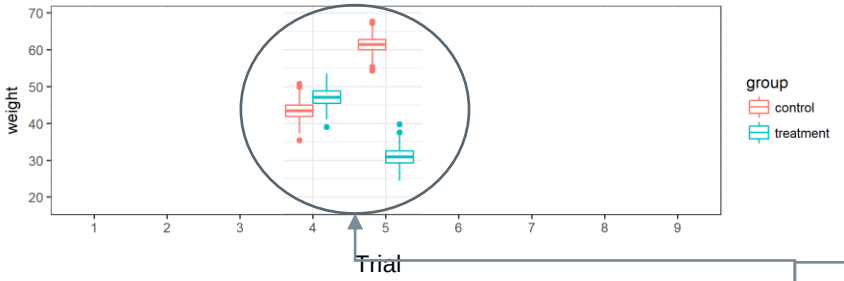
If you do one trial you may get one of those outcomes….

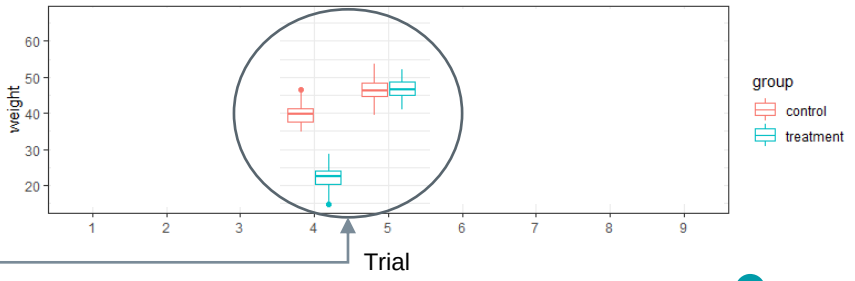
Groups vary independently



Groups vary with some dependencies





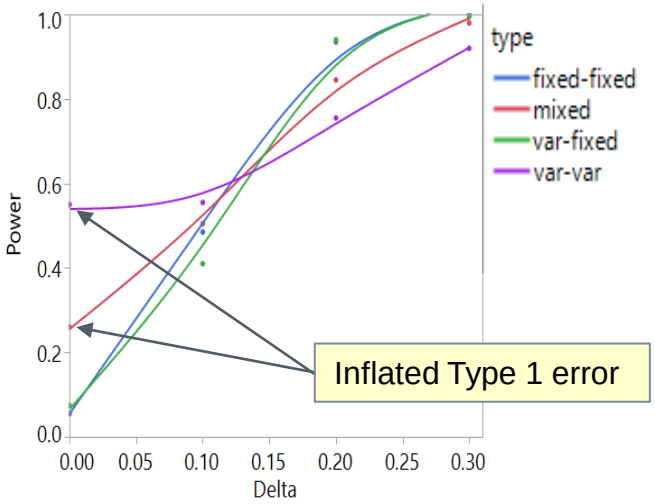


Does this new treatment works ?

Impact of study-to-study variability (and lab-to-lab)

- Everyone know there are such variabilities but this is ignored in design, power calculation, evaluation, …..
- It is even consciously avoided !
  - To have a “Better precision” !
- It is related to the “replicability” issue, achieving a robust conclusion regardless of the study

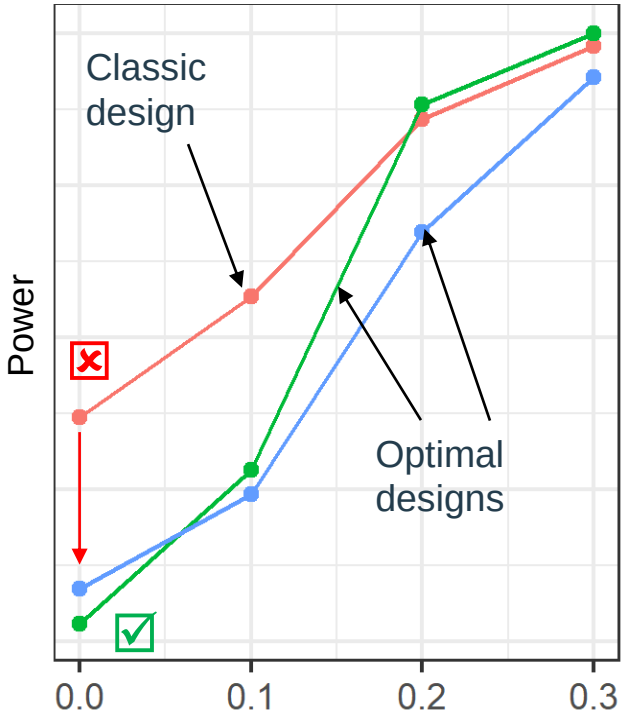
- If ignored and existing:
  - then there is a major risk of type I error-inflation!
  - the estimates are biased
  - It violates fundamental DoE practices

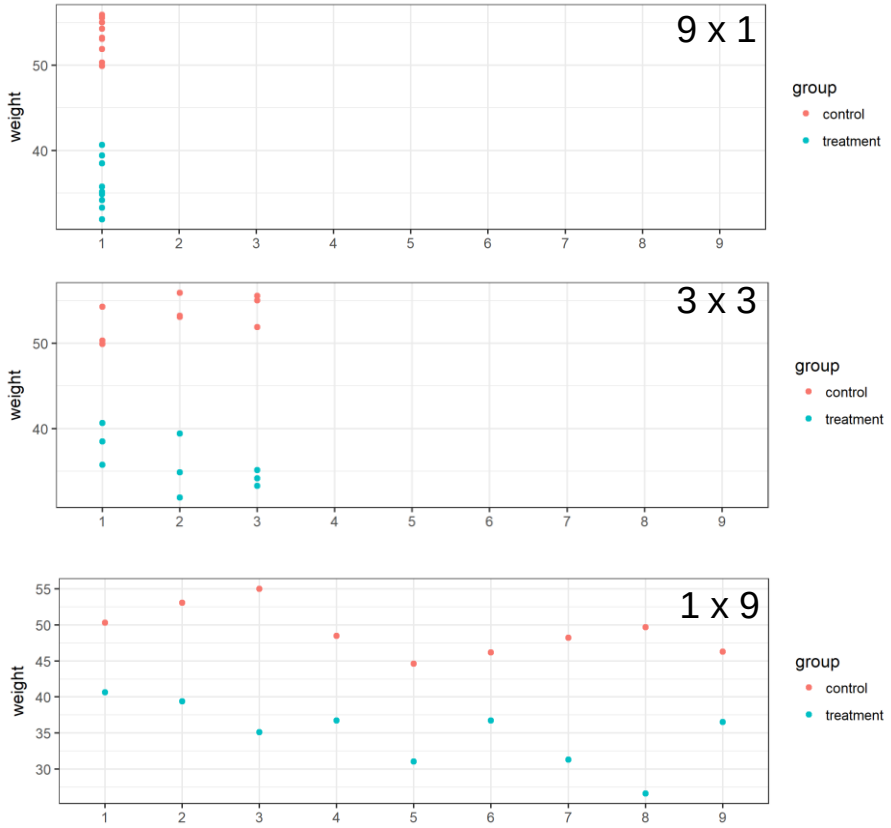


Bayesian method

➔ Novick S., and Zhang T. Mean Comparisons and Power Calculations to Ensure Replicability in Preclinical Drug Discovery, Stat. in Medicine, 2020.

Study “formats”: example in pre-clinical pharmacology





Improving precision of measurements

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

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

But in reality, it is:

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

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

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

1 study, 10 animals

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

2 studies, 5 animals/study

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

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

1 study, 10 animals

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

2 studies, 5 animals/study

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

Bayesian Methods

Received: 18 July 2020 | Revised: 18 November 2020 | Accepted: 21 November 2020

DOI: 10.1002/sim.8848

RESEARCH ARTICLE

Statistics in Medicine WILEY

Mean comparisons and power calculations to ensure reproducibility in preclinical drug discovery

Steven Novick | Tianhui Zhang

- Use predictive distribution
- Use of informative priors is justified in preclinical research

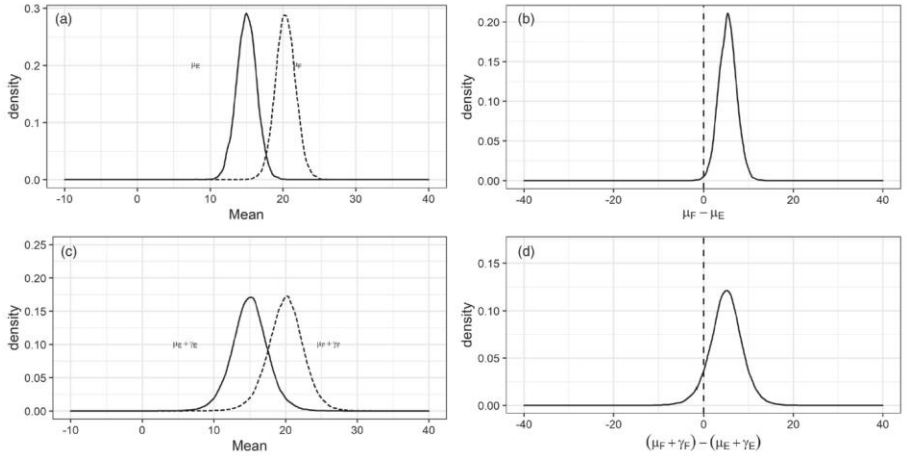


FIGURE 3 Panels (A) and (B) show the marginal posterior distributions for μ_E, μ_F, and μ_F - μ_E from the single-study analysis. Panels (C) and (D) show the predictive distributions for $\mu_E + \gamma_E$, $\mu_F + \gamma_F$, and $(\mu_F + \gamma_F) - (\mu_E + + \gamma_E)$

Conclusions

- What's the question ?
- In discovery the prior probability of success is low
- Broad use of Bayesian statistics in discovery and preclinical research will help to tackle the replicability crisis
- ….combined with better design of experiments as well
- Assurance instead of Power