Tackling The Reproducibility Crisis In Ecology

ESA Annual Meeting 2024

Peter Levy, UK Centre for Ecology & Hydrology

2024-07-25

The Reproducibility Crisis

1. What it means
2. The cause - high “false discovery” rates
3. An ecological example: soil carbon change
4. Solutions: a change in perspective in statistical thinking

The Reproducibility Crisis

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on *p*-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. "Negative" research is also very useful. "Negative" is actually a misnomer, and is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R+1)$. The probability of a study finding a true relationship reflects the power $1-\beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2×2 table, one gets PPV = $(1-\beta)R/(R$

"Reproducibility" of results, not methods.

Ioannidis, 2005, PLOS Medicine

The Reproducibility Crisis

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

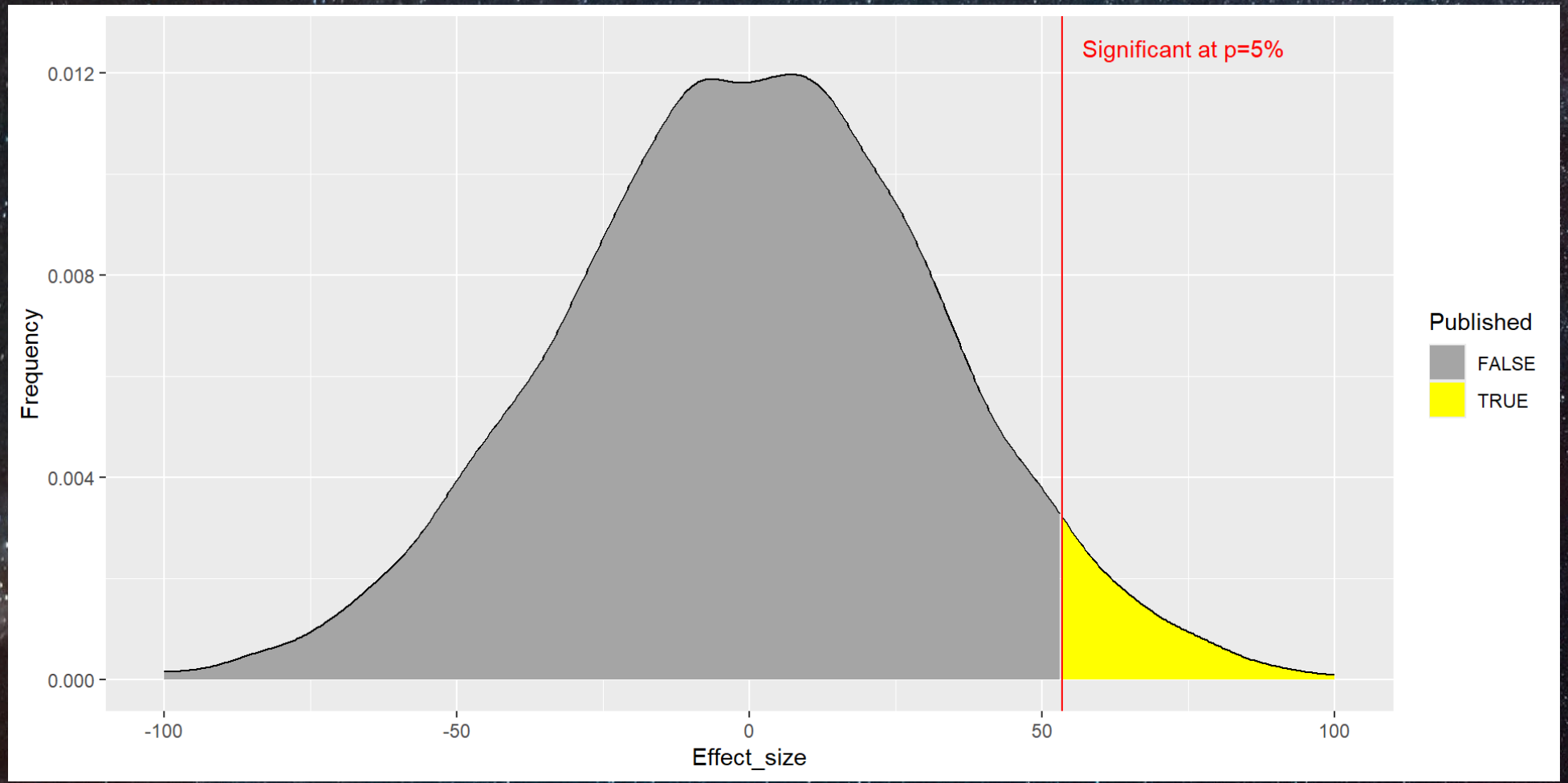
should be interpreted based only on *p*-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. "Negative" research is also very useful. "Negative" is actually a misnomer, and is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R+1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2×2 table, one gets PPV = $(1 - \beta)R/(R$

"The real purpose of scientific method is to make sure Nature hasn't misled you into thinking you know something you don't actually know."

Robert Pirsig

https://doi.org/10.1371/journal.pmed.0020124

Unpublished work as Dark Matter



We can only see 5% of the universe.

"Why most published research findings are false"

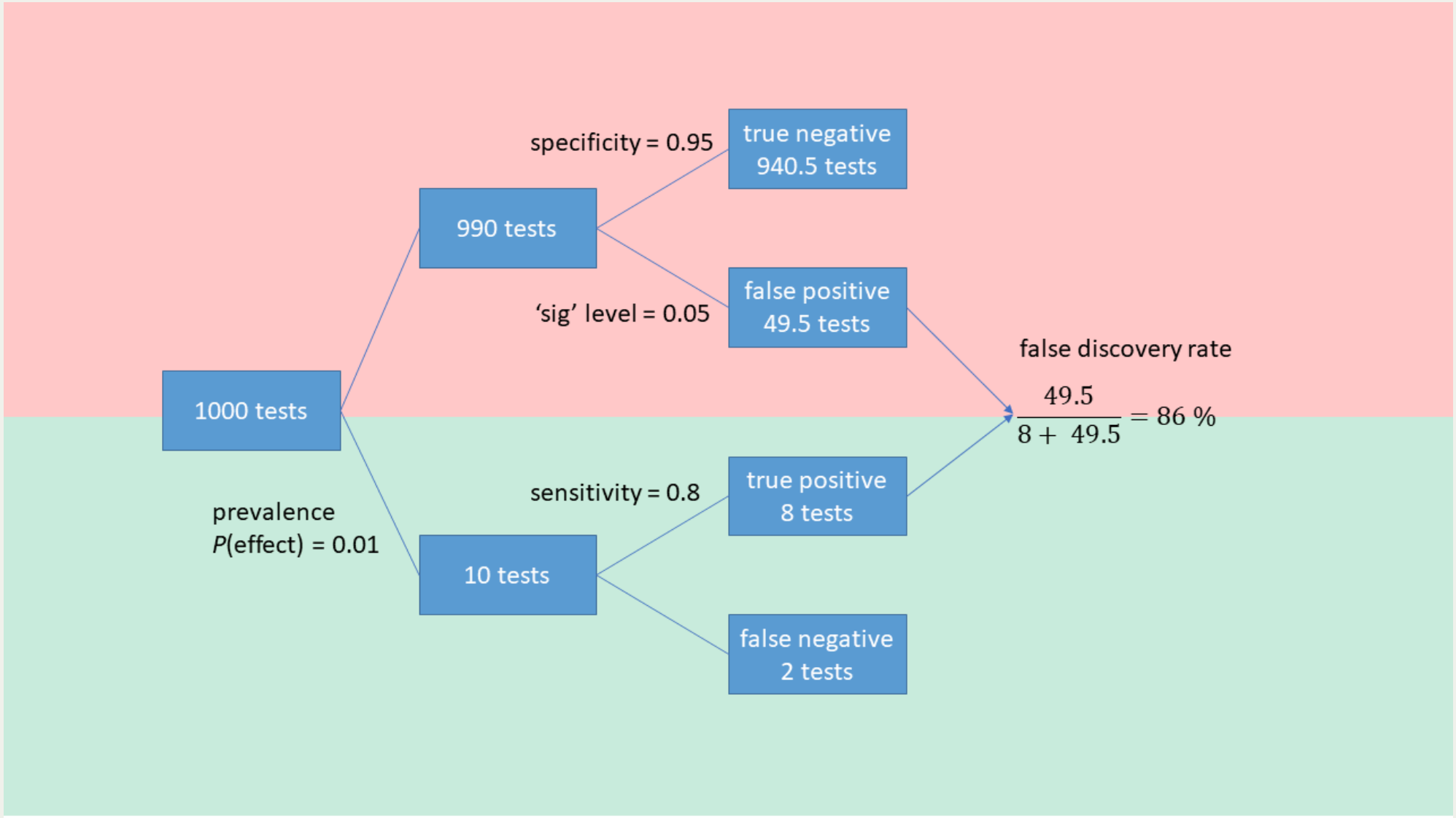
High "false discovery rates", often much higher than 5 %

because of:

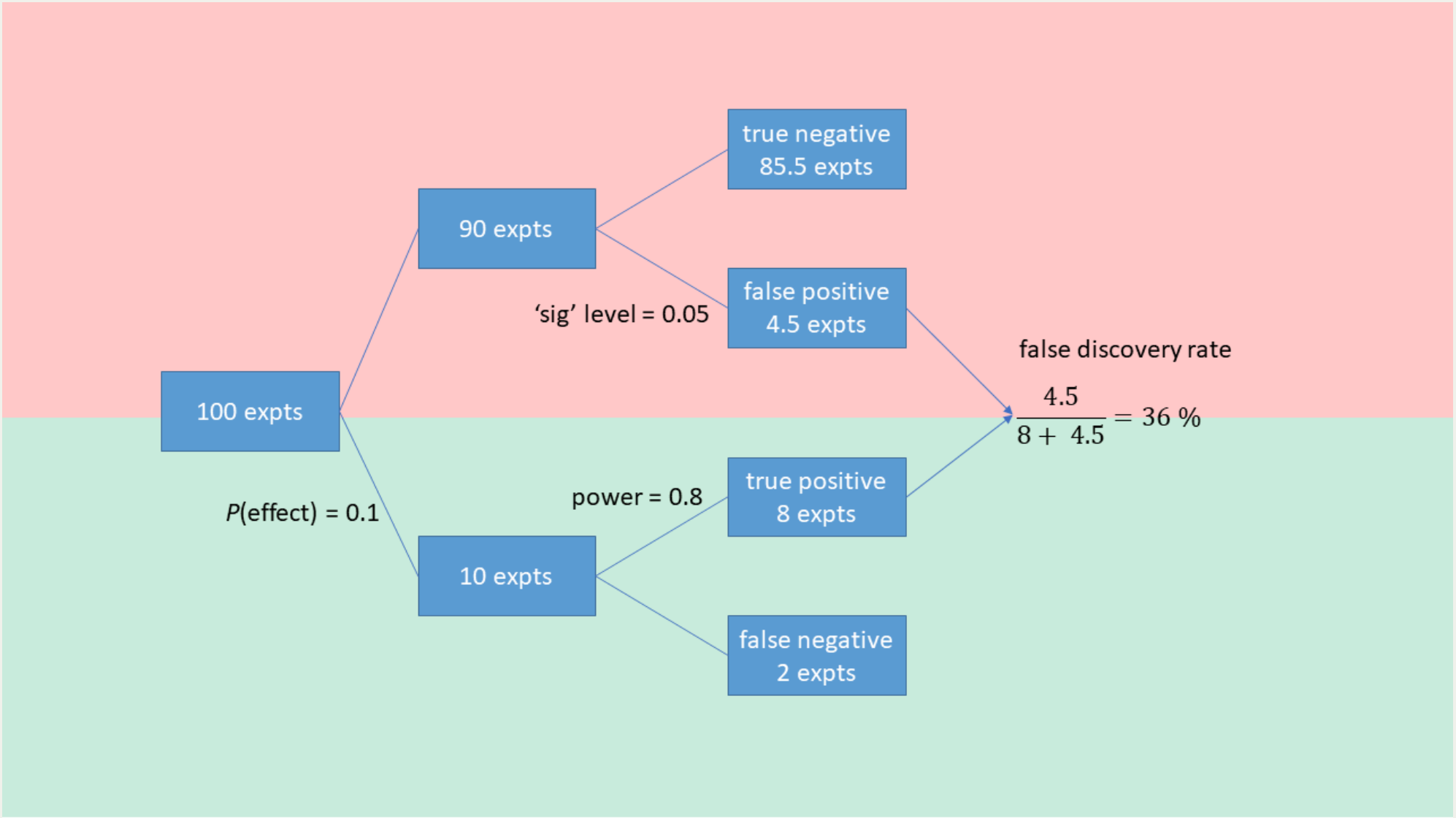
- low prior $P(\mathrm{effect})$ - unlikely / rare effects
- low statistical power - low signal:noise
- bias - systematic uncertainty

Understanding false discovery rates

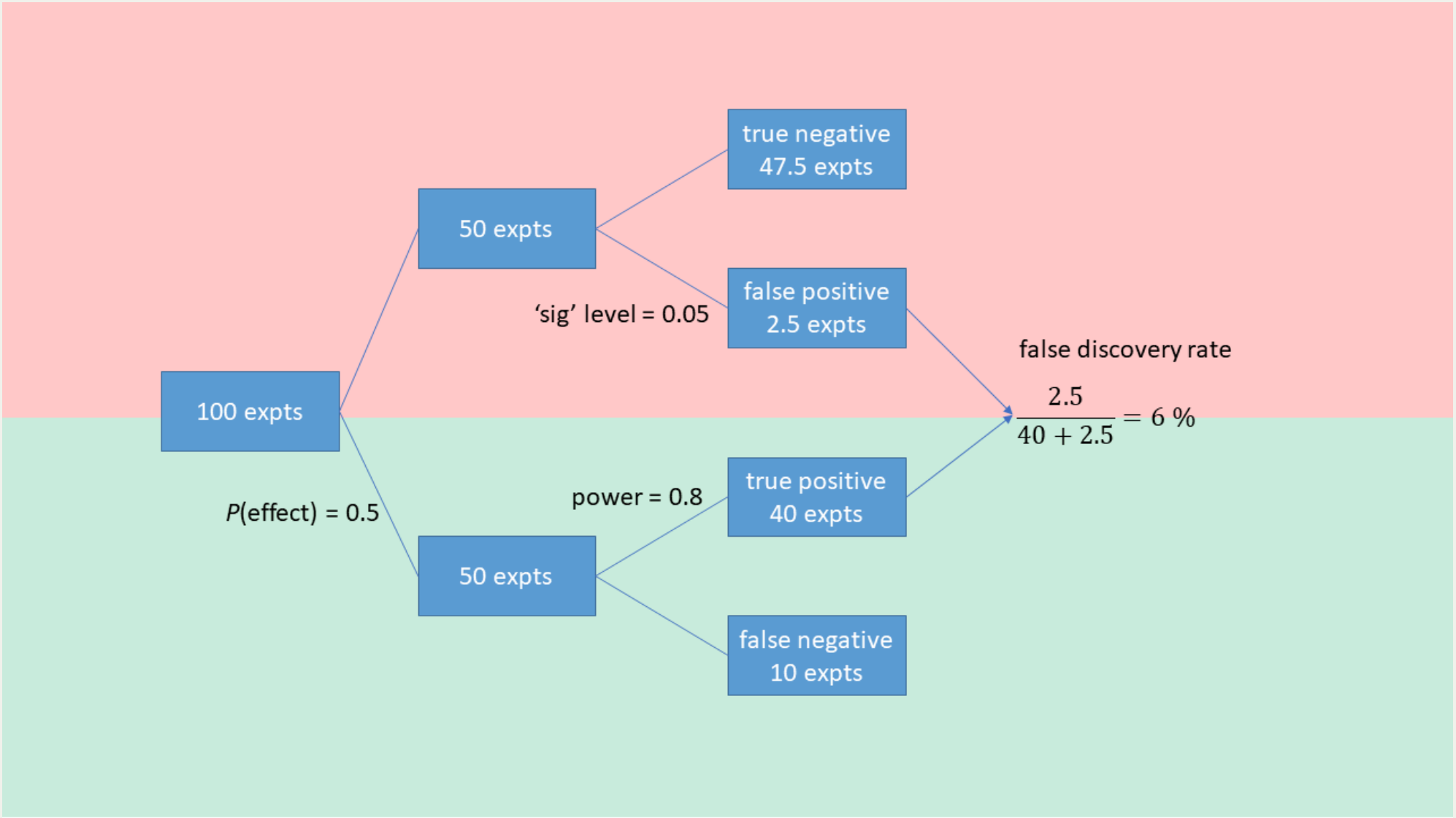
Disease testing



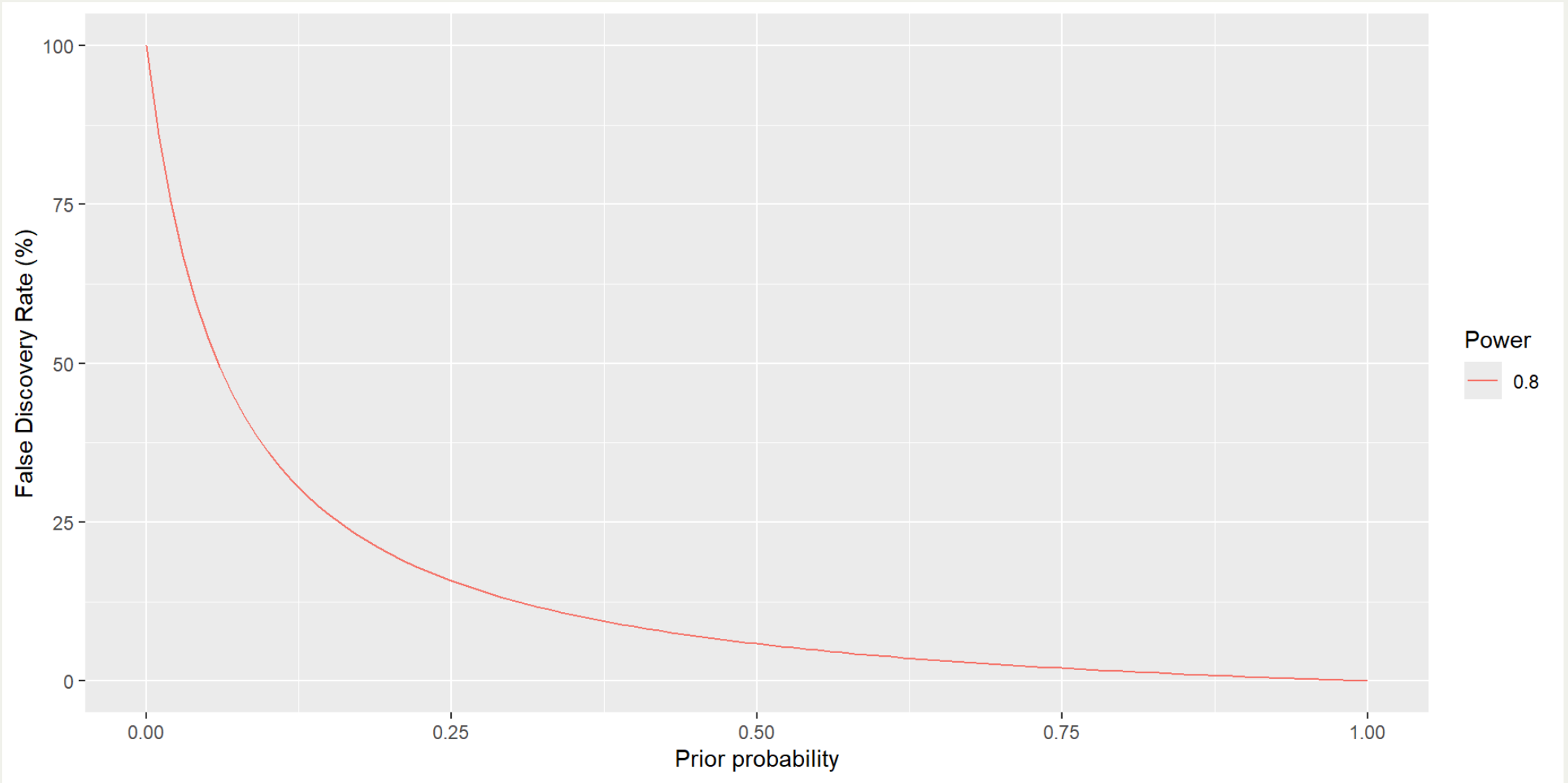
Generic experiments



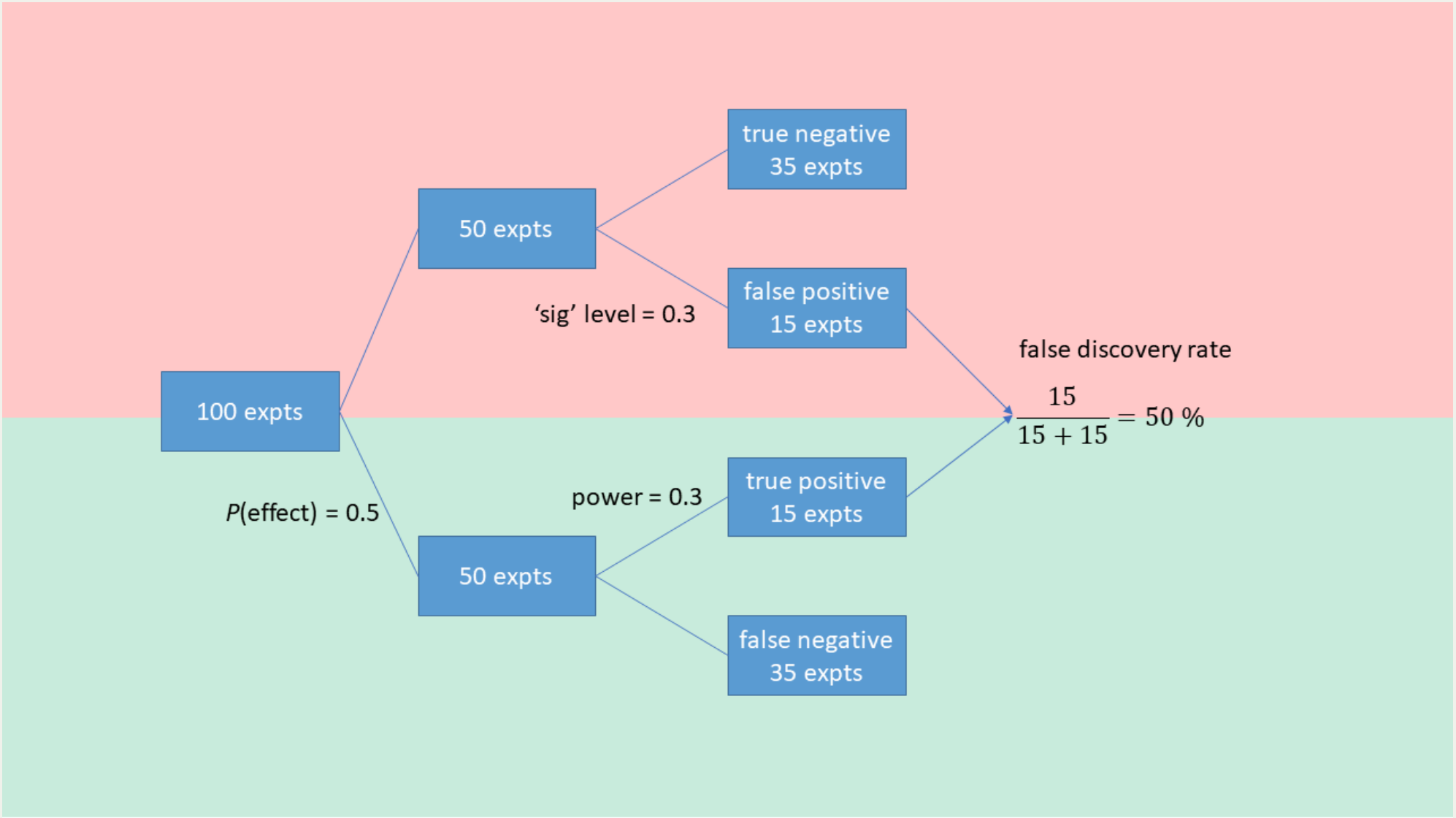
Generic experiments



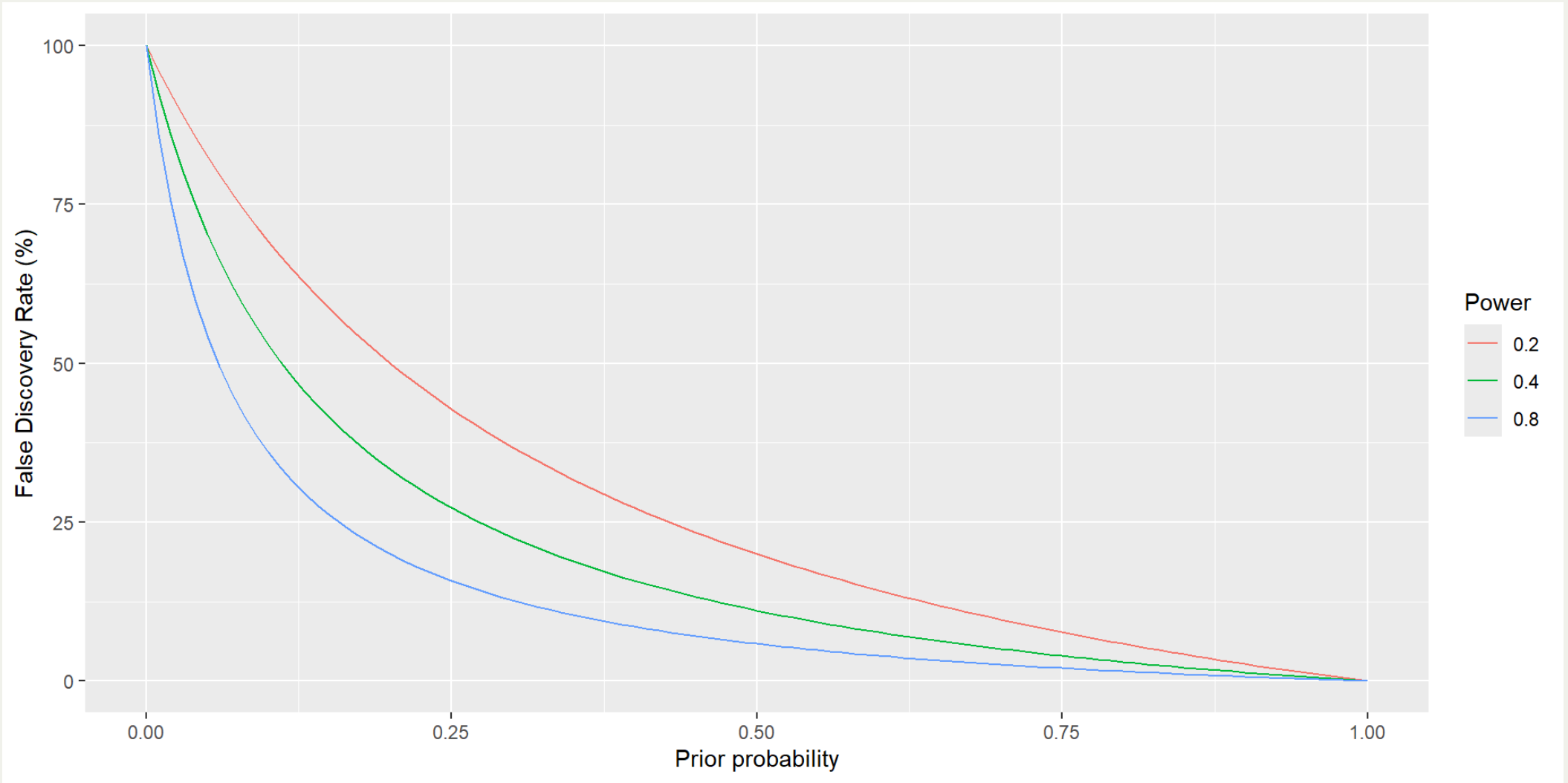
Effect of prior probability



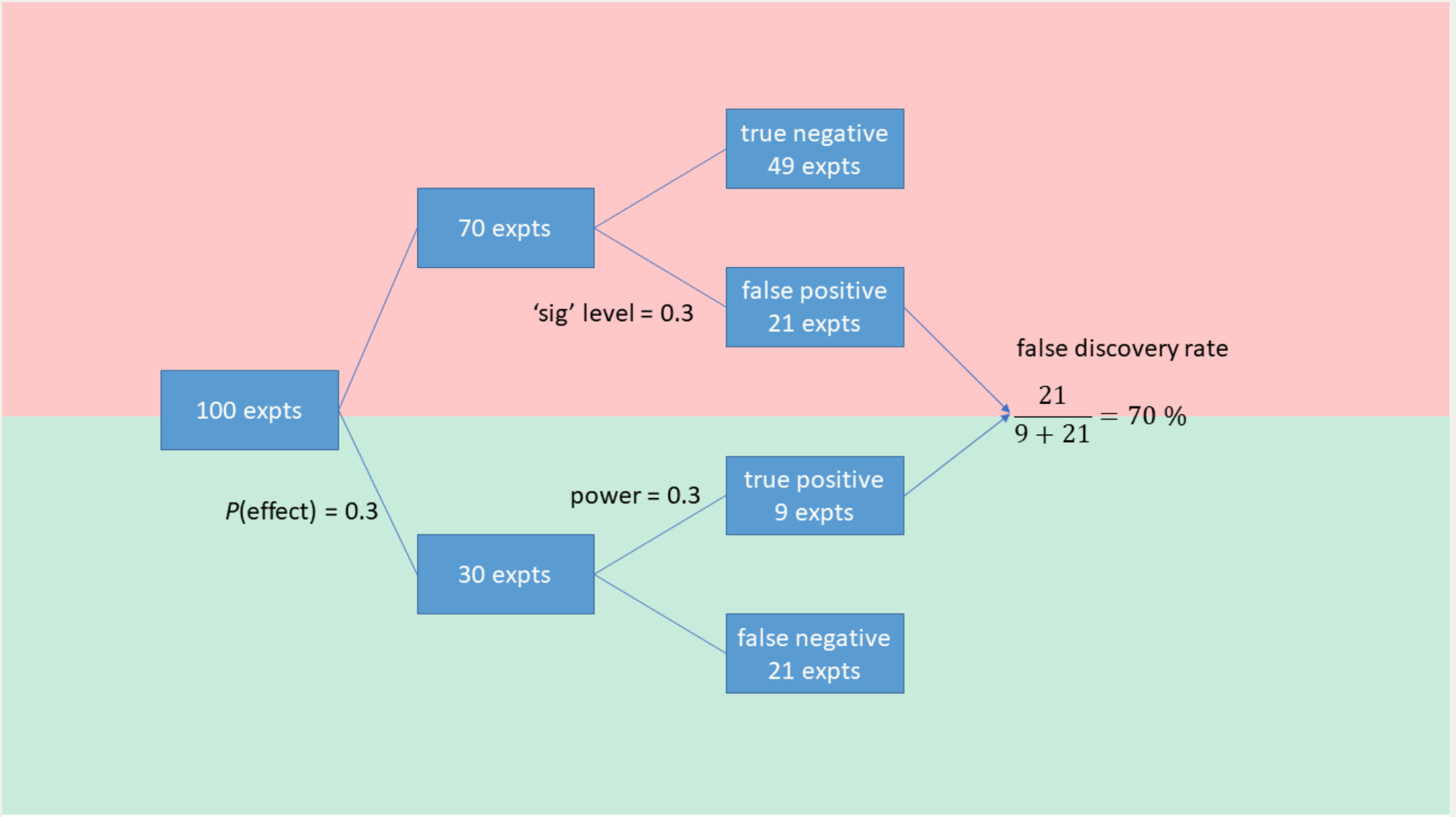
Generic experiments: low power



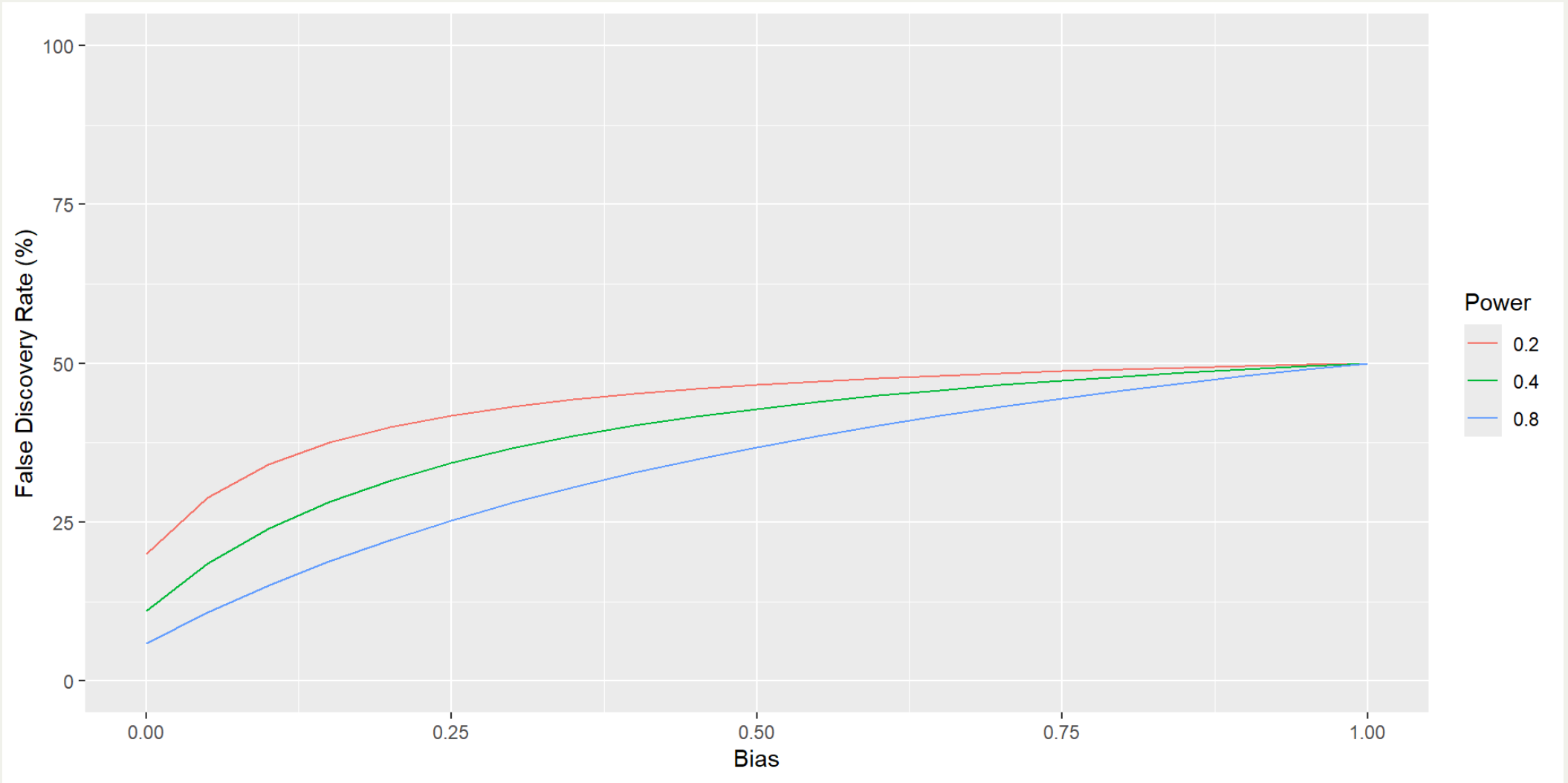
Effect of low power



Generic experiments: bias



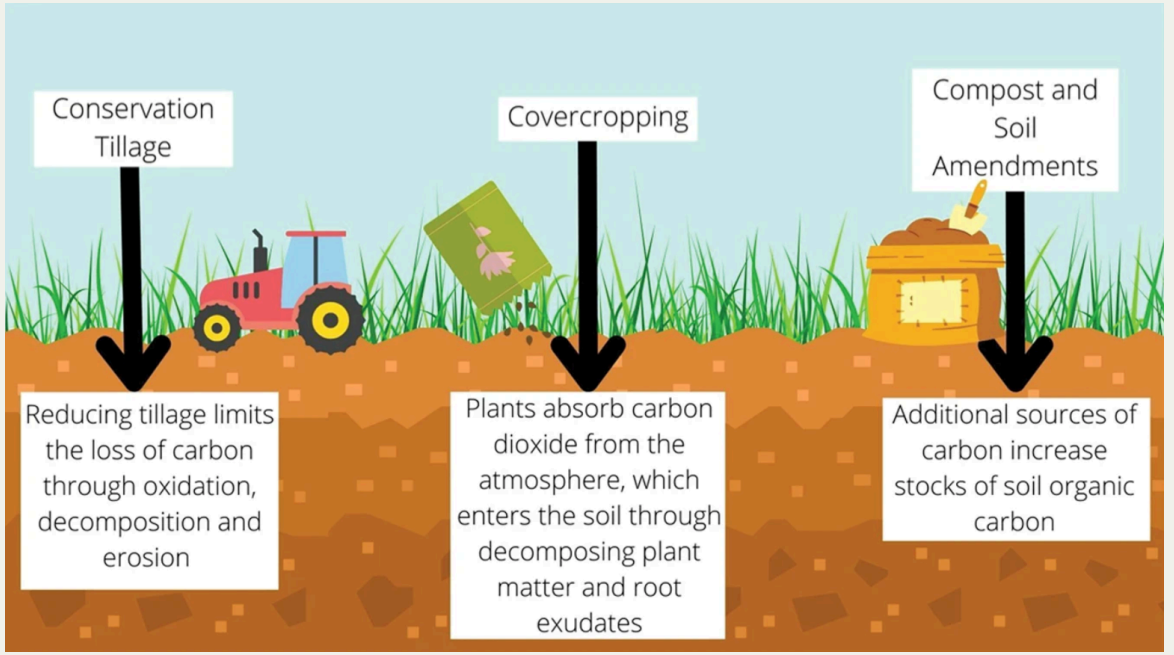
Effect of bias



An ecological example: soil carbon change

“Nature-based solutions” to climate change

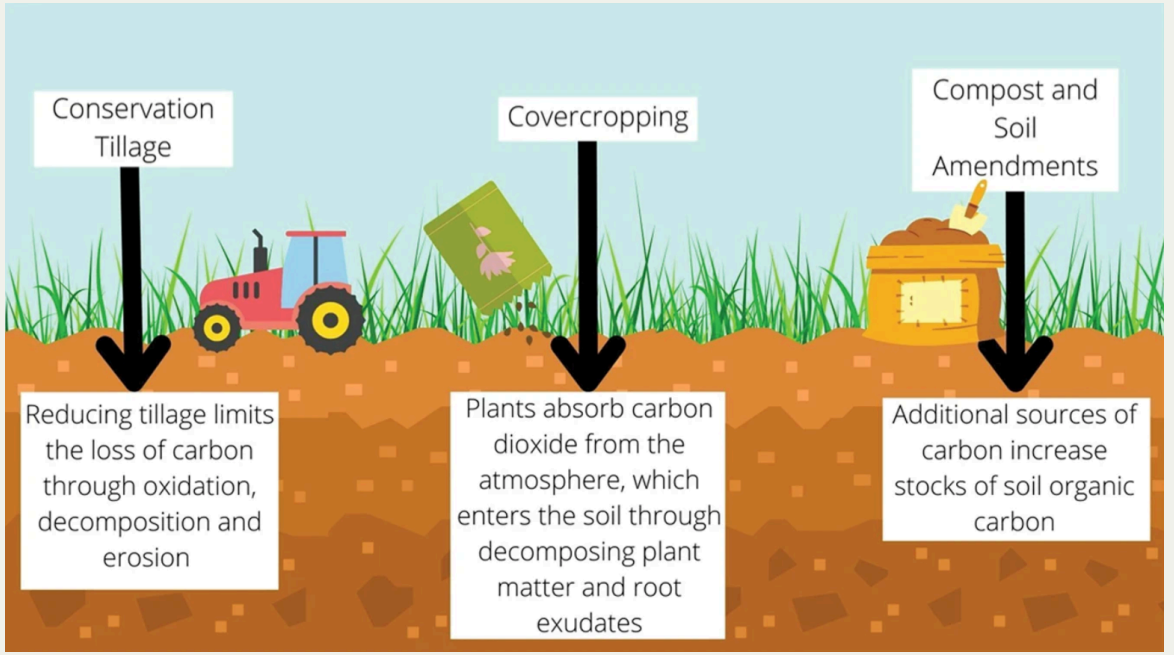
- “rejuvenative farming”
- biochar
- reduced stocking
- rewilding



“Nature-based solutions” to climate change

BUT:

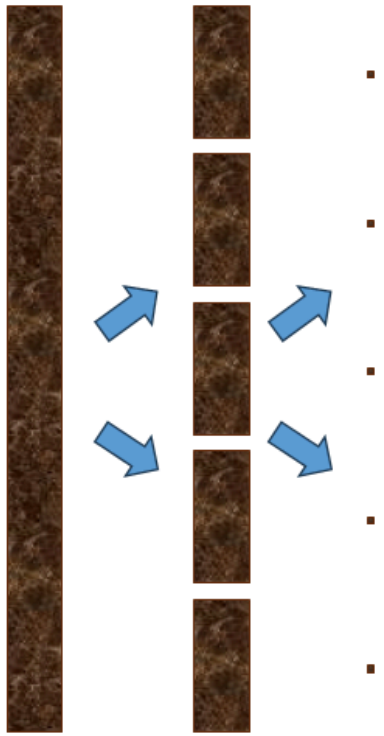
- change in soil carbon very hard to verify
- huge potential for “greenwashing”



Measuring soil carbon: field

Take soil cores

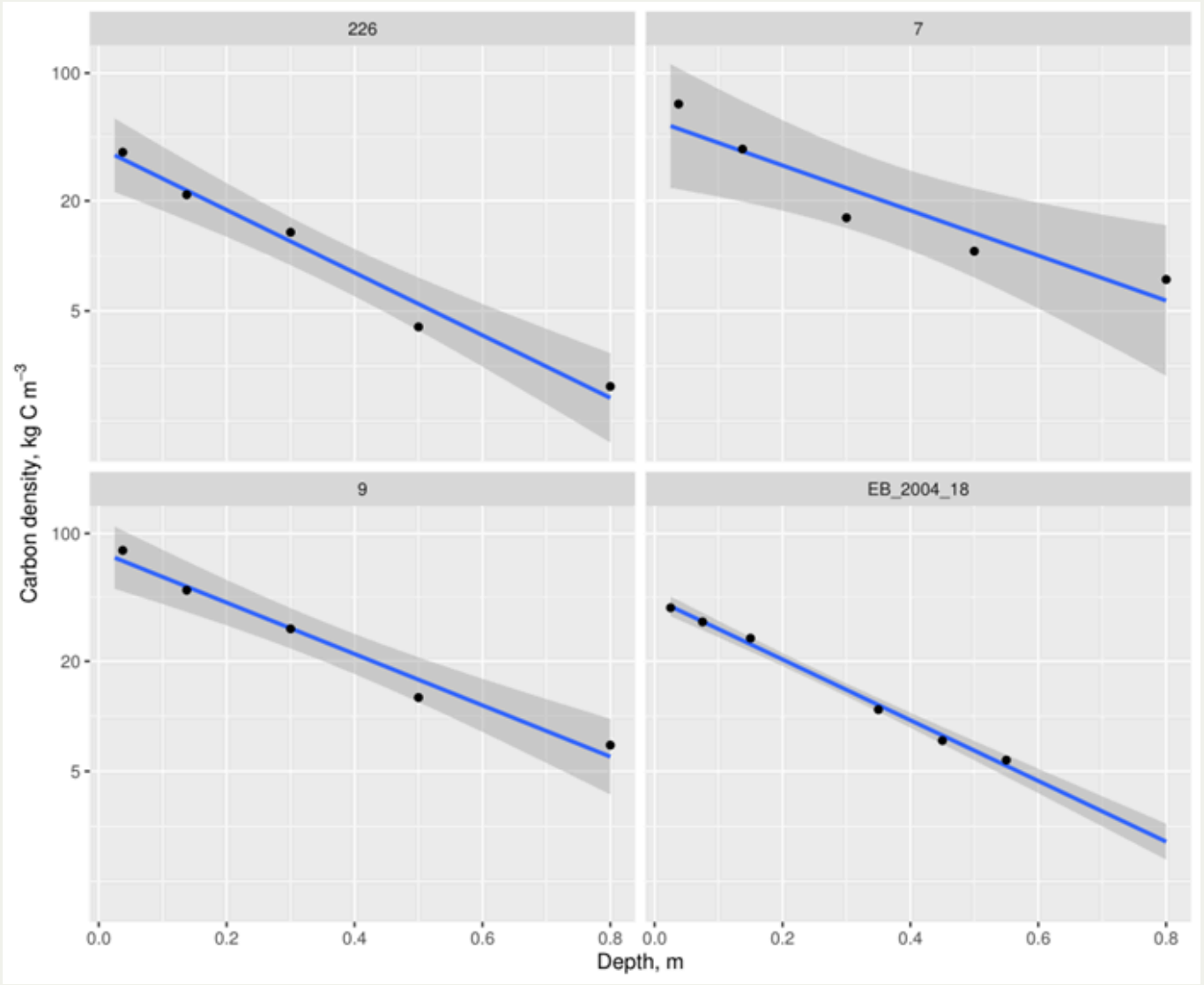
Measuring soil carbon: lab



soil core → depth section → dry, sieve, grind, sub-sample → weigh → combust → weigh

carbon = β × weight loss

Soil carbon: integrate over depth



$$\log C = \alpha + \beta \times \text{dep}$$

UKCEH

Soil carbon: integrate over space

Extrapolate samples to whole field with spatial model

$$\mu_{\text{field}} = f(\theta, C_{\text{samples}}$$

Soil carbon: uncertainties

- We would typically say we have “measurements of soil carbon”.
- But we actually have measurements of weight loss
- from a sample of a sample …
- from which we predict carbon fraction with a model …
- and predict carbon stock to depth with a model …
- and extrapolate this in space with a model …
- and we (usually) ignore most of the uncertainties!

Soil carbon: uncertainties

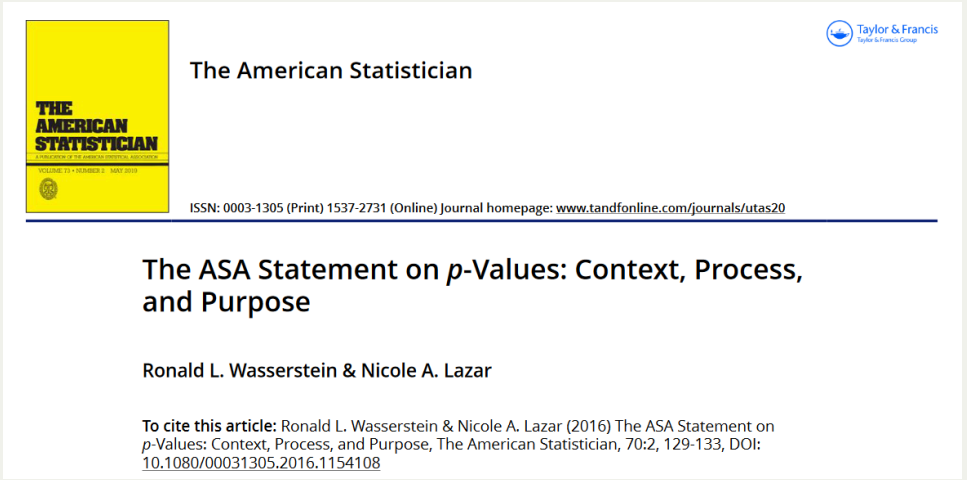
It gets worse.

- Systematic uncertainties (bias) from different:
- surveyors
- sampling protocols
- areas & depths sampled
- labs & instruments
- lab protocols
- very hard to be consistent over ~20 years.

Results

Plot of FDR for typical experiments with typical power and bias Bias is the killer

The ASA statement



Taylor & Francis

The American Statistician

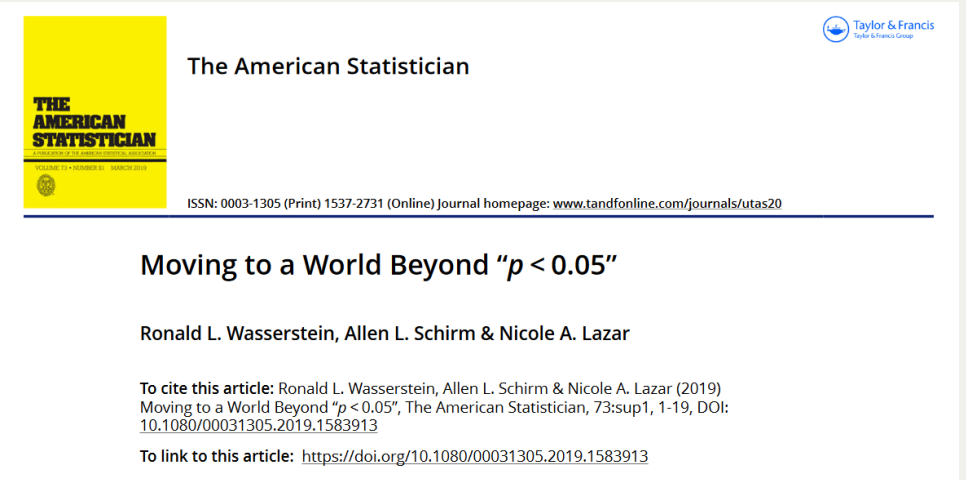
ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: www.tandfonline.com/journals/utas20

The ASA Statement on *p*-Values: Context, Process, and Purpose

Ronald L. Wasserstein & Nicole A. Lazar

To cite this article: Ronald L. Wasserstein & Nicole A. Lazar (2016) The ASA Statement on *p*-Values: Context, Process, and Purpose, The American Statistician, 70:2, 129-133, DOI: 10.1080/00031305.2016.1154108

2016



Taylor & Francis

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: www.tandfonline.com/journals/utas20

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

Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar

To cite this article: Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar (2019) Moving to a World Beyond "$p < 0.05$", The American Statistician, 73:sup1, 1-19, DOI: 10.1080/00031305.2019.1583913

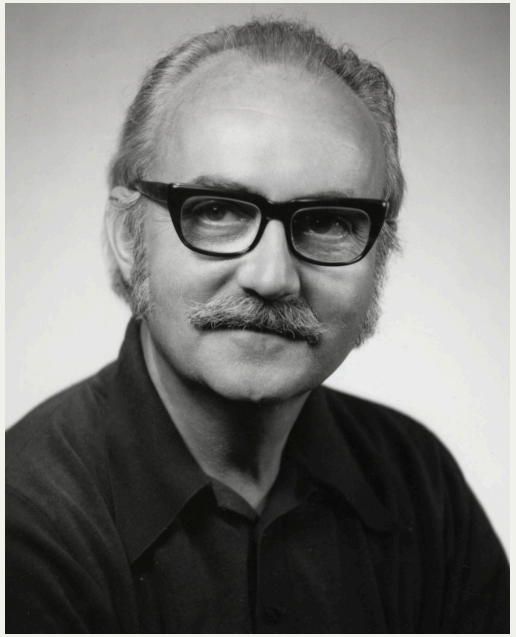
To link to this article: https://doi.org/10.1080/00031305.2019.1583913

2019

Stop using *p* values and "statistical significance".

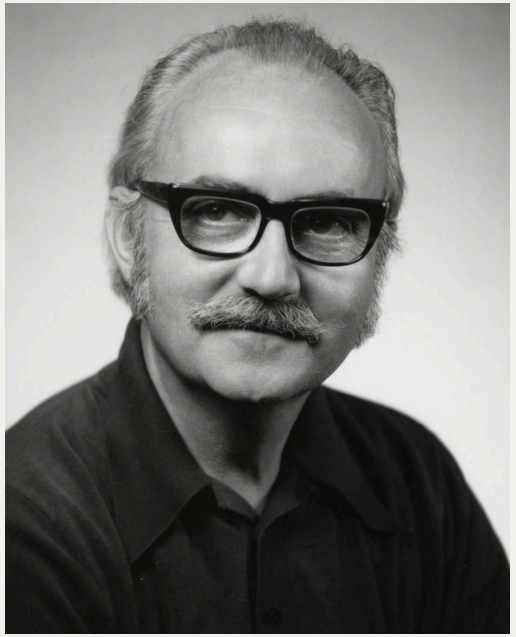
“All models are wrong, but some are useful.”

George Box, 1976.



“All models are wrong, but some are useful.”

George Box, 1976.



“All data are wrong, but some are useful.”

Patterson & Gimlin, 1967

A Copernican Revolution