
www.forschung-und-lehre.de
29. Jahrgang | 7,–€
Forschung & Lehre
10 | 22
ALLES WAS DIE WISSENSCHAFT BEWEGT

Statistisch signifikant bzw. nicht signifikant

Warum die Reform der statistischen Berichtspraxis nicht vorankommt

| NORBERT HIRSCHAUER | SVEN GRÜNER | OLIVER MUSSHOFF |

Bei der Debatte über statistische Signifikanztests geht es um die wohl grundlegendste Frage der datenbasierten Wissenschaften: Was können Wissenschaftlerinnen und Wissenschaftler aus einem bestimmten Datensatz lernen und wie kommen sie angesichts der verbleibenden Unsicherheit zu einer möglichst vernünftigen Aussage über einen unbekannten realen Sachverhalt? Es geht um das Verständnis, wann und wie statistische Größen helfen können, den jeweils erzielbaren Erkenntnisgewinn zu bewerten.

Seit den 50er Jahren des 20. Jahrhunderts haben sich „statistische Signifikanztests“ in fast allen empirischen Disziplinen als *die* inferenzstatistische Norm durch

oder nicht. Oft findet man sogar schon beim Forschungsziel die Formulierung, man wolle herausfinden, ob ein statistisch signifikanter Effekt vorliegt oder nicht. Die allgegenwärtige und nicht

bei einer Vielzahl nicht signifikanter kleiner Studien, die alle Größenunterschiede zwischen 100 und 140 mm finden, wird der Informationsstand oft wie folgt zusammengefasst: „*Eine Vielzahl von Studien hat gezeigt, dass kein statistisch signifikanter Unterschied in der Körpergröße besteht.*“

An dieser Stelle erübrigt sich vermutlich jeder Kommentar zu den irreführenden Wirkungen von Signifikanzaussagen. Entgegen aller alltagssprachlichen Assoziationen bedeutet „statistisch signifikant“ lediglich, dass der

(Forschung und Lehre 10/2022)

Critical debate about use and misuse of p-values (Hirschauer et al. 2022)

- Null Hypothesis Significance Testing (NHST)
- Based on ideas of significance testing (Fisher 1925) and hypothesis testing (Neyman & Pearson 1933)
- Methodological warning American Statistical Association 2016, special issue

  "Statistical Inference in the 21st Century: A World Beyond $p < 0.05$."
- Call to retire significance testing in Nature 2019 (Amrhein et al.)
- Reproducibility crisis (Ioannidis 2005 "Why most research findings are false")
- Publication bias

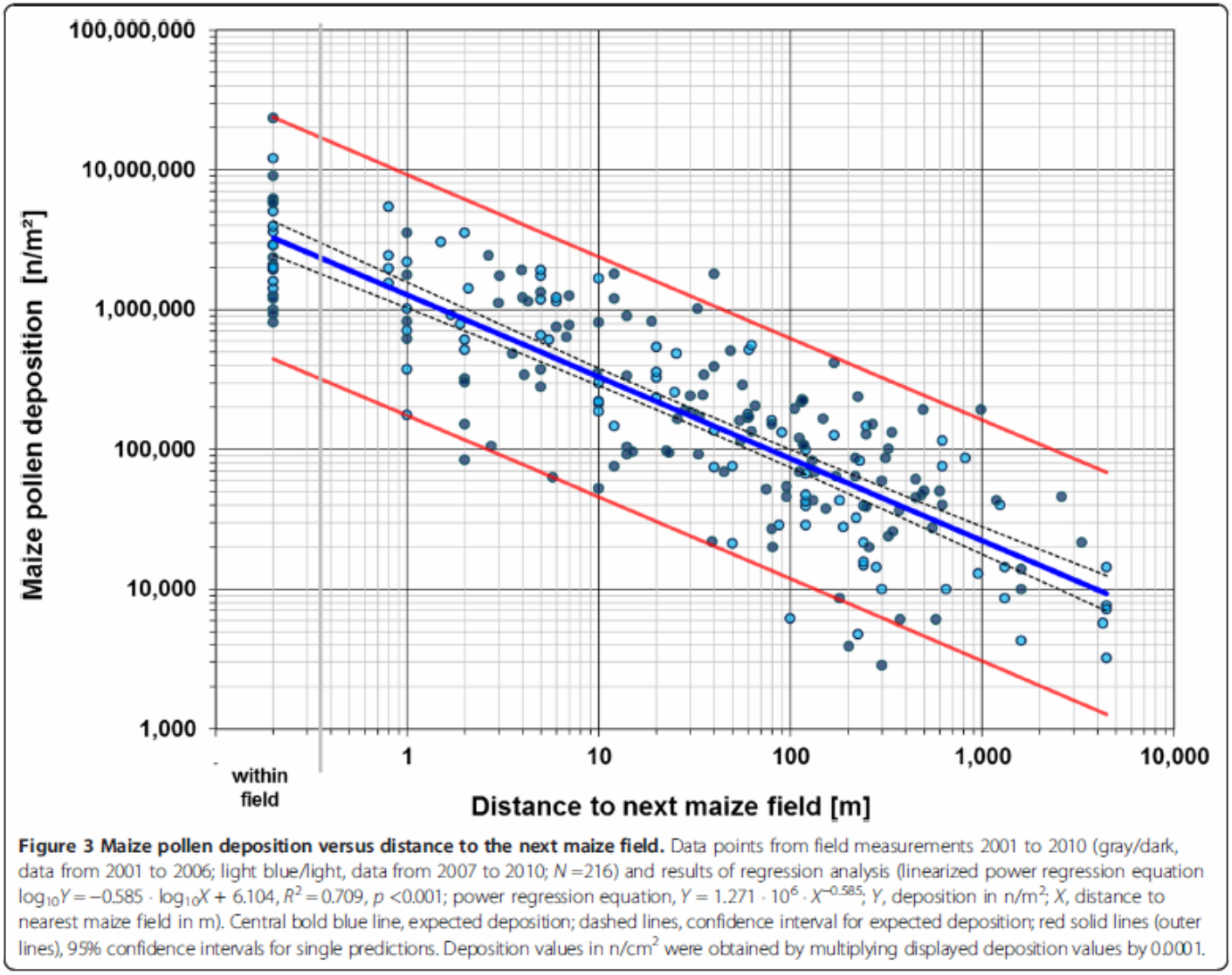


Figure 3 Maize pollen deposition versus distance to the next maize field. Data points from field measurements 2001 to 2010 (gray/dark, data from 2001 to 2006; light blue/light, data from 2007 to 2010; $N = 216$) and results of regression analysis (linearized power regression equation $\log_{10}Y = -0.585 \cdot \log_{10}X + 6.104$, $R^2 = 0.709$, $p < 0.001$; power regression equation, $Y = 1.271 \cdot 10^6 \cdot X^{-0.585}$; Y, deposition in n/m^2; X, distance to nearest maize field in m). Central bold blue line, expected deposition; dashed lines, confidence interval for expected deposition; red solid lines (outer lines), 95% confidence intervals for single predictions. Deposition values in n/cm^2 were obtained by multiplying displayed deposition values by 0.0001.

Hofmann et al. (2014): *Environmental Sciences Europe* **26**:24

Maize pollen deposition (y; log-scale) in relation to distance from the nearest pollen source under common cultivation (x; log-scale) - results of 10 years of monitoring (2001 to 2010)

$$E(y) = \alpha + \beta x$$

α = intercept

β = slope

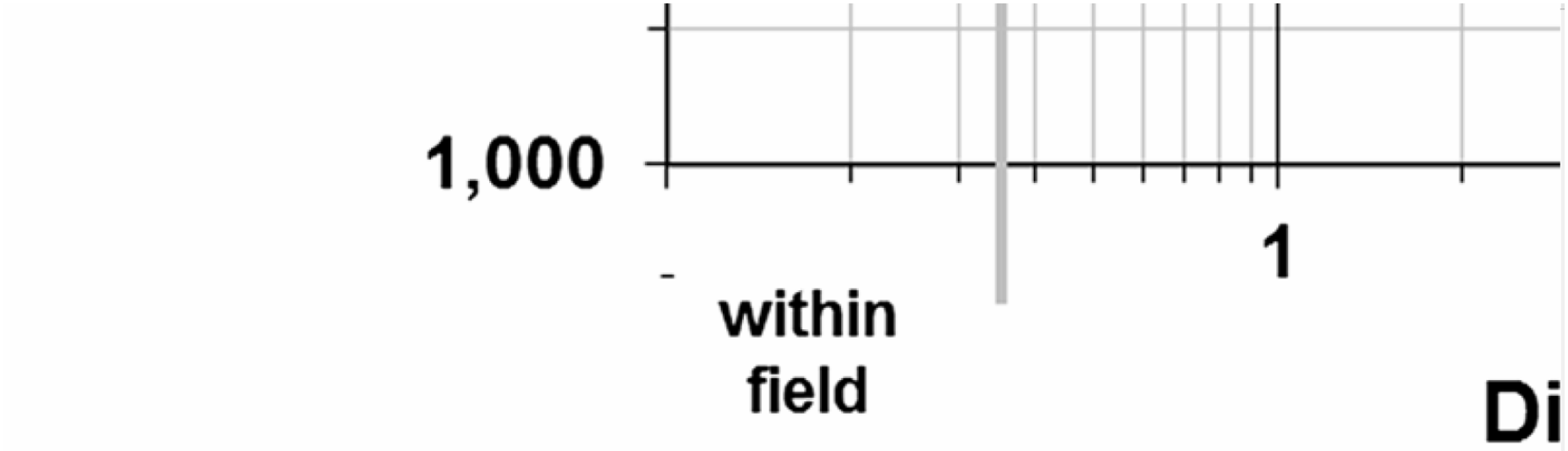


Figure 3 Maize pollen deposition versus distance to th

data from 2001 to 2006; light blue/light, data from 2007 to

$\log_{10}Y = -0.585 \cdot \log_{10}X + 6.104$, $R^2 = 0.709$, $p < 0.001$; powe

nearest maize field in m). Central bold blue line, expected c

lines), 95% confidence intervals for single predictions. Depos

Basic concepts

Example: linear regression, slope parameter β

Sample estimator $\hat{\beta}$

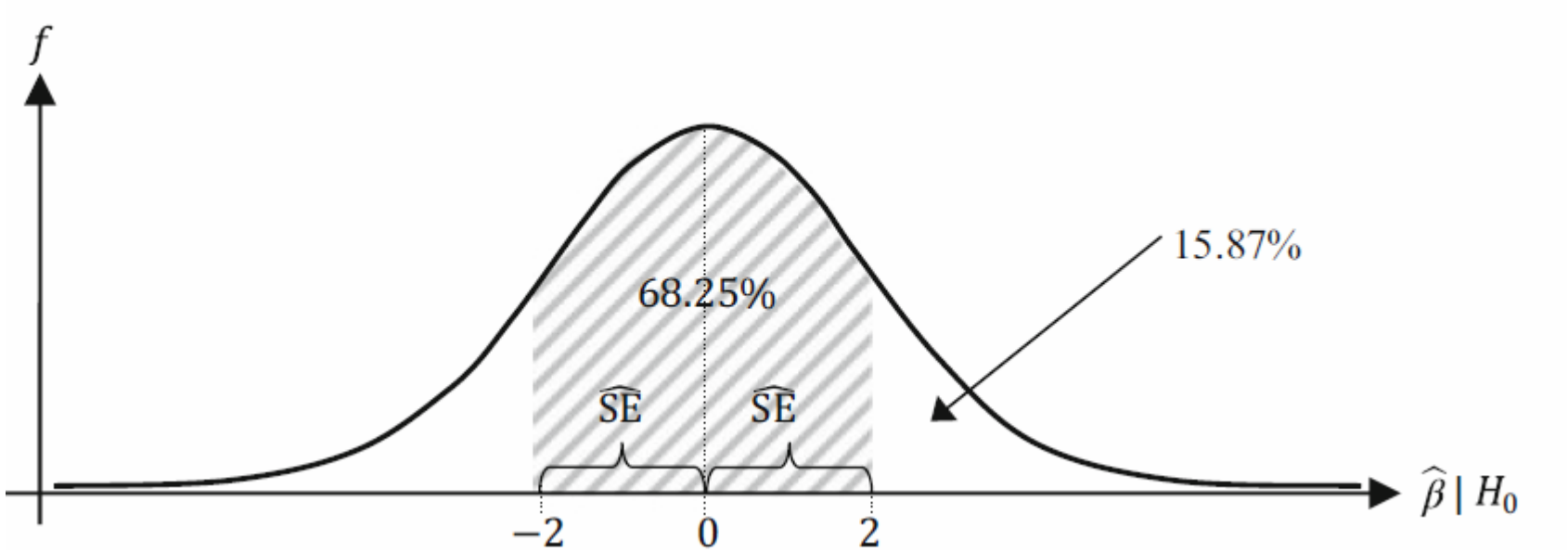
Signal-to-noise ratio $z = \frac{\hat{\beta}}{SE}$

SE = standard error

Null hypothesis H_0: $\beta = 0$, Alternative hypothesis H_A: $\beta \neq 0$

Test statistic: z !

Distribution of $\hat{\beta}$ under H_0



(Hirschauer et al. 2022)

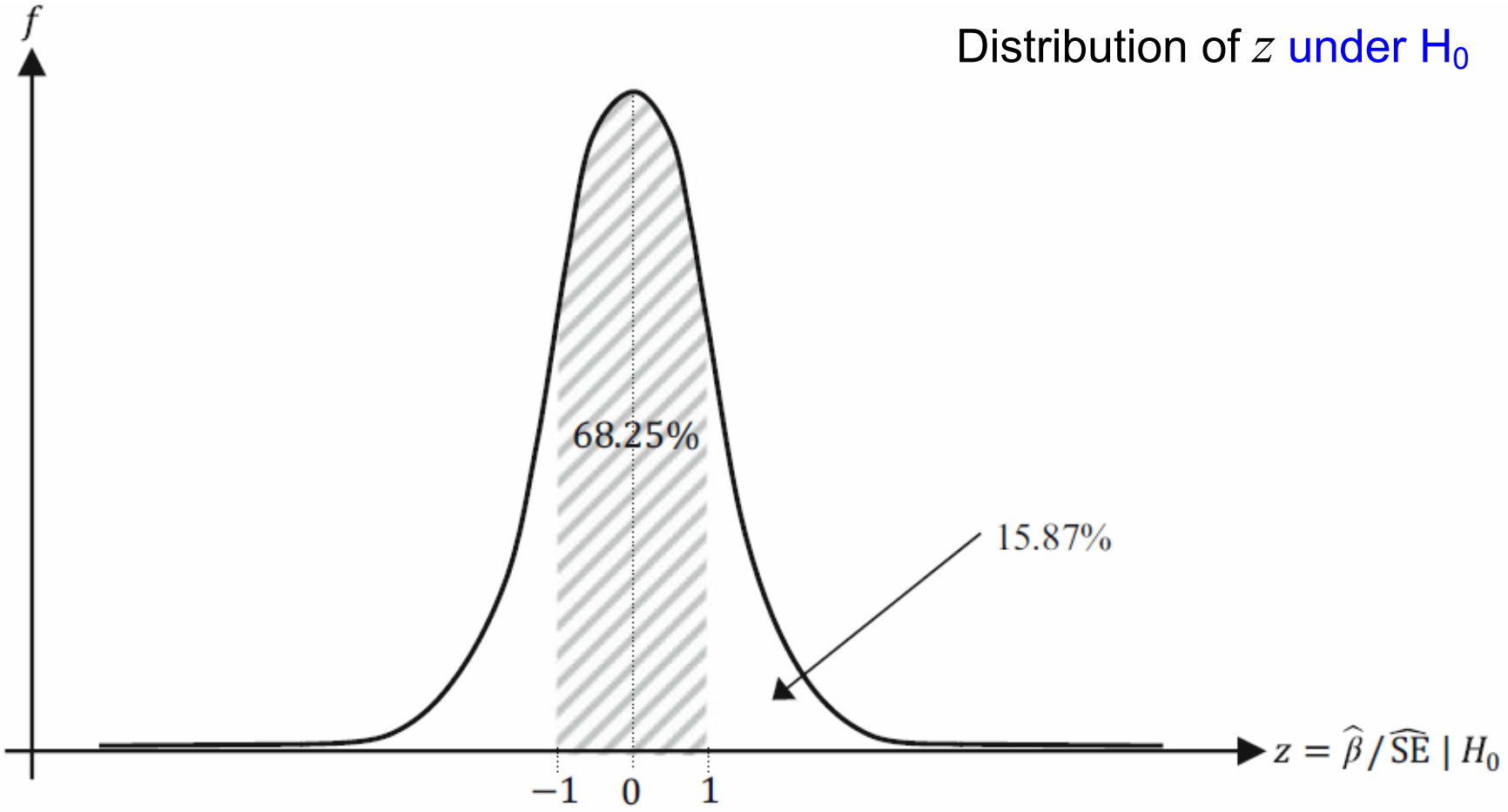
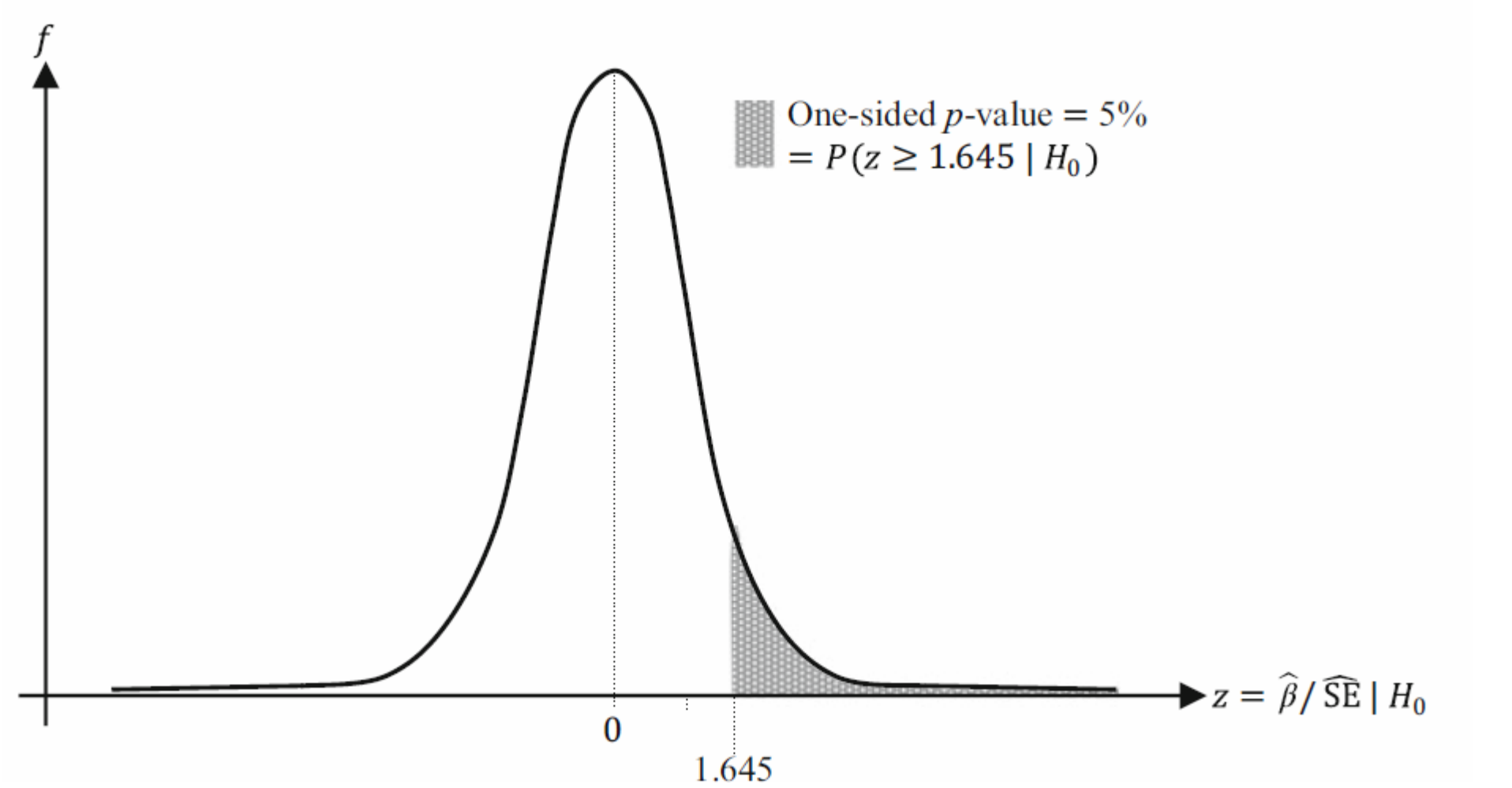


Fig. 6.1 Sampling distribution of a slope estimate $\widehat{\beta}$ (upper part) and distribution of the z-ratio (lower part) conditional on the null hypothesis $H_0: \beta_{H_0} = 0$



(Hirschauer et al. 2022)

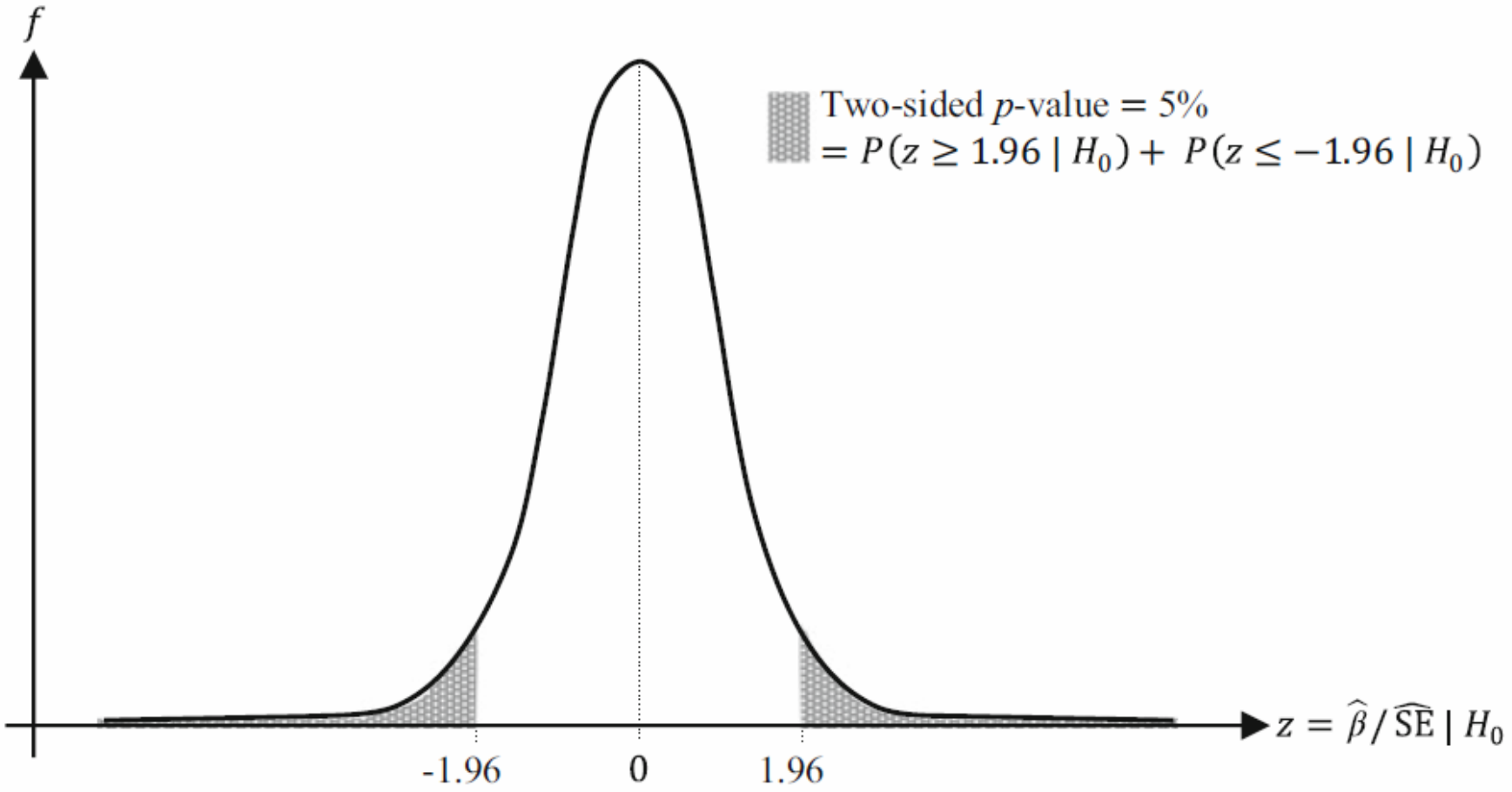


Fig. 6.2 One-sided and two-sided p-value of 5% derived from the distribution of the signal-to-noise ratio $z = \widehat{\beta}/\widehat{\mathrm{SE}}$ under the null hypothesis H_0: $\beta_{H_0} = 0$

p-value

Probability of observing z as large as in this study, or larger, assuming H_0 is true

This is a *conditional* probability

$\Rightarrow$ If p-value is small (e.g. ≤ 0.05), we reject H_0

Criticism

- p-value is just a one-to-one function of the signal-to-noise ratio z
- Better report signal (effect size) $\hat{\beta}$ and noise (precision) SE separately
- p-value is a conditional probability that assumes H_0 to be true, *but* does *not* say anything about the probability that H_0 is true
- A non-significant test does *not* prove the H_0 to be true
  ("Absence of evidence is not evidence of absence", Altman and Bland 1995)
- Significance ≠ Relevance
- *p*-hacking

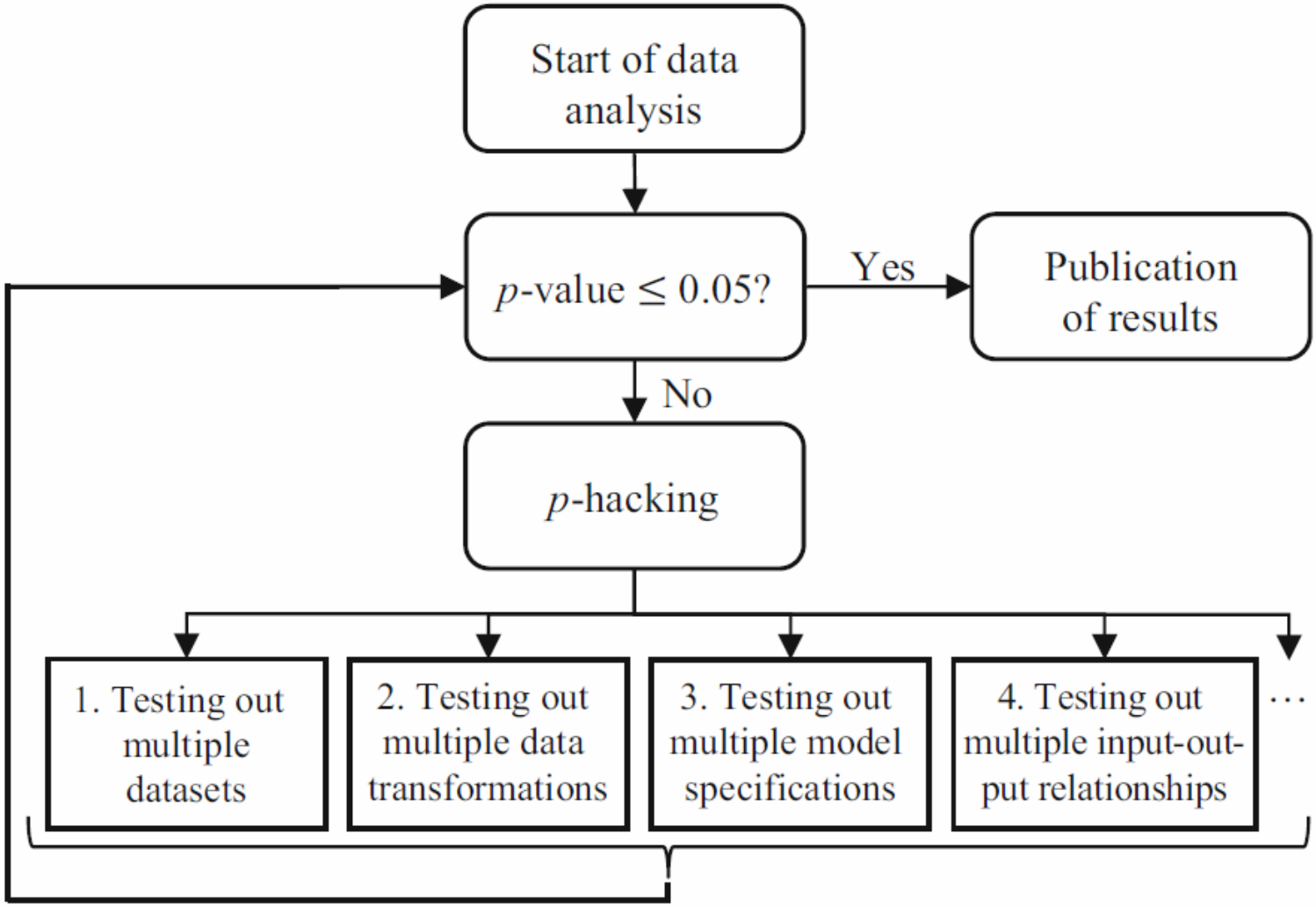


Fig. 6.4 The manifold possibilities of *p*-hacking. Source: Own representation based on Motulsky (2014: Fig. 1).

(Hirschauer et al. 2022)

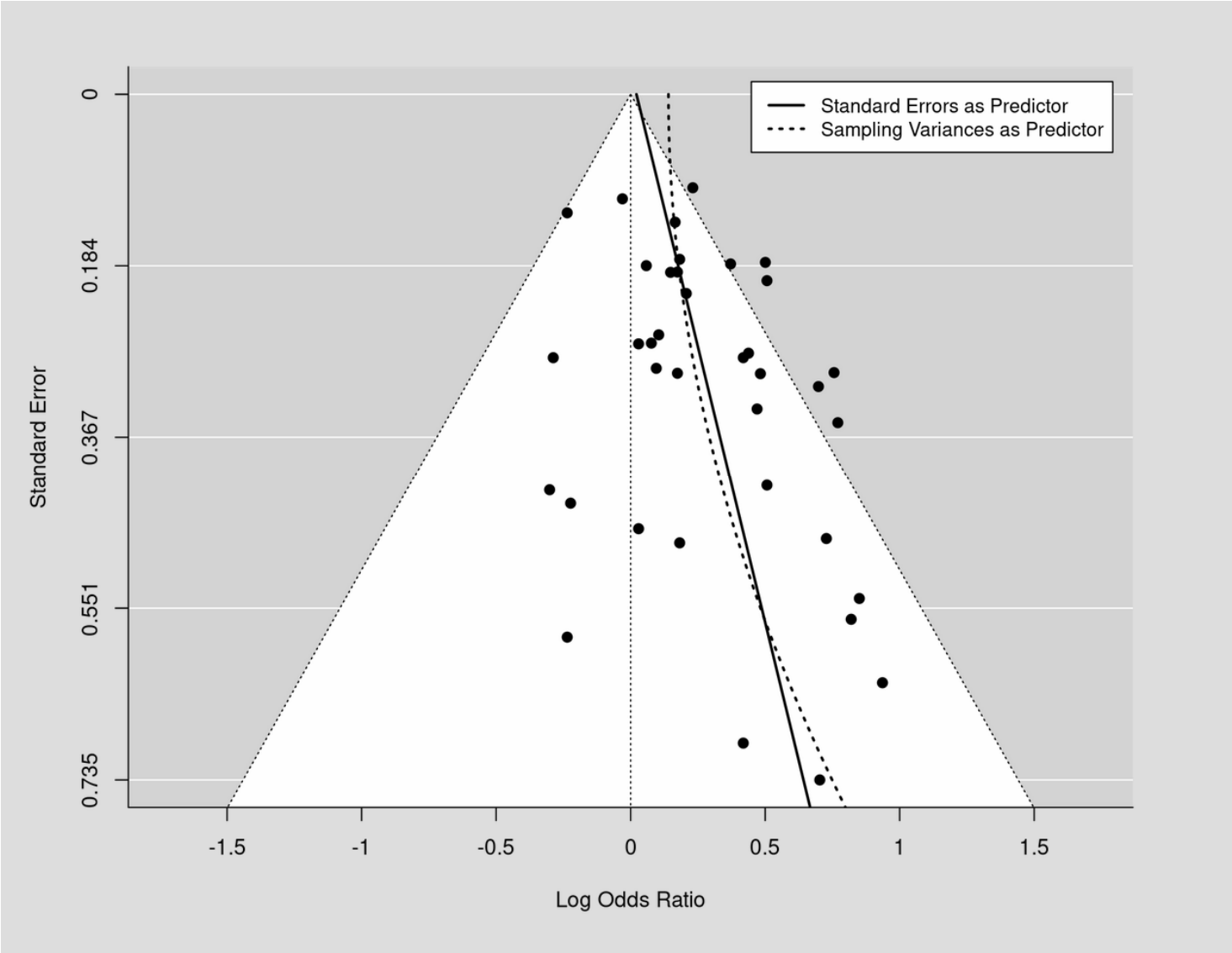


Figure: A funnel plot

$\Rightarrow$ publication bias

(www.metafor-project.org)

Table 7.1 Inferential cases in mean comparisons depending on data generation

	Non-randomization: subjects' membership in pre-existing groups (e.g., male/female) is observed	**Randomization:** subjects are randomly assigned to experimental treatment groups (RCTs)
Non-random selection of study subjects (convenience sample)	(1) *Neither generalizing inference nor causal inference* can be supported by *statistics*	(3) *Causal inference* regarding a treatment effect can be supported by *statistics*
Random selection of study subjects (random sample)	(2) *Generalizing inference* from the sample to the population can be supported by *statistics*	(4) *Causal inference and generalizing inference* can be supported by *statistics*

(Hirschauer et al. 2022)

Main recommendation by Hirschauer et al. (2022)

Go back to basics

$\Rightarrow$ just report point estimates $\hat{\beta}$ and standard errors (*SE*)

Confidence intervals[1]

Example: linear regression, slope parameter β

$$\hat{\beta} \pm t \times SE$$

Sample estimator $\hat{\beta}$

SE = standard error

t = (1−α/2)×100%-quantile of t-distribution

$\Rightarrow$ (1−α)×100% coverage probability

---

[1] Hirschauer et al. (2022) only mention CI once, in a footnote on page 67

Link between confidence intervals and significance tests

If the $(1-\alpha)\times 100\%$ confidence interval (CI)

$$\hat{\beta} \pm t \times SE$$

covers $\beta = 0$, then H_0 cannot be rejected at significance level α

Example: Maize pollen deposition

$$\hat{\beta} = -0.585$$
$$SE = 0.092$$

$$\text{CI: } -0.0585 \pm 2 \times 0.092 \quad \Rightarrow \quad (-0.769;\ -0.401)$$

Tests of equivalence

One key point of criticism of NHST: Can't prove H_0: $\beta = 0$

The way out: Prove that β is "close" to 0

$\Rightarrow$ Reverse roles of H_0 and H_A

H_0: β is *not* close to 0 $(|\beta| \geq \delta)$
H_A: β *is* close to 0 $(|\beta| < \delta)$

$\Rightarrow$ Two one-sided tests (TOST; Schuirmann 1987)

$\Rightarrow$ Show that $(1-2\alpha)100\%$ confidence Interval for β is inside interval $(-\delta, +\delta)$

$\Rightarrow$ Plan sample size accordingly (Piepho et al. 2022)

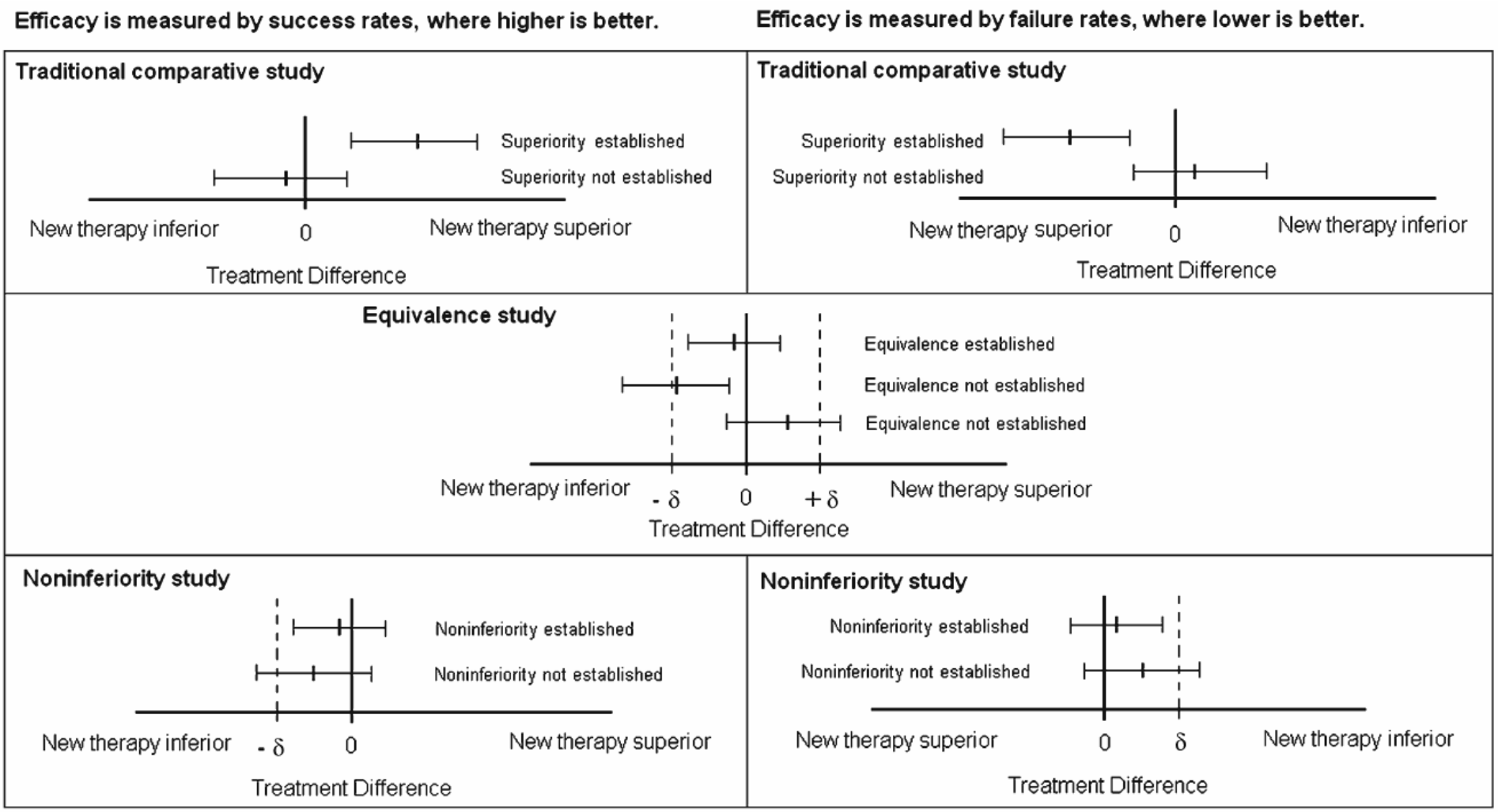


Figure 1. Two one-sided test procedure (TOST) and the equivalence margin in equivalence/noninferiority testing.

(Walker & Nowacki 2011)

Model selection

p-values are often used for model selection (variable selection)

There are alternatives, such as

- Use of information criteria (Akaike Information Criterion, AIC; Bayesian Information Criterion, BIC)
- Model averaging / multi-model inference

(Burnham & Anderson 2002; Heinze et al. 2018)

Bayesian methods

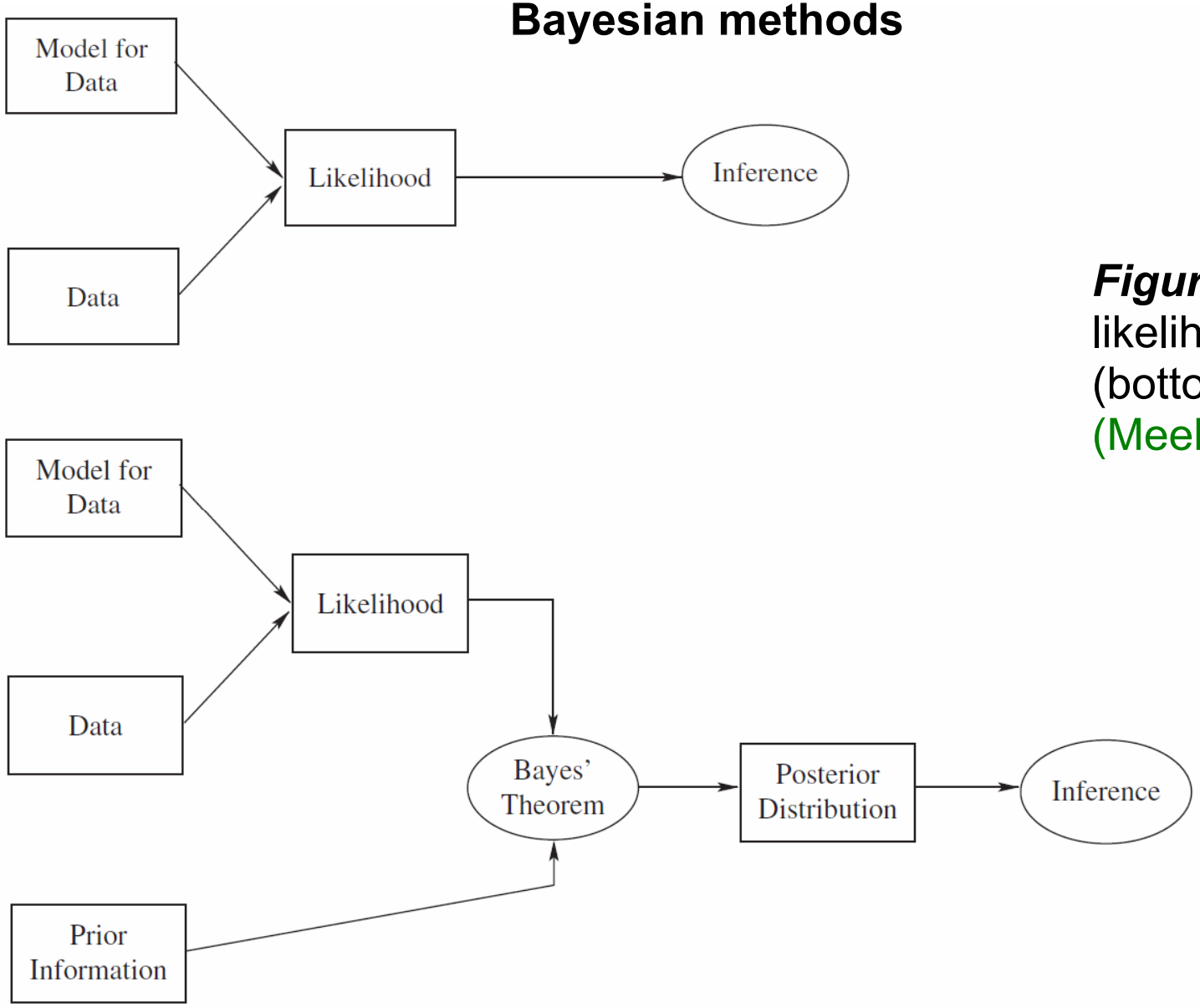
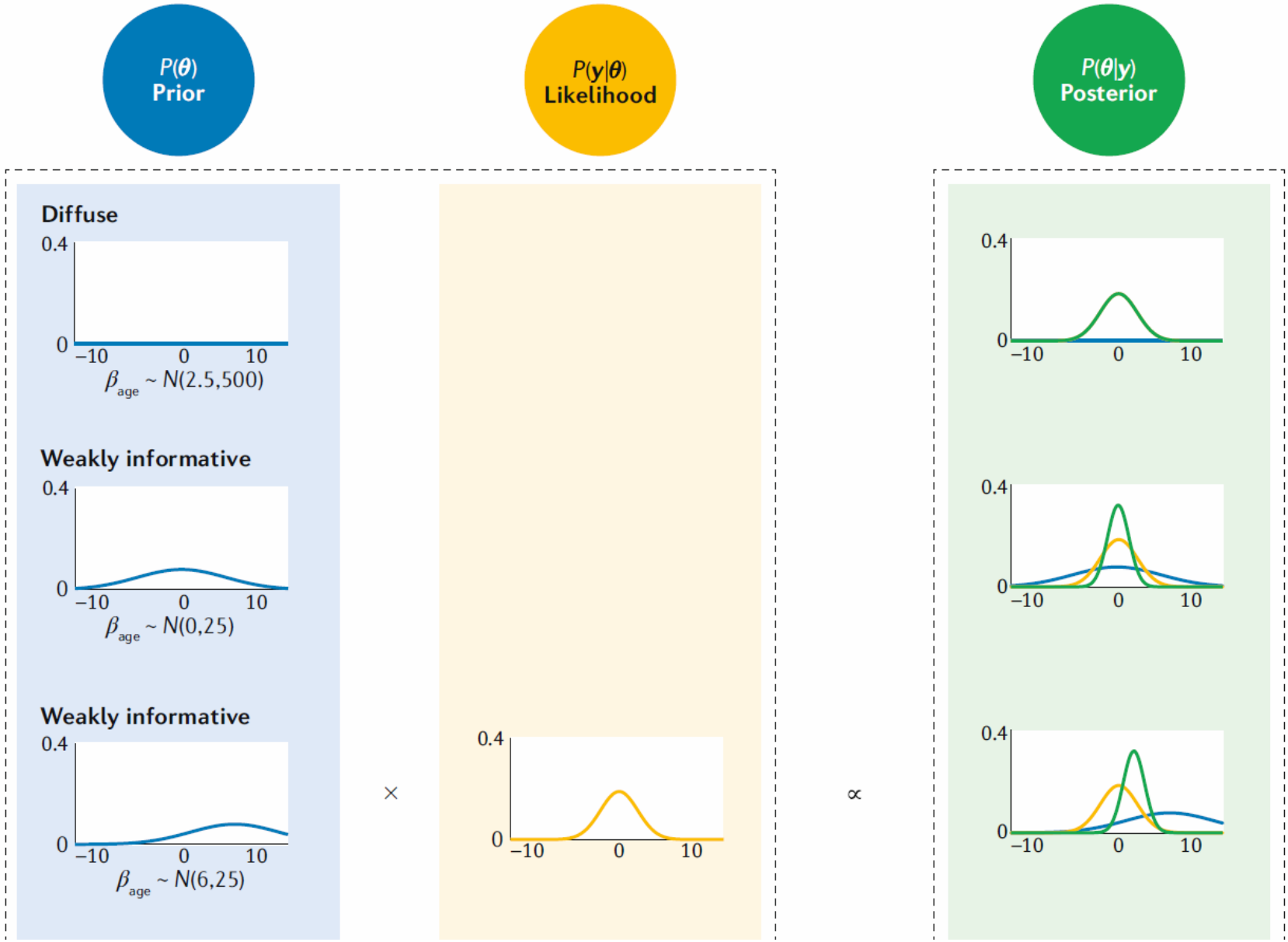


***Figure*: Comparison of likelihood (top) and Bayesian (bottom) inference methods (Meeker et al. 2017).**



(van de Schott et al. 2021)

Conclusion

- Much of the criticism raised against the common usage p-values is valid
- Abandoning p-values altogether is throwing out the baby with the bath water
- Better increase awareness of the meaning and limitation of p-values
- Emphasize importance of sample size and design
- Focus on point estimates and standard errors is good but leaving it at that is not usually sufficient

  $\Rightarrow$ confidence intervals
  $\Rightarrow$ equivalence tests
  $\Rightarrow$ model selection criteria such as AIC
  $\Rightarrow$ go Bayesian (great but not so easy)

Table: Summary of options mentioned

Method		**A good reference**		**An R-package**
p-values (NHST[$])		(your favourite classical text)		
Confidence intervals		Meeker et al. (2017)		‘multcomp’
Equivalence test		Walker & Nowacki (2011)		‘equivalence’
Multi-model inference		Burnham & Anderson (2002)		‘MuMIn’
Bayesian methods		van de Schott et al. (2021)		‘rjags’, ‘stan’

$ Null Hypothesis Significance Testing

References

Altman DG, Bland JM 1995 Absence of evidence is not evidence of absence. BMJ 311: 485-486. https://doi.org/10.1136/bmj.311.7003.485

Burnham KP, Anderson DR 2002 Model selection and multimodel inference. A practical information-theoretic approach. 2nd edition. Springer. https://link.springer.com/book/10.1007/b97636

Heinze G, Wallish C, Dunkler D 2018 Variable selection – A review and recommendations for practicing statisticians. Biometrical Journal 60:431-449. DOI: 10.1002/bimj.201700067

Hirschhauer N, Grüner S, Mußhoff O 2022 Fundamentals of statistical inference. What is the meaning of random error. Springer. https://link.springer.com/chapter/10.1007/978-3-030-99091-6_1

Meeker WQ, Hahn GJ, Escobar LA 2017 Statistical intervals. 2d edition. Wiley. https://www.wiley-vch.de/de/fachgebiete/mathematik-und-statistik/statistical-intervals-978-0-471-68717-7

Piepho HP, Gabriel D, Hartung J, Büchse A, Grosse M, Kurz S, Laidig F, Michel V, Proctor I, Sedlmeier JE, Toppel K, Wittenburg D 2022 One, two, three: Portable sample size in agricultural research. Journal of Agricultural Science 160:459-482. https://doi.org/10.1017/S0021859622000466

van de Schott et al. 2021 Bayesian statistics and modelling. Nature Reviews Methods Primers 1:1. https://doi.org/10.1038/s43586-020-00001-2

Walker E, Nowacki AS 2011 Understanding equivalence and noninferiority testing. Journal of General Internal Medicine 26:192-196. 10.1007/s11606-010-1513-8