Servan Grüninger, MSc Biostatistics

Weight of Statistical Evidence

Detection and Correction of Publication Bias

Master Project (2019), Mathematics Section

to achieve the university degree of

Master of Science in Computational Science and Engineering

submitted to

École polytechnique fédérale de Lausanne

Supervisor

Prof. Dr. Stephan Morgenthaler

Chair of Applied Statistics, EPFL

Écublens, June 21 2019

This document was written with Overleaf and is set in Palatino, compiled with pdfLaTeX and `Biber`.

It is an adapted version of the KOMA-script-based template created by Karl Voit. The template can be found online on Github.

Abstract

Assessing the statistical evidence of scientific findings is challenging. Firstly, the construction of robust evidence measures can be challenging and often hinges on a range of theoretical assumptions that might not be fulfilled in practice. Secondly, there are many different procedures to construct evidence measures which makes comparisons across studies difficult. Thirdly, the landscape of publication in science is heavily distorted by non-scientific incentives which gives rise to so-called 'publication bias' and thus makes aggregation of evidence across studies even more challenging.

In the first part of this thesis, I describe different methods to construct robust statistical evidence measures based on the idea of variance stabilising transformations. I then use these estimators in the second part to analyse and improve various methods for the detection and correction of publication. I used theoretical arguments in combination with results from simulations to show that the construction of robust evidence measures remains challenging and how statistical methods for the detection and correction of biases in scientific findings can be further improved.

Acknowledgements

I thank my supervisor, Prof. Dr. Stephan Morgenthaler, for his guidance and feedback in the course of this thesis as well as the intellectual freedom I enjoyed. In addition, I thank the Swiss Study Foundation, and the Werner Siemens Foundation for their financial and non-material support. It has empowered me and given me the freedom to conduct my studies the way I wanted. Finally, I want to thank my fiancée for the many enthralling discussions about the philosophical aspects of scientific evidence.

Contents

List of Figures

List of Tables

1. Introduction

'I've been studying statistics for over 40 years & I still don't understand it. The ease with which non-statisticians master it is staggering.'

— Stephen Senn, Twitter, (November 27, 2014)

Complaints about statistics (and statisticians) are legion. In the best case, scientists simply find it boring, tedious and obfuscated, but recognise its worth in uncovering scientific facts. In the worst case, they see it as a mere means to an end, the end often being significant findings or findings going into the 'desired direction'. Everyone wants evidence, but nobody wants to work for it—or at least collaborate with the ones who are willing to work the stats.

This might seem a bit caustic but it exemplifies the reactions that statistics usually evokes. And it it should make clear that many of the problems of scientific research—including the eponymous bias of this thesis—have less to do with statistics and are more deeply rooted in distorting external and internal incentives. Hence, it would be short-sighted to believe that statistical methods will be able to solve problems in research practice without major changes in research culture and specifically the current reward structures within scientific research.

Nevertheless, it is crucial to provide a solid statistical basis for the analysis and interpretation of scientific evidence in the face of uncertainty. Therefore, Chapter 2 provides an overview of basic methods to test hypotheses and compare statistical evidence from different sources, primarily focusing on the idea of variance stabilising transformations and their advantageous properties for testing and comparing scientific findings.

Whereas these methods primarily focus on how to calculate comparable evidence measures from a single study, the remainder of the second part of the thesis is concerned with how to combine statistical evidence from multiple studies in the presence of bias. This is crucial to draw the correct

inferences about statistical parameters, which is, as William Sealy Gosset puts it in his seminal paper 'The probable error of a mean' (Student, 1908, p. 1), the main purpose of scientific experiments: 'Any series of experiments is only of value in so far as it enables us to form a judgement as to the statistical constants of the population to which the experiments belong'.

It follows then that incomplete and biased publication records heavily undermine the value of scientific research because they prevent a reliable 'judgement as to the statistical constants of the population'. Therefore, Chapter 3 introduces the reader to a variety of sources behind publication bias in general and significance-driven publication bias in particular (Section 3.1).

In addition, I present a selection of methods to detect (see Section 3.2) and correct (see Section 3.3) publication bias.

Finally, Chapter 4 gives the reader a quick overview of additional steps and methods to be used for the detection and correction of publication bias.

2. Analysing and Testing Evidence

'Use the CRS database to size the market.' – 'That data is wrong.'
'Then use the SIBS database.' – 'That data is also wrong.'
'Can you average them?' – 'Sure. I can multiply them too.'

— Dilbert by Scott Adams May 07, 2008

Good statistics is like strong coffee: bitter at times but always sobering. And there are few things more sobering than taking a cherished hypothesis and putting it to the test—especially, if the test turns out to be negative. If used correctly, rigorous statistics is the ultimate tool for the 'skillful interrogation of Nature', as Fisher Box (1978, p. 140) puts it: It helps to uncover relationships and connections which do not immediately catch the eye and prevents over-excited researchers from clinging to spurious findings.

Clever designs and nimble mathematical manipulations can make complex problems much more accessible and prevent common pitfalls. The next few sections are dedicated to a short overview of the basics of some of the many methods used for analysing and testing statistical evidence stemming from scientific data.

2.1. Hypothesis testing

Researchers often have to address questions such as the following:

Is treatment A superior to treatment B?

In order to address this question, one usually formulates two distinct hypotheses:

H_0 : Treatment A is not superior to treatment B.
H_1 : Treatment A is superior to treatment B.

To translate these hypotheses into a mathematically feasible language, one needs to operationalise them by establishing the criteria for 'superiority' of a treatment. This can be done by defining primary (and potentially secondary) outcomes that measure the treatment effect of treatments A and B. What these primary outcomes are should be decided based on the specific question at hand and the expert knowledge of the involved researchers.

In other words, one must define a parameter θ with which one can quantify 'superiority' and that allows us to reformulate the hypothesis as follows:

$$H_0 : \theta_A \leq \theta_B$$
$$H_1 : \theta_A > \theta_B$$

For illustrative purposes, let us assume that we are dealing with two different cancer treatments. One primary outcome could be the proportion of patients still alive after a given time period. This example is assessed in Section 2.3. Another primary outcome could be the average survival time of the patients after receiving the treatment. This example is further commented on in Section 2.4.

2.1.1. Types of hypotheses and hypothesis tests

Before I can address the question of how to test these hypotheses, I first need to make a small detour to delve deeper into the specific properties a hypothesis can have. Above—and for the remainder of this Master thesis—I let the hypotheses in question be 'composite'. However, hypotheses can also be 'simple', with the distinction lying in the number of possible values for θ that a hypothesis specifies.

A hypothesis is called simple if it uniquely identifies the parameter θ (or the probability distribution specified by θ). Conversely, if a hypothesis states more than one possible parameter value, it is called composite because it is composed of multiple simple hypotheses.

For illustration, let us consider a set of hypotheses that is different from the one stated above:

$$H_0 : \theta = \theta_0 = 0$$
$$H_1 : \theta > \theta_0 = 0$$

In this case, the null hypothesis H_0 is simple: For it to be true, θ can only correspond to a single value. Conversely, the alternative hypothesis H_1 is composite because it specifies more than one value for θ: H_1 is true for any $\theta > 0$, regardless of whether this corresponds to $\theta = 10^{-3}$ or $\theta = 10^{3}$.

Now, in order to test the hypotheses, be they simple or composite, there are, in principle, three options:

- Perform a one-sided test for superiority: Is θ larger than θ_0?
- Perform a one-sided test for inferiority: Is θ smaller than θ_0?
- Perform a two-sided test: Is θ either smaller or larger than θ_0?

These tests can be performed for three different scenarios: for two simple, a simple and a composite or two composite hypotheses. For the remainder of this Master thesis, I will let both H_0 and H_1 be composite hypotheses and explain all methods using the example of a one-sided test for superiority.

2.1.2. Test statistics and decision functions

To test a specific set of hypotheses, one needs to construct a test statistic $S_n(\cdot)$, that is, a function that takes as input random variable $X_1 \dots X_n$ and provides as output another random variable whose realisation serves to distinguish between H_0 and H_1 given a decision function $\delta(\cdot)$. Note that the test statistic can but does not have to be an estimator of θ. For example, the exact binomial test described in Section 2.3.1 is based on a test statistic that is at the same time an estimator of θ. In many cases, however, a test statistic consists of a transformation of the estimator, see for example the Z-statistic in Section 2.3.3 or the Student's t-statistic in Section 2.4.1.

To perform a one-sided test for superiority on the set of hypotheses described at the beginning of this chapter, I define the following decision function:

$$\delta(S_n(X_1, \dots, X_n)) = \begin{cases} 1, & \text{if } S_n(X_1, \dots, X_n) > q; \\ 0, & \text{otherwise.} \end{cases}$$

If the output of $\delta(S_n)$ is 1, H_0 is rejected, if it is 0, the test fails to reject H_0. The exact value of the decision threshold q depends on the test statistic in question as well as on the pre-defined Type I error rate α—that is, the false rejection rate of H_0—that one is willing to accept on average (see Figure 2.1 for a visualisation).

The goal is then to find the smallest value of q for which the significance level is equal to α, that is,

$$\begin{aligned} & \Pr(\delta = 1 \mid H_0) = \Pr(S_n(X_1, \ldots, X_n) > q \mid H_0) = \alpha \\ \iff & 1 - F_{S_n|H_0}(q) = \alpha \\ \iff & F^{-1}_{S_n|H_0}(1-\alpha) = q \end{aligned}$$

where $F_{S_n|H_0}(\cdot)$ and $F^{-1}_{S_n|H_0}(\cdot)$ are the cumulative distribution function and the quantile function, respectively, of S_n under the null hypothesis and $\Pr(\delta = 1 \mid H_0)$ is the probability of the decision function $\delta(\cdot)$ returning 1 under the null hypothesis. Both q and α are often referred to as the 'significance threshold' because they set the threshold for a test statistic or its corresponding cumulative tail probability, respectively, to pass.

2.1.3. The power of a statistical test

If a specific significance threshold α is defined, we can calculate the Type II error rate $\beta = \Pr(\delta = 0 \mid H_1)$ of a test, that is, the probability of not rejecting the null hypothesis in the presence of a real effect (see Figure 2.1):

$$\beta = \Pr(\delta = 0 \mid H_1) = \Pr(S_n(X_1, \ldots, X_n \mid H_1) < q) = F_{S_n|H_1}(q).$$

Here, $F_{S_n|H_1}(\cdot)$ denotes the cumulative distribution function of S_n under the alternative hypothesis. From this, we can calculate the power of a hypothesis test, which is given by $1 - \beta = 1 - F_{S_n|H_1}(q)$. Hence, the power of a test is simply the cumulative probability of a test statistic S_n being larger than its significance threshold, q_α assuming the alternative hypothesis H_1 is true.

2.1.4. The p-value

For a given observation x, the p-value is defined as the probability of observing a value that is at least as extreme as x given that the null hypothesis H_0 is true. The 'extremity' of x is usually assessed by calculating the cumulative probability of randomly drawing values from the null distribution that are at least as far away from the distribution mean as x. For one-sided directional hypotheses, the p-value is thus defined as

$$p = \begin{cases} \Pr(X < x \mid H_0), & \text{if testing for inferiority;} \\ \Pr(X > x \mid H_0), & \text{if testing for superiority.} \end{cases}$$

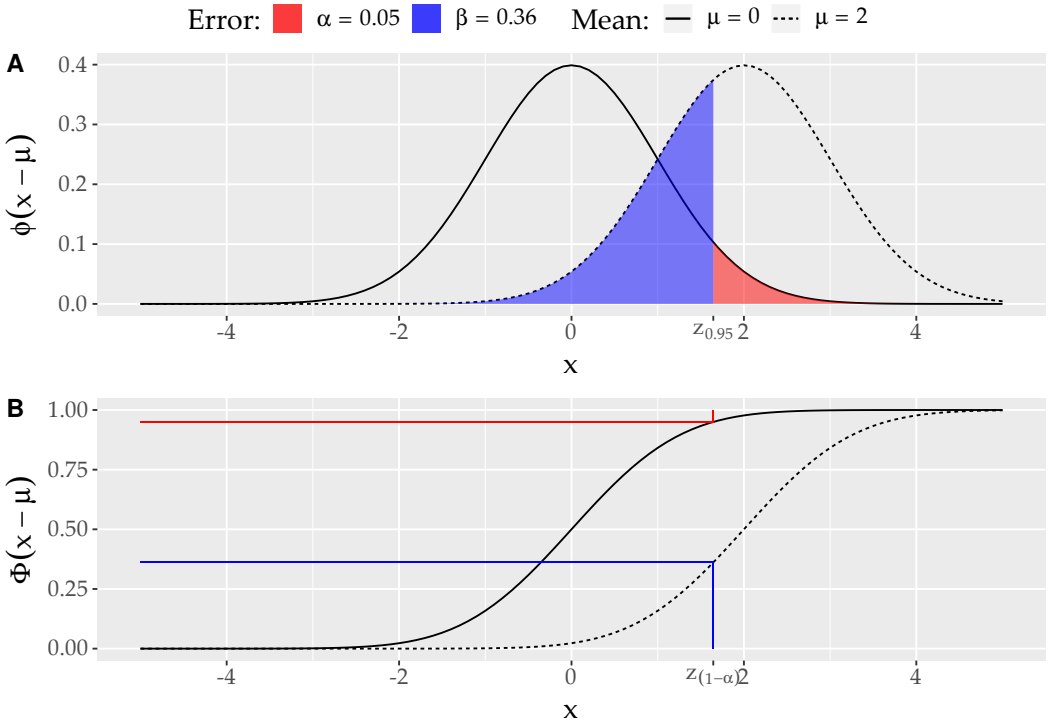

Figure 2.1.: (A) The probability density functions of two normal distributions with variance 1 and mean μ according to the null ($H_0 : \mu = 0$) and alternative hypothesis ($H_1 : \mu = 2$), respectively. The critical value $z_{1-\alpha}$ denotes the $(1-\alpha)$-quantile of the standard normal distribution, above which the null hypothesis is rejected. (B) The cumulative density functions of the same two normal distributions are displayed in panel (A). Under H_0, the curve crosses the critical value $z_{1-\alpha}$ where $\Phi(x - \mu_0) = 1 - \alpha$. Under H_1, the curve crosses the threshold at $\Phi(x - \mu_1) = \beta$.

For two-sided non-directional hypotheses—and if the null distribution is symmetric—the p-value can simply be calculated by setting $x = |x|$ and then doubling the one-sided p-value for a superiority test, that is, $p = 2 \cdot \Pr(X > |x| \mid H_0)$.

If the null distribution is not symmetric, however, the calculation of the two-sided p-value does not immediately follow from its definition. One solution is to double the minimum of the p-values of both one-sided hypotheses, that is,

$$p = 2 \cdot \min(\Pr(X > x \mid H_0), \Pr(X < x \mid H_0)).$$

Another is to 'mirror' x around a central tendency ct of the distribution (for example the mean or the median) and then sum up the p-values of both

one-sided hypotheses:

$$p = \Pr(X < ct - |x| \mid H_0) + \Pr(X > ct + |x| \mid H_0).$$

Given a continuous probability distribution under a simple null hypothesis, the p-value is uniformly distributed between 0 and 1. This can be shown by simulations or by the following proof[1].

To show that P is uniformly distributed, I will use the probability integral transform, which states that for a continuous cumulative distribution function $F_X(x)$, the random variable $Y = F_X(X)$ is uniformly distributed as $Y \sim \text{Unif}(0,1)$, that is, $\Pr(Y < y) = y$ for $y \in [0,1]$ (see for example Casella and Berger, 2002, p. 54f.).

Let $P = \Pr(X > S_n \mid H_0) = 1 - F_{H_0}(S_n)$ where $F_{H_0}(\cdot)$ denotes the cumulative distribution function of $S_n \sim \mathcal{N}(0, \sigma^2)$. Hence, we can write

$$\begin{aligned}
\Pr(P \leq p) &= \Pr(1 - F_{H_0}(S_n) \leq p) = \Pr(F_{H_0}(S_n) \geq 1 - p) \\
&= \Pr(F_{H_0}^{-1}(F_{H_0}(S_n)) \geq F_{H_0}^{-1}(1-p)) = \Pr(S_n \geq F_{H_0}^{-1}(1-p)) \\
&= 1 - F_{H_0}(F_{H_0}^{-1}(1-p)) = 1 - (1-p) \\
&= p
\end{aligned}$$

which concludes the proof.

For a composite null hypothesis, e. g. $H_0 : \mu < \mu_0 = 0$, uniformity of P only holds with regard to the least favourable value of μ under the null, that is, $\mu = \mu_0 = 0$. For any $\mu < \mu_0$, the distribution of P is left skewed. Similarly, the distribution of P under a (simple or composite) alternative hypothesis is not uniform either, but right-skewed (see for example Murdoch, Tsai and Adcock, 2008).

2.2. What properties should an evidence measure have?

As mentioned in the previous section, the test statistic S_n depends on the set of hypotheses to be tested and the distributional nature of the data. A test

[1] Note that p-values are in fact realisations of a random variable (Murdoch, Tsai and Adcock, 2008). Thus, I will use the uppercase letter P in the following paragraphs to distinguish the random variable from its realisation, even though lowercase letters are usually used for both.

statistic that might be suited for binomially distributed data might lead to false inference if the data are normally distributed. In addition, the values of different test statistics are usually not directly comparable.

To alleviate this, we can transform test statistics S_n into an evidence measure $V_n = h_n(S_n)$ which is also a statistic but for which comparisons are easier to make. Kulinskaya, Morgenthaler and Staudte (2008, p. 115) stated that four desirable properties a one-sided evidence measure V_n should have:

E_1 The one-sided evidence V_n is a monotonically increasing function of the test statistic S_n, that is, $h_n(S_{n_1}) \geq h_n(S_{n_2})$ if $S_{n_1} \geq S_{n_2}$. This ensures that optimal values of S_n are retained by the transformation to V_n and facilitates interpretation.

E_2 The distribution of V_n is normal for all values of the unknown parameters. The normal distribution has a variety of desirable statistical properties which facilitate mathematical manipulation and interpretation.

E_3 The variance equals one ($\text{Var}[V_n] = 1$) for all values of the unknown parameters. Stabilising the variance to 1 allows for direct comparison of different evidence measures.

E_4 The expected evidence $\tau = \tau(\theta) = \text{E}_\theta[V_n]$ is monotonically increasing in θ from $\tau(0) = 0$, that is, larger parameter values θ indicate larger expectations of the evidence measure V_n and vice versa.

For some combinations of distributions and statistics, such as the Z-score based on normally distributed variables (Section 2.4), all properties E_1 to E_4 are exactly fulfilled. For others, such as the Z-score based on binomially distribute variables (Section 2.3.3), the properties are only fulfilled asymptotically. Hence, for most combinations of distributions and statistics we need to find a transformation $V_n = h_n(S_n)$, so that V_n fulfils properties E_1 to E_4 stated above. To fulfil E_1, any monotonically increasing function will do. To ensure that E_2, E_3, and E_4 are met, however, slightly more effort is needed.

2.2.1. E_2: Making sure everything is normal

To meet criterion E_2, one usually needs to resort to a distribution-dependent transformation that turns S_n into a normally distributed variable. If S_n follows a lognormal distribution—the simplest case—taking the natural logarithm $\ln(S_n)$ yields a normally distributed variable.

If S_n follows a different continuous distribution, an approximate solution can be found by using the transformation $h_n = \phi^{-1}(F_{S_n}(\cdot))$, with ϕ^{-1} denoting the quantile function of the standard normal distribution and F_{S_n} denoting the cumulative distribution function of S_n. Since ϕ^{-1} does not have a closed form description, solutions have to be approximated using numerical methods.

Yet another possibility to ensure E_2 is to exploit the central limit theorem. For example, if $S_n \sim \text{Bin}(n, p)$ and p fixed, then $S_n \overset{n\to\infty}{\sim} \mathcal{N}(np, np(1-p))$ (deMoivre-Laplace theorem; see Section 2.3.3 for more details on the usefulness of the central limit theorem in this context).

2.2.2. E_3: The variance stabilising transformation

To ensure that property E_3 is given, we need to make sure that h_n is variance stabilising, so that $\text{Var}[h_n(S_n)] = 1$. If the variance of S_n can be expressed in terms of its expectation passed to a known function g_n, that is, $\text{Var}[S_n] = g_n(\text{E}[S_n])$, then h_n is defined—up to an additive constant—by

$$h_n(s_n) = \int^{s_n} [g_n(t)]^{-1/2} dt$$

if the indefinite integral exists (Kulinskaya, Morgenthaler and Staudte, 2008, p. 126–127). Hence,

$$\{h_n'(\text{E}[S_n])\}^2 = \{g_n(\text{E}[S_n])\}^{-1} = \{\text{Var}[S_n]\}^{-1}. \tag{2.1}$$

For $\text{Var}[S_n]$ to be small enough, we can use a first order Taylor approximation around $\text{E}[S_n]$ to approximate $\text{Var}[h_n(S_n)] = \text{Var}[V_n]$. This is done as follows:

$$V_n = h_n(\text{E}[S_n]) + (S_n - \text{E}[S_n])h_n'(\text{E}[S_n]) + R_1.$$

Calculating the variance on both sides of the equation yields

$$\begin{aligned}\mathrm{Var}[V_n] &= \mathrm{Var}[h_n(\mathrm{E}[S_n]) + (S_n - \mathrm{E}[S_n])h_n'(\mathrm{E}[S_n]) + R_1] \\ &= \mathrm{Var}[S_n h_n'(\mathrm{E}[S_n]) + R_1] \\ &\simeq \mathrm{Var}[S_n]\{h_n'(\mathrm{E}[S_n])\}^2. \end{aligned} \tag{2.2}$$

Combining Eq. 2.1 with Eq. 2.2 yields $\mathrm{Var}[V_n] \simeq 1$. Even though finding a variance stabilising function h_n might seem straightforward in theory, it is usually more tedious in practice, because it often depends on unknown parameters, as Kulinskaya, Morgenthaler and Staudte (2008, p. 127) point out.

2.2.3. E_4: Monotonically increasing expectation of evidence

E_4 requires that $\tau(\theta) = \mathrm{E}_\theta[V_n]$ is monotonically increasing in θ, starting from $\tau(0) = 0$. In order to find the expectation of V_n, we can again resort to a Taylor approximation around $\mathrm{E}[S_n]$, but this time expanding the series up to the second order:

$$\begin{aligned} V_n = h_n(\mathrm{E}[S_n]) &+ (S_n - \mathrm{E}[S_n])h_n'(\mathrm{E}[S_n]) \\ &+ (S_n - \mathrm{E}[S_n])^2 \frac{h_n''(\mathrm{E}[S_n])}{2} + R_2. \end{aligned}$$

Taking the expected values on both sides yields

$$\begin{aligned}\mathrm{E}[V_n] &= h_n(\mathrm{E}[S_n]) + \mathrm{Var}(S_n)\frac{h_n''(\mathrm{E}[S_n])}{2} + R2 \\ &\simeq h_n(\mathrm{E}[S_n]) + \mathrm{Var}(S_n)\frac{h_n''(\mathrm{E}[S_n])}{2}. \end{aligned}$$

In order to make sure that $\tau(0) = 0$, we can subtract $h_n(\mathrm{E}[S_n \mid H_0]) = h_n(\theta_0)$ from V_n.

2.2.4. The key inferential function

If an evidence statistics V_n fulfils all criteria E_1 to E_4, it is often possible to rewrite $\tau(\theta) = \mathrm{E}_\theta[V_n - h_n(\theta_0)]$ as $\tau(\theta) = \sqrt{n}K_{\theta_0}(\theta)$, with K_{θ_0}—the so-called

'key inferential function' of θ with respect to a constant θ_0— independent of the sample size n. As outlined in Kulinskaya, Morgenthaler and Staudte (2008, p. 127–128), K is a useful tool to solve some routine problems which arise in statistical testing, such as:

Sample size calculation: Recall the hypotheses stated in Section 2.1.1 in which we wanted to test $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$. To obtain a desired expected evidence τ_1 in favour of a specific θ_1, we need to find

$$n_1 \geq \left(\frac{\tau_1}{K_{\theta_0}(\theta_1)} \right)^2 .$$

Power calculation: The power of a test with significance threshold α and $z_{1-\alpha}$ denoting the $(1-\alpha)$-quantile of the standard normal distribution is given by

$$\begin{aligned} 1 - \beta &= \Pr(V_n \geq z_{1-\alpha} \mid H_1) && (2.3) \\ &= \Phi(\tau_1 - z_{1-\alpha}) && (2.4) \\ &= \Phi(\sqrt{n}\{K_{\theta_0}(\theta_1)\} - z_{1-\alpha}). && (2.5) \end{aligned}$$

Confidence intervals: Confidence intervals can easily be found by using the inverse of the key inferential function K^{-1}:

$$\left[K_{\theta_0}^{-1}\left(\frac{V_n - z_{1-\alpha/2}}{\sqrt{n}} \right), K_{\theta_0}^{-1}\left(\frac{V_n - z_{1-\alpha/2}}{\sqrt{n}} \right) \right]. \qquad (2.6)$$

2.3. Statistical evidence in binomial variables

Recall the set of hypotheses stated at the beginning of this chapter. Let A and B be two different cancer treatment whose efficacy we want to assess, and let us assume that the primary outcome of interest is the survival rate of the patients; that is, we want to know the proportion of patients p who are still alive twelve months after undergoing one of the two treatments.

We can describe the survival of an individual patient after twelve months as a Bernoulli distributed random variable $L \sim \text{Ber}(p)$ where p is the survival probability and $L \in \{0, 1\}$. The null and alternative hypothesis can then be formulated as

$$\begin{aligned} H_0 &: p_A \leq p_B \\ H_1 &: p_A > p_B \end{aligned}$$

where p_A and p_B are the surviving probabilities of participants receiving treatment A and B, respectively. For the sake of the argument, let us assume that treatment B is already well established and leads to a well-known survival probability p_B twelve months after the treatment. Hence, we only need to estimate p_A to perform our hypothesis test. We can do so by randomly assigning treatment A to n participants and to then count the number of participants m who are still alive twelve months later[2]. Then, we can use this estimate to perform an exact binomial test.

2.3.1. The exact binomial test

Since the sum of independently and identically distributed $\text{Ber}(p)$-random variables is a binomial random variable $M = \sum_{i=1}^{n} L_i \sim \text{Bin}(n, p)$, we know that m is a realisation of the binomial random variable $M \sim \text{Bin}(n, p_A)$. Hence, the ratio $\hat{p}_A = m/n$ serves as an estimate of the twelve-month survival probability of the participants who received treatment A.

To assess whether the ratio of surviving participants is greater in the group that received treatment A than in the group that received treatment B, we need a test for statistic S_n. In our simplified case, this is simply the total number of surviving participants, hence $S_n(L_1, \ldots, L_n) = M = \sum_{i=1}^{n} L_i$.

We can now use this statistic together with the following decision function $\delta(\cdot)$ to perform a significance test:

$$\delta(S_n(L_1, \ldots, L_n)) = \begin{cases} 1, & \text{if } S_n(L_1, \ldots, L_n) > q_{B,1-\alpha}; \\ 0, & \text{otherwise.} \end{cases}$$

Here, $q_{B,1-\alpha}$ denotes the $(1-\alpha)$-quantile of the $\text{Bin}(n, p_B)$-distribution. Hence, if the number of surviving participants who received treatment A is larger than the number of surviving participants that would be expected in at least $(1-\alpha)\%$ of the cases in which n patients received treatment B, we would reject H_0. This is the so-called exact binomial test.

[2] Please note that this example serves illustrative purposes only. It is in no way representative of the way clinical trials are conducted in reality. Both fortunately and unfortunately, real-life clinical trials are more complex endeavours.

2.3.2. Evidence of one-sample binomial tests

Testing hypotheses using the exact binomial test has multiple drawbacks, for example:

1. The variance of a random variable following the binomial distribution is dependent on the expected value p, which makes the comparison of the results of different test statistics S_n non-trivial.
2. The variance of the random variable tends to zero at the extremes of p.
3. The calculation of the critical values as well as other properties of the test, such as the power, is computationally very costly when compared to other test statistics.
4. For each n a different critical value needs to be calculated, even if H_0 remains the same.
5. With increasing n, calculating the cumulative probabilities of a binomial test statistic becomes computationally more taxing even though practically this is unproblematic in most use cases.

Recalling the desirable properties for an evidence measure stated in Section 2.2, it is clear the test statistic of the binomial test clearly violates properties E_2 and E_3, since it follows a binomial distribution and has a non-constant variance for all S_n, respectively. In order to make the test statistic fulfil all four properties, we need to transform it. When the sample size n is large enough, we can use the well-known Z-score or standard score to do so.

2.3.3. The Z-score

The Z-score is defined as follows:

$$Z = \frac{X - \mu}{\sigma}$$

If X is a normally distributed random variable with expectation μ and variance σ^2, (i. e., $X \sim \mathcal{N}(\mu, \sigma^2)$), the Z-score is a random variable following a standard normal distribution ($Z \sim \mathcal{N}(0,1)$), thus fulfilling properties E_1 to E_4.

In order to test a hypothesis, we can simply fix our α-level and compare the Z-score with the critical values defined as $\Phi^{-1}(1-\alpha)$, where $\Phi^{-1}(\cdot)$ is the quantile function of the standard normal distribution.

If the sample size n is large enough and by virtue of the central limit theorem (CLT), we can use the Z-score to transform the binomial test statistic into a test statistic that fulfils properties E_1 to E_4. The Lindeberg-Lévy CLT states that if $X_1, \dots X_n$ are identically and independently distributed random variables with $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$, then $\sqrt{n}(\bar{X}_n - \mu)$ converges in distribution to $\mathcal{N}(0, \sigma^2)$. Hence,

$$Z_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0,1).$$

In our example from above, we have $\mu = p_B$, $\sigma^2 = \hat{p}_A(1-\hat{p}_A)$ under the null hypothesis, and $\bar{X}_n = \bar{L}_n = \sum_{i=1}^{n} L_i/n = M/n = \hat{p}_A$. Hence

$$Z_n = \frac{\sqrt{n}(\hat{p}_A - p_B)}{\sqrt{\hat{p}_A(1-\hat{p}_A)}} \xrightarrow[H_0]{d} \mathcal{N}(0,1).$$

If the null hypothesis does not hold, we have

$$Z_n = \frac{\sqrt{n}(\hat{p}_A - p_B)}{\sqrt{\hat{p}_A(1-\hat{p}_A)}} \xrightarrow[H_1]{d} \mathcal{N}(\sqrt{n}\frac{\hat{p}_A - p_B}{\sqrt{\hat{p}_A(1-\hat{p}_A)}}, 1). \tag{2.7}$$

However, if n is not large enough or if p_A lies too close to either zero or one, a wide range of problems arises:

- The probability distribution of Z_n becomes leptokurtic with positive skew (p_A too close to 0) or negative skew (p_A too close to 1 (see Figure 2.2 and Figure 2.3).
- The empirical mean of Z_n diverges starkly from its theoretical expectation (see Figure 2.4 and Figure 2.5).
- The empirical variance of Z_n is not stabilised anymore and diverges from 1 (see Figure 2.4 and Figure 2.5).
- The Type I error grows beyond the pre-defined α-threshold and thus is not controlled anymore (see Figure 2.6 and Figure 2.7).
- The empirical coverage probabilities of the (1-α)-confidence intervals lies below the nominal probabilities and deteriorates to 0 for p_A going to zero or one (see Figure 2.8).

In these cases, the normal distribution cannot serve as a good approximation of the variable M/n, so it is more prudent to use the key inferential function of the binomial model to transform the binomial test statistic.

2.3.4. The key inferential function for the binomial model

For the set of hypotheses $H_0 : p_A \leq p_B$ and $H_1 : p_A > p_B$ with p_B assumed to be known, the key inferential function for the binomial model is given by

$$K_{p_B}(p_A) = 2\{\sin^{-1}(\sqrt{p_A}) - \sin^{-1}(\sqrt{p_B})\}.$$

The expectation of the transformed test statistic is then $\mathrm{E}[V_n] = \sqrt{n}K_{p_B}(p_A)$ (see Kulinskaya, Morgenthaler and Staudte (2008, p. 139–140) for the derivation) and the variance stabilised test statistic V_n is given by

$$V_n = 2\sqrt{n}\{\sin^{-1}(\sqrt{\hat{p}_A}) - \sin^{-1}(\sqrt{p_B})\}.. \tag{2.8}$$

Simulations have shown that the empirical mean and variance of V_n are much closer to their theoretical counterparts than it is the case for Z_n (see Figure 2.4). In addition, confidence intervals for V_n are much closer to their nominal levels than it is the case for Z_n (see Figure 2.8). However, V_n is in general not better at controlling the Type I error: As can be seen in Figure 2.6, the power curves of Z_n and V_n coincide in most scenarios with V_n outperforming Z_n only for very small n and p_1.

For both V_n and Z_n, the normal approximation can be improved by applying the Anscombe continuity correction (Kulinskaya, Morgenthaler and Staudte, 2008, p. 140–141):

$$\tilde{p}_A = (M + 3/8)/(n + 3/4) \tag{2.9}$$

This can also be seen visually by comparing Figure 2.2 (no continuity correction) to Figure 2.3 (Anscombe correction) and—even more clearly—by comparing Figure 2.4 (no continuity correction) to Figure 2.5 (Anscombe correction).

The Anscombe correction also improves the empirical coverage probability of nominal confidence intervals for both Z_n and V_n (see Figure 2.8), but does not help with controlling the Type I error (see Figure 2.7). Notably, for very small n and p_1, the Type I error is greater when using the Anscombe-corrected V_n as test statistic.

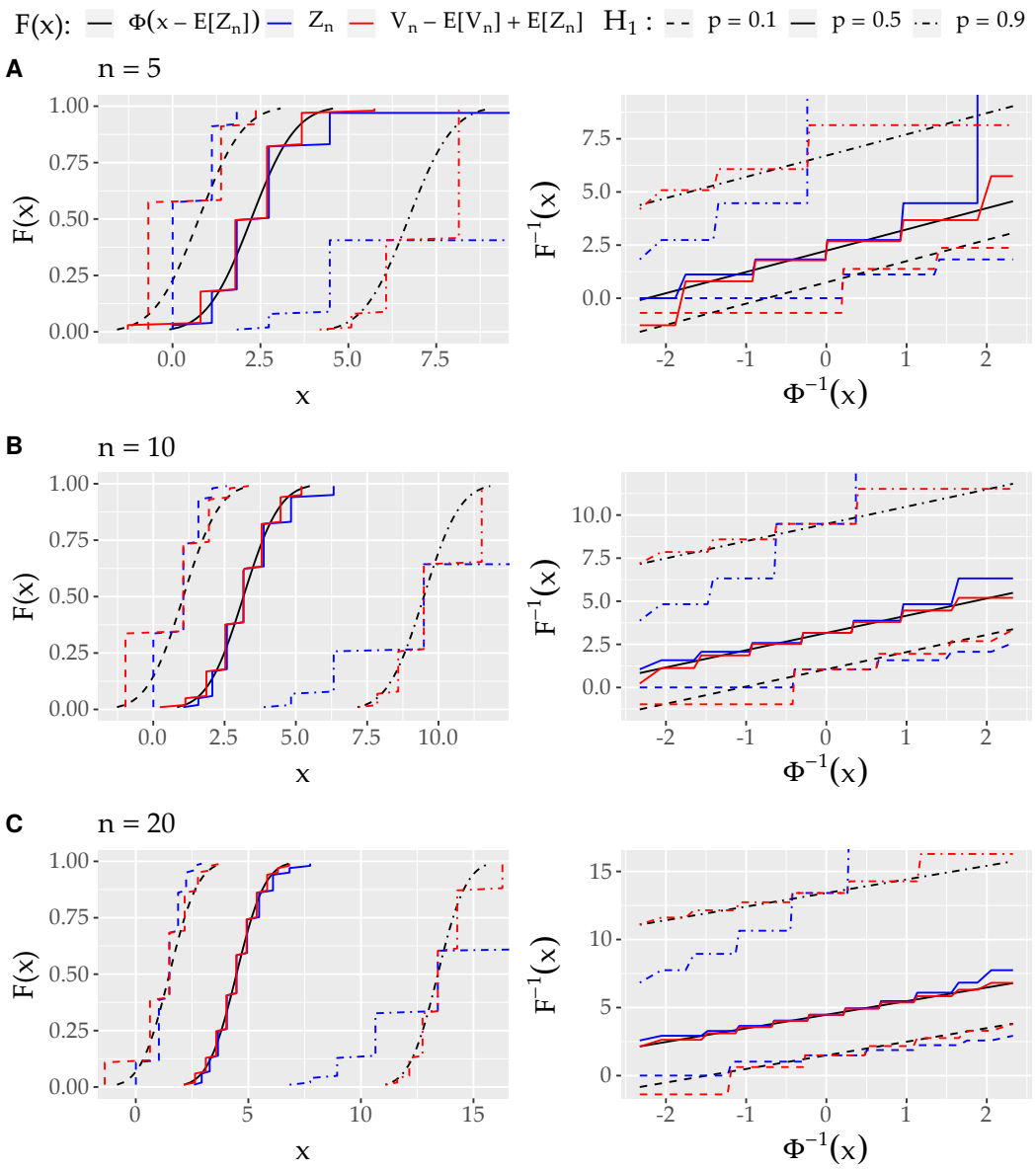

Figure 2.2.: The standard normal cdf Φ (black) approximated by Z_n (blue, Eq. 2.7) and V_n (red, Eq. 2.8) based on binomial variables. The comparison is made for $n = 5$ (A), $n = 10$ (B), and $n = 20$ (C) as well as for three different sets of hypotheses: $H_1 : p = 0.1$ (dashed line), $H_1 : p = 0.5$ (solid line), and $H_1 : p = 0.9$ (dot-dashed line) with $H_0 : p = 0$ in all three cases. The left column shows cumulative distribution functions, the right column shows quantile-quantile-plots. The approximation is worse for smaller n and for p_1 close to either one or zero but V_n outperforms Z_n in all cases. The empirical probability distributions are based on a Gaussian kernel density estimate of V_n and T_n based on 5000 individual draws from a Bin(n, p)-distribution.

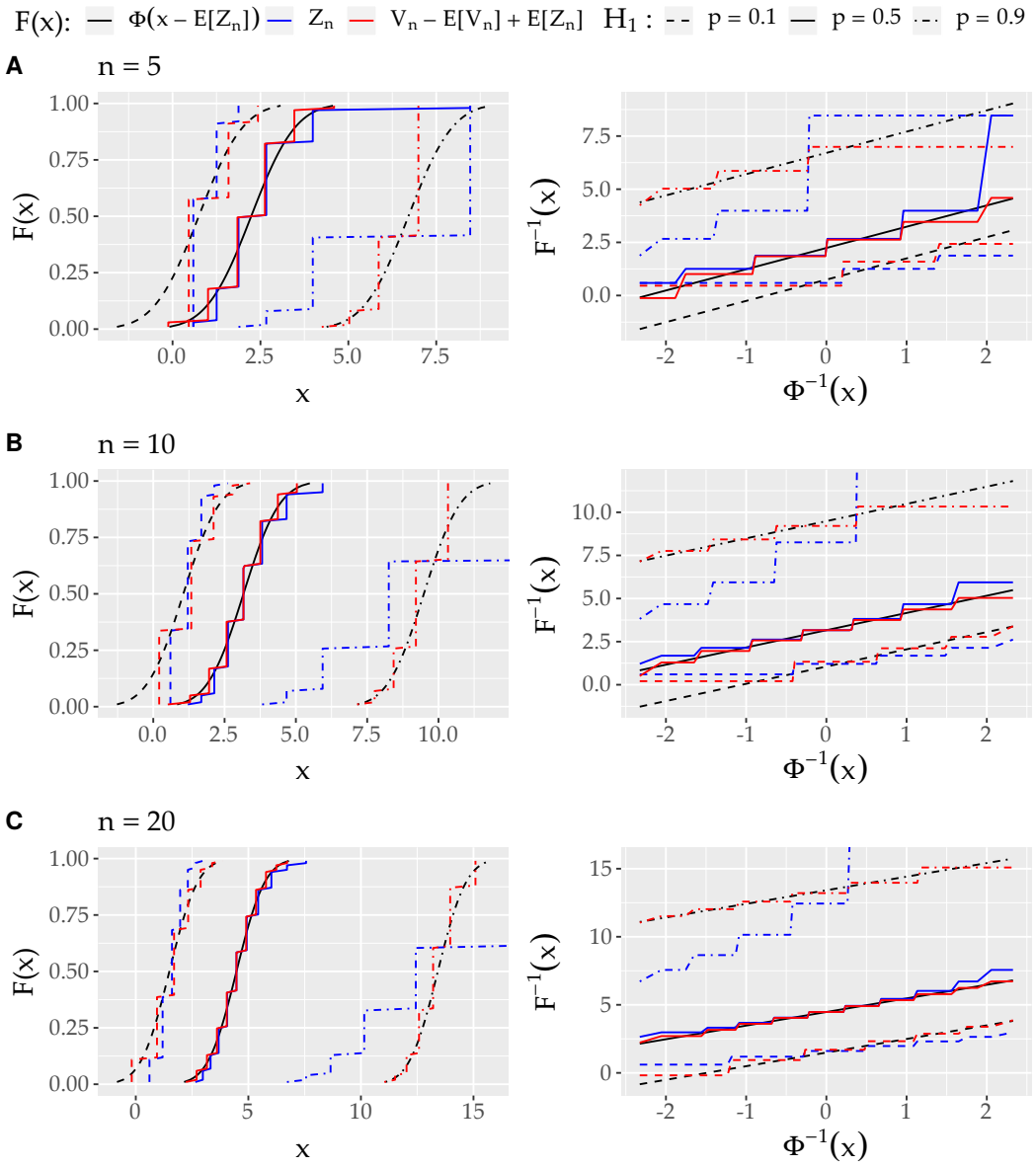

Figure 2.3.: The same description as for Figure 2.2 applies. However, normal approximations Z_n and V_n shown here are based on binomial variables and including Anscombe continuity corrections as described in Eq. 2.9.

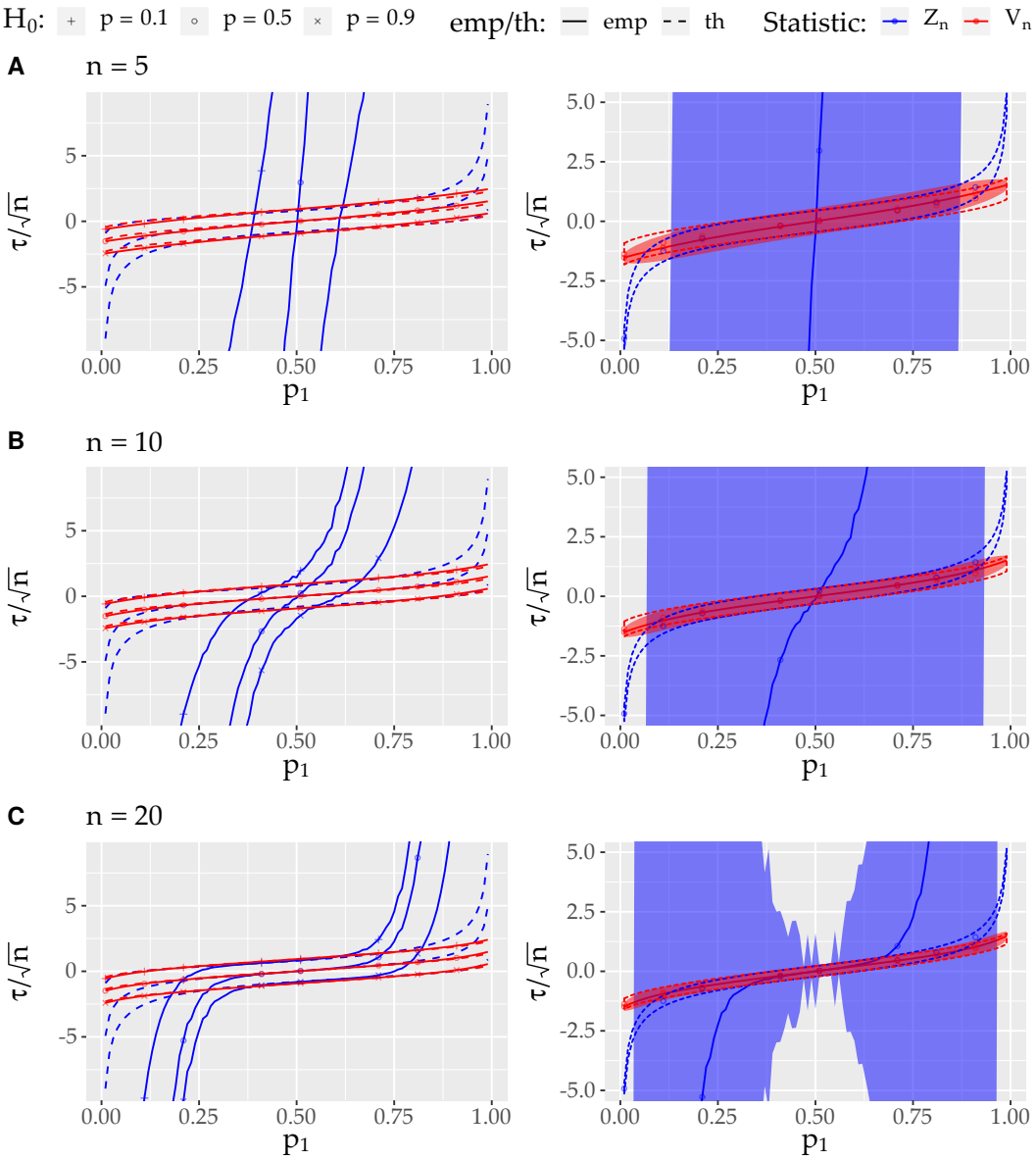

Figure 2.4.: (Left) Empirical means of Z_n (solid blue, Eq. 2.7) and V_n (solid red, Eq. 2.8) based on binomial variables compared to theoretical expectations (dashed blue and dashed red, respectively). The comparison is made for $n = 5$ (A), $n = 10$ (B), and $n = 20$ (C) as well as for three different sets of hypotheses: $H_0 : p = 0.1$ (plus), $H_0 : p = 0.5$ (circle), and $H_0 : p = 0.9$ (cross) with $H_1 : p \in [0.01, 0.99]$ in all three cases. (Right) Empirical standard deviations shown for Z_n and V_n for $H_0 : p = 0.5$. The dashed curves indicate the theoretically expected standard deviations. All empirical values are based on 100′000 independent draws from a binomial distribution. Empirical and theoretical evidence value were weighted by study size for comparison.

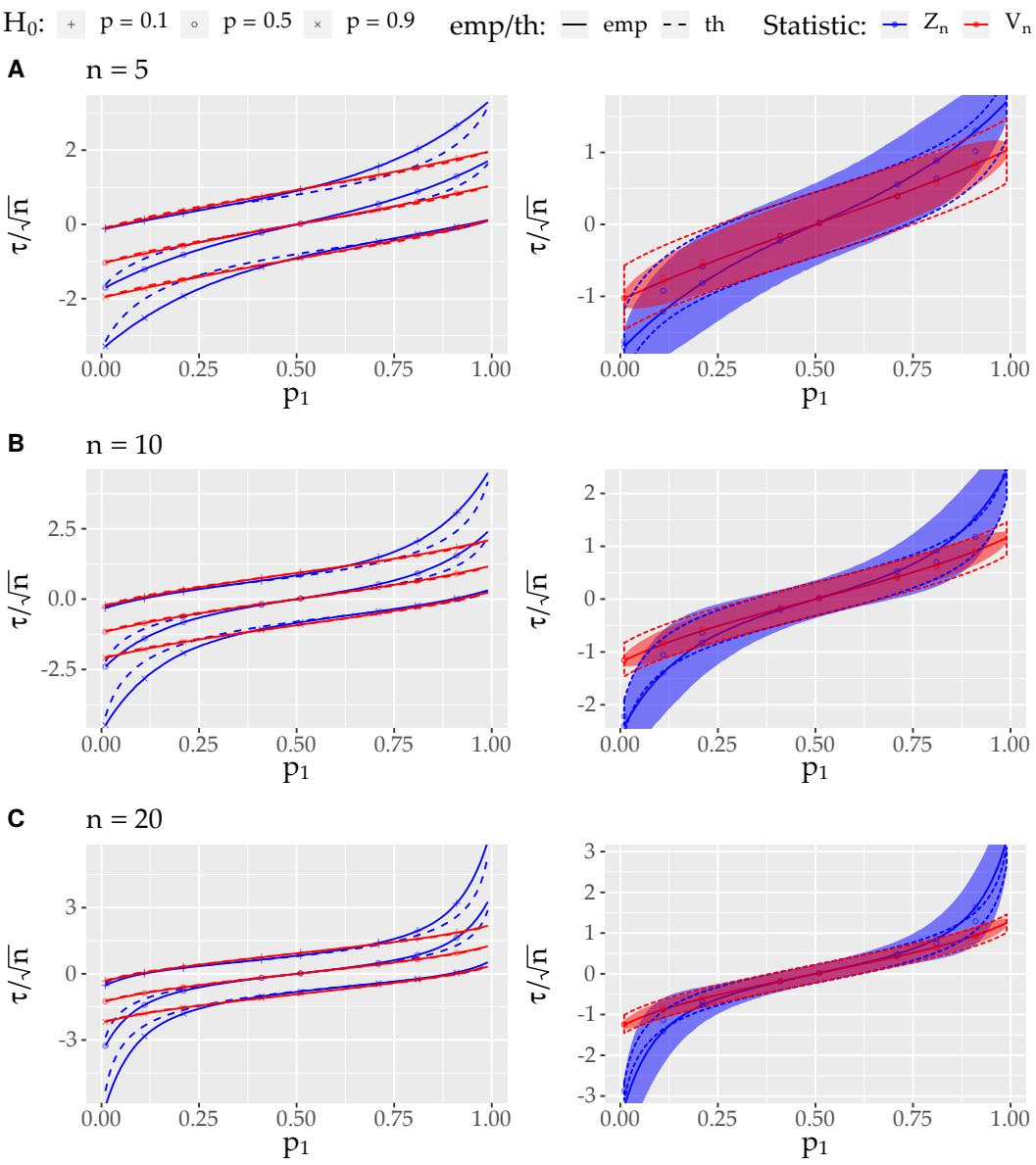

Figure 2.5.: The same description as for Figure 2.4. However, normal approximations Z_n and V_n shown here are based on binomial variables and including Anscombe continuity corrections as described in Eq. 2.9. The correction clearly improves the approximation of the empirical expectations to their theoretical counterpart and improves the variance stabilisation for both Z_n and V_n.

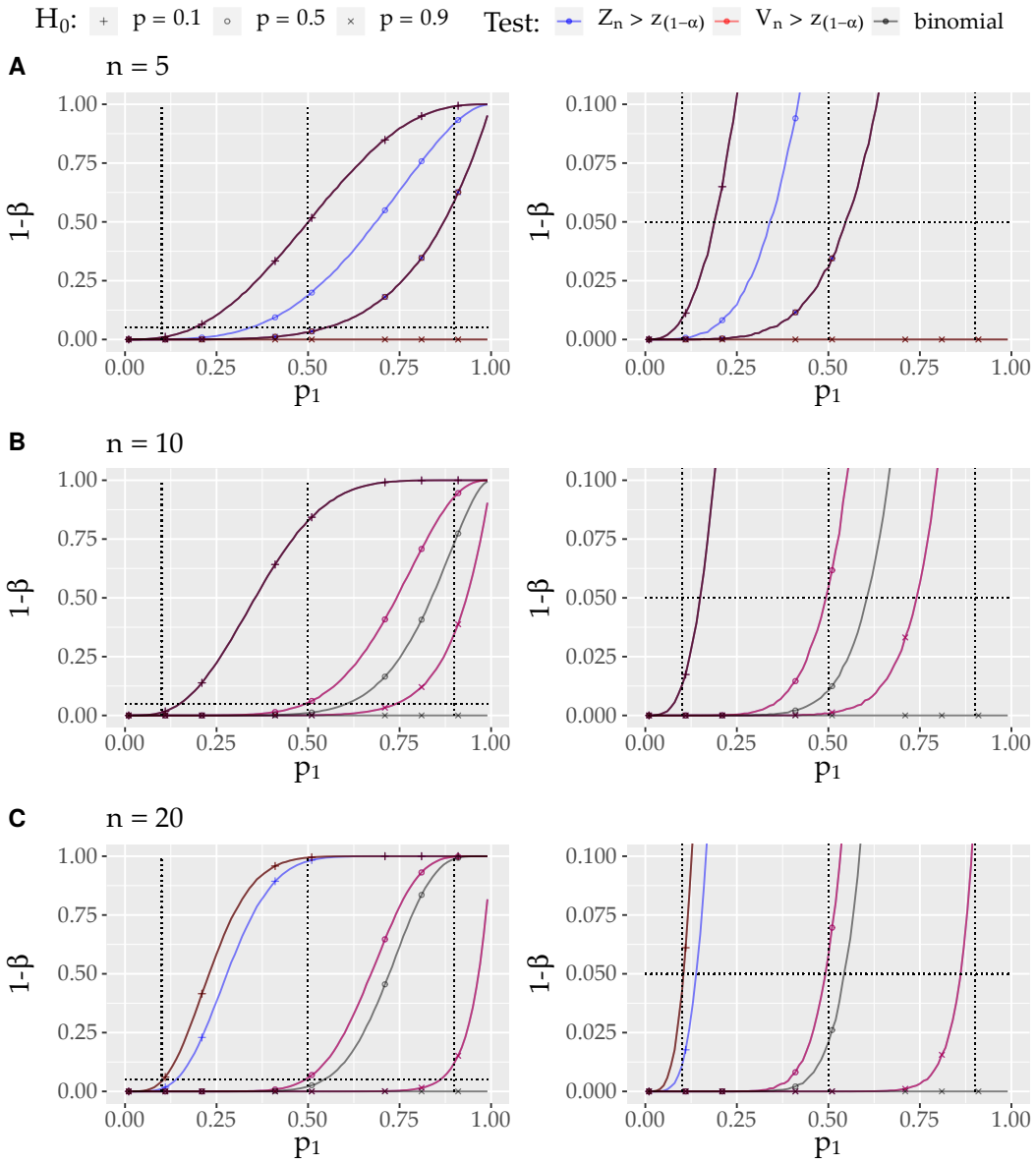

Figure 2.6.: (Left) Empirical power curves for $Z_n > z_{0.95}$ (blue, Eq. 2.7) and $V_n > z_{0.95}$ (red, Eq. 2.8) based on binomial variables compared to the exact binomial test (black). The comparison is made for $n = 5$ (A), $n = 10$ (B), and $n = 20$ (C) as well as for three different sets of hypotheses: $H_0 : p = 0.1$ (plus), $H_0 : p = 0.5$ (circle), and $H_0 : p = 0.9$ (cross) with $H_1 : p \in [0.01, 0.99]$ in all three cases. In most cases, the blue and red power curves coincide. The dotted line on the horizontal axis denotes $1 - \beta = 0.05$. The dotted lines on the vertical axis denote $p_1 = p_0$. (Right) Same curves as in the left column but zoomed in around $1 - \beta \in [0, 0.1]$. All empirical values are based on 100′000 independent draws from a binomial distribution.

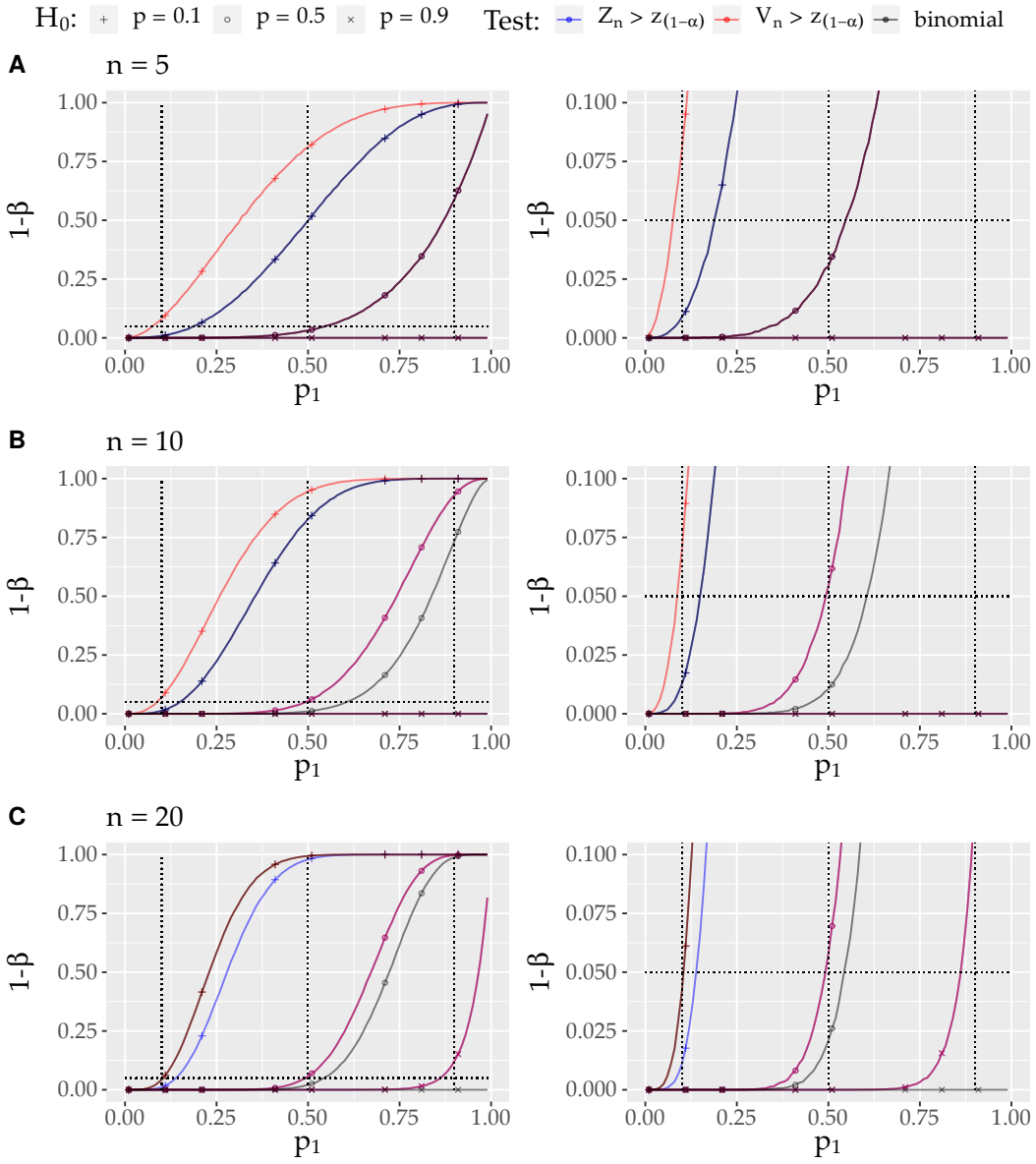

Figure 2.7.: The same description as for Figure 2.6 applies. However, the power curves shown here are based on Anscombe-corrected Z_n and V_n values as described in Eq. 2.9. The correction does in general not improve the power or the control of the Type I error and even increases the Type I error for V_n when n and p_1 are small.

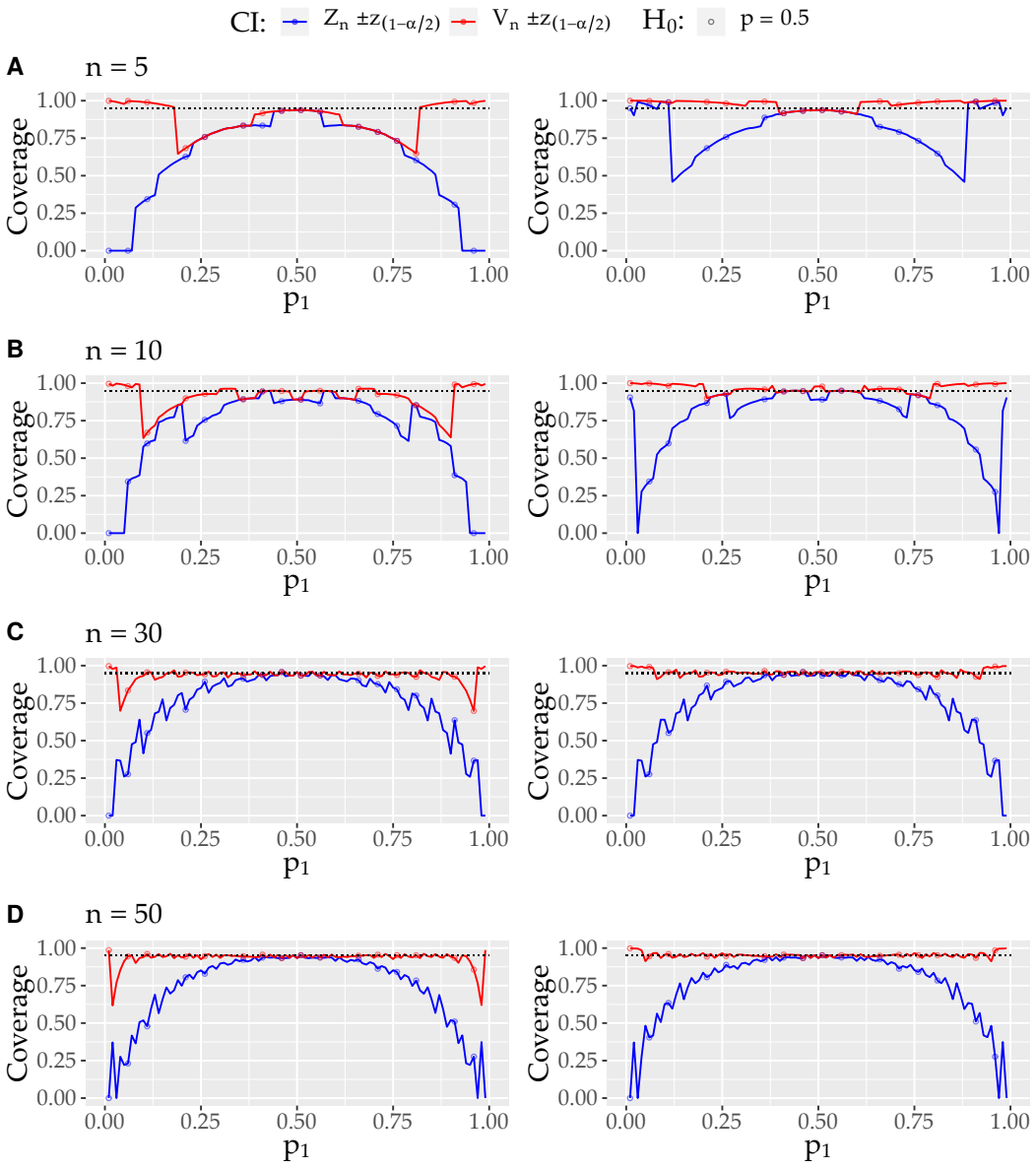

Figure 2.8.: (Left) Empirical coverage probabilities for nominal 95% confidence intervals around Z_n (blue, Eq. 2.7) and V_n (red, Eq. 2.8) based on binomial variables. The comparison is made for $n = 5$ (A), $n = 10$ (B), $n = 30$ (C), and $n = 50$ (D) and for $H_0 : p = 0.5$ (circle) with $H_1 : p \in [0.01, 0.99]$. The dotted black line on the horizontal axis denotes a nominal coverage probability of 95%. (Right) Same comparisons shown as in the left column but for Anscombe-corrected Z_n and V_n as described in Eq. 2.9. All values are based on 100′000 independent draws from a binomial distribution.

2.4. Evidence for the difference in means

Instead of choosing the proportion of surviving patients after a given time period as the primary outcome, another measure to assess the efficacy of a treatment could be the survival time in months after receiving the treatment.

To assess whether treatment A is in fact superior to treatment B, we would randomly and independently assign patients to one of the two treatments and measure their survival time. Let us furthermore assume that the survival times Y_A and Y_B of patients in treatment group A and B, respectively, are log-normally distributed[3], that is, $Y_A \sim \text{Lognormal}(\mu_A, \sigma_A^2)$ and $Y_B \sim \text{Lognormal}(\mu_B, \sigma_B^2)$.

By log transforming Y_A and Y_B we get two normally distributed variables $X_A = \ln(Y_A) \sim \mathcal{N}(\mu_A, \sigma_A^2)$ and $X_B = \ln(Y_B) \sim \mathcal{N}(\mu_B, \sigma_B^2)$. For testing, our hypotheses need to be reformulated as

$$H_0 : \mu \leq \mu_0$$
$$H_1 : \mu > \mu_0$$

with $\mu = \mu_A - \mu_B$ and $\mu_0 = 0$. To test these hypotheses, we define our test statistic as

$$S_n = \bar{X}_{n_A} - \bar{X}_{n_B}.$$

$\bar{X}_{n_A}$ and $\bar{X}_{n_B}$ represent the sampled arithmetic mean of the log transformed survival time of n_A and n_B patients receiving treatment A and B, respectively. Both means are normally distributed with $\bar{X}_{n_A} \sim \mathcal{N}(\mu_A, \sigma^2/n_A)$ and $\bar{X}_{n_B} \sim \mathcal{N}(\mu_B, \sigma^2/n_B)$. The difference of these two means is normally distributed as well, namely $(\bar{X}_A - \bar{X}_B) \sim \mathcal{N}(\mu_A - \mu_B, \sigma_A^2/n_A + \sigma_B^2/n_B)$. The corresponding decision function is then

$$\delta(S_n(\bar{X}_{n_A}, \bar{X}_{n_B})) = \begin{cases} 1, & \text{if } S_n(\bar{X}_{n_A}, \bar{X}_{n_B}) > q_{H_0, 1-\alpha}; \\ 0, & \text{otherwise;} \end{cases}$$

with the threshold value $q_{H_0,1-\alpha}$ denoting the $(1-\alpha)$-quantile of a normal distribution with mean $(\mu_A - \mu_B) = \mu_0 = 0$ and variance $\sigma_A^2/n_A + \sigma_B^2/n_B$ under the null hypothesis.

[3] See for example Royston (2001) or Chapman et al. (2013) for cases in which this assumption can be justified

This statistic fulfills properties E_1, E_2, and E_4, but has $\text{Var}[S_n] \neq 1$. If we know the variance $\sigma^2(1/n_A + 1/n_B)$, we can again use the Z-statistic

$$Z_n = \frac{S_n - \mu_0}{\sigma_{n_A+n_B}} = \frac{(\bar{X}_A - \bar{X}_B) - \mu_0}{\sigma\sqrt{(1/n_A + 1/n_B)}} \overset{H_0}{\sim} \mathcal{N}(0,1),$$

described in Section 2.3.3 to stabilise the variance of the test statistic, thereby fulfilling property E_3 as well.

2.4.1. Student's t-statistic

If the population variance σ^2 is unknown it can be estimated by the empirical variance

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Plugging this into the Z-statistic yields the so-called Student's t-statistic:

$$T_n = \frac{\sqrt{n}\,(\bar{X}_n - \mu_0)}{s_n}. \tag{2.10}$$

In the specific case of the hypotheses stated at the beginning of this section, the t-statistic is given by

$$T_n = \frac{\bar{X}_N - \mu_0}{s_N(1/n_A + 1/n_B)} \sim t(\nu = N-2)$$

with

$$N = n_A + n_b, \quad \bar{X}_N = \bar{X}_A - \bar{X}_B, \quad s_N = \sqrt{\frac{(n_A-1)s_{n_A}^2 + (n_B-1)s_{n_B}^2}{(n_A-1)+(n_B-1)}}.$$

s_N is called the pooled sample variance and serves to fix the variance of the test statistic to 1. Note that this assumes that the true variance σ^2 is the same across both populations. If this is not the case, it is better to use Welch's generalisation of the t-test (Welch, 1947).

As in the examples outlined in Section 2.3, I am going to treat the distributional parameters of treatment B as known and thus will use the t-statistic described in Eq. 2.10 for the following calculations and simulations.

If X is a random variable with normal distribution, T_n is a random variable with a central Student's t-distribution and $\nu = n - 1$ degrees of freedom denoted as $T_n \sim t(\nu)$ (Student, 1908), hence the decision function of the so-called t-test is equal to

$$\delta(T_n(X_1, \dots X_n) = \begin{cases} 1, & \text{if } T_n(X_1, \dots, X_n) > q_{H_0, 1-\alpha}; \\ 0, & \text{otherwise;} \end{cases}$$

with $q_{H_0,1-\alpha}$ equal to the $(1-\alpha)$-quantile of the $t(\nu)$-distribution.

Since s_n^2 converges to σ^2 in probability, we can treat $\sigma \simeq s_n$ and $T_n \overset{H_0}{\simeq} Z_n$ and just use the Z-test. However, if n is not large enough or if μ deviates too much from μ_0 the normal approximation of T_n does not hold anymore and problems similar to those described in Section 2.3.3 appear:

- The probability distribution of T_n becomes leptokurtic with a positive skew ($\mu > \mu_0$) or negative skew ($\mu < \mu_0$ with symmetry only attained if $\mu = \mu_0$ (see Figure 2.9 and Figure 2.10).
- The empirical mean of T_n diverges from its theoretical expectation assuming normality (see Figure 2.11 and Figure 2.12).
- The empirical variance of T_n is not stabilised anymore and diverges from 1 (see Figure 2.11 and Figure 2.12).
- The Type I error grows beyond the pre-defined α-threshold and thus is not controlled anymore (see Figure 2.13 and Figure 2.14).
- The empirical coverage probabilities of the (1-α)-confidence intervals lies clearly below the nominal probabilities and deteriorates further with μ diverging from μ_0 (see Figure 2.15).

For small sample sizes it is therefore inappropriate to treat T_n as normally distributed to test for differences in means. Instead, we should apply the variance stabilised test statistic V_n based on Student's t-distribution.

Under the nul hypothesis H_0, we are dealing with a central t-distribution, but in to calculate the evidence for a set of hypotheses, we also need to know the distribution of T_n under the alternative hypothesis $H_1 : \mu > \mu_0$. To do so, we can rewrite T_n as

$$T_n = \frac{\sqrt{n}\,(\bar{X}_n - \mu) + \sqrt{n}\,(\mu - \mu_0)}{s_n}$$

which follows a noncentral t-distribution with $\nu = n - 1$ degrees of freedom and non-centrality parameter $\lambda = \sqrt{n}(\mu - \mu_0)/\sigma$, that is, $T_n \sim t(\lambda, \nu)$ (Kulinskaya, Morgenthaler and Staudte, 2008, p. 159–160).

2.4.2. The key inferential function for the t-statistic

As before with the binomial test statistical, we can transform the t-statistic into an evidence measure V_n. To do so, we can apply the following transformation function (see Kulinskaya, Morgenthaler and Staudte (2008, p. 160–161) for derivation):

$$h_n(T_n) = \sqrt{2(n-1)}\sinh^{-1}\left(T_n/\sqrt{2(n-1)}\right) \simeq \sqrt{2n}\sinh^{-1}\left(T_n/\sqrt{2n}\right)$$

The evidence V_n in a t-statistic for testing $H_0 : \mu \leq \mu_0$ against $H_1 : \mu > \mu_0$ is therefore given by

$$V_n = \sqrt{2n}\sinh^{-1}\left(\frac{T_n}{\sqrt{2n}}\right) = \sqrt{2n}\sinh^{-1}\left(\frac{(\bar{X}_n - \mu_0)/s_n}{\sqrt{2}}\right) \tag{2.11}$$

and its expectation is

$$\mathrm{E}[V_n] = \tau(\mu) = \sqrt{n}\sqrt{2}\sinh^{-1}\left(\frac{\mu - \mu_0}{\sqrt{2}}\right) = \sqrt{n}\,K_{\mu_0}(\mu).$$

Simulations confirm that the normal approximation and variance stabilisation are better for V_n than for T_n, especially if $\mu \neq \mu_0$ (see Figure 2.9). The same holds true for the correspondence of the empirical mean and variance to their theoretical counterparts (see Figure 2.11).

Whereas V_n does not completely control Type I error rates when n is low but is consistently better than T_n when using standard normal quantiles as critical values. Similarly, the empirical coverage probability of the confidence intervals around V_n lies below the nominal probability for small n but consistently outperforms the empirical coverage probability of confidence intervals around T_n when using using standard normal quantiles.

The normal approximation and performance of both V_n and Z_n can be improved by applying the following finite sample correction (Kulinskaya, Morgenthaler and Staudte, 2008, p. 161):

$$V_n^* = \left(\frac{n - 1.7}{n - 1}\right)\sqrt{2n}\sinh^{-1}\left(\frac{T_n}{\sqrt{2n}}\right) \tag{2.12}$$

The improvement in normal approximation and variance stabilisation is best in the tails of the distribution and is also observable visually by comparing Figure 2.9 (no correction applied) to Figure 2.10 (including finite sample

correction) as well as by comparing Figure 2.11 (no correction applied) to Figure 2.12 (including finite sample correction).

The finite sample correction also helps to control the Type I error. The corrected T_n meets the nominal α-threshold almost, the corrected V_n meets it completely, even for very small n (see Figure 2.14). Finally, the correction does also markedly improve the empirical coverage probability of confidence intervals around V_n and T_n (see Figure 2.15).

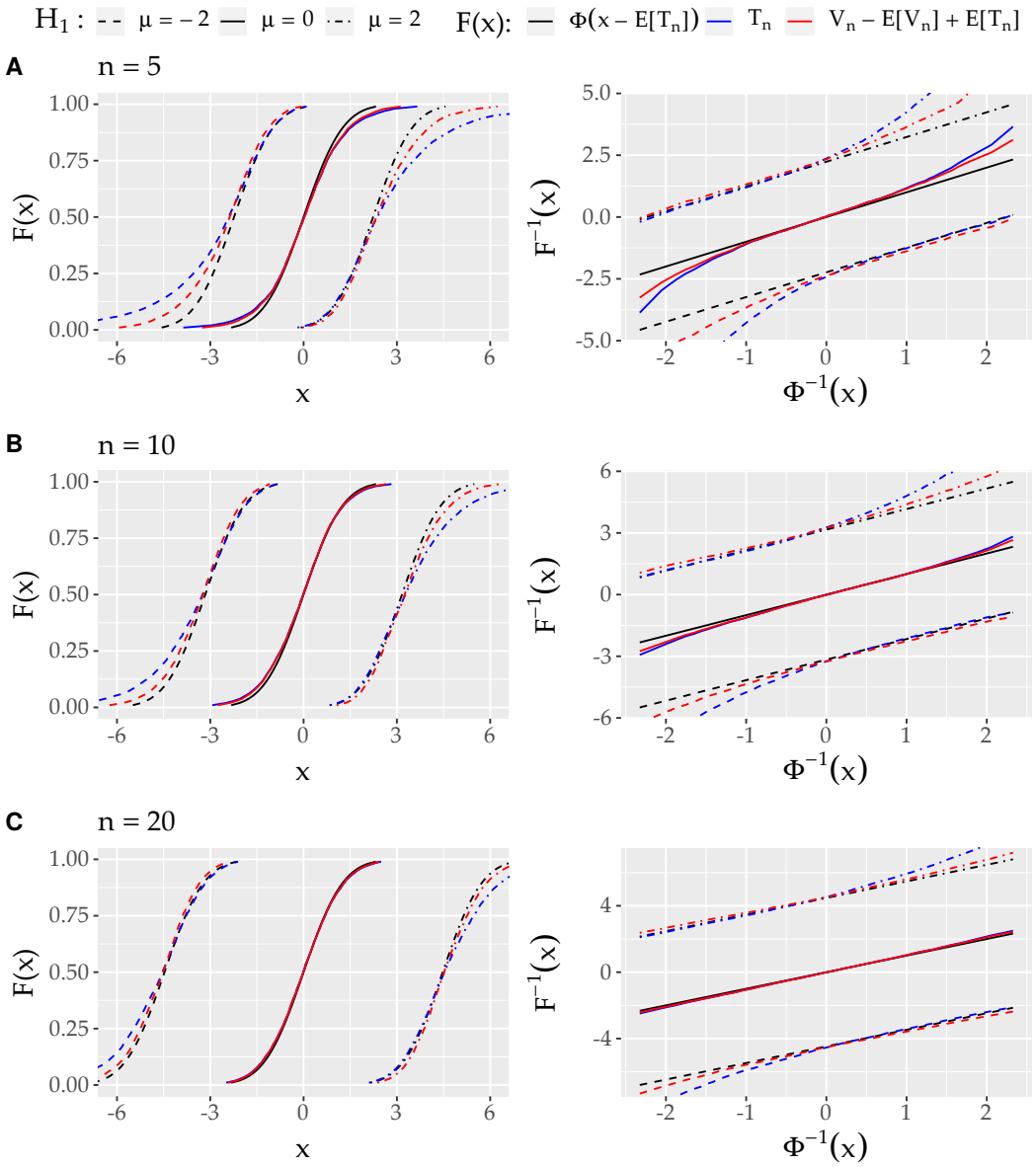

Figure 2.9.: The standard normal cdf Φ (black) approximated by T_n (blue, Eq. 2.10) and V_n (red, Eq. 2.11) based on normally distributed variables with variance $\sigma^2 = 4$. The comparison is made for $n = 5$ (A), $n = 10$ (B), and $n = 20$ (C) as well as for three different sets of hypotheses: $H_1 : \mu = -2$ (dashed line), $H_1 : \mu = 0$ (solid line), and $H_1 : \mu = 2$ (dot-dashed line) with $H_0 : \mu = 0$ in all three cases. The left column shows cumulative distribution functions, the right column shows quantile-quantile-plots. The approximation is worse for smaller n and for mu_1 farther away from μ_0 but V_n outperforms Z_n in all cases. The empirical probability distributions are based on a Gaussian kernel density estimate of V_n and T_n based on 5000 individual draws from a $\mathcal{N}(\mu, 4)$-distribution.

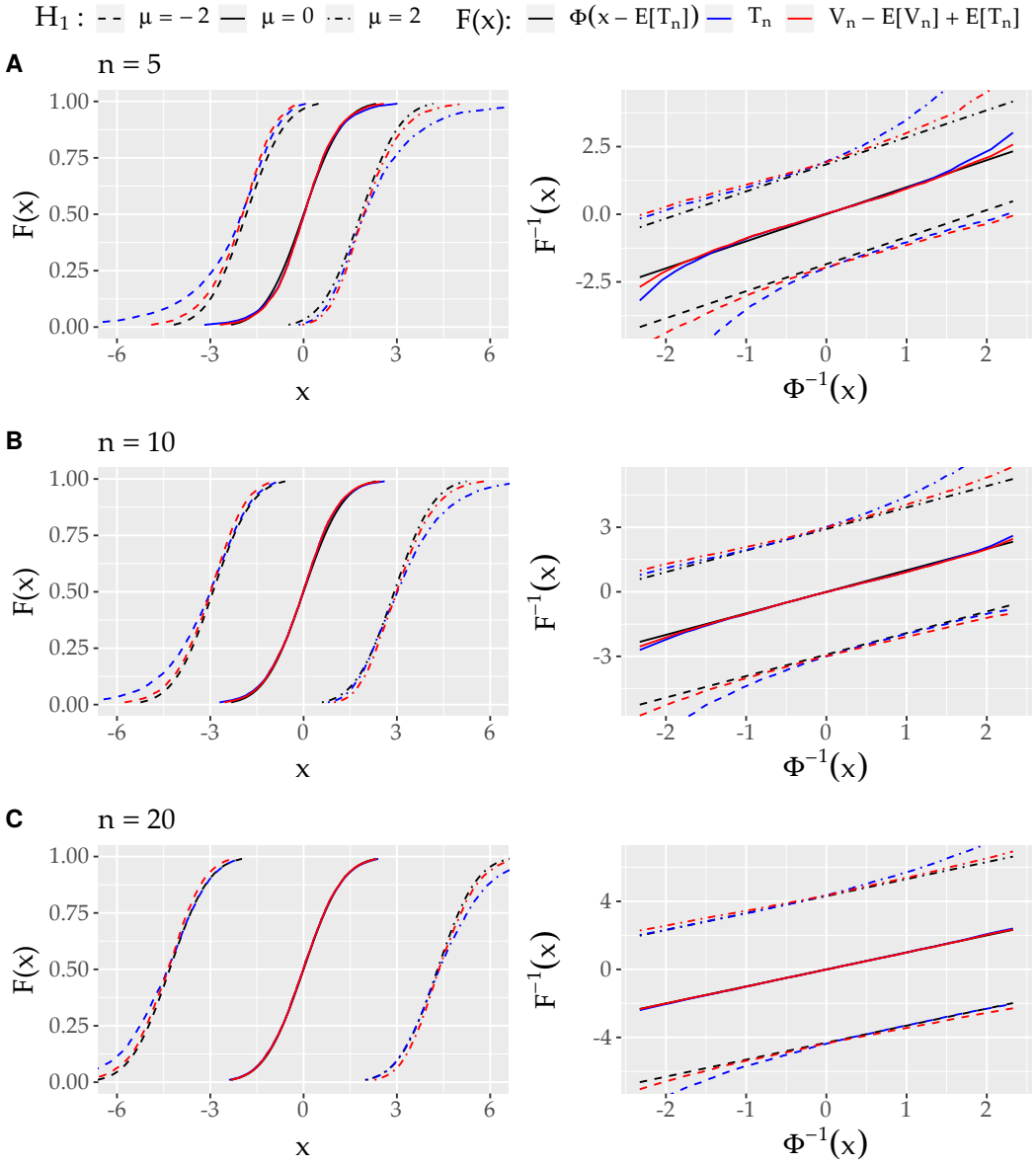

Figure 2.10.: The same description as for Figure 2.9 applies. However, Z_n and V_n shown here include the finite sample correction described in Eq. 2.12.

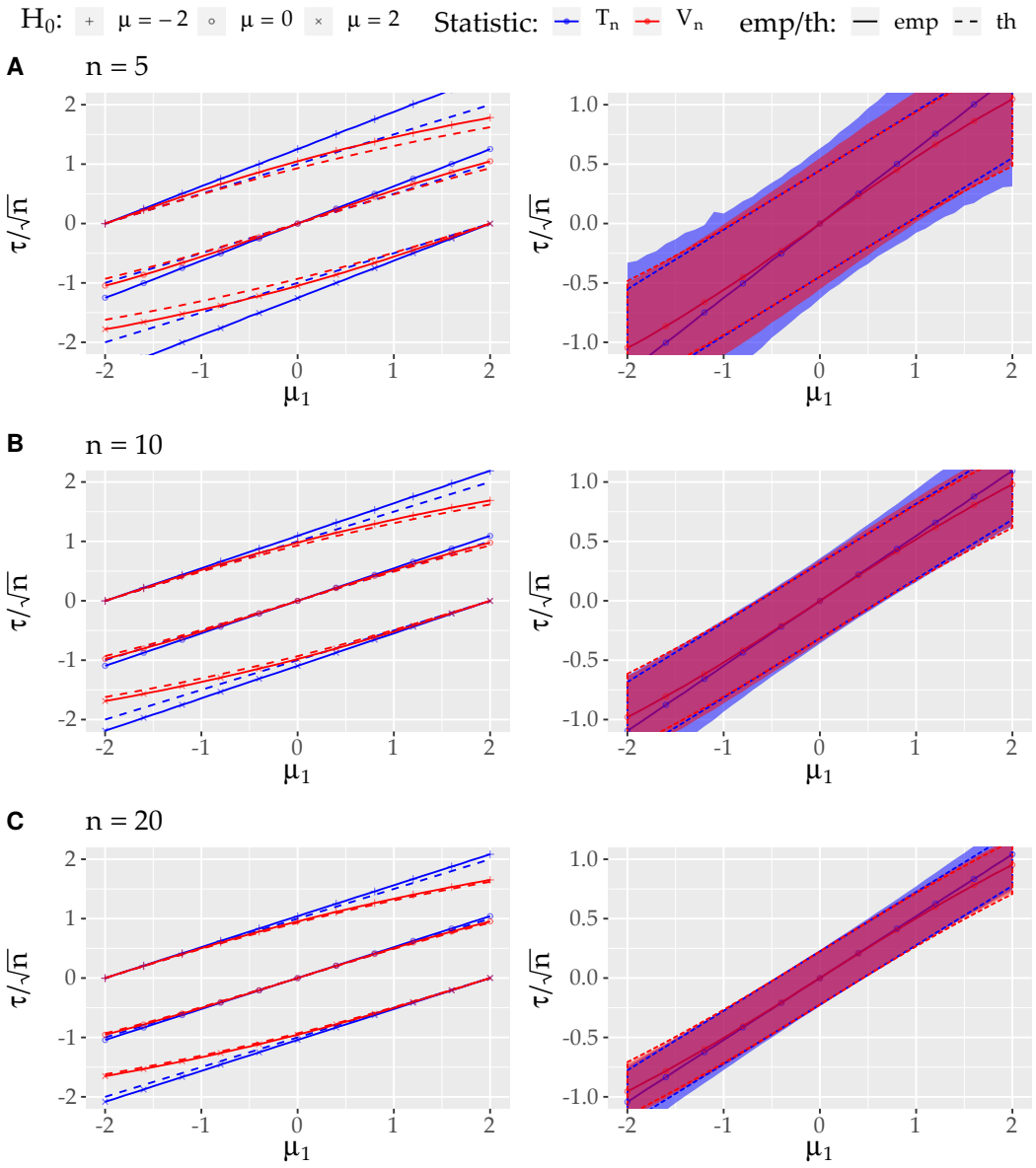

Figure 2.11.: (Left) Empirical means of T_n (solid blue, Eq. 2.10) and V_n (solid red, Eq. 2.11) based on normally distributed variables with variance $\sigma^2 = 4$ compared to theoretical expectations (dashed blue and dashed red, respectively). The comparison is made for $n = 5$ (A), $n = 10$ (B), and $n = 20$ (C) as well as for three different sets of hypotheses: $H_0 : \mu = -2$ (plus), $H_0 : \mu = 0$ (circle), and $H_0 : \mu = 2$ (cross) with $H_1 : \mu \in [-2, 2]$ in all three cases. (Right) Empirical standard deviations shown for T_n and V_n for $H_0 : \mu = 0$. The dashed curves indicate the theoretically expected standard deviations. All empirical values are based on 100′000 independent draws from a $\mathcal{N}(\mu, 4)$-distribution. Empirical and theoretical evidence value were weighted by study size for comparison.

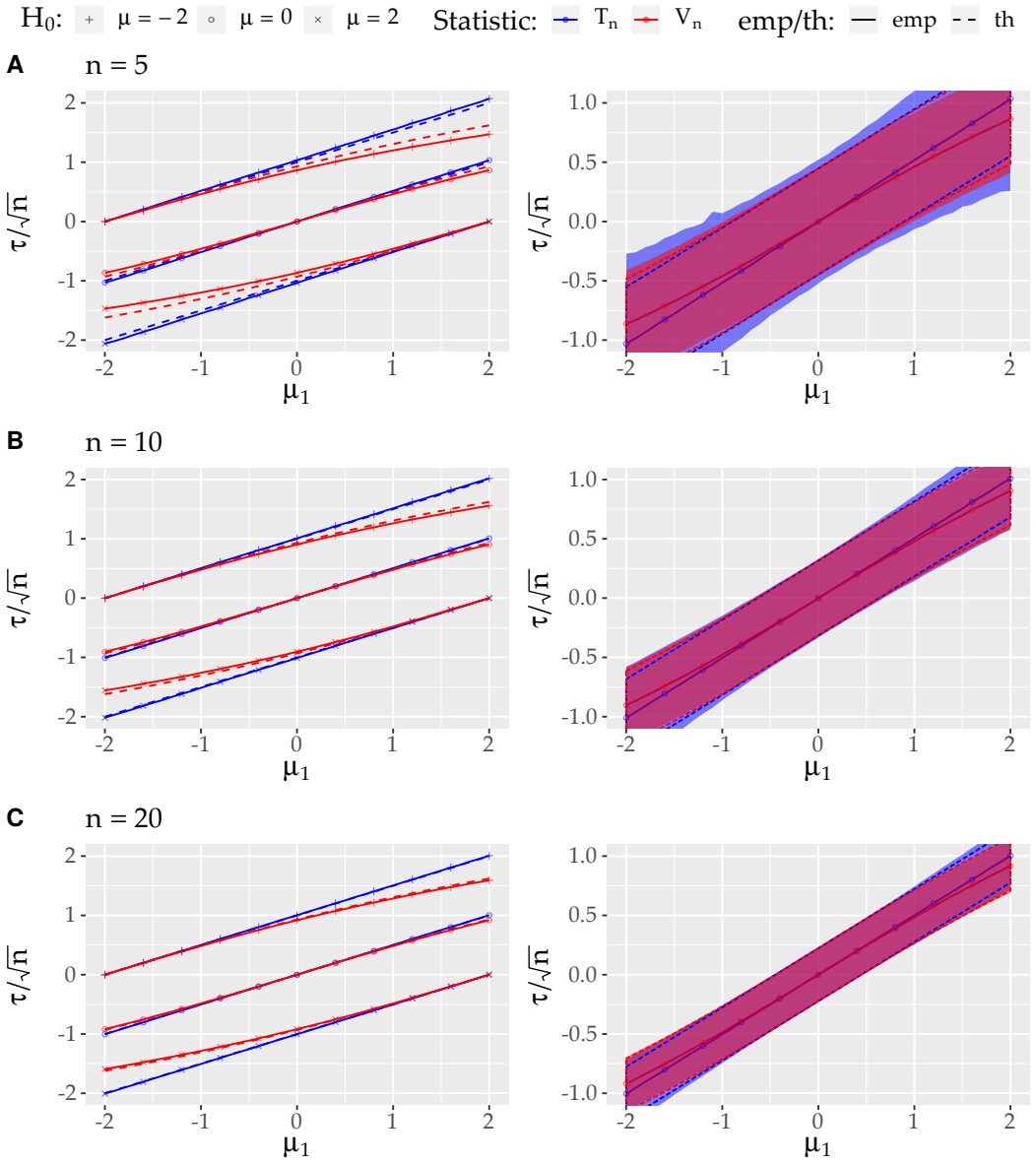

Figure 2.12.: The same description as for Figure 2.11. However, Z_n and V_n shown here include the finite sample correction described in Eq. 2.12. The correction clearly improves the approximation of the empirical expectations to their theoretical counterpart and improves the variance stabilisation for both T_n and V_n, especially in the tails.

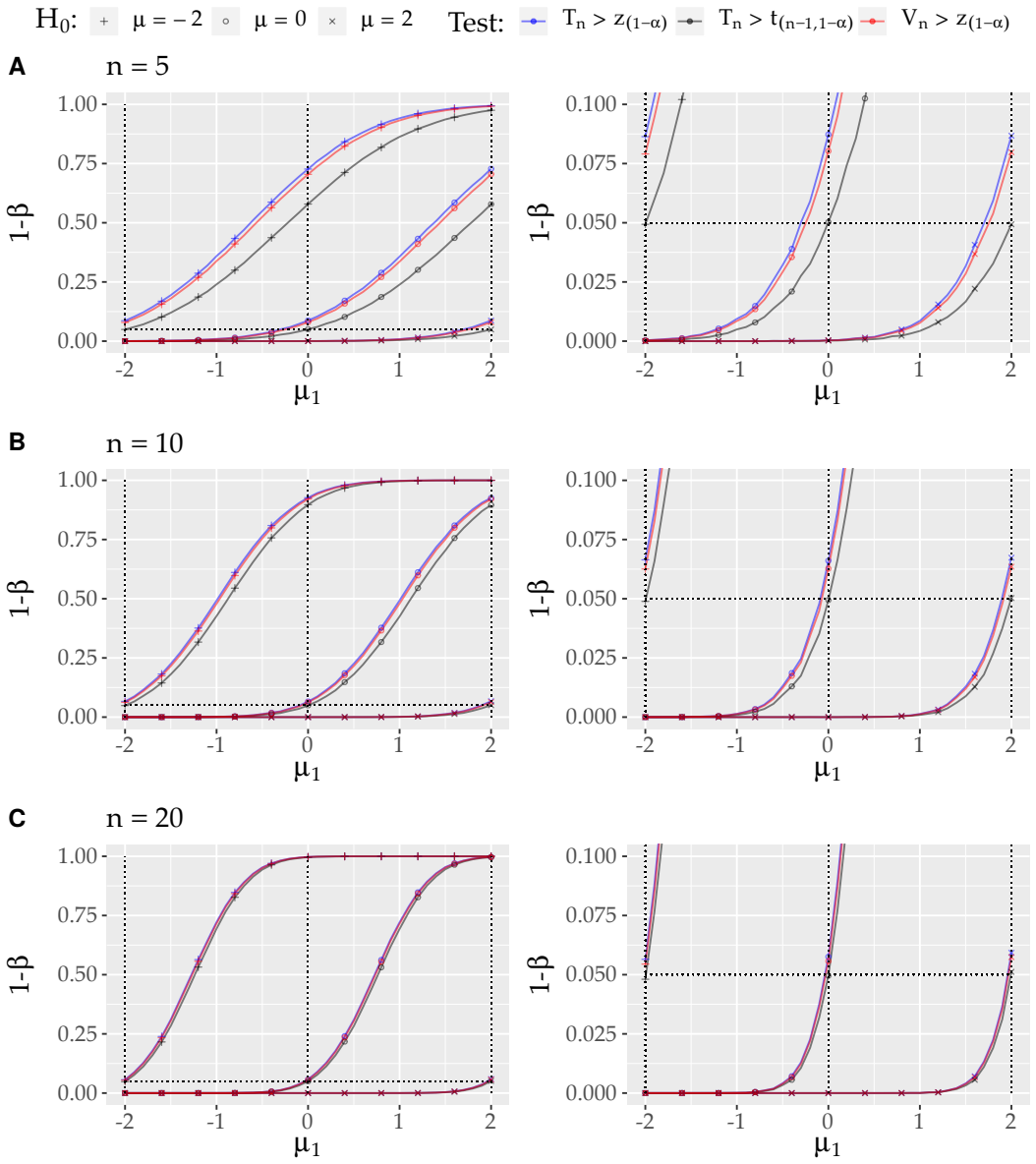

Figure 2.13.: (Left) Empirical power curves for $T_n > z_{0.95}$ (blue) and $V_n > z_{0.95}$ (red) compared to $T_n > t_{n-1,0.95}$ (black). The comparison is made for $n = 5$ (A), $n = 10$ (B), and $n = 20$ (C) as well as for three different sets of hypotheses: $H_0 : \mu = -2$ (plus), $H_0 : \mu = 0$ (circle), and $H_0 : \mu = 2$ (cross) with $H_1 : \mu \in [-2, 2]$ in all three cases. In most cases, the blue and red power curves coincide. The dotted line on the horizontal axis denotes $1 - \beta = 0.05$. The dotted lines on the vertical axis denote $mu_1 = mu_0$. (Right) Same curves as in the left column but zoomed in around $1 - \beta \in [0, 0.1]$. V_n and T_n are calculated according to Eq. 2.11 and Eq. 2.10, respectively, based on 100′000 independent draws from a $\mathcal{N}(\mu, 4)$-distribution.

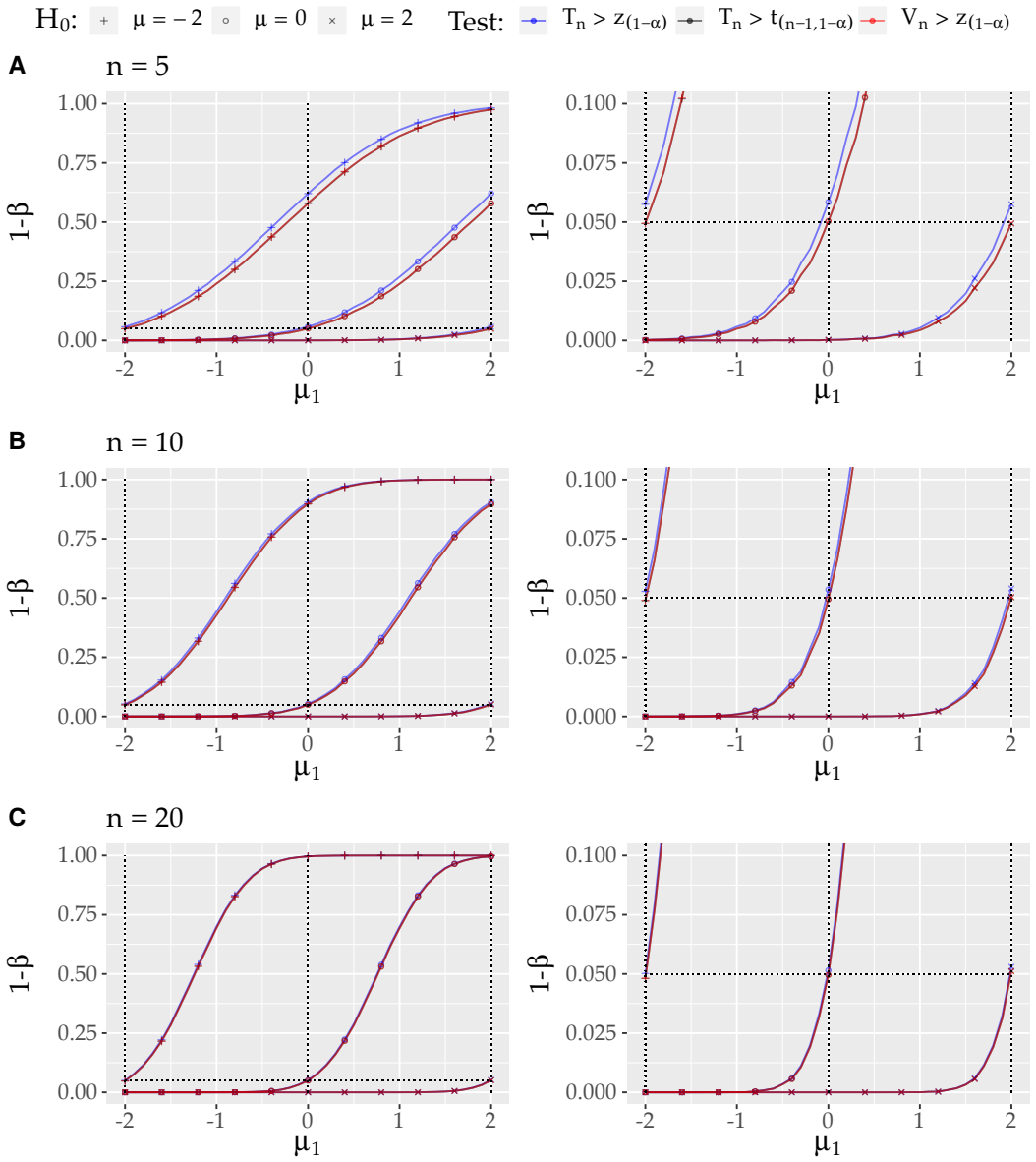

Figure 2.14.: The same description as for Figure 2.13 applies. However, the power curves shown here include the finite sample correction described in Eq. 2.12. The correction clearly improves control of Type I error rates for both T_n and V_n with the latter meeting the nominal α-threshold.

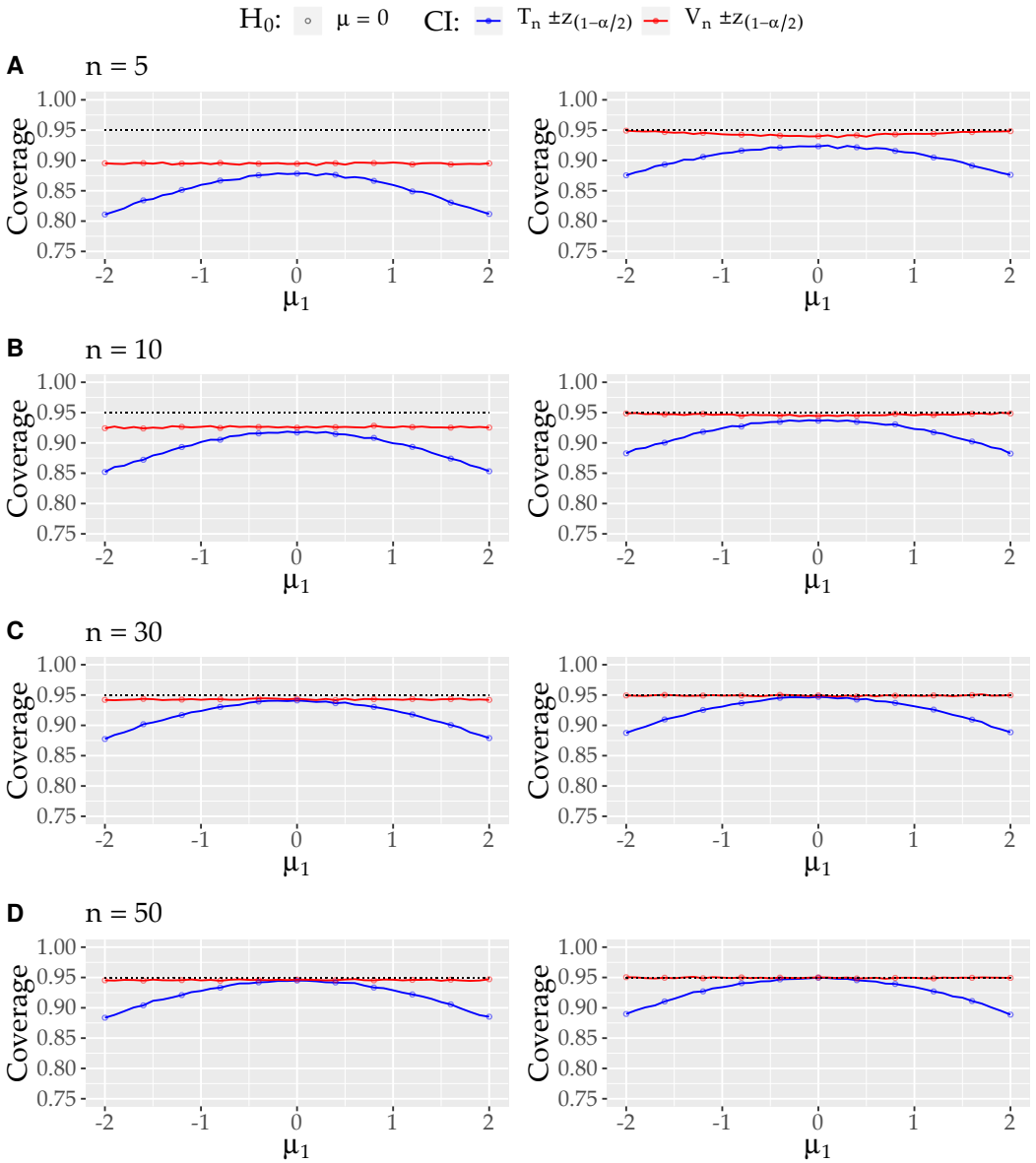

Figure 2.15.: (Left) Empirical coverage probabilities for nominal 95% confidence intervals around T_n (blue) and around V_n (red) using standard normal quantiles. The comparison is made for $n = 5$ (A), $n = 10$ (B), $n = 30$ (C), and $n = 50$ (D) and for $H_0 : \mu = 0$ (circle) with $H_1 : \mu \in [-2, 2]$. The dotted black line on the horizontal axis denotes a nominal coverage probability of 95%. (Right) Same comparisons shown as in the left column but with T_n and V_n including the finite sample correction described in Eq. 2.12. V_n and T_n are calculated according to Eq. 2.11 and Eq. 2.10, respectively, based on 100′000 independent draws from a $\mathcal{N}(\mu, 4)$-distribution.

3. Detecting and Correcting Publication Bias

'The political principle that anything can be proved by statistics arises from the practice of presenting only a selected sub-set of the data available.'

— Ronald A. Fisher,
Statistical Methods and Scientific Induction, (1955, p. 75)

Aggregating evidence measures from different studies is quite simple—theoretically. Given access to the raw data of each study and assuming that each study was designed and conducted in the same manner, one can simply calculate summary and test statistics based on the total aggregate of the data. Of course, statistical practice is hardly ever so straightforward.

Firstly, easy access to indicative summary statistics such as mean, standard deviation and sample size, let alone the raw data, is still very rare. Secondly, study designs and protocols often deviate heavily between different sites even if the intention is to measure the same primary outcome. Thirdly, even if access to raw data and consistent and rigorous study designs and protocols are given, attempts to find accurate global evidence measures might be hampered since the available body of literature can be biased in favour of certain study properties other than its methodological quality.

This is not a particularly novel insight. Sterling (1959) already pointed out in 1959 that 'when a fixed level of significance is used as critical criterion for selecting reports for dissemination in professional journals, it may result in embarrassing and unanticipated results'. One of these 'embarrassing results' might be that the majority of journals are 'filled with the 5% of the studies that show Type I errors' while the other 95% of studies with non-significant test results remain largely unpublished (Rosenthal, 1979).

To make things worse, selection for statistical significance is by far not the only reason for the occurrence of this so-called 'publication bias'. Other

properties that might influence the publication probability of a result include its novelty (Auspurg and Hinz, 2011), its concordance with prior knowledge (Cooper, DeNeve and Charlton, 1997), its newsworthiness (Auspurg and Hinz, 2011), its economic value (Chalmers, Frank and Reitman, 1990) or its political content (Eitan et al., 2018). In addition, other outcomes of the same study (Dickersin, 1990), such as the funding source of said study (Dickersin, 1990), economic or ideological conflicts of interests (Chalmers, Frank and Reitman, 1990; Eitan et al., 2018), ignorance about previous studies (Chalmers, Frank and Reitman, 1990) and even the motivation (Chalmers, Frank and Reitman, 1990; Cooper, DeNeve and Charlton, 1997; Auspurg and Hinz, 2011; Franco, Malhotra and Simonovits, 2014) or reputation (Dickersin, 1990) of the study authors have been proposed as potential influences on the probability of publication.

However, since *post hoc* correction of publication bias is only possible if the probability of publishing a study is influenced by the statistical properties of the result, this chapter will exclusively focus on methods to detect and—at least theoretically—correct for publication bias arising from using a fixed level of significance as critical criterion for publication.

3.1. Significance: The fickle gatekeeper of scientific publishing

Using a pre-defined threshold with which to compare the outcome of a statistical test has been commonplace since the early days of modern statistical inference (Cowles and Davis, 1982). William Sealy Gosset wrote already in 1908 that a deviation of 'three times the probable error in the normal curve [...] would be considered significant [for most purposes]' (Student, 1908, p. 13).

Three times the probable error roughly corresponds to two standard deviations of a normal distribution or an α-level of 0.05—a threshold that was repeatedly adopted by Fisher (1925) in his seminal book 'Statistical Methods for Research Workers' to assess whether a particular result warrants further investigation. Fisher called it 'convenient to take this point as a limit in judging whether a deviation [from the null] is to be considered significant or not' (p. 45), but simply used this threshold simply as a rough filter and never as a definitive decision rule. 'If one in twenty does not seem high

enough odds, we may, if we prefer it, draw the line at one in fifty [...], or one in a hundred [...]. Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment *rarely fails* to give this level of significance. The very high odds sometimes claimed for experimental results should usually be discounted, for inaccurate methods of estimating error have far more influence than has the particular standard of significance chosen' (Fisher, 1926, p. 85–86).

Similarly, the hypothesis testing approach developed by Jerzy Neyman and Egon Pearson hinged on fixing a 'level of rejection' (Neyman, Pearson and Yule, 1933) or 'critical region' (Neyman and Pearson, 1933). They also made abundantly clear that the exact choice of these significance levels should be instructed by the researcher's assessment of how damaging the false rejection of the null hypothesis would be. 'In making a decision upon which subsequent action will be based we are influenced by the consequences which follow from a wrong decision; some errors will matter more than others' (Neyman, Pearson and Yule, 1933, p. 509).

Nevertheless, most users of statistics quickly started to treat significance thresholds as rigid decision boundaries with which to distinguish 'good' from 'bad' results. 'Significance' has in many ways become a synonym for 'relevance', 'importance', or even 'scientific quality'. It was shown that even journal reviewers tend to assess the methodological quality of a study more favourably if it reports significant findings (Mahoney, 1977), so it should not come as a surprise either that statistically significant studies are heavily overrepresented in publications across a wide range of disciplines and generally have a higher probability of being submitted and considered for publication in the first place (Hedges, 1984; Begg and Berlin, 1988; Dickersin, 1990; Easterbrook et al., 1991; Cooper, DeNeve and Charlton, 1997; Gerber, Green and Nickerson, 2001; Dickersin, 2005; Gerber and Malhotra, 2006; Ioannidis and Trikalinos, 2007; Gerber and Malhotra, 2008; Weiß and Wagner, 2011; Fanelli, 2012; Franco, Malhotra and Simonovits, 2014; Kühberger, Fritz and Scherndl, 2014; Flint et al., 2015; Berning and Weiß, 2016).

These distorting effects on scientific research in general and scientific publishing in particular have sparked heated debates among statisticians about the role of significance for statistical inference, with some scholars calling for abandoning the usage of statistical significance as decision criterion altogether (McShane et al., 2019; Amrhein, Greenland and McShane, 2019) and

others defending the merits of predefined decision rules such as significance tests (Ioannidis, 2019).

There is consensus, however, that an exaggerated fixation on statistically significant results or 'significosis', as Antonakis (2017) dubbed it, can heavily bias meta-analyses of the published body of literature by inflating effect sizes and distorting the accuracy of estimates. The following sections are therefore dedicated to methods for the detection and correction of publication bias.

3.2. Detecting publication bias: How much significance is too much?

Even though an increased amount of significant findings in the research literature should definitely give meta-analysts pause, it cannot be taken as proof for the existence of publication bias. After all, researchers usually do not embark haphazardly on scientific endeavours but rather base their studies on prior knowledge about which ideas are worth pursuing. Hence, in the ideal world of carefully designed and executed experiments, one would expect researchers to find real effects (as opposed to statistical blips) on a regular basis, therefore increasing the relative amount of statistically significant findings in the published literature. Hence, it is important to have methods with which to distinguish over-representation of significant findings in the literature due to publication bias from over-representation because of real effects.

3.2.1. The file drawer problem

Rosenthal (1979) offered a crude, but straightforward way to calculate the number of studies that theoretically got stuck in the so-called 'file drawer' in a worst-case scenario; that is, a scenario in which all published inferences about hypotheses consist exclusively of Type I errors.

To estimate how many unpublished studies would be needed to bring the overall p-value of all studies combined (both published and unpublished) to a certain significance level, Rosenthal suggested combining all published

results by using the standard normal deviates z_j associated with the p-values of each result j so that

$$z_c = k\bar{z}_k / \sqrt{k} = \sqrt{k}\bar{z}_k.$$

Here, $\bar{z}_k$ denotes the arithmetic mean of all standard normal deviates combined. Since $Z_j \sim \mathcal{N}(0,1)$, it holds that $\frac{1}{k}\sum_{i=1}^{k} Z_j = \bar{Z}_k \sim \mathcal{N}(0,1/k)$ and $\sqrt{k}\bar{Z}_k \sim \mathcal{N}(0,1)$, hence z_c is the realisation of a standard normally distributed variable. By assuming that all k studies are independent from each other, we can calculate the number of studies o needed to achieve the desired significance threshold α with standard normal quantile $z_{1-\alpha}$ for a one-sided superiority test:

$$\begin{aligned} z_{1-\alpha} &= \frac{k\bar{z}_k + o\bar{z}_o}{\sqrt{k+o}} \\ \iff \quad o &= \frac{(k\bar{z}_k + o\bar{z}_o)^2}{z_{1-\alpha}^2} - k. \end{aligned}$$

If one assumes—as Rosenthal did—that the omitted studies o show a null effect on average ($\bar{z}_o = 0$), the expression can be simplified to

$$o = \frac{(k\bar{z}_k)^2}{z_{1-\alpha}^2} - k.$$

However, this heavily overestimates the number of potentially omitted studies because the mean of the standard normal distribution truncated to the right at $z_{1-\alpha} < \infty$ is negative. It is therefore better to calculate o by using $\bar{z}_o = \mathrm{E}[Z \mid Z < z_{1-\alpha}]$ which leads to

$$o^* = \frac{-2k\bar{z}_k\bar{z}_o + z_{1-\alpha}^2 - z_{1-\alpha}\sqrt{4k\bar{z}_o^2 - 4k\bar{z}_k\bar{z}_o + z_{1-\alpha}^2}}{2\bar{z}_o^2}.$$

In both cases, if the number of studies needed to bring the combined p-value to the significance threshold α is rather low (Rosenthal suggests $5k + 10$ as a tentative threshold), one should be wary of potential publication bias. If the number is very high, one can be more confident that there is indeed an effect, but even in this latter case one cannot rule out publication bias. Rosenthal's file drawer estimator and the correction outlined above only yield worst-case estimates for the number of omitted studies assuming the null effect is true. If there is a real effect, one needs to resort to other methods to gauge the number of potentially omitted studies.

Table 3.1 shows that the detection of publication bias not only fails if there is a real effect but also if the worst-case assumption holds, that is, if one assumes that all published studies consist exclusively of Type I errors. In the example simulations shown in Figure 3.1, the file drawer method only detects publication bias if the true effect is zero and the body of published studies consists of all significant results plus 10% of non-significant results.

	k	μ_1	o	o^*	bias detected?
full sample (no bias, A)	200	0	−141	−181	o : no o^*: no
	200	0.3	8123	884	o : no o^*: no
significant studies (B)	13	0	320	109	o : no o^*: no
	38	0.3	2705	457	o : no o^*: no
significant studies and 10% of non-significant studies (C)	32	0	99	43	o : yes o^*: yes
	55	0.3	3086	494	o : no o^*: no

Table 3.1.: The number of potentially suppressed studies o (uncorrected) and o^* (corrected) calculated for the examples shown in Figure 3.1. To test for the presence of publication bias, the number of studies in the file drawer was compared to the threshold value proposed by Rosenthal ($5k + 10$).

In addition to the potential shortcomings explained above, it is important to point out that calculating the file drawer in the manner described above hinges on the assumption that the primary outcomes, the hypotheses as well as the testing procedure have been defined by the authors at the onset of the study. If they resorted to any form of 'p-hacking' (Head et al., 2015) or 'HARKing: Hypothesising After the Results are Known' (Kerr, 1998), however, the expected Type I error would be much larger than indicated by the predefined significance threshold because it is rather easy to turn spurious results into significant findings by resorting to the strategies outlined above (Simmons, Nelson and Simonsohn, 2011).

Hence, the number of theoretically omitted studies that is needed to turn published results non-significant would be considerably lower because the source of bias is not primarily the omission of studies who show non-significant findings but rather the violation of fundamental principles of significance testing.

3.2.2. How many significant studies should be expected?

One approach to circumvent the shortcomings of the file drawer estimation in the presence of p-hacking and HARKing was developed by Ioannidis and Trikalinos (2007). It hinges on the idea that—given a real effect θ_j—the probability of detecting this effect using a significance test equals the power $1 - \beta_j$ of said test. If we furthermore assume that the effect is the same for all studies assessing the same question, the expected number of significant results is given by

$$E = \sum_{j=1}^{k} (1 - \beta_j).$$

This expected number can then be compared against the actually observed number of studies reporting significant results O by means of the χ^2 test statistic

$$A = [(O - E)^2 / E + (O - E)^2 / (k - E)] \sim \chi^2_1$$

and the corresponding decision function

$$\delta(A) = \begin{cases} 1, & \text{if } A > q_{\chi^2_1, 1-\alpha}; \\ 0, & \text{otherwise.} \end{cases}$$

To calculate the power β_j of each study, Ioannidis and Trikalinos suggested using the observed aggregate effect size $\hat{\theta}_k$, while admitting that this overestimates E, because 'most biases that increase the proportion of 'positive' results may also inflate the observed summary effect size' (p. 246). To alleviate this, they recommend considering a range of effect sizes derived from external evidence.

Table 3.2 shows the results of this test for the simulated examples shown in Figure 3.1. In this case, the observed aggregate effect size $\hat{\theta}_k = \bar{X}_N$ was

calculated according to equation Eq. 3.1. Assuming σ^2 to be known, the power of the jth study is given by:

$$1 - \beta_j = \Phi(\sqrt{n_j}\frac{\bar{X}_N}{\sigma} - z_{1-\alpha})$$

as outlined in Eq. 2.5. All instances of publication bias are correctly detected.

	k	μ_1	$\bar{x}_N$	A	bias detected?
full sample (no bias, A)	200 200	0 0.3	−0.02 0.27	1.70 0.08	no no
significant studies (B)	13 38	0 0.3	0.69 0.61	11.67 19.94	yes yes
significant studies and 10% of non-significant studies (C)	32 55	0 0.3	0.15 0.51	28.64 8.21	yes yes

Table 3.2.: Results of the χ^2-test to test for concordance between the expected and observed number of significant studies. The test was conducted for the studies shown in Figure 3.1 and was performed at $\alpha = 0.05$. The global estimate $\bar{x}_N$ was calculated according to Eq. 3.1 but with theoretical parameter values replaced by empirical estimates. The power of each study j was calculated using Eq. 2.5 with τ_1 equal to the global estimate divided by the theoretical standard error of each study j.

Alternatively, Ioannidis and Trikalinos suggest to use a binomial probability test, which is only advisable, however, if the power values β_j are roughly the same across studies j. If this is not the case, it is advisable to use the Poisson binomial test statistic instead with $B_j \sim \text{Ber}(1 - \beta_j)$ and therefore $E = \sum_{j=1}^{k} B_j \sim \text{Poisbin}(k, [(1 - \beta_1), \ldots, (1 - \beta_k)])$ so that the decision function would be

$$\delta(E) = \begin{cases} 1, & \text{if } E > q_{1-\alpha}; \\ 0, & \text{otherwise;} \end{cases}$$

with $q_{1-\alpha}$ the $(1 - \alpha)$-quantile of the Poisson binomial distribution outlined as above.

3.2.3. Funnelling statistical evidence

The so-called 'funnel plot' is probably the best known and most widely adopted instrument to detect publication bias. Described as early as 1984 by Light and Pillemer (1984, p. 64–69), it has since become a standard method for meta-analysts to visually assess the extent of publication bias present in the published literature.

Funnel plots are usually constructed by plotting some measure of effect size against some measure of precision. Without publication bias, the study results should be distributed around the true effect size θ, thereby creating an inverted funnel with the most precise studies at the top of the funnel and the rest of the studies symmetrically spreading out around the population mean towards the bottom of the graph (see Figure 3.1). If publication bias is present, however, one would expect a 'gap' in the funnel where those studies with little precision and non-significant findings were expected to be.

The correct choice of the axes often depends on the question at hand (see for example Sterne, Becker and Egger (2005, p. 81–89)). Often, a summary statistic of the effect size such as the mean or the log odds ratio is used for the abscissa whereas the values on the ordinate correspond to the standard error, the variance, their respective inverse, or simply the sample size (Sterne and Egger, 2001). Researchers have shown that the variety of possible choices for the axes can severely impact the robustness of the funnel plot as a diagnostic tool, since a different axis often leads to a different assessment of the presence or absence of publication bias (Tang and Liu, 2000; Lau et al., 2006). In addition, it was shown that mere visual inspection of funnel plots is, in general, not enough to reliably detect publication bias and might even be misleading, even for experienced systematic reviewers (Terrin, Schmid and Lau, 2005). This can also be observed in Figure 3.1: Even though both funnel plots on the last row are generated from a biased sample, the proclaimed funnel asymmetry is only observable if there are no non-significant studies published (B). Adding just 10% of randomly selected non-significant studies makes the funnels appear much more symmetric. This shows that visual inspection of funnel plots can be rather misleading.

With 'Egger regression' and the rank correlation test developed by Begg and Mazdumar, two non-visual tools to detect publication bias based on the rationale of the funnel plot exist but they both suffer from similar drawbacks as the funnel plot itself if the publication bias is not very pronounced or

the number of studies is low (Sterne and Egger, 2005). The following two sections contain a quick overview over both of them.

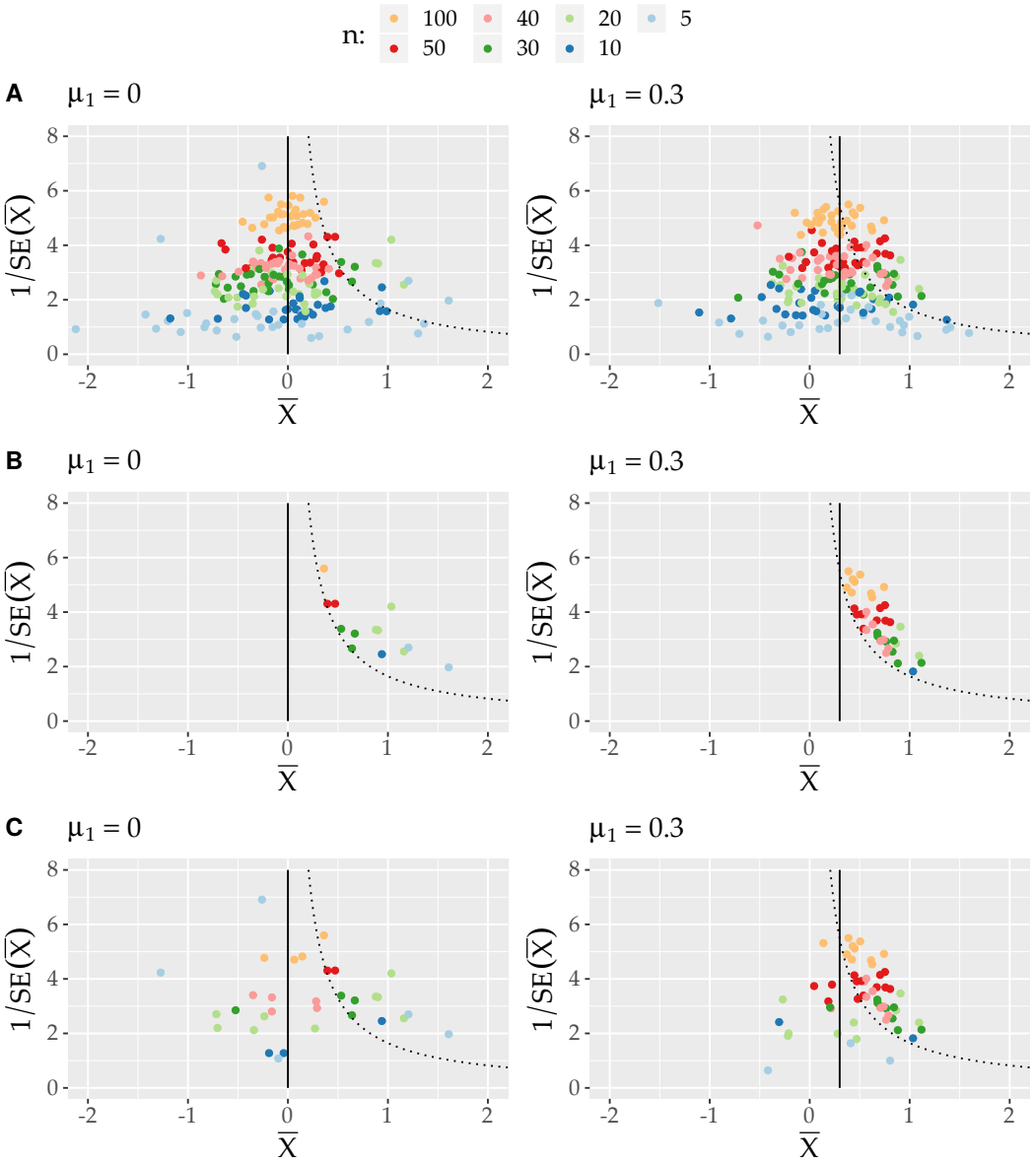

Figure 3.1.: Funnel plots of 200 randomly simulated studies with sample sizes $n \in [5, 10, 20, 30, 40, 50, 100]$ and equal sampling probability for each n. The funnel plots are shown for $H_1 : \mu = 0$ (left column) and $H_1 : \mu = 0.3$ (right column) with $H_0 : \mu = 0$ in both cases. Row A shows the complete sample, row B only significant studies and row C significant studies plus 10% of non-significant studies. The solid black line on the vertical axis denotes the true population mean whereas the dotted black line denotes the significance threshold at different levels of precision. Individual data points in each simulated study were independently drawn from a $\mathcal{N}(\mu_1, 4)$-distribution. To determine whether a given $\bar{x}_n$ was significant, V_n was calculated as described in Eq. 2.11 and then compared to the $z_{1-\alpha}$-quantile with $\alpha = 0.05$.

3.2.4. Rank correlation between effect size and standard error

Begg and Mazumdar (1994) developed a simple non-parametric test for the detection of publication bias, calling it 'a direct statistical analogue of the popular "funnel-graph" ' (p. 1088). They exploited the fact that in the presence of publication bias, 'if negative studies are less likely to be published, the [funnel] graph will tend to be skewed, inducing a negative correlation in the graph, or, expressed differently, a positive correlation between estimates of effects and their variances' (p. 1088).

Let X be a normally distributed random variable with unknown expectation μ and known variance σ^2. Suppose researchers estimate μ by the sample mean $\hat{\mu}_j = \bar{X}_{n_j} \sim \mathcal{N}(\mu, \sigma^2/n_j)$, where n_j denotes the sample size of the jth study. The global estimate of the effect size across all k studies (assuming fixed effects) can be calculated by averaging the sample mean $\bar{X}_{n_j}$ of each study j weighted by the inverse of the sampling variance of that study:

$$\bar{X}_N = \frac{\sum_{j=1}^{k} \bar{X}_{n_j} w_j}{\sum_{j=1}^{k} w_j} \sim \mathcal{N}(\mu, \sigma^2/N) \tag{3.1}$$

where $N = \sum_{j=1}^{k} n_j$ and $w_j = n_j/\sigma^2$. With this, one can center each study estimate around the global estimate of the effect size

$$(\bar{X}_{n_j} - \bar{X}_N) \sim \mathcal{N}(0, v_j)$$

with $v_j = 1/w_j - 1/\sum_{i=1}^{k} w_i$ denoting the sampling variance of the centred effect estimate $(\bar{X}_{n_j} - \bar{X}_N)$. Finally, one can stabilise the variance to 1 by dividing the centred effect estimate by the square root of its sampling variance

$$Z_j = \frac{\bar{X}_{n_j} - \bar{X}_N}{\sqrt{v_j}} \sim \mathcal{N}(0, 1) \tag{3.2}$$

which yields a sample statistic following a standard normal distribution. Both Z_j and v_j can then be used to calculate Kendall's τ (Kendall, 1938) to test for correlation between the observed standardised effect sizes z_j and their corresponding variances v_j:

$$\tau = \frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \operatorname{sgn}(z_i - z_j) \operatorname{sgn}(v_i - v_j)}{k(k-1)/2} = \frac{c - d}{k(k-1)/2}$$

where c denotes the number of concordant pairs—that is, those pairs of studies i and j for which z_i and v_i are both larger or both smaller than z_j and v_j. Conversely, d denotes the number of discordant pairs, that is, those pairs of studies i and j for which z_i is larger than z_j whereas v_i is smaller than v_j or vice versa. In other words, τ gives the ratio of concordant pairs minus discordant pairs divided by the total number of possible permutations of pairs $k(k-1)/2$. Under the null hypothesis of no correlation between effect size z and sampling variance v and with k sufficiently large, the sampling distribution of τ is approximately normal with standard error $\sigma_\tau = \sqrt{(2(2k+5))/(9k(k-1))}$ (Kendall, 1938). Hence, we can construct the following statistic to test whether one should assume publication bias

$$Z_k = \frac{\tau}{\sigma_\tau} = \frac{c-d}{\sqrt{(2k+5)k(k-1)/18}} \dot{\sim} \mathcal{N}(0,1).$$

The test statistic can then be passed to the following two-sided decision function

$$\delta(Z_k) = \begin{cases} 1, & \text{if } |Z_k| > z_{1-\alpha/2}; \\ 0, & \text{otherwise.} \end{cases}$$

with $z_{1-\alpha/2}$ denoting the $(1-\alpha/2)$-quantile of the standard normal distribution.

Going back to the simulated studies presented in Figure 3.1, we can now calculate the test statistic for each simulation. As Table 3.3 shows, the test detects publication bias when it is also clearly recognizable visually but fails to do so when the bias is less obvious.

3.2.5. Egger regression

Egger et al. (1997) used a regression-based approach to assess the correlation between the standardised effect size of a study (z_j, as described in Eq. 3.2) and the corresponding precision defined as the inverse of the empirical or theoretical standard error ($1/\sqrt{v_j}$):

$$Z_j \sim \beta_0 + \beta_1 / \sqrt{v_j}.$$

If there is no publication bias, the relationship between Z_j and $1/\sqrt{v_j}$ should be completely determined by β_1, because standardised effect sizes should

	k	μ_1	$\bar{x}_N$	$\lvert\tau/\sigma_\tau\rvert$	bias detected?
full sample (no bias, A)	200 200	0 0.3	−0.02 0.27	1.33 0.51	no no
significant studies (B)	13 38	0 0.3	0.69 0.61	2.81 5.70	yes yes
significant studies and 10% of non-significant studies (C)	32 55	0 0.3	0.15 0.51	0.36 1.60	no no

Table 3.3.: The rank correlation between effect size and variance calculated for the studies shown in Figure 3.1. The significance test was performed at $\alpha = 0.05$ and the global estimate $\bar{x}_N$ was calculated according to Eq. 3.1 but with theoretical parameter values replaced by empirical estimates.

be low, on average, for studies with high standard errors (low precision) and high, on average, for studies with low standard errors (high precision). Hence, the regression line should pass through the origin, that is, intercept $\beta_0 = 0$. However, if there is publication bias present, we should expect higher standardised effect sizes for studies with low precision.

To test whether $\hat{\beta}_0$ differs significantly from $\beta_0 = 0$, the following t-statistic can be constructed:

$$T_k = \frac{\hat{\beta}_0 - \beta_0}{s_{\hat{\beta}_0}} = \frac{\hat{\beta}_0}{s_{\hat{\beta}_0}} \sim t(\nu = k - 2)$$

with $s_{\hat{\beta}_0}$ denoting the standard error of $\hat{\beta}_0$, k denoting the total number of studies and $\nu = k - 2$ the degrees of freedom of the t-distribution. The statistic can be passed to the following two-sided decision function:

$$\delta(T_k) = \begin{cases} 1, & \text{if } |T_k| > q_{(t_{k-2}, 1-\alpha/2)}; \\ 0, & \text{otherwise.} \end{cases}$$

where $q_{(t_{k-2}, 1-\alpha/2)}$ denotes the $(1 - \alpha/2)$-quantile of the central $t(k-2)$-distribution. The Table 3.4 shows the results of the Egger regression test for the examples displayed in Figure 3.1. As before, the test is able to detect publication bias that is visually recognisable, but often fails to detect publication bias if it is less pronounced.

	k	μ_1	$\bar{x}_N$	$\lvert t \rvert$	bias detected?
full sample (no bias, A)	200	0	−0.02	1.30	no
	200	0.3	0.27	0.95	no
significant studies (B)	13	0	0.69	3.77	yes
	38	0.3	0.61	8.54	yes
significant studies and 10% of non-significant studies (C)	32	0	0.15	1.02	no
	55	0.3	0.51	4.05	yes

Table 3.4.: The Egger regression test to detect correlation between effect size and variance calculated for the studies shown in Figure 3.1. The significance test was performed at $\alpha = 0.05$ and the global estimate $\bar{x}_N$ was calculated according to Eq. 3.1 but with theoretical parameter values replaced by empirical estimates.

3.2.6. The calliper test: Pinching significance thresholds

A calliper is a mechanical instrument used to precisely measure the diameter of (small) objects by clamping them in between the two jaws of callipers. Inspired by this measuring device, Gerber and Malhotra (2006) and Gerber and Malhotra (2008) developed the so-called 'calliper test', which serves to detect distributional discontinuities of the p-value around significance levels. Without significance-driven publication bias, the number of publications reporting barely significant findings should be roughly the same as the number of publications reporting p-values just above the α-threshold.

Let us assume that there is a unknown effect size μ and known variance σ^2 so that $X \sim \mathcal{N}(\mu, \sigma^2)$. Then, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}(\mu, \sigma^2/n)$ and its variance stabilisation is $Z = \sqrt{n}\bar{X}/\sigma \sim \mathcal{N}(\sqrt{n}\mu/\sigma, 1)$. If X is not normally distributed or the variance σ^2 is unknown, these distributional properties hold asymptotically.

Let $\Phi(\cdot)$ be the cumulative distribution function of the standard normal distribution. For the limits $c < c' < c''$, the conditional probability that a Z-score lies in the interval $[c, c']$, given that it is drawn from the interval $[c, c'']$, corresponds to

$$\Pr[Z \in [c, c'] \mid Z \in [c, c'']] = \frac{\Phi(c') - \Phi(c)}{\Phi(c'') - \Phi(c)}. \tag{3.3}$$

Since Φ is continuous, it holds that $\Phi(z+e) - \Phi(z) = e\phi(z) + \epsilon$ where ϵ is an approximation error which goes to zero as e approaches zero. Therefore, the ratio described in Eq. 3.3 can be approximated by

$$\Pr[Z \in [c, c'] \mid Z \in [c, c'']] \simeq \frac{(c' - c)\phi(c)}{(c'' - c)\phi(c)} = \frac{c' - c}{c'' - c} \tag{3.4}$$

for small intervals $[c, c'']$ (that is, small e) and small changes in $\phi(\cdot)$ within the interval. If we choose c, c' and c'' so that $c' - c = c'' - c'$, then

$$p = \Pr(Z \in [c, c'] \mid Z \in [c, c'']) = 0.5 \tag{3.5}$$

that is, the conditional probability of a Z-score being an element of either the upper or lower half of the interval is fifty per cent. Eq. 3.4 and Eq. 3.5 can now be used to construct an exact binomial test to test for publication bias.

Let k be the total number of published studies and let k' be the number of studies that fall into the interval $[c, c'']$. If the Z-scores reported in the published studies represent k independent draws from $Z \sim \mathcal{N}(\sqrt{n}\frac{\mu}{\sigma}, 1)$, then the number of studies $k'' \leq k'$ falling into the upper (or lower) half of $[c, c'']$ is a realisation of a binomially distributed variable $K'' \sim \text{Bin}(k', p = 0.5)$.

To test for publication bias of results from one-sided superiority tests, one can set $c' = z_{1-\alpha}$ and $[c, c''] = [c' - e, c + e]$ with $z_{1-\alpha}$ the $(1-\alpha)$-quantile of the standard normal distribution and e small. The test hypotheses are

$$\begin{aligned} H_0 &: p = 0.5 \\ H_1 &: p \neq 0.5. \end{aligned}$$

with test statistic k'' and decision function

$$\delta(k'') = \begin{cases} 1, & \text{if } k'' > q_{1-\alpha/2}; \\ 1, & \text{if } k'' < -q_{1-\alpha/2}, \\ 0, & \text{otherwise;} \end{cases}$$

where $q_{1-\alpha/2}$ denotes the $(1-\alpha/2)$-quantile of the binomial distribution $\text{Bin}(k', p = 0.5)$.

Table 3.5 shows the results of the calliper test for the simulated studies depicted in Figure 3.1. Due to the small number of studies lying in the test

	k	μ_1	k'	k''	bias detected?
full sample (no bias, A)	200	0	4	3	no
	200	0.3	15	6	no
significant studies (B)	13	0	3	3	no
	38	0.3	6	6	yes
significant studies and 10% of non-significant studies (C)	32	0	3	3	no
	55	0.3	7	6	no

Table 3.5.: The calliper test to detect publication bias for the examples shown in Figure 3.1. The window size e was set to 0.2 and the exact binomial test was performed at $\alpha = 0.05$.

window $[c, c'']$, the test has low power in these examples and publication bias is only detected in one case. As opposed to the file drawer calculations proposed by Rosenthal (1979), the calliper test can theoretically also detect publication bias due to p-hacking and HARKing because these manipulations directly influence the conditional probabilities of finding studies in the upper or lower part of a small interval around the significance threshold.

It should be noted, however, that the test is only approximately accurate if the sampling distribution of Z has a mean value close to the critical value, that is, the $(1-\alpha)$-quantile of the standard normal distribution. If the sampling distribution is not symmetric around the critical value, $\Pr(z \in [c, c+e] \mid z \in [c-e, c+e])$ does not equal 0.5 anymore. However, for very large deviations of the mean value from the critical value, the overall probability of any values falling into $[c-e, c+e]$ is negligible. Coincidentally, in those cases a test for significance-driven publication bias is rendered moot, since all or most of the results are expected to be significant because of the presence of a strong real effect.

3.2.7. The p-curve

As already mentioned in Section 3.2.1, not all methods that try to detect significance-driven publication bias are robust against p-hacking and HARKing. The p-curve test was specifically designed by Simonsohn, Nelson and

Simmons (2014a) to provide a robust method to detect publication bias in the presence of such manipulations. The rationale of the test is based on the observation that the distribution of p-values is uniform if the null hypothesis is true[1] but should be right-skewed if the alternative holds (see Section 2.1.4 for more details).

Simonsohn *et al.* proposed using these properties by aggregating all significant p-values pertinent to a specific set of hypotheses in order to assess whether a body of published studies contained any evidential value or not. Let

$$\gamma = \frac{1/n \sum_{j=1}^{l} (p_j - \bar{p}_l)^3}{[1/(n-1) \sum_{j=1}^{l} (p_j - \bar{p}_l)^2]^{3/2}}$$

be the sample skewness of p below the significance threshold. We can then distinguish between the following four cases:

1. H_0 true and p-hacking absent:	$P \sim \text{Unif}(0,1)$	(no skew);
2. H_0 true and p-hacking present:	$\gamma < 0$	(left skew);
3. H_0 false and p-hacking absent:	$\gamma > 0$	(right skew);
4. H_0 false and p-hacking present:	$\gamma < 0$	(left skew) or
	$\gamma > 0$	(right skew).

Cases 1 to 3 can relatively easy be distinguished, for example by visual inspection of the empirical distribution of p-values or by calculating the sample skewness and testing whether it significantly deviates from zero skew expected under $P \sim \text{Unif}(0,1)$. Distinguishing Case 4 from Cases 2 and 3 is more subtle, however. Simnsohn *et al.* did not offer a method to do so, but instead suggested to reframe the problem as a test for the presence or absence of evidential value.

To test whether a certain set of significant findings l contains real evidence against the null, one first calculates for each observed p-value the probability of observing a p-value that is at least as extreme if the null were true and given that the observed p-value is significant, that is, $\Pr(P \leq p \mid p < \alpha)$. Simonsohn *et al.* call this the 'pp-value'—the p-value of the p-value. They then combine the pp-values from all k significant findings to construct the

[1] This only holds for simple null hypothesis and composite null hypotheses evaluated at the most unfavourable null parameter value.

test statistic for Fisher's combined probability test statistic

$$pp_{\text{comb}} = -2 \sum_{j=1}^{l} \ln(pp_j)$$

which follows a χ^2_{2s}-distribution if the pp-values are drawn from a Unif$(0,1)$-distribution. Hence, the test statistic can be passed to the following decision function to test for significant skewness:

$$\delta(pp_{\text{comb}}) = \begin{cases} 1, & \text{if } pp_{\text{comb}} > q_{\chi^2_{2s},1-\alpha}; \\ 0, & \text{otherwise.} \end{cases}$$

If the distribution p-values are significantly left skewed, that is, $\gamma < 0$ and $\delta(pp_{\text{comb}}) = 1$, the analysed results were probably p-hacked. If the distribution of p-values is significantly right-skewed, ($\gamma > 0$ and $\delta = 1$), p-hacking might still have occurred, but one can at least conclude that the studies reporting significant findings did indeed contain some evidential value against the null. However, if Fisher's combined probability test turns out to be non-significant, this may indicate either that there are not enough findings to draw any conclusion or that the findings did not contain evidential value, that is, that they originated from p-hacking.

To distinguish between these two cases, Simonsohn *et al.* suggested conducting a second test, but this time against a null hypothesis that assumes a very small effect instead of no effect. Specifically, they proposed to compare the observed p-curve to a p-curve that would be expected for studies with a low power of 0.33. This curve can be constructed recalculating the pp-values under the assumption that the underlying p-values originated from studies with a small real effect and corresponding power of 0.33.

Let us follow through with the example inspired by Simonsohn *et al.* and assume that the original p-values were calculated using Student's t-statistic (see 2.4.1 for more details). Under the null hypothesis H_0, the t-statistic follows a central $t(\nu)$-distribution, with $\nu = n - 1$ degrees of freedom and n denoting the sample size used to calculate the statistic. The corresponding p-value is given by

$$p_{H_0} = \Pr(T > t_n \mid H_0) = 1 - F_T(t_n \mid H_0)$$

with $F_T(\cdot \mid H_0)$ the cumulative distribution function of the central $t(\nu)$-distribution.

Under the alternative hypothesis H_1, the t-statistic follows a non-central $t(\lambda, \nu)$-distribution, with non-centrality parameter $\lambda = \sqrt{n}(\mu - \mu_0)/\sigma$ and ν and n same as above.

Given a t-statistic t_n, sample size n and the $(1-\alpha)$-quantile of the central $t(\nu)$-distribution $q_{1-\alpha}$, we can find the non-centrality parameter λ corresponding to a power of $1-\beta$ as follows:

$$\begin{aligned} & \Pr(T > q_{1-\alpha} \mid H_1) && = 1-\beta \\ \iff & 1 - F_T(q_{1-\alpha} \mid H_1) && = 1-\beta \\ \iff & F_T^{-1}(F_T(q_{1-\alpha} \mid H_1) \mid H_1) && = F_T^{-1}(\beta \mid H_1) \\ \iff & F_T^{-1}(\beta \mid H_1) && = q_{1-\alpha} \end{aligned}$$

with $F_T(\cdot \mid H_1)$ and $F_T^{-1}(\cdot \mid H_1)$ the cumulative distribution function and the quantile function of T under the alternative hypothesis. In other words, we need to find the non-central $t(\lambda, \nu)$-distribution whose β-quantile is equal to the $q_{1-\alpha}$-quantile of the central $t(\nu)$-distribution. Having done that, we can calculate the p-value of t_n under the alternative which amounts to

$$p_{H_1} = \Pr(T > t_n \mid H_1) = 1 - F_T(t_n \mid H_1).$$

The corresponding pp-value is then

$$\begin{aligned} \Pr(P < p_{H_1} \mid p_{H_0} < \alpha, H_1) &= \frac{\Pr(P < p_{H_1}, p_{H_0} < \alpha \mid H_1)}{\Pr(p_{H_0} < \alpha) \mid H_1)} \\ &= \frac{\Pr(P < p_{H_1}, p_{H_1} < 1-\beta \mid H_1)}{\Pr(p_{H_1} < 1-\beta \mid H_1)} \\ &= \frac{\Pr(P < p_{H_1} \mid H_1) - (\Pr(p_{H_1} \geq 1-\beta \mid H_1))}{\Pr(p_{H_1} < 1-\beta \mid H_1)} \\ &= \frac{p_{H_1} - \beta}{1-\beta}. \end{aligned}$$

By calculating these alternative pp-values for each observation, one can construct a reference p-curve for assumed power $1-\beta$ that can be tested against the null hypothesis of uniform distribution of the p-values. If the reference p-curve is significantly right-skewed and thus more right skewed than the observed p-curve, one can conclude that the analysed studies contain less evidential value than the same amount of studies with power $1-\beta$. If one picks $1-\beta$ so that the studies would be clearly underpowered (such as $1-\beta = 0.33$), this would mean that evidential value is rather low.

However, if the observed p-curve is not significantly less right-skewed than the newly constructed reference curve, one cannot make a judgement about the presence or absence of any evidential value based on the p-curve.

As Simonsohn, Nelson and Simmons (2014a) show, their p-curve method has high power to detect that a set of studies has evidential value or lack thereof, even if the individual studies have rather low power or are even slightly p-hacked, respectively. However, an important prerequisite of these results is the judicious definition and use of selection criteria with which to select p-values from individual studies. For example, the p-curve analysis does only work if the analysed p-values are independent from each other, if they have a uniform distribution under the null and if they are linked to the same hypothesis of interest.

	k	μ_1	γ	pp_{comb}	bias detected?
full sample (no bias, A)	200 200	0 0.3	0.55 0.45	44.94 116.67	no no
significant studies (B)	13 38	0 0.3	0.55 0.45	44.94 116.67	no no
significant studies and 10% of non-significant studies (C)	32 55	0 0.3	0.55 0.45	44.94 116.67	no no

Table 3.6.: Results of the p-curve test to check for uniformity of significant p-values under the null hypothesis. The test was conducted for the studies shown in Figure 3.1 at an α-level of 0.05.

In addition, if p-hacking and HARKing are completely absent and publication bias originates exclusively from an increased publication probability for significant findings, the p-curve fails to detect any bias. This happens because the p-curve method only looks at p-values below the significance threshold whose distribution remains unchanged in the absence of p-hacking and HARKing regardless of how many non-significant findings are suppressed. The p-curve tests conducted for the simulated studies shown in Figure 3.1 confirm this. As can be seen in Table 3.6, the tests yields exactly the same result in all three scenarios.

3.3. Correcting biased estimates

The detection of publication bias is an important endeavour in itself, lest biased meta-analyses are used as a basis for future research or even policy decision. This is especially important in the case of the most extreme form of publication bias, in which all published findings about a specific questions are simply a false positive. However, having detected both the presence of a real effect as well as publication bias, it would be desirable to have tools at hand with which to correct biased effect size estimates. The following sections present a selection of methods to do so.

3.3.1. Publication probabilities and truncated distributions

In the absence of publication bias—and assuming no other major source of bias—the distribution of published findings is given by the sampling distribution of the corresponding summary statistic. Let $X \sim \mathcal{N}(\mu, \sigma^2)$ be a random variable from which data points are drawn and let the corresponding summary statistic be $S_{n_j} = \bar{X}_{n_j} = \frac{1}{n_j} \sum_{i=1}^{n_j} X_i \sim \mathcal{N}(\mu, \sigma^2/n_j)$ for each study j.

If we calculate the mean of $\bar{X}_{n_j}$ weighted by the inverse of the standard error we get the global effect size estimate

$$\bar{X}_N = \frac{\sum_{j=1}^{k} \bar{X}_{n_j} w_j}{\sum_{j=1}^{k} w_j}$$

which was already outlined in Eq. 3.1. $N = \sum_{j=1}^{k} n_j$ is again the total number of individual samples summed over all studies and w_j is the squared inverse of the theoretical ($w_j = n/\sigma_j^2$) or estimated ($w_j = n/s_j^2$) standard error of each estimate j.

In the presence of publication bias, however, the global estimate is biased upwards. The following sections introduce methods to correct this, most of which rely on the following definitions.

Publication probability: Let $X_1, \ldots, X_n$ be a set of i.i.d random variables with expectation μ. I define the publication probability of a finding based on a test statistic $S_n(X_1, \ldots, X_n)$ and given μ as

$$\text{ppr}(S_n, \pi \mid \mu) = \pi + (1 - \pi)\delta(S_n(X_1, \ldots, X_n \mid \mu)) \tag{3.6}$$

where $\delta(\cdot)$ is a decision function taking S_n as input and returning 1 if S_n lies beyond the predefined significance threshold $q_{1-\alpha}$ and 0 otherwise. Hence,

$$\delta(S_n(X_1, \ldots, X_n \mid \mu)) = D \sim \text{Ber}(p = \Pr(S_n(X_1, \ldots, X_n \mid \mu) > q_{1-\alpha})).$$

In other words, if a finding is significant, its publication probability is assumed to be equal to 1, if it is not, it is assumed to be equal to π. For the methods outlined below, π is assumed to be a constant, but it could easily be replaced by function $\pi(\cdot)$, dependent on effect size, sample size, or other study properties.

Expected publication probability given n and μ: The expectation of the publication probability ppr is given by

$$\text{E}[\text{ppr}(S_n, \pi \mid \mu)] = \int_0^1 p f_{\text{ppr}}(p \mid \mu) dp \tag{3.7}$$

with $f_{\text{ppr}}(\cdot \mid \mu)$ denoting the probability density function of ppr given μ. If we assume π to be constant, this simplifies further to

$$\text{E}[\text{ppr}(S_n \mid \pi, \mu] = \pi \Pr(\text{ppr} = \pi \mid \mu) + \Pr(\text{ppr} = 1 \mid \mu) \tag{3.8}$$

Truncated probability density function of S_n: Let $f_{S_n|\mu}(\cdot \mid \mu)$ be the probability density function of S_n given μ. If publication bias is present, the truncated probability density function $f^*_{S_n|\mu}(\cdot \mid \mu)$ given μ is defined by

$$f^*_{S_n|\mu}(s_n \mid \mu) = \frac{\text{ppr}(s_n, \pi \mid \mu)}{\text{E}[\text{ppr}(S_n, \pi \mid \mu)]} f_{S_n|\mu}(s_n \mid \mu). \tag{3.9}$$

Truncated likelihood of μ: The likelihood of μ given the observed test statistics $s_{n_1}, \ldots, s_{n_k}$ is defined as $\mathcal{L}(\mu \mid s_{n_1}, \ldots, s_{n_k}) = \prod_{j=1}^k f_{S_{n_j}|\mu}(s_{n_j} \mid \mu)$ with $f_{S_{n_j}|\mu}(s_{n_j} \mid \mu)$ denoting the probability density function of S_{n_j} given μ. In the presence of publication bias, I define the truncated likelihood of μ as

$$\begin{aligned} \mathcal{L}^*(\mu \mid s_{n_1}, \ldots, s_{n_k}) &= \prod_{j=1}^k \frac{\text{ppr}(s_{n_j}, \pi \mid \mu)}{\text{E}[\text{ppr}(S_{n_j}, \pi \mid \mu)]} f_{S_{n_j}|\mu}(s_{n_j} \mid \mu) \\ &= \frac{\mathcal{L}(\mu \mid s_{n_1}, \ldots, s_{n_k})}{\text{E}[\text{ppr}(S_{n_j}, \pi \mid \mu)]} \prod_{j=1}^k \text{ppr}(S_{n_j}, \pi \mid \mu) \end{aligned} \tag{3.10}$$

The measures described above are inspired by Andrews and Kasy (2017), but I use them for bias corrections methods that are different from the ones proposed by the two authors.

3.3.2. Reweighting estimates by publication probabilities

Let $S_{n_j} = \bar{X}_{n_j} = \frac{1}{n_j}\sum_{i=1}^{n_j} X_i \sim \mathcal{N}(\mu, \sigma^2/n_j)$ be the summary statistic reported by a study j. To estimate the true effect size μ in the absence of publication bias, one can simply apply Eq. 3.1. However, if publication bias is present, this estimator is biased upwards.

In order to find the distribution of the effect size under publication bias, we first need to know the distribution of $\bar{X}_{n_j}$ in the absence of publication bias given that the corresponding test statistic S_{n_j} is above or below the significance threshold, respectively.

Let $D = \delta(S_n(X_1, \ldots, X_n \mid \mu))$ be Bernoulli distributed random variable as described in Eq. 3.6 so that $\mu_{\text{sig}} = \mathrm{E}[\bar{X}_{n_j} \mid D = 1]$ and $\mu_{\text{ns}} = \mathrm{E}[\bar{X}_{n_j} \mid D = 0]$ denote the expected value of $\bar{X}_{n_j}$ given that the effect is significant and non-significant, respectively. Similarly, let the variances be $\sigma^2_{\text{sig}} = \mathrm{Var}[\bar{X}_{n_j} \mid D = 1]$ and $\sigma^2_{\text{ns}} = \mathrm{Var}[\bar{X}_{n_j} \mid D = 0]$.

In the absence of publication bias, the distribution of $\bar{X}_{n_j} \mid S_n$ is thus given by

$$\bar{X}_{n_j} \mid D_j \sim \begin{cases} \mathcal{N}\left(\mu_{\text{sig}}, \frac{\sigma^2_{\text{sig}}}{n_j}\right), & \text{if } D_j = 1 \text{ (i.e. } S_{n_j} > q_{1-\alpha}); \\ \mathcal{N}\left(\mu_{\text{ns}}, \frac{\sigma^2_{\text{ns}}}{n_j}\right), & \text{if } D_j = 0 \text{ (i.e. } S_{n_j} \leq q_{1-\alpha}). \end{cases}$$

Next, we need to find the distribution of $ppr(S_{n_j}, \pi \mid \mu)$ given S_{n_j}. Recalling Eq.3.6, we know that $ppr(S_{n_j}, \pi \mid \mu)$ follows a Bernoulli distribution. However, by conditioning on D_j, it turns into a constant and we can write

$$ppr_j \mid D_j = ppr(S_{n_j}, \pi \mid D_j, \mu) = \begin{cases} 1, & \text{if } D_j = 1; \\ \pi, & \text{if } D_j = 0. \end{cases}$$

In the presence of publication bias, non-significant results are published less often then significant results. This can be modelled by multiplying each result $\bar{X}_{n_j}$ by its corresponding publication probability ppr_j. Hence, the distribution of $\bar{X}_{n_j|S_{n_j}} = \bar{X}_{n_j} \mid S_{n_j}$ in the presence of publication bias can be written as

$$\bar{X}_{n_j|D_j}\,\mathrm{ppr}_j \sim \begin{cases} \mathcal{N}\left(\mu_{\text{sig}}\,\mathrm{ppr}_j, \sigma^2_{\text{sig}}\frac{\mathrm{ppr}_j^2}{n_j}\right), & \text{if } D_j = 1; \\ \mathcal{N}\left(\mu_{\text{ns}}\,\mathrm{ppr}_j, \sigma^2_{\text{ns}}\frac{\mathrm{ppr}_j^2}{n_j}\right), & \text{if } D_j = 0. \end{cases}$$

Without loss of generality, we can assume that all significant results have indices $j \in [1, \ldots, s]$ and all non-significant results have indices $j \in [s+1, \ldots, k]$. We can then write the biased weight $w_j^* = w_j \, \text{ppr}_j$ and the global estimator given in Eq. 3.1 then turns into

$$\bar{X}_N = \frac{\sum_{j=1}^{k} \bar{X}_{n_j} w_j^*}{\sum_{j=1}^{k} w_j^*} \sim \mathcal{N}(\mu^*, \sigma^*) \tag{3.11}$$

with

$$\mu^* = \frac{\sum_{j=1}^{s} w_j^* \mu_{\text{sig}} + \sum_{j=s+1}^{k} w_j^* \mu_{\text{ns}}}{\sum_{j=1}^{k} w_j^*}$$

$$\sigma^* = \frac{\sum_{j=1}^{s} (w_j^*)^2 \sigma_{\text{sig}} / n_j + \sum_{j=s+1}^{k} (w_j^*)^2 \sigma_{\text{ns}} / n_j}{(\sum_{j=1}^{k} w_j^*)^2}$$

Eq. 3.11 yields a biased estimator as soon as the publication probability for significant results differs from the publication probability of non-significant results. To correct the bias, on can simply reweight each biased weight w^* by the inverse of the publication probability of the corresponding study, that is, $w_j = w_j^* / \text{ppr}_j$ and then calculate the unbiased global estimate using Eq. 3.1 which yields the following unbiased estimator:

$$\bar{X}_N^* = \frac{\sum_{j=1}^{k} \bar{X}_{n_j} w_j^* / \text{ppr}_j}{\sum_{j=1}^{k} w_j^* / \text{ppr}_j} \sim \mathcal{N}\left(\mu, \sigma^2 / N\right) \tag{3.12}$$

This method—inspired by the finite sample estimator developed by Hansen and Hurwitz (1943)—performs well when there is a real but biased effect but underestimates the global effect size the assumed or estimated publication probability is smaller than the real one. For biased null effects the estimator also tends to underestimate the true effect size if In scenarios in which there are only significant findings observed, it completely fails (see Table 3.7 for examples). In addition, the method is only applicable if the publication probability for a given study is known or can be estimated.

3.3.3. Trim-and-fill: Closing gaps in funnel plots

The 'trim-and-fill' method developed by Duval and Tweedie (2000b) and Duval and Tweedie (2000a) leverages on a similar rationale as the funnel plot

	k	μ_1	$\bar{x}_N$	$\bar{x}^*_N$
full sample (no bias, A)	200 200	0 0.3	−0.02 0.27	−0.07 0.15
significant studies (B)	13 38	0 0.3	0.69 0.61	0.69 0.61
significant studies and 10% of non-significant studies (C)	32 55	0 0.3	0.15 0.51	−0.17 0.27

Table 3.7.: The results of the bias correction based on reweighting each result by the inverse of its respective publication probability. The correction was applied to the studies shown in Figure 3.1, the global estimate $\bar{x}_N$ was calculated according to Eq. 3.1 and the corrected estimate $\bar{x}^*_N$ was calculated according to Eq. 3.12—in both cases with theoretical parameter values replaced by empirical estimates.

introduced in Section 3.2.3. It assumes that publication bias suppresses those findings which are most 'negative' in the sense of 'pointing in the opposite direction of significant findings'. As a result, we would expect a considerable gap in the lower left part of the funnel plot, where the those studies with low precision and negative effect sizes are located, and—conversely—a global effect size estimate that is biased upwards.

To correct for this, Duval and Tweedie propose an expectation-maximisation algorithm based on trimming the most 'positive' findings, calculating a corrected global effect size estimate, again trimming the most positive findings with regard to the corrected estimate, re-correcting said estimate and so on, until convergence is achieved. Then, the most positive findings with regard to the converged global effect size estimate are mirrored around said estimate and a final global effect size estimate is calculated. In detail, the algorithm consists of the following stages (adapted from Duval, 2005):

1. Transform the individual findings $\hat{\theta}_1, \ldots, \hat{\theta}_k$ into standard normally distributed test statistics $V_{n_1}, \ldots, V_{n_k}$ using a variance stabilising transformation.
2. Set $i = 1$, $k_0^{(0)} = 0$ and calculate a global estimate $V_N^{(i)}$ of the individual test statistics V_{n_j}.
3. Centre the findings around $V_N^{(i)}$, that is, $V^*_{n_j} = V_{n_j} - V_N^{(i)}$.
4. Rank all centred estimates according to their absolute value from smal-

lest to largest and assign each rank the sign of its corresponding $V^*_{n_j}$. This yields signed ranks $\text{sr}_{V^*_{n_j}} = \text{sgn}(V^*_{n_j}) \cdot \text{rank}_{|V^*_{n_1}|,\dots,|V^*_{n_k}|}(|V^*_{n_j}|)$, with $\text{rank}_{|V^*_{n_1}|,\dots,|V^*_{n_k}|}(\cdot)$ denoting the rank function with regard to absolute values of all centred estimates.

5. Sum all ranks with positive sign to obtain the rank sum

$$S_{\text{rank}} = \sum_{j=1}^{k} \text{sr}_{V^*_{n_j}} \cdot \mathbb{1}(\text{sr}_{V^*_{n_j}} > 0).$$

6. Estimate the number of potentially omitted studies

$$k_0^{(i)} = \left\lceil \frac{4S_{\text{rank}} - n(n+1)}{2n-1} \right\rceil.$$

7. If $k_0^{(i)} = k_0^{(i-1)}$ go to step 10. Otherwise, continue with step 8.
8. Trim off the $k_0^{(i)}$ most positive findings V_{n_j} and recalculate the global estimate of the trimmed sample $V_N^{(i+1)}$.
9. Set $i = i + 1$ and restart from step 3 using all original findings (including the ones that were trimmed in step 8).
10. Take the $k_0^{(i)}$ most positive findings V_{n_j}, 'mirror' them around $V_N^{(i)}$ and add them to the data set along with the corresponding standard error.
11. Transform all V_{n_j} back into effect size estimates $\hat{\theta}_j$ (including the newly filled data) and calculate a global effect size estimate $\hat{\theta}_N$.

Since the trim-and-fill method hinges on the assumption that only the most extreme negative findings are omitted in the presence of publication bias, the correction will fail if the publication probability of a finding $\hat{\theta}_i$ does not decay with increasing negative distance from the significance threshold. If—for example—the publication probability of a finding is given by $\text{ppr}_j = \text{ppr}(V_{n_j}, \pi_j)$ as described in Section 3.3.1, Eq. 3.6, with $V_{n_j} = V_{n_j}(\hat{\theta}_j)$ a variance stabilised and normally distributed test statistic based on $\hat{\theta}_j$ and π_j either constant across all j or a function dependent on some study characteristic of interest, then the publication probability will be equal to 1 above the significance threshold and π_j below. In such cases, the trim-and-fill estimator will overestimate the true effect size as can be seen in Table 3.8.

If there is reason to belief that the ppr_j represents the publication probability of a finding j below the significance threshold better than the decaying probability assumed by Duval and Tweedie, I propose the following adjustment

	k	μ_1	$\bar{x}_N$	$\bar{x}_N^{(\text{t\&f})}$	$\bar{x}_N^{(\text{t\&f-ppr})}$
full sample (no bias, A)	200 200	0 0.3	−0.02 0.27	−0.11 0.27	−0.10 0.10
significant studies (B)	13 38	0 0.3	0.69 0.61	0.46 0.45	0.46 0.45
significant studies and 10% of non-significant studies (C)	32 55	0 0.3	0.15 0.51	0.11 0.42	−0.002 0.40

Table 3.8.: Results for the trim-and-fill method and the adapted trim-and-fill method to correct for effect size estimates. The correction was applied to the studies shown in Figure 3.1. The global estimate for V_{n_j} was calculated by using the arithmetic mean. The global estimate $\bar{x}_N$ was calculated according to Eq. 3.1 but with theoretical parameter values replaced by empirical estimates.

to their trim-and-fill method, to be executed after the first execution of step 1 and before the first execution of step 3 described above:

2.1. Set $i = 1$, but instead of starting out with $k_0^{(0)} = 0$, set it equal to the minimum of the number of all significant findings and the number of all non-significant findings reweighted by their respective publication probability ppr_j, that is,

$$k_0^{(0)} = \min\left(\sum_{j=1}^{k} \delta(V_{n_j}), \sum_{j=1}^{k} (1 - \delta(V_{n_j}))/\,\text{ppr}_j\right)$$

where $\delta(\cdot)$ is a decision function taking S_n as input and returning 1 if V_n lies beyond the predefined significance threshold and 0 otherwise.

2.2. Trim off the $k_0^{(0)}$ must positive findings and calculate the global estimate $\hat{\theta}_{\text{tot}}^{(i)}$ based on the remaining data.

2.3. Enter the trim-and-fill algorithm at step 3 and iterate until convergence.

Unless there is no publication present or the publication probability of non-significant studies is larger than assumed, the corrections outlined above improve the performance of the trim-and-fill method as can also be observed in Table 3.8.

3.3.4. Effect size correction based on p-curves

The p-curve described in Section 3.2.7 cannot only be used for the detection of publication bias, but also for its correction. The correction method put forward by Simonsohn, Nelson and Simmons (2014b) leverages again on the different distributions of p-values depending on whether the null hypothesis is true or false, respectively, and whether p-hacking is present or absent, respectively (see Section 3.2.7 for details). In short, the researcher calculates the pp-values for the published findings for a range of possible true effect sizes and then checks to resulting p-curves for correspondence to a uniform distribution. In detail, the procedure goes as follows (assuming that findings result from a one-sided superiority test):

1. Let $s_{n_1}, \ldots, s_{n_l}$ be a set of published and significant test statistics with $S_{n_j} \sim \mathcal{N}(\mu, \sigma^2/n_j)$. Create a set of possible candidates $\{\mu_1, \ldots, \mu_m\}$ for μ.
2. For each μ_i calculate the corresponding p-values for $s_{n_1}, \ldots, s_{n_l}$, that is, $p_{ij} = \Pr(S_{n_j} > s_{n_j} \mid \mu_i) = 1 - F_{S_n}(\cdot \mid \mu_i)$, with $F_{S_{n_j}}(s_{n_j} \mid \mu_i)$ denoting the cumulative distribution function of S_{n_j} given $\mu = \mu_i$.
3. Transform the p_{ij}-values into pp_{ij}-values by conditioning on the fact that all observed p-values were significant and thus below the predefined Type I error rate α:

$$pp_{ij} = \Pr(P < p_{ij} \mid p_{ij} < 1 - \beta_{ij}, \mu_i) = \frac{p_{ij} - \beta_{ij}}{1 - \beta_{ij}}.$$

4. For each μ_i, compare the corresponding p-curve consisting of pp-values $pp_{i1}, \ldots, pp_{il}$ to the uniform distribution. Simonsohn *et al.* suggested using the Kolmogorov-Smirnov statistic (see for example Barlow, 1989, p. 155), as a measure of distance between of the empirical p-curve and the uniform distribution, that is,

$$D_i = \max_j(|pp_{ij} - 1/\alpha|).$$

5. Pick the μ_i with the lowest D_i as corrected estimate of the true effect size μ.

Table 3.9 shows the results for the corrections applied to the studies shown in Figure 3.1. As shown before (see Table 3.6), the p-curve discards all information from non-significant studies and therefore is not able to distinguish between different extents of 'pure' publication bias, that is, publication

bias that only originates from a higher publication probability for significant results. Simonsohn *et al.* showed that effect size corrections using the p-curve

	k	μ_1	$\bar{x}_N$	$\bar{\mu}$
full sample (no bias, A)	200 200	0 0.3	−0.02 0.27	0.41 0.25
significant studies (B)	13 38	0 0.3	0.69 0.61	0.41 0.25
significant studies and 10% of non-significant studies (C)	32 55	0 0.3	0.15 0.25	0.41 0.27

Table 3.9.: The results of the bias correction based on the p-curve method. The correction was applied to the studies shown in Figure 3.1, the global estimate $\bar{x}_N$ was calculated according to Eq. 3.1 with the theoretical parameter values replaced by empirical estimates.

outperforms the standard trim-and-fill method in the case of publication bias. If p-hacking or HARKing is the cause of the publication bias (and not only bias based on preferential publishing of significant findings alone), then the p-curve correction still performs well, but underestimates the true effect size.

3.3.5. Maximising the truncated likelihood

Yet another possibility to correct biased effect size estimates is possible by exploiting the truncated likelihood function as outlined in Section 3.3.1, Eq. 3.10. Let $v_{n_1}, \ldots, v_{n_k}$ be the variance stabilised and standard normally distributed test statistics for the true population mean μ in k studies with n_j data points each. Furthermore, let the publication probability of significant and non-significant studies be 1 and π, respectively, with π being constant. We can then write the truncated likelihood for the μ as:

$$\mathcal{L}^*(\mu \mid v_{n_1}, \ldots, v_{n_k}) = \frac{\mathcal{L}(\mu \mid v_{n_1}, \ldots, v_{n_k})}{\mathrm{E}[\mathrm{ppr}(V_{n_j}, \pi) \mid \mu]} \prod_{j=1}^{k} \mathrm{ppr}(V_{n_j}, \pi).$$

Estimated values or both μ and π can now easily be computed by maximising the likelihood through grid-search optimisation:

1. Define a set of candidate values $\{\mu_1, \ldots, \mu_m\}$ and $\{\pi_1, \ldots, \pi_n\}$ for μ and π, respectively.
2. For each combination of μ and π, calculate the likelihood.
3. Choose the candidate value for μ and π that yields the highest likelihood.

As Table 3.10 shows, this approach yields corrected estimates for μ and π that are rather close to the unbiased effect sizes and performs robustly across different scenarios.

	k	μ_1	$\bar{x}_N$	$\hat{\mu}$	$\hat{\pi}$
full sample (no bias, A)	200 200	0 0.3	−0.02 0.27	−0.02 0.27	0.66 1
significant studies (B)	13 38	0 0.3	0.69 0.61	0.31 0.31	0 0
significant studies and 10% of non-significant studies (C)	32 55	0 0.3	0.15 0.51	−0.07 0.33	0.05 0.16

Table 3.10.: Results for the effect size corrections and publication probability estimates based on the maximisation of the truncated likelihood. The correction was applied to the studies shown in Figure 3.1. The global estimate $\bar{x}_N$ was calculated according to Eq. 3.1 but with theoretical parameter values replaced by empirical estimates.

4. Conclusion

'In theory, practice is simple.'

— Trygve M. H. Reenskaug

In this thesis, I presented an overview of a range of methods to construct robust and comparable measures for statistical evidence. In addition, I presented different approaches to detect and correct publication bias in effect size measures. Even though all of these methods and approaches are backed up by solid theoretical arguments, they often hinge on assumptions about the nature of scientific studies and the publication process. Assumptions which might be true but which might also be completely wrong.

For example, all tests introduced in this thesis as well as my adaptions thereof assume that individual studies are independent, that there is no between-study heterogeneity and that there is a fixed global effect. All these assumptions might be and are usually violated in real-life scientific practice.

Hence, it is important to keep in mind that none of the methods explained can be regarded as a 'silver bullet' to fight publication bias—this is especially true, when one only relies on one of these methods alone. For example, it should have become clear that the reliance on the funnel plot as a diagnostic tool and the trim-and-fill approach as a corrective method is woefully inadequate to capture the wide range of forms in which publication bias can appear. Hence, it should become common practice to use more than one approach with different underlying assumptions to detect and correct publication bias.

In this spirit, I would like to pursue this topic with the following three goals in mind:

Systematic evaluation of performance The scope of this did not allow me to systematically test the performance of all methods described. Hence, I would like to do this for a range of different scenarios of publication bias

based on different assumptions and empirical observations. This is important since many of the published methods only perform well in very specific scenarios and completely fail if certain theoretical assumptions are not met. This also holds for my own methods presented in this thesis. It is therefore crucial to systematically asses the performance of all of these methods in order to construct reliable usage recommendations for statisticians and non-statisticians alike.

Extend assessment to additional methods This thesis is not an exhaustive summary of methods pertaining to the detection and correction of publication bias. For example, I did not go into methods accounting for between-study heterogeneity and applying random effects models (Piao et al., 2018) or Bayesian approaches (Cleary and Casella, 1997; Andrews and Kasy, 2017), and hardly touched maximum likelihood-based approach such as the ones proposed by Copas (1999). Since these approaches often hinge on assumptions that differ from those underlying the methods described in this thesis, it is most likely advantageous to include them at a later step in order to increase the range of scenarios in which publication bias can be detected.

Creating ensemble models My simulations have shown that many standard methods to detect and correct publication bias fail in certain situations and can even be further improved from a practical and sometimes even from a theoretical point of view. Given the fact that different methods hinge on different assumptions and thus show different performances depending on the concrete scenario, it should be clear that reliance on only one approach to detect or correct publication bias is prone to yield misleading results. It is therefore desirable to combine some of these methods into ensemble models for the detection and correction of publication bias. Since many of the methods described in this thesis are based on similar or related statistical measure, aggregating results across different methods should be possible.

However, I must point out the observation that the best statistical tools to correct biased estimates pale in comparison with non-statistical methods when it comes to efficacy to prevent bias in scientific results. For example, a strict distinction between exploratory and confirmatory studies—with mandatory pre-registration for the latter—could alleviate a lot of the problems outlined in this thesis. For a start, it would be possible to select only those studies that were intended to be confirmatory for meta-analysis and ignore

the rest. In addition, it would be possible to calculate the publication probability of non-significant findings to correct for any kind of publication bias. And last but definitely not least it would help increasing the overall quality of scientific work if researchers are forced to think about study designs and statistical methods before the onset of the study—hopefully also by consulting a statistician before and not, as is often the case, after a study. Because, as Fisher (1938, p. 17) already remarked: 'To consult a statistician after an experiment is finished is often merely to ask him to conduct a *post mortem* examination. He can perhaps say what the experiment died of."

Affidavit

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used.

Date

Signature

Appendix

Appendix A.

Package Overview and Code Repository

A.1. Repository on Github

All simulations and figures presented in this thesis were produced using the R programming language. The corresponding code can be found on Github under https://github.com/segrue/2019_WeightOfStatisticalEvidence.git. For remarks, corrections or inquiries about the thesis you can reach me under the contact information given on my website www.servangrueninger.ch.

A.2. R-packages

```
R version 3.6.0 (2019-04-26)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 16.04.6 LTS

Matrix products: default
BLAS:   /usr/lib/libblas/libblas.so.3.6.0
LAPACK: /usr/lib/lapack/liblapack.so.3.6.0

Random number generation:
 RNG:     Mersenne-Twister
 Normal:  Inversion
 Sample:  Rounding

attached base packages:
[1] grid stats graphics grDevices utils datasets methods base

other attached packages:
 [1] reshape2_1.4.3      RColorBrewer_1.1-2  poibin_1.3          NMOF_1.6-0
 [5] metasens_0.3-2      meta_4.9-5          gridExtra_2.3       ggpubr_0.2
 [9] magrittr_1.5        fitdistrplus_1.0-14 npsurv_0.4-0        lsei_1.2-0
[13] survival_2.44-1.1   MASS_7.3-51.4       extrafont_0.17      data.table_1.12.2
[17] cowplot_0.9.4       ggplot2_3.1.1

loaded via a namespace (and not attached):
 [1] Rcpp_1.0.1          pillar_1.4.0        compiler_3.6.0      plyr_1.8.4          tools_3.6.0
 [6] nlme_3.1-140        tibble_2.1.1        gtable_0.3.0        lattice_0.20-38     pkgconfig_2.0.2
[11] rlang_0.3.4         Matrix_1.2-17       rstudioapi_0.10     parallel_3.6.0      Rttf2pt1_1.3.7
[16] stringr_1.4.0       metafor_2.1-0       withr_2.1.2         dplyr_0.8.1         tidyselect_0.2.5
[21] glue_1.3.1          R6_2.4.0            purrr_0.3.2         extrafontdb_1.0     scales_1.0.0
[26] splines_3.6.0       assertthat_0.2.1    colorspace_1.4-1    stringi_1.4.3       lazyeval_0.2.2
[31] munsell_0.5.0       crayon_1.3.4
```

Bibliography

Amrhein, Valentin, Sander Greenland and Blake McShane (2019). 'Retire statistical significance'. In: p. 3 (cit. on p. 38).

Andrews, Isaiah and Maximilian Kasy (2017). 'Identification of and correction for publication bias'. In: *National Bureau of Economic Research* w23298, pp. 1–46 (cit. on pp. 58, 68).

Antonakis, John (2017). 'On doing better science: From thrill of discovery to policy implications'. In: *The Leadership Quarterly* 28.1, pp. 5–21 (cit. on p. 39).

Auspurg, Katrin and Thomas Hinz (2011). 'What Fuels Publication Bias?' In: *Jahrbücher für Nationalökonomie und Statistik / Journal of Economics and Statistics* 231.5, pp. 636–660 (cit. on p. 37).

Barlow, Roger J (1989). *Statistics: a guide to the use of statistical methods in the physical sciences*. Vol. 29. John Wiley & Sons (cit. on p. 64).

Begg, Colin B. and Jesse A. Berlin (1988). 'Publication Bias: A Problem in Interpreting Medical Data'. In: *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 151.3, p. 419 (cit. on p. 38).

Begg, Colin B. and Madhuchhanda Mazumdar (1994). 'Operating Characteristics of a Rank Correlation Test for Publication Bias'. In: *Biometrics* 50.4, p. 1088 (cit. on p. 47).

Berning, Carl C. and Bernd Weiß (2016). 'Publication bias in the German social sciences: an application of the caliper test to three top-tier German social science journals'. In: *Quality & Quantity* 50.2, pp. 901–917 (cit. on p. 38).

Casella, George and Roger L. Berger (2002). *Statistical inference*. 2nd ed. Australia ; Pacific Grove, CA: Thomson Learning. 660 pp. (cit. on p. 8).

Chalmers, Thomas C., Cynthia S. Frank and Dinah Reitman (1990). 'Minimizing the Three Stages of Publication Bias'. In: *JAMA* 263.10, pp. 1392–1395 (cit. on p. 37).

Chapman, J W et al. (2013). 'Innovative estimation of survival using log-normal survival modelling on ACCENT database'. In: *British Journal of Cancer* 108.4, pp. 784–790 (cit. on p. 24).

Cleary, Richard J and George Casella (1997). 'An Application of Gibbs Sampling to Estimation in Meta-Analysis: Accounting for Publication Bias'. In: *Journal of Educational and Behavioral Statistics* 22.2, pp. 141–154 (cit. on p. 68).

Cooper, Harris, Kristina DeNeve and Kelly Charlton (1997). 'Finding the missing science: The fate of studies submitted for review by a human subjects committee.' In: *Psychological Methods* 2.4, pp. 447–452 (cit. on pp. 37, 38).

Copas, John (1999). 'What Works?: Selectivity Models and Meta-Analysis'. In: *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 162.1, pp. 95–109 (cit. on p. 68).

Cowles, Michael and Caroline Davis (1982). 'On the Origins of the .05 Level of Statistical Significance'. In: p. 6 (cit. on p. 37).

Dickersin, K. (1990). 'The existence of publication bias and risk factors for its occurrence'. In: *JAMA: The Journal of the American Medical Association* 263.10, pp. 1385–1389 (cit. on pp. 37, 38).

Dickersin, Kay (2005). 'Publication Bias: Recognizing the Problem, Understanding Its Origins and Scope, and Preventing Harm'. In: *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*. Ed. by Hannah R. Rothstein, Alexander J. Sutton and Michael Borenstein. John Wiley & Sons Ltd, pp. 11–34 (cit. on p. 38).

Duval, Sue (2005). 'The Trim and Fill Method'. In: *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*. Ed. by Hannah R. Rothstein, Alexander J. Sutton and Michael Borenstein. John Wiley & Sons Ltd, pp. 127–144 (cit. on p. 61).

Duval, Sue and Richard Tweedie (2000a). 'A Nonparametric "Trim and Fill" Method of Accounting for Publication Bias in Meta-Analysis'. In: *Journal of the American Statistical Association* 95.449, pp. 89–98 (cit. on p. 60).

Duval, Sue and Richard Tweedie (2000b). 'Trim and Fill: A Simple Funnel-Plot-Based Method of Testing and Adjusting for Publication Bias in Meta-Analysis'. In: *Biometrics* 56.2, pp. 455–463 (cit. on p. 60).

Easterbrook, P.J et al. (1991). 'Publication bias in clinical research'. In: *The Lancet* 337.8746, pp. 867–872 (cit. on p. 38).

Egger, M. et al. (1997). 'Bias in meta-analysis detected by a simple, graphical test'. In: *BMJ* 315.7109, pp. 629–634 (cit. on p. 48).

Eitan, Orly et al. (2018). 'Is research in social psychology politically biased? Systematic empirical tests and a forecasting survey to address the controversy'. In: *Journal of Experimental Social Psychology* 79, pp. 188–199 (cit. on p. 37).

Fanelli, Daniele (2012). 'Negative results are disappearing from most disciplines and countries'. In: *Scientometrics* 90.3, pp. 891–904 (cit. on p. 38).
Fisher Box, Joan (1978). *R. A. Fisher, The Life of a Scientist*. John Wiley & Sons (cit. on p. 3).
Fisher, Ronald A (1925). *Statistical methods for research workers*. Vol. 6. Oliver and Boyd (cit. on p. 37).
Fisher, Ronald A. (1926). 'The arrangements of field experiments'. In: (cit. on p. 38).
Fisher, Ronald A (1938). 'Presidential address to the first Indian statistical congress'. In: *Sankhya* 4.14, pp. 14–17 (cit. on p. 69).
Flint, J. et al. (2015). 'Is there an excess of significant findings in published studies of psychotherapy for depression?' In: *Psychological Medicine* 45.2, pp. 439–446 (cit. on p. 38).
Franco, A., N. Malhotra and G. Simonovits (2014). 'Publication bias in the social sciences: Unlocking the file drawer'. In: *Science* 345.6203, pp. 1502–1505 (cit. on pp. 37, 38).
Gerber, Alan S., Donald P. Green and David Nickerson (2001). 'Testing for Publication Bias in Political Science'. In: *Political Analysis* 9.4, pp. 385–392 (cit. on p. 38).
Gerber, Alan S. and Neil Malhotra (2008). 'Publication Bias in Empirical Sociological Research: Do Arbitrary Significance Levels Distort Published Results?' In: *Sociological Methods & Research* 37.1, pp. 3–30 (cit. on pp. 38, 50).
Gerber, Alan and Neil Malhotra (2006). 'Can political science literatures be believed? A study of publication bias in the APSR and the AJPS'. In: *Annual Meeting of the Midwest Political Science Association*. CiteseerX (cit. on pp. 38, 50).
Hansen, Morris H. and William N. Hurwitz (1943). 'On the Theory of Sampling from Finite Populations'. In: *The Annals of Mathematical Statistics* 14.4, pp. 333–362 (cit. on p. 60).
Head, Megan L. et al. (2015). 'The Extent and Consequences of P-Hacking in Science'. In: *PLOS Biology* 13.3, e1002106 (cit. on p. 41).
Hedges, Larry V. (1984). 'Estimation of Effect Size under Nonrandom Sampling: The Effects of Censoring Studies Yielding Statistically Insignificant Mean Differences'. In: *Journal of Educational Statistics* 9.1, p. 61 (cit. on p. 38).
Ioannidis, J. P. and T. A Trikalinos (2007). 'An exploratory test for an excess of significant findings'. In: *Clinical Trials* 4.3, pp. 245–253 (cit. on pp. 38, 42).

Ioannidis, John P. A. (2019). 'The Importance of Predefined Rules and Prespecified Statistical Analyses: Do Not Abandon Significance'. In: *JAMA* (cit. on p. 39).

Kendall, M G (1938). 'A New Measure of Rank Correlation'. In: *Biometrika* 30.1, pp. 81–93 (cit. on pp. 47, 48).

Kerr, Norbert L. (1998). 'HARKing: Hypothesizing After the Results are Known'. In: *Personality and Social Psychology Review* 2.3, pp. 196–217 (cit. on p. 41).

Kühberger, Anton, Astrid Fritz and Thomas Scherndl (2014). 'Publication Bias in Psychology: A Diagnosis Based on the Correlation between Effect Size and Sample Size'. In: *PLoS ONE* 9.9. Ed. by Daniele Fanelli, e105825 (cit. on p. 38).

Kulinskaya, Elena, Stephan Morgenthaler and Robert G. Staudte (2008). *Meta analysis: a guide to calibrating and combining statistical evidence*. Wiley series in probability and statistics. Chichester: Wiley. 260 pp. (cit. on pp. 9–12, 16, 26, 27).

Lau, Joseph et al. (2006). 'The case of the misleading funnel plot'. In: *BMJ* 333.7568, pp. 597–600 (cit. on p. 44).

Light, Richard J. and David B. Pillemer (1984). *Summing up: The Science of Reviewing Research*. Harvard University Press (cit. on p. 44).

Mahoney, Michael J. (1977). 'Publication prejudices: An experimental study of confirmatory bias in the peer review system'. In: *Cognitive Therapy and Research* 1.2, pp. 161–175 (cit. on p. 38).

McShane, Blakeley B. et al. (2019). 'Abandon Statistical Significance'. In: *The American Statistician* 73 (sup1), pp. 235–245 (cit. on p. 38).

Murdoch, Duncan J, Yu-Ling Tsai and James Adcock (2008). '*P* -Values are Random Variables'. In: *The American Statistician* 62.3, pp. 242–245 (cit. on p. 8).

Neyman, J. and E. S. Pearson (1933). *On The Problem of the most Efficient Tests of Statistical Hypotheses* (cit. on p. 38).

Neyman, J., E. S. Pearson and G. U. Yule (1933). 'The testing of statistical hypotheses in relation to probabilities a priori'. In: *Mathematical Proceedings of the Cambridge Philosophical Society* 29.4, p. 492 (cit. on p. 38).

Piao, Jin et al. (2018). 'Copas-like selection model to correct publication bias in systematic review of diagnostic test studies'. In: *Statistical Methods in Medical Research*, p. 096228021879160 (cit. on p. 68).

Rosenthal, Robert (1979). 'The "File Drawer Problem" and Tolerance for Null Results'. In: *Psychological Bulletin* 86.3, pp. 638–641 (cit. on pp. 36, 39, 52).

Royston, P. (2001). 'The Lognormal Distribution as a Model for Survival Time in Cancer, With an Emphasis on Prognostic Factors'. In: *Statistica Neerlandica* 55.1, pp. 89–104 (cit. on p. 24).

Simmons, Joseph P., Leif D. Nelson and Uri Simonsohn (2011). 'False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant'. In: *Psychological Science* 22.11, pp. 1359–1366 (cit. on p. 41).

Simonsohn, Uri, Leif D Nelson and Joseph P Simmons (2014a). 'P-curve: a key to the file-drawer.' In: *Journal of Experimental Psychology: General* 143.2, p. 534 (cit. on pp. 52, 56).

Simonsohn, Uri, Leif D Nelson and Joseph P Simmons (2014b). 'P-curve and effect size: Correcting for publication bias using only significant results'. In: *Perspectives on Psychological Science* 9.6, pp. 666–681 (cit. on p. 64).

Sterling, Theodore D. (1959). 'Publication Decisions and their Possible Effects on Inferences Drawn from Tests of Significance—or Vice Versa'. In: *Journal of the American Statistical Association* 54.285, pp. 30–34 (cit. on p. 36).

Sterne, Jonathan A C, Betsy Jane Becker and Matthias Egger (2005). 'The Funnel Plot'. In: *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*. Ed. by Hannah R. Rothstein, Alexander J. Sutton and Michael Borenstein. John Wiley & Sons Ltd, pp. 75–98 (cit. on p. 44).

Sterne, Jonathan A C and Matthias Egger (2001). 'Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis'. In: *Journal of Clinical Epidemiology*, p. 10 (cit. on p. 44).

Sterne, Jonathan A C and Matthias Egger (2005). 'Regression Methods to Detect Publication and Other Bias in Meta-Analysis'. In: *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*. Ed. by Hannah R. Rothstein, Alexander J. Sutton and Michael Borenstein. John Wiley & Sons Ltd, pp. 99–110 (cit. on p. 45).

Student (1908). 'The Probable Error of a Mean'. In: *Biometrika* 6.1, pp. 1–25 (cit. on pp. 2, 26, 37).

Tang, Jin-Ling and Joseph LY Liu (2000). 'Misleading funnel plot for detection of bias in meta-analysis'. In: *Journal of Clinical Epidemiology* 53.5, pp. 477–484 (cit. on p. 44).

Terrin, Norma, Christopher H. Schmid and Joseph Lau (2005). 'In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias'. In: *Journal of Clinical Epidemiology* 58.9, pp. 894–901 (cit. on p. 44).

Weiß, Bernd and Michael Wagner (2011). 'The Identification and Prevention of Publication Bias in the Social Sciences and Economics'. In: *Jahrbücher*

für Nationalökonomie und Statistik / Journal of Economics and Statistics 231.5, pp. 661–684 (cit. on p. 38).

Welch, B. L. (1947). 'The Generalization of 'Student's' Problem when Several Different Population Variances are Involved'. In: *Biometrika* 34.1, p. 28 (cit. on p. 25).