

# **Improving Your Statistical Inferences**

Daniël Lakens

# Table of contents

<b>Overview</b>	<b>9</b>
<b>1 Using <i>p</i>-values to test a hypothesis</b>	<b>11</b>
1.1 Philosophical approaches to <i>p</i> -values . . . . .	12
1.2 Creating a null model . . . . .	14
1.3 Calculating a <i>p</i> -value . . . . .	16
1.4 Which <i>p</i> -values can you expect? . . . . .	17
1.5 Lindley's paradox . . . . .	21
1.6 Correctly reporting and interpreting <i>p</i> -values . . . . .	22
1.7 Preventing common misconceptions about <i>p</i> -values . . . . .	25
1.7.1 Misconception 1: A non-significant <i>p</i> -value means that the null hypothesis is true. . . . .	32
1.7.2 Misconception 2: A significant <i>p</i> -value means that the null hypothesis is false. . . . .	35
1.7.3 Misconception 3: A significant <i>p</i> -value means that a practically important effect has been discovered. . . . .	37
1.7.4 Misconception 4: If you have observed a significant finding, the probability that you have made a Type 1 error (a false positive) is 5%. . . . .	38
1.7.5 Misconception 5: One minus the <i>p</i> -value is the probability that the effect will replicate when repeated. . . . .	40
1.8 Test Yourself . . . . .	40
1.8.1 Questions about which <i>p</i> -values you can expect . . . . .	40
1.8.2 Questions about <i>p</i> -value misconceptions . . . . .	47
1.8.3 Open Questions . . . . .	53
<b>2 Error control</b>	<b>55</b>
2.1 Which outcome can you expect if you perform a study? . . . . .	55
2.2 Positive predictive value . . . . .	58
2.3 Type 1 error inflation . . . . .	61
2.4 Optional stopping . . . . .	62
2.5 Justifying Error Rates . . . . .	69
2.6 Why you don't need to adjust your alpha level for all tests you'll do in your lifetime. . . . .	73
2.7 Power Analysis . . . . .	75

2.8	Test Yourself . . . . .	77
2.8.1	Questions about the positive predictive value . . . . .	77
2.8.2	Questions about optional stopping . . . . .	79
2.8.3	Open Questions . . . . .	85
<b>3</b>	<b>Likelihoods</b>	<b>86</b>
3.1	Likelihood ratios . . . . .	93
3.2	Likelihood of mixed results in sets of studies . . . . .	96
3.3	Likelihoods for <i>t</i> -tests . . . . .	99
3.4	Test Yourself . . . . .	101
3.4.1	Questions about likelihoods . . . . .	101
3.4.2	Questions about mixed results . . . . .	103
3.4.3	Open Questions . . . . .	106
<b>4</b>	<b>Bayesian statistics</b>	<b>108</b>
4.1	Bayes factors . . . . .	109
4.2	Updating our belief . . . . .	112
4.3	Preventing common misconceptions about Bayes Factors . . . . .	117
4.3.1	Misunderstanding 1: Confusing Bayes Factors with Posterior Odds. . . . .	117
4.3.2	Misunderstanding 2: Failing to interpret Bayes Factors as relative evidence. . . . .	118
4.3.3	Misunderstanding 3: Not specifying the null and/or alternative model. . . . .	119
4.3.4	Misunderstanding 4: Claims based on Bayes Factors do not require error control. . . . .	119
4.3.5	Misunderstanding 5: Interpreting Bayes Factors as effect sizes. . . . .	121
4.4	Bayesian Estimation . . . . .	121
4.5	Test Yourself . . . . .	123
4.5.1	Open Questions . . . . .	129
<b>5</b>	<b>Asking Statistical Questions</b>	<b>131</b>
5.1	Description . . . . .	131
5.2	Prediction . . . . .	132
5.3	Explanation . . . . .	132
5.4	Loosening and Tightening . . . . .	133
5.5	Three Statistical Philosophies . . . . .	135
5.6	Falsification . . . . .	136
5.7	Severe Tests . . . . .	138
5.8	Risky Predictions . . . . .	139
5.9	Do You Really Want to Test a Hypothesis? . . . . .	142
5.10	Directional (One-Sided) versus Non-Directional (Two-Sided) Tests . . . . .	144
5.11	Systematic Noise, or the Crud Factor . . . . .	147
5.12	Dealing with Inconsistencies in Science . . . . .	149
5.13	Verisimilitude and Progress in Science . . . . .	154

<b>6 Effect Sizes</b>	<b>157</b>
6.1 Effect sizes . . . . .	158
6.2 The Facebook experiment . . . . .	159
6.3 The Hungry Judges study . . . . .	160
6.4 Standardised Mean Differences . . . . .	162
6.5 Interpreting effect sizes . . . . .	168
6.6 Correlations and Variance Explained . . . . .	170
6.7 Correcting for Bias . . . . .	173
6.8 Effect Sizes for Interactions . . . . .	174
6.9 Why Effect Sizes Selected for Significance are Inflated . . . . .	178
6.10 The Minimal Statistically Detectable Effect . . . . .	183
6.11 Test Yourself . . . . .	183
6.11.1 Open Questions . . . . .	189
<b>7 Confidence Intervals</b>	<b>190</b>
7.1 Population vs. Sample . . . . .	190
7.2 What is a Confidence Interval? . . . . .	191
7.3 Interpreting a single confidence interval . . . . .	194
7.4 The relation between confidence intervals and <i>p</i> -values . . . . .	195
7.5 The Standard Error and 95% Confidence Intervals . . . . .	197
7.6 Overlapping Confidence Intervals . . . . .	198
7.7 Prediction Intervals . . . . .	198
7.8 Capture Percentages . . . . .	201
7.9 Calculating Confidence Intervals around Standard Deviations . . . . .	202
7.10 Computing Confidence Intervals around Effect Sizes . . . . .	202
7.11 Test Yourself . . . . .	204
7.11.1 Open Questions . . . . .	217
<b>8 Sample Size Justification</b>	<b>218</b>
8.1 Six Approaches to Justify Sample Sizes . . . . .	219
8.2 Six Ways to Evaluate Which Effect Sizes are Interesting . . . . .	220
8.3 The Value of Information . . . . .	221
8.4 Measuring (Almost) the Entire Population . . . . .	222
8.5 Resource Constraints . . . . .	222
8.6 A-priori Power Analysis . . . . .	224
8.7 Planning for Precision . . . . .	231
8.8 Heuristics . . . . .	232
8.9 No Justification . . . . .	234
8.10 What is Your Inferential Goal? . . . . .	234
8.11 What is the Smallest Effect Size of Interest? . . . . .	235
8.12 The Minimal Statistically Detectable Effect . . . . .	235
8.13 What is the Expected Effect Size? . . . . .	237
8.14 Using an Estimate from a Meta-Analysis . . . . .	237

8.15 Using an Estimate from a Previous Study . . . . .	239
8.16 Using an Estimate from a Theoretical Model . . . . .	242
8.17 Compute the Width of the Confidence Interval around the Effect Size . . . . .	242
8.18 Plot a Sensitivity Power Analysis . . . . .	244
8.19 The Distribution of Effect Sizes in a Research Area . . . . .	248
8.20 Additional Considerations When Designing an Informative Study . . . . .	249
8.21 Compromise Power Analysis . . . . .	249
8.22 What to do if Your Editor Asks for Post-hoc Power? . . . . .	253
8.23 Sequential Analyses . . . . .	255
8.24 Increasing Power Without Increasing the Sample Size . . . . .	255
8.25 Know Your Measure . . . . .	259
8.26 Conventions as meta-heuristics . . . . .	260
8.27 Sample Size Justification in Qualitative Research . . . . .	261
8.28 Discussion . . . . .	262
8.29 Test Yourself . . . . .	262
8.29.1 Open Questions . . . . .	267
<b>9 Equivalence Testing and Interval Hypotheses</b>	<b>269</b>
9.1 Equivalence tests . . . . .	273
9.2 Reporting Equivalence Tests . . . . .	279
9.3 Minimum Effect Tests . . . . .	280
9.4 Power Analysis for Interval Hypothesis Tests . . . . .	283
9.5 The Bayesian ROPE procedure . . . . .	285
9.6 Which interval width should be used? . . . . .	287
9.7 Setting the Smallest Effect Size of Interest . . . . .	288
9.8 Specifying a SESOI based on theory . . . . .	289
9.9 Anchor based methods to set a SESOI . . . . .	290
9.10 Specifying a SESOI based on a cost-benefit analysis . . . . .	291
9.11 Specifying the SESOI using the small telescopes approach . . . . .	291
9.12 Setting the Smallest Effect Size of Interest to the Minimal Statistically Detectable Effect . . . . .	297
9.13 Test Yourself . . . . .	300
9.13.1 Questions about equivalence tests . . . . .	300
9.13.2 Questions about the small telescopes approach . . . . .	307
9.13.3 Questions about specifying the SESOI as the Minimal Statistically Detectable Effect . . . . .	309
9.13.4 Open Questions . . . . .	311
<b>10 Sequential Analysis</b>	<b>313</b>
10.1 Choosing alpha levels for sequential analyses. . . . .	315
10.2 The Pocock correction . . . . .	316
10.3 Comparing Spending Functions . . . . .	317
10.4 Alpha spending functions . . . . .	321

10.5 Updating boundaries during a study . . . . .	322
10.6 Sample Size for Sequential Designs . . . . .	325
10.7 Stopping for futility . . . . .	332
10.8 Reporting the results of a sequential analysis . . . . .	336
10.9 Test Yourself . . . . .	341
10.9.1 Open Questions . . . . .	346
<b>11 Meta-analysis</b>	<b>349</b>
11.1 Random Variation . . . . .	349
11.2 A single study meta-analysis . . . . .	361
11.3 Simulating meta-analyses of mean standardized differences . . . . .	364
11.4 Fixed Effect vs Random Effects . . . . .	369
11.5 Simulating meta-analyses for dichotomous outcomes . . . . .	369
11.6 Heterogeneity . . . . .	373
11.7 Exploring heterogeneity through subgroup analyses . . . . .	378
11.8 Strengths and weaknesses of meta-analysis . . . . .	379
11.9 Which results should you report to be included in a future meta-analysis? . . . . .	381
11.10 Improving the reproducibility of meta-analyses . . . . .	382
11.11 Test Yourself . . . . .	384
11.11.1 Open Questions . . . . .	389
<b>12 Bias detection</b>	<b>390</b>
12.1 Publication bias . . . . .	392
12.2 Bias detection in meta-analysis . . . . .	396
12.3 Trim and Fill . . . . .	405
12.4 PET-PEESE . . . . .	407
12.5 <i>P</i> -value meta-analysis . . . . .	409
12.6 Conclusion . . . . .	417
12.7 Test Yourself . . . . .	418
12.7.1 Open Questions . . . . .	424
<b>13 Preregistration and Transparency</b>	<b>426</b>
13.1 Preregistration of the Statistical Analysis Plan . . . . .	427
13.2 The value of preregistration . . . . .	430
13.3 How to preregister . . . . .	432
13.4 Journal Article Reporting Standards . . . . .	432
13.5 Deviating from a Preregistration . . . . .	436
13.6 What Does a Formalized Analytic Strategy Look Like? . . . . .	438
13.7 Are you ready to preregister a hypothesis test? . . . . .	440
13.8 Test Yourself . . . . .	441
13.8.1 Practical Aspects of an Online Preregistration . . . . .	446
13.8.2 Pre-registering on PsychArchives by ZPID . . . . .	446
13.8.3 Pre-registering on the Open Science Framework . . . . .	449

13.8.4 Pre-registering on AsPredicted . . . . .	451
<b>14 Computational Reproducibility</b>	<b>453</b>
14.1 Step 1: Setting up a GitHub repository . . . . .	454
14.2 Step 2: Cloning your GitHub repository into RStudio . . . . .	457
14.3 Step 3: Creating an R Markdown file . . . . .	465
14.4 Step 4: Reproducible Data Analysis in R Studio . . . . .	469
14.5 Step 5: Committing and Pushing to GitHub . . . . .	474
14.6 Step 6: Reproducible Data Analysis . . . . .	478
14.6.1 Extra: APA formatted manuscripts in papaja . . . . .	484
14.7 Step 7: Organizing Your Data and Code . . . . .	486
14.8 Step 8: Archiving Your Data and Code . . . . .	486
14.8.1 EXTRA: Sharing Reproducible Code on Code Ocean . . . . .	493
14.9 Some points for improvement in computational reproducibility . . . . .	499
14.10 Conclusion . . . . .	500
<b>15 Research Integrity</b>	<b>501</b>
15.1 Questionable Research Practices . . . . .	503
15.2 Fabrication, Falsification, and Plagiarism . . . . .	504
15.3 Informed consent and data privacy . . . . .	507
15.4 Conflicts of Interest . . . . .	508
15.5 Research ethics . . . . .	508
15.6 Test Yourself . . . . .	509
15.7 Grade Yourself . . . . .	511
<b>16 Confirmation Bias and Organized Skepticism</b>	<b>514</b>
16.1 Confirmation bias in science . . . . .	519
16.2 Organized Skepticism . . . . .	522
16.2.1 Error control . . . . .	522
16.2.2 Preregistration . . . . .	523
16.2.3 Independent Replication Studies . . . . .	523
16.2.4 Peer Review . . . . .	524
16.2.5 Double-Checking Errors . . . . .	526
16.2.6 The Devil's Advocate . . . . .	527
16.2.7 Adversarial Collaborations . . . . .	527
16.2.8 Red Team Science . . . . .	528
16.2.9 Blinding . . . . .	529
16.2.10 Separating Theorists from Experimentalists . . . . .	529
16.2.11 Method of multiple working hypotheses . . . . .	530
16.3 Conclusion . . . . .	531
<b>17 Replication Studies</b>	<b>533</b>
17.1 Why replication studies are important . . . . .	537

17.2 Direct versus conceptual replications . . . . .	539
17.3 Analyzing Replication Studies. . . . .	540
17.4 Replication studies or lower alpha levels? . . . . .	549
17.5 When replication studies yield conflicting results . . . . .	553
17.6 Why are replication studies so rare? . . . . .	557
<b>References</b>	<b>558</b>
<b>Change Log</b>	<b>591</b>

# Overview

“No book can ever be finished. While working on it we learn just enough to find it immature the moment we turn away from it.” *Karl Popper, The Open Society and its Enemies*

This open educational resource integrates information from my [blog](#), my MOOCs [Improving Your Statistical Inferences](#) and [Improving Your Statistical Questions](#), and my [scientific work](#). The goal is to make the information more accessible, and easier to update in the future.

I have re-used and adapted (parts of) my own open access articles, without adding quotation marks. Immense gratitude to my collaborators Casper Albers, Farid Anvari, Aaron Caldwell, Harlan Cambell, Nicholas Coles, Lisa DeBruine, Marie Delacre, Zoltan Dienes, Noah van Dongen, Alexander Etz, Ellen Evers, Jaroslav Gottfriend, Seth Green, Christopher Harms, Arianne Herrera-Bennett, Joe Hilgard, Peder Isager, Maximilian Maier, Neil McLatchie, Brian Nosek, Friedrich Pahlke, Pepijn Obels, Amy Orben, Anne Scheel, Janneke Staaks, Leo Tiokhin, Mehmet Tunç, Duygu Uygun Tunç, and Gernot Wassmer, who have contributed substantially to the ideas in this open educational resource. I would also like to thank Zeki Akyol, Emrah Er, Max Ditroilo, Lewis Halsey, Kyle Hamilton, David Lane, Elen LeFoll, Jeremiah Lewis, Mike Smith, and Leong Utek who gave comments on GitHub or Twitter to improve this textbook. The first version of this textbook was created during a sabbatical at Padova University, with thanks to the Advanced Data Analysis for Psychological Science students, and Gianmarco Altoè and Ughetta Moscardino for their hospitality.

Thanks to Dale Barr and Lisa DeBruine for the [webexercises](#) package that is used to create the interactive questions at the end of each chapter. Thanks to Nick Brown for his editing service.

If you find any mistakes, or have suggestions for improvement, you can [submit an issue on the GitHub page](#) of this open educational resource. You can also download a pdf or epub version (click the download button in the menu on the top left). This work is shared under a [CC-BY-NC-SA License](#). You can cite this resource as:

Lakens, D. (2022). Improving Your Statistical Inferences. Retrieved from [https://lakens.github.io/statistical\\_inference/](https://lakens.github.io/statistical_inference/)  
<https://doi.org/10.5281/zenodo.6409077>

You can check the [Change Log](#) at the end of this book to track updates over time, or to find the version number if you prefer to cite a specific version of this regularly updated textbook.

This work is dedicated to Kyra, the love of my life.



Daniël Lakens

Eindhoven University of Technology

# 1 Using *p*-values to test a hypothesis

Scientists can attempt to answer a wide range of questions by collecting data. One question that interests scientists is whether measurements that have been collected under different conditions differ, or not. The answer to such a question is an *ordinal claim*, where a researcher states the average of the measurements is larger, or smaller, or the same, when comparing conditions. For example, a researcher might be interested in the hypothesis that students learn better if they do tests, during which they need to retrieve information they have learned (condition A), compared to not doing tests, but spending all of their time studying (condition B). After collecting data, and observing that the mean grade is higher for students who spent part of their time doing tests, the researcher can make the ordinal claim that student performance was *better* in condition A compared to condition B. Ordinal claims can only be used to state that there is a difference between conditions. They do not quantify the **size of the effect**.

To make ordinal claims, researchers typically rely on a methodological procedure known as a **hypothesis test**. One part of a hypothesis test consists of computing a ***p*-value** and examining whether there is a statistically **significant** difference. ‘Significant’ means that something is worthy of attention. A hypothesis test is used to distinguish a signal (that is worth paying attention to) from random noise in empirical data. It is worth distinguishing **statistical significance**, which is only used to claim whether an observed effect is a signal or noise, from **practical significance**, which depends on whether the size of the effect is large enough to have any worthwhile consequences in real life. Researchers use a methodological procedure to decide whether to make an ordinal claim or not as a safeguard against **confirmation bias**. In an internal report for Guinness brewery on the use of statistical tests in an applied setting, William Gosset (or ‘Student’, who developed the *t*-test) already wrote (1904):

On the other hand, it is generally agreed that to leave the rejection of experiments entirely to the discretion of the experimenter is dangerous, as he is likely to be biassed. Hence it has been proposed to adopt a criterion depending on the probability of such a wide error occurring in the given number of observations.

Depending on their desires, scientists might be tempted to interpret data as support for their hypothesis, even when it is not. As Benjamini (2016) notes, a *p*-value “offers a first line of defense against being fooled by randomness, separating signal from noise”. There are indications that banning the use of *p*-values increases the ability of researchers to present erroneous claims. Based on qualitative analyses of scientific articles published after the null hypothesis significance ban in the journal Basic and Applied Social Psychology, Fricker et al. (2019)

conclude: “When researchers only employ descriptive statistics we found that they are likely to overinterpret and/or overstate their results compared to a researcher who uses hypothesis testing with the  $p < 0.05$  threshold”. A hypothesis test, when used correctly, controls the amount of time researchers will fool themselves when they make ordinal claims.

## 1.1 Philosophical approaches to $p$ -values

Before we look at how  $p$ -values are computed, it is important to examine how they are supposed to help us make ordinal claims when testing hypotheses. The definition of a  $p$ -value is the probability of observing the sample data, or more extreme data, assuming the null hypothesis is true. But this definition does not tell us much about how we should interpret a  $p$ -value.

The interpretation of a  $p$ -value depends on the statistical philosophy one subscribes to. [Ronald Fisher](#) published ‘Statistical Methods for Research Workers’ in 1925 which popularized the use of  $p$ -values. In a Fisherian framework a  $p$ -value is interpreted as a descriptive continuous measure of compatibility between the observed data and the null hypothesis (or the null model). The null hypothesis is a model of the data that is expected if there is no effect. For example, if we assume scores from two groups of students are the same, and an intervention in one of the two groups has no effect on the scores, the null model is that the difference between the scores is distributed around zero with a certain standard deviation. In a statistical test a researcher would examine if this null hypothesis can be rejected based on the observed data, or not.

According to Fisher (1956),  $p$ -values “do not generally lead to any probability statement about the real world, but “a rational and well-defined measure of reluctance to the acceptance of the hypotheses they test” (p. 44). The lower the  $p$ -value, the greater the reluctance to accept the null hypothesis. Spanos (1999) refers to the Fisherian approach as **misspecification testing**, as used when testing model assumptions. In a misspecification test (such as measures of goodness of fit) the null hypothesis is tested against any unspecified alternative hypothesis. Without a specified alternative hypothesis, the  $p$  value can only be used to describe the probability of the observed or more extreme data under the null model.

Fisher tried to formalize his philosophy in an approach called ‘fiducial inference’, but this has not received the same widespread adoption of other approaches, such as decision theory, likelihoods, and Bayesian inference. Indeed, Zabell (1992) writes “The fiducial argument stands as Fisher’s one great failure”, although others have expressed the hope that it might be developed into a useful approach in the future (Schweder & Hjort, 2016). A Fisherian  $p$ -value describes the incompatibility of the data with a single hypothesis, and is known as *significance testing*. The main reason a *significance test* is limited is because researchers only specify a null hypothesis ( $H_0$ ), but not the alternative hypothesis ( $H_1$ ). According to Spanos (1999), after the model assumptions have been checked, when “testing theoretical restrictions within a statistically adequate model, however, the Neyman–Pearson approach becomes the procedure of choice” (p. xxiii).

Neyman and Pearson (1933) built on insights about  $p$ -values by Gosset and Fisher, and developed an approach called *statistical hypothesis testing*. The main difference with the significance testing approach developed by Fisher is that in a statistical hypothesis test both a null hypothesis and an alternative hypothesis are specified. In a Neyman-Pearson framework, the goal of statistical tests is to guide the behavior of researchers with respect to these two hypotheses. Based on the results of a statistical test, and without ever knowing whether the hypothesis is true or not, researchers choose to tentatively act as if the null hypothesis or the alternative hypothesis is true. In psychology, researchers often use an imperfect hybrid of the Fisherian and Neyman-Pearson frameworks, but the Neyman-Pearson approach is, according to Dienes (2008), “the logic underlying all the statistics you see in the professional journals of psychology”.

When a Neyman-Pearson hypothesis test is performed, the observed  $p$ -value is only used to check if it is smaller than the chosen alpha level, but it does not matter how much smaller it is. For example, if an alpha level of 0.01 is used, both a  $p = 0.006$  and a  $p = 0.000001$  will lead researchers to decide to act as if the state of the world is best described by the alternative hypothesis. This differs from a Fisherian approach to  $p$ -values, where the lower the  $p$ -value, the greater the psychological reluctance of a researcher to accept the null hypothesis they are testing. A Neyman-Pearson hypothesis test does not see the goal of an inference as quantifying a continuous measure of compatibility or evidence. Instead, as Neyman (1957) writes:

The content of the concept of inductive behavior is the recognition that the purpose of every piece of serious research is to provide grounds for the selection of one of several contemplated courses of action.

Intuitively, one might feel that decisions about how to act should not be based on the results of a single statistical test, and this point is often raised as a criticism of the Neyman-Pearson approach to statistical inferences. However, such criticisms rarely use the same definition of an ‘act’ as Neyman used. It is true that, for example, the decision to implement a new government policy should not be based on a single study result. However, Neyman considered making a scientific claim an ‘act’ as well, and wrote (1957, p. 10) that the concluding phase of a study involves: “an act of will or a decision to take a particular action, perhaps to assume a particular attitude towards the various sets of hypotheses”.

Cox (1958) writes:

[I]t might be argued that in making an inference we are ‘deciding’ to make a statement of a certain type about the populations and that therefore, provided that the word decision is not interpreted too narrowly, the study of statistical decisions embraces that of inference. The point here is that one of the main general problems of statistical inference consists in deciding what types of statement can usefully be made and exactly what they mean.

Thus, in a Neyman-Pearson approach,  $p$ -values form the basis of decisions about which claims to make. In science, such claims underlie most novel experiments in the form of **auxiliary**

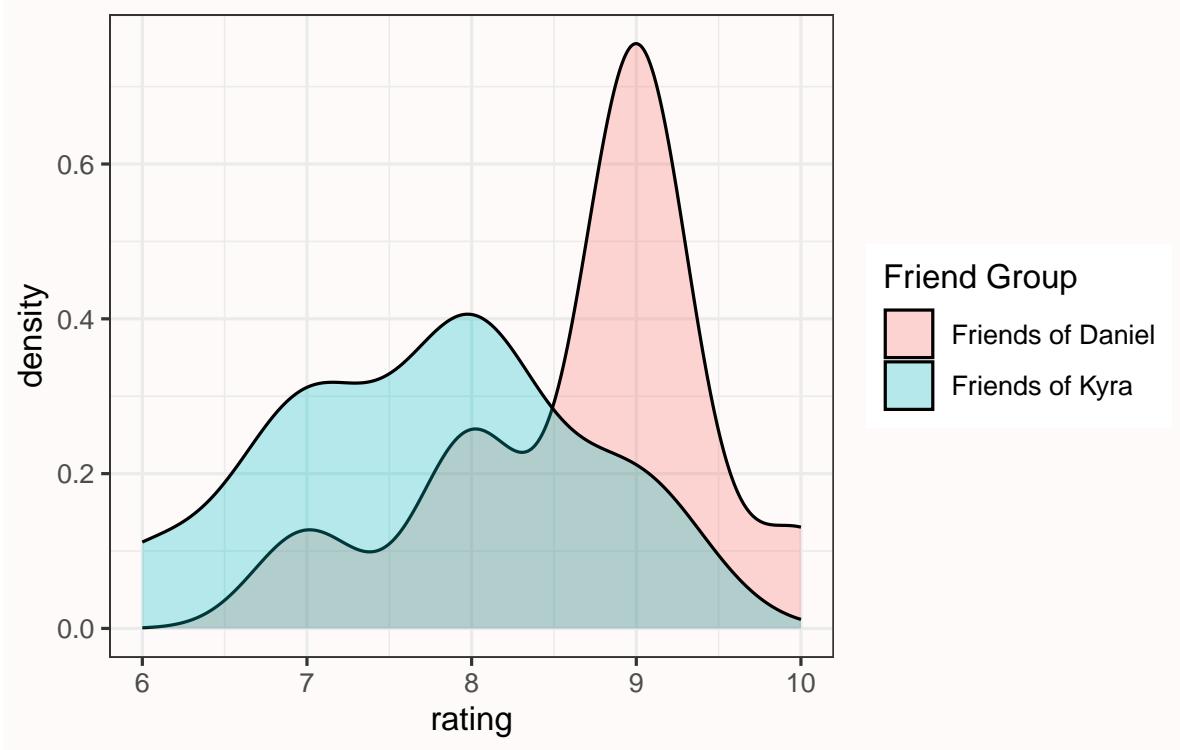
**hypotheses**, or the assumptions about underlying hypotheses that are assumed to be accurate in order for a test to work as planned (Hempel, 1966). For example, if it is important that participants can see color in a planned experiment, we assume it is true that the [Ishihara test](#) successfully identifies which participants are colorblind.

## 1.2 Creating a null model

Assume I ask two groups of 10 people how much they liked the extended director's cut of the Lord of the Rings (LOTR) trilogy. This means our **total sample size** ( $N$ ) is 20, and the sample size in each group ( $n$ ) is 10. The first group consists of my friends, and the second group consists of friends of my wife. Our friends rate the trilogy on a score from 1 to 10. We can calculate the average rating by my friends, which is 8.7, and the average rating by my wife's friends, which is 7.7. We can compare the scores in both groups by looking at the raw data, and by plotting the data.

Table 1.1: Ratings for the Lord of the Rings extended trilogy by two groups of friends.

	Friends of Daniel	Friends of Kyra
friend_1	9	9
friend_2	7	6
friend_3	8	7
friend_4	9	8
friend_5	8	7
friend_6	9	9
friend_7	9	8
friend_8	10	8
friend_9	9	8
friend_10	9	7



We can see the groups overlap but the mean ratings differ by 1 whole point. The question we are now faced with is the following: Is the difference between the two groups just random variation, or can we claim that my friends like the extended director's cut of the LOTR trilogy more than my wife's friends?

In a **null hypothesis significance test** we try to answer this question by calculating the probability of the observed difference (in this case, a mean difference of 1) or a more extreme difference, under the assumption that there is no real difference between how much my friends and my wife's friends like the extended director's cut of LOTR, and we are just looking at random noise. This probability is called the *p*-value. If this probability is low enough, we decide to claim there is a difference. If this probability is not low enough, we refrain from making a claim about a difference.

The null hypothesis assumes that if we would ask an infinite number of my friends and an infinite number of my wife's friends how much they like LOTR, the difference between these huge groups is exactly 0. However, in any sample drawn from the population, random variation is very likely to lead to a difference somewhat larger or smaller than 0. We can create a **null model** that quantifies the expected variation in the observed data, just due to random noise, to tell us what constitutes a reasonable expectation about how much the differences between groups can vary if there is no difference in the population.

It is practical to create a null model in terms of a **standardized** distribution, as this makes it easier to calculate the probability that specific values will occur, regardless of the scale that is used to collect the measurements. One version of a null model for differences is the *t*-distribution, which can be used to describe which differences should be expected when drawing samples from a population. Such a null model is built on **assumptions**. In the case of the *t*-distribution, the assumption is that scores are normally distributed. In reality, the assumptions upon which statistical methods are built are never met perfectly, which is why statisticians examine the impact of violations of assumptions on methodological procedures (e.g., the procedure to reject a hypothesis based on a certain test value). Statistical tests are still useful in practice when the impact of violations on statistical inferences is small enough.

We can quantify the distribution of *t*-values that is expected when there is no difference in the population by a *probability density function*. Below is a plot of the probability density function for a *t*-distribution with 18 **degrees of freedom** (df), which corresponds to our example where we collect data from 20 friends (df = N - 2 for two independent groups). For a continuous distribution, where probabilities are defined for an infinite number of points, the probability of observing any single point (e.g.,  $t = 2.5$ ) is always zero. Probabilities are measured over intervals. For this reason, when a *p*-value is computed, it is not defined as ‘the probability of observing the data’, but as ‘the probability of observing the data, *or more extreme data*’. This creates an interval (a tail of a distribution) for which a probability can be calculated.

### 1.3 Calculating a *p*-value

A *t*-value for a two-sample *t*-test can be computed from the means in the two samples, the standard deviations in the two samples, and the sample sizes in each group. By then computing the probability of observing a *t*-value as extreme or more extreme as the one observed, we get a *p*-value. For the comparison of the movie ratings for the two groups of friends above, performing a two-sided Student’s *t*-test yields a *t*-value of 2.5175 and a *p*-value of 0.02151.

```
t.test(df_long$rating ~ df_long$`Friend Group`, var.equal = TRUE)
```

Two Sample *t*-test

```
data: df_long$rating by df_long$`Friend Group`
t = 2.5175, df = 18, p-value = 0.02151
alternative hypothesis: true difference in means between group Friends of Daniel and group F
95 percent confidence interval:
 0.1654875 1.8345125
sample estimates:
mean in group Friends of Daniel   mean in group Friends of Kyra
                           8.7                           7.7
```

We can plot the  $t$ -distribution (for  $\text{df} = 18$ ) and highlight the two tail areas that start at the  $t$ -values of 2.5175 and -2.5175.

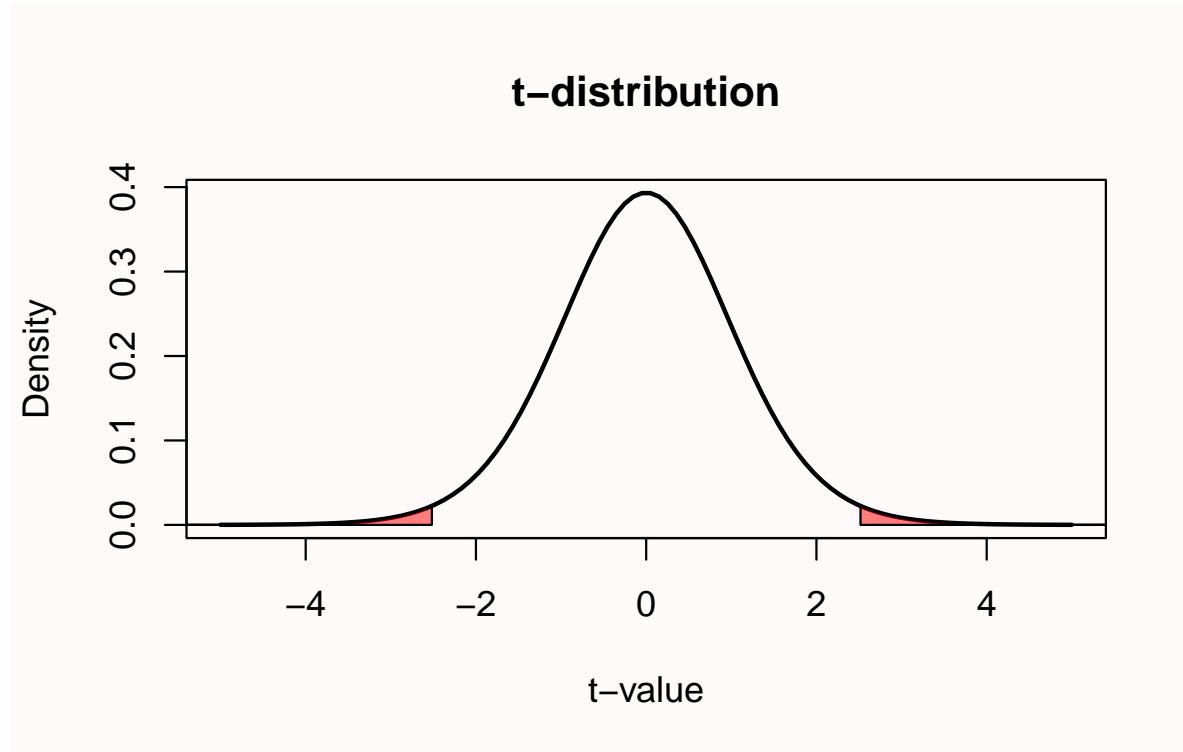


Figure 1.1: A  $t$ -distribution with 18 degrees of freedom.

## 1.4 Which $p$ -values can you expect?

In a very educational video about the ‘[Dance of the  \$p\$ -values](#)’, Geoff Cumming explains that  $p$ -values vary from experiment to experiment. However, this is not a reason to ‘not trust  $p$ ’ as he mentions in the video. Instead, it is important to clearly understand  **$p$ -value distributions** to prevent misconceptions. Because  $p$ -values are part of frequentist statistics, we need to examine what we can expect *in the long run*. Because we never do the same experiment hundreds of times, and we do only a very limited number of studies in our lifetime, the best way to learn about what we should expect in the long run is through computer simulations.

Take a moment to try to answer the following two questions. Which  $p$ -values can you expect to observe if there is a true effect, and you repeat the same study 100000 times? And which  $p$ -values can you expect if there is no true effect, and you repeat the same study one-hundred thousand times? If you don’t know the answer, don’t worry - you will learn it now. But if you don’t know the answer, it is worth reflecting on why you don’t know the answer about such

an essential aspect of  $p$ -values. If you are like me, you were simply never taught this. But as we will see, it is essential to a solid understanding of how to interpret  $p$ -values.

Which  $p$ -values you can expect is completely determined by the statistical power of the study, or in other words the probability that you will observe a significant effect if there is a true effect. The statistical power ranges from 0 to 1. We can illustrate this by simulating independent  $t$ -tests. The idea is that we simulate IQ scores for a group of people. We know the standard deviation of IQ scores is 15. For now, we will set the mean IQ score in one simulated group to 100, and in the other simulated group to 105. We are testing if the people in one group have an IQ that differs from the other group (and we know the correct answer is ‘yes’, because we made it so in the simulation).

```
p <- numeric(100000) # store all simulated *p*-values

for (i in 1:100000) { # for each simulated experiment
  x <- rnorm(n = 71, mean = 100, sd = 15) # Simulate data
  y <- rnorm(n = 71, mean = 105, sd = 15) # Simulate data
  p[i] <- t.test(x, y)$p.value # store the *p*-value
}

(sum(p < 0.05) / 100000) # compute power

hist(p, breaks = 20) # plot a histogram
```

In the simulation, we generate  $n = 71$  normally distributed IQ scores with means of  $M$  (100 and 105 by default) and a standard deviation of 15. We then perform an independent  $t$ -test, store the  $p$ -value, and generate a plot of the  $p$ -value distribution.

On the x-axis we see  $p$ -values from 0 to 1 in 20 bars, and on the y-axis we see how frequently these  $p$ -values were observed. There is a horizontal red dotted line that indicates an alpha of 5% (located at a frequency of  $100000 * 0.05 = 5000$ ) – but you can ignore this line for now. In the title of the graph, the statistical power that is achieved in the simulated studies is given (assuming an alpha of 0.05): The studies have 50% power.

The simulation result illustrates the **probability density function** of  $p$ -values for a  $t$ -test. A probability density function provides the probability that a random variable has a specific value (such as Figure 1.1) of the  $t$ -distribution). Because the  $p$ -value is a random variable, we can use its probability density function to plot the  $p$ -value distribution (Hung et al., 1997; Ulrich & Miller, 2018), as in Figure 1.3). In [this online Shiny app](#) you can vary the sample size, effect size, and alpha level to examine the effect on the  $p$ -value distribution. Increasing the sample size or the effect size will increase the steepness of the  $p$ -value distribution, which means that the probability to observe small  $p$ -values increases. The  $p$ -value distribution is a function of the statistical power of the test.

## p-value Distribution with 50% power

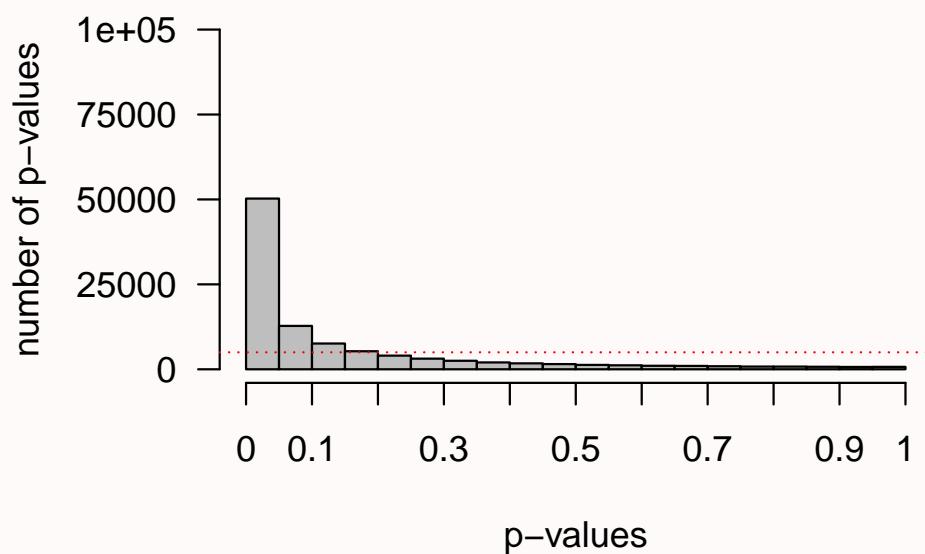


Figure 1.2: Distribution of  $p$ -values when power = 50%.

## p-value distribution

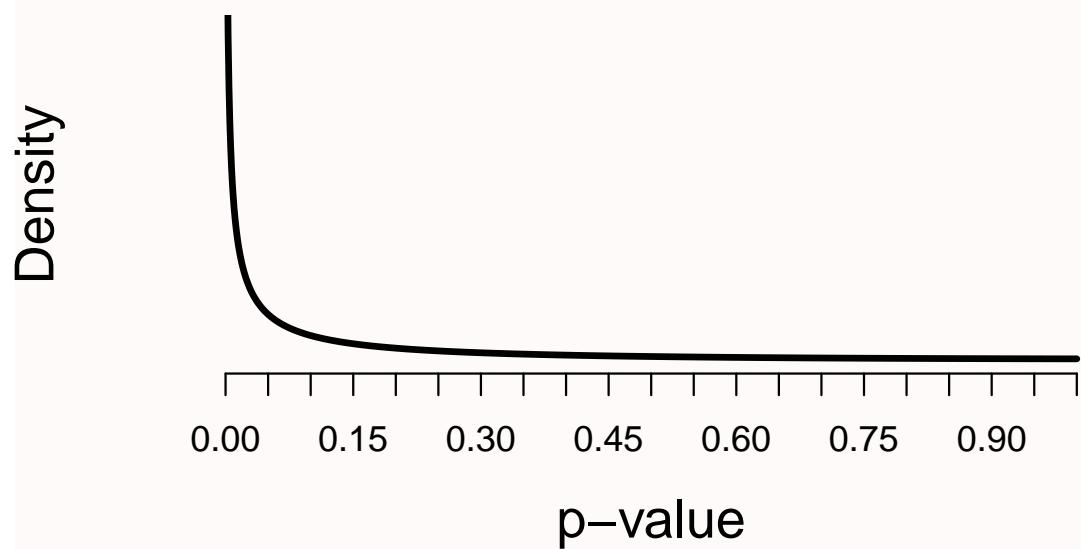


Figure 1.3: Probability density function for  $p$ -values from a two-sided  $t$ -test.

When there is no true effect, and test assumptions are met,  $p$ -values for a  $t$ -test are **uniformly distributed**. This means that every  $p$ -value is equally likely to be observed when the null hypothesis is true. In other words, when there is no true effect, a  $p$ -value of 0.08 is just as likely as a  $p$ -value of 0.98. I remember thinking this was very counterintuitive when I first learned about uniform  $p$ -value distributions (after completing my PhD). But it makes sense that  $p$ -values are uniformly distributed when we remember that, when  $H_0$  is true, alpha % of the  $p$ -values should fall below the alpha level, whichever alpha level we set. Hence, if we set alpha to 0.01, 1% of the observed  $p$ -values should fall below 0.01, and if we set alpha to 0.12, 12% of the observed  $p$ -values should fall below 0.12. This can only happen if  $p$ -values are uniformly distributed when the null hypothesis is true. Note that the uniform  $p$ -value distribution only emerges if the distribution of the test statistic is continuous (e.g., such as for the  $t$ -test), and not if the distribution of the test statistic is discrete (e.g., in the case of the chi-squared test see Wang et al. (2019)).

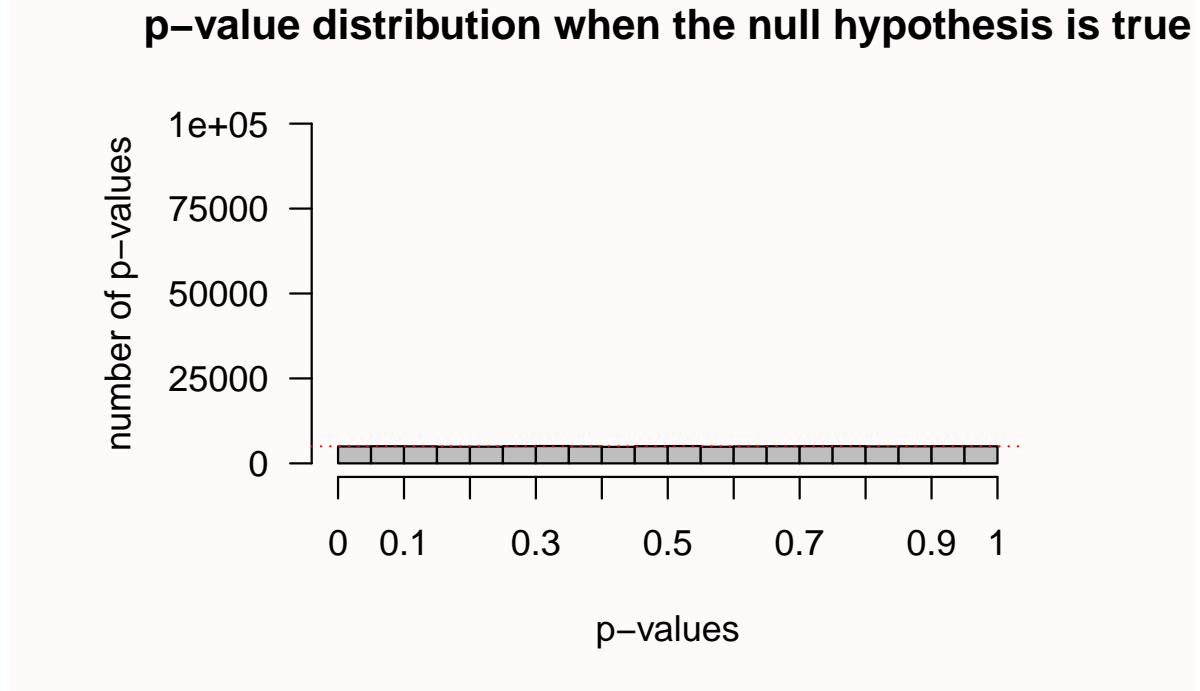


Figure 1.4: Distribution of  $p$ -values when the null hypothesis is true.

## 1.5 Lindley's paradox

As statistical power increases, some  $p$ -values below 0.05 (e.g.,  $p = 0.04$ ) can be more likely when there is *no* effect than when there *is* an effect. This is known as Lindley's paradox (Lindley,

1957), or sometimes the Jeffreys-Lindley paradox (Spanos, 2013). Because the distribution of  $p$ -values is a function of the statistical power (Cumming, 2008), the higher the power, the more right-skewed the  $p$ -value distribution becomes (i.e., the more likely it becomes that small  $p$ -values will be observed). When there is no true effect,  $p$ -values are uniformly distributed, and 1% of observed  $p$ -values fall between 0.04 and 0.05. When the statistical power is extremely high, not only will most  $p$ -values fall below 0.05; in fact, most  $p$ -values will fall well below 0.01. In Figure 1.5 we see that, with high power, very small  $p$ -values (e.g., 0.001) are more likely to be observed when there *is* an effect than when there is *no* effect (e.g., the dotted black curve representing 99% power falls above the grey horizontal line representing the uniform distribution when the null is true for a  $p$ -value of 0.01).

Yet perhaps surprisingly, observing a  $p$ -value of 0.04 is more likely when the null hypothesis ( $H_0$ ) is true than when the alternative hypothesis ( $H_1$ ) is true and we have very high power, as illustrated by the fact that in Figure 1.5 the density of the  $p$ -value distribution at  $p = 0.04$  is higher if the null hypothesis is true than when the alternative hypothesis is true and a test has 99% power. Lindley's paradox shows that a  $p$ -value of for example 0.04 can be statistically significant, but at the same time provides evidence for the null hypothesis. From a Neyman-Pearson approach we have made a claim that has a maximum error rate of 5%, but from a likelihood or Bayesian approach, we should conclude that our data provides evidence in favor of the null hypothesis, relative to the alternative hypothesis. Lindley's paradox illustrates when different statistical philosophies would reach different conclusions, and why a  $p$ -value cannot directly be interpreted as a measure of evidence, without taking the power of the test into account. Although a strict application of the Neyman-Pearson approach does not require it, researchers might desire to prevent situations where a frequentist rejects the null hypothesis based on  $p < 0.05$ , when the evidence in the test favors the null hypothesis over the alternative hypothesis. This can be achieved by lowering the alpha level as a function of the sample size (Good, 1992; Leamer, 1978; Maier & Lakens, 2022), as explained in the chapter on [error control](#).

## 1.6 Correctly reporting and interpreting $p$ -values

Although from a strict Neyman-Pearson perspective it is sufficient to report that  $p < \alpha$  or that  $p > \alpha$ , researchers should always report exact  $p$ -values. This facilitates the re-use of results for secondary analyses (Appelbaum et al., 2018), and allows other researchers to compare the  $p$ -value to an alpha level they would have preferred to use (Lehmann & Romano, 2005). Because claims are made using a methodological procedure with known maximum error rates, a  $p$ -value never allows you to state anything with certainty. Even if we set the alpha level to 0.000001, any single claim can be an error, Fisher (1935) reminds us, ‘for the “one chance in a million” will undoubtedly occur, with no less and no more than its appropriate frequency, however surprised we may be that it should occur to *us*’. This is the reason that **replication studies** are important in science. Any single finding could be a fluke, but this probability quickly becomes very small if several replication studies observe the same finding. This uncertainty

### **p-value distribution for $d = 0$ , 50% power, and 99% power**

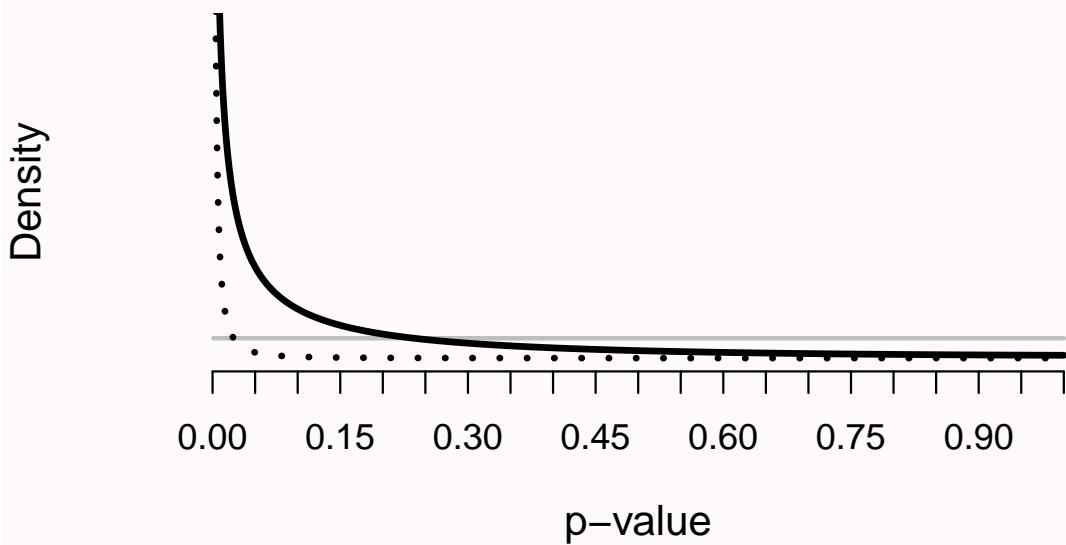


Figure 1.5:  $P$ -value distribution for 0 (grey horizontal line, 50 percent power (black solid curve), and 99 percent power (black dotted curve, where  $p$ -values just below 0.05 are more likely when  $H_0$  is true than when  $H_1$  is true).

is sometimes not reflected in academic writing, where researchers can be seen using words as 'prove', 'show', or 'it is known'. A slightly longer but more accurate statement after a statistically significant null hypothesis test would read:

We claim there is a non-zero effect, while acknowledging that if scientists make claims using this methodological procedure, they will be misled, in the long run, at most alpha % of the time, which we deem acceptable. We will, for the foreseeable future, and until new data or information emerges that proves us wrong, assume this claim is correct.

After a statistically non-significant null hypothesis test one could write:

We can not claim there is a non-zero effect, while acknowledging that if scientists refrain from making claims using this methodological procedure, they will be misled, in the long run, at most beta % of the time, which we deem acceptable.

Note that after a non-significant null hypothesis test we can not claim the absence of an effect, which is why it is advisable to always also perform an [equivalence test](#). If you have performed a [minimum effect test](#), replace 'non-zero' in the sentences above with minimum effect that you test against.

Remember that in a Neyman-Pearson framework researchers make claims, but do not necessarily *believe* in the truth of these claims. For example, the OPERA collaboration reported in 2011 that they had observed data that seemed to suggest neutrinos traveled faster than the speed of light. This claim was made with a 0.2-in-a-million Type 1 error rate, *assuming the error was purely due to random noise*. However, none of the researchers actually believed this claim was true, because it is theoretically impossible for neutrinos to move faster than the speed of light. Indeed, it was later confirmed that equipment failures were the cause of the anomalous data: a fiber optic cable had been attached improperly, and a clock oscillator was ticking too fast. Nevertheless, the claim was made with the explicit invitation to the scientific community to provide new data or information that would prove this claim wrong.

When researchers "accept" or "reject" a hypothesis in a Neyman-Pearson approach to statistical inferences, they do not communicate any belief or conclusion about the substantive hypothesis. Instead, they utter a Popperian **basic statement** based on a prespecified decision rule that the observed data reflect a certain state of the world. Basic statements describe an observation that has been made (e.g., "I have observed a black swan") or an event that has occurred (e.g., "students performed better at the exam when they spread out their revision over multiple days, than when they revised everything in one day").

The claim is about the data we have observed, but not about the theory we used to make our predictions, which requires a theoretical inference. Data never 'proves' a theory is true or false. A basic statement can **corroborate** a prediction derived from a theory, or not. If many predictions deduced from a theory are corroborated, we can become increasingly convinced

that the theory is close to the truth. This ‘truth-likeness’ of theories is called **verisimilitude** (Niiniluoto, 1998; Popper, 2002). A shorter statement when a hypothesis test is presented would therefore read ‘ $p = .xx$ , which corroborates our prediction, at an alpha level of y%’, or ‘ $p = .xx$ , which does not corroborate our prediction, at a statistical power of y% for our effect size of interest’. Often, the alpha level or the statistical power is only mentioned in the experimental design section of an article, but repeating them in the results section might remind readers of the error rates associated with your claims.

Even when we have made correct claims, the underlying theory can be false. Popper (2002) reminds us that “The empirical basis of objective science has thus nothing ‘absolute’ about it”. He argues that science is not built on a solid bedrock, but on piles driven in a swamp, and notes that “We simply stop when we are satisfied that the piles are firm enough to carry the structure, at least for the time being.” As Hacking (1965) writes: “Rejection is not refutation. Plenty of rejections must be only tentative.” So when we reject the null model, we do so tentatively, aware of the fact we might have done so in error, without necessarily believing the null model is false, and without believing the theory we have used to make predictions is true. For Neyman (1957) inferential behavior is an “act of will to behave in the future (perhaps until new experiments are performed) in a particular manner, conforming with the outcome of the experiment.” All knowledge in science is provisional.

Some statisticians recommend interpreting  $p$ -values as measures of *evidence*. For example, Bland (2015) teaches that  $p$ -values can be interpreted as a “rough and ready” guide for the strength of evidence, and that  $p > 0.1$  indicates ‘little or no evidence’,  $.05 < p < 0.1$  indicates ‘weak evidence’,  $0.01 < p < 0.05$  indicates ‘evidence’,  $p < 0.001$  is ‘very strong evidence’. This is incorrect (Johansson, 2011; Daniël Lakens, 2022b), as is clear from the previous discussions of Lindley’s paradox and uniform  $p$ -value distributions. Researchers who claim  $p$  values are measures of evidence typically do not *define* the concept of evidence. In this textbook I follow the mathematical theory of evidence as developed by Shafer (1976, p. 144), who writes “An adequate summary of the impact of the evidence on a particular proposition  $A$  must include at least two items of information: a report on how well  $A$  is supported and a report on how well its negation  $\bar{A}$  is supported.” According to Shafer, evidence is quantified through support functions, and when assessing statistical evidence, support is quantified by the likelihood function. If you want to quantify *evidence*, see the chapters on [likelihoods](#) or [Bayesian statistics](#).

## 1.7 Preventing common misconceptions about $p$ -values

A  $p$ -value is the probability of the observed data, or more extreme data, under the assumption that the null hypothesis is true. To understand what this means, it might be especially useful to know what this doesn’t mean. First, we need to know what ‘the assumption that the null hypothesis is true’ looks like, and which data we should expect if the null hypothesis is true. Although the null hypothesis can be any value, in this assignment we will assume the null hypothesis is specified as a mean difference of 0. For example, we might be interested

in calculating the difference between a control condition and an experimental condition on a dependent variable.

It is useful to distinguish the null hypothesis (the prediction that the mean difference in the population is exactly 0) and the null model (a model of the data we should expect when we collect data when the null hypothesis is true). The null hypothesis is a point at 0, but the null model is a distribution. It is visualized in textbooks or power analysis software using pictures as you can see in Figure 1.1, with  $t$ -values on the horizontal axis, and a critical  $t$ -value somewhere between 1.96 – 2.15 (depending on the sample size). This is done because the statistical test when comparing two groups is based on the  $t$ -distribution, and the  $p$ -value is statistically significant when the  $t$ -value is larger than a critical  $t$ -value.

I personally find things become a lot clearer if you plot the null model as mean differences instead of  $t$ -values as in Figure 1.6. This plot shows a null model for the mean differences that we can expect when comparing two groups of 50 observations where the true difference between the two groups is 0, and the standard deviation in each group is 1. Because the standard deviation is 1, you can also interpret the mean differences as a Cohen's  $d$  effect size. So this is also the distribution you can expect for a Cohen's  $d$  of 0, when collecting 50 observations per group in an independent  $t$ -test.

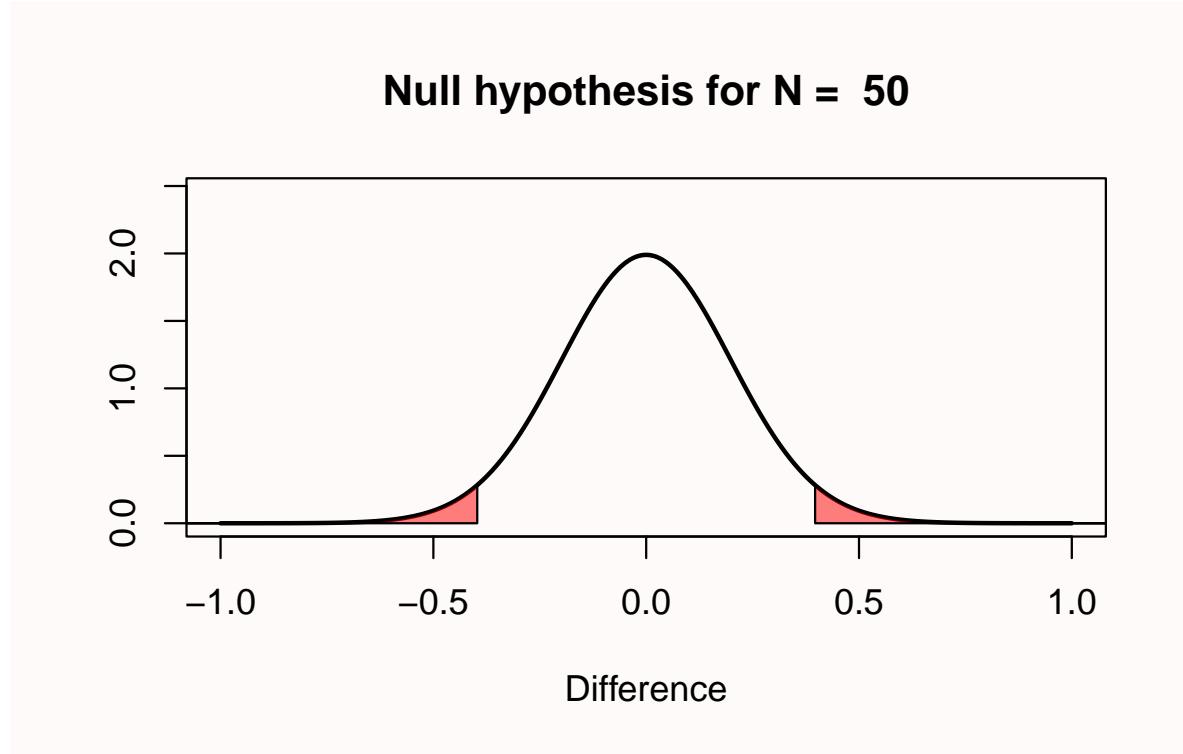


Figure 1.6: Distribution of observed Cohen's  $d$  effect sizes when collecting 50 observations per group in an independent  $t$ -test.

The first thing to notice is that we expect that the mean of the null model is 0. Looking at the x-axis, we see the plotted distribution is centered on 0. But even if the mean difference in the population is 0 that does not imply that every pair of samples that we draw from the population will result in a mean difference of exactly zero. There is variation around the population value, as a function of the standard deviation and the sample size.

The y-axis of the graph represents the density, which provides an indication of the relative likelihood of measuring a particular value from a continuous distribution. We can see that the most likely mean difference is the true population value of zero, and that differences with a magnitude greater than zero become increasingly less likely. The graph has two areas that are colored red. These areas represent the most extreme 2.5% of values in the left tail of the distribution, and the most extreme 2.5% of values in the right tail of the distribution. Together, they make up the 5% most extreme mean differences that we would expect to observe, given our number of observations, when the true mean difference is exactly 0. When a mean difference in the red area is observed, the corresponding statistical test will return a statistically significant result at a 5% alpha level, hence with a  $p$ -value of  $> 0.05$ . In other words, no more than 5% of the observed mean differences will be far enough away from 0 to be considered ‘surprising’ or statistically significant. Because the null hypothesis is true, observing a ‘surprising’ mean difference in either of the two red areas is a Type 1 error.

Let’s assume that the null model in Figure 1.6 above is true, and that we observe a mean difference of 0.5 between the two groups. This observed difference falls in the red area in the right tail of the distribution. This means that the observed mean difference is relatively surprising, under the assumption that the true mean difference is 0. If the true mean difference is 0, the probability density functions shows that we should not expect a mean difference of 0.5 very often. If we calculate a  $p$ -value for this observation, it would be lower than 5%. The probability of observing a mean difference that is at least as far away from 0 as 0.5 (either to the left of the mean, or to the right, when we do a two-tailed test) is less than 5%.

One reason why I prefer to plot the null model in raw scores instead of  $t$ -values is that you can see how the null model changes when the sample size increases. When we collect 5000 instead of 50 observations, we see that the null model is still centered on 0 – but in our null model we now expect most values will fall very closely around 0 (see Figure 1.7).

The distribution is much narrower because the distribution of mean differences is based on the standard error of the difference between means. The standard error is calculated based on the standard deviation and the sample size, as follows:

$$SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

This formula shows that the standard deviations of each group ( $\sigma$ ) are squared and divided by the sample size of that group, then added together, after which the square root is taken. The

### Null hypothesis for N = 5000

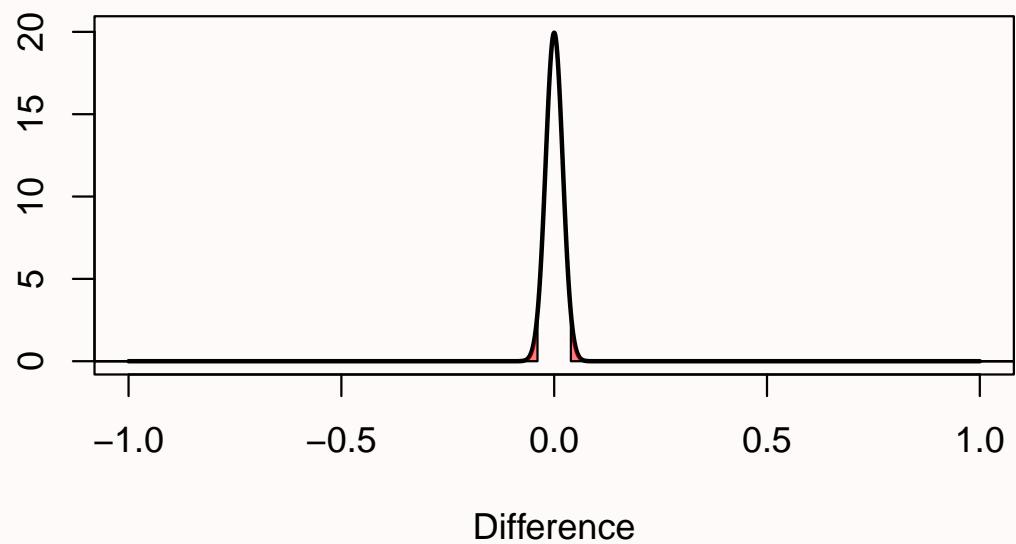


Figure 1.7: Distribution of observed Cohen's  $d$  effect sizes when collecting 5000 observations per group in an independent  $t$ -test when  $d = 0$ .

larger the sample size, the bigger the number we divide by, and thus the smaller the standard error of the difference between means. In our  $n = 50$  example this is:

$$\sqrt{\frac{1^2}{50} + \frac{1^2}{50}}$$

The standard error of the differences between means is thus 0.2 for  $n = 50$  in each group. For  $n = 5000$ , it is 0.02. Assuming a normal distribution, 95% of the observations fall between  $-1.96 * SE$  and  $1.96 * SE$ . So, with 50 observations per group, the mean differences should fall between  $-1.96 * 0.2 = -0.392$ , and  $+1.96 * 0.2 = 0.392$ , and we can see the red areas start from approximately -0.392 to 0.392 for  $n = 50$  (see Figure 1.6). With 5000 observations per group, the mean differences should fall between  $-1.96 * 0.02$ , and  $+1.96 * 0.02$ ; in other words, between -0.0392 to 0.0392 (see Figure 1.7). Due to the larger sample size of  $n = 5000$  observations per group, we can expect to observe mean differences closer to 0 compared to when we only had 50 observations. If we collected  $n = 5000$  observations, and we were again to observe a mean difference of 0.5, it should be clear that this difference is even more surprising than it was when we collected 50 observations.

We are now almost ready to address common misconceptions about  $p$ -values, but before we can do this, we need to introduce a model of the data when the null is not true. If we are not sampling data from a model where the true mean difference is 0, what does our alternative model look like? Some software (such as G\*power Faul et al. (2007), see Figure 1.8) will visualize both the null model (red curve) and the alternative model (blue curve) in their output.

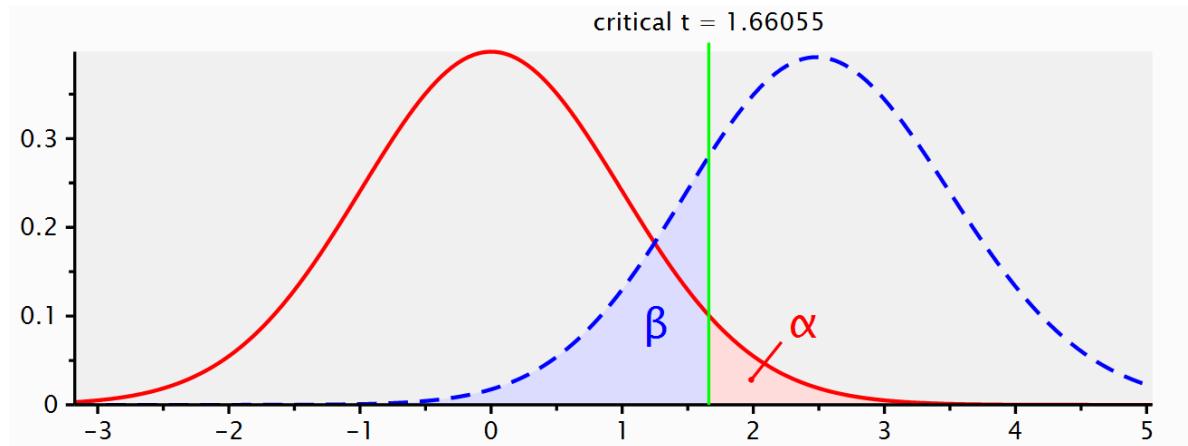


Figure 1.8: Screenshot from G\*Power software visualizing the distributions of  $t$ -values of the null model (in red) and the alternative model (in blue) and the critical  $t$ -value (1.66055), i.e. the threshold distinguishing significant from non-significant results.

When we do a study, we rarely know in advance what the true mean difference is (if we already knew, why would we do the study?). But let's assume there is an all-knowing entity. Following Paul Meehl, we will call this all-knowing entity 'Omniscient Jones'. Before we collect our sample of 50 observations, Omniscient Jones already knows that the true mean difference in the population is 0.5. Again, we expect some variation around 0.5 in this alternative model. The figure below shows the expected data pattern when the null hypothesis is true (now indicated by a grey line) and it shows an alternative model, assuming a true mean difference of 0.5 exists in the population (indicated by a black line).

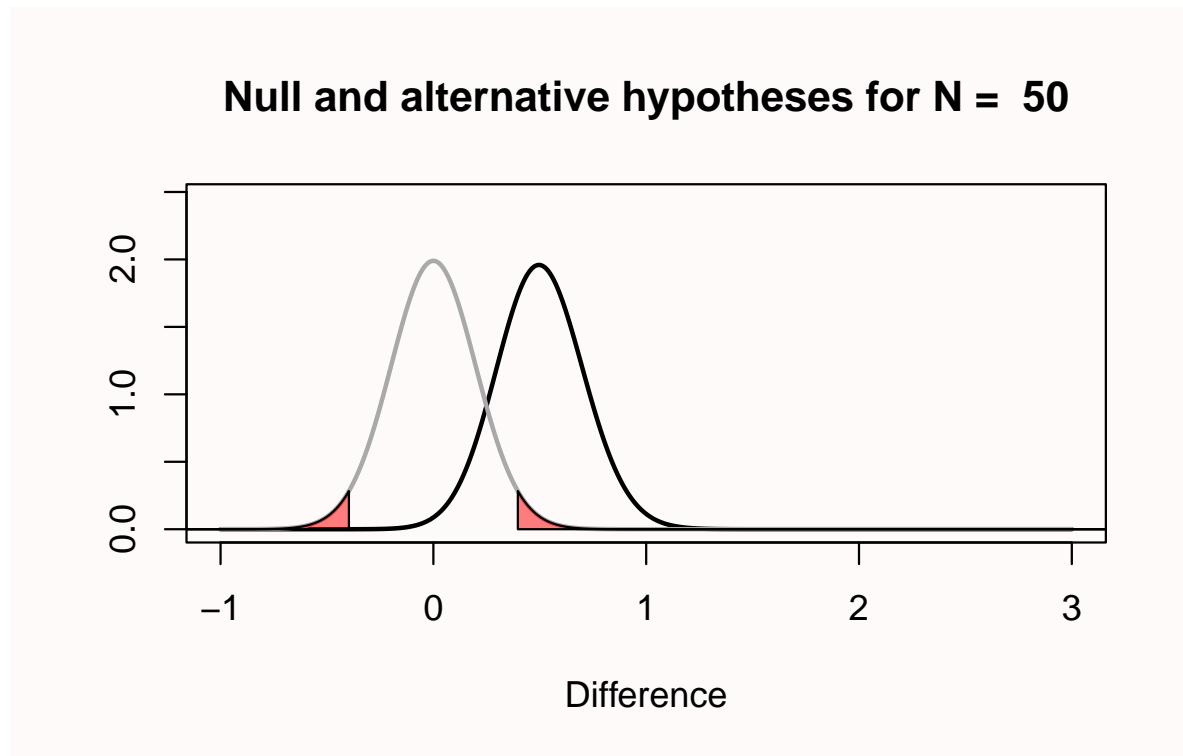


Figure 1.9: Distribution of observed Cohen's  $d$  effect sizes when collecting 50 observations per group in an independent  $t$ -test when  $d = 0$  (null model, left) or  $d = 0.5$  (alternative model, right).

But Omniscient Jones could have revealed that the true difference was much larger. Let's assume we do another study but now, before we collect our 50 observations, Omniscient Jones tells us that the true mean difference is 1.5. The null model does not change, but the alternative model now moves over to the right.

You can play around with the alternative and null models in [this online app](#). The app allows you to specify the sample size in each group of an independent  $t$ -test (from 2 to infinity), the mean difference (from 0 to 2), and the alpha level. In the plot, the red areas visualize Type 1 errors. The blue area visualizes the Type 2 error rate (which we will discuss below). The

app also tells you the critical value: There is a vertical line (with  $n = 50$  this line falls at a mean difference of 0.4) and a sentence that says: “Effects larger than 0.4 will be statistically significant”. Note that the same is true for effects smaller than -0.4, even though there is no second label there, but the app shows the situation for a two-sided independent  $t$ -test.

The plots in the app also include a blue area, which is to the left of the vertical line that indicates the critical mean difference and is part of the alternative model. This is the Type 2 error rate (or 1 minus the power of the study). If a study has 80% power, 80% of the mean differences we will observe should fall on the right of the critical value indicated by the line. If the alternative model is true, but we observe an effect smaller than the critical value, the observed  $p$ -value will be larger than 0.05, even when there is a true effect. You can check in the app that the larger the sample size, the further to the right the entire alternative distribution falls, and thus the higher the power. You can also see that the larger the sample size, the narrower the distribution, and the less of the distribution will fall below the critical value (as long as the true population mean is larger than the critical value). Finally, the larger the alpha level, the further to the left the critical mean difference moves, and the smaller the area of the alternative distribution that falls below the critical value.

The app also plots three graphs that illustrate the power curves as a function of different alpha levels, sample sizes, or true mean differences. Play around with the app by changing the values. Get a feel for how each variable impacts the null and alternative models, whether the mean difference is statistically significant, and the Type 1 and Type 2 error rates.

So far, several aspects of null models should have become clear. First of all, the population value in a traditional null hypothesis is a value of 0, but in any sample you draw, the observed difference falls in a distribution centered around 0, and will thus most often be slightly larger or smaller than 0. Second, the width of this distribution depends on the sample size and the standard deviation. The larger the sample size in the study, the narrower the distribution will be around 0. Finally, when a mean difference is observed that falls in the tails of the null model, this can be considered surprising. The further away from the null value, the more surprising this result is. But when the null model is true, these surprising values will happen with a probability specified by the alpha level (and are called Type 1 errors). Remember that a Type 1 error occurs when a researcher concludes that there is a difference in the population, even though the true mean difference in the population is zero.

We are now finally ready to address some common misconceptions about  $p$ -values. Let’s go through a list of common misconceptions that have been reported in the scientific literature. Some of these examples might sound like semantics. At first glance, it is easy to think that the statement communicates the right idea, even if the written version is not formally correct. However, when a statement is not formally correct, it is wrong. And exactly because people so often misunderstand  $p$ -values, it is worth insisting on being formally correct about how they should be interpreted.

### 1.7.1 Misconception 1: A non-significant $p$ -value means that the null hypothesis is true.

A common version of this misconception is reading a sentence such as ‘because  $p > 0.05$  we can conclude that there is no effect’. Another version of such a sentence is ‘there was no difference, ( $p > 0.05$ )’.

Before we look at this misconception in some detail, I want to remind you of one fact that is easy to remember, and will enable you to recognize many misconceptions about  $p$ -values:  $p$ -values are a statement about the probability of data, not a statement about the probability of a hypothesis or the probability of a theory. Whenever you see  $p$ -values interpreted as a probability of a theory or a hypothesis, you know something is not right. Examples of statements about a hypothesis are ‘The null hypothesis is true’, or ‘The alternative hypothesis is true’, because both these statements say that the probability that the null or alternative model is true is 100%. A subtler version is a statement such as ‘the observed difference is not due to chance’. The observed difference is only ‘due to chance’ (instead of due to the presence of a real difference) when the null hypothesis is true, and as before, this statement implies it is 100% probable that the null hypothesis is true.

When you conclude that ‘there is no effect’ or that ‘there is no difference’ you are similarly claiming that it is 100% probable that the null hypothesis is true. But since  $p$ -values are statements about the probability of data, you should refrain from making statements about the probability of a hypothesis solely based on a  $p$ -value. That’s ok.  $p$ -values were designed to help you identify surprising results from a noisy data generation process (aka the real world). They were not designed to quantify the probability that a hypothesis is true.

Let’s take a concrete example that will illustrate why a non-significant result does not mean that the null hypothesis is true. In Figure 1.10 below, Omniscient Jones tells us that the true mean difference is 0.5. We can see this because the alternative distribution which visualizes the probability of the mean differences we expect when the alternative hypothesis is true is centered on 0.5. We have observed a mean difference of 0.35. This value is not extreme enough to be statistically different from 0. We can see this because the value does not fall within the red area of the null model (and hence, the  $p$ -value is not smaller than our alpha level).

Nevertheless, we see that observing a mean difference of 0.35 is not only quite likely given that the true mean difference is 0.5, but observing a mean difference of 0.35 is much more likely under the alternative model than under the null model. You can see this by comparing the height of the density curve at a difference of 0.35 for the null model, which is approximately 0.5, and the height of the density curve for the alternative model, which is approximately 1.5. See the chapter on [likelihoods](#) for further details.

All the  $p$ -value tells us is that a mean difference of 0.35 is not extremely surprising if we assume that the null hypothesis is true. There can be many reasons for this. In the real world, where we have no Omniscient Jones to tell us about the true mean difference, it is possible that there is a true effect, as illustrated in the figure above.

## Null and alternative hypotheses for $N = 50$

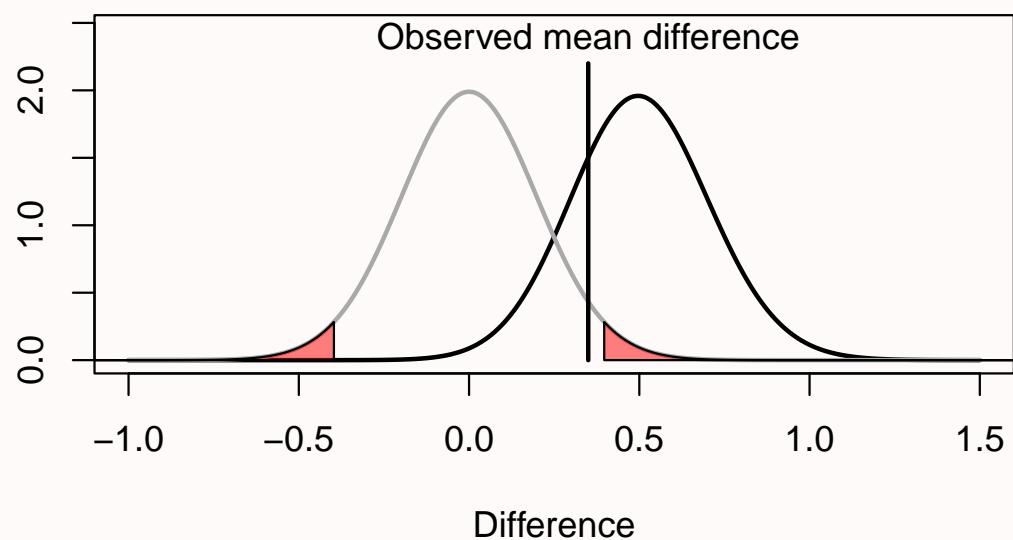


Figure 1.10: Distribution of observed Cohen's  $d$  effect sizes when collecting 50 observations per group in an independent  $t$ -test for  $d = 0$  and  $d = 0.5$  when observing  $d = 0.35$ .

So what should we say instead? The solution is subtle, but important. Let's revisit the two examples of incorrect statements we made earlier. First, 'because  $p > 0.05$  we can conclude that there is no effect' is incorrect, because there might very well be an effect (and remember  $p$ -values are statements about data, not about the probability that there is an effect or is no effect). Fisher's interpretation of a  $p$ -value was that we can conclude a rare event has happened, or that the null hypothesis is false (the exact quote is: "Either an exceptionally rare chance has occurred, or the theory of random distribution is not true"). This might sound like it is a statement about the probability of a theory, but it is really just stating the two possible scenarios under which low  $p$ -values occur (when you have made a Type 1 error, or when the alternative hypothesis is true). Both a true positive and a false positive remain possible, and we do not quantify the probability of either possibilities (e.g., we are not saying that it is 95% probable that the null hypothesis is false). From a Neyman-Pearson perspective, a  $p > .05$  means that we cannot act as if the null hypothesis can be rejected, without maintaining our desired error rate of 5%.

If you are interested in concluding that an effect is absent, null hypothesis testing is not the tool to use. A null hypothesis test answers the question 'can I reject the null hypothesis with a desired error rate?'. If you cannot do this, and  $p > 0.05$ , no conclusion can be drawn based only on the  $p$ -value. It might be useful to think of the answer to the question whether an effect is absent after observing  $p > 0.05$  as (mu), used as a non-dualistic answer, neither yes nor no, or 'unasking the question'. It is simply not possible to answer the question whether a meaningful effect is absent based on  $p > 0.05$ . Luckily, statistical approaches have been developed to ask questions about the absence of an effect such as [equivalence testing](#), Bayes factors, and Bayesian estimation (see Harms & Lakens (2018), for an overview).

The second incorrect statement was 'there was no difference'. This statement is somewhat easier to correct. You can instead write 'there was no statistically significant difference'. Granted, this is a bit tautological, because you are basically saying that the  $p$ -value was larger than the alpha level in two different ways, but at least this statement is formally correct. The difference between 'there was no difference' and 'there was no statistically significant difference' might sound like semantics, but in the first case you are formally saying 'the difference was 0' while in the second you are saying 'there was no difference large enough to yield a  $p < .05$ '. Although I have never seen anyone do this, a more informative message might be : 'Given our sample size of 50 per group, and our alpha level of 0.05, only observed differences more extreme than 0.4 could be statistically significant. Our observed mean difference was 0.35, hence we could not reject the null hypothesis'. If this feels like a very unsatisfactory conclusion, remember that a null hypothesis test was not designed to draw interesting conclusions about the absence of effects – you will need to learn about equivalence tests to get a more satisfactory answers about null effects.

### 1.7.2 Misconception 2: A significant $p$ -value means that the null hypothesis is false.

This is the opposite misconception from the one we discussed previously. Examples of incorrect statements based on this misconception are ‘ $p < .05$ , therefore there is an effect’, or ‘there is a difference between the two groups,  $p < .05$ ’. As before, both these statements imply it is 100% probable that the null model is false, and an alternative model is true.

As a simple example of why such extreme statements are incorrect, imagine we generate a series of numbers in R using the following command:

```
rnorm(n = 50, mean = 0, sd = 1)
```

```
[1] -0.93096996  0.87602457 -1.17535048  0.61642383 -0.62674866  0.35695708  
[7]  0.19884117 -1.84412366 -0.09389757  1.06559589 -0.48747976  1.11997271  
[13] -0.53145256 -0.73820361 -1.12222532  1.10362936 -0.72225071 -0.72554184  
[19]  0.13208951 -0.19758461  0.80255719 -0.57882865 -0.36473721  0.73429015  
[25] -0.52460098  0.28248052  1.07488620 -0.54680591  0.55069865  0.99859103  
[31]  1.24006579  1.41327371  0.38151730 -0.22844152  1.44957118 -1.20506213  
[37] -0.24074123  0.45020793 -0.51632645  1.54219406 -0.44706277  1.10194419  
[43] -2.37844853 -1.72325956 -0.25689718  0.68220098  1.14163862  0.62945655  
[49]  1.61260600  1.31530129
```

This command generates 50 random observations from a distribution with a mean of 0 and a standard deviation of 1 (in the long run – the mean and standard deviation will vary in each sample that is generated). Imagine we run this command once, and we observe a mean of 0.5. The figure below visualizes this scenario. We can perform a one-sample  $t$ -test against 0, and this test tells us, with a  $p < .05$ , that the data we have observed is surprisingly different from 0, assuming the random number generator in R functions as it should and generates data with a true mean of 0.

The significant  $p$ -value does not allow us to conclude that the null hypothesis (“the random number generator works”) is false. It is true that the mean of the 50 samples we generated was surprisingly extreme. But a low  $p$ -value simply tells us that an observation is surprising. We should observe such surprising observations with a low probability when the null hypothesis is true – when the null is true, they still happen. Therefore, a significant result does not mean an alternative hypothesis is true – the result can also be a Type 1 error, and in the example above, Omniscient Jones knows that this is the case.

Let’s revisit the incorrect statement ‘ $p < .05$ , therefore there is an effect’. A correct interpretation of a significant  $p$ -value requires us to acknowledge the possibility that our significant result might be a Type 1 error. Remember that Fisher would conclude that “Either an exceptionally rare chance has occurred, or the theory of random distribution is not true”. A

## Null hypothesis for N = 50

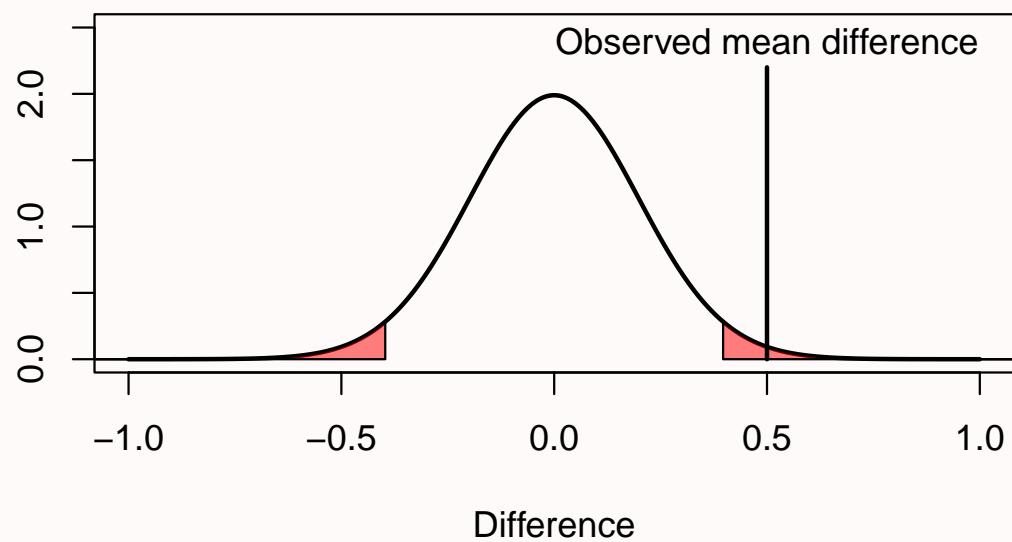


Figure 1.11: Distribution of observed Cohen's  $d$  effect sizes when collecting 50 observations per group in an independent  $t$ -test when  $d = 0$  and observing  $d = 0.5$ .

correct interpretation in terms of Neyman-Pearson statistics would be: ‘we can act as if the null hypothesis is false, and we would not be wrong more than 5% of the time in the long run’. Note the specific use of the word ‘act’, which does not imply anything about whether this specific hypothesis is true or false, but merely states that if we act as if the null hypothesis is false any time we observe  $p < \alpha$ , we will not make an error more than  $\alpha$  percent of the time.

Both these formally correct statements are a bit long. In scientific articles, we often read a shorter statement such as: ‘we can reject the null hypothesis’, or ‘we can accept the alternative hypothesis’. These statements might be made with the assumption that readers will themselves add ‘with a 5% probability of being wrong, in the long run’. But it might be useful to add ‘with a 5% long run error rate’ at least the first time you make such a statement in your article to remind readers.

In the example above we have a very strong subjective prior probability that the random number generator in R works. Alternative statistical procedures to incorporate such prior beliefs are [Bayesian statistics](#) or [false positive report probabilities](#). In frequentist statistics, the idea is that you need to replicate your study several times. You will observe a Type 1 error every now and then, but you are unlikely to observe a Type 1 error three times in a row. Alternatively, you can lower the alpha level in a single study to reduce the probability of a Type 1 error rate.

### **1.7.3 Misconception 3: A significant $p$ -value means that a practically important effect has been discovered.**

A common concern when interpreting  $p$ -values is that ‘significant’ in normal language implies ‘important’, and thus a ‘significant’ effect is interpreted as an ‘important’ effect. However, the question whether an effect is important is completely orthogonal to the question whether it is different from zero, or even how large the effect is. Not all effects have practical impact. The smaller the effect, the less likely such effects will be noticed by individuals, but such effects might still have a large impact on a societal level. Therefore, the general take home message is that statistical significance does not answer the question whether an effect matters in practice, or is ‘practically important’. To answer the question whether an effect matters, you need to present a cost-benefit analysis.

This issue of practical significance most often comes up in studies with a very large sample size. As we have seen before, with an increasing sample size, the width of the density distribution around the null value becomes more and more narrow, and the values that are considered surprising fall closer and closer to zero.

If we plot the null model for a very large sample size (e.g.,  $n = 10000$  per group) we see that even very small mean differences (those larger than 0.03) will be considered ‘surprising’. This still means that if there really is no difference in the population, you will observe differences larger than 0.04 less than 5% of the time, in the long run, and 95% of the observed differences

will be smaller than a mean difference of 0.04. But it becomes more difficult to argue for the practical significance of such effects. Imagine that a specific intervention is successful in changing people's spending behavior, and when implementing some intervention all 18 million people in the Netherlands will save 12 cents per year. It is difficult to argue how this effect will make any individual happier. However, if this money is combined, it will yield over 2 million, which could be used to treat diseases in developing countries, where it would have a real impact. The cost of the intervention might be considered too high if the goal is to make individuals happier, but it might be considered worthwhile if the goal is to raise 2 million for charity.

Not all effects in psychology are additive (we cannot combine or transfer an increase in happiness of 0.04 scale points), so it is often more difficult to argue for the importance of small effects in subjective feelings (Anvari et al., 2021). A cost-benefit analysis might show small effects matter a lot, but whether or not this is the case cannot be inferred from a  $p$ -value.

Note that nothing about this is a problem with the interpretation of a  $p$ -value per se: A  $p < 0.05$  still correctly indicates that, if the null hypothesis is true, we have observed data that should be considered surprising. However, just because data are surprising, does not mean we need to care about it. It is mainly the word 'significant' that causes confusion here – it is perhaps less confusing to think of a 'significant' effect as a 'surprising' effect, but not necessarily as an 'important' effect.

#### **1.7.4 Misconception 4: If you have observed a significant finding, the probability that you have made a Type 1 error (a false positive) is 5%.**

This misinterpretation is one possible explanation of the incorrect statement that a  $p$ -value is 'the probability that the data are observed by chance.' Assume we collect 20 observations, and Omniscient Jones tells us the null hypothesis is true (as in the example above where we generated random numbers in R). This means we are sampling from the distribution in Figure 1.12.

If this is our reality, it means that 100% of the time that we observe a significant result, it is a false positive (or Type I error). Thus, 100% of our significant results are Type 1 errors.

It is important to distinguish probabilities before collecting the data and analyzing the result, and probabilities after collecting data and analyzing the results. The Type 1 error rate controls the probability that no more than 5% of our observed mean differences will fall in the red tail areas from all studies we will perform in the future where the null hypothesis is true. But when our data falls in the tail areas with  $p < \alpha$ , and we know that the null hypothesis is true, these observed significant effects are always a Type 1 error. If you read carefully, you will notice that this misconception is caused by differences in the question that is asked. "If I have observed a  $p < .05$ , what is the probability that the null hypothesis is true?" is a different question than "If the null hypothesis is true, what is the probability of observing this (or more extreme) data?". Only the latter question is answered by a  $p$ -value. The first question

## Null hypothesis for N = 20

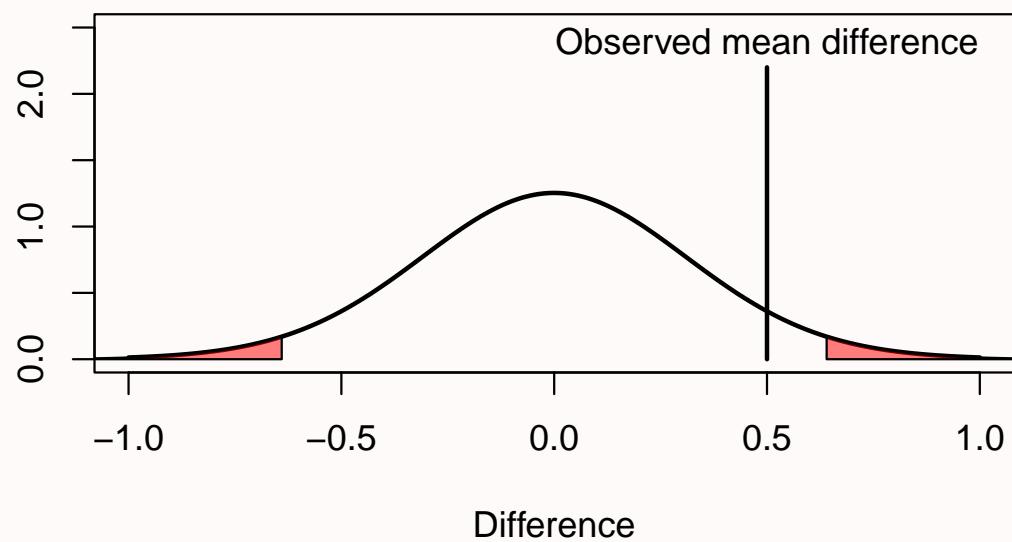


Figure 1.12: Distribution of observed Cohen's  $d$  effect sizes when collecting 20 observations per group in an independent  $t$ -test when  $d = 0$ .

cannot be answered without making a subjective judgment about the probability that the null hypothesis is true prior to collecting the data.

### **1.7.5 Misconception 5: One minus the *p*-value is the probability that the effect will replicate when repeated.**

It is impossible to calculate the probability that an effect will replicate (Miller, 2009). There are too many unknown factors to accurately predict the replication probability, and one of the main factors is the true mean difference. If we were Omniscient Jones, and we knew the true mean difference (e.g., a difference between the two groups of 0.5 scale points), we would know the statistical power of our test. The statistical power is the probability that we will find a significant result, if the alternative model is true (i.e., if there is a true effect). For example, reading the text in the left bar in the app, we see that with  $N = 50$  per group, an alpha level of 0.05, and a true mean difference of 0.5, the probability of finding a significant result (or the statistical power) is 69.69%. If we were to observe a significant effect in this scenario (e.g.,  $p = 0.03$ ) it is not true that there is a 97% probability that an exact replication of the study (with the same sample size) would again yield a significant effect. The probability that a study yields a significant effect when the alternative hypothesis is true is determined by the statistical power — not by the *p*-value in a previous study.

What we can generally take away from this last misconception is the fact that the probability of replication depends on the presence versus the absence of a true effect. In other words, as stated above, if a true effect exists then the level of statistical power informs us about how frequently we should observe a significant result (i.e., 80% power means we should observe significant result 80% of the time). On the other hand, if the null hypothesis is true (e.g., the effect is 0) then significant results will be observed only with a frequency approaching the chosen alpha level in the long run (i.e., a 5% Type 1 error rate if an alpha of 0.05 is chosen). Therefore, if the original study correctly observed an effect, the probability of a significant result in a replication study is determined by the statistical power, and if the original study correctly observed no significant effect, the probability of a significant effect in a replication study is determined by the alpha level. In practice, many other factors determine whether an effect will replicate. The only way to know if an effect will replicate, is to replicate it. If you want to explore how difficult it is to predict whether findings in the literature will replicate you can perform [this test by 80.000 Hours](#).

## **1.8 Test Yourself**

### **1.8.1 Questions about which *p*-values you can expect**

Answer each question. Then click the ‘Show Answers’ button at the bottom of this set of questions to check which questions you answered correctly.

Copy the code below to R and run the code. You can click the ‘clipboard’ icon on the top right of the code section to copy all the code to your clipboard, so you can easily paste it into R.

```
nsims <- 100000 # number of simulations

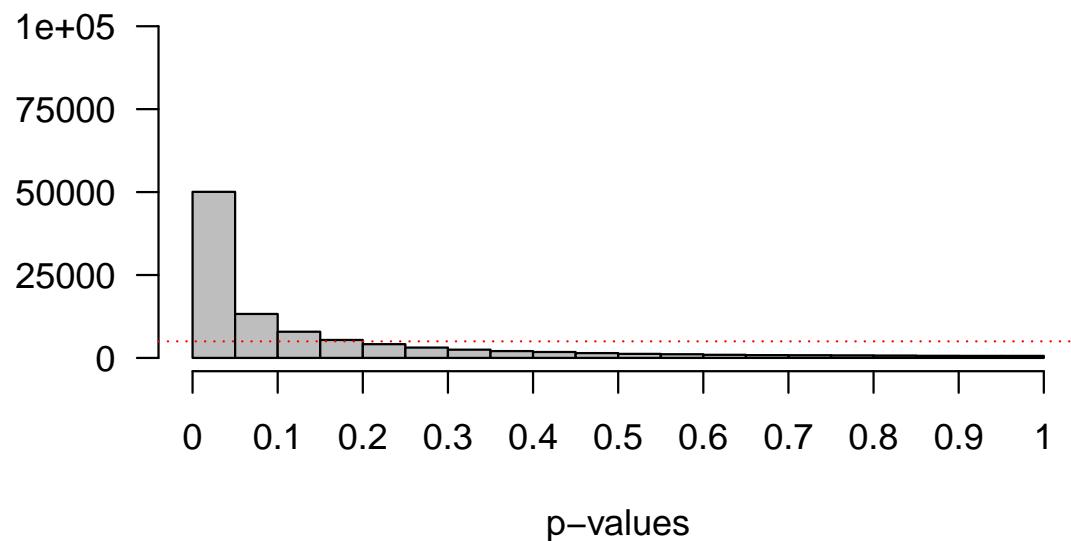
m <- 106 # mean sample
n <- 26 # set sample size
sd <- 15 # SD of the simulated data

p <- numeric(nsims) # set up empty vector
bars <- 20

for (i in 1:nsims) { # for each simulated experiment
  x <- rnorm(n = n, mean = m, sd = sd)
  z <- t.test(x, mu = 100) # perform the t-test
  p[i] <- z$p.value # get the p-value
}
power <- round((sum(p < 0.05) / nsims), 2) # power

# Plot figure
hist(p,
  breaks = bars, xlab = "p-values", ylab = "number of p-values\n",
  axes = FALSE, main = paste("p-value distribution with",
    round(power * 100, digits = 1), "% Power"),
  col = "grey", xlim = c(0, 1), ylim = c(0, nsims))
axis(side = 1, at = seq(0, 1, 0.1), labels = seq(0, 1, 0.1))
axis(side = 2, at = seq(0, nsims, nsims / 4),
  labels = seq(0, nsims, nsims / 4), las = 2)
abline(h = nsims / bars, col = "red", lty = 3)
```

## p-value distribution with 50 % Power



On the x-axis we see  $p$ -values from 0 to 1 in 20 bars, and on the y-axis we see how frequently these  $p$ -values were observed. There is a horizontal red dotted line that indicates an alpha of 5% (located at a frequency of  $100000 \cdot 0.05 = 5000$ ) – but you can ignore this line for now. In the title of the graph, the statistical power that is achieved in the simulated studies is given (assuming an alpha of 0.05): The studies have 50% power (with minor variations for each simulation).

**Q1:** Since the statistical power is the probability of observing a statistically significant result assuming there is a true effect, we can also see the power in the figure itself. Where?

- (A) We can calculate the number of  $p$ -values larger than 0.5, and divide them by the number of simulations.
- (B) We can calculate the number of  $p$ -values in the first bar (which contains all 'significant'  $p$ -values from 0.00 to 0.05) and divide the  $p$ -values in this bar by the total number of simulations.
- (C) We can calculate the difference between  $p$ -values above 0.5 minus the  $p$ -values below 0.5, and divide this number by the total number of simulations.

- (D) We can calculate the difference between  $p$ -values above 0.5 minus the  $p$ -values below 0.05, and divide this number by the number of simulations.

**Q2:** Change the sample size on line 4 in the code from  $n <- 26$  to  $n <- 51$ . Run the simulation by selecting all lines and pressing CTRL+Enter. What is the power in the simulation now that we have increased the sample size from 26 people to 51 people? Remember that simulations can sometimes yield slightly varying answers, so choose the answer option closest to the simulation results.

- (A) 55%
- (B) 60%
- (C) 80%
- (D) 95%

**Q3:** If you look at the distribution of  $p$ -values, what do you notice?

- (A) The  $p$ -value distribution is exactly the same as with 50% power
- (B) The  $p$ -value distribution is much steeper than with 50% power
- (C) The  $p$ -value distribution is much flatter than with 50% power
- (D) The  $p$ -value distribution is much more normally distributed than with 50% power

Feel free to increase and decrease the sample size and see what happens if you run the simulation. When you are done exploring, make sure that  $n <- 51$  again in line 4 before you continue.

**Q4:** What would happen when there were no true difference between our simulated samples and the average score? In this situation, we have no probability to observe an effect, so you might say we have ‘0 power’. Formally, power is not defined when there is no true effect. However, we can casually refer to this as 0 power. Change the mean in the sample to 100 (i.e., set  $m <- 106$  to  $m <- 100$ ). There is now no difference between the mean in our sample, and the population value we are testing against in the one-sample  $t$ -test. Run the script again. What do you notice?

- (A) The  $p$ -value distribution is exactly the same as with 50% power

- (B) The  $p$ -value distribution is much steeper than with 50% power
- (C) The  $p$ -value distribution is basically completely flat (ignoring some minor variation due to random noise in the simulation)
- (D) The  $p$ -value distribution is normally (i.e., bell-shaped) distributed

The question below builds on the simulation above where there was no true difference between the groups.

**Q5:** Look at the leftmost bar in the plot produced for Q4, and look at the frequency of  $p$ -values in this bar. What is the formal name for this bar?

- (A) The power (or true positives)
- (B) The true negatives
- (C) The Type 1 error (or false positives)
- (D) The Type 2 error (or false negatives)

Let's take a look at just the  $p$ -values below 0.05. Bear with me for the next few steps – it will be worth it. Find the variable that determines how many bars there are, in the statement `bars <- 20` in line 8. Change it to `bars <- 100`. We will now get 1 bar for  $p$ -values between 0 and 0.01, one bar for  $p$ -values between 0.01 and 0.02, and 100 bars in total. The red dotted line will now indicate the frequency of  $p$ -values when the null hypothesis is true, where every bar contains 1% of the total number of  $p$ -values. We only want to look at  $p$ -values below 0.05, and we will cut off the plot at 0.05. Change `xlim = c(0, 1)` to `xlim = c(0, 0.05)`. Instead of seeing all  $p$ -values between 0 and 1, we will only see  $p$ -values between 0 and 0.05. Re-run the simulation (still with `m <- 100`). We see the same uniform distribution, but now every bar contains 1% of the  $p$ -values, so the  $p$ -value distribution is very flat (we will zoom in on the y-axis later this assignment). The red line now clearly gives the frequency for each bar, assuming the null hypothesis is true.

Change the mean in the simulation in line 3 to `m <- 107` (remember, `n` is still 51). Re-run the simulation. It's clear we have very high power. Most  $p$ -values are in the left-most bar, which contains all  $p$ -values between 0.00 and 0.01.

**Q6:** The plot from the last simulation tells you that we have approximately 90.5% power (the number in your simulation might vary a bit due to random variation). This is the power if we use an alpha of 5%. But we can also use an alpha of 1%. What is the statistical power we have in the simulated studies when we would use an alpha of 1%, looking at the graph? Pick the answer closest to the answer from your simulations. Note that you can also compute the

power for an alpha of 0.01 by changing  $p < 0.05$  to  $p < 0.01$  in line 15, just make sure to set it back to 0.05 before your continue.

- (A) ~90%
- (B) ~75%
- (C) ~50%
- (D) ~5%

To be able to look at the  $p$ -values around 0.03 and 0.04, we will zoom in on the y-axis as well. In the part of the code where the plot is draw, change  $ylim = c(0, nSims)$  to  $ylim = c(0, 10000)$ . Re-run the script.

Change the mean in the sample to 108 ( $m <- 108$ ), and leave the sample size at 51. Run the simulation. Look at how the distribution has changed compared to the graph above.

Look at the fifth bar from the left. This bar now contains all the  $p$ -values between 0.04 and 0.05. You will notice something peculiar. Remember that the red dotted line indicates the frequency in each bar, assuming the null hypothesis is true. See how the bar with  $p$ -values between 0.04 and 0.05 is lower than the red line. We have simulated studies with 96% power. When power is very high,  $p$ -values between 0.04 and 0.05 are very rare – they occur less than 1% of the time (most  $p$ -values are smaller than 0.01). When the null hypothesis is true,  $p$ -values between 0.04 and 0.05 occur exactly 1% of the time (because  $p$ -values are uniformly distributed). Now ask yourself: When you have very high power, and you observe a  $p$ -value between 0.04 and 0.05, is it more likely that the null hypothesis is true, or that the alternative hypothesis is true? Given that you are more likely to observe  $p$ -values between 0.04 and 0.05 when the null hypothesis is true, than when the alternative hypothesis is true, in such a scenario, you should interpret a  $p$ -value significant with an alpha of 0.05 as more likely when the null hypothesis is true than when the alternative hypothesis is true.

In our simulations, we know whether there is a true effect or not, but in the real world, you don't know. When you have very high power, use an alpha level of 0.05, and find a  $p$ -value of  $p = .045$ , the data is surprising, assuming the null hypothesis is true, but it is even *more* surprising, assuming the alternative hypothesis is true. This shows how a significant  $p$ -value is not always evidence for the alternative hypothesis.

**Q7:** When you know you have very high (e.g., 98%) power for the smallest effect size you care about, and you observe a  $p$ -value of 0.045, what is the correct conclusion?

- (A) The effect is significant, and provides strong support for the alternative hypothesis.

- (B) The effect is significant, but it is without any doubt a Type 1 error.
- (C) With high power, you should use an alpha level that is smaller than 0.05, and therefore, this effect cannot be considered significant.
- (D) The effect is significant, but the data are more likely under the null hypothesis than under the alternative hypothesis.

**Q8:** Play around with the sample size ( $n$ ) and/or the mean ( $m$ ) by changing the numerical values (and thus, vary the statistical power in the simulated studies). Look at the simulation result for the bar that contains  $p$ -values between 0.04 and 0.05. The red line indicates how many  $p$ -values would be found in this bar if the null hypothesis was true (and is always at 1%). At the very best, how much more likely is a  $p$ -value between 0.04 and 0.05 to come from a  $p$ -value distribution representing a true effect, than it is to come from a  $p$ -value distribution when there is no effect? You can answer this question by seeing how much higher the bar of  $p$ -values between 0.04 and 0.05 can become. If at best the bar in the simulation is five times as high at the red line (so the bar shows that 5% of  $p$ -values end up between 0.04 and 0.05, while the red line remains at 1%), then at best  $p$ -values between 0.04 and 0.05 are five times as likely when there is a true effect than when there is no true effect.

- (A) At best,  $p$ -values between 0.04 and 0.05 are equally likely under the alternative hypothesis, and under the null hypothesis.
- (B) At best,  $p$ -values between 0.04 and 0.05 are approximately 4 times more likely under the alternative hypothesis, than under the null hypothesis.
- (C) At best,  $p$ -values between 0.04 and 0.05 are ~10 times more likely under the alternative hypothesis, than under the null hypothesis.
- (D) At best,  $p$ -values between 0.04 and 0.05 are ~30 times more likely under the alternative hypothesis, than under the null hypothesis.

For this reason, statisticians warn that  $p$ -values just below 0.05 (e.g., between 0.04 and 0.05) are at the very best weak support for the alternative hypothesis. If you find  $p$ -values in this range, consider replicating the study, or if that's not possible, interpret the result at least a bit cautiously. Of course, you can make a claim in a Neyman-Pearson approach that has at most a 5% Type 1 error rate. The Lindley's paradox therefore nicely illustrates the difference between different philosophical approaches to statistical inferences.

## 1.8.2 Questions about *p*-value misconceptions

Answer each question. Then click the ‘Show Answers’ button at the bottom of this set of questions to check which questions you answered correctly.

**Q1:** When the sample size in each group of an independent *t*-test is 50 observations (see Figure 1.6), which statement is correct?

- (A) The mean of the differences you will observe between the two groups is always 0.
- (B) The mean of the differences you will observe between the two groups is always different from 0.
- (C) Observing a mean difference of +0.5 or -0.5 is considered surprising, assuming the null hypothesis is true.
- (D) Observing a mean difference of +0.1 or -0.1 is considered surprising, assuming the null hypothesis is true.

**Q2:** In what sense are the null models in Figure 1.6 and Figure 1.7 similar, and in what sense are they different? Note that these plots do not contain *t*-values, but you need to infer what the *t*-values are for these distributions.

- (A) In both cases, the distributions are centered on zero, and the critical *t*-value is between 1.96 and 2 (for a two-sided test, depending on the sample size). But the larger the sample size, the closer to 0 the mean differences fall that are considered ‘surprising’.
- (B) In both cases, a *t*-value of 0 is the most likely outcome, but the critical *t*-value is around 0.4 for  $n = 50$ , and around 0.05 for  $n = 5000$ .
- (C) In both cases, means will vary in exactly the same way around 0, but the Type 1 error rate is much smaller when  $n = 5000$  than when  $n = 50$ .
- (D) Because the standard error is much larger for  $n = 50$  than for  $n = 5000$ , it is much more likely that the null hypothesis is true for  $n = 50$ .

**Q3:** You can play around with the alternative and null models in this online app: [http://shiny.ieis.tue.nl/d\\_p\\_power/](http://shiny.ieis.tue.nl/d_p_power/). The app allows you to specify the sample size in each group of an independent *t*-test (from 2 to infinity), the mean difference (from 0 to 2), and the alpha level. In the plot, the red areas visualize Type 1 errors. The blue area visualizes the Type 2 error rate. The app also tells you the critical value: There is a vertical line (with  $n = 50$

this line falls at a mean difference of 0.4) and a verbal label that says: “Effects larger than 0.4 will be statistically significant”. Note that the same is true for effects smaller than -0.4, even though there is no second label there, but the app shows the situation for a two-sided independent  $t$ -test.

You can see that on the left of the vertical line that indicates the critical mean difference there is a blue area that is part of the alternative model. This is the Type 2 error rate (or  $1 -$  the power of the study). If a study has 80% power, 80% of the mean differences we will observe should fall on the right of the critical value indicated by the line. If the alternative model is true, but we observe an effect smaller than the critical value, the observed  $p$ -value will be larger than 0.05, even when there is a true effect. You can check in the app that the larger the effect size, the further to the right the entire alternative distribution falls, and thus the higher the power. You can also see that the larger the sample size, the narrower the distribution, and the less of the distribution will fall below the critical value (as long as the true population mean is larger than the critical value). Finally, the larger the alpha level, the further to the left the critical mean difference moves, and the smaller the area of the alternative distribution that falls below the critical value.

The app also plots three graphs that illustrate the power curves as a function of different alpha levels, sample sizes, or true mean differences. Play around in the app by changing the values. Get a feel for how each variable impacts the null and alternative models, the mean difference that will be statistically significant, and the Type 1 and Type 2 error rates.

Open the app, and make sure it is set to the default settings of a sample size of 50 and an alpha level of 0.05. Look at the distribution of the null model. Set the sample size to 2. Set the sample size to 5000. The app will not allow you to plot data for a ‘group’ size of 1, but with  $n = 2$  you will get a pretty good idea of the range of values you can expect when the true effect is 0, and when you collect single observations ( $n = 1$ ). Given your experiences with the app as you change different parameters, which statement is true?

- (A) When the null hypothesis is true and the standard deviation is 1, if you randomly take 1 observation from each group and calculate the difference score, the differences will fall between -0.4 and 0.4 for 95% of the pairs of observations you will draw.
- (B) When the null hypothesis is true and the standard deviation is 1, with  $n = 50$  per group, 95% of studies where data is collected will observe in the long run a mean difference between -0.4 and 0.4.
- (C) In any study with  $n = 50$  per group, even when the SD is unknown and it is not known if the null hypothesis is true, you should rarely observe a mean difference more extreme than -0.4 or 0.4.
- (D) As the sample size increases, the expected distribution of means become narrower for the null model, but not for the alternative model.

**Q4:** Open the app once more with the default settings. Set the slider for the alpha level to 0.01 (while keeping the mean difference at 0.5 and the sample size at 50). Compared to the critical value when alpha = 0.05, which statement is true?

- (A) Compared to an alpha of 0.05, only *less* extreme values are considered surprising when an alpha of 0.01 is used, and only differences larger than 0.53 scale points (or smaller than -0.53) will now be statistically significant.
- (B) Compared to an alpha of 0.05, only *less* extreme values are considered surprising when an alpha of 0.01 is used, and only differences larger than 0.33 scale points (or smaller than -0.33) will now be statistically significant.
- (C) Compared to an alpha of 0.05, only *more* extreme values are considered surprising when an alpha of 0.01 is used, and only differences larger than 0.53 scale points (or smaller than -0.53) will be statistically significant.
- (D) Compared to an alpha of 0.05, only *more* extreme values are considered surprising when an alpha of 0.01 is used, and only differences larger than 0.33 scale points (or smaller than -0.33) will now be statistically significant.

**Q5:** Why can't you conclude that the null hypothesis is true, when you observe a statistically non-significant *p*-value ( $p > \text{alpha}$ )?

- (A) When calculating *p*-values you always need to take the prior probability into account.
- (B) You need to acknowledge the probability that you have observed a Type 1 error.
- (C) The alternative hypothesis is never true.
- (D) You need to acknowledge the probability that you have observed a Type 2 error.

**Q6:** Why can't you conclude that the alternative hypothesis is true, when you observe a statistically significant *p*-value ( $p < \text{alpha}$ )?

- (A) When calculating *p*-values you always need to take the prior probability into account.
- (B) You need to acknowledge the probability that you have observed a Type 1 error.

- (C) The alternative hypothesis is never true.
- (D) You need to acknowledge the probability that you have observed a Type 2 error.

**Q7:** A common concern when interpreting  $p$ -values is that ‘significant’ in normal language implies ‘important’, and thus a ‘significant’ effect is interpreted as an ‘important’ effect. However, **the question whether an effect is important is completely orthogonal to the question whether it is different from zero, or even how large the effect is**. Not all effects have practical impact. The smaller the effect, the less likely such effects will be noticed by individuals, but such effects might still have a large impact on a societal level. Therefore, the general take home message is that **statistical significance does not answer the question whether an effect matters in practice, or is ‘practically important’**. To answer the question whether an effect matters, you need to present a **cost-benefit analysis**.

Go to the app: [http://shiny.ieis.tue.nl/d\\_p\\_power/](http://shiny.ieis.tue.nl/d_p_power/). Set the sample size to 50000, the mean difference to 0.5, and the alpha level to 0.05. Which effects will, when observed, be statistically different from 0?

- (A) Effects more extreme than -0.01 and 0.01
- (B) Effects more extreme than -0.04 and 0.04
- (C) Effects more extreme than -0.05 and 0.05
- (D) Effects more extreme than -0.12 and 0.12

If we plot the null model for a very large sample size (e.g.,  $n = 10000$  per group) we see that even very small mean differences (differences more extreme than a mean difference of 0.04) will be considered ‘surprising’. This still means that if there really is no difference in the population, you will observe differences larger than 0.04 less than 5% of the time, in the long run, and 95% of the observed differences will be smaller than a mean difference of 0.04. But it becomes more difficult to argue for the practical significance of such effects. Imagine that a specific intervention is successful in changing people’s spending behavior, and when implementing some intervention people save 12 cents per year. It is difficult to argue how this effect will make any individual happier. However, if this money is combined, it will yield over 2 million, which could be used to treat diseases in developing countries, where it would have a real impact. The cost of the intervention might be considered too high if the goal is to make individuals happier, but it might be consider worthwhile if the goal is to raise 2 million for charity.

Not all effects in psychology are additive (we cannot combine or transfer an increase in happiness of 0.04 scale points), so it is often more difficult to argue for the importance of small

effects in subjective feelings. A cost-benefit analysis might show small effects matter a lot, but whether or not this is the case cannot be inferred from a  $p$ -value. Instead, you need to report and interpret the [effect size](#),

**Q8:** Let's assume that the random number generator in R works, and we use `rnorm(n = 50, mean = 0, sd = 1)` to generate 50 observations, and the mean of these observations is 0.5, which in a one-sample  $t$ -test against an effect of 0 yields a  $p$ -value of 0.03, which is smaller than the alpha level (which we have set to 0.05). What is the probability that we have observed a significant difference ( $p < \text{alpha}$ ) just by chance?

- (A) 3%
- (B) 5%
- (C) 95%
- (D) 100%

**Q9:** Which statement is true?

- (A) The probability that a replication study will yield a significant result is  $1-p$ .
- (B) The probability that a replication study will yield a significant result is  $1-p$  multiplied by the probability that the null hypothesis is true.
- (C) The probability that a replication study will yield a significant result is equal to the statistical power of the replication study (if there is a true effect), or the alpha level (if there is no true effect).
- (D) The probability that a replication study will yield a significant result is equal to the statistical power of the replication study + the alpha level.

This question is conceptually very similar to that asked by Tversky and Kahneman (1971) in article 'Belief in the law of small numbers':

Suppose you have run an experiment on 20 subjects, and have obtained a significant result which confirms your theory ( $z = 2.23$ ,  $p < .05$ , two-tailed). You now have cause to run an additional group of 10 subjects. What do you think the probability is that the results will be significant, by a one-tailed test, separately for this group?

Tversky and Kahneman argue a reasonable answer is 48%, but the only correct response is the same as the correct response to question 9, and the exact probability cannot be known (Miller, 2009).

“Suppose you have run an experiment on 20 subjects, and have obtained a significant result which confirms your theory ( $z = 2.23$ ,  $p < .05$ , two-tailed). You now have cause to run an additional group of 10 subjects. What do you think the probability is that the results will be significant, by a one-tailed test, separately for this group?”

Figure 1.13: Screenshot of first paragraph in Tversky and Kahneman, 1971.

**Q10:** Does a non-significant  $p$ -value (i.e.,  $p = 0.65$ ) mean that the null hypothesis is true?

- (A) No - the result could be a Type 2 error, or a false negative.
- (B) Yes, because it is a true negative.
- (C) Yes, if the  $p$ -value is larger than the alpha level then the null hypothesis is true.
- (D) No, because you need at least two non-significant  $p$ -values to conclude that the null hypothesis is true.

**Q11:** What is a correct way to present a non-significant  $p$ -value (e.g.,  $p = 0.34$  assuming an alpha level of 0.05 is used in an independent  $t$ -test)?

- (A) The null hypothesis was confirmed,  $p > 0.05$ .
- (B) There was no difference between the two conditions,  $p > 0.05$ .
- (C) The observed difference was not statistically different from 0.
- (D) The null hypothesis is true.

**Q12:** Does observing a significant  $p$ -value ( $p < .05$ ) mean that the null hypothesis is false?

- (A) No, because  $p < .05$  only means that the alternative is true, not that the null hypothesis is wrong.
- (B) No, because  $p$ -values are never a statement about the probability of a hypothesis or theory.
- (C) Yes, because an exceptionally rare event has occurred.
- (D) Yes, because the difference is statistically significant.

**Q13:** Is a statistically significant effect always a practically important effect?

- (A) No, because in extremely large samples, extremely small effects can be statistically significant, and small effects are never practically important.
- (B) No, because the alpha level could in theory be set to 0.20, and in that case a significant effect is not practically important.
- (C) No, because how important an effect is depends on a cost-benefit analysis, not on how surprising the data is under the null hypothesis.
- (D) All of the above are true.

**Q14:** What is the correct definition of a  $p$ -value?

- (A) A  $p$ -value is the probability that the null hypothesis is true, given data that is as extreme or more extreme than the data you have observed.
- (B) A  $p$ -value is the probability that the alternative hypothesis is true, given data that is as extreme or more extreme than the data you have observed.
- (C) A  $p$ -value is the probability of observing data that is as extreme or more extreme than the data you have observed, assuming the alternative hypothesis is true.
- (D) A  $p$ -value is the probability of observing data that is as extreme or more extreme than the data you have observed, assuming the null hypothesis is true.

### 1.8.3 Open Questions

1. What determines the shape of the  $p$ -value distribution?
2. How does the shape of the  $p$ -value distribution change when there is a true effect and the sample size increases?

3. What is Lindley's paradox?
4. How are  $p$ -values of continuous test statistics (e.g.,  $t$ -tests) distributed when there is no true effect?
5. What is the correct definition of a  $p$ -value?
6. Why is it incorrect to think that a non-significant  $p$ -value means that the null hypothesis is true?
7. Why is it incorrect to think that a significant  $p$ -value means that the null hypothesis is false?
8. Why is it incorrect to think that a significant  $p$ -value means that a practically important effect has been discovered?
9. Why is it incorrect to think that if you have observed a significant finding, the probability that you have made a Type 1 error (a false positive) is 5%?
10. Why is it incorrect to think that  $1 - p$  (e.g.,  $1 - 0.05 = 0.95$ ) is the probability that the effect will replicate when repeated?
11. What are differences between the Fisherian and Neyman-Pearson approach to interpreting  $p$ -values?
12. What does the null model, or the null hypothesis, represent in a null-hypothesis significance test?
13. We cannot use a null hypothesis significance test to conclude there is no (meaningful) effect. What are some statistical approaches we can use to examine if there is no (meaningful) effect?

## 2 Error control

In the previous chapter on *p-values* we learned that in the Neyman-Pearson approach to hypothesis testing the goal is to make scientific claims while controlling how often you will make a fool of yourself in the long-run. At the core of this **frequentist** approach to statistics lies the idea of **error control**: the desire to make scientific claims based on a methodological procedure that, when the assumptions are met, limits the percentage of incorrect claims to a desired maximum value. Frequentist statistics differs from *Bayesian* approaches to statistics, which focus on the probability of an event given some prior knowledge or personal belief. By focusing on long-run probabilities, frequentist statistical approaches that rely on error control can not make statements about the probability that a hypothesis is true based on the data from a single study. As Neyman and Pearson (1933) write:

But we may look at the purpose of tests from another view-point. Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long-run of experience, we shall not be too often wrong.

Researchers often do not control error rates when they make claims, and sometimes intentionally use flexibility in the data analysis to ‘p-hack’ or cherry-pick one out of many performed analyses that shows the results they wanted to see. From an error-statistical approach to statistical inferences, this is problematic behavior, as Mayo (2018) writes:

The problem with cherry picking, hunting for significance, and a host of biasing selection effects – the main source of handwringing behind the statistics crisis in science – is they wreak havoc with a method’s error probabilities. It becomes easy to arrive at findings that have not been severely tested.

### 2.1 Which outcome can you expect if you perform a study?

If you perform a study and plan to make a claim based on the statistical test you plan to perform, the long-run probability of making a correct claim or an erroneous claim is determined by three factors, namely the Type 1 error rate, the Type 2 error rate, and the probability that the null hypothesis is true. There are four possible outcomes of a statistical test, depending on whether or not the result is statistically significant, and whether or not the null hypothesis is true.

**False Positive (FP):** Concluding there is a true effect, when there is no true effect ( $H_0$  is true). This is also referred to as a **Type 1 error**, and indicated by  $\alpha$ .

**False Negative (FN):** Concluding there is no true effect, when there is a true effect ( $H_1$  is true). This is also referred to as a **Type 2 error**, and indicated by  $\beta$ .

**True Negative (TN):** Concluding there is no true effect, when there is indeed no true effect ( $H_0$  is true). This is the complement of a False Positive, and is thus indicated by  $1 - \alpha$ .

**True Positive (TP):** Concluding there is a true effect, when there is indeed a true effect ( $H_1$  is true). This is the complement of a False Negative, and is thus indicated by  $1 - \beta$ .

The probability of observing a true positive when there is a true effect is, in the long-run, equal to the **statistical power** of your study. The probability of observing a false positive when the null hypothesis is true is, in the long-run, equal to the **alpha level** you have set, or the **Type 1 error rate**.

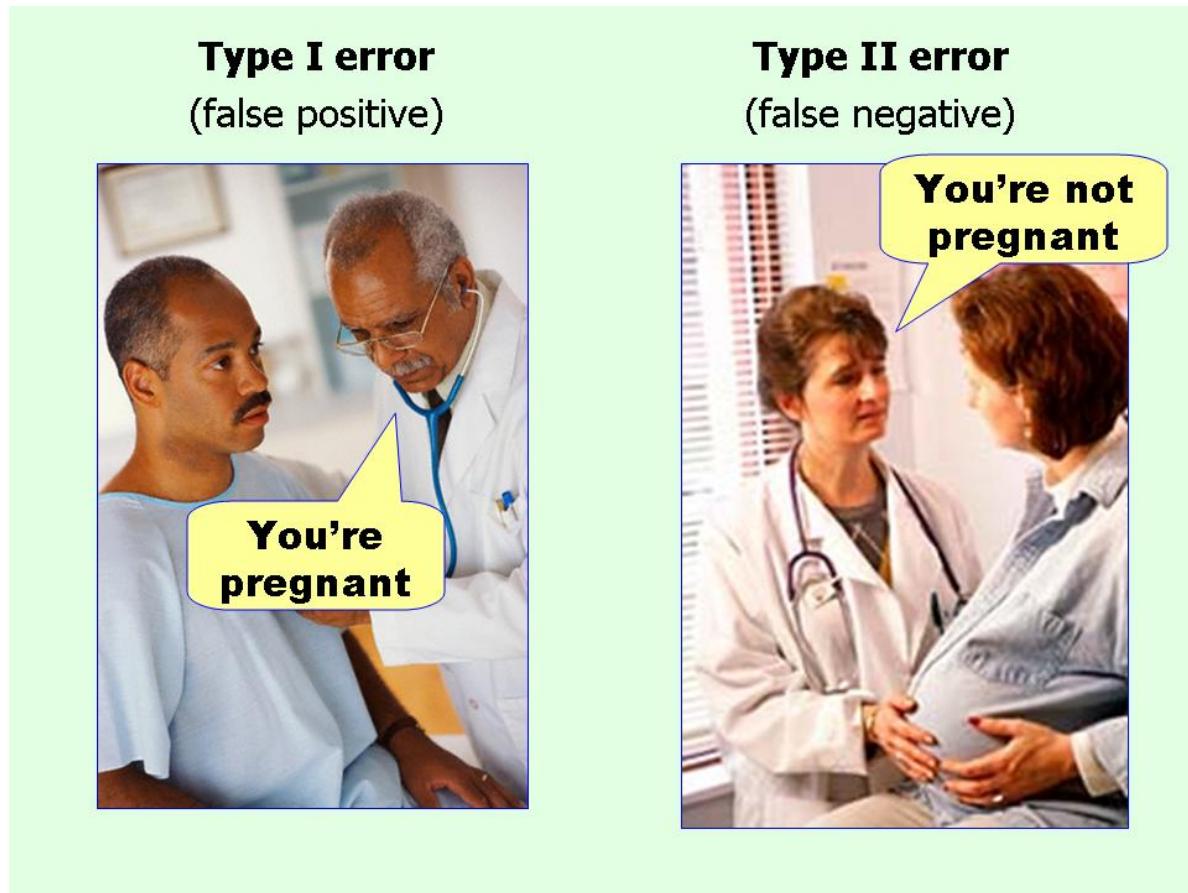


Figure 2.1: Difference between Type 1 and Type 2 errors. Figure made by Paul Ellis

So, for the next study you will perform, which of the four possible outcomes is most likely? First, let's assume you have set the alpha level to 5%. Furthermore, let's assume you have designed a study so that it will have 80% power (and for this example, let's assume that Omniscient Jones knows you indeed have exactly 80% power). The last thing to specify is the probability that the null hypothesis is true. Let's assume for this next study you have no idea if the null hypothesis is true or not, and that it is equally likely that the null hypothesis is true, or the alternative hypothesis is true (both have a probability of 50%). We can now calculate what the most likely outcome of such a study is.

Before we perform this calculation, take a moment to think if you know the answer. You might have designed studies with a 5% alpha level and 80% power, where you believed it was equally likely that  $H_0$  or  $H_1$  was true. Surely, it is useful to have reasonable expectations about which result to expect, when we perform such a study? Yet in my experience, many researchers perform without thinking about these probabilities at all. They often hope to observe a true positive, even when in the situation described above, the most likely outcome is a true negative. Let's now calculate these probabilities.

Let's assume we perform 200 studies with a 5% alpha level, 80% power, and a 50% probability that  $H_0$  is true. How many false positives, true positives, false negatives, and true negatives should we expect in the long-run?

	$H_0$ True (50%)	$H_1$ True (50%)
Significant Finding (Positive result) $\alpha = 5\%$ , $1-\beta = 80\%$	<b>False Positive 5% × 50% = 2.5% (5 studies)</b>	<b>True Positive 80% × 50% = 40% (80 studies)</b>
Non-Significant Finding (Negative result) $1-\alpha = 95\%$ , $\beta = 20\%$	<b>True Negative 95% × 50% = 47.5% (95 studies)</b>	<b>False Negative 20% × 50% = 10% (20 studies)</b>

In the table above we see that 2.5% of all studies will be a false positive (a 5% Type 1 error rate, multiplied by a 50% probability that  $H_0$  is true). 40% of all studies will be a true positive (80% power multiplied by a 50% probability that  $H_1$  is true). The probability of a false negative is 10% (a 20% Type 2 error rate multiplied by a 50% probability that  $H_1$  is true). The most likely outcome is a true negative, with 47.5% (a 95% probability observing a non-significant result, multiplied by a 50% probability that  $H_0$  is true). You can check that these percentages sum to 100, so we have covered all of the possibilities.

It might be that you are not too enthusiastic about this outlook, and you would like to perform studies that have a higher probability of observing a true positive. What should you do? We can reduce the alpha level, increase the power, or increase the probability that  $H_1$  is true. As the probability of observing a true positive depends on the power, multiplied by the probability that  $H_1$  is true, we should design studies where both of these values are high. Statistical power can be increased by changes in the design of the study (e.g., by increasing the sample size).

The probability that  $H_1$  is true depends on the hypothesis you are testing. If the probability that  $H_1$  is true is very high from the outset, you are at the risk of testing a hypothesis that is already established with enough certainty. A solution, which might not happen that often in your career, is to come up with the test of a hypothesis that is not trivial, but that which, when you explain it to your peers, makes a lot of sense to them. In other words, they would not have come up with the idea themselves, but after explaining it to them, they think it is extremely plausible. Such creative research ideas will most likely be very rare in your academic career, if you ever have any at all. Not all research needs to be this ground-breaking. It is also extremely valuable to perform **replication and extension studies** where it is relatively likely that  $H_1$  is true, but the scientific community still benefits from knowing that findings generalize to different circumstances.

## 2.2 Positive predictive value

John Ioannidis wrote a well known article titled “Why Most Published Research Findings Are False” (Ioannidis, 2005). At the same time, we have learned that if you set your alpha at 5%, the Type 1 error rate will not be higher than 5% (in the long-run). How are these two statements related? Why aren’t 95% of published research findings true? The key to understanding this difference is that two different probabilities are calculated. The Type 1 error rate is the probability of saying there is an effect, when there is no effect. Ioannidis calculates the *positive predictive value* (PPV), which is the conditional probability that if a study turns out to show a statistically significant result, there is actually a true effect. This probability is useful to understand, because people often selectively focus on significant results, and because due to publication bias, in some research areas only significant results are published.

A real-life example where it is useful to understand the concept of the positive predictive value concerns the number of vaccinated and unvaccinated people admitted to hospital with COVID-19 symptoms. In some places, official statistics showed that 50% of people who were hospitalized with COVID-19 were vaccinated. If you do not understand the concept of a positive predictive value, you might believe this reveals that you are equally likely to end up in the hospital, whether you are vaccinated or not. This is incorrect. As Figure 2.2 nicely visualizes, the probability that a person is vaccinated is very high, and the probability that a vaccinated person ends up in the hospital is much lower than the probability that an unvaccinated person ends up in the hospital. However, if we select only those individuals who end up in the hospital, we are computing a probability *conditional* on being in the hospital.

It is useful to understand what the probability is that, if you have observed a significant result in an experiment, the result is actually a true positive. In other words, in the long-run, how many *true positives* can we expect, among all positive results (both true positives and false positives)? This is known as the **Positive Predictive Value** (PPV). We can also calculate how many *false positives* we can expect, among all positive results (again, both true positives

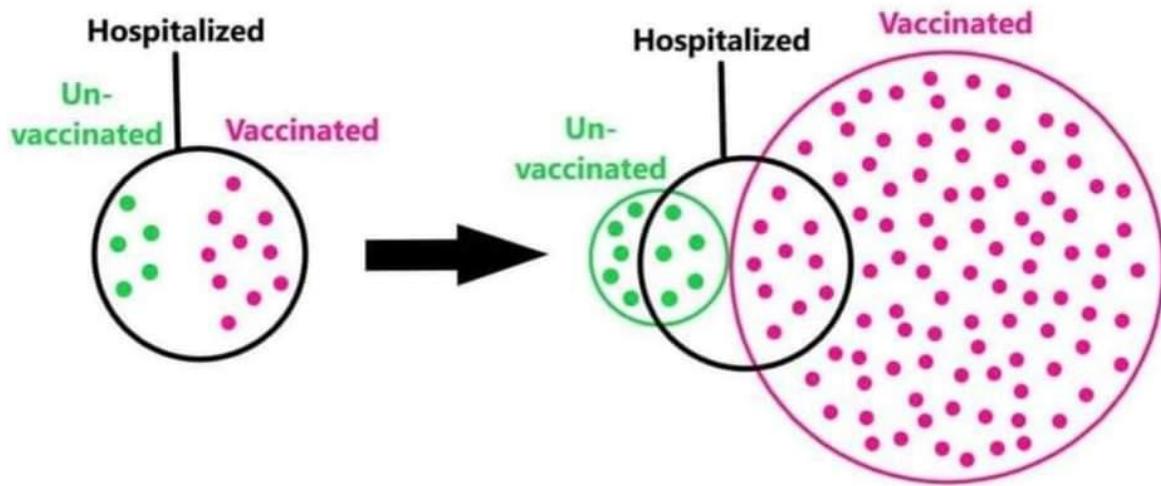


Figure 2.2: The positive predictive value can be used to explain why there are more vaccinated people in the hospital than unvaccinated people.

and false positives). This is known as the **False Positive Report Probability** (Wacholder et al., 2004), sometimes also referred to as the False Positive Risk (Colquhoun, 2019).

$$PPV = \frac{\text{True Positives}}{(\text{True Positives} + \text{False Positives})}$$

$$FPRP = \frac{\text{False Positives}}{(\text{True Positives} + \text{False Positives})}$$

The PPV and FPRP combine classic Frequentist concepts of statistical power and alpha levels with prior probabilities that  $H_0$  and  $H_1$  are true. They depend on the proportion of studies that you run where there is an effect ( $H_1$  is true), and where there is no effect ( $H_0$  is true), in addition to the statistical power, and the alpha level. After all, you can only observe a false positive if the null hypothesis is true, and you can only observe a true positive if the alternative hypothesis is true. Whenever you perform a study, you are either operating in a reality where there is a true effect, or you are operating in a reality where there is no effect – but you don't know in which reality you are.

When you perform studies, you will be aware of all outcomes of your studies (both the significant and the non-significant findings). In contrast, when you read the literature, there is publication bias, and you often only have access to significant results. This is when thinking about the PPV (and the FPRP) becomes important. If we set the alpha level to 5%, in the long-run 5% of studies where  $H_0$  is true (FP + TN) will be significant. But in a literature

with only significant results, we do not have access to all of the true negatives, and so it is possible that the proportion of false positives in the literature is much larger than 5%.

If we continue the example above, we see there are 85 positive results ( $80 + 5$ ) in the 200 studies. The false positive report probability is  $5/85 = 0.0588$ . At the same time, the alpha of 5% guarantees that (in the long-run) 5% of the 100 studies where the null hypothesis is true will Type 1 errors:  $5\% \times 100 = 5$ . When we do 200 studies, at most  $0.05 \times 200 = 10$  could possibly be false positives (if  $H_0$  was true in all experiments). In the 200 studies we performed (and where  $H_0$  was true in only 50% of the studies), the **proportion of false positives for all experiments** is only 2.5%. Thus, for all experiments you do, the proportion of false positives will, in the long-run, never be higher than the Type I error rate set by the researcher (e.g., 5% when  $H_0$  is true in all experiments), but it can be lower (when  $H_0$  is true in less than 100% of the experiments).

true positives: 40%; false negatives: 10%; true negatives: 47.5%; false positives: 2.5%

**Positive predictive value (PPV):** 94.1% of claimed findings are true

**False discovery rate (FDR):** 5.9% of claimed findings are false

Figure 2.3: Screenshot of the output of the results of the PPV Shiny app by Michael Zehetleitner and Felix Schönbrodt

(Note: FDR and FPRP are different abbreviations for the same thing.)

People often say something like: “Well, we all know 1 in 20 results in the published literature are Type 1 errors”. You should be able to understand this is not true in practice, after learning about the positive predictive value. When in 100% of the studies you perform, the null hypothesis is true, and all studies are published, only *then* are 1 in 20 studies, in the long-run, false positives (and the rest correctly reveal no statistically significant difference). It also explains why the common *p*-value misconception “If you have observed a significant finding, the probability that you have made a Type 1 error (a false positive) is 5%.” is not correct, because in practice the null hypothesis is not true in all tests that are performed (sometimes the alternative hypothesis is true). Importantly, as long as there is publication bias (where findings with desired results end up in the scientific literature, and for example non-significant results are not shared) then even if researchers use a 5% alpha level, it is quite reasonable to assume much more than 5% of significant findings in the published literature are false positives. In the scientific literature, the false positive report probability can be quite high, and under specific circumstances, it might even be so high that most published research findings are false. This will happen when researchers examine mostly studies where 1) the null hypothesis is true, 2) with low power, or 3) when the Type 1 error rate is inflated due to *p*-hacking or other types of bias.

## 2.3 Type 1 error inflation

*Of Cooking.* This is an art of various forms, the object of which is to give to ordinary observations the appearance and character of those of the highest degree of accuracy.

One of its numerous processes is to make multitudes of observations, and out of these to select those only which agree, or very nearly agree. If a hundred observations are made, the cook must be very unlucky if he cannot pick out fifteen or twenty which will do for serving up.

Figure 2.4: Quote from the 1830 book by Babbage, “Reflections on the Decline of Science in England And on Some of Its Causes.”

If you perform multiple comparisons, there is a risk that the Type 1 error rate may become inflated. When multiple comparisons are planned, in some cases it is possible to control the Type 1 error rate by lowering the alpha level for each individual analysis. The most widely known approach to control for multiple comparisons is the Bonferroni correction, where the alpha level is divided by the number of tests that is performed. However, researchers also often use informal data analysis strategies that inflate the Type 1 error rate. Babbage (1830) already complained about these problematic practices in 1830, and two centuries later, they are still common. Barber (1976) provides an in-depth discussion of a range of approaches, such as eyeballing the data to decide which hypotheses to test (sometimes called “double dipping”); selectively reporting only those analyses that confirm predictions and ignoring non-significant results, collecting many variables and performing multitudes of tests, or performing sub-group analyses when the planned analysis yields nonsignificant results; or after a nonsignificant prediction, deriving a new hypothesis that is supported by the data, and testing the hypothesis on the data that the hypothesis was derived from (sometimes called **HARKing**, Hypothesizing

After Results are Known (Kerr, 1998)). Many researchers admit to having used practices that inflate error rates (see section about [questionable research practices](#) in Chapter 15 on research integrity). I myself have used such practices in the first scientific article I published, before I was fully aware of how problematic this was - for an article that my co-authors and I published several years later in which we reflect on this, see Jostmann et al. (2016).

For some paradigms, researchers have a lot of flexibility in how to compute the main dependent variable. Elson and colleagues examined 130 publications that used the Competitive Reaction Time Task, in which participants select the duration and intensity of blasts of an unpleasant noise to be delivered to a competitor (Elson et al., 2014). The task is used to measure ‘aggressive behavior’ in an ethical manner. To compute the score, researchers can use the duration of a noise blast, the intensity, or a combination thereof, averaged over any number of trials, with several possible transformations of the data. The 130 publications that were examined reported 157 different quantification strategies in total, showing that most calculations of the dependent variable were unique, used only in a single article. One might wonder why the same authors sometimes used different computations across articles. One possible explanation is that they used this flexibility in the data analysis to find statistically significant results.

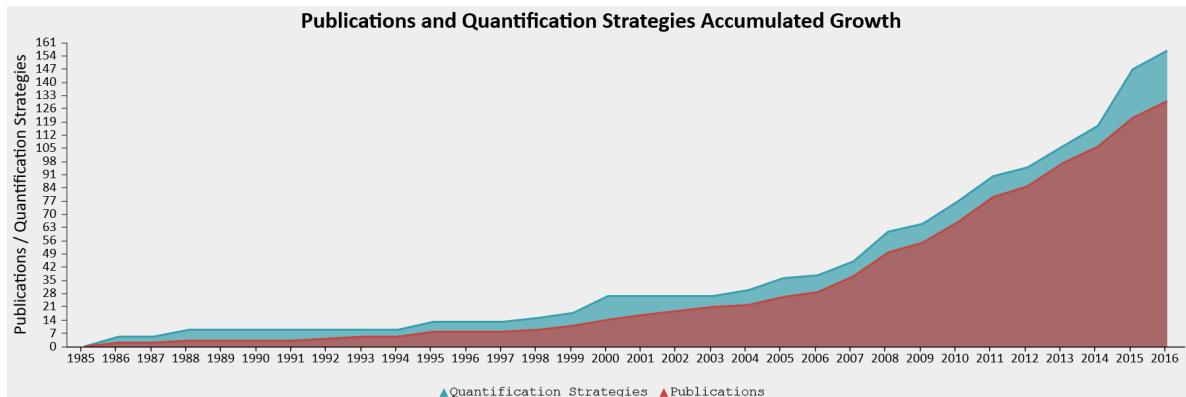


Figure 2.5: Plot of publications using CRTT (blue) and unique quantifications of the measure (red). Figure from [FlexibleMeasures.com](#) by Malte Elson.

## 2.4 Optional stopping

One practice that inflates the Type 1 error rate is known as **optional stopping**. In optional stopping a researcher repeatedly analyzes the data, continues the data collection when the test result is not statistically significant, but stops when a significant effect is observed. The quote from a published article in Figure 2.6 is an example where researchers transparently report they used optional stopping, but more commonly people do not disclose the use of optional stopping in their methods sections. In recent years, many researchers have learned that optional stopping is problematic. This has led some to the general idea that you should

**Sample sizes.** For optogenetic activation experiments, cell-type-specific ablation experiments, and in vivo recordings (optrode recordings and calcium imaging), we continuously increased the number of animals until statistical significance was reached to support our conclusions. For rabies-mediated and anterograde tracing

Figure 2.6: Screenshot a scientific paper explicitly admitting to using optional stopping.

never collect data, look at whether the results are significant, and stop data collection when the result is significant, or if not, continue data collection. That is not the correct conclusion, and is an example of becoming too inflexible. The correct approach — to collect data in batches, called **sequential analysis** — has been extensively developed by statisticians, and is used in many medical trials. We discuss [sequential analyses in Chapter 10](#). The main lesson is that certain research practices can increase the flexibility and efficiency of studies you perform, when done right, but the same practices can inflate the Type 1 error rate when done wrong. Let’s therefore try to get a better understanding of when and how we risk inflating our Type 1 error rate with optional stopping, and how to do this correctly using sequential analysis.

Copy the code below into R and run it. This script will simulate an ongoing data collection. After 10 participants in each condition, a  $p$ -value is calculated by performing an independent  $t$ -test, and this  $t$ -test is then repeated after every additional participant that is collected. Then, all these  $p$ -values are plotted as a function of the increasing sample size.

```
n <- 200 # total number of datapoints (per condition) after initial 10
d <- 0.0 # effect size d

p <- numeric(n) # store p-values
x <- numeric(n) # store x-values
y <- numeric(n) # store y-values

n <- n + 10 # add 10 to number of datapoints

for (i in 10:n) { # for each simulated participants after the first 10
  x[i] <- rnorm(n = 1, mean = 0, sd = 1)
  y[i] <- rnorm(n = 1, mean = d, sd = 1)
  p[i] <- t.test(x[1:i], y[1:i], var.equal = TRUE)$p.value
}

p <- p[10:n] # Remove first 10 empty p-values

# Create the plot
par(bg = "#fffffa")
plot(0, col = "red", lty = 1, lwd = 3, ylim = c(0, 1), xlim = c(10, n),
```

```

    type = "l", xlab = "sample size", ylab = "p-value")
lines(p, lwd = 2)
abline(h = 0.05, col = "darkgrey", lty = 2, lwd = 2) # draw line at p = 0.05

min(p) # Return lowest p-value from all looks
cat("The lowest p-value was observed at sample size", which.min(p) + 10)
cat("The p-value dropped below 0.05 for the first time at sample size:",
    ifelse(is.na(which(p < 0.05)[1] + 10), "NEVER", which(p < 0.05)[1] + 10))

```

For example, in Figure 2.7 you see the  $p$ -value plotted on the y-axis (from 0 to 1) and the sample size plotted on the x-axis (from 0 to 200). For this simulation, the true effect size was  $d = 0$ , meaning there is no true effect. We can thus only observe true negatives or false positives. As the sample size increases, the  $p$ -value slowly moves up and down (remember from the Chapter 1 on  $p$ -values that when there is no true effect,  $p$ -values are uniformly distributed). In Figure 2.7 the  $p$ -value drops below the grey line (indicating an alpha level 0.05) after collecting 83 participants in each condition, only to drift back upwards to larger  $p$ -values. From this figure, it becomes clear that the more often we look at the data, and the larger the total sample size, the higher the probability that one of the analyses will yield a  $p < \alpha$ . If resources are infinite, the Type 1 error rate will be 1, and a researcher can always find a significant result through optional stopping.

When there *is* a true effect, we see that  $p$ -values also vary, but they will eventually drop below the alpha level. We just do not know exactly when this will happen due to sampling error. When we perform an a-priori power analysis, we can compute the probability that looking at a specific sample size will yield a significant  $p$ -value. In Figure 2.8 we see the same simulation, but now when there is a true but small effect of  $d = 0.3$ . With 200 observations per condition, a sensitivity power analysis reveals that we have 85% power. If we were to analyze the data at an interim analysis (e.g., after 150 observations) we would often already find a statistically significant effect (as we would have 74% power). This illustrates a benefit of sequential analyses, where we control error rates, but can stop early at an interim analysis. Sequential analyses are especially useful in large or expensive studies where there is uncertainty about the true effect size.

Let's more formally examine the inflation of the Type 1 error rate through optional stopping in a **simulation study**. Copy the code below into R and run the code. Note that the 50000 simulations (needed to get the error rates reasonably accurate) take some time to run.

```

N <- 100 # total datapoints (per condition)
looks <- 5 # set number of looks at the data
nsims <- 50000 # number of simulated studies
alphalevel <- 0.05 # set alphalevel

if(looks > 1){

```

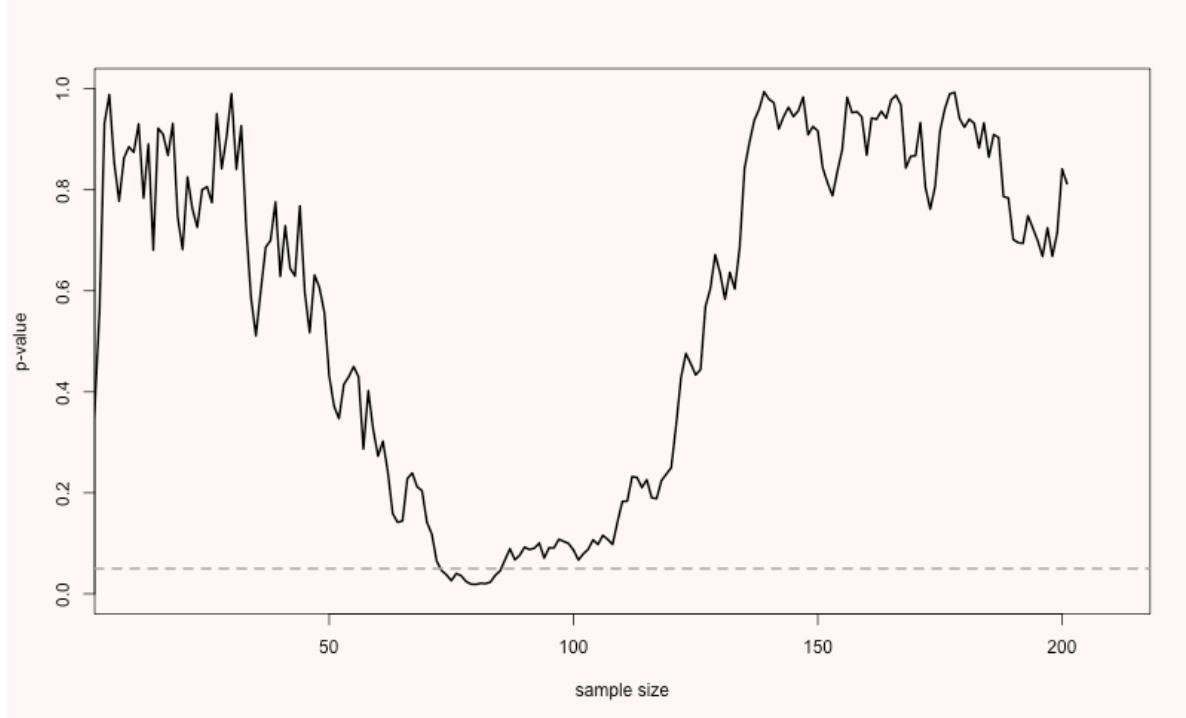


Figure 2.7: Simulated  $p$ -values for each additional observation when the null is true.

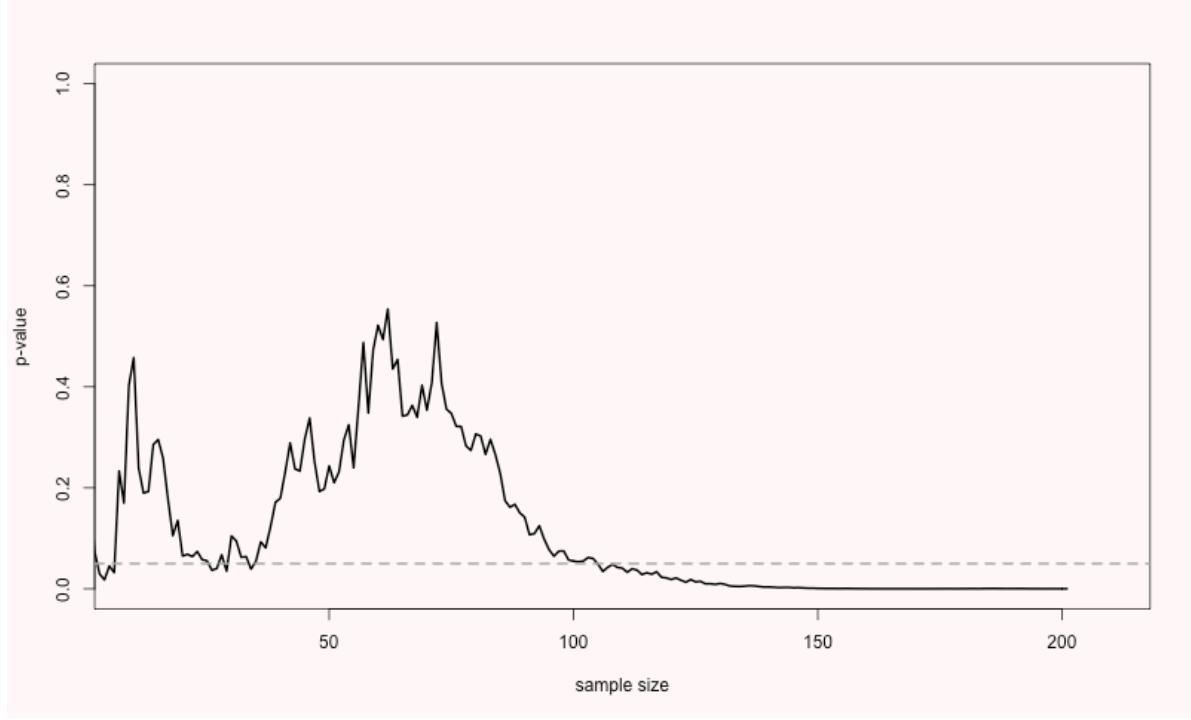


Figure 2.8: Simulated  $p$ -values for each additional observation when  $d = 0.3$ .

```

look_at_n <- ceiling(seq(N / looks, N, (N - (N / looks)) / (looks - 1)))
} else {
  look_at_n <- N
}
look_at_n <- look_at_n[look_at_n > 2] # Remove looks at N of 1 or 2
looks<-length(look_at_n) # if looks are removed, update number of looks

matp <- matrix(NA, nrow = nsims, ncol = looks) # Matrix for p-values 1 tests
p <- numeric(nsims) # Variable to save pvalues

# Loop data generation for each study, then loop to perform a test for each N
for (i in 1:nsims) {
  x <- rnorm(n = N, mean = 0, sd = 1)
  y <- rnorm(n = N, mean = 0, sd = 1)
  for (j in 1:looks) {
    matp[i, j] <- t.test(x[1:look_at_n[j]], y[1:look_at_n[j]],
                           var.equal = TRUE)$p.value # perform the t-test, store
  }
  cat("Loop", i, "of", nsims, "\n")
}

# Save Type 1 error rate smallest p at all looks
for (i in 1:nsims) {
  p[i] <- ifelse(length(matp[i,which(matp[i,] < alphalevel)]) == 0,
                 matp[i,looks], matp[i,which(matp[i,] < alphalevel)])
}

hist(p, breaks = 100, col = "grey") # create plot
abline(h = nsims / 100, col = "red", lty = 3)

cat("Type 1 error rates for look 1 to", looks, ":" ,
    colSums(matp < alphalevel) / nsims)
cat("Type 1 error rate when only the lowest p-value for all looks is reported:", 
    sum(p < alphalevel) / nsims)

```

This simulation will perform multiple independent  $t$ -tests on simulated data, looking multiple times until the maximum sample size is reached. In the first four lines, you can set the most important parameters of the simulation. First, the maximum sample size in each condition (e.g., 100). Then, the number of looks (e.g., 5). At best, you can look at the data after every participant (e.g., with 100 participants, you can look 100 times – or actually 98 times, because you need more than 2 participants in each condition for a  $t$ -test!). You can set the number of simulations (the more, the clearer the pattern will be, but the longer the simulation takes),

and the alpha level (e.g., 0.05). Since you can only make a Type 1 error when there is no true effect, the effect size is set to 0 in these simulations.

When you perform only a single test, the Type 1 error rate is the probability of finding a  $p$ -value lower than your alpha level, when there is no effect. In an optional stopping scenario where you look at the data twice, the Type 1 error rate is the probability of finding a  $p$ -value lower than your alpha level at the first look, plus the probability of **not** finding a  $p$ -value lower than your alpha level at the **first** look, but finding a  $p$ -value lower than your alpha level at the **second** look. This is a *conditional probability*, which makes error control a little bit more complex than when multiple looks are completely independent.

So how much does optional stopping inflate the Type 1 error rate? And which  $p$ -values can we expect under optional stopping?

Start by running the simulation without changing any values, so simulating 100 participants in each condition, looking 5 times at your data, with an alpha of 0.05. Note the 50.000 simulations take a while! You should see something similar to Figure 2.9 below (which is based on 500.000 simulations to make the pattern very clear).

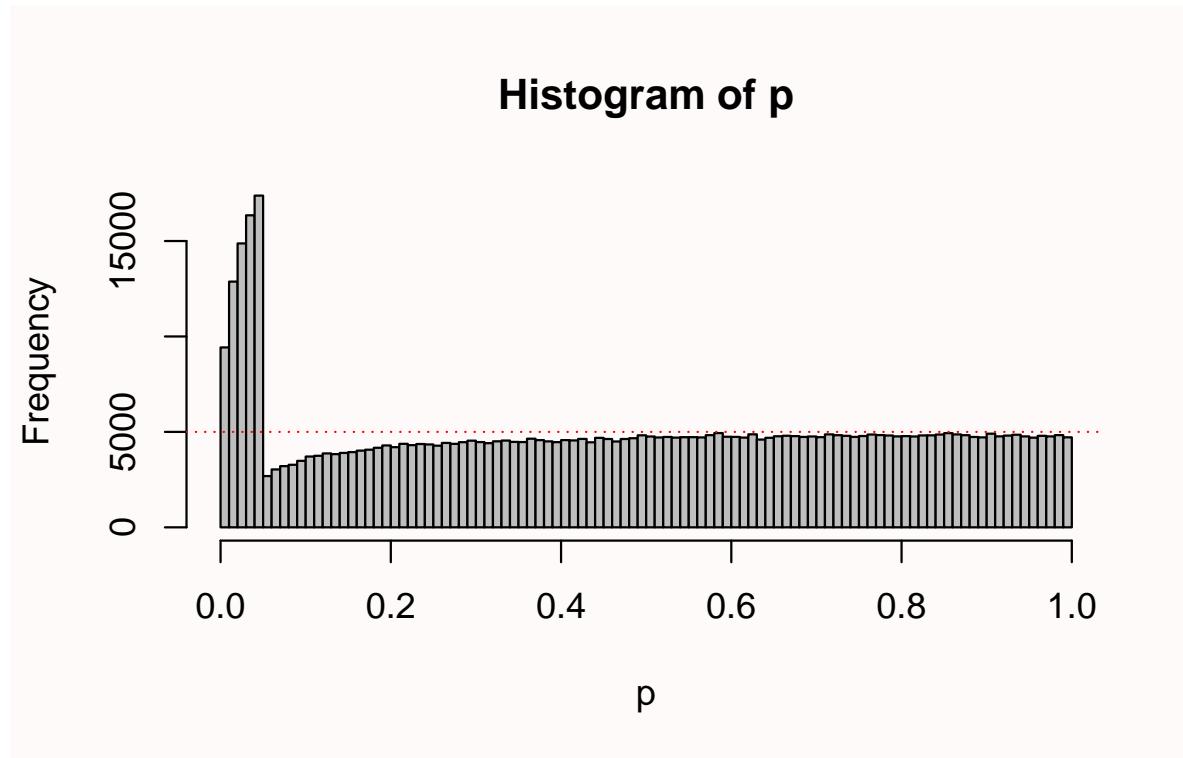


Figure 2.9: Simulation of 500000 studies performing 5 interim analyses at an alpha level of 5%.

We see 100 bars, one for each percentile (so one for all  $p$ -values between 0.00 and 0.01, one for  $p$ -values between 0.01 and 0.02, etc.). There is a horizontal line that indicates where all

$p$ -values would fall, if they were uniformly distributed (as they should be when there is no true effect, as explained in Chapter 1 on  $p$ -values).

The distribution of  $p$ -values is peculiar. We see that compared to a uniform distributions, a bunch of results just above the alpha threshold of 0.05 are missing, and they seem to have been pulled just below 0.05, where there is a much higher frequency of outcomes compared to when data is not analyzed multiple times as it comes in. Notice how relatively high  $p$ -values (e.g.,  $p = 0.04$ ) are more common than lower  $p$ -values (e.g., 0.01). We will see in Chapter 12 on [bias detection](#) that statistical techniques such as  $p$ -curve analysis can pick up on this pattern.

When using an alpha level of 5% with 5 looks at the data, the overall Type 1 error rate has inflated to 14%. If we lower the alpha level at each interim analysis, the overall Type 1 error rate can be controlled. The shape of the  $p$ -value distribution will still look peculiar, but the total number of significant test results will be controlled at the desired alpha level. The well-known Bonferroni correction (i.e., using an alpha level of  $\alpha / \text{the number of looks}$ ), but the [Pocock correction](#) is slightly more efficient. For more information on how to perform interim analyses while controlling error rates, see Chapter 10 on [sequential analysis](#).

## 2.5 Justifying Error Rates

If we reject  $H_0$ , we may reject it when it is true; if we accept  $H_0$ , we may be accepting it when it is false, that is to say, when really some alternative  $H_t$  is true. These two sources of error can rarely be eliminated completely; in some cases it will be more important to avoid the first, in others the second. We are reminded of the old problem considered by Laplace of the number of votes in a court of judges that should be needed to convict a prisoner. Is it more serious to convict an innocent man or to acquit a guilty? That will depend upon the consequences of the error; whether the punishment is death or a fine; what the danger is to the community of released criminals; and what are the current ethical views on punishment. From the point of view of mathematical theory, all that we can do is to show how the risk of the errors may be controlled and minimised. The use of these statistical tools in any given case, in determining just how the balance should be struck, must be left to the investigator.

Even though in *theory* the Type 1 and Type 2 error rate should be justified by the researcher (as Neyman and Pearson (1933) write above), in *practice* researchers tend to imitate others. The default use of an alpha level of 0.05 can already be found in the work by Gosset on the  $t$ -distribution (Cowles & Davis, 1982; Kennedy-Shaffer, 2019), who believed that a difference of two standard errors (a z-score of 2) was sufficiently rare. The default use of 80% power (or a 20% Type 2 error rate) is similarly based on personal preferences by Cohen (1988), who writes:

It is proposed here as a convention that, when the investigator has no other basis for setting the desired power value, the value .80 be used. This means that beta is set at .20. This value is offered for several reasons (Cohen, 1965, pp. 98-99). The chief among them takes into consideration the implicit convention for alpha of .05. The beta of .20 is chosen with the idea that the general relative seriousness of these two kinds of errors is of the order of .20/.05, i.e., that Type I errors are of the order of four times as serious as Type II errors. This .80 desired power convention is offered with the hope that it will be ignored whenever an investigator can find a basis in his substantive concerns about his specific research investigation to choose a value ad hoc.

We see that conventions are built on conventions: the norm to aim for 80% power is built on the norm to set the alpha level at 5%. Although there is nothing special about an alpha level of 5%, it is interesting to reflect on why it has become so widely established. Irwin Bross (1971) argues the use of an alpha level is functional and efficient when seen as an aspect of communication networks among scientists, and writes “Thus the specification of the critical levels [...] has proved in practice to be an effective method for controlling the noise in communication networks.” Bross believes the 0.05 threshold is *somewhat*, but not *completely* arbitrary, and asks us to imagine what would have happened had an alpha level of 0.001 been proposed, or an alpha level of 0.20. In both cases, he believes the convention would not have spread – in the first case because in many fields there are not sufficient resources to make claims at such a low error rate, and in the second case because few researchers would have found that alpha level a satisfactory quantification of ‘rare’ events. Uygun Tunç et al. (2023) argue that one possible reason is that, as far as conventions go, an alpha level of 5% might be low enough that peers take any claims made with this error rate seriously, while at the same time being high enough that peers will be motivated to perform an independent replication study to increase or decrease our confidence in the claim. Although lower error rates would establish claims more convincingly, this would also require more resources. One might speculate that in research areas where not every claim is important enough to warrant a careful justification of costs and benefits, 5% has a pragmatic function in facilitating conjectures and refutations in fields that otherwise lack a coordinated approach to knowledge generation, but are faced with limited resources.

Nevertheless, some researchers have proposed to move away from the default use of a 5% alpha level. For example, Johnson (2013) proposes a default significance level of 0.005 or 0.001. Others have cautioned against such blanket recommendation because the additional resources required to reduce the Type 1 error rate might not be worth the costs (Daniël Lakens, Adolfs, et al., 2018). A lower alpha level requires a larger sample size to achieve the same statistical power. If the sample size cannot be increased, a lower alpha level reduces the statistical power, and increases the Type 2 error rate. Whether that is desirable should be evaluated on a case by case basis.

There are two main reasons to abandon the universal use of a 5% alpha level. The first is that decision-making becomes more efficient (Gannon et al., 2019; Mudge et al., 2012). If

researchers use hypothesis tests to make dichotomous decisions from a methodological falsificationist approach to statistical inferences, and have a certain maximum sample size they are willing or able to collect, it is typically possible to make decisions more efficiently by choosing error rates such that the combined cost of Type 1 and Type 2 errors is minimized. If we aim to either minimize or balance Type 1 and Type 2 error rates for a given sample size and effect size, the alpha level should be set not based on convention, but by weighting the relative cost of both types of errors (Maier & Lakens, 2022).

```
Warning in ggplot2::geom_point(ggplot2::aes(x = res$minimum, y = (costT1T2 * : All aesthetics
i Please consider using `annotate()` or provide this layer with data containing
a single row.
```

For example, imagine a researcher plans to collect 64 participants per condition to detect a  $d = 0.5$  effect, and weighs the cost of Type 1 errors 4 times as much as Type 2 errors. This is exactly the scenario Cohen (1988) described, and with 64 participants per condition the relative weight of Type 1 and Type 2 errors yields a 5% Type 1 error rate and a 20% Type 2 error rate. Now imagine that this researcher realizes they have the resources to collect 80 observations instead of just 64. With an interest in an effect size of  $d = 0.5$ , the relative weight of Type 1 and Type 2 errors of 4 (as suggested by Cohen) would be satisfied if they were to set the alpha level to 0.037, as the Type 2 error rate would be 0.147. Alternatively, the researcher might have decided to collect 64 observations, but rather than balance the error rates, instead set the alpha level such that the weighted combined error rate is minimized, which is achieved when the alpha level is set to 0.033, as visualized in Figure 2.10 (for further information, see Maier & Lakens (2022)).

```
Warning in ggplot2::geom_point(ggplot2::aes(x = res$minimum, y = (costT1T2 * : All aesthetics
i Please consider using `annotate()` or provide this layer with data containing
a single row.
```

Justifying error rates can lead to situations where the alpha level is increased above 0.05, because this leads to better decision making. Winer (1962) writes:

The frequent use of the .05 and .01 levels of significance is a matter of convention having little scientific or logical basis. When the power of tests is likely to be low under these levels of significance, and when Type 1 and Type 2 errors are of approximately equal importance, the .30 and .20 levels of significance may be more appropriate than the .05 and .01 levels.

The reasoning here is that a design that has 70% power for the smallest effect size of interest would not balance the Type 1 and Type 2 error rates in a sensible manner. Of course, such an increase of the alpha level should only be deemed acceptable when authors can justify that the cost of the increase in the Type 1 error rate is sufficiently compensated by the benefit of the

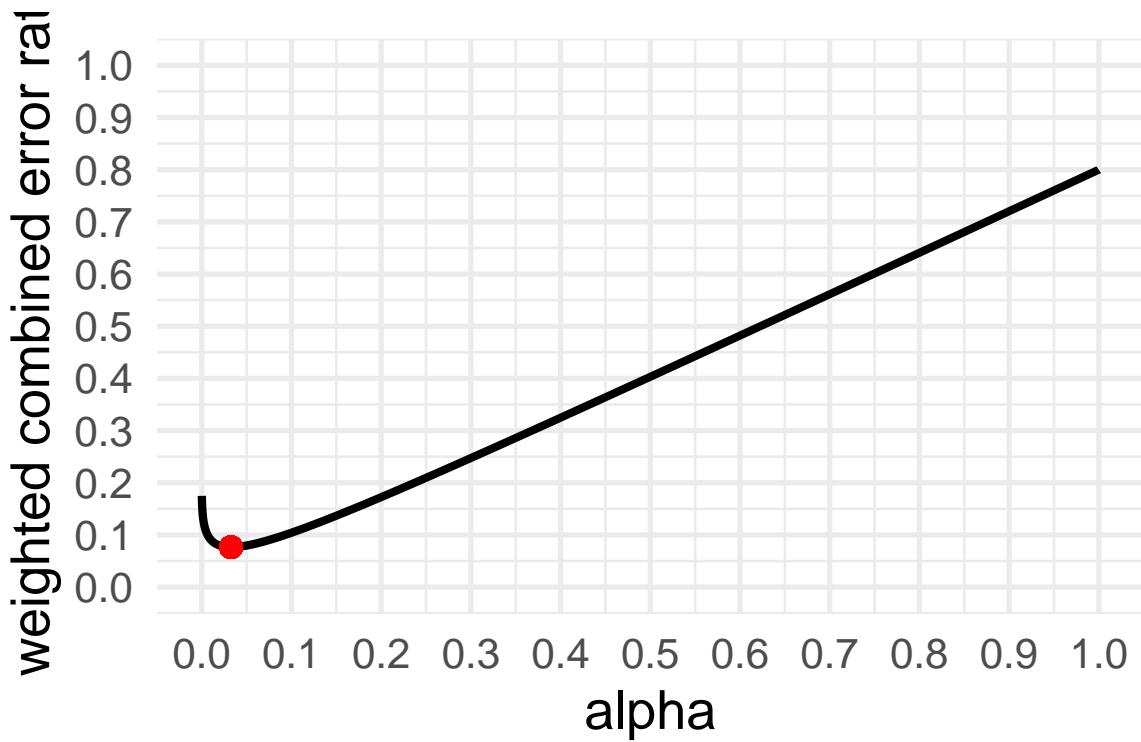


Figure 2.10: Weighted combined error rate, minimized at  $\alpha = 0.037$ .

decreased Type 2 error rate. This will encompass cases where (1) the study will have practical implications that require decision making, (2) a cost-benefit analysis is provided that gives a clear rationale for the relatively high costs of a Type 2 error, (3) the probability that  $H_1$  is false is relatively low, and (4) it is not feasible to reduce overall error rates by collecting more data.

One should also carefully reflect on the choice of the alpha level when an experiment achieves very high statistical power for all effect sizes that are considered meaningful. If a study has 99% power for effect sizes of interest, and thus a 1% Type 2 error rate, but uses the default 5% alpha level, it also suffers from a lack of balance, and the use of a lower alpha level would lead to a more balanced decision, and increase the severity of the test.

The second reason for making a study-specific choice of alpha level is most relevant for large data sets, and is related to [Lindley's paradox](#). As the statistical power increases, some  $p$ -values below 0.05 (e.g.,  $p = 0.04$ ) can be more likely when there is *no* effect than when there *is* an effect. To prevent situations where a frequentist rejects the null hypothesis based on  $p < 0.05$ , when the evidence in the test favors the null hypothesis over the alternative hypothesis, it is recommended to lower the alpha level as a function of the sample size. The need to do so is discussed by Leamer (1978), who writes “The rule of thumb quite popular now, that is, setting the significance level arbitrarily to .05, is shown to be deficient in the sense that from every reasonable viewpoint the significance level should be a decreasing function of sample size.” The idea of this approach is to reduce the alpha level such that a Bayes factor or likelihood computed for a significant result would never be evidence *for* the null hypothesis (for an online Shiny app to perform such calculations, see [here](#)).

## 2.6 Why you don't need to adjust your alpha level for all tests you'll do in your lifetime.

Some researchers criticize corrections for multiple comparisons because one might as well correct for all of the tests you will do in your lifetime (Perneger, 1998). If you choose to use a Neyman-Pearson approach to statistics the only reason to correct for all tests you perform in your lifetime is when all the work you have done in your life tests a single theory, and you would use your last words to decide to accept or reject this theory, as long as only one of all individual tests you have performed yielded a  $p < \alpha$ . Researchers rarely work like this.

Instead, in a Neyman-Pearson approach to hypothesis testing, the goal is to use data to make decisions about how to act. Neyman (1957) calls his approach **inductive behavior**. The outcome of an experiment leads one to take different possible actions, which can be either practical (e.g., implement a new procedure, abandon a research line) or scientific (e.g., claim there is or is not an effect). From an error-statistical approach (Mayo, 2018), inflated Type 1 error rates mean that it has become very likely that you will be able to claim support for your

hypothesis, even when the hypothesis is wrong. This reduces the **severity of the test**. To prevent this, we need to control our error rate at the level of our claim.

A useful distinction in the literature on multiple testing is a **union-intersection** testing approach, and an **intersection-union** testing approach (Dmitrienko & D'Agostino Sr, 2013). In a union-intersection approach, a claim is made when *at-least-one* test is significant. In these cases, a correction for multiple comparisons is required to control the error rate. In an intersection-union approach, a claim is made when all performed tests are statistically significant, and no correction for multiple comparisons is required (indeed, under some assumptions researchers could even *increase* the alpha level in a intersection-union approach).

Let's assume we collect data from 100 participants in a control and treatment condition. We collect 3 dependent variables (dv1, dv2, and dv3). In the population there is no difference between groups on any of these three variables (the true effect size is 0). We will analyze the three dv's in independent *t*-tests. This requires specifying our alpha level, and thus deciding whether we need to correct for multiple comparisons. For some reason I do not fully understand, several researchers believe it is difficult to decide when you need to correct for multiple comparisons. As Bretz, Hothorn, & Westfall (2011) write in their excellent book "Multiple Comparisons Using R": "The appropriate choice of null hypotheses being of primary interest is a controversial question. That is, it is not always clear which set of hypotheses should constitute the family  $H_1, \dots, H_m$ . This topic has often been in dispute and there is no general consensus." In one of the best papers on controlling for multiple comparisons out there, Bender & Lange (2001) write: "Unfortunately, there is no simple and unique answer to when it is appropriate to control which error rate. Different persons may have different but nevertheless reasonable opinions. In addition to the problem of deciding which error rate should be under control, it has to be defined first which tests of a study belong to one experiment."

I have never understood this confusion, at least not when working within a Neyman-Pearson approach to hypothesis testing, where the goal is to control error rates at the level of a *statistical claim*. How we control error rates depends on the claim(s) we want to make. We might want to act as if (or claim that) our treatment works if there is a difference between the treatment and control conditions on any of the three variables. This means we consider the prediction corroborated when the *p*-value of the first *t*-test is smaller than alpha level, the *p*-value of the second *t*-test is smaller than the alpha level, or the *p*-value of the third *t*-test is smaller than the alpha level. This falls under the union-intersection approach, and a researcher should correct the alpha level for multiple comparisons.

We could also want to make three different predictions. Instead of one hypothesis ("something will happen") we have three different hypotheses, and predict there will be an effect on dv1, dv2, and dv3. Each of these claims can be corroborated, or not. As these are three tests, that inform three claims, there are no multiple comparisons, and no correction for the alpha level is required.

It might seem that researchers can get out of Performing corrections for multiple comparisons by formulating a hypothesis for every possible test they will perform. Indeed, they can. For

a  $10 \times 10$  correlation matrix, a researcher might state they are testing 45 unique predictions, each at an uncorrected alpha level. However, readers might reasonably question whether these 45 tests were all predicted by a sensible theory, or if the author is just making up predictions in order to not have to correct the alpha level. Distinguishing between these two scenarios is not a *statistical* question, but a *theoretical* question. If only a few of the 45 tests corroborate the prediction, the meager track record of the predictions should make readers doubt whether the body of work that was used to derive the predictions has anything going for it.

There are different ways to control for error rates, the easiest being the Bonferroni correction and the ever-so-slightly less conservative Holm-Bonferroni sequential procedure. When the number of statistical tests becomes substantial, it is sometimes preferable to control false discovery rates, instead of error rates (Benjamini & Hochberg, 1995).

## 2.7 Power Analysis

So far we have largely focused on Type 1 error control. As was clear from Figure 2.8, when there is a true effect  $p$ -values will eventually become smaller than any given alpha level as the sample size becomes large enough. When designing an experiment, one goal might be to choose a sample size that provides a desired Type 2 error rate for an effect size of interest. This can be achieved by performing an a-priori power analysis. The statistical power of a test (and, hence, the Type 2 error rate) depends on the effect size, the sample size, and the alpha level. All else equal, the larger the effect size, the sample size, and alpha level, the higher the statistical power, and the smaller the effect size, sample size, and alpha level, the lower the statistical power.

It is important to highlight that the goal of an a-priori power analysis is *not* to achieve sufficient power for the true effect size. The true effect size is always unknown when designing a study. The goal of an a-priori power analysis is to achieve sufficient power, given a specific *assumption* of the effect size a researcher wants to detect. Just as a Type I error rate is the maximum probability of making a Type I error conditional on the assumption that the null hypothesis is true, an a-priori power analysis is computed under the assumption of a specific effect size. It is unknown if this assumption is correct. All a researcher can do is to make sure their assumptions are well justified. Statistical inferences based on a test where the Type II error is controlled are conditional on the assumption of a specific effect size. They allow the inference that, assuming the true effect size is at least as large as that used in the a-priori power analysis, the maximum Type II error rate in a study is not larger than a desired value.

In Figure 2.11 we see the expected distribution of observed standardized effect sizes (Cohen's  $d$ ) for an independent  $t$ -test with 50 observations in each condition. The bell-shaped curve on the left represents the expectations if the null is true, and the red areas in the tail represent Type 1 errors. The bell-shaped curve on the right represents the expectations if the alternative hypothesis is true, and an effect size of  $d = 0.5$ . The vertical line at  $d = 0.4$  represents the **critical effect size**. With this sample size and an alpha level of 0.05, observed effect sizes

smaller than  $d = 0.4$  will not be statistically significant. The critical effect size is independent of the true effect size (you can change  $d = 0.5$  to any other value). If there is a true effect, these outcomes will be Type 2 errors, illustrated by the blue shaded area. The remainder of the curve reflects true positives, when there is a true effect, and the observed effect sizes are statistically significant. The power of the test is the percentages of the distribution on the right that is larger than the critical value.

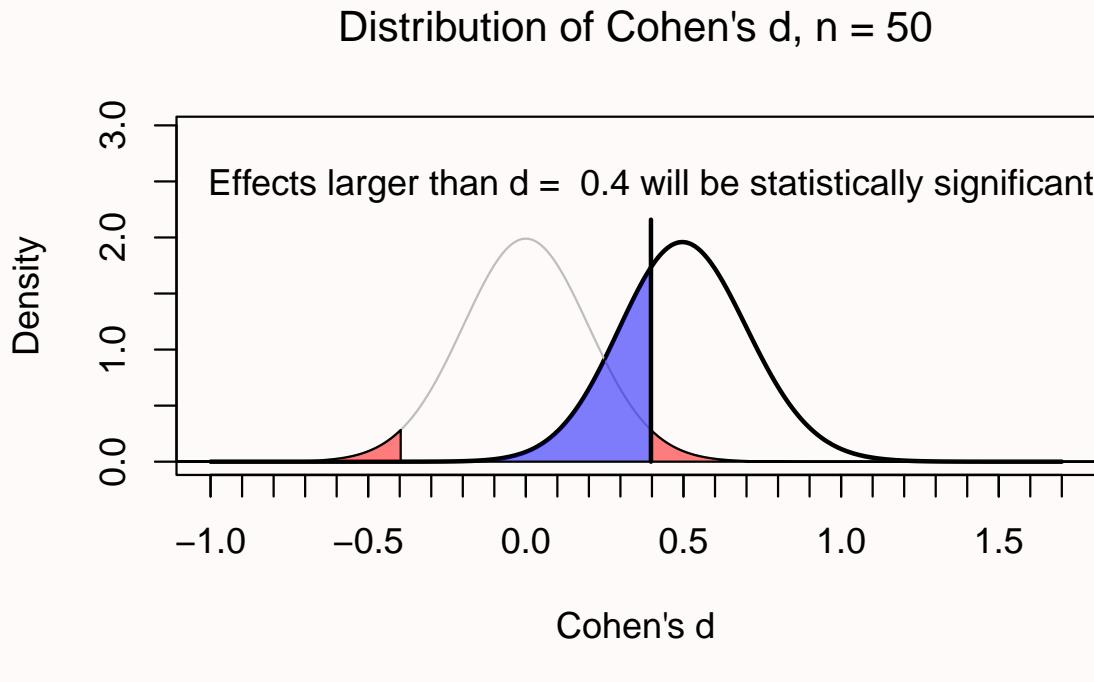


Figure 2.11: Distribution of  $d = 0$  and  $d = 0.5$  for an independent  $t$ -test with  $n = 50$ .

The issue of Type 2 error control will be discussed in more detail in Chapter 8 on [sample size justification](#). Even though the topic of Type 2 error control is only briefly discussed here, it is at least as important as Type 1 error control. An informative study should have a high probability of observing an effect if there is an effect. Indeed, the default recommendation to aim for 80% power leaves a surprisingly large (20%) probability of a Type 2 error. If a researcher only cares about not making a decision error, but the researcher does not care about whether this decision error is a false positive or a false negative, an argument could be made that Type 1 and Type 2 errors are weighed equally. Therefore, designing a study with balanced error rates (e.g., a 5% Type 1 error rate and 95% power) would make sense.

## 2.8 Test Yourself

### 2.8.1 Questions about the positive predictive value

**Q1:** In the example at the start of this chapter, we saw that we can control the Type 1 error rate at 5% by using an alpha of 0.05. Still, when there is a 50% probability that  $H_0$  is true, the proportion of false positives for all experiments performed turns out to be much lower, namely 2.5%, or 0.025. Why?

- (A) The proportion of false positives for all experiments we have performed is a variable with a distribution around the true error rate – sometimes it's higher, sometimes it's lower, due to random variation.
- (B) The proportion of false positives for all experiments we have performed is only 5% when  $H_0$  is true for all 200 studies.
- (C) The proportion of false positives for all experiments we have performed is only 5% when you have 50% power – if power increases above 50%, the proportion of false positives for all experiments we have performed becomes smaller.
- (D) The proportion of false positives for all experiments we have performed is only 5% when you have 100% power, and it becomes smaller if power is lower than 100%.

**Q2:** What will make the biggest difference in improving the probability that you will find a true positive? Check your answer by shifting the sliders in the online PPV app at <http://shinyapps.org/apps/PPV/> or <https://shiny.ieis.tue.nl/PPV/>

- (A) Increase the % of a-priori true hypotheses
- (B) Decrease the % of a-priori true hypotheses
- (C) Increase the alpha level
- (D) Decrease the alpha level
- (E) Increase the power
- (F) Decrease the power

Increasing the power requires bigger sample sizes, or studying larger effects. Increasing the % of a-priori true hypotheses can be done by making better predictions – for example building

on reliable findings, and relying on strong theories. These are useful recommendations if you want to increase the probability of performing studies where you find a statistically significant result.

**Q3:** Set the “% of a priori true hypotheses” slider to 50%. Leave the ‘ $\alpha$  level’ slider at 5%. Leave the ‘% of p-hacked studies’ slider at 0. The title of Ioannidis’ paper is ‘Why most published research findings are false’. One reason might be that studies often have low power. At which value for power is the PPV 50%? In other words, at which level of power is a significant result just as likely to be true, as that it is false?

- (A) 80%
- (B) 50%
- (C) 20%
- (D) 5%

It seems that low power alone is not the best explanation for why most published findings might be false, as it is unlikely that power is low enough in the scientific literature. Ioannidis (2005) discusses some scenarios under which it becomes likely that most published research findings are false. Some of these assume that ‘p-hacked studies’, or studies that show a significant result due to bias, enter the literature. There are good reasons to believe this happens, as we discussed in this chapter. In the ‘presets by Ioannidis’ dropdown menu, you can select some of these situations. Explore all of them, and pay close attention to the ones where the PPV is smaller than 50%.

**Q4:** In general, when are most published findings false? Interpret ‘low’ and ‘high’ in the answer options below in relation to the values in the first example in this chapter of 50% probability  $H_1$  is true, 5% alpha, 80% power, and 0% bias.

- (A) When the probability of examining a true hypothesis is low, combined with either low power or substantial bias (e.g., p-hacking).
- (B) When the probability of examining a true hypothesis is high, combined with either low power or substantial bias (e.g., p-hacking).
- (C) When the alpha level is high, combined with either low power or substantial bias (e.g., p-hacking).
- (D) When power is low and p-hacking is high (regardless of the % of true hypotheses one examines).

**Q5:** Set the “% of a priori true hypotheses” slider to 0%. Set the “% of p-hacked studies” slider to 0%. Set the “ $\alpha$  level” slider to 5%. Play around with the power slider. Which statement is true? Without  $p$ -hacking, when the alpha level is 5%, and when 0% of the hypotheses are true,

- (A) the proportion of false positives for all experiments we have performed is 100%.
- (B) the PPV depends on the power of the studies.
- (C) regardless of the power, the PPV equals the proportion of false positives for all experiments we have performed.
- (D) regardless of the power, the proportion of false positives for all experiments we have performed is 5%, and the PPV is 0% (all significant results are false positives).

### 2.8.2 Questions about optional stopping

For Questions 1 to 4, use the script below:

```
n <- 200 # total number of datapoints (per condition) after initial 10
d <- 0.0 # effect size d

p <- numeric(n) # store p-values
x <- numeric(n) # store x-values
y <- numeric(n) # store y-values

n <- n + 10 # add 10 to number of datapoints

for (i in 10:n) { # for each simulated participants after the first 10
  x[i] <- rnorm(n = 1, mean = 0, sd = 1)
  y[i] <- rnorm(n = 1, mean = d, sd = 1)
  p[i] <- t.test(x[1:i], y[1:i], var.equal = TRUE)$p.value
}

p <- p[10:n] # Remove first 10 empty p-values

# Create the plot
par(bg = "#ffffafa")
plot(0, col = "red", lty = 1, lwd = 3, ylim = c(0, 1), xlim = c(10, n),
      type = "l", xlab = "sample size", ylab = "p-value")
lines(p, lwd = 2)
```

```

abline(h = 0.05, col = "darkgrey", lty = 2, lwd = 2) # draw line at p = 0.05

min(p) # Return lowest p-value from all looks
cat("The lowest p-value was observed at sample size", which.min(p) + 10)
cat("The p-value dropped below 0.05 for the first time at sample size:",
    ifelse(is.na(which(p < 0.05)[1] + 10), "NEVER", which(p < 0.05)[1] + 10))

```

**Q1:** The script above plots the  $p$ -value as the sample size increases. Run it 20 times, and count how often the lowest  $p$ -value ends up below 0.05 (we will calculate the long-run probability of this happening through more extensive simulations later). Remember that you can click the ‘clipboard’ icon on the top right of the code section to copy all the code to your clipboard, and paste it into RStudio.

**Q2:** If there is a true effect, we can only observe a true positive or a false negative. Change the effect size in the second line of the script from  $d <- 0.0$  to  $d <- 0.3$ . This is a relatively small true effect, and with 200 participants in each condition, we have 85% power (that is, an 85% probability of finding a significant effect). Run the script again. Below is one possible example of the trajectory of  $p$ -values as the sample size increases. Run the script 20 times. Take a good look at the variation in the  $p$ -value trajectory. Remember that with  $N = 200$ , in 85% of the cases (17 out of 20), the  $p$ -value should have ended up below 0.05. The script returns the sample size the  $p$ -value is the lowest (which is often, but not always, at the maximum sample size, when there is a true effect) and the sample size at which the  $p$ -value drops below 0.05 for the first time. Which statement is true?

- (A) If the  $p$ -value drops below 0.05, it stays below 0.05.
- (B) The  $p$ -value randomly moves between 0 and 1, and will every now and then end up below 0.05.
- (C) The  $p$ -value often drops below 0.05 well before 200 participants in each condition. In around 50% of the simulations, this already happens at  $N = 100$ .
- (D) The  $p$ -value will typically move below 0.05 and stay there for some time, but given a large enough sample, it will always move back up to  $p > 0.05$ .

**Q3:** Change the effect size in the second line of the script to  $d <- 0.8$ , which can be regarded as a large effect. Run the script 20 times. Take a good look at the variation in the  $p$ -value trajectory. Which statement is true?

- (A) The  $p$ -value randomly moves between 0 and 1, and will every now and then end up below 0.05.

- (B) The  $p$ -values drop below and stay below 0.05 much earlier than when the true effect size is 0.3.
- (C)  $p$ -values are meaningful when effect sizes are large (e.g.,  $d = 0.8$ ), but meaningless when effect sizes are small (e.g.,  $d = 0.3$ ).
- (D) When you examine a large effect, whenever a  $p$ -value drops below 0.05, it will always stay below 0.05 as the sample size increases.

**Q4:** Looking at Figure 2.9, which statement is true?

- (A) Optional stopping does not impact the Type 1 error rate.
- (B) Optional stopping inflates the Type 1 error rate. We can see this in the first five bars ( $p$ -values between 0.00 and 0.05), which are substantially higher than the horizontal line.
- (C) Optional stopping inflates the Type 1 error rate. We can see this in the bars just above 0.05, which dip substantially below the uniform distribution that should be present if there is no true effect.

For Questions 5 to 8, use the script below:

```
N <- 100 # total datapoints (per condition)
looks <- 5 # set number of looks at the data
nsims <- 50000 # number of simulated studies
alphalevel <- 0.05 # set alphalevel

if(looks > 1){
  look_at_n <- ceiling(seq(N / looks, N, (N - (N / looks)) / (looks - 1)))
} else {
  look_at_n <- N
}
look_at_n <- look_at_n[look_at_n > 2] # Remove looks at N of 1 or 2
looks<-length(look_at_n) # if looks are removed, update number of looks

matp <- matrix(NA, nrow = nsims, ncol = looks) # Matrix for p-values 1 tests
p <- numeric(nsims) # Variable to save pvalues

# Loop data generation for each study, then loop to perform a test for each N
for (i in 1:nsims) {
```

```

x <- rnorm(n = N, mean = 0, sd = 1)
y <- rnorm(n = N, mean = 0, sd = 1)
for (j in 1:looks) {
  matp[i, j] <- t.test(x[1:look_at_n[j]], y[1:look_at_n[j]],
                        var.equal = TRUE)$p.value # perform the t-test, store
}
cat("Loop", i, "of", nsims, "\n")
}

# Save Type 1 error rate smallest p at all looks
for (i in 1:nsims) {
  p[i] <- ifelse(length(matp[i,which(matp[i,] < alphalevel)]) == 0,
                  matp[i,looks], matp[i,which(matp[i,] < alphalevel)])
}

hist(p, breaks = 100, col = "grey") # create plot
abline(h = nsims / 100, col = "red", lty = 3)

cat("Type 1 error rates for look 1 to", looks, ":" ,
    colSums(matp < alphalevel) / nsims)
cat("Type 1 error rate when only the lowest p-value for all looks is reported:",
    sum(p < alphalevel) / nsims)

```

**Q5:** The script to simulate optional stopping provides written output. The first line of output gives you the Type 1 error rate for each individual look at the results, and the second summary gives the Type 1 error rate when optional stopping is used. When running the script with the default values, which statement is true?

- (A) At each look, the Type 1 error rate is higher than the alpha level (0.05). When using optional stopping (and reporting only the lowest  $p$ -value), the Type 1 error rate is higher than 0.05.
- (B) At each look, the Type 1 error rate is approximately equal to the alpha level (0.05). When using optional stopping (and reporting only the lowest  $p$ -value), the alpha level also approximately equals the alpha level (0.05).
- (C) At each look, the Type 1 error rate is approximately equal to the alpha level (0.05). When using optional stopping, the Type 1 error rate is higher than the alpha level (0.05).

**Q6:** Change the number of looks in the simulation to **2** (change ‘looks <- 5’ to ‘looks <- 2’), and leave all other settings the same. Run the simulation again. What is the Type 1 error

rate using optional stopping with only 1 interim analysis, rounded to 2 digits? (Note that due to the small number of simulations, the exact alpha level you get might differ a little bit from the answer options below).

- (A) approximately 0.05
- (B) approximately 0.08
- (C) approximately 0.12
- (D) approximately 0.18

**Q7:** As Wagenmakers (2007) notes: “*a user of NHST could always obtain a significant result through optional stopping (i.e., analyzing the data as they accumulate and stopping the experiment whenever the p-value reaches some desired significance level)*”. This is correct. It’s true that the *p*-value will always drop below the alpha level at some point in time. But, we need a rather large number of observations. We can calculate the maximum Type 1 error rate due to optional stopping for any maximum sample size. For example, what is the maximum Type 1 error rate when optional stopping is used when collecting 200 participants in each condition, and looking 200 times (or 198 times, given that you can’t perform a *t*-test on a sample size of 1 or 2 people)? Set the number of participants to **200**, the number of looks to **200**, the number of simulations to **10000** (this simulation will take even longer!), and the alpha to **0.05**.

What is maximum Type 1 error rate when collecting 200 participants in each condition of an independent *t*-test, using optional stopping, rounded to 2 digits? (Note that the simulation will take a while, but still, due to the relatively small number of simulations, the exact alpha level you get might differ a little bit from the answer options below – choose the answer option closest to your result).

- (A) 0.05
- (B) 0.11
- (C) 0.20
- (D) 0.41

**Q8:** Read the Wikipedia entry about the Pocock boundary: [https://en.wikipedia.org/wiki/Pocock\\_boundary](https://en.wikipedia.org/wiki/Pocock_boundary). There can be good ethical reasons to look at the data, while it is being collected. These are clear in medicine, but similar arguments can be made for other research areas (see Lakens, 2014). Researchers often want to look at the data multiple times. This is

perfectly fine, as long as they design a study with a number of looks in advance, and control their Type 1 error rate.

The Pocock boundary provides a very easy way to control the type 1 error rate in sequential analyses. Sequential analysis is the formal way to do optional stopping. Researchers should use a slightly lower alpha level for each look, to make sure the overall alpha level (after all looks) is not larger than 5%.

Set the number of participants to **100**, the number of looks to **5**, and the number of simulations to **50000** (so back to the original script). In the Wikipedia article on the Pocock boundary, find the corrected alpha level for 5 looks at the data. Change the alpha level in the simulation to this value. Run the simulation. Which of the following statements is true?

- (A) The Type 1 error rate at each look is approximately 0.03, and the overall alpha level is approximately 0.05.
- (B) The Type 1 error rate at each look is approximately 0.03, and the overall alpha level is approximately 0.15.
- (C) The Type 1 error rate at each look is approximately 0.016, and the overall alpha level is approximately 0.05.
- (D) The Type 1 error rate at each look is approximately 0.016, and the overall alpha level is approximately 0.08.

**Q9:** Look at the graph of the *p*-value distribution when using the Pocock boundary, and compare it to the graph you obtained when not using the Pocock boundary. You can flip back and forth between plots you have generated in RStudio using the blue arrows on the Plots tab. Which statement is true?

- (A) **Without** Pocock's boundary, **small** *p*-values (e.g.,  $p = 0.01$ ) are **more** likely than slightly **higher** *p*-values ( $p = 0.04$ ). **With** Pocock's boundary, **small** *p*-values (e.g.,  $p = 0.01$ ) are **also more** likely than slightly **higher** *p*-values ( $p = 0.04$ ).
- (B) **Without** Pocock's boundary, **small** *p*-values (e.g.,  $p = 0.01$ ) are **more** likely than slightly **higher** *p*-values ( $p = 0.04$ ). **With** Pocock's boundary, **small** *p*-values (e.g.,  $p = 0.01$ ) are **less** likely than slightly **higher** *p*-values ( $p = 0.04$ ).
- (C) **Without** Pocock's boundary, **small** *p*-values (e.g.,  $p = 0.01$ ) are **less** likely than slightly **higher** *p*-values ( $p = 0.04$ ). **With** Pocock's boundary, **small** *p*-values (e.g.,  $p = 0.01$ ) are **more** likely than slightly **higher** *p*-values ( $p = 0.04$ ).

- (D) **Without** Pocock's boundary, **small**  $p$ -values (e.g.,  $p = 0.01$ ) are **less** likely than slightly **higher**  $p$ -values ( $p = 0.04$ ). **With** Pocock's boundary, **small**  $p$ -values (e.g.,  $p = 0.01$ ) are **also less** likely than slightly **higher**  $p$ -values ( $p = 0.04$ ).

### 2.8.3 Open Questions

1. What is the definition of the positive predictive value?
2. What is the definition of a false positive?
3. What is the definition of a false negative?
4. What is the definition of a true positive?
5. What is the definition of a true negative?
6. If you perform 200 studies, where there is a 50% probability  $H_0$  is true, you have 80% power, and use a 5% Type 1 error rate, what is the most likely outcome of a study?
7. How can you increase the positive predictive value in lines of research you decide to perform?
8. Why is it incorrect to think that “1 in 20 results in the published literature are Type 1 errors”?
9. What is the problem with optional stopping?
10. How do multiple tests inflate the Type 1 error rate, and what can be done to correct for multiple comparisons?
11. What is the difference between a union-intersection testing approach, and an intersection-union testing approach, and under which testing approach is it important to correct for multiple comparisons to not inflate the Type 1 error rate?
12. In a replication study, what determines the probability that you will observe a significant effect?
13. Which approach to statistical inferences is the Neyman-Pearson approach part of, and what is the main goal of the Neyman-Pearson approach?
14. How should error rates (alpha and beta) in a statistical test be determined?

## 3 Likelihoods

In addition to frequentist and Bayesian approaches to statistical inferences, likelihoods provide a third approach to statistical inferences (Dienes, 2008; Pawitan, 2001). Like [Bayesian approaches](#), which will be discussed in the next chapter, likelihoodists are interested in quantifying a measure of relative evidence when comparing two models or hypotheses. Unlike Bayesians, however, they are not too enthusiastic about the idea of incorporating prior information into their statistical inferences. As the likelihoodists Taper and Lele (2011) write:

It is not that we believe that Bayes' rule or Bayesian mathematics is flawed, but that from the axiomatic foundational definition of probability Bayesianism is doomed to answer questions irrelevant to science. We do not care what you believe, we barely care what we believe, what we are interested in is what you can show.

Likelihoodists are interested in a measure of relative evidence,. Unlike the Fisherian frequentist approach, where only  $H_0$  is specified, and lower  $p$ -values that are less compatible with the null model are interpreted as evidence against the null, likelihoodists specify a null and an alternative model, and quantify the relative likelihood of the data under both models. The Neyman-Pearson approach, in which  $H_0$  and  $H_1$  are specified, is concerned with making decisions about how to act, and does not aim to quantify evidence. At the same time, likelihood functions are an important part of both frequentist and Bayesian approaches. In the Neyman-Pearson approach, likelihoods play an important role through the Neyman-Pearson lemma, which shows that the likelihood ratio test is the most powerful test of  $H_0$  against  $H_1$ . The Neyman-Pearson lemma is used to determine the critical value to reject a hypothesis. In Bayesian approaches, the likelihood is combined with a prior to compute a posterior probability distribution.

We can use likelihood functions to make inferences about unknown quantities. Let's imagine you flip a coin 10 times, and it turns up heads 8 times. What is the true probability (which is sometimes indicated by the Greek letter  $\theta$  (theta), but we will use  $p$  in this chapter) of this coin landing on heads?

The **binomial probability** of observing  $k$  successes in  $n$  studies is:

$$Pr(k; n, p) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

where  $p$  is the probability of a success,  $k$  is the observed number of successes, and  $n$  is the number of trials. The first term indicates the number of possible combinations of results

(e.g., you could start out with eight successes, end with eight successes, or observe any of the other possible combinations of eight successes and two failures), which is multiplied by the probability of observing one success in each of the trials, which is then multiplied by the probability of observing no success in each of the remaining trials.

Let's assume you expect this is a fair coin. What is the binomial probability of observing 8 heads out of 10 coin flips, when  $p = 0.5$ ? The answer is:

$$Pr(8; 10, 0.5) = \frac{10!}{8!(10-8)!} * 0.5^8 * (1 - 0.5)^{10-8}$$

In R this probability is computed as:

```
factorial(10) / (factorial(8) * (factorial(10 - 8))) * 0.5^8 * (1 - 0.5)^(10 - 8)
```

or by using the function:

```
dbinom(x = 2, size = 10, prob = 0.5)
```

Let's assume we don't have any other information about this coin. (You might believe most coins are fair; such priors will be discussed when we talk about [Bayesian statistics](#) in the next chapter). When computing a probability, we assume the model to be known, and compute the probability of observing a specific outcome. The equation  $Pr(k;n,p)$  gives the probability of observing  $k$  successes from  $n$  trials when a coin's probability of success is  $p$ . But based on the data we have observed, we can ask the reversed question: which value of  $p$  will make the observed data **most likely**? When computing a likelihood, we assume the data to be known, and make an inference about the most likely parameter for the model. To answer this question, we can plug in the values for  $k$  and  $n$  and find which value of  $p$  maximizes this function. [Ronald Fisher](#) called this **maximum likelihood estimation**. This is considered one of the most important developments in statistics in 20th century, and Fisher published his first paper on this topic in 1912 as a third-year undergraduate when he was 22 years old (Aldrich, 1997)). Since  $p$  can be any value between 0 and 1, we can plot all values in what is known as the *likelihood function*, so that we can see the maximum more easily.

The likelihood is plotted for all possible values of  $p$  (from 0 to 1). It should not be surprising that given the data we have observed, the most likely value for the true parameter is 8 out of 10, or  $p = 0.8$ , with a likelihood of 0.30 (the highest point on the y-axis). In this example,  $p = 0.8$  is called the **maximum likelihood estimator**. It is important to know that the likelihood itself has no meaning in isolation. In this sense, it differs from a probability. But we can compare likelihoods of the same function across different values of  $p$ . You can read off any other value of the likelihood for any other  $p$ , and see that given the observed data, low values of  $p$  (e.g., 0.2) are not very likely.

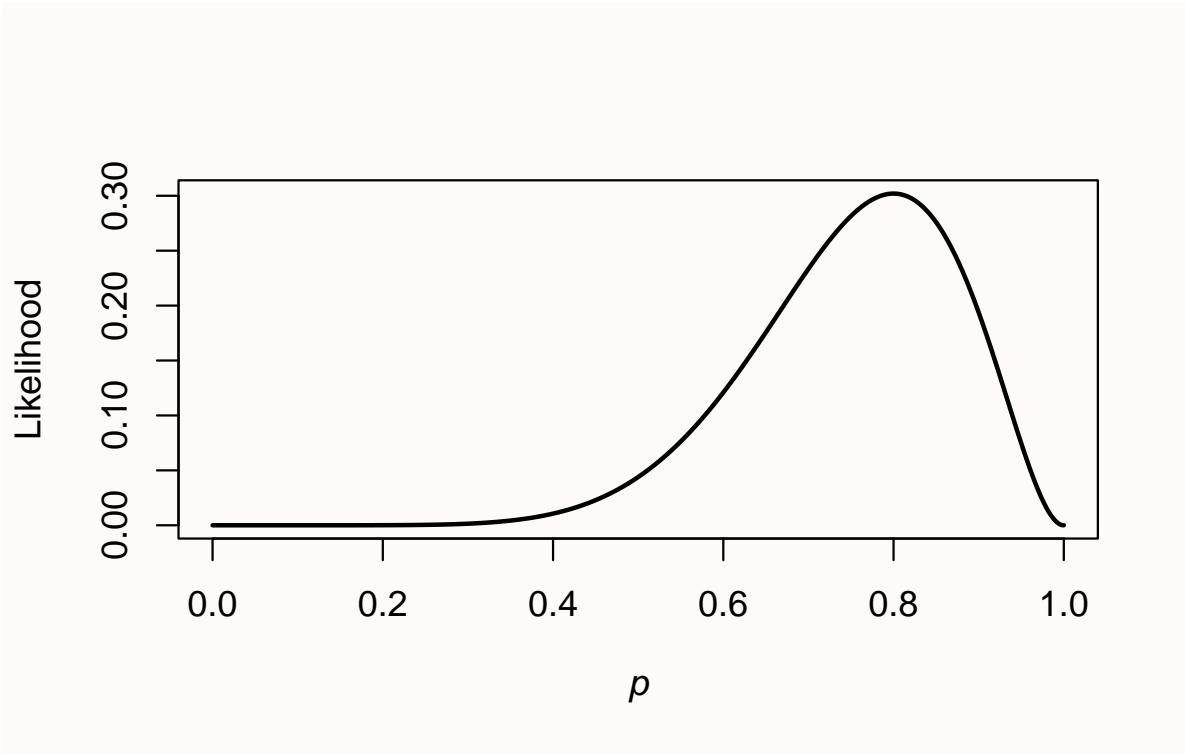


Figure 3.1: Binomial likelihood function for 8 successes in 10 trials.

There is a subtle difference between a probability and a likelihood. In colloquial language, you can use both terms to mean the same thing, but in statistics they refer to different ways of looking at the same problem. Note how the equation for  $Pr$  involves both information about the data ( $k, n$ ) and information about the parameter ( $p$ ). To compute a **probability**, we view  $p$  as fixed (for instance, for a fair coin, we plug in  $p = 0.5$ ) and then estimate the probability of different outcomes ( $k, n$ ). The resulting function is the probability mass function. To compute the **likelihood**, we instead view the observed data as fixed (e.g., observing 5 heads out of 10 coin tosses), and we view  $Pr$  as a function of  $p$ , estimating the value that maximizes the likelihood of a particular sample.

Likelihoods are an example of statistical inference: We have observed some data, and we use this data to draw an inference about different population parameters. More formally, the likelihood function is the (joint) density function evaluated at the observed data. Likelihood functions can be calculated for many different models (for example, binomial distributions, or normal distributions; see Millar (2011)). This approach is called **likelihoodist statistics**, or **likelihoodism**, and it is distinct from frequentist and Bayesian approaches to statistics, as it directly uses the likelihood function to make inferences.

When a mix of heads and tails has been observed, the likelihood curve rises and falls, as it is not possible that the coin can only come up heads or tails (after all, both have already been observed). If 10 heads or 0 heads are observed, the likelihood curve peaks at the far left or right of the x-axis. When we plot the likelihood curves for 0 heads in 10 coin flips, the likelihood curve looks like Figure 3.2.

Likelihoods can easily be combined. Imagine we have two people flipping the same coin independently. One person observes 8 heads out of 10 flips, and the other observes 4 heads out of 10 flips. You might expect that this should give the same likelihood curve as one person flipping a coin 20 times, and observing 12 heads, and indeed, it does. In the plot below, all likelihood curves are standardized by dividing each curve by its maximum likelihood. This is why all curves now have a maximum of 1, and we can more easily compare different likelihood curves.

The curve on the left is for 4 out of 10 heads, while the one on the right is for 8 out of 10 heads. The black dotted curve in the middle is for 12 out of 20 heads. The grey curve, directly beneath the 12 out of 20 heads curve, is calculated by multiplying the likelihood curves:  $L(p_{\text{combined}}) = L(p = 0.8)/L(p = 0.4)$ .

In Figure 3.4 we see likelihood curves for 10, 100, and 1000 coin flips, which yield 5, 50, and 500 heads, respectively. The likelihood curves are again standardized to make them more easily comparable. As the sample size increases, the curves become more narrow (the dashed line is for  $n = 10$ , the dotted line is for  $n = 100$ , and the solid line is for  $n = 1000$ ). This means that as the sample size increases, our data become increasingly less likely under population parameters further removed from the observed number of heads. In other words, we have collected increasingly strong evidence for  $p = 0.5$ , compared to most other possible population parameters.

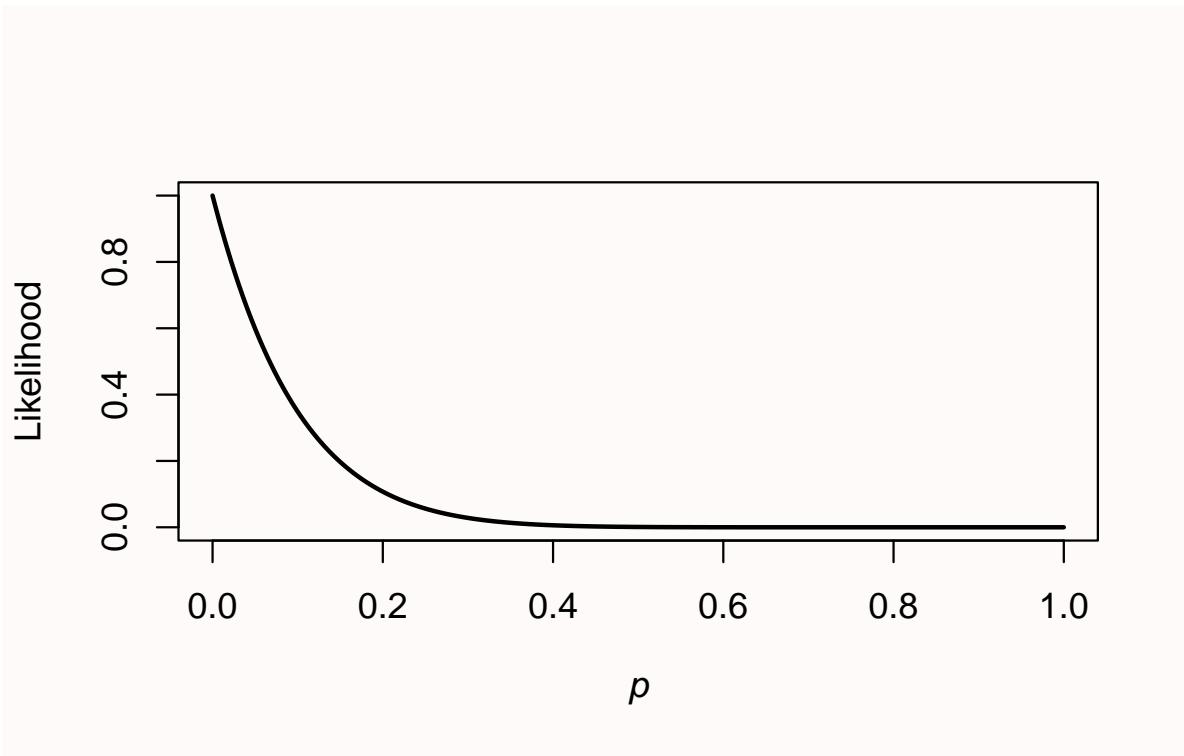


Figure 3.2: Binomial likelihood function for 0 successes in 10 trials.

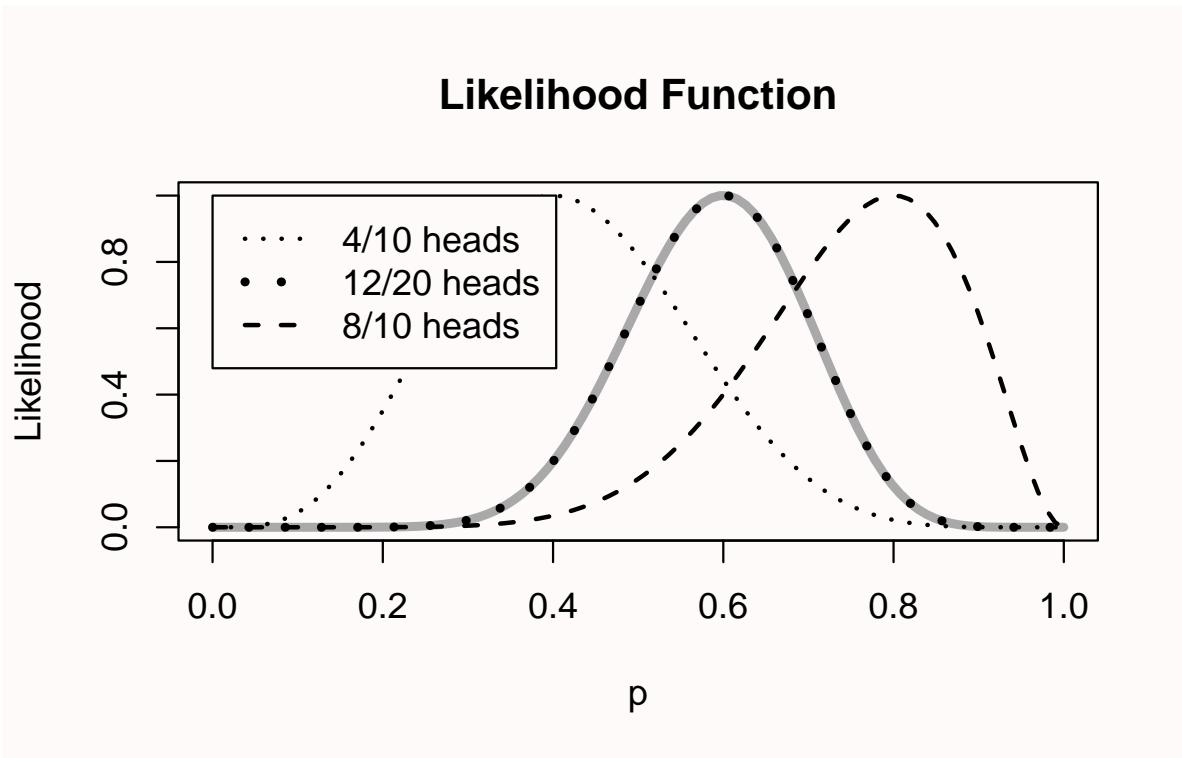


Figure 3.3: Combining likelihoods.

### Likelihood Curve for N = 10, 100, and 1000

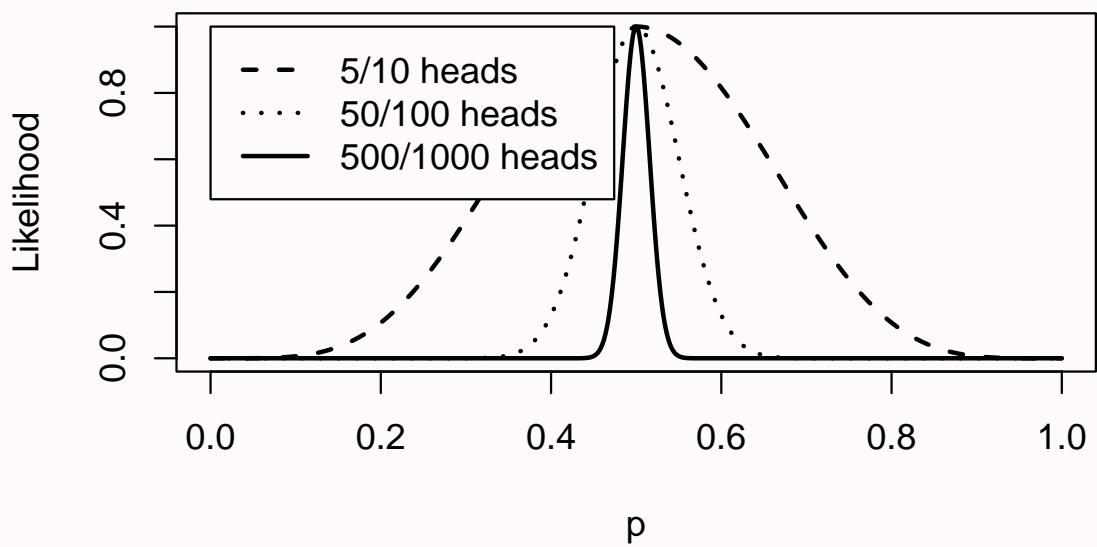


Figure 3.4: Likelihood function for 5/10, 50/100 and 500/1000 heads in coin flips.

### 3.1 Likelihood ratios

We can use the likelihood function to compare possible values of  $p$ . For example, we might believe the coin we flipped was fair, even though we flipped eight out of ten heads. A fair coin will have  $p = 0.5$ , while we observed  $p = 0.8$ . The likelihood function allows us to compute the relative likelihood for different possible parameters. How much more likely is our observed data under the hypothesis that this is an unfair coin that will on average turn up heads 80% of the time, compared to the alternative theory that this is a fair coin which should turn up heads 50% of the time?

We can calculate the likelihood ratio:

$$\frac{L(p = 0.8)}{L(p = 0.5)}$$

Which is  $0.302/0.044 = 6.87$ . In the plot, both circles show the points on the likelihood curve for  $L(p = 0.5)$  and  $L(p = 0.8)$ .

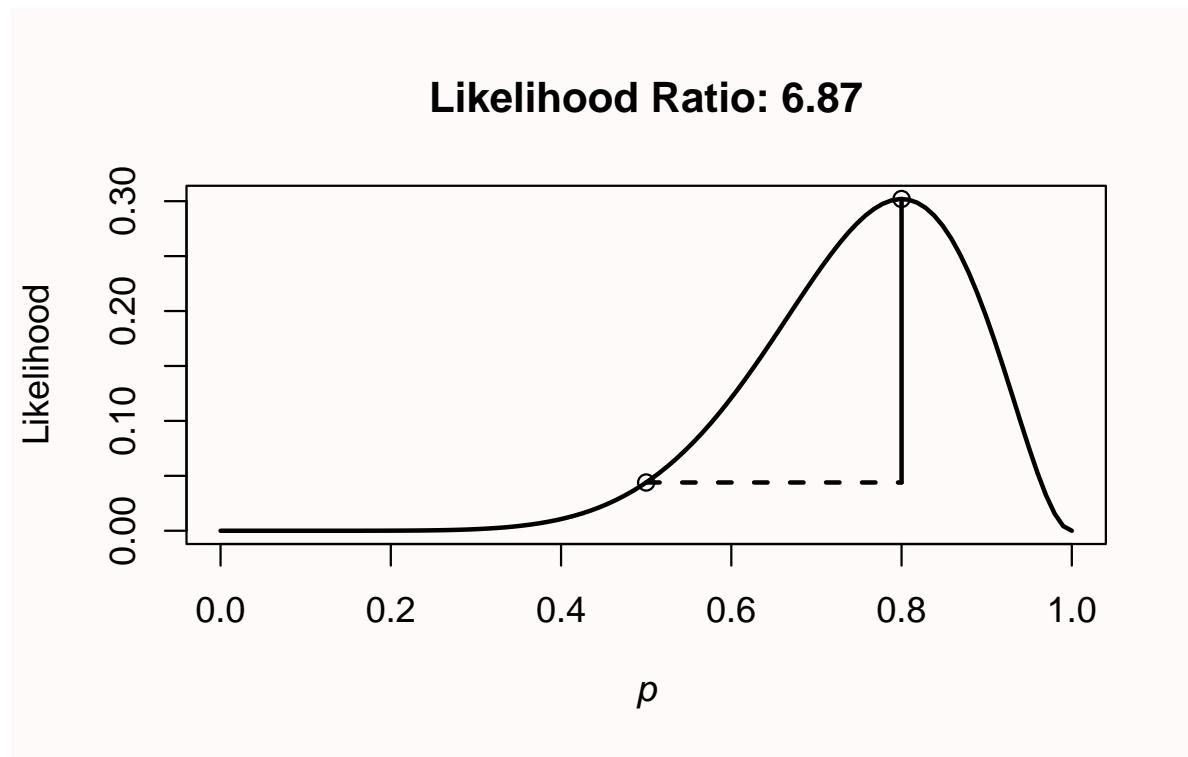


Figure 3.5: Computing a likelihood ratio for  $p = 0.5$  relative to  $p = 0.8$  when observing  $p = 0.8$ .

We can subjectively interpret this likelihood ratio, which tells us that our observed data is 6.87 times more likely under the hypothesis that this coin is unfair and will turn up heads 80% of the time, than under the hypothesis that this is a fair coin. How convincing is this? Let's round the likelihood ratio to 7, and imagine two bags of marbles. One bag contains 7 blue marbles. The second contains 7 marbles, each one a different color of the rainbow, so violet, indigo, blue, green, yellow, orange, and red. Someone randomly picks one of the two bags, draws a marble, and shows it to you. The marble is blue: How certain are you this marble came from the bag with all blue marbles, compared to the bag with rainbow coloured marbles? This is how strongly the likelihood ratio tells us to believe our data were generated by an unfair coin that turns up heads 80% of the time, relative to a fair coin, given that we have observed 8 heads in 10 tosses. After this explanation, which is intended to not make you rely too much on benchmarks, it might still be useful to know that Royall (1997) considered likelihood ratios of 8 as moderately strong evidence, and likelihood ratios of 32 as strong evidence.

Note that likelihood ratios give us the relative evidence for one specified hypothesis, over another specified hypothesis. The likelihood ratio can be calculated for any two hypothesized values. For example, in Figure 3.6 below, the likelihood ratio is calculated that compares the hypothesis for a fair coin ( $p = 0.5$ ) with the alternative hypothesis that the coin comes up heads 80% of the time ( $p = 0.8$ ), when we have observed 4 heads out of 10 coin flips. We see that the observed data are  $0.2050/0.0055 = 37.25$  times more likely (ignoring rounding differences – try to calculate these numbers by hand using the formula provided earlier) under the hypothesis that this is a fair coin than under the hypothesis that this is a coin that turns up heads 80% of the time.

A likelihood ratio of 1 means the data are equally likely under both hypotheses. Values further away from 1 indicate that the data are more likely under one hypothesis than the other. The ratio can be expressed in favor of one hypothesis over the other (for example  $L(p = 0.5)/L(p = 0.8)$  or vice versa  $(L(p = 0.8)/L(p = 0.5))$ ). This means the likelihood ratio of 37.25 for  $H_0$  relative to  $H_1$  is equivalent to a likelihood ratio of  $1/37.25 = 0.02685$  for  $H_1$  relative to  $H_0$ . Likelihood ratios range from 0 to infinity, and the closer to zero or infinity, the stronger the relative evidence for one hypothesis over the other. We will see in the chapter on [Bayesian statistics](#) that likelihood ratios are in this sense very similar (and a special case of) a Bayes Factor.

Likelihoods are relative evidence. Just because the data are more likely under one possible value of  $p$  than another value of  $p$  doesn't mean that the data have come from either of these two distributions. Other values might generate even higher likelihood values. For example, consider the situation where we flip a coin 100 times, and observe 50 heads. We compare  $p = 0.3$  versus  $p = 0.8$ , and find that the likelihood ratio is 803462, implying that there is 803462 times more evidence in the data for  $p = 0.3$  than for  $p = 0.8$ . That might sound pretty conclusive evidence for  $p = 0.3$ . But it is only relative evidence for  $p = 0.3$  compared to  $p = 0.8$ . If we look at the likelihood function, we clearly see that, not surprisingly,  $p = 0.5$  is the value that maximizes the likelihood function. Just because one hypothesis is more likely

**Likelihood Ratio: 37.25**

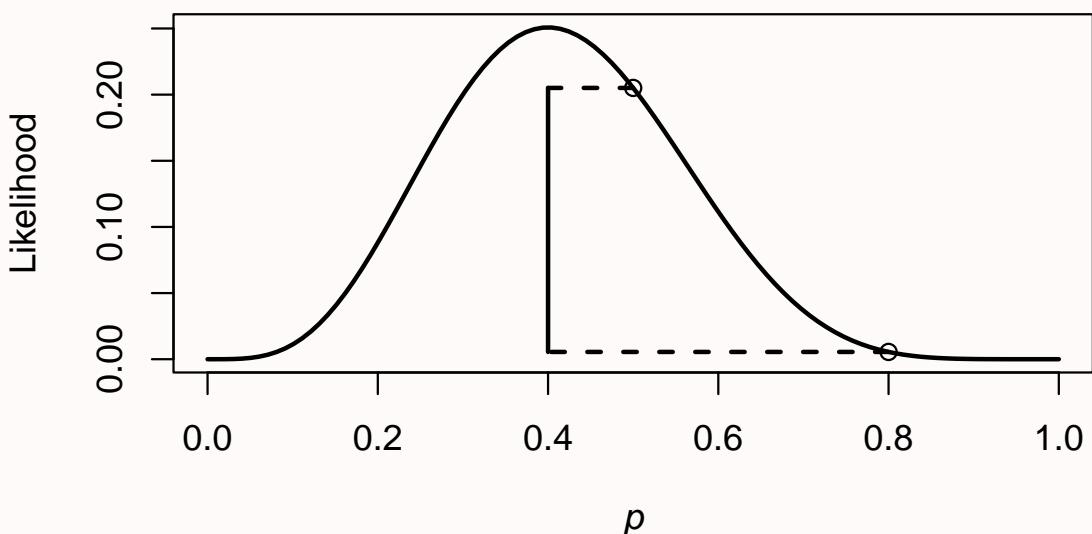


Figure 3.6: Computing a likelihood ratio for  $p = 0.5$  relative to  $p = 0.8$  when observing  $p = 0.4$ .

than another hypothesis, does not mean that there isn't a third hypothesis that is even more likely.

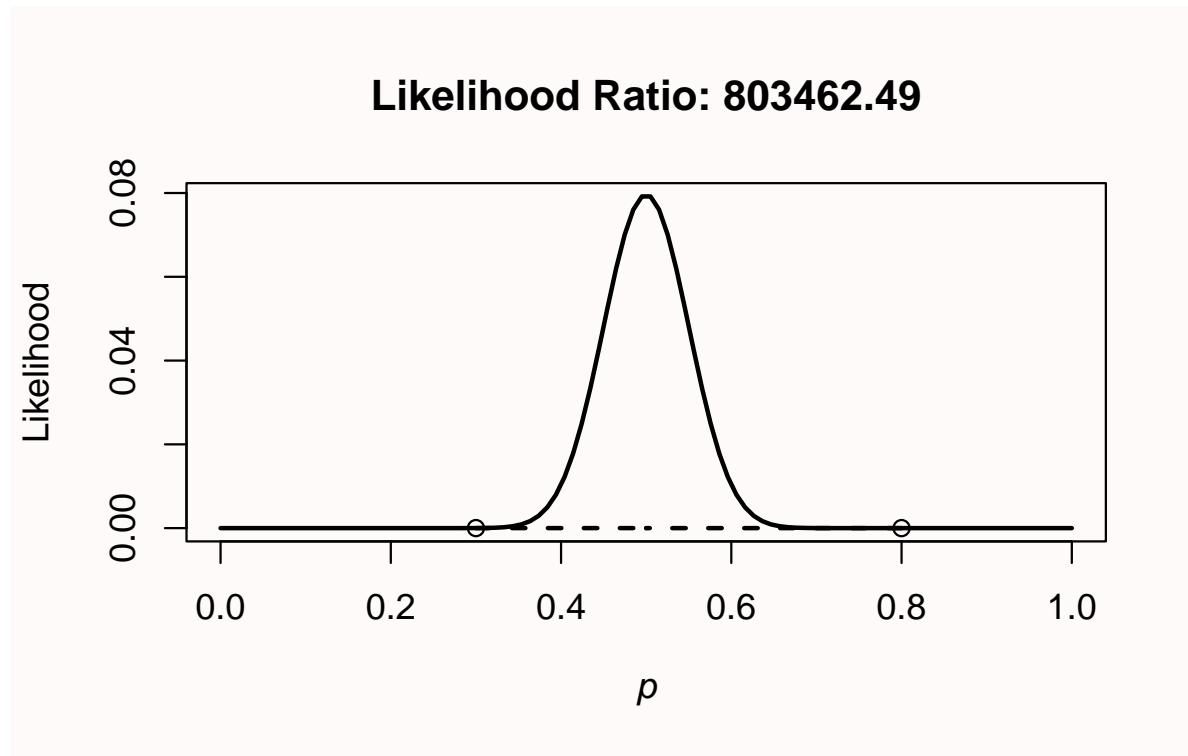


Figure 3.7: Computing a likelihood ratio for  $p = 0.3$  relative to  $p = 0.8$  when observing  $p = 0.5$  in 100 coin flips.

### 3.2 Likelihood of mixed results in sets of studies

Science is a cumulative process, and we should evaluate lines of research, not single studies. One big problem in the scientific literature is that nonsignificant results are often never published (Fanelli, 2010; Franco et al., 2014). At the same time, because the statistical power of hypothesis tests is never 100% (and often much lower), it is a mathematical reality that it is unlikely (or “too good to be true”) that a set of multiple studies yields exclusively significant results. (Francis, 2014; Schimmack, 2012). We can use binomial likelihoods to examine how likely it is to observe mixed results, and understand when mixed results are nevertheless strong evidence for the presence of an effect. The following is largely based on Daniël Lakens & Etz (2017).

The probability of observing a significant or nonsignificant result in a study depends on the Type 1 error rate ( $\alpha$ ), the statistical power of the test ( $1 - \beta$ ), and the probability that the

null hypothesis is true (Wacholder et al., 2004). There are four possible outcomes of a study: a true positive, a false positive, a true negative, and a false negative. When  $H_0$  is true, the probability of observing a false positive depends on the  $\alpha$  level or the Type 1 error rate (e.g., 5%). When  $H_1$  is true, the probability of observing a true positive depends on the statistical power of the performed test (where an often recommended minimum is 80%), which in turn depends on the  $\alpha$  level, the true effect size, and the sample size. With an  $\alpha$  level of 5%, and when  $H_0$  is true, a false positive will occur with a 5% probability (as long as error rates are controlled, e.g., in preregistered studies) and a true negative will occur with a 95% probability. When a test has 80% power, and  $H_1$  is true, a true positive has a probability of 80%, and a false negative has a probability of 20%.

If we perform multiple studies, we can calculate the binomial probability that we will observe a specific number of significant and non-significant findings (Hunt, 1975; Ioannidis & Trikalinos, 2007). For example, we can calculate the probability of finding exactly two significant results out of three studies assuming the null hypothesis is true. When  $H_0$  is true, the probability of significant results equals the  $\alpha$  level, and thus when the  $\alpha$  level is carefully controlled (e.g., in preregistered studies) the probability of observing a significant result ( $p$ ) = 0.05. That is, when  $k = 2$ ,  $n = 3$ , and  $p = .05$ , the binomial probability function tells us that the probability of finding exactly two significant results in three studies is 0.007 ( $0.05 \times 0.05 \times 0.95 = 0.002375$ , and there are three orders in which two of the three results can be observed, so  $0.002375 \times 3 = 0.007$ ).

To calculate the likelihood assuming  $H_1$  is true, we need to make an assumption about the power in each study. Let's provisionally assume all studies were powered at 80% and thus  $p = .80$ . The probability of observing exactly two significant results in three studies, assuming a power of 0.8, is 0.384 ( $0.8 \times 0.8 \times 0.2 = 0.128$ , and with three orders in which two of the three results can be significant,  $0.128 \times 3 = 0.384$ ). In other words, if you set out to perform 3 studies, your hypothesis is correct, and you test your hypothesis with 80% power, there is a 38.4% probability of observing 2 out of 3 significant results, and a 9.6% probability of observing 1 out of 3 significant results (and for an extremely unlucky individual, a 0.8% probability of not finding any significant results in three studies, even though there is a true effect). Unless power is extremely high, mixed results should be expected in sets of studies.

Both likelihoods at  $p = .05$  and  $p = .80$  are highlighted in Figure 3.8 by the circles on the dotted vertical lines. We can use the likelihood of the data assuming  $H_0$  or  $H_1$  is true to calculate the likelihood ratio,  $0.384/0.007 = 53.89$ , which tells us that the observed outcome of exactly two significant results out of three studies is 53.89 times more likely when  $H_1$  is true and studies had 80% power, than when  $H_0$  is true and studies have a carefully controlled 5% Type 1 error rate. Using Royall's (1997) proposed values for likelihood ratios of 8 and 32 as benchmarks of moderately strong and strong evidence, respectively, this implies that finding two significant results out of the three studies could be considered strong evidence for  $H_1$ , assuming 80% power. A Shiny app to perform these calculations is available [here](#).

In sets of studies, the likelihood ratio in favor of  $H_1$  versus  $H_0$  after observing a mix of significant and nonsignificant findings can become surprisingly large. Even though the evidence

**Likelihood Ratio: 53.89**

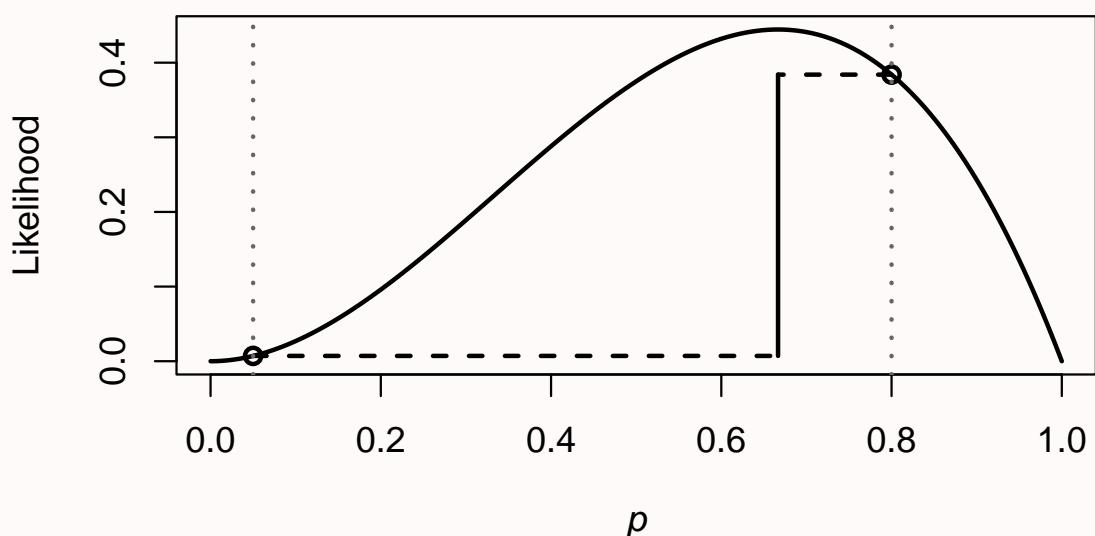


Figure 3.8: Computing a likelihood ratio for two out of three significant results, assuming an alpha of 5% and 80% power.

appears to be mixed, there is actually strong evidence in favor of a true effect. For example, when a researcher performs six studies with 80% power and a 5% alpha level and finds three significant outcomes and three nonsignificant outcomes, the cumulative likelihood ratio is convincingly large at 38-to-1 in favor of  $H_1$  to consider the set of studies strong evidence for a true effect. Intuitively, researchers might not feel convinced by a set of studies where three out of six results were statistically significant. But if we do the math, we see that such a set of studies can be very strong evidence in favor of a true effect. A better understanding of these probabilities might be an important step in mitigating the negative effects of publication bias.

Hopefully, researchers will become more inclined to submit nonsignificant findings for publication when they have a better understanding of the evidential value in lines of research with mixed results. Publishing all studies that were performed in any given line will reduce publication bias, and increase the informational value of the data in the scientific literature. Expecting all studies in lines of research to be statistically significant is not reasonable, and it is important that researchers develop more realistic expectations if they are to draw meaningful inferences from lines of research. We don't have a very good feeling for what real patterns of studies look like, because we are continuously exposed to a scientific literature that does not reflect reality. Almost all multiple study papers in the scientific literature present only statistically significant results, even though this is unlikely given the power of these studies, and the probability that we would only study correct predictions (Scheel, Schijen, et al., 2021). Educating researchers about binomial probabilities and likelihood ratios is a straightforward way to develop more realistic expectations about what research lines that contain evidential value in favor of  $H_1$  actually look like.

### 3.3 Likelihoods for $t$ -tests

So far we have computed likelihoods for binomial probabilities, but likelihoods can be computed for any statistical model (Glover & Dixon, 2004; Pawitan, 2001). For example, we can compute the relative likelihood of observing a particular  $t$ -value under the null and an alternative hypothesis as illustrated in Figure 3.9. Of course, the observed data is most likely if we assume that the observed effect equals the true effect, but examining the likelihood reveals that there are many alternative hypotheses that are relatively more likely than the null hypothesis. This also holds when observing non-significant results, which can be more likely under an alternative hypothesis of interest, than under the null hypothesis. This is one of the reasons why it is incorrect to say that there is no effect when  $p > \alpha$  (see [p-value misconception 1](#)).

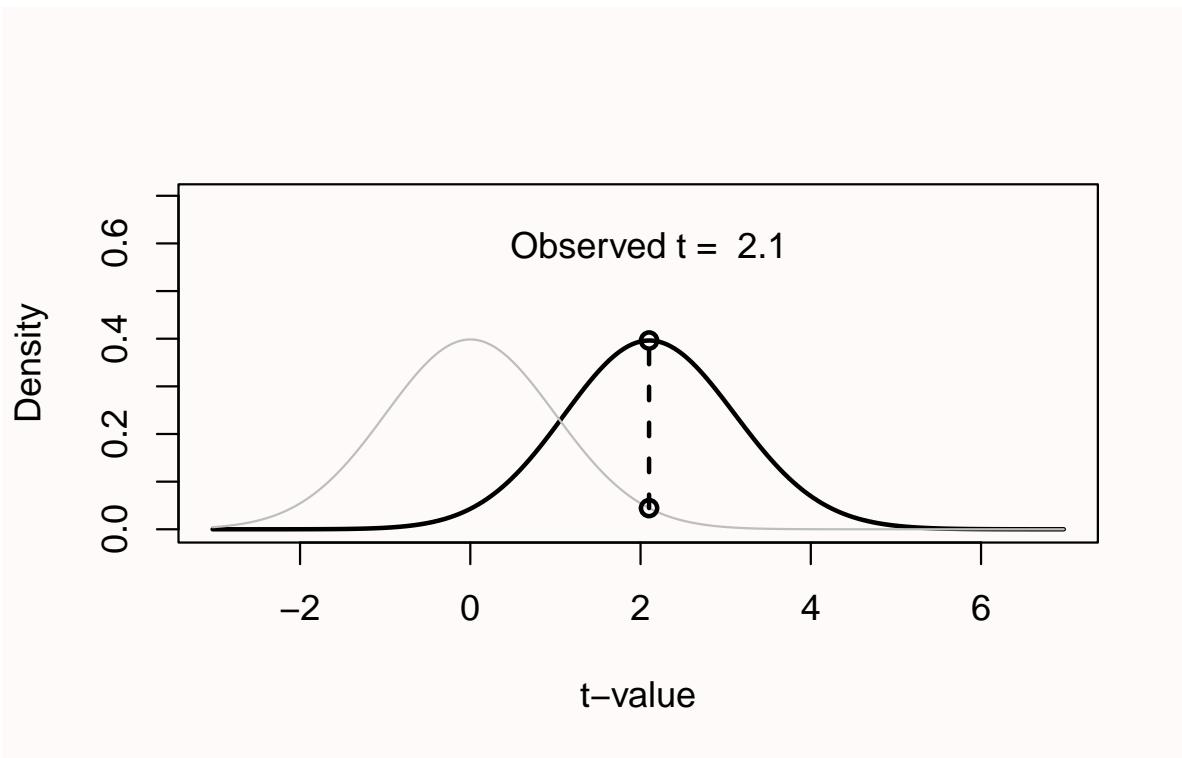


Figure 3.9: Likelihood ratio for observed  $t$ -value under  $H_0$  and  $H_1$ .

## 3.4 Test Yourself

### 3.4.1 Questions about likelihoods

**Q1:** Let's assume that you flip what you believe to be a fair coin. What is the binomial probability of observing 8 heads out of 10 coin flips, when  $p = 0.5$ ? (You can use the functions in the chapter, or compute it by hand).

- (A) 0.044
- (B) 0.05
- (C) 0.5
- (D) 0.8

**Q2:** The likelihood curve rises and falls, except in the extreme cases where 0 heads or only heads are observed. Copy the code below (remember that you can click the 'clipboard' icon on the top right of the code section to copy all the code to your clipboard), and plot the likelihood curves for 0 heads ( $x <- 0$ ) out of 10 flips ( $n <- 10$ ) by running the script. What does the likelihood curve look like?

```
n <- 10 # set total trials
x <- 5 # set successes
H0 <- 0.5 # specify one hypothesis you want to compare
H1 <- 0.4 # specify another hypothesis you want to compare
dbinom(x, n, H0) / dbinom(x, n, H1) # Returns the H0/H1 likelihood ratio
dbinom(x, n, H1) / dbinom(x, n, H0) # Returns the H1/H0 likelihood ratio

theta <- seq(0, 1, len = 100) # create probability variable from 0 to 1
like <- dbinom(x, n, theta)

plot(theta, like, type = "l", xlab = "p", ylab = "Likelihood", lwd = 2)
points(H0, dbinom(x, n, H0))
points(H1, dbinom(x, n, H1))
segments(H0, dbinom(x, n, H0), x / n, dbinom(x, n, H0), lty = 2, lwd = 2)
segments(H1, dbinom(x, n, H1), x / n, dbinom(x, n, H1), lty = 2, lwd = 2)
segments(x / n, dbinom(x, n, H0), x / n, dbinom(x, n, H1), lwd = 2)
title(paste("Likelihood Ratio H0/H1:", round(dbinom(x, n, H0) / dbinom(x, n, H1), digits = 2)))
```

- (A) The likelihood curve is a horizontal line.

- (B) The script returns an error message: It is not possible to plot the likelihood curve for 0 heads.
- (C) The curve starts at its highest point at  $p = 0$ , and then the likelihood decreases as  $p$  increases.
- (D) The curve starts at its lowest point at  $p = 0$ , and then the likelihood increases as  $p$  increases.

**Q3:** Get a coin out of your pocket or purse Flip it 13 times, and count the number of heads. Using the code above, calculate the likelihood of your observed results under the hypothesis that your coin is fair, compared to the hypothesis that the coin is not fair. Set the number of successes ( $x$ ) to the number of heads you observed. Change  $H_1$  to the number of heads you have observed (or leave it at 0 if you didn't observe any heads at all!). You can just use 4/13, or enter 0.3038. Leave  $H_0$  at 0.5. Run the script to calculate the likelihood ratio. What is the likelihood ratio of a fair compared to a non-fair coin (or  $H_0/H_1$ ) that flips heads as often as you have observed, based on the observed data? Round your answer to 2 digits after the decimal.

---

**Q4:** Earlier we mentioned that with increasing sample sizes, we had collected stronger relative evidence. Let's say we would want to compare  $L(p = 0.4)$  with  $L(p = 0.5)$ . What is the likelihood ratio if  $H_1$  is 0.4,  $H_0$  is 0.5, and you flip 5 heads in 10 trials? From the two possible ways to calculate the likelihood ratio ( $H_1/H_0$  and  $H_0/H_1$ ), report the likelihood that is 1, and round to 2 digits after the decimal point.

---

**Q5:** What is the likelihood ratio if  $H_1$  is 0.4,  $H_0$  is 0.5, and you flip 50 heads in 100 trials? From the two possible ways to calculate the likelihood ratio ( $H_1/H_0$  and  $H_0/H_1$ ), report the likelihood that is 1, and round to 2 digits after the decimal point.

---

**Q6:** What is the likelihood ratio if  $H_1$  is 0.4,  $H_0$  is 0.5, and you flip 500 heads in 1000 trials? From the two possible ways to calculate the likelihood ratio ( $H_1/H_0$  and  $H_0/H_1$ ), report the likelihood that is 1, and round to 2 digits after the decimal point.

---

**Q7:** When comparing two hypotheses ( $p = X$  vs  $p = Y$ ), a likelihood ratio of:

- (A) 0.02 means that there is not enough evidence in the data for either of the two hypotheses.
- (B) 5493 means that hypothesis  $p = X$  is most supported by the data.
- (C) 5493 means that hypothesis  $p = X$  is much more supported by the data than  $p = Y$ .
- (D) 0.02 means that the hypothesis that the data are 2% more likely under the hypothesis that  $p = X$  than under the hypothesis that  $p = Y$ .

### 3.4.2 Questions about mixed results

A Shiny app to perform the calculations is available [here](#).

**Q8:** Which statement is correct when you perform 3 studies?

- (A) When  $H_1$  is true, alpha = 0.05, and power = 0.80, you are almost as likely to observe one or more non-significant results (48.8%) as to observe only significant results (51.2%).
- (B) When alpha = 0.05 and power = 0.80, it is extremely rare that you will find 3 significant results (0.0125%), regardless of whether  $H_0$  is true or  $H_1$  is true.
- (C) When alpha = 0.05 and power = 0.80, 2 out of 3 statistically significant results is the most likely outcome of all possible outcomes (0 out of 3, 1 out of 3, 2 out of 3, or 3 out of 3), and occurs 38.4% of the time when  $H_1$  is true.
- (D) When alpha = 0.05 and power = 0.80, the probability of finding at least one false positive (a significant result when  $H_0$  is true) in three studies is 5%.

**Q9:** Sometimes in a set of three studies, you'll find a significant effect in one study, but there is no effect in the other two related studies. Assume the two related studies were not exactly the same in every way (e.g., you changed the manipulation, or the procedure, or some of the questions). It could be that the two other studies did not work because of minor differences that had some effect that you do not fully understand yet. Or it could be that the single significant result was a Type 1 error, and  $H_0$  was true in all three studies. Which statement below is correct, assuming a 5% Type 1 error rate and 80% power?

- (A) All else being equal, the probability of a Type 1 error in one of three studies is 5% when there is no true effect in all three studies, and the probability of finding exactly 1 out of 3 significant effects, assuming 80% power in all three studies, is 80%, which is substantially more likely.
- (B) All else being equal, the probability of a Type 1 error in one of three studies is 13.5% when there is no true effect in all three studies, and the probability of finding exactly 1 out of 3 significant effects, assuming 80% power in all three studies (and thus a true effect), is 9.6%, which is slightly, but not substantially less likely.
- (C) All else being equal, the probability of a Type 1 error in one of three studies is 85.7% when there is no true effect in all three studies, and the probability of finding exactly 1 out of 3 significant effects, assuming 80% power in all three studies (and thus a true effect), is 0.8%, which is substantially less likely.
- (D) It is not possible to know the probability that you will observe a Type 1 error if you perform 3 studies.

The idea that most studies have 80% power is slightly optimistic. **Examine the correct answer to the previous question across a range of power values** (e.g., 50% power, and 30% power).

**Q10:** Several papers suggest it is a reasonable assumption that the power in the psychological literature might be around 50%. Set the number of studies to 4, the number of successes also to 4, and the assumed power slider to 50%, and look at the table at the bottom of the app. How likely is it that you will observe 4 significant results in 4 studies, assuming there is a true effect?

- (A) 6.25%
- (B) 12.5%
- (C) 25%
- (D) 37.5%

Imagine you perform 4 studies, and 3 show a significant result. **Change these numbers in the online app. Leave the power at 50%.** The output in the text tells you:

*When the observed results are equally likely under  $H_0$  and  $H_1$ , the likelihood ratio is 1. Benchmarks to interpret Likelihood Ratios suggest that when  $1 < LR < 8$  there is weak evidence, when  $8 < LR < 32$  there is moderate evidence, and when  $LR > 32$ , there is strong evidence.*

*The data are more likely under the alternative hypothesis than the null hypothesis with a likelihood ratio of 526.32.*

These calculations show that, assuming you have observed three significant results out of four studies, and assuming each study had 50% power, you are 526 times more likely to have observed these data when the alternative hypothesis is true, than when the null hypothesis is true. In other words, your are 526 times more likely to find a significant effect in three studies when you have 50% power, than to find three Type 1 errors in a set of four studies.

**Q11:** Maybe you don't think 50% power is a reasonable assumption. How low can the power be (rounded to 2 digits), for the likelihood to remain higher than 32 in favor of  $H_1$  when observing 3 out of 4 significant results?

- (A) 5% power
- (B) 17% power
- (C) 34% power
- (D) 44% power

The main take-home message of these calculations is to understand that 1) mixed results are supposed to happen, and 2) mixed results can contain strong evidence for a true effect, across a wide range of plausible power values. The app also tells you how much evidence, in a rough dichotomous way, you can expect. This is useful for our educational goal. But when you want to evaluate results from multiple studies, the formal way to do so is by performing a meta-analysis.

The above calculations make a very important assumption, namely that the Type 1 error rate is controlled at 5%. If you try out many different tests in each study, and only report the result that yielded  $p < 0.05$ , these calculations no longer hold.

**Q12:** Go back to the default settings of 2 out of 3 significant results, but now set the Type 1 error rate to 20%, to reflect a modest amount of  $p$ -hacking. Under these circumstances, what is the **highest** likelihood in favor of  $H_1$  you can get if you explore all possible values for the true power?

- (A) Approximately 1
- (B) Approximately 4.63
- (C) Approximately 6.70

- (D) Approximately 62.37

As the scenario above shows, *p*-hacking makes studies extremely uninformative. **If you inflate the error rate, you quickly destroy the evidence in the data.** You can no longer determine whether the data are more likely when there is no effect, than when there is an effect. Sometimes researchers complain that people who worry about *p*-hacking and try to promote better Type 1 error control are missing the point, and that other things (better measurement, better theory, etc.) are more important. I fully agree that these aspects of scientific research are at least as important as better error control. But better measures and theories will require decades of work. Better error control could be accomplished today, if researchers would stop inflating their error rates by flexibly analyzing their data. And as this assignment shows, inflated rates of false positives very quickly make it difficult to learn what is true from the data we collect. Because of the relative ease with which this part of scientific research can be improved, and because we can achieve this today (and not in a decade), I think it is worth stressing the importance of error control, and publish more realistic-looking sets of studies.

**Q13:** Some ‘prestigious’ journals (which, when examined in terms of scientific quality such as reproducibility, reporting standards, and policies concerning data and material sharing, are quite low-quality despite their prestige) only publish manuscripts with a large number of studies, which should all be statistically significant. If we assume an average power in psychology of 50%, only 3.125% of 5-study articles should contain exclusively significant results. If you pick up a random issue from such a prestigious journal, and see 10 articles, each reporting 5 studies, and all manuscripts have exclusively significant results, would you trust the reported findings more, or less, than when all these articles had reported mixed results? Why?

**Q14:** Unless you manage to power all your studies at 99.99% for the rest of your career (which would be slightly inefficient, but great if you don’t like insecurity about erroneous statistical claims), you will observe mixed results within any given line of research. How do you plan to deal with these mixed results?

### 3.4.3 Open Questions

1. What is the difference between a probability and a likelihood?
2. Why is it important to remember that a likelihood ratio is relative evidence?
3. If we compare 2 hypotheses,  $H_0$  and  $H_1$ , and the likelihood ratio of  $H_1$  compared to  $H_0$  is 77, what does this mean?
4. What are benchmarks for medium and strong evidence according to Royall (1997)?
5. How can it be the case that we have observed that a likelihood ratio of 200, but both hypotheses are incorrect?

6. If we perform multiple studies and find that only 2 out of 3 studies show a significant result, how can this actually be strong evidence for  $H_1$ ?

## 4 Bayesian statistics

“Logic!” said the Professor half to himself. “Why don’t they teach logic at these schools? There are only three possibilities. Either your sister is telling lies, or she is mad, or she is telling the truth. You know she doesn’t tell lies and it is obvious that she is not mad. For the moment then and unless any further evidence turns up, we must assume that she is telling the truth.”

*The Lion, The Witch, and The Wardrobe. A Story for Children* by C. S. Lewis.

In the children’s book *The Lion, The Witch, and The Wardrobe*, Lucy and Edmund go through a wardrobe into a country called Narnia. Lucy tells her older brother and sister, Peter and Susan, about Narnia, but Edmund wants to keep it a secret, and tells Peter and Susan he and Lucy were just pretending Narnia exists. Peter and Susan don’t know what to believe — does Narnia exist, or not? Is Lucy telling the truth, or is Edmund? Thinking about probabilities in the long run will not help much - this is a unique event, and we will need to think about the probability that Narnia exists, or not, based on the information we have available.

They ask the Professor, who lives in the house with the wardrobe, for advice. The Professor asks Susan and Peter if in their past experience, Lucy or Edward has been more truthful, to which Peter answers “Up till now, I’d have said Lucy every time.” So, they have a stronger prior belief Lucy is telling the truth, relative to Edward telling the truth. The Professor then replies with the quote above. From the three possible options, we don’t believe Lucy is lying, as she has not done so in the past, and the Professor believes it is clear just from talking to Lucy that she is not mad. Therefore, the most plausible option is that Lucy is telling the truth. If new evidence is uncovered, these beliefs can be updated in the future. This approach to knowledge generation, where the prior probability of different hypotheses is quantified, and if possible updated in light of new data, is an example of *Bayesian inference*.

Although frequentist statistics is by far the dominant approach in science, it is important to have had at least rudimentary exposure to Bayesian statistics during any statistics training. Bayesian statistics is especially useful when inferences are made in cases where the data under investigation are unique, and there is no frequentist probability, which is typically defined as the limit of a variable averaged over many trials. For example, the question might not be how often Lucy lies *on average*, but whether Lucy is lying *in this specific instance* about the existence of Narnia. When we do research, we often start with a prior belief that a hypothesis is true. After collecting data, we can use this data to update our prior beliefs. Bayesian statistics allows you to update prior beliefs into posterior probabilities in a logically consistent

manner. Before we have collected data, the **prior odds** of Hypothesis 1 ( $H_1$ ) over the null-hypothesis ( $H_0$ ) are  $P(H_1)/P(H_0)$ . After we have collected data, we have the **posterior odds**  $P(H_1|D)/P(H_0|D)$ , which you can read as the probability of  $H_1$ , given the data, divided by the probability of  $H_0$ , given the data. There are different approaches to Bayesian statistics. We will first discuss Bayes factors, and then Bayesian estimation.

## 4.1 Bayes factors

One approach in Bayesian statistics focuses on the comparison of different models that might explain the data. In this model comparison approach to Bayesian statistics the probability of data under a specified model ( $P(D|H_0)$ ) is a number that expresses what is sometimes referred to as the absolute **evidence**, and more formally referred to as a marginal likelihood. The marginal likelihood uses prior probabilities to average the likelihood across the parameter space. For example, assume we have a simple model  $M$  that is based on a single parameter, that can take on two values,  $X$  and  $Y$ , and that *a priori* we believe the probability of both values is  $P(X) = 0.4$  and  $P(Y) = 0.6$ . We collect data, and calculate the likelihood for both these parameter values, which is  $P(D|X) = 0.02$  and  $P(D|Y) = 0.08$ . The marginal likelihood of our model  $M$  is then  $P(D|M) = 0.4 \times 0.02 + 0.6 \times 0.08 = 0.056$ . Most often, models have continuously varying parameters, and the marginal likelihood formula is based on an integral, but the idea remains the same.

A comparison of two models is based on the relative evidence that the data provides for each model. The relative evidence is calculated by dividing the marginal likelihood for one model by the marginal likelihood for the other, and this ratio of relative evidence based on these marginal likelihoods is called the **Bayes factor**. Bayes factors are the Bayesian equivalent of hypothesis tests (Dienes, 2008; Kass & Raftery, 1995). The Bayes factor represents how much we have updated our beliefs, based on observing the data. We can express Bayes factors to indicate how much more likely  $H_1$  has become given the data compared to  $H_0$  (often indicated by  $BF_{10}$ ) or as how much more likely  $H_0$  has become compared to  $H_1$  ( $BF_{01}$ ), and  $BF_{10} = 1/BF_{01}$ . Similar to a likelihood ratio of 1, a Bayes factor of 1 does not change our beliefs in favor of one model compared to the other model. A very large Bayes factor for  $H_1$  over  $H_0$  increases our belief in  $H_1$  relative to  $H_0$ , and a Bayes factor close to 0 increases our belief in  $H_0$  relative to  $H_1$ . If our prior belief in  $H_1$  was very, very low (e.g., your belief in unicorns) even a large Bayes Factor that supports the presence of a unicorn might not yet convince you that unicorns are real – but you have updated your belief in unicorns, and now believe they are at least more likely than they were before (even if you still think unicorns are very unlikely to exist). The contribution of the Bayes Factor and the prior in calculating the posterior odds is clear in the following formula:

$$\frac{P(H_1|D)}{P(H_0|D)} = \frac{P(D|H_1)}{P(D|H_0)} \times \frac{P(H_1)}{P(H_0)}$$

$$\text{Posterior Probability} = \text{Bayes Factor} \times \text{Prior Probability}$$

A Bayesian analysis of data requires specifying the prior. Here, we will continue our example based on a binomial probability, such as a coin flip. In the likelihood example, we compared two point hypotheses (e.g.,  $p = 0.5$  vs.  $p = 0.8$ ). In Bayesian statistics, parameters are considered to be random variables, and the uncertainty or degree of belief with respect to the parameters is quantified by **probability distributions**.

A binomial probability lies between 0 and 1. You could draw any probability density you want over 0 and 1, and turn it into a prior, but for good reasons (simplicity, mostly) a beta-prior is often used for binomial probabilities. The shape of the beta-prior depends on two parameters,  $\alpha$  and  $\beta$ . Note that these are the same Greek letters used for the Type 1 error rate and Type 2 error rate, but that is purely coincidental! The  $\alpha$  and  $\beta$  in binomial probabilities are unrelated to error rates, and the use of the same letters is mainly due to a lack of creativity among statisticians and the limited choice the alphabet gives us. It also does not help that the distribution of which  $\beta$  is one of the parameters is called the Beta distribution. Try to keep these different betas apart! The probability density function is:

$$f(x; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}$$

where  $B(\alpha, \beta)$  is the beta function. Understanding the mathematical basis of this function is beyond the scope of this chapter, but you can read more on [Wikipedia](#) or Kruschke's book on Doing Bayesian Data Analysis (Kruschke, 2014). The beta prior for a variety of values for  $\alpha$  and  $\beta$  can be seen in Figure 4.1.

These beta densities reflect different types of priors. Let's imagine that you are approached by a street merchant who tries to sell you a special coin with heads and tails that, when flipped, will almost always turn up heads. The  $\alpha = 1, \beta = 1$  prior is what a newborn baby would have as a prior, without any idea of what to expect when you flip a coin, and thus every value of  $p$  is equally likely. The  $\alpha = 1, \beta = 1/2$  prior is what a true believer would have as a prior. The sales merchant tells you the coin will turn up heads almost every time, and thus, you believe it will turn up heads almost every time. The  $\alpha = 4, \beta = 4$ , and the  $\alpha = 100, \beta = 100$  priors are for slightly and extremely skeptical people. With an  $\alpha = 4, \beta = 4$  prior, you expect the coin will be fair, but you are willing to believe a wide range of other true values is possible (the curve is centered on 0.5, but the curve is wide, allowing for very high and low values of  $p$ ). With the  $\alpha = 100, \beta = 100$  prior you are really convinced coins are fair, and believe there will be only a very slight bias, at most (the curve is again centered on 0.5, and a skeptic believes that  $p$  will lie between 0.4 and 0.6 – a much narrower range compared to the slightly skeptical individual).

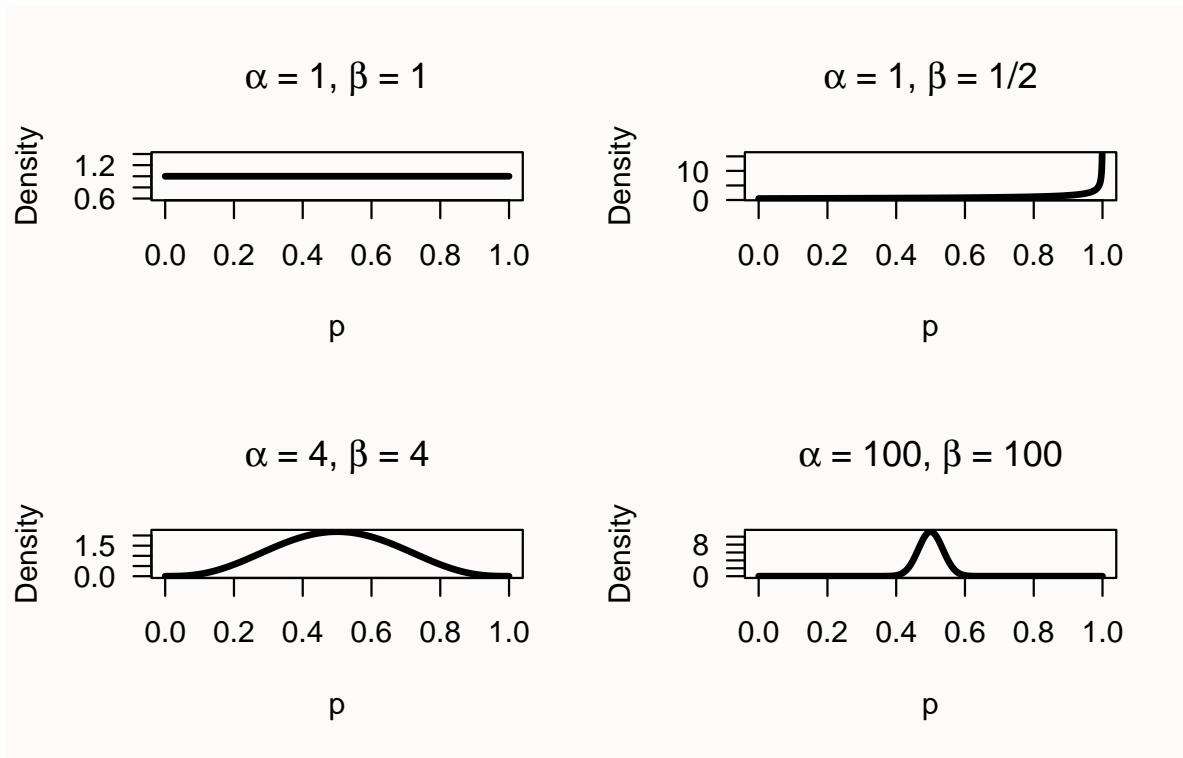


Figure 4.1: Four examples of Bayesian priors.

Let's assume the newborn baby, the true believer, the slight skeptic, and the extreme skeptic all buy the coin, flip it  $n = 20$  times, and observe  $x = 10$  heads. This outcome can be plotted as a binomial distribution with 10 heads out of 20 trials, or as a Beta(11, 11) distribution.

The newborn baby had a prior Beta distribution with  $\alpha = 1$  and  $\beta = 1$ , which equals a binomial likelihood distribution for 0 heads out of 0 trials. The posterior is a Beta distribution with  $\text{Beta}(\alpha^*, \beta^*)$ , where:

$$\alpha^* = \alpha + x = 1 + 10 = 11$$

$$\beta^* = \beta + n - x = 1 + 20 - 10 = 11$$

Or calculating these values more directly from the  $\alpha$  and  $\beta$  of the prior and likelihood:

$$\alpha^* = \alpha_{\text{prior}} + \alpha_{\text{likelihood}} - 1 = 1 + 11 - 1 = 11$$

$$\beta^* = \beta_{\text{prior}} + \beta_{\text{likelihood}} - 1 = 1 + 11 - 1 = 11$$

Thus, the posterior distribution for the newborn is a Beta(11,11) distribution. This equals a binomial likelihood function for 10 heads out of 20 trials, or Beta(11,11) distribution. In other words, the posterior distribution is identical to the likelihood function when a uniform prior is used.

Take a look at Figure 4.2. For the newborn baby, given 10 heads out of 20 coin flips, we see the prior distribution (the horizontal grey line), the likelihood (the blue dotted line), and the posterior (the black line).

For the true believer the posterior distribution is not centered on the maximum likelihood of the observed data, but just a bit in the direction of the prior. The slight skeptic and the skeptic end up with a much stronger belief in a fair coin after observing the data than the newborn and believer, but mainly because they already had a stronger prior that the coin was fair.

## 4.2 Updating our belief

Now that we have a distribution for the prior, and a distribution for the posterior, we can see in the graphs below for which values of  $p$  our belief has increased. Everywhere where the black line (of the posterior) is higher than the grey line (of the prior) our belief in that  $p$  has increased.

The Bayes Factor is used to quantify this increase in relative evidence. Let's calculate the Bayes Factor for the hypothesis that the coin is fair for the newborn. The Bayes Factor is simply the value of the posterior distribution at  $p = 0.5$ , divided by the value of the prior distribution at  $p = 0.5$ :

$$BF_{10} = \text{Beta}(p = 0.5, 11, 11) / \text{Beta}(p = 0.5, 1, 1) = 3.70 / 1 = 3.70$$

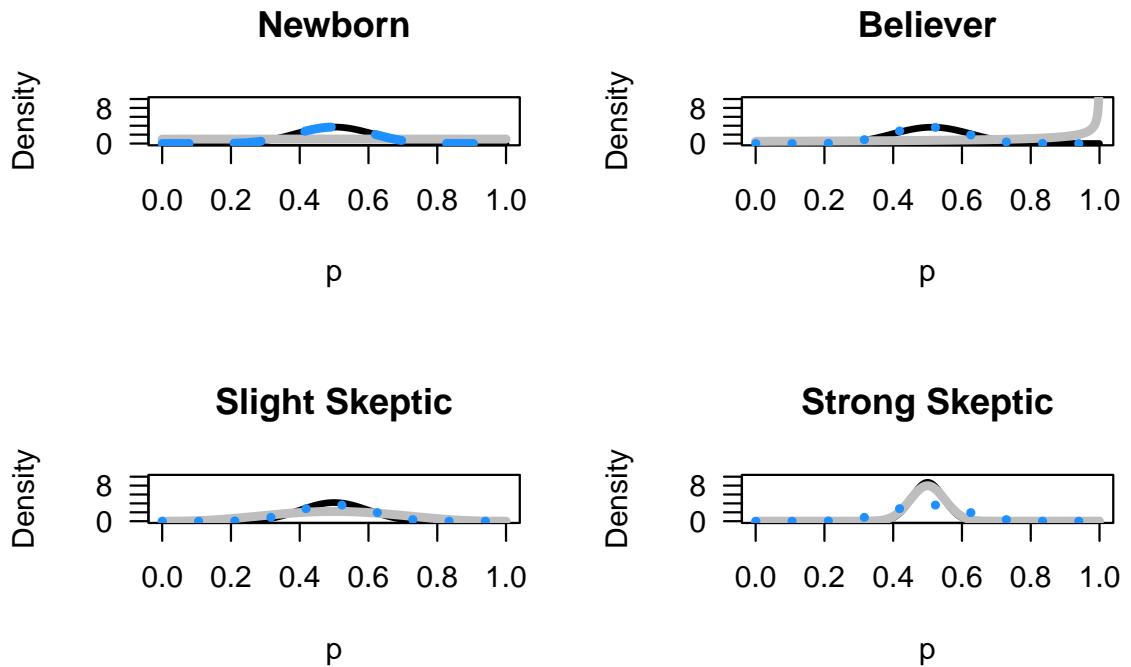


Figure 4.2: Four examples of how different priors are updated to the posterior based on data.

**Mean posterior: 0.5 , 95% Credible Interval: 0.3 ; 0.7**

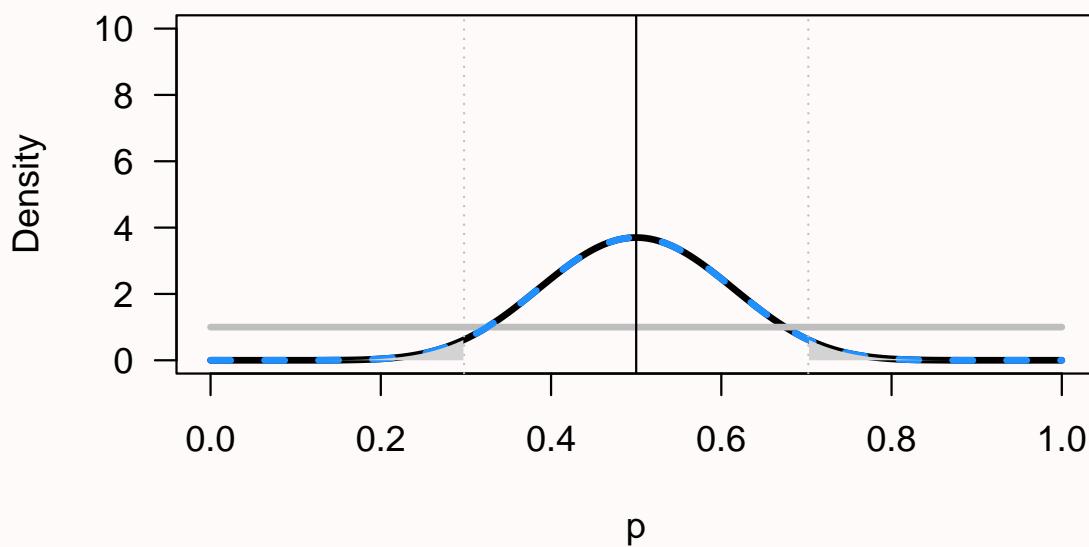


Figure 4.3: Plot for the prior, likelihood, and posterior.

We can calculate and plot the Bayes Factor, and show the prior (grey), likelihood (dashed blue) and posterior (black). For the example of 20 flips, 10 heads, and the newborn baby's prior, the plot looks like this:

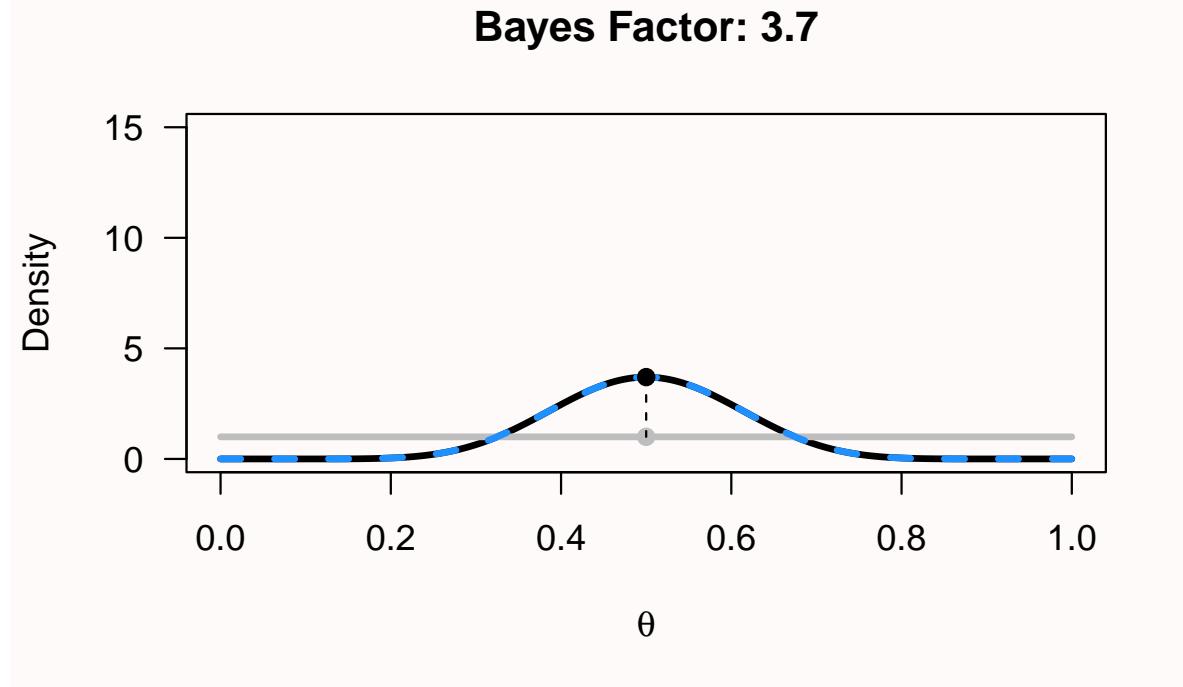


Figure 4.4: Plot for a uniform prior, the likelihood, and the posterior.

We see that for the newborn,  $p = 0.5$  has become more probable, but so has  $p = 0.4$ . Now let's examine the curves for the extreme skeptic, who believes the coin is fair with a prior of Beta(100, 100), buys the coin, and flips it 100 times. Surprisingly, the coin comes up heads 90 out of 100 flips. The plot of the prior, likelihood, and posterior now looks much more extreme, because we had a very informed prior, and extremely different data. We see the grey prior distribution, the dashed blue likelihood based on the data, and the posterior distribution in black. The Bayes factor of 0 (note that the value is rounded, and is extremely small, but not exactly zero) represents the substantial drop in belief that the coin is fair – indeed, this now seems an untenable hypothesis, even for the extreme skeptic. It shows how data can update your belief. Where a newborn would now completely believe that the true  $p$  for the coin is somewhere around 0.9, the extreme skeptic has more reason to believe the  $p$  is around 0.65, due to the strong prior conviction that the coin is fair. Given enough data, even this extreme skeptic will become convinced that the coin will return heads most of the time as well.

We can now also see the difference between a likelihood inference approach and a Bayesian inference approach. In likelihood inference, you can compare different values of  $p$  for the same

## Bayes Factor: 0

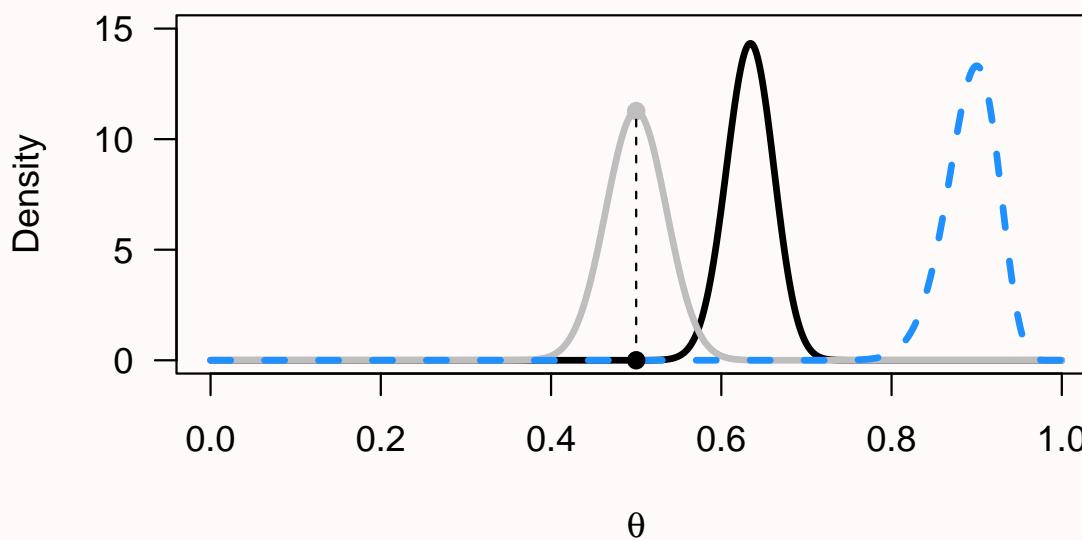


Figure 4.5: Plot for a strongly informed prior, the likelihood, and the posterior.

likelihood curve (e.g.,  $p = 0.5$  vs  $p = 0.8$ ) and calculate the likelihood ratio. In Bayesian inference, you can compare the difference between the prior and the posterior for the same value of  $p$ , and calculate the Bayes Factor.

If you have never seen Bayes Factors before, you might find it difficult to interpret the numbers. As with any guideline (e.g., interpreting effect sizes as small, medium, and large) there is criticism on the use of benchmarks. On the other hand, you have to start somewhere in getting a feel for what Bayes Factors mean. A Bayes factor between 1 and 3 is considered ‘not worth more than a bare mention’, larger than 3 (or smaller than 1/3) is considered ‘substantial’, and larger than 10 (or smaller than 1/10) is considered ‘strong’ (Jeffreys, 1939). These labels refer to the increase in how much you believe a specific hypothesis, not in the posterior belief in that hypothesis. If you think extra-sensory perception is extremely implausible, a single study with a  $\text{BF} = 14$  will increase your belief, but only to the point where you think extra-sensory perception is “pretty much extremely implausible”.

Bayes factors are often promoted as an alternative to  $p$ -values. One stated benefit is that they can provide support both for the alternative and the null (Dienes, 2014). However, the same can be achieved with frequentist equivalence tests, as we will see in the chapter on [equivalence testing](#), and inferences based on Bayes factors and equivalence tests typically lead to the same conclusions (Daniël Lakens et al., 2020). Another reason that some people give to switch to Bayes factors instead of  $p$ -values is that, as we saw in Chapter 1 on [p-values](#),  $p$ -values are often misunderstood. However, not surprisingly, Bayes factors are at least as often misunderstood and misused (Tendeiro et al., 2024; Wong et al., 2022). Statistical inferences are hard, and thinking about probabilities is not something we get right by trusting our intuition. We need to train ourselves to draw correct inferences, and switching to a different approach to statistics will not prevent misuse.

## 4.3 Preventing common misconceptions about Bayes Factors

As more people have started to use Bayes Factors, we should not be surprised that misconceptions about Bayes Factors have become common. A recent study shows that the percentage of scientific articles that draw incorrect inferences based on observed Bayes Factors is distressingly high (Tendeiro et al., 2024; Wong et al., 2022), with 92% of articles demonstrating at least one misconception about Bayes Factors.

### 4.3.1 Misunderstanding 1: Confusing Bayes Factors with Posterior Odds.

One common criticism by Bayesians of null hypothesis significance testing (NHST) is that NHST quantifies the probability of the data (or more extreme data), given that the null hypothesis is true, but that scientists should be interested in the probability that the hypothesis is true, given the data. Cohen (1994) wrote:

What's wrong with NHST? Well, among many other things, it does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does! What we want to know is “Given these data, what is the probability that  $H_O$  is true?”

One might therefore believe that Bayes factors tell us something about the probability that a hypothesis true, but this is incorrect. A Bayes factor merely quantifies how much we should update our belief in a hypothesis. If this hypothesis was extremely unlikely (e.g., the probability that people have telepathy) then we might still believe it to be very unlikely, even after computing a large Bayes factor in a single study demonstrating telepathy. If we believed the hypothesis that people have telepathy was unlikely to be true (e.g., we thought it was 99.9% certain telepathy was not true), then evidence for telepathy might only increase our belief in telepathy to the extent that we now believe it is 98% unlikely. The Bayes factor only corresponds to our posterior belief if we were initially perfectly uncertain about the hypothesis being true or not. If both hypotheses were equally likely, and a Bayes factor indicates we should update our belief in such a way that the alternative hypothesis is three times more likely than the null hypothesis, only then would we end up believing the alternative hypothesis is exactly three times more likely than the null hypothesis. One should therefore not conclude that, for example, given a BF of 10, the alternative hypothesis is more likely to be true than the null hypothesis. The correct claim is that people should update their belief in the alternative hypothesis by a factor of 10.

#### **4.3.2 Misunderstanding 2: Failing to interpret Bayes Factors as relative evidence.**

One benefit of Bayes factors that is often mentioned by Bayesians is that, unlike NHST, Bayes factors can provide support for the null hypothesis, and thereby falsify predictions. It is true that NHST can only reject the null hypothesis (that is, it can never accept the null hypothesis), although it is important to add that in frequentist statistics [equivalence tests](#) can be used to reject the alternative hypothesis, and therefore there is no need to switch to Bayes factors to meaningfully interpret the results of non-significant null hypothesis tests.

Bayes factors quantify support for one hypothesis relative to another hypothesis. As with likelihood ratios (and as illustrated in Figure 3.7), it is possible that one hypothesis is supported more than another hypothesis, while both hypotheses are actually false. It is incorrect to interpret Bayes factors in an absolute manner, for example by stating that a Bayes factor of 0.09 provides support for the null hypothesis. The correct interpretation is that the Bayes factor provides relative support for  $H_0$  compared to H1. With a different alternative model, the Bayes factor would change. As with a significant equivalence test, even a Bayes factor strongly supporting  $H_0$  does not mean there is no effect at all; there could be a true, but small, effect.

For example, after Daryl Bem (2011) published 9 studies demonstrating support for pre-cognition (that is, conscious cognitive awareness of a future event that could not otherwise be

known) a team of Bayesian statisticians re-analyzed the studies, and concluded “Out of the 10 critical tests, only one yields “substantial” evidence for  $H_1$ , whereas three yield “substantial” evidence in favor of  $H_0$ . The results of the remaining six tests provide evidence that is only “anecdotal”” (Wagenmakers et al., 2011). In a reply, Bem and Utts (2011) reply by arguing that the set of studies provide convincing evidence for the alternative hypothesis, if the Bayes factors are computed as relative evidence between the null hypothesis and a more realistically specified alternative hypothesis, where the effects of pre-cognition are expected to be small. This back-and-forth illustrates how Bayes factors are relative evidence, and a change in the alternative model specification changes whether the null or the alternative hypothesis receives relatively more support given the data.

#### **4.3.3 Misunderstanding 3: Not specifying the null and/or alternative model.**

Given that Bayes factors are relative evidence for or against one model compared to another model, it might be surprising that many researchers fail to specify the alternative model to begin with when reporting their analysis. And yet, in a systematic review of how psychologists use Bayes factors, van de Schoot et al. (2017) found that “31.1% of the articles did not even discuss the priors implemented”. Whereas in a null hypothesis significance test researchers do not need to specify the model that the test is based on, as the test is by definition a test against an effect of 0, and the alternative model consists of any non-zero effect size (in a two-sided test), this is not true when computing Bayes factors. The null model when computing Bayes factors is often (but not necessarily) a point null as in NHST, but the alternative model is typically only one of many possible alternative hypotheses that a researcher could test against. It has become common to use ‘default’ priors, but as with any heuristic, defaults will most often give an answer to a nonsensical question, and quickly become a form of mindless statistics. When introducing Bayes factors as an alternative to frequentist  $t$ -tests, Rouder et al. (2009) write:

This commitment to specify judicious and reasoned alternatives places a burden on the analyst. We have provided default settings appropriate to generic situations. Nonetheless, these recommendations are just that and should not be used blindly. Moreover, analysts can and should consider their goals and expectations when specifying priors. Simply put, principled inference is a thoughtful process that cannot be performed by rigid adherence to defaults.

The priors used when computing a Bayes factor should therefore be both specified and justified.

#### **4.3.4 Misunderstanding 4: Claims based on Bayes Factors do not require error control.**

In a paper with the provocative title “Optional stopping: No problem for Bayesians” Rouder (2014) argues that “Researchers using Bayesian methods may employ optional stopping in

their own research and may provide Bayesian analysis of secondary data regardless of the employed stopping rule.” A reader who merely read the title and abstract of that paper might come to the conclusion that Bayes factors are a wonderful solution to the error inflation due to optional stopping in the frequentist framework, but this is not correct (de Heide & Grünwald, 2017).

There is a big caveat about the type of statistical inferences that is unaffected by optional stopping. Optional stopping is no problem for Bayesians only if they refrain from a) making a dichotomous claim about the presence or absence of an effect, or b) when they refrain from drawing conclusions about a prediction being supported or falsified. Rouder notes how “Even with optional stopping, a researcher can interpret the posterior odds as updated beliefs about hypotheses in light of data.” In other words, even after optional stopping, a Bayes factor tells researchers how much they should update their belief in a hypothesis. Importantly, when researchers make dichotomous claims based on Bayes factors (e.g., “The effect did not differ significantly between the condition,  $BF_{10} = 0.17$ ”) then this claim can be either correct or an error, so that error rates become a relevant consideration, unlike when researchers simply present the Bayes factor for readers to update their personal beliefs.

Bayesians disagree among each other about whether Bayes factors should be the basis of dichotomous claims, or not. Those who promote the use of Bayes factors to make claims often refer to thresholds proposed by Jeffreys (1939), where a  $BF > 3$  is “substantial evidence”, and a  $BF > 10$  is considered “strong evidence”. Some journals, such as *Nature Human Behavior*, have the following requirement for researchers who submit a Registered Report (a novel article publication format where a preregistration is peer reviewed before the data is analyzed, and authors receive a decision about whether their article will be published before the results are known): “For inference by Bayes factors, authors must be able to guarantee data collection until the Bayes factor is at least 10 times in favour of the experimental hypothesis over the null hypothesis (or vice versa).” When researchers decide to collect data until a specific threshold is crossed to make a claim about a test, their claim can be correct or wrong, just as when  $p$ -values are the statistical basis for a claim. As both the Bayes factor and the  $p$ -value can be computed based on the sample size and the  $t$ -value (Francis, 2016; Rouder et al., 2009), there is nothing special about using Bayes factors as the basis of an ordinal claim. The exact long-run error rates can not be directly controlled when computing Bayes factors, and the Type 1 and Type 2 error rate depends on the choice of the prior and the choice for the cut-off used to decide to make a claim. Simulations studies show that for commonly used priors and a  $BF > 3$  cut-off to make claims, the Type 1 error rate is somewhat smaller, but the Type 2 error rate is considerably larger (Kelter, 2021).

In summary, whenever researchers make claims, they can make erroneous claims, and error control should be a worthy goal. Error control is not a consideration when researchers do not make ordinal claims (e.g., X is larger than Y, there is a non-zero correlation between X and Y, etc.). If Bayes factors are used to quantify how much researchers should update personal beliefs in a hypothesis, there is no need to consider error control, but, as a corollary, researchers should also refrain from making any ordinal claims based on Bayes factors in the

Results section or Discussion sections of their paper. Giving up error control also means giving up claims dichotomous claims about the presence or absence of effects.

#### 4.3.5 Misunderstanding 5: Interpreting Bayes Factors as effect sizes.

Bayes factors are not statements about the size of an effect. It is therefore not appropriate to conclude that the effect size is small or large purely based on the Bayes factor. Depending on the priors used when specifying the alternative and null model, the same Bayes factor can be observed for very different effect size estimates. The reverse is also true: The same effect size can correspond to Bayes factors supporting the null or the alternative hypothesis, depending on how the null and alternative model are specified. Researchers should therefore always report and interpret effect size measures separately from their test statistics. Statements about the size of effects should only be based on these effect size measures, and not on Bayes factors.

### 4.4 Bayesian Estimation

The posterior distribution summarizes our belief about the expected number of heads when flipping a coin after seeing the data, by averaging over our prior beliefs and the data (or the likelihood). The mean of a Beta distribution can be calculated by  $\alpha/(\alpha+\beta)$ . We can thus easily calculate the mean of a posterior distribution, which is the expected value based on our prior beliefs and the data.

We can also calculate a **credible interval** around the mean, which is a Bayesian version of a **confidence interval** with a slightly different interpretation. Instead of the frequentist interpretation where a parameter has one (unknown) true value, the Bayesian approach considers the data fixed, but allow the parameter to vary. In Bayesian approaches, probability distributions represent our degree of belief. When calculating a credible interval, one is saying ‘I believe it is 95% probable (given my prior and the data) that the true parameter falls within this credible interval’. A 95% credible interval is simply the area of the posterior distribution between the 0.025 and 0.975 quantiles.

A credible interval and a confidence interval are the same, when a uniform prior such as Beta(1,1) is used. In this case, credible interval is numerically identical to the confidence interval. For an example, see Figure 4.6 where the mean and 95% credible interval are plotted for the posterior when 10 heads out of 20 coin flips are observed, given a uniform prior. In this example, the credible interval is identical to the confidence interval. Only the interpretation differs. Whenever an informed prior is used, the credible interval and confidence interval differ. If the chosen prior is not representative of the truth, the credible interval will not be representative of the truth, but it is always a correct formalization of your beliefs. As will be explained in more detail in Chapter 7, for a **single confidence interval**, the probability that it contains the true population parameter is either 0 or 1. Only in the long run will 95% of

confidence intervals contain the true population parameter. These are important differences between Bayesian credible intervals and frequentist confidence intervals to keep in mind.

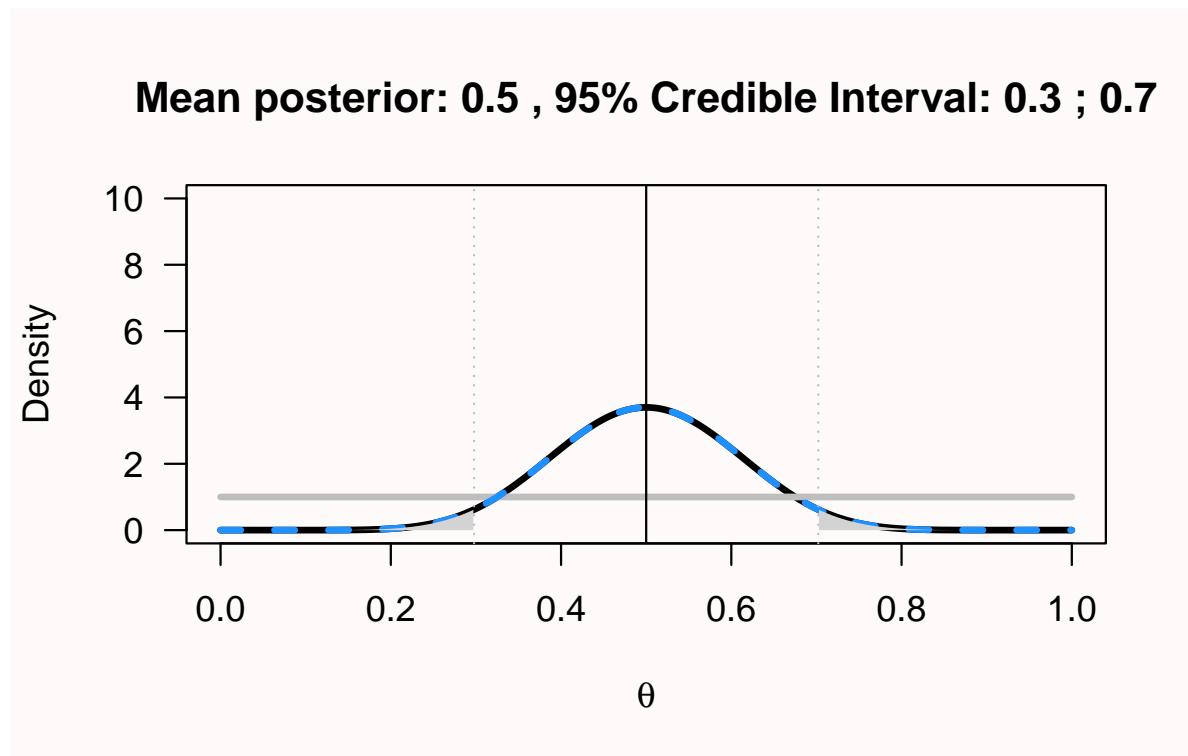


Figure 4.6: Plot for the mean of the posterior when 10 out of 20 heads are observed given a uniform prior.

We can also use the ‘binom’ package in R to calculate the posterior mean, credible interval, and **highest density interval (HDI)**. The highest density interval is an alternative to the credible interval that works better when the posterior beta distribution is skewed (and is identical when the posterior distribution is symmetrical). We won’t go into the calculations of the HDI here.

```
library(binom)

n <- 20 # set total trials
x <- 10 # set successes
aprior <- 1 # Set the alpha for the Beta distribution for the prior
bprior <- 1 # Set the beta for the Beta distribution for the prior

binom.bayes(x, n, type = "central", prior.shape1 = aprior, prior.shape2 = bprior)
binom.bayes(x, n, type = "highest", prior.shape1 = aprior, prior.shape2 = bprior)
```

method	x	n	shape1	shape2	mean	lower	upper	sig
bayes	10	20	11	11	0.5	0.2978068	0.7021932	0.05

method	x	n	shape1	shape2	mean	lower	upper	sig
bayes	10	20	11	11	0.5	0.2978068	0.7021932	0.05

The posterior mean is identical to the Frequentist mean, but this is only the case when the mean of the prior equals the mean of the likelihood (Albers et al., 2018). In your research, you will most likely need other calculations than the binomial example we have used here, and a lot of Bayesian tests are now available in the free open source software package [JASP](#). The math and the priors become more complex, but the basic idea remains the same. You can use Bayesian statistics to quantify relative evidence, which can inform you how much you should believe, or update your beliefs, in theories.

This chapter showed the essence of Bayesian inference, where we decide upon a prior distribution, collect data and calculate a marginal likelihood, and use these to calculate a posterior distribution. From this posterior distribution, we can estimate the mean and the 95% credible interval. For any specific hypothesis, we can calculate the relative evidence for a posterior model, compared to a prior model, through the Bayes Factor. There are many different flavors of Bayesian statistics. This means there are disagreements among Bayesians themselves about what the best approach to statistical inferences is, which are at least as vehement as the disagreements between frequentists and Bayesians. For example, many Bayesians dislike Bayes factors (McElreath, 2016). Some Bayesians dislike subjective priors as used in **subjective Bayesian analysis**, and instead prefer what is known as **objective Bayesian analysis** (Berger & Bayarri, 2004). Teaching material on Bayesian statistics will often present it as superior to frequentist statistics. For a more balanced educational lecture on Bayesian vs. frequentist statistics that more honestly highlights the strengths and weaknesses of both approaches, see the first 50 minutes of [this lecture by Michael I. Jordan](#).

## 4.5 Test Yourself

**Q1:** The true believer had a prior of Beta(1,0.5). After observing 10 heads out of 20 coin flips, what is the posterior distribution, given that  $\alpha = \alpha + x$  and  $\beta = \beta + n - x$ ?

- (A) Beta(10, 10)
- (B) Beta(11, 10.5)

- (C) Beta(10, 20)
- (D) Beta(11, 20.5)

**Q2:** The extreme skeptic had a prior of Beta(100,100). After observing 50 heads out of 100 coin flips, what is the posterior distribution, given that  $\alpha = \alpha + x$  and  $\beta = \beta + n - x$ ?

- (A) Beta(50, 50)
- (B) Beta(51, 51)
- (C) Beta(150, 150)
- (D) Beta(11, 20.5)

Copy the R script below into R. This script requires 5 input parameters (identical to the Bayes Factor calculator website used above). These are the hypothesis you want to examine (e.g., when evaluating whether a coin is fair,  $p = 0.5$ ), the total number of trials (e.g., 20 flips), the number of successes (e.g., 10 heads), and the  $\alpha$  and  $\beta$  values for the Beta distribution for the prior (e.g.,  $\alpha = 1$  and  $\beta = 1$  for a uniform prior). Run the script. It will calculate the Bayes Factor, and plot the prior (grey), likelihood (dashed blue), and posterior (black).

```
H0 <- 0.5 # Set the point null hypothesis you want to calculate the Bayes Factor for
n <- 20 # set total trials
x <- 10 # set successes
aprior <- 1 # Set the alpha for the Beta distribution for the prior
bprior <- 1 # Set the beta for the Beta distribution for the prior

alikelihood <- x + 1 # Calculate the alpha for the Beta distribution for the likelihood
blikelihood <- n - x + 1 # Calculate the beta for the Beta distribution for the likelihood
aposterior <- aprior + alikelihood - 1 # Calculate the alpha for the Beta distribution for the posterior
bposterior <- bprior + blikelihood - 1 # Calculate the beta for the Beta distribution for the posterior

theta <- seq(0, 1, 0.001) # create probability range p from 0 to 1
prior <- dbeta(theta, aprior, bprior)
likelihood <- dbeta(theta, alikelihood, blikelihood)
posterior <- dbeta(theta, aposterior, bposterior)

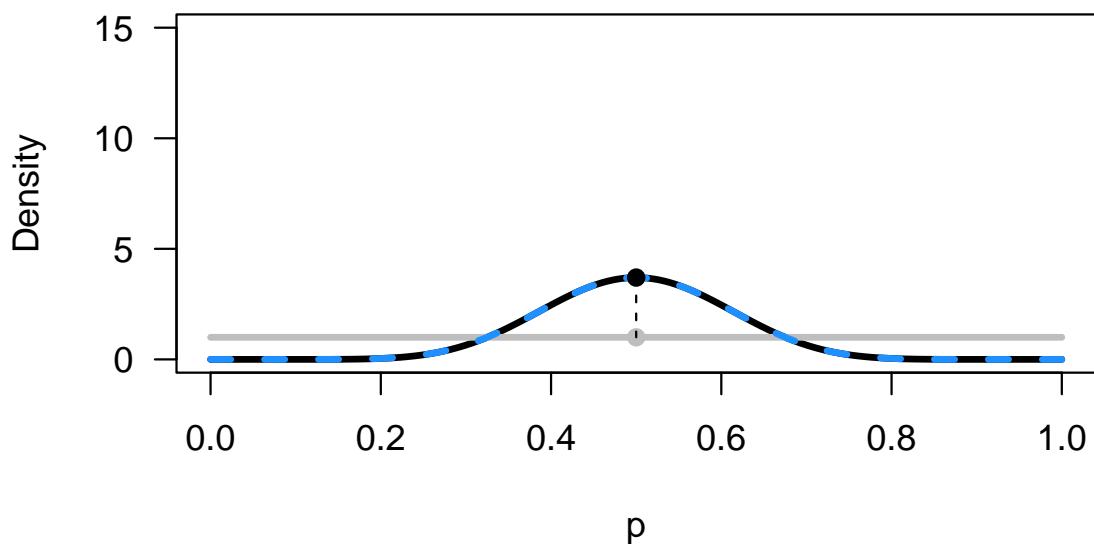
# Create plot
plot(theta, posterior, ylim = c(0, 15), type = "l", lwd = 3, xlab = "p", ylab = "Density", lty = 1)
lines(theta, prior, col = "grey", lwd = 3)
```

```

lines(theta, likelihood, lty = 2, lwd = 3, col = "dodgerblue")
BF10 <- dbeta(H0, aposterior, bposterior) / dbeta(H0, aprior, bprior)
points(H0, dbeta(H0, aposterior, bposterior), pch = 19)
points(H0, dbeta(H0, aprior, bprior), pch = 19, col = "grey")
segments(H0, dbeta(H0, aposterior, bposterior), H0, dbeta(H0, aprior, bprior), lty = 2)
title(paste("Bayes Factor:", round(BF10, digits = 2)))

```

## Bayes Factor: 3.7



We see that for the newborn baby,  $p = 0.5$  has become more probable, but so has  $p = 0.4$ .

**Q3:** Change the hypothesis in the first line from 0.5 to 0.675, and run the script. If you were testing the idea that this coin returns 67.5% heads, which statement is true?

- (A) Your belief in this hypothesis, given the data, would have decreased.
- (B) Your belief in this hypothesis, given the data, would have stayed the same.
- (C) Your belief in this hypothesis, given the data, would have increased.

**Q4:** Change the hypothesis in the first line back to 0.5. Let's look at the increase in the belief of the hypothesis  $p = 0.5$  for the extreme skeptic after 10 heads out of 20 coin flips. Change the  $\alpha$  for the prior in line 4 to 100 and the  $\beta$  for the prior in line 5 to 100. Run the script. Compare the figure from R to the increase in belief for the newborn baby. Which statement is true?

- (A) The belief in the hypothesis that  $p = 0.5$ , given the data, has **increased** for the extreme skeptic, but **not** as much as it has for the newborn.
- (B) The belief in the hypothesis that  $p = 0.5$ , given the data, has **increased** for the extreme skeptic, **exactly as much** as it has for the newborn.
- (C) The belief in the hypothesis that  $p = 0.5$ , given the data, has **increased** for the extreme skeptic, and **much more** than it has for the newborn.
- (D) The belief in the hypothesis that  $p = 0.5$ , given the data, has **decreased** for the extreme skeptic.

Copy the R script below and run it. The script will plot the mean for the posterior when 10 heads out of 20 coin flips are observed, given a uniform prior (as in Figure 4.6). The script will also use the ‘binom’ package to calculate the posterior mean, credible interval, and highest density interval is an alternative to the credible interval.

```
n <- 20 # set total trials
x <- 10 # set successes
aprior <- 1 # Set the alpha for the Beta distribution for the prior
bprior <- 1 # Set the beta for the Beta distribution for the prior

ymax <- 10 # set max y-axis

alikelihood <- x + 1 # Calculate the alpha for the Beta distribution for the likelihood
blikelihood <- n - x + 1 # Calculate the beta for the Beta distribution for the likelihood
aposterior <- aprior + alikelihood - 1 # Calculate the alpha for the Beta distribution for the posterior
bposterior <- bprior + blikelihood - 1 # Calculate the beta for the Beta distribution for the posterior

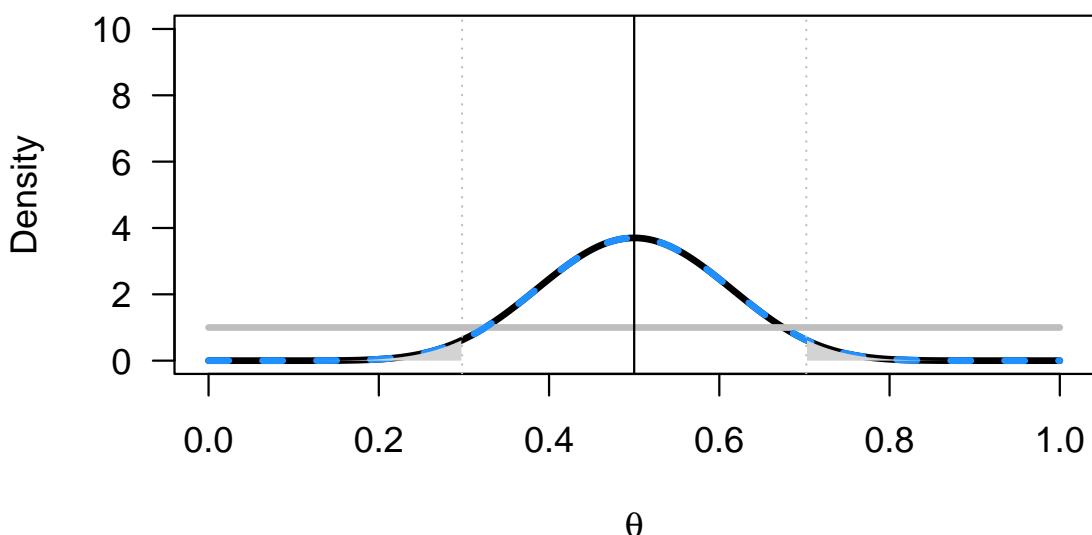
theta <- seq(0, 1, 0.001) # create probability range p from 0 to 1
prior <- dbeta(theta, aprior, bprior) # determine prior distribution
likelihood <- dbeta(theta, alikelihood, blikelihood) # determine likelihood distribution
posterior <- dbeta(theta, aposterior, bposterior) # determine posterior distribution
plot(theta, posterior, ylim = c(0, ymax), type = "l", lwd = 3, xlab = bquote(theta), ylab = "Posterior Probability")
lines(theta, prior, col = "grey", lwd = 3) # draw prior distribution
lines(theta, likelihood, lty = 2, lwd = 3, col = "dodgerblue") # draw likelihood distribution
```

```

LL <- qbeta(.025, aposterior, bposterior) # calculate lower limit credible interval
UL <- qbeta(.975, aposterior, bposterior) # calculate upper limit credible interval
abline(v = aposterior / (aposterior + bposterior)) # draw line mean
abline(v = LL, col = "grey", lty = 3) # draw line lower limit
abline(v = UL, col = "grey", lty = 3) # draw line upper limit
polygon(c(theta[theta < LL], rev(theta[theta < LL])), c(posterior[theta < LL], rep(0, sum(theta[theta < LL]))), col = "grey")
polygon(c(theta[theta > UL], rev(theta[theta > UL])), c(posterior[theta > UL], rep(0, sum(theta[theta > UL]))), col = "grey")
title(paste("Mean posterior: ", round((aposterior / (aposterior + bposterior)), digits = 5),

```

**Mean posterior: 0.5, 95% Credible Interval: 0.3;0.7**



```

if (!require(binom)) {
  install.packages("binom")
}
library(binom)
binom.bayes(x, n, type = "central", prior.shape1 = aprior, prior.shape2 = bprior)
binom.bayes(x, n, type = "highest", prior.shape1 = aprior, prior.shape2 = bprior)

```

method	x	n	shape1	shape2	mean	lower	upper	sig
bayes	10	20	11	11	0.5	0.2978068	0.7021932	0.05

method	x	n	shape1	shape2	mean	lower	upper	sig
bayes	10	20	11	11	0.5	0.2978068	0.7021932	0.05

The posterior mean is identical to the Frequentist mean, but this is only the case when the mean of the prior equals the mean of the likelihood.

**Q5:** Assume the outcome of 20 coin flips had been 18 heads. Change x to 18 in line 2 and run the script. Remember that the mean of the prior Beta(1,1) distribution is  $\alpha / (\alpha + \beta)$ , or  $1/(1+1) = 0.5$ . The Frequentist mean is simply  $x/n$ , or  $18/20=0.9$ . Which statement is true?

- (A) The frequentist mean is **higher** than the mean of the posterior, because by combining the prior with the data, the mean of the posterior is **closer** to the mean of the prior distribution.
- (B) The frequentist mean is **lower** than the mean of the posterior, because by combining the prior with the data, the mean of the posterior is **closer** to the mean of the prior distribution.
- (C) The frequentist mean is **higher** than the mean of the posterior, because by combining the prior with the data, the mean of the posterior is **further from** the mean of the prior distribution.
- (D) The frequentist mean is **lower** than the mean of the posterior, because by combining the prior with the data, the mean of the posterior is **further from** the mean of the prior distribution.

**Q6:** What is, today, your best estimate of the probability that the sun will rise tomorrow? Assume you were born with an uniform Beta(1,1) prior. The sun can either rise, or not. Assume you have seen the sun rise every day since you were born, which means there has been a continuous string of successes for every day you have been alive. It is OK to estimate the days you have been alive by just multiplying your age by 365 days. What is your best estimate of the probability that the sun will rise tomorrow?

**Q7:** What would have been the best estimate of the probability from Q6 from a frequentist perspective?

---

**Q8:** What do you think the goal of science is? Rozeboom (1960) has criticized Neyman-Pearson hypothesis testing by stating:

But the primary aim of a scientific experiment is not to precipitate decisions, but to make an appropriate adjustment in the degree to which one accepts, or believes, the hypothesis or hypotheses being tested”.

Frick (1996) has argued against Rozeboom, by stating:

Rozeboom (1960) suggested that scientists should not be making decisions about claims, they should be calculating and updating the probability of these claims. However, this does not seem practical. If there were only a handful of potential claims in any given area of psychology, it would be feasible to assign them probabilities, to be constantly updating the probabilities, and to expect experimenters to keep track of these ever-changing probabilities. In fact, just the number of claims in psychology is overwhelming. It would probably be impossible for human beings to keep track of the probability for each claim, especially if these probabilities were constantly changing. In any case, scientists do not assign probabilities to claims. Instead, scientists act like the goal of science is to collect a corpus of claims that are considered to be established (Giere, 1972).

When it comes to philosophy of science, there are no right or wrong answers. Reflect in 250 words on your thoughts about the two goals of science outlines by Rozeboom and Frick, and how these relate to your philosophy of science.

#### 4.5.1 Open Questions

1. What is a Bayes factor?
2. What is the difference between a Bayes factor and a likelihood ratio?
3. What does a Bayes factor of 1 mean?
4. What is the prior in Bayesian inference, and is it possible that different people have different priors?
5. Give a definition of a credible interval.
6. What is the difference between a frequentist confidence interval and a Bayesian credible interval?

7. What is the difference between a uniform and an informed prior when we compute the posterior distribution?
8. When computing a Bayes factor to, for example, analyze the mean difference between two independent groups, why is it incorrect to write “The Bayes factor of 0.2 indicated that there was no effect”?
9. When computing a Bayes factor to, for example, analyze the mean difference between two independent groups, why is it incorrect to write “The Bayes factor of 8 indicated that the alternative hypothesis was more likely than the null hypothesis”?

# 5 Asking Statistical Questions

Beware of the man of one method or one instrument, either experimental or theoretical. He tends to become method-oriented rather than problem-oriented. The method-oriented man is shackled; the problem-oriented man is at least reaching freely toward what is most important. *Platt, (1964), Strong Inference.*

At the core of the design of a new study is the evaluation of its **information quality**: the potential of a particular dataset for achieving a given analysis goal by employing data analysis methods and considering a given utility (Kenett et al., 2016). The goal of data collection is to gain information through **empirical research** where observations are collected and analyzed, often through statistical models. Three approaches to statistical modelling can be distinguished Shmueli (2010): Description, explanation, and prediction, which are discussed below. The utility often depends on which effects are deemed interesting. A thorough evaluation of the information quality of a study therefore depends on clearly specifying the goal of data collection, the statistical modelling approach that is chosen, and the usefulness of the data to draw conclusions about effects of interest with the chosen analysis method. A study with low information quality might not be worth performing, as the data that will be collected has low potential to achieve the analysis goal.

## 5.1 Description

Description aims to answer questions about features of the empirical manifestation of some phenomenon. Description can involve unique events (e.g., case studies of single patients) and classes of events (e.g., patients with a certain disease). Examples of features of interest are duration (how long), quantity (how many), location (where), etc.

An example of a descriptive question is research by [Kinsey](#), who studied the sexual behavior and experiences of Americans in a time that very little scientific research was available on this topic. He used interviews that provided the statistical basis to draw conclusions about sexuality in the United States, which, at the time, challenged conventional beliefs about sexuality.

Descriptive research questions are answered through **estimation statistics**. The informational value of an estimation study is determined by the amount of observations (the more observations, the higher the **precision** of the estimates) and the sampling plan (the more representative the sample, the lower the **sample selection bias**, which increases the ability to generalize from the sample to the population), and the reliability of the measure. It is

also important to create reliable and valid measures. For a free open educational resource on psychological measurement, see “[Introduction to Educational and Psychological Measurement Using R](#)”.

Descriptive research questions are sometimes seen as less exciting than explanation or prediction questions (Gerring, 2012), but they are an essential building block of theory formation (Scheel, Tiokhin, et al., 2021). Although estimation question often focus on the mean score of a measure, accurate estimates of the variance of a measure are extremely valuable as well. The variance of a measure is essential information in a well-informed sample size justification, both when planning for accuracy, as when performing an a-priori power analysis.

## 5.2 Prediction

The goal in predictive modeling is to apply an algorithm or a statistical model to predict future observations (Shmueli, 2010). For example, during the COVID-19 pandemic a large number of models were created that combined variables to estimate the risk that people would be infected with COVID, or that people who were infected would experience negative effects on their health (Wynants et al., 2020). Ideally, the goal is to develop a prediction model that accurately captures the regularities in its training data, and that generalizes well to unseen data. There is a **bias-variance trade off** between these two goals, and researchers need to decide how much bias should be reduced which increases the variance, or vice-versa (Yarkoni & Westfall, 2017). The goal in prediction is to minimize prediction error. A common method to evaluate prediction errors is **cross-validation**, where it is examined whether a model developed on a training dataset generalizes to a holdout dataset. The development of prediction models is becoming increasingly popular with the rise of machine learning approaches.

## 5.3 Explanation

The use of statistical models concerns tests of explanatory theories. In this case, statistical models are used to test causal assumptions, or explanations that we derive from theories. Meehl (1990a) reminds us of the important distinction between a substantive theory, a statistical hypothesis, and observations. Statistical inference is only involved in drawing conclusions about the statistical hypothesis. Observations can lead to the conclusion that the statistical hypothesis is confirmed (or not), but this conclusion does not directly translate into corroboration for the theory. Platt (1964) refers to the systematic application of statistical tests to accumulate knowledge as **strong inference**. It consists of 1) specifying alternative hypotheses, 2) designing an experiment that can corroborate one hypothesis and falsify another, and 3) performing the experiment. This cycle can be repeated to test a number of hypotheses until one hypothesis that can explain the observed data remains. Platt notes how entertaining multiple alternative hypotheses prevents researchers from becoming too attached to a single

hypothesis. When designing a new experiment, researchers should ask themselves what Platt calls **'The Question'**: "But sir, what hypothesis does your experiment disprove?".

We never test a theory in isolation, but always include auxiliary hypotheses about the measures and instruments that are used in a study, the conditions realized in the experiment, up to the **ceteris paribus** clause that assumes "all other things are equal". The best experimental set-up can rarely be 'deduced' from theory, and requires premisses that are tacitly taken for granted. As Hempel (1966) states: "Reliance on auxiliary hypotheses, as we shall see, is the rule rather than the exception in the testing of scientific hypotheses; and it has an important consequence for the question whether an unfavorable test finding, i.e., one that shows  $H$  to be false, can be held to disprove the hypothesis under investigation." Therefore, it is never clear if a failure to corroborate a theoretical prediction should be blamed on the theory or the auxiliary hypotheses. To generate reliable explanatory theories, researchers therefore have to perform lines of research in which auxiliary hypotheses are systematically tested (Uygun Tunç & Tunç, 2022).

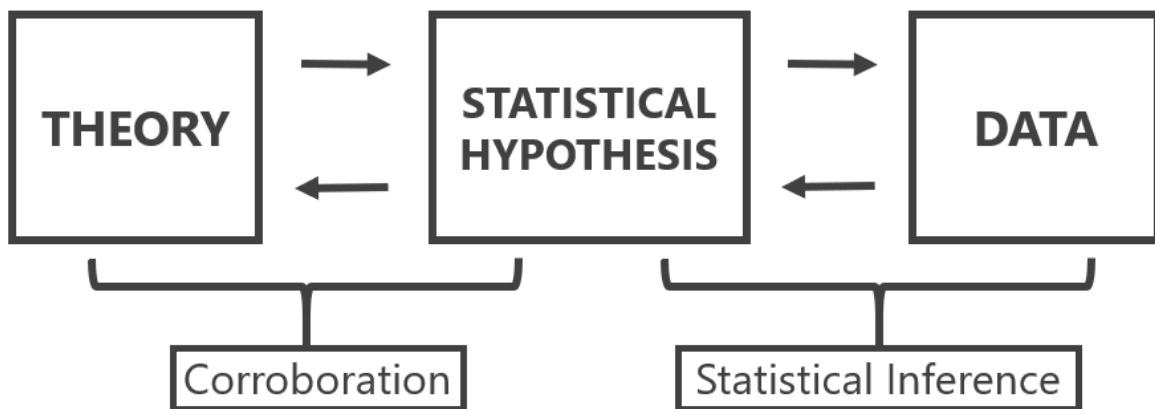


Figure 5.1: Distinction between a theoretical hypothesis, a statistical hypothesis, and observations. Figure based on Meehl, 1990.

## 5.4 Loosening and Tightening

For each of the three questions above, we can ask questions about description, prediction, and explanation during a **loosening** phase when doing research, or during a **tightening** phase (Fiedler, 2004). The distinction is relative. During the loosening stage, the focus is on creating variation that provides the source for new ideas. During the tightening stage, selection takes place with the goal to distinguish useful variants from less useful variants. In descriptive research, an unstructured interview is more aligned with the loosening phase, while a structured interview is more aligned with the tightening phase. In prediction, building a prediction model based on the training set is the loosening phase, while evaluating the prediction error in the

holdout dataset is the tightening phase. In explanation, exploratory experimentation functions to generate hypotheses, while hypothesis tests function to distinguish theories that make predictions that are corroborated from those theories which predictions are not corroborated.

It is important to realize whether your goal is to generate new ideas, or to test new ideas. Researchers are often not explicit about the stage their research is in, which runs the risk of trying to test hypotheses prematurely (Scheel, Tiokhin, et al., 2021). Clinical trials research is more explicit about the different phases of research, and distinguishes Phase 1, Phase 2, Phase 3, and Phase 4 trials. In a Phase 1 trial researchers evaluate the safety of a new drug or intervention in a small group of non-randomized (often healthy) volunteers, by examining how much of a drug is safe to give, while monitoring a range of possible side effects. A phase 2 trial is often performed with patients as participants, and can focus in more detail on finding the definite dose. The goal is to systematically explore a range of parameters (e.g., the intensity of a stimulus) to identify boundary conditions (Dubin, 1969). A phase 3 trial is a large randomized controlled trial with the goal to test the effectiveness of the new intervention in practice. Phase 3 trials require a prespecified statistical analysis plan that strictly controls error rates. Finally, a Phase 4 trial examines long term safety and generalizability. Compared to a Phase 3 trial, there is more loosening, as researchers explore the possibility of interactions with other drugs, or moderating effects in certain subgroups of the population. In clinical trials, a Phase 3 trial requires a huge amount of preparation, and is not undertaken lightly.

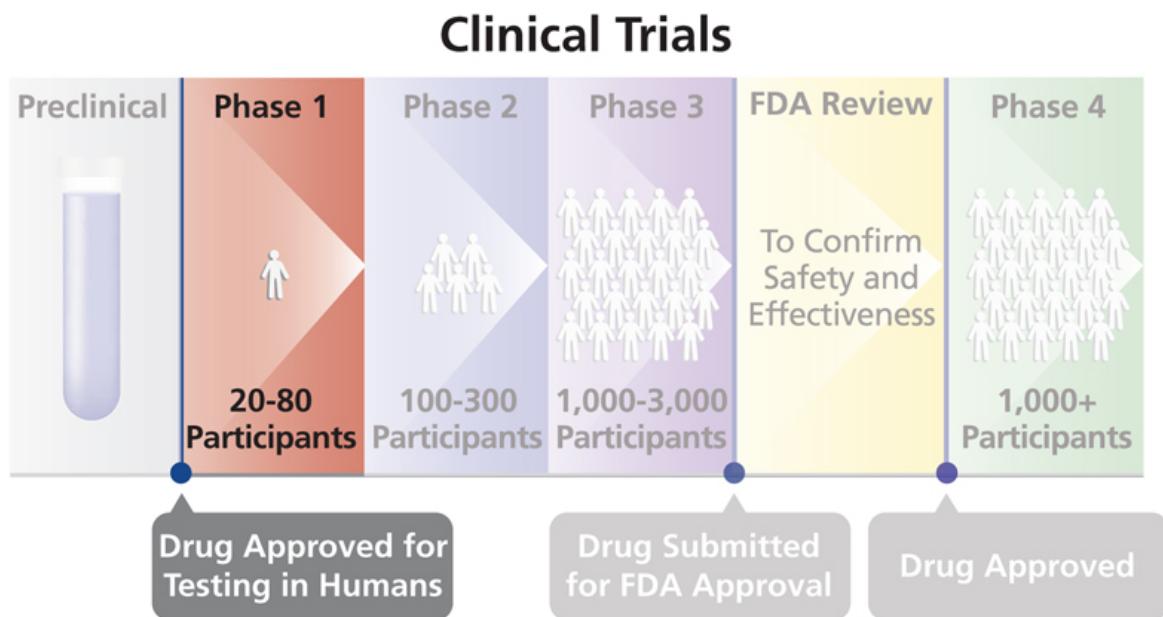


Figure 5.2: Four phases of clinical research. Source.

## 5.5 Three Statistical Philosophies

Royall (1997) distinguishes three questions one can ask:

1. What do I believe, now that I have this observation?
2. What should I do, now that I have this observation?
3. What does this observation tell me about A versus B? (How should I interpret this observation as evidence regarding A versus B?)

One useful metaphor for thinking about these differences is if we look at Hinduism, where there are three ways to reach enlightenment: The Bhakti yoga, or the Path of Devotion, the Karma yoga, or the Path of Action, the Jnana yoga, or the Path of Knowledge. The three corresponding statistical paths are Bayesian statistics, which focuses on updating beliefs, Neyman-Pearson statistics, which focuses on making decisions about how to act, and likelihood approaches, which focus on quantifying the evidence or knowledge gained from the data. Just like in Hinduism the different paths are not mutually exclusive, and the emphasis on these three yoga's differs between individuals, so will scientists differ in their emphasis of their preferred approach to statistics.

The three approaches to statistical modelling (description, prediction, and explanation) can be examined from each the three statistical philosophies (e.g., frequentist estimation, maximum likelihood estimation, and Bayesian estimation, or Neyman-Pearson hypothesis tests, likelihood ratio tests, and Bayes factors). Bayesian approaches start from a specified prior belief, and use the data to update their belief. Frequentist procedures focus on methodological procedures that allow researchers to make inferences that control the probability of error in the long run. Likelihood approaches focus on quantifying the evidential value in the observed data. When used knowledgeably, these approaches often yield very similar inferences (Dongen et al., 2019; Daniël Lakens et al., 2020; Tendeiro & Kiers, 2019). Jeffreys (1939), who developed a Bayesian hypothesis test, noted the following when comparing his Bayesian hypothesis test against frequentist methods proposed by Fisher:

I have in fact been struck repeatedly in my own work, after being led on general principles to a solution of a problem, to find that Fisher had already grasped the essentials by some brilliant piece of common sense, and that his results would be either identical with mine or would differ only in cases where we should both be very doubtful. As a matter of fact I have applied my significance tests to numerous applications that have also been worked out by Fisher's, and have not yet found a disagreement in the actual decisions reached.

At the same time, each approach is based on different principles, and allows for specific inferences. For example, a Neyman-Pearson approach does not quantify evidence, and a Bayesian approach can lead conclusions about the relative support for one over another hypothesis, given specified priors, while ignoring the rate at which such a conclusion would be misleading. Understanding these basic principles is useful, as criticisms on statistical practices (e.g.,

computing  $p$ -values) always boil down to a disagreement about the principles that different statistical philosophies are built on. However, when we survey the literature, we rarely see the viewpoint that all approaches to statistical inferences, including  $p$  values, provide answers to specific questions a researcher might want to ask. Instead, statisticians often engage in what I call the **statistician's fallacy** — a declaration of what they believe researchers really “want to know” without limiting the usefulness of their preferred statistical question to a specific context (Daniël Lakens, 2021). The most well-known example of the statistician’s fallacy is provided by Cohen (1994) when discussing null-hypothesis significance testing:

What’s wrong with NHST? Well, among many other things, it does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does! What we want to know is ‘Given these data, what is the probability that  $H_0$  is true?’

Different statisticians will argue what you actually “want to know” is the posterior probability of a hypothesis, the false-positive risk, the effect size and its confidence interval, the likelihood, the Bayes factor, or the severity with which a hypothesis has been tested. However, it is up to you to choose a statistical strategy that matches the question you want to ask (Hand, 1994).

## 5.6 Falsification

As we have seen above scientists can adopt a Bayesian perspective, and try to quantify their belief in the probability that a hypothesis is true, or they can make claims based on frequentist long run probabilities that have a low probability of being an error. The falsificationist philosophy of Karl Popper is built on this second approach:

Instead of discussing the ‘probability’ of a hypothesis we should try to assess what tests, what trials, it has withstood; that is, we should try to assess how far it has been able to prove its fitness to survive by standing up to tests. In brief, we should try to assess how far it has been ‘corroborated’.

It is important to distinguish *dogmatic falsificationism* - which Karl Popper and Imre Lakatos criticized in their philosophical work - from *naïve falsificationism* and *sophisticated methodological falsificationism*. Dogmatic falsificationism proposes a clear distinction between theory and facts, and argues that facts (observations) can falsify theories. Lakatos (1978) (p. 13) summarizes this view as: “the theoretician proposes, the experimenter - in the name of Nature - disposes”. Lakatos argues against this idea, because “there are and can be no sensations unimpregnated by expectation and therefore there is no natural (i.e. psychological) demarcation between observational and theoretical propositions.” The facts we observe are themselves influenced, at least to some extent, by our theories. Dogmatic falsificationism also argues that the truth-value of observational statements can be derived from facts alone. Popper (2002)

criticized this view, and argued that our direct experiences can not logically justify statements (p. 87): “Experiences can motivate a decision, and hence an acceptance or a rejection of a statement, but a basic statement cannot be justified by them — no more than by thumping the table.” Finally, Lakatos criticizes the demarcation criterion of dogmatic falsificationists, that “only those theories are ‘scientific’ which forbid certain observable states of affairs and therefore are factually disprovable”. Instead, he argues “exactly the most admired scientific theories simply fail to forbid any observable state of affairs.” The reason for this is that theories often only make predictions in combination with a **ceteris paribus** clause (as discussed above), and one therefore has to decide if failed predictions should be relegated to the theory, or the **ceteris paribus** clause.

What is the difference between dogmatic falsificationism and naïve or methodological falsificationism as proposed by Popper? First, Popper accepts there is never a strict distinction between theories and facts, but relegates the influence of theories to “unproblematic background knowledge” that is (tentatively) accepted while testing a theory. These are ‘auxiliary hypotheses’ that, according to Popper, should be used as sparingly as possible. Second, methodological falsificationism separates rejection and disproof. In methodological falsificationism the truth-value of statements is not disproven by facts, but it can be rejected based on agreed upon methodological procedures. These methodological procedures are never certain. As explained in the section on interpreting [p-values](#), Popper argues:

Science does not rest upon solid bedrock. The bold structure of its theories rises, as it were, above a swamp. It is like a building erected on piles. The piles are driven down from above into the swamp, but not down to any natural or ‘given’ base; and if we stop driving the piles deeper, it is not because we have reached firm ground. We simply stop when we are satisfied that the piles are firm enough to carry the structure, at least for the time being.

In methodological falsificationism the demarcation criterion is much more liberal than in dogmatic falsificationism. For example, probabilistic theories are now deemed ‘scientific’ because these can be made ‘falsifiable’ by “specifying certain rejection rules which may render statistically interpreted evidence ‘inconsistent’ with the probabilistic theory” (Lakatos, 1978, p. 25).

Popper and especially Lakatos developed methodological falsification further into **sophisticated falsificationism**. Sophisticated methodological falsificationism stresses that science is often not simply about testing a theory in an experiment, but testing different theories or a series of theories against each other in lines of experiments. Furthermore, it acknowledges that in practice confirmation also plays an important role in deciding between competing theories. Lakatos attempted to integrate views by Thomas Kuhn (1962) on how scientific knowledge was generated in practice, but replaced Kuhn’s social and psychological processes by logical and methodological processes. In sophisticated methodological falsificationism a theory is falsified if the novel theory 1) predicts novel facts, 2) is able to explain the success of the previous theory, and 3) some of the novel predictions are corroborated. Falsification no longer occurs in

single tests of predictions, but through *progressive and degenerative* research lines. Of course, it is difficult to know if a research line is progressing or degenerating in a short time scale. According to Meehl (2004) progressive research lines lead to theories appearing in textbooks, discussion meetings about the theory disappear from conferences, and the theory is no longer tested but mainly improved. Meehl refers to this endpoint as ‘ensconce’ment and suggests fifty-year ensconce as a good proxy for the truth (even though some theories, as those by Newton, can take longer to be falsified). Note that scientists untrained in philosophy of science often incorrectly characterize Popper’s ideas about falsification as dogmatic falsificationism, without realizing Popper’s sophisticated methodological falsificationism was a direct criticism of dogmatic falsificationism.

## 5.7 Severe Tests

A central feature of methodological falsificationism is to design experiments that provide severe tests of hypotheses. According to Mayo (2018) “a claim is severely tested to the extent it has been subjected to and passed a test that probably would have found flaws, were they present.” Severe tests are not the only goal in science - after all, tautologies can be severely tested - and the aim of severe tests should be pursued together with the goal to test interesting theoretical or practical questions. But they are seen as a desireable feature, as nicely expressed by the physicist Richard Feynman (1974): “I’m talking about a specific, extra type of integrity that is not lying, but bending over backwards to show how you’re maybe wrong, that you ought to do when acting as a scientist.” The idea of severe (or ‘risky’) tests is well explained in the article “Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it” by Paul Meehl (1990a):

A theory is corroborated to the extent that we have subjected it to such risky tests; the more dangerous tests it has survived, the better corroborated it is. If I tell you that Meehl’s theory of climate predicts that it will rain sometime next April, and this turns out to be the case, you will not be much impressed with my “predictive success.” Nor will you be impressed if I predict more rain in April than in May, even showing three asterisks (for  $p < .001$ ) in my *t*-test table! If I predict from my theory that it will rain on 7 of the 30 days of April, and it rains on exactly 7, you might perk up your ears a bit, but still you would be inclined to think of this as a “lucky coincidence.” But suppose that I specify which 7 days in April it will rain and ring the bell; then you will start getting seriously interested in Meehl’s meteorological conjectures. Finally, if I tell you that on April 4th it will rain 1.7 inches (.66 cm), and on April 9th, 2.3 inches (.90 cm) and so forth, and get seven of these correct within reasonable tolerance, you will begin to think that Meehl’s theory must have a lot going for it. You may believe that Meehl’s theory of the weather, like all theories, is, when taken literally, false, since probably all theories are false in the eyes of God, but you will at least say, to use Popper’s language,

that it is beginning to look as if Meehl's theory has considerable verisimilitude, that is, "truth-likeness."

To appreciate the concept of severe tests, it is worth reflecting on what **insevere** tests look like. Imagine a researcher who collects data, and after looking at which statistical tests yield a statistically significant result, thinks up a theory. What is the problem of this practice, known as hypothesizing after results are known, or HARKing (Kerr, 1998)? After all, the hypothesis this researcher comes up with could be correct! The reason that HARKing in science is problematic is that the statistical test is completely insevere: There is no way that the statistical test could have proven the claim wrong, if it was wrong. Again, the claim may be correct, but the test does not increase our confidence in this in any way. Mayo (2018) calls this: Bad Evidence, No Test (BENT). A similar problem occurs when researchers engage in [questionable research practices](#). As these practices can substantially inflate the Type 1 error rate, they greatly increase the probability a test will corroborate a prediction, even if that prediction is wrong. Again, the severity of the test is impacted. Of course, you can use questionable research practices and reach a correct conclusion. But after *p*-hacking, the test has a greatly reduced capacity to prove the researcher wrong. If this lack of a severe test is not transparently communicated, readers are fooled into believing a claim has been severely tested, when it has not. These problems can be mitigated by preregistering the statistical analysis plan (Daniël Lakens, 2019).

## 5.8 Risky Predictions

The goal of a hypothesis test is to carefully examine whether predictions that are derived from a scientific theory hold up under scrutiny. Not all predictions we can test are equally exciting. For example, if a researcher asks two groups to report their mood on a scale from 1 to 7, and then predicts the difference between these groups will fall within a range of -6 to +6, we know in advance that it must be so. No result can **falsify** the prediction, and therefore finding a result that **corroborates** the prediction is completely trivial and a waste of time.

The most common division of states of the world that are predicted and that are not predicted by a theory in null-hypothesis significance testing is the following: An effect of exactly zero is *not* predicted by a theory, and all other effects are taken to corroborate the theoretical prediction. Here, I want to explain why this is a very weak hypothesis test. In certain lines of research, it might even be a pretty trivial prediction. It is quite easy to perform much stronger tests of hypotheses. One way would be to reduce the alpha level of a test, as this increases the probability of being proven wrong, when the prediction is wrong. But it is also possible to increase the riskiness of a test by reducing which outcomes are still considered support for the prediction.

Take a look at the three circles below. Each circle represents all possible outcomes of an empirical test of a theory. The blue line illustrates the state of the world that was observed

in a (hypothetical) perfectly accurate study. The line could have fallen anywhere on the circle. We performed a study and found one specific outcome. The black area in the circle represents the states of the world that will be interpreted as *falsifying* our prediction, whereas the white area illustrates the states in the world we predicted, and that will be interpreted as *corroborating* our prediction.

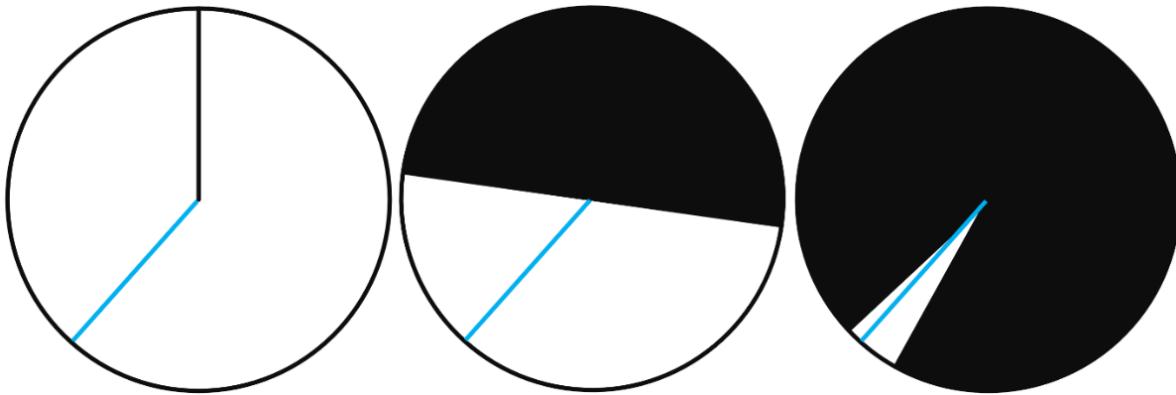


Figure 5.3: Three circles visualizing predictions that exclude different parts of the world.

In the figure on the left, only a tiny fraction of states of the world will falsify our prediction. This represents a hypothesis test where only an infinitely small portion of all possible states of the world is not in line with the prediction. A common example is a two-sided null-hypothesis significance test, which forbids (and tries to reject) only the state of the world where the true effect size is exactly zero.

In the middle circle, 50% of all possible outcomes falsify the prediction, and 50% corroborates it. A common example is a one-sided null-hypothesis test. If you predict the mean is *larger than zero*, this prediction is falsified by all states of the world where the true effect is either *equal to zero*, or *smaller than zero*. This means that half of all possible states of the world can no longer be interpreted as corroborating your prediction. The blue line, or observed state of the world in the experiment, happens to fall in the white area for the middle circle, so we can still conclude the prediction is supported. However, our prediction was already slightly riskier than in the circle on the left representing a two-sided test.

In the scenario in the right circle, almost all possible outcomes are not in line with our prediction – only 5% of the circle is white. Again, the blue line, our observed outcome, falls in this white area, and our prediction is confirmed. However, now our prediction is confirmed in a very risky test. There were many ways in which we could be wrong – but we were right regardless.

Although our prediction is confirmed in all three scenarios above, philosophers of science such as Popper and Lakatos would be most impressed after your prediction has withstood the most severe test (i.e., in the scenario illustrated by the right circle). Our prediction was most specific: 95% of possible outcomes were judged as falsifying our prediction, and only 5% of

possible outcomes would be interpreted as support for our theory. Despite this high hurdle, our prediction was corroborated. Compare this to the scenario on the left – almost any outcome would have supported our theory. That our prediction was confirmed in the scenario in the left circle is hardly surprising.

Making more risky range predictions has some important benefits over the widespread use of null-hypothesis tests. These benefits mean that even if a null-hypothesis test is defensible, it would be preferable if you could test a range prediction. Making a more risky prediction gives your theory higher **verisimilitude**. You will get more credit in darts when you correctly predict you will hit the bullseye, than when you correctly predict you will hit the board. Many sports work like this, such as figure skating or gymnastics. The riskier the routine you perform, the more points you can score, since there were many ways the routine could have failed if you lacked the skill. Similarly, you get more credit for the predictive power of your theory when you correctly predict an effect will fall within 0.5 scale points of 8 on a 10 point scale, than when you predict the effect will be larger than the midpoint of the scale.

Meehl (1967) compared the use of statistical tests in psychology and physics and notes that in physics researchers make point predictions. One way to test point predictions is to examine whether the observed mean falls between an upper and lower bound. In chapter 9 we will discuss how to perform such tests, such as [equivalence tests](#) or [minimum effect tests](#), in practice. Although equivalence tests are often used to test whether an effect falls within a specified range around 0, and interval hypothesis test can be performed around any value, and thereby used to perform more risky tests of hypotheses.

Although Meehl prefers point predictions that lie within a certain range, he doesn't completely reject the use of null-hypothesis significance testing. When he asks 'Is it ever correct to use null-hypothesis significance tests?' his own answer is 'Of course it is' (Meehl, 1990a). There are times, such as very early in research lines, where researchers do not have good enough models, or reliable existing data, to make point or range predictions. Other times, two competing theories are not more precise than that one predicts rats in a maze will learn *something*, while the other theory predicts the rats will learn *nothing*. As Meehl (1990a) writes: "When I was a rat psychologist, I unabashedly employed significance testing in latent-learning experiments; looking back I see no reason to fault myself for having done so in the light of my present methodological views."

There are no good or bad statistical approaches – all statistical approaches provide an answer to a specific question. It makes sense to allow traditional null-hypothesis tests early in research lines, when theories do not make more specific predictions than that 'something' will happen. But we should also push ourselves to develop theories that make more precise range predictions, and then test these more specific predictions in interval hypothesis tests. More mature theories should be able to predict effects in some range – even when these ranges are relatively wide.

## 5.9 Do You Really Want to Test a Hypothesis?

A hypothesis test is a very specific answer to a very specific question. We can use a dart game as a metaphor for the question a hypothesis test aims to answer. In essence, both a dart game and a hypothesis test are a methodological procedure to make a directional prediction: Is A better or worse than B? In a dart game we very often compare two players, and the question is whether we should act as if player A is the best, or player B is the best. In a hypothesis test, we compare two hypotheses, and the question is whether we should act as if the null hypothesis is true, or whether the alternative hypothesis is true.

Historically, researchers have often been interested in testing hypotheses to examine whether predictions that are derived from a scientific theory hold up under scrutiny. Some philosophies of science (but not all) value theories that are able to make predictions. If a darter wants to convince you they are a good player, they can make a prediction ('the next arrow will hit the bulls-eye'), throw a dart, and impress you by hitting the bulls-eye. When a researcher uses a theory to make a prediction, collects data, and observes can claim based on a predefined methodological procedure that the results confirm their prediction, the idea is you are impressed by the **predictive validity of a theory** (de Groot, 1969). The test supports the idea that the theory is a useful starting point to generate predictions about reality. Philosophers of science such as Popper call this 'verisimilitude' – the theory is in some way related to the truth, and it has some 'truth-likeness'.

In order to be impressed when a prediction is confirmed, the prediction must be able to be wrong. In other words, a theoretical prediction needs to be falsifiable. If our predictions concern the presence or absence of clearly observable entities (e.g., the existence of a black swan) it is relatively straightforward to divide all possible states of the world into a set that is predicted by our theory (e.g., all swans are white), and a set that is not predicted by our theory (e.g., swans can have other colors than white). However, many scientific questions concern probabilistic events where single observations contain noise due to random variation – rats have a certain probability to develop a tumor, people have a certain probability to buy a product, or particles have a certain probability to appear after a collision. If we want to forbid certain outcomes of our test when measuring probabilistic events, we can divide the states of the world based on the probability that some result will be observed.

Just because a hypothesis test can be performed, does not mean it is interesting. A hypothesis test is most useful when 1) both data generating models that are decided between have some plausibility, and 2) it is possible to apply an informative methodological procedure.

First, the two competing models should both be good players. Just as in a dart game there would be very little interest if I played Michael van Gerwen (the world champion at the time of writing) to decide who the better dart player is. Since I do not play darts very well, a game between the two of us would not be interesting to watch. Similarly, it is sometimes completely uninteresting to compare two data generating models, one representing the state of the world

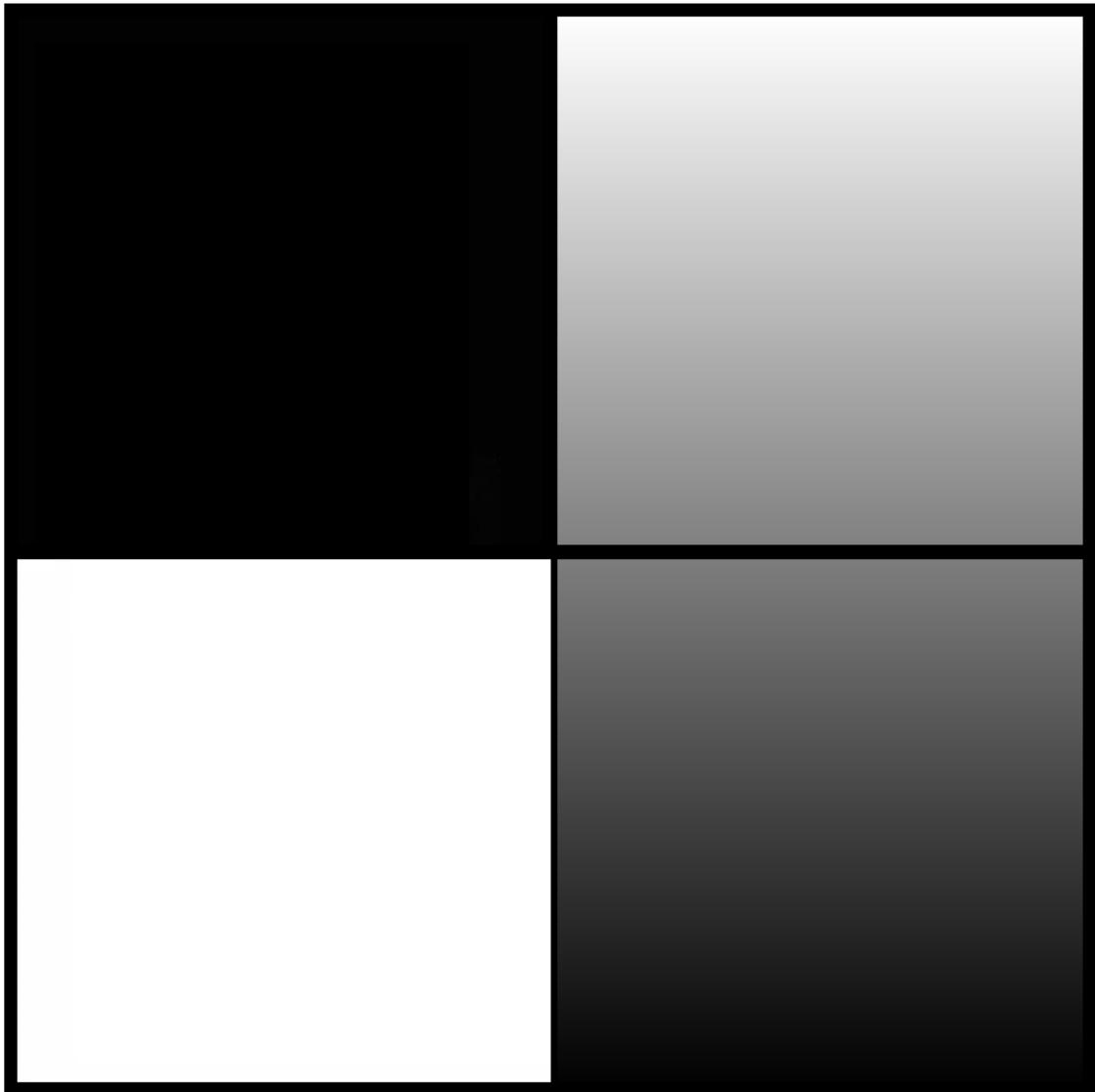


Figure 5.4: Some fields make black and white predictions about the presence or absence of observables, but in many sciences, predictions are probabilistic, and shades of grey.

when there is no effect, and another representing the state of the world when there is some effect, because in some cases the absence of an effect is extremely implausible.

Second, for a hypothesis test to be interesting you need to have designed an informative study. When designing a study, you need to be able to make sure that the methodological rule provides a severe test, where you are likely to corroborate a prediction if it is correct, while at the same time fail to corroborate a prediction when it is wrong (Mayo, 2018). If the world champion in darts and I stand 20 inches away from a dart board and can just push the dart in the location where we want it to end up, it is not possible to show my lack of skill. If we are both blindfolded and throwing the darts from 100 feet, it is not possible for the world champion to display their skill. In a hypothesis test, the statistical severity of a test is determined by the error rates. Therefore, a researcher needs to be able to adequately control error rates to perform a test of a hypothesis with high informational value.

By now it is hopefully clear that hypothesis tests are a very specific tool, that answer a very specific question: After applying a methodological rule to observed data, which decision should I make if I do not want to make incorrect decisions too often? If you have no desire to use a methodological procedure to decide between competing theories, there is no real reason to report the results of a hypothesis test. Even though it might feel like you should test a hypothesis when doing research, carefully thinking about the statistical question you want to ask might reveal that alternative statistical approaches, such as describing the data you have observed, quantifying your personal beliefs about hypotheses, or reporting the relative likelihood of data under different hypotheses might be the approach that answers the question you really want to know.

## 5.10 Directional (One-Sided) versus Non-Directional (Two-Sided) Tests

As explained above, one way to increase the riskiness of a prediction is by performing a directional test. Interestingly, there is quite some disagreement about whether the statistical question you ask in a study should be **directional** (meaning that only effects in a predicted direction will lead to rejection of the null hypothesis) or **non-directional** (meaning that effects in either direction will lead to the rejection of the null-hypothesis). For example, Baguley (2012) writes “one-sided tests should typically be avoided” because researchers are rarely willing to claim an effect in the non-predicted direction is non-significant, regardless of how large it is. At the same time, Jones (1952) has stated: “Since the test of the null hypothesis against a one-sided alternative is the most powerful test for all directional hypotheses, it is strongly recommended that the one-tailed model be adopted wherever its use is appropriate”, and Cho & Abe (2013) complain about the “widespread overuse of two-tailed testing for directional research hypotheses tests”. Let’s reflect on some arguments for or against the choice to perform a one-sided test.

First, it is clear that a directional test provides a clear advantage in statistical power. As Figure 5.5 shows, the ratio of the sample for a non-directional versus a directional test means that approximately 80% of the sample size of a non-directional test is required to achieve the same power in a directional test (the exact benefit depends on the power and effect size, as seen in the figure below).

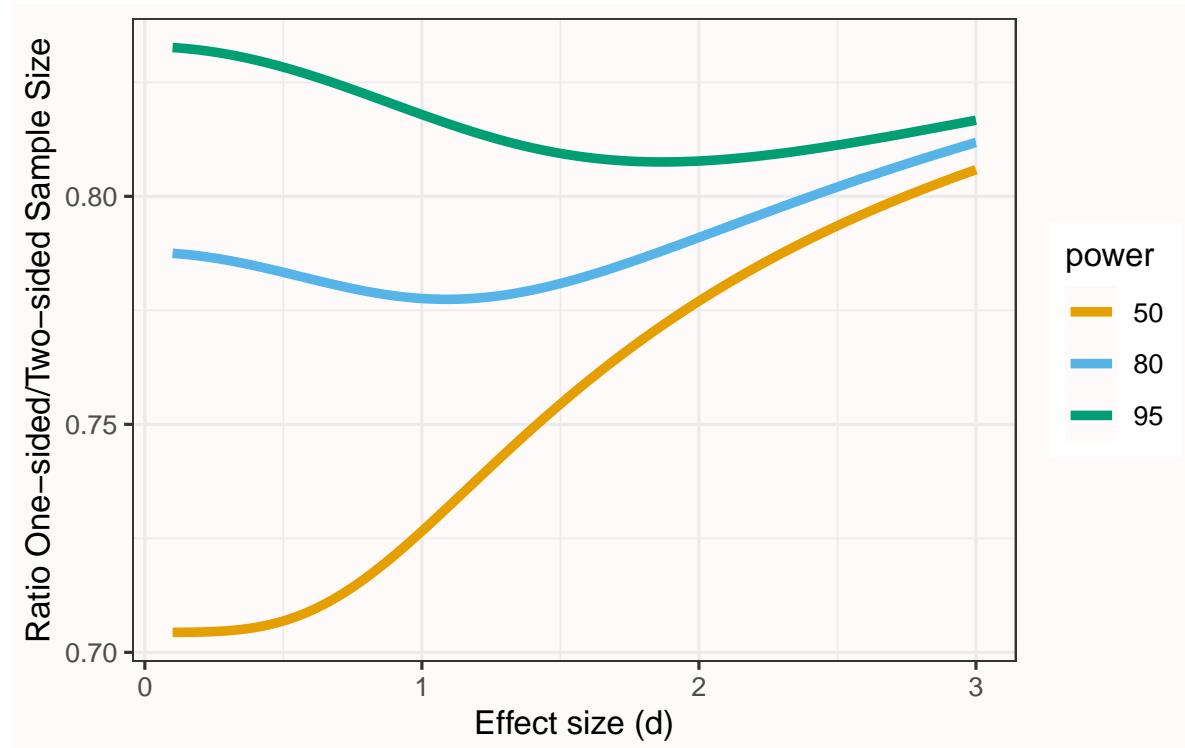


Figure 5.5: Ratio of the required sample size for a one-sample  $t$ -test for a non-directional/directional test to achieve 50%, 80% or 95% power.

Because in a directional test the alpha level is used for only one tail of the distribution, the critical test value is lower, and all else equal, power is higher. This reduction of the critical value required to declare a statistically significant effect has been criticized because it leads to weaker evidence. For example, Schulz & Grimes (2005) write: “Using a one-sided test in sample size calculations to reduce required sample sizes stretches credulity.”. This is trivially true: Any change to the design of a study that requires a smaller sample size reduces the strength of the evidence you collect, since the strength of evidence is inherently tied to the total number of observations. However, it conflates two types of statistical philosophies, namely a likelihoodist approach, which aims to quantify relative evidence, and a frequentist approach, which aims to provide a procedure to make claims with a maximum error rate. There is a difference between designing a study that yields a certain level of evidence, and a study that adequately controls the error rates when performing a hypothesis test. If you desire a specific level of evidence,

design a study that provides this desired level of evidence. If you desire to control the error rate of claims, then that error rate is at most 5% as long as the alpha level is 5%, regardless of whether a one-sided or two-sided test is performed.

Note that there is a subtle distinction between a directional and a one-sided test (Baguley, 2012). Although the two terms overlap when performing a  $t$ -test, they do not overlap for an  $F$ -test. The  $F$ -value and the  $t$ -value are related:  $t^2 = F$ . This holds as long as the  $df1 = 1$  (e.g.,  $F(1, 100)$ , or in other words as long as only two groups are compared. We can see in Figure 5.6 that the two distributions touch at  $t = 1$  (as  $1^2 = 1$ ), and that the  $F$ -test has no negative values due to the squared nature of the distribution. The critical  $t$ -value, squared, of a non-directional  $t$ -test with a 5% error rate equals the critical  $F$ -value for an  $F$ -test, which is always one-sided, with a 5% error rate. Due to the ‘squared’ nature of an  $F$ -test, an  $F$ -test is always non-directional. You can logically not halve the  $p$ -value in an  $F$ -test to perform a ‘one-sided’ test, because you can’t have a directional  $F$ -test. When comparing two groups, you can use a  $t$ -test instead of an  $F$ -test, which can be directional.

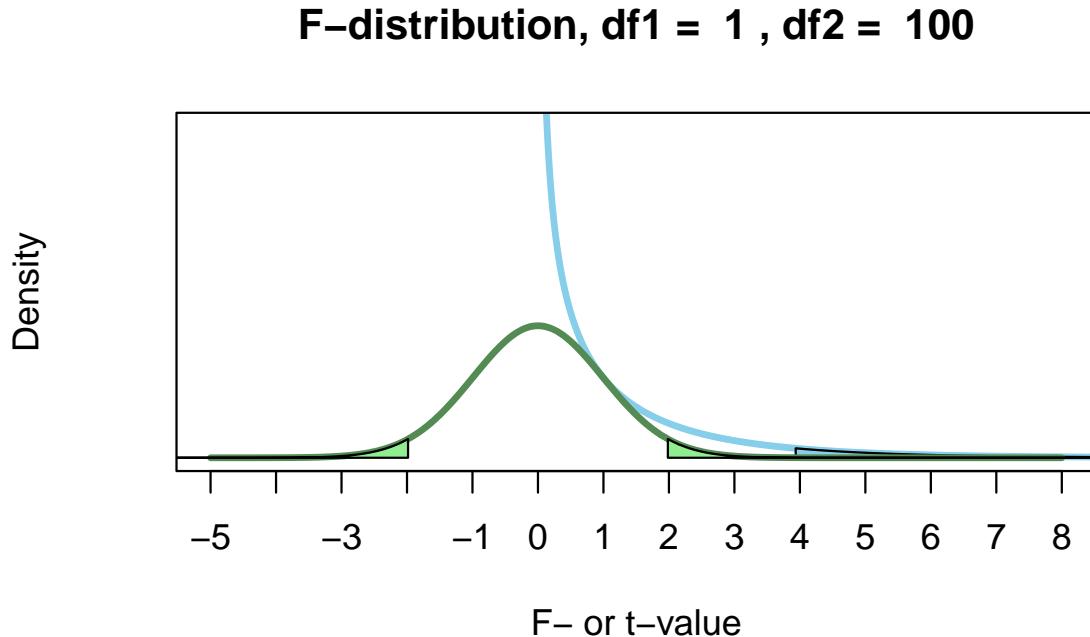


Figure 5.6: Distribution and rejection areas for a two-sided  $t$ -test and the corresponding  $F$ -test with  $df1 = 1$  and  $df2 = 100$ .

A final concern raised against one-sided tests is that surprising findings in the opposite direction might be meaningful, and should not be ignored. I agree, but this is not an argument against one-sided testing. The goal in hypothesis testing is, not surprisingly, to test a hypothesis. If

you have a directional hypothesis, a result in the opposite direction can never confirm your hypothesis. It can lead one to create a new hypothesis, but this new hypothesis should be tested on a new dataset (de Groot, 1969). It makes sense to *describe* an unexpected effect in the opposite direction of your prediction, but there is a difference between describing data, and testing a hypothesis. A one-sided hypothesis test does not prohibit researchers from describing unexpected data patterns. And if you really want to test if there is an effect in either direction, simply preregister a two-sided test.

## 5.11 Systematic Noise, or the Crud Factor

Meehl (1978) believes “the almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories in the soft areas is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology”. At the same time, he also wrote: “When I was a rat psychologist, I unabashedly employed significance testing in latent-learning experiments; looking back I see no reason to fault myself for having done so in the light of my present methodological views” (Meehl, 1990a). When he asks ‘Is it ever correct to use null-hypothesis significance tests?’ his own answer is:

Of course it is. I do not say significance testing is never appropriate or helpful; there are several contexts in which I would incline to criticize a researcher who failed to test for significance.

Meehl is not of the opinion that null hypothesis significance tests are not useful at all, but that the question if *any* difference from zero exists is sometimes not a very interesting question to ask. Crucially, Meehl is especially worried about the widespread use of null hypothesis significance tests where there is room for **systematic noise**, or the **crud factor** in the data that are analyzed. The presence of systematic noise in data means that it is extremely unlikely that the null hypothesis is true, and combined with a large enough dataset, the question whether the null hypothesis can be rejected is uninteresting.

Systematic noise can only be excluded in an ideal experiment. In this ideal experiment, only one single factor can lead to an effect, such as in a perfect **randomized controlled trial**. Perfection is notoriously difficult to achieve in practice. In any not perfect experiment, there can be tiny causal factors that, although not being the main goal of the experiment, lead to differences between the experimental and control condition. Participants in the experimental condition might read more words, answer more questions, need more time, have to think more deeply, or process more novel information. Any of these things could slightly move the true effect size away from zero – without being related to the independent variable the researchers aimed to manipulate. The difference is reliable, but not caused by anything the researcher is **theoretically interested** in. In real life, experiments are not even close to

perfect. Consequently, there is always some room for systematic noise, although there is no way to know how large this systematic noise is in any specific study.

Systematic noise is especially a problem in studies where there is no randomization, such as in correlational studies. As an example of correlational data, think about research that examines differences between women and men. In such a study the subjects cannot be randomly assigned to each condition. In such non-experimental studies, it is possible that '**everything is correlated to everything**'. Or slightly more formally, crud can be defined as the epistemological concept that, in correlational research, all variables are connected through multivariate causal structures which result in real non-zero correlations between all variables in any given dataset (Orben & Lakens, 2020). For example, men are on average taller than women, and as a consequence men will be asked by strangers to pick an object from a high shelf in a supermarket a bit more often than women. If we ask men and women 'how often do you help strangers' this average difference in height has some tiny but systematic effect on their responses, even though a researcher might be theoretically interested in differences unrelated to height. In this specific case, systematic noise moves the mean difference from zero to a slightly higher value for men – but an unknown number of other sources of systematic noise are at play, and these all interact, leading to an unknown final true population difference that is very unlikely to be exactly zero.

As a consequence, some scientific fields find tests of correlations relatively uninteresting. Researchers in these fields might find it interesting to *estimate* the size of correlations, but they might not find it worthwhile to perform a null hypothesis significance *test* for a correlation, as with a large enough dataset, statistical significance is practically guaranteed. This is increasingly true, the bigger the dataset. As an anecdote, while working on a paper on [sequential analysis](#), I asked my collaborator Prof. Wassmer why the `rpact` package did not have a module for tests of correlations. He replied that there was not enough interest in null hypothesis significance tests for correlations in biopharmaceutical statistics, because as everything correlates with everything anyway, why would anyone want to test it?

When you perform a nil null hypothesis test, you should justify why the nil null hypothesis is an interesting hypothesis to test against. This is not always self-evident, and sometimes the nil null hypothesis is simply not very interesting. Is it plausible that the nil null hypothesis is true? If not, then it is more interesting to perform a [minimal effect test](#). For a concrete example of how to determine if the presence of crud warrants the use of minimal effect tests in a literature, see C. J. Ferguson & Heene (2021).

Several Many Lab Registered Replication Reports in psychology, where randomized experiments with very large sample sizes are performed that revisit published findings, have shown that for all practical purposes, and given the sample sizes psychologists are able to collect, it has proven surprisingly difficult to find significant effects. A multilab replication study examining the action-sentence compatibility effect showed an average effect on the logarithm of the lift-off times close to 0 [-0.006, 0.004] in 903 native English speakers (Richard D. Morey et al., 2021). A Registered Replication Report examining the effect of priming participants with either professor or hooligan related concepts yielded a non-significant difference in the

number of general knowledge questions answered of a difference of 0.14% [-0.71%, 1.00%] in a sample of 4493 participants (O'Donnell et al., 2018). A Registered Replication Report examining the effect of recalling the ten commandments or 10 books read in highschool on how often people cheated on a problem-solving task showed a non-significant difference of 0.11 [-0.09; 0.31] matrices in a sample of 4674 participants (Verschueren et al., 2018). A Registered Replication Report testing the facial feedback hypothesis showed a non-significant effect on funniness ratings between conditions where participants were manipulated to move muscles related to smiling or pouting of 0.03 [-0.11; 0.16] scale units in a sample of 1894 participants (Wagenmakers et al., 2016). A multi-lab replication study of the ego-depletion effect (which will feature more prominently in the chapter on [bias](#)) observed an effect of  $d = 0.04$  [-0.07, 0.15] in a sample of 2141 participants (Hagger et al., 2016). These studies suggest that sometimes the nil null hypothesis is a plausible model to test against, and that even with sample sizes much larger than are typically collected in psychological research, the nil null is surprisingly difficult to reject.

Other multi-lab studies provide indications of tiny true effects, which could be due to the crud factor. Colling et al. (2020) observed congruency effects in the attentional SNARC effect for four inter-stimulus interval conditions (250, 500, 750, and 1000 ms) of -0.05 ms [-0.82]; 0.71], 1.06 ms [0.34; 1.78], 0.19 ms [-0.53; 0.90], and 0.18 ms [-0.51; 0.88] with a sample size of 1105 participants. For the statistically significant effect in the 500 ms ISI condition (which might be crud) they conclude: “we view a difference of about 1 ms, even if “real,” as too small for any neurally or psychologically plausible mechanism—particularly one constrained to operate only within a narrow time window of 500 ms after the stimulus.” McCarthy et al. (2018) observed a difference of 0.08 [0.004; 0.16] in how hostile ambiguous behavior in a vignette was rated after a priming task where more or less words were related to hostility, and conclude “Our results suggest that the procedures we used in this replication study are unlikely to produce an assimilative priming effect that researchers could practically and routinely detect.” In these instances, the null-hypothesis can be rejected, but the observed effect size is deemed too small to matter. As discussed in the chapter on equivalence testing and interval hypotheses, the solution to this problem is to specify a [smallest effect size of interest](#).

## 5.12 Dealing with Inconsistencies in Science

We might prefer clear answers from scientific research, but in practice we are often presented with inconsistent results in a scientific literature. What should we do when ‘even scientists can't agree’?

According to Karl Popper the ability of scientists to reach consensus about basic statements is key criteria of science:

If some day it should no longer be possible for scientific observers to reach agreement about basic statements this would amount to a failure of language as a means

of universal communication. It would amount to a new ‘Babel of Tongues’: scientific discovery would be reduced to absurdity. In this new Babel, the soaring edifice of science would soon lie in ruins.

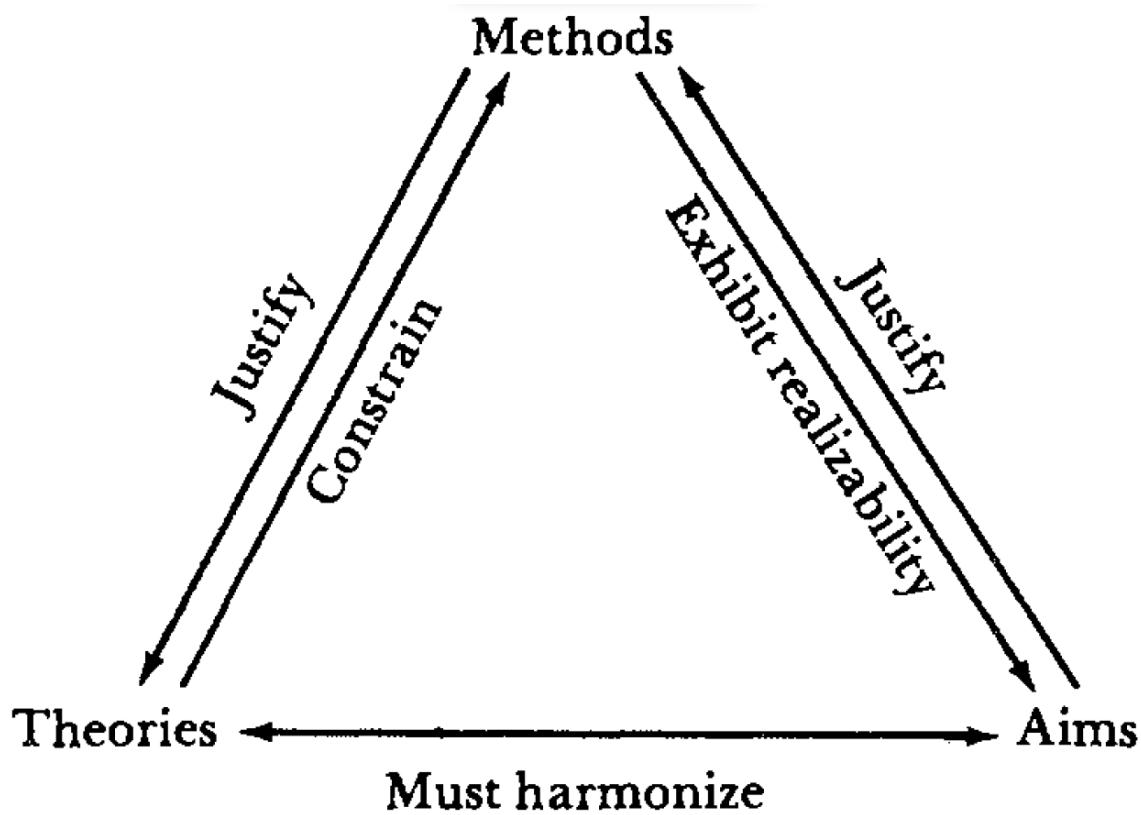
Other philosophers of science, such as Thomas Kuhn, viewed different paradigms in science as **incommensurable**. Because research Kuhn believed paradigms change dramatically over time (which he calls scientific revolutions) advocates of competing theories can not directly compare and discuss their theories (Kuhn, 1962). Kuhn acknowledges that scientists do reach consensus within a particular research tradition (which he calls ‘normal science’):

Men whose research is based on shared paradigms are committed to the same rules and standards for scientific practice. That commitment and the apparent consensus it produces are prerequisites for normal science, i.e., for the genesis and continuation of a particular research tradition.

Harry Laudan aims to resolve these different views on whether scientists can or can not reach consensus by distinguishing disagreements on three levels (Laudan, 1986). The first level involves claims about theoretical or observable entities, where scientists can have factual disagreements or factual consensus. These can be resolved by methodological rules. However, scientists can also have disagreements about which methods or procedures should be used. These disagreements on the methodological level can only be resolved by discussing the aims of science, as the methods we use should be optimal techniques to achieve our aims in science. Laudan calls this the axiological level. According to Laudan, there is a mutual justification process between these three levels, and even though there are different aims, methods, and theories, scientists need to be able to justify how their approach is coherent.

Factual inconsistencies can emerge in different ways. First, the support for a specific scientific claim can be mixed, in that some studies show statistically significant results ( $p < .05$ ), while other studies do not ( $p > 0.05$ ). We have seen that **mixed results** are expected in sets of studies. It is possible (and sometimes likely) that the statistical power of studies is low. If 60% of studies yield  $p < .05$  and 40% of studies yield  $p > .05$  this might seem inconsistent, but in reality the pattern of results would be perfectly consistent with the expected long run Type 2 error rates in a set of studies with low statistical power. We will see later that combining all studies in a **meta-analysis** can yield more clarity when individual studies have low statistical power.

As Popper (2002) writes: “a few stray basic statements contradicting a theory will hardly induce us to reject it as falsified. We shall take it as falsified only if we discover a reproducible effect which refutes the theory.” Remember that any claim of rejecting or accepting a hypothesis is done with a certain error rate, and that close replication studies are the only way to distinguish erroneous from correct dichotomous claims in statistical hypothesis tests. If the null hypothesis is true, the alpha level determines how many false positive results will be observed. Again, these errors should occur as often as the Type 1 error rate a study was designed to have. In a two-sided test performed with an alpha level of 5%, 2.5% of all studies



**Fig. 2. The Triadic Network of Justification**

Figure 5.7: The interrelationship between the methodological level, theories that explain factual observation, and the aims of science according to Laudan's reticulated model of scientific rationality.

will lead to a claim about an effect in the positive direction, and 2.5% of the studies will lead to a claim about an effect in the negative direction (when in reality the null hypothesis is true). Seeing statistically significant effects in the literature in both the positive and negative direction might seem inconsistent, but if all these findings are Type 1 errors, they should occur exactly as often as expected based on the chosen alpha level.

Even if there is a true effect, just because of random variation it is possible to very rarely observe a statistically significant effect in the opposite direction, which has been called an ‘error of the third kind’ (Kaiser, 1960) or a Type S error (Altoè et al., 2020; Gelman & Carlin, 2014). Although such results are rare, you are much more likely to hear about them because a newspaper article that reads ‘as opposed to what researchers believed for the last decades, a new study suggests that spending *less* time studying might lead to better exam results’ makes for a very attention-grabbing headline. Because there is a real risk that counter-intuitive findings are actually just flukes, it would be good if science journalists spent more time reporting on meta-analyses, and less time reporting on surprising novel findings.

If all research results are transparently reported, multiple studies should quickly indicate whether we were dealing with a relatively rare Type 1 error, or a true finding. However, as we will see in the chapter on [bias](#) not all research findings are shared. As explained in the section on the [positive predictive value](#) this can lead to a literature where many Type 1 errors are published, which makes it difficult to determine if there is a true effect or not. The combination of random variation and bias in the scientific literature can make it easy to find a single study that can be used to support any viewpoint or argument. To prevent confirmation bias, you should actively search for studies that contradict the point you want to make, and evaluate the evidence across multiple studies. If this larger literature shows inconsistencies, bias detection tests might provide a first indication that the cause of the inconsistency is a biased literature. To resolve inconsistencies due to bias in the literature new studies should be performed - preferably Registered Reports that have a preregistered statistical analysis plan and are published regardless of whether results are significant or not.

A second type of inconsistency occurs when two conflicting claims have been supported by an unbiased literature. In this case, different researchers might argue that one or the other claim is true, but it is most likely that both are false, as both are only true *under specific conditions*. One might argue that in some research fields, like psychology, there are always some conditions under which a claim is true, and some conditions under which the same claim is false. Indeed, if one wanted to summarize all knowledge generated by psychologists in two words, it would be “it depends”. McGuire (2004) refers to this as ‘perspectivism’, and proposed it as a fruitful approach when theorizing: “all hypotheses, even pairs of contraries, are true (at least from some perspective).” Thinking in advance about when a prediction might hold and when not is a good approach to theorize about boundary conditions and other types of **moderators**. If two conflicting claims have received reliable support, the presence of a moderator means that a statistical relationship between two variables depends on a third variable. In Figure 5.8 we see that the effect of X and Y depends on the level of Z (Z impacts the relationship between X and Y). For example, an effect of winning the lottery on how happy you are depends on whether

your friends and family are happy for you (let's call this condition  $Z = 0$ ), or whether arguments about money ruin your personal relationships ( $Z = 1$ ). The effect (indicated as  $a$  and  $b$ ) might be positive in one condition of  $Z$ , and absent or even negative in another condition of  $Z$ . As there are many possible moderators, and studying moderation effects typically requires more resources than studying main effects, it is possible there is relatively little empirical research that examines moderators, in which case inconsistencies remain unresolved.

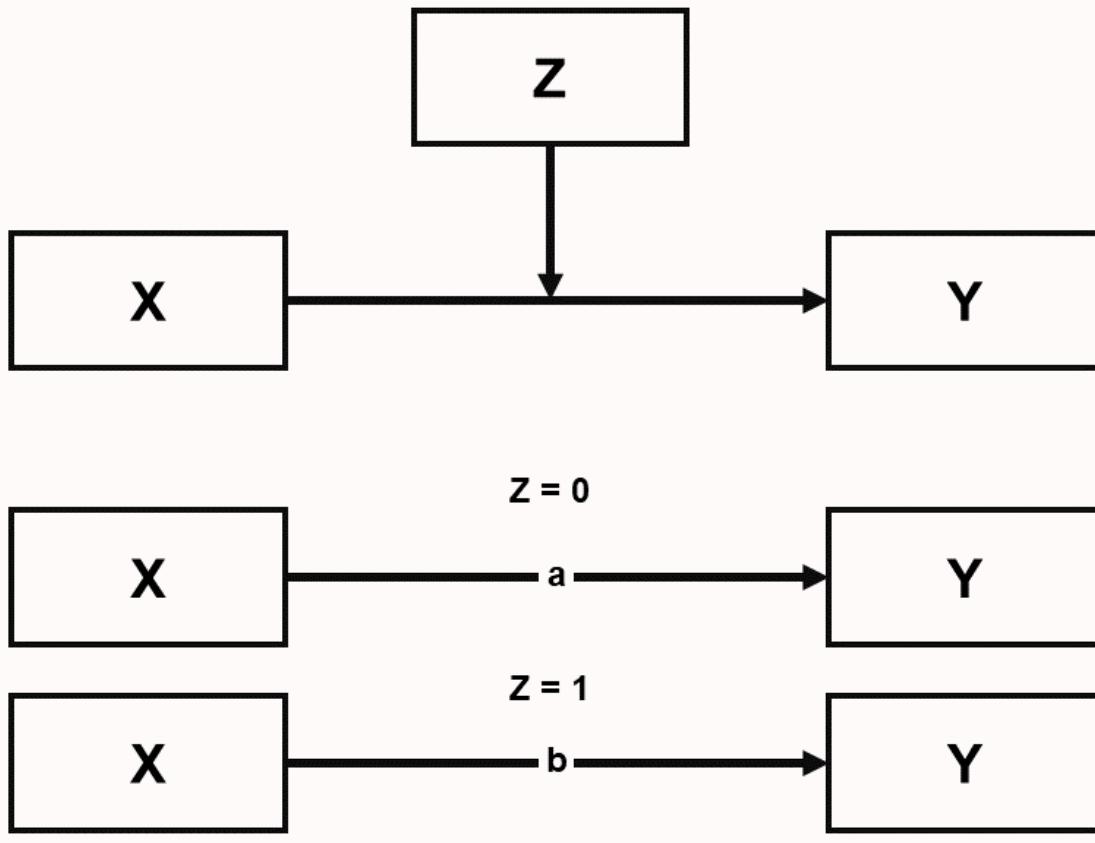


Figure 5.8: Path model of a moderation effect where the effect of  $X$  on  $Y$  depends on  $Z$ , where the effect sizes  $a$  and  $b$  differ from each other depending on the level of  $Z$ .

Some researchers strongly believe failures to replicate published findings can be explained by the presence of hitherto unidentified, or ‘hidden’, moderators (Stroebe & Strack, 2014). There has been at least one example of researchers who were able to provide modest support for the idea that a previous failure to replicate a finding was due to how personally relevant a message in the study was (Luttrell et al., 2017). It is difficult to reliably identify moderator variables that explain failures to replicate published findings, but easy to raise them as an explanation when replication studies do not observe the same effect as the original study. Especially in the social sciences it is easy to point to moderators that are practically impossible to test, such

as the fact that society has changed over time, or that effects that work in one culture might not replicate in different cultures. This is an age-old problem, already identified by Galileo in [The Assayer](#), one of the first books on the scientific method. In this book, Galileo discusses the claim that Babylonians cooked eggs by whirling them in a sling, which is impossible to replicate, and writes:

‘If we do not achieve an effect which others formerly achieved, it must be that we lack something in our operation which was the cause of this effect succeeding, and if we lack one thing only, then this alone can be the true cause. Now we do not lack eggs, or slings, or sturdy fellows to whirl them, and still they do not cook, but rather cool down faster if hot. And since we lack nothing except being Babylonians, then being Babylonian is the cause of the egg hardening.’

Resolving inconsistencies in science is an effortful process that can be facilitated by engaging in an **adversarial collaboration**, where two teams join forces to resolve inconsistencies (Mellers et al., 2001). It requires first establishing a reliable empirical basis by reducing the probability of Type 1 and Type 2 errors and bias, and then systematically testing hypotheses that are proposed to explain inconsistencies (Uygun Tunç & Tunç, 2022).

## 5.13 Verisimilitude and Progress in Science

It makes a fellow cheery To be cooking up a theory; And it needn’t make him blue  
That it’s not exactly true If at least it’s getting neary. Verisimilitude — Meehl,  
11/7/1988

Does science offer a way to learn what is true about our world? According to the perspective in philosophy of science known as *scientific realism*, the answer is ‘yes’. Scientific realism is the idea that successful scientific theories that have made novel predictions give us a good reason to believe these theories make statements about the world that are at least partially true. Known as the *no miracle argument*, only realism can explain the success of science, which consists of repeatedly making successful predictions (Duhem, 1954), without requiring us to believe in miracles.

Not everyone thinks that it matters whether scientific theories make true statements about the world, as scientific realists do. Laudan (1981) argues against scientific realism based on a pessimistic meta-induction: If theories that were deemed successful in the past turn out to be false, then we can reasonably expect all our current successful theories to be false as well. Van Fraassen (1980) believes it is sufficient for a theory to be ‘empirically adequate’, and make true predictions about things we can observe, irrespective of whether these predictions are derived from a theory that describes how the unobservable world is in reality. This viewpoint is known as *constructive empiricism*. As Van Fraassen summarizes the constructive empiricist perspective (1980, p.12): “Science aims to give us theories which are empirically adequate; and acceptance of a theory involves as belief only that it is empirically adequate”.

The idea that we should ‘believe’ scientific hypotheses is not something scientific realists can get behind. Either they think theories make true statements about things in the world, but we will have to remain completely agnostic about when they do (Feyerabend, 1993), or they think that corroborating novel and risky predictions makes it reasonable to believe that a theory has some ‘truth-likeness’, or *verisimilitude*. The concept of verisimilitude is based on the intuition that a theory is closer to a true statement when the theory allows us to make more true predictions, and less false predictions. When data is in line with predictions, a theory gains verisimilitude, when data are not in line with predictions, a theory loses verisimilitude (Meehl, 1978). Popper clearly intended verisimilitude to be different from belief (Niiniluoto, 1998). Importantly, verisimilitude refers to how close a theory is to the truth, which makes it an ontological, not epistemological question. That is, verisimilitude is a function of the degree to which a theory is similar to the truth, but it is not a function of the degree of belief in, or the evidence for, a theory (Meehl, 1978, 1990a). It is also not necessary for a scientific realist that we ever know what is true – we just need to be of the opinion that we can move closer to the truth (known as comparative scientific realism, Kuipers (2016)).

Attempts to formalize verisimilitude have been a challenge, and from the perspective of an empirical scientist, the abstract nature of this ongoing discussion does not really make me optimistic it will be extremely useful in everyday practice. On a more intuitive level, verisimilitude can be regarded as the extent to which a theory makes the most correct (and least incorrect) statements about specific features in the world. One way to think about this is using the ‘possible worlds’ approach (Niiniluoto, 1999), where for each basic state of the world one can predict, there is a possible world that contains each unique combination of states.

For example, consider the experiments by Stroop (1935), where color related words (e.g., RED, BLUE) are printed either in congruent colors (i.e., the word RED in red ink) or incongruent colors (i.e., the word RED in blue ink). We might have a very simple theory predicting that people automatically process irrelevant information in a task. When we do two versions of a Stroop experiment, one where people are asked to read the words, and one where people are asked to name the colors, this simple theory would predict slower responses on incongruent trials, compared to congruent trials. A slightly more advanced theory predicts that congruency effects are dependent upon the salience of the word dimension and color dimension (Melara & Algom, 2003). Because in the standard Stroop experiment the *word* dimension is much more salient in both tasks than the *color* dimension, this theory predicts slower responses on incongruent trials, but only in the color naming condition. We have four possible worlds, two of which represent predictions from either of the two theories, and two that are not in line with either theory.

	Responses Color Naming	Responses Word Naming
World 1	Slower	Slower
World 2	Slower	Not Slower
World 3	Not Slower	Slower
World 4	Not Slower	Not Slower

Meehl (1990b) discusses a ‘box score’ of the number of successfully predicted features, which he acknowledges is too simplistic. No widely accepted formalized measure of verisimilitude is available to express the similarity between the successfully predicted features by a theory, although several proposals have been put forward (Cevolani et al., 2011; Niiniluoto, 1998; Oddie, 2013). However, even if formal measures of verisimilitude are not available, it remains a useful concept to describe theories that are assumed to be closer to the truth because they make novel predictions (Psillos, 1999).

# 6 Effect Sizes

Effect sizes are an important statistical outcome in most empirical studies. Researchers want to know whether an intervention or experimental manipulation has an effect greater than zero, or (when it is obvious that an effect exists) how big the effect is. Researchers are often reminded to report effect sizes, because they are useful for three reasons. First, they allow you to present the magnitude of the reported effects, which in turn allows you to reflect on the **practical significance** of the effects, in addition to the *statistical* significance. Second, effect sizes allow researchers to draw meta-analytic conclusions by comparing standardized effect sizes across studies. Third, effect sizes from previous studies can be used when planning a new study in an a-priori power analysis.

A measure of effect size is a quantitative description of the strength of a phenomenon. It is expressed as a number on a scale. For **unstandardized effect sizes**, the effect size is expressed on the scale that the measure was collected on. This is useful whenever people are able to intuitively interpret differences on a measurement scale. For example, children grow on average 6 centimeters a year between the age of 2 and puberty. We can interpret 6 centimeters a year as an effect size, and many people in the world have an intuitive understanding of how large 6 cm is. Whereas a *p*-value is used to make a claim about whether there is an effect, or whether we might just be looking at random variation in the data, an effect size is used to answer the question of how large the effect is. This makes an effect size estimate an important complement to *p*-values in most studies. A *p*-value tells us that we can claim that children grow as they age; effect sizes tell us what size clothes we can expect children to wear when they are a certain age, and how long it will take before their new clothes are too small.

For people in parts of the world that do not use the metric system, it might be difficult to understand what a difference of 6 cm is. Similarly, a psychologist who is used to seeing scores of 0–20 on their preferred measure of depression might not be able to grasp what a change of 3 points means on a different measure, which could have a scale of 0–10 or 0–50. To facilitate a comparison of effect sizes across situations where different measurement scales are used, researchers can report **standardized effect sizes**. A standardized effect size, such as **Cohen's *d***, is computed by dividing the difference on the raw scale by the standard deviation, and is thus scaled in terms of the variability of the sample from which it was taken. An effect of  $d = 0.5$  means that the difference is the size of half a standard deviation of the measure. This means that standardized effect sizes are determined both by the magnitude of the observed phenomenon and the size of the standard deviation. As standardized effect sizes are a ratio of the mean difference divided by the standard deviation, different standardized effect sizes can indicate the mean difference is not identical, or the standard deviations are not identical,

or both. It is possible that two studies find the same unstandardized difference, such as a 0.5-point difference on a 7-point scale, but because the standard deviation is larger in Study A (e.g., SD = 2) than in Study B (e.g., SD = 1) the standardized effect sizes differ (e.g., Study 1:  $0.5/2 = 0.25$ , Study B:  $0.5/1 = 0.5$ ).

Standardized effect sizes are common when variables are not measured on a scale that people are familiar with, or are measured on different scales within the same research area. If you ask people how happy they are, an answer of ‘5’ will mean something very different if you ask them to answer on a scale from 1 to 5 versus a scale from 1 to 9. Standardized effect sizes can be understood and compared regardless of the scale that was used to measure the dependent variable. Despite the ease of use of standardized effect size measures, there are good arguments to prefer to report and interpret unstandardized effect sizes over standardized effect sizes wherever possible (Baguley, 2009).

Standardized effect sizes can be grouped in two families (Rosnow & Rosenthal (2009)): The  $d$  family (consisting of standardized mean differences) and the  $r$  family (consisting of measures of strength of association). Conceptually, the  $d$  family effect sizes are based on the difference between observations, divided by the standard deviation of these observations, while the  $r$  family effect sizes describe the proportion of variance that is explained by group membership. For example, a correlation ( $r$ ) of 0.5 indicates that 25% of the variance ( $r^2$ ) in the outcome variable is explained by the difference between groups. These effect sizes are calculated from the sum of squares of the residuals (the differences between individual observations and the mean for the group, squared and summed) for the effect, divided by the total sum of squares in the design.

## 6.1 Effect sizes

What is the most important outcome of an empirical study? You might be tempted to say it’s the  $p$ -value of the statistical test, given that it is almost always reported in articles, and determines whether we call something ‘significant’ or not. However, as Cohen (1990) writes in his ‘Things I’ve learned (so far)’:

I have learned and taught that the primary product of a research inquiry is one or more measures of effect size, not  $p$ -values.

Although what you want to learn from your data is different in every study, and there is rarely any single thing that you always want to know, effect sizes are a very important part of the information we gain from data collection.

One reason to report effect sizes is to facilitate future research. It is possible to perform a meta-analysis or a power analysis based on unstandardized effect sizes and their standard deviation, but it is easier to work with standardized effect sizes, especially when there is variation in the measures that researchers use. But the main goal of reporting effect sizes is to reflect on the

question whether the observed effect size is meaningful. For example, we might be able to reliably measure that, on average, 19-year-olds will grow 1 centimeter in the next year. This difference would be statistically significant in a large enough sample, but if you go shopping for clothes when you are 19 years old, it is not something you need care about. Let's look at two examples of studies where looking at the effect size, in addition to its statistical significance, would have improved the statistical inferences.

## 6.2 The Facebook experiment

In the summer of 2014 there were some concerns about an experiment that Facebook had performed on its users to examine 'emotional mood contagion', or the idea that people's moods can be influenced by the mood of people around them. You can read the article [here](#). For starters, there was substantial concern about the ethical aspects of the study, primarily because the researchers who performed the study had not asked for **informed consent** from the participants in the study, nor did they ask for permission from the **institutional review board** (or ethics committee) of their university.

One of the other criticisms of the study was that it could be dangerous to influence people's mood. As Nancy J. Smyth, dean of the University of Buffalo's School of Social Work wrote on her [Social Work blog](#): "There might even have been increased self-harm episodes, out of control anger, or dare I say it, suicide attempts or suicides that resulted from the experimental manipulation. Did this experiment create harm? The problem is, we will never know, because the protections for human subjects were never put into place".

If this Facebook experiment had such a strong effect on people's mood that it made some people commit suicide who would otherwise not have committed suicide, this would obviously be problematic. So let us look at the effects the manipulation Facebook used had on people a bit more closely.

From the article, let's see what the researchers manipulated:

Two parallel experiments were conducted for positive and negative emotion: One in which exposure to friends' positive emotional content in their News Feed was reduced, and one in which exposure to negative emotional content in their News Feed was reduced. In these conditions, when a person loaded their News Feed, posts that contained emotional content of the relevant emotional valence, each emotional post had between a 10% and 90% chance (based on their User ID) of being omitted from their News Feed for that specific viewing.

Then what they measured:

For each experiment, two dependent variables were examined pertaining to emotionality expressed in people's own status updates: the percentage of all words

produced by a given person that was either positive or negative during the experimental period. In total, over 3 million posts were analyzed, containing over 122 million words, 4 million of which were positive (3.6%) and 1.8 million negative (1.6%).

And then what they found:

When positive posts were reduced in the News Feed, the percentage of positive words in people's status updates decreased by  $B = -0.1\%$  compared with control [ $t(310,044) = -5.63$ ,  $P < 0.001$ , Cohen's  $d = 0.02$ ], whereas the percentage of words that were negative increased by  $B = 0.04\%$  ( $t = 2.71$ ,  $P = 0.007$ ,  $d = 0.001$ ). Conversely, when negative posts were reduced, the percent of words that were negative decreased by  $B = -0.07\%$  [ $t(310,541) = -5.51$ ,  $P < 0.001$ ,  $d = 0.02$ ] and the percentage of words that were positive, conversely, increased by  $B = 0.06\%$  ( $t = 2.19$ ,  $P < 0.003$ ,  $d = 0.008$ ).

Here, we will focus on the negative effects of the Facebook study (so specifically, the increase in negative words people used) to get an idea of whether there is a risk of an increase in suicide rates. Even though apparently there was a negative effect, it is not easy to get an understanding about the size of the effect from the numbers as mentioned in the text. Moreover, the number of posts that the researchers analyzed was really large. With a large sample, it becomes important to check if the size of the effect is such that the finding is substantially interesting, because with large sample sizes even minute differences will turn out to be statistically significant (we will look at this in more detail below). For that, we need a better understanding of "effect sizes".

### 6.3 The Hungry Judges study

In Figure 6.1 we see a graphical representation of the proportion of favorable parole decisions that real-life judges are making as a function of the number of cases they process across the day in Figure 6.1. The study from which this plot is taken is mentioned in many popular science books as an example of a finding that shows that people do not always make rational decisions, but that "judicial rulings can be swayed by extraneous variables that should have no bearing on legal decisions" (Danziger et al., 2011). We see that early on in the day, judges start by giving about 65% of people parole, which basically means, "All right, you can go back into society." But then very quickly, the number of favorable decisions decreases to basically zero. After a quick break which, as the authors say, "may replenish mental resources by providing rest, improving mood, or by increasing glucose levels in the body" the parole decisions are back up at 65%, and then again quickly drop down to basically zero. They take another break, and the percentage of positive decisions goes back up to 65%, only to drop again over the course of the day.

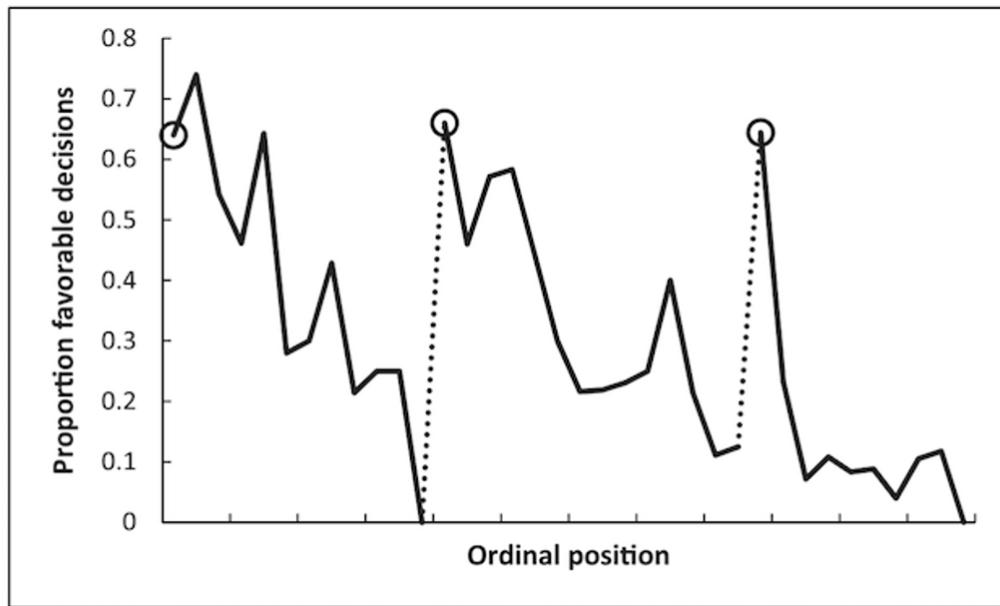


Figure 6.1: Proportion of rulings in favor of the prisoners by ordinal position. Circled points indicate the first decision in each of the three decision sessions; tick marks on x axis denote every third case; dotted line denotes food break. From Danziger, S., Levav, J., Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. Proceedings of the National Academy of Sciences, 108(17), 6889–6892. <https://doi.org/10.1073/PNAS.1018033108>

If we calculate the effect size for the drop after a break, and before the next break (Glöckner, 2016), the effect represents a Cohen's  $d$  of approximately 2, which is incredibly large. There are hardly any effects in psychology this large, let alone effects of mood or rest on decision making. And this surprisingly large effect occurs not just once, but three times over the course of the day. If mental depletion actually has such a huge real-life impact, society would basically fall into complete chaos just before lunch break every day. Or at the very least, our society would have organized itself around this incredibly strong effect of mental depletion. Just like manufacturers take size differences between men and women into account when producing items such as golf clubs or watches, we would stop teaching in the time before lunch, doctors would not schedule surgery, and driving before lunch would be illegal. If a psychological effect is this big, we don't need to discover it and publish it in a scientific journal - you would already know it exists.

We can look at a meta-meta-analysis (a paper that meta-analyzes a large number of meta-analyses in the literature) by Richard, Bond, & Stokes-Zoota (2003) to see which effect sizes in law psychology are close to a Cohen's  $d$  of 2. They report two meta-analyzed effects that are slightly smaller. The first is the effect that a jury's final verdict is likely to be the verdict a majority initially favored, which 13 studies show has an effect size of  $r = .63$ , or  $d = 1.62$ . The second is that when a jury is initially split on a verdict, its final verdict is likely to be lenient, which 13 studies show to have an effect size of  $r = .63$  as well. In their entire database, some effect sizes that come close to  $d = 2$  are the finding that personality traits are stable over time ( $r = .66$ ,  $d = 1.76$ ), people who deviate from a group are rejected from that group ( $r = .6$ ,  $d = 1.5$ ), or that leaders have charisma ( $r = .62$ ,  $d = 1.58$ ). You might notice the almost tautological nature of these effects. And that is, supposedly, the effect size that the passing of time (and subsequently eating lunch) has on parole hearing sentencing.

We see how examining the size of an effect can lead us to identify findings that cannot be caused by their proposed mechanisms. The effect reported in the hungry judges study must therefore be due to a confound. Indeed, such confounds have been identified, as it turns out the ordering of the cases is not random, and it is likely the cases that deserve parole are handled first, and the cases that do not deserve parole are handled later (Chatziathanasiou, 2022; Weinshall-Margel & Shapard, 2011). An additional use of effect sizes is to identify effect sizes that are too large to be plausible. Hilgard (2021) proposes to build in 'maximum positive controls', experimental conditions that show the largest possible effect in order to quantify the upper limit on plausible effect size measures.

## 6.4 Standardised Mean Differences

Conceptually, the  $d$  family effect sizes are based on a comparison between the difference between the observations, divided by the standard deviation of these observations. This means that a Cohen's  $d = 1$  means the standardized difference between two groups equals one standard deviation. The size of the effect in the Facebook study above was quantified with Cohen's

*d*. Cohen's *d* (the *d* is always italicized) is used to describe the standardized mean difference of an effect. This value can be used to compare effects across studies, even when the dependent variables are measured with different scales, for example when one study uses 7-point scales to measure the dependent variable, while the other study uses a 9-point scale. We can even compare effect sizes across completely different measures of the same construct, for example when one study uses a self-report measure, and another study uses a physiological measure. Although we can compare effect sizes across different measurements, this does not mean they are comparable, as we will discuss in more detail in the section on [heterogeneity](#) in the chapter on meta-analysis.

Cohen's *d* ranges from minus infinity to infinity (although in practice, the mean difference in the positive or negative direction that can be observed will never be infinite), with the value of 0 indicating that there is no effect. Cohen (1988) uses subscripts to distinguish different versions of *d*, a practice I will follow because it prevents confusion (without any specification, the term 'Cohen's *d*' denotes the entire family of effect sizes). Cohen refers to the standardized mean difference between two groups of independent observations for the *sample* as  $d_s$ . Before we get into the statistical details, let's first visualize what a Cohen's *d* of 0.001 (as was found in the Facebook study) means. We will use a visualization from <http://rpsychologist.com/d3/cohend/>, a website made by Kristoffer Magnusson, that allows you to visualize the differences between two measurements (such as the increase in negative words used by the Facebook user when the number of positive words on the timeline was reduced). The visualization actually shows two distributions, one dark blue and one light blue, but they overlap so much that the tiny difference in distributions is not visible (click the settings button to change the slider settings, and set the step size to 0.001 to reproduce the figure below in the online app).

The four numbers below the distribution express the effect size in different ways to facilitate the interpretation. For example, the **probability of superiority** expresses the probability that a randomly picked observation from one group will have a larger score than a randomly picked observation from the other group. Because the effect is so small, this probability is 50.03% - which means that people in the experimental write almost the same number of positive or negative words as people in the control condition. The **number needed to treat** index illustrates that in the Facebook study a person needs to type 3,570 words before we will observe one additional negative word, compared to the control condition. I don't know how often you type this many words on Facebook, but I think we can agree that this effect is not noticeable on an individual level.

To understand how Cohen's *d* for two independent groups is calculated, let's first look at the formula for the *t*-statistic:

$$t = \frac{\bar{M}_1 - \bar{M}_2}{SD_{pooled} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

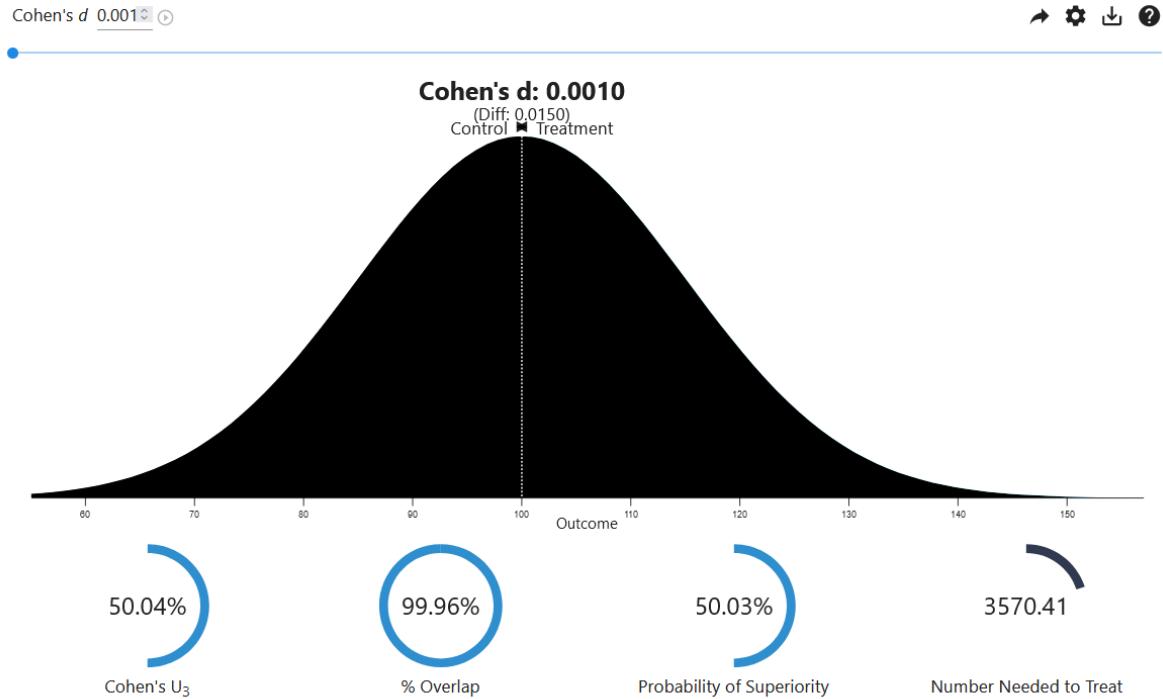


Figure 6.2: A visualization of 2 groups (although the difference is hardly visible) representing  $d = 0.001$ .

Here  $\overline{M}_1 - \overline{M}_2$  is the difference between the means, and  $SD_{pooled}$  is the pooled standard deviation (Daniël Lakens, 2013), and  $n_1$  and  $n_2$  are the sample sizes of the two groups that are being compared. The  $t$ -value is used to determine whether the difference between two groups in a  $t$ -test is statistically significant (as explained in the chapter on *p-values*). The formula for Cohen's  $d$  is very similar:

$$d_s = \frac{\overline{M}_1 - \overline{M}_2}{SD_{pooled}}$$

As you can see, the sample size in each group ( $n_1$  and  $n_2$ ) is part of the formula for a  $t$ -value, but it is not part of the formula for Cohen's  $d$  (the pooled standard deviation is computed by weighing the standard deviation in each group by the sample size, but it cancels out if groups are of equal size). This distinction is useful to know, because it tells us that the  $t$ -value (and consequently, the  $p$ -value) is a function of the sample size, but Cohen's  $d$  is independent of the sample size. If there is a true effect (i.e., a non-zero effect size in the population) the  $t$ -value for a null hypothesis test against an effect of zero will on average become larger (and the  $p$ -value will become smaller) as the sample size increases. The effect size, however, will not increase or decrease, but will become more accurate, as the standard error decreases as the sample size increases. This is also the reason why  $p$ -values cannot be used to make a statement about whether an effect is **practically significant**, and effect size estimates are often such an important complement to  $p$ -values when making statistical inferences.

You can calculate Cohen's  $d$  for independent groups from the independent samples  $t$ -value (which can often be convenient when the results section of the paper you are reading does not report effect sizes):

$$d_s = t \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

A  $d$  of 0.001 is an extremely tiny effect, so let's explore an effect size that is a bit more representative of what you would read in the literature. In the meta-meta-analysis mentioned earlier, the median effect size in published studies included in meta-analyses in the psychological literature is  $d = 0.43$  (Richard et al., 2003). To get a feeling for this effect size, let's use the online app and set the effect size to  $d = 0.43$ .

One example of a meta-analytic effect size in the meta-meta-analysis that is exactly  $d_s = 0.43$  is the finding that people in a group work less hard to achieve a goal than people who work individually, a phenomenon called *social loafing*. This is an effect that is large enough that we notice it in daily life. Yet, if we look at the overlap in the two distributions, we see that the amount of effort that people put in overlaps considerably between the two conditions (in the case of social loafing, working individually versus working in a group). We see in Figure 6.3 that the **probability of superiority**, or the probability that if we randomly draw one person from the group condition and one person from the individual condition, the person working in

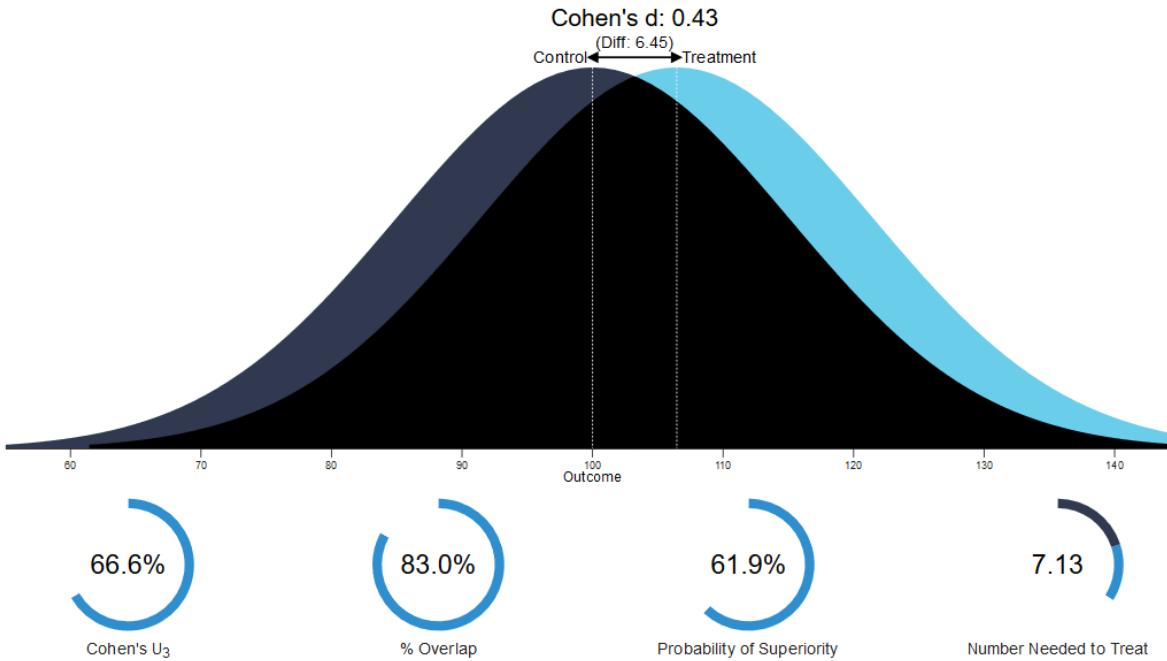


Figure 6.3: A visualization of 2 groups representing  $d = 0.43$ .

a group puts in less effort, is only 61.9%. This interpretation of differences between groups is also called the **common language effect size** (McGraw & Wong, 1992).

Based on [this data](#), the difference between the height of 21-year old men and women in The Netherlands is approximately 13 centimeters (in an unstandardized effect size), or a standardized effect size of  $d_s = 2$ . If I pick a random man and a random woman walking down the street in my hometown of Rotterdam, how likely is it that the man will be taller than the woman? We see this is quite likely, with a probability of superiority of 92.1%. But even with such a huge effect, there is still considerable overlap in the two distributions. If we conclude that the height of people in one group is greater than the height of people in another group, this does not mean that everyone in one group is taller than everyone in the other group.

Sometimes when you try to explain scientific findings at a birthday party, a skeptical aunt or uncle might remark ‘well I don’t believe that is true because *I* never experience this’. With probabilistic observations, there is a distribution of observed effects. In the example about social loafing, *on average* people put in less effort to achieve a goal when working in a group than working by themselves. For any individual in the population, the effect might be larger, smaller, absent, or even in the opposite direction. If your skeptical aunt or uncle never experiences a particular phenomenon, this does not contradict the claim that the effect exists *on average* in the population. Indeed, it is even expected that there will be no effect for some people in the population, at least some of the time. Although there might be some exceptions (e.g., almost every individual will experience the [Stroop effect](#)), many effects are smaller, or

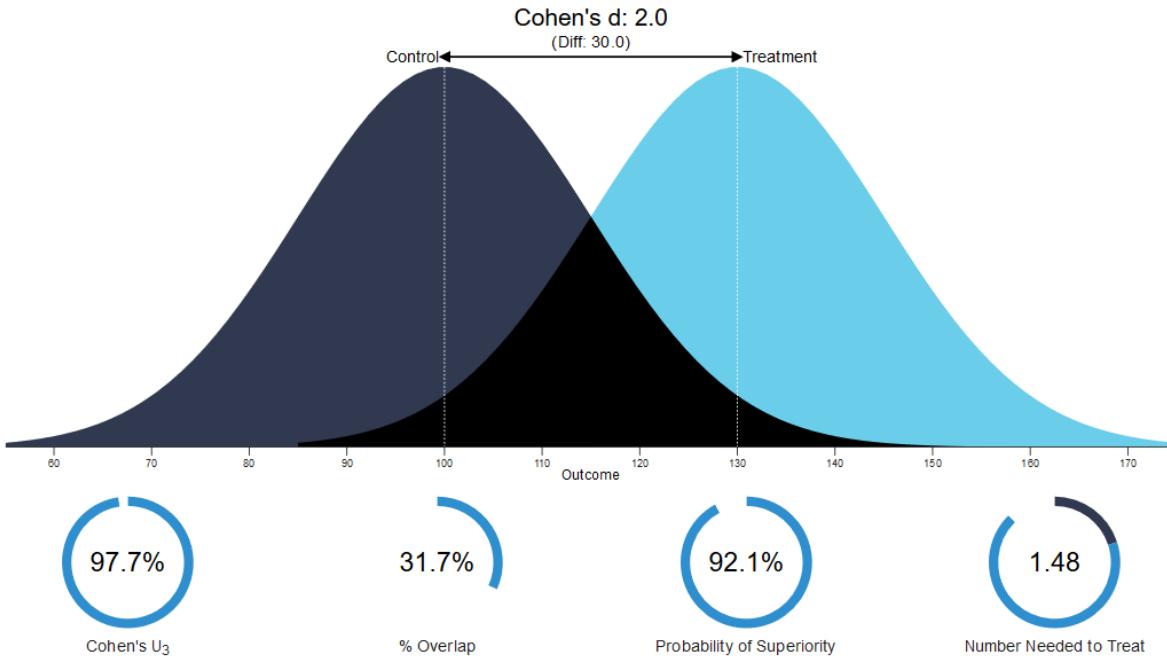


Figure 6.4: A visualization of 2 groups representing  $d = 2$ .

have sufficient variation, such that the effect is not present for every single individual in the population.

Conceptually, calculating Cohen's  $d$  for within-subjects comparisons is based on the same idea as for independent groups, where the differences between two observations are divided by the standard deviation within the groups of observations. However, in the case of correlated samples the most common standardizer is the standard deviation of the difference scores. Testing whether two correlated means are significantly different from each other with a paired samples  $t$ -test is the same as testing whether the difference scores of the correlated means is significantly different from 0 in a one-sample  $t$ -test. Similarly, calculating the effect size for the difference between two correlated means is similar to the effect size that is calculated for a one sample  $t$ -test. The standardized mean difference effect size for within-subjects designs is referred to as Cohen's  $d_z$ , where the  $z$  alludes to the fact that the unit of analysis is no longer  $x$  or  $y$ , but their difference,  $z$ , and can be calculated with:

$$d_z = \frac{M_{dif}}{\sqrt{\frac{\sum (X_{dif} - M_{dif})^2}{N-1}}}$$

The effect size estimate Cohen's  $d_z$  can also be calculated directly from the  $t$ -value and the number of participants using the formula:

$$d_z = \frac{t}{\sqrt{n}}$$

Given the direct relationship between the  $t$ -value of a paired-samples  $t$ -test and Cohen's  $d_z$ , it is not surprising that software that performs power analyses for within-subjects designs (e.g., G\*Power) relies on Cohen's  $d_z$  as input.

Maxwell & Delaney (2004) remark: ‘a major goal of developing effect size measures is to provide a standard metric that meta-analysts and others can interpret across studies that vary in their dependent variables as well as types of designs.’ Because Cohen's  $d_z$  takes the correlation between the dependent measures into account, it cannot be directly compared with Cohen's  $d_s$ . Some researchers prefer to use the average standard deviation of both groups of observations as a standardizer (which ignores the correlation between the observations), because this allows for a more direct comparison with Cohen's  $d_s$ . This effect size is referred to as Cohen's  $d_{av}$  (Cumming, 2013), and is simply:

$$d_{av} = \frac{M_{dif}}{\frac{SD_1+SD_2}{2}}$$

## 6.5 Interpreting effect sizes

A commonly used interpretation of Cohen's  $d$  is to refer to effect sizes as small ( $d = 0.2$ ), medium ( $d = 0.5$ ), and large ( $d = 0.8$ ) based on benchmarks suggested by Cohen (1988). However, these values are arbitrary and should not be used. In practice, you will only see them used in a form of circular reasoning: The effect is small, because it is  $d = 0.2$ , and  $d = 0.2$  is small. We see that using the benchmarks adds nothing, beyond covering up the fact that we did not actually interpret the size of the effect. Furthermore, benchmarks for what is a ‘medium’ and ‘large’ effect do not even correspond between Cohen's  $d$  and  $r$  (as explained by McGrath & Meyer (2006); see the ‘Test Yourself’ Q12). Any verbal classification based on benchmarks ignores the fact that any effect can be practically meaningful, such as an intervention that leads to a reliable reduction in suicide rates with an effect size of  $d = 0.1$ . In other cases, an effect size of  $d = 0.1$  might have no consequence at all, for example because such an effect is smaller than the just noticeable difference, and is therefore too small to be noticed by individuals in the real world.

Psychologists rely primarily on standardized effect sizes, where difference scores are divided by the standard deviation. Standardized effect sizes are convenient to compare effects across studies with different measures, and to combine effects in meta-analyses Daniël Lakens (2013). However, standardized effect metrics hinder the meaningful interpretation of effects in psychology, as they can reflect either a difference in means, or a difference in standard deviation, or any combination of the two. As an illustrative example, the ratio effect reveals that people find it easier to indicate which of two numbers represents the larger quantity when the ratio

between the numbers is large (e.g., 2 vs. 8) than small (e.g., 4 vs. 5). These numerical comparisons tasks have been used to study the development of numerical processing in children. As noted by Lyons and colleagues (2015), and illustrated in Figure 6.5, the average effect in raw scores (reaction times in milliseconds) declines over grades (pane a). But because the variability declines even more (pane b), the standardized effect size shows the opposite pattern than the raw effect size (pane c). Not surprisingly given this conflicting pattern, the authors ask: “Hence the real question: what is a meaningful effect-size?” (Lyons et al., 2015, p. 1032).

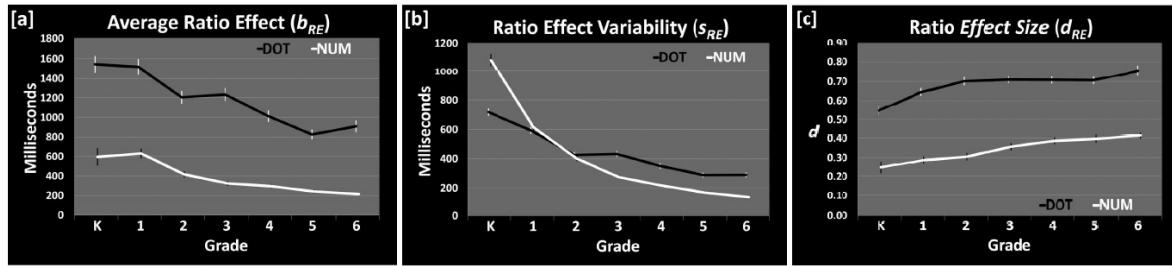


Figure 6.5: Figure from Lyons et al (2015) plotting mean differences (a), variance (b), and standardized effect size (c).

Publication bias and flexibility in the data analysis inflate effect size estimates. Innovations such as **Registered Reports** (Chambers & Tzavella, 2022; Nosek & Lakens, 2014) increasingly lead to the availability of unbiased effect size estimates in the scientific literature. Registered Reports are scientific publications which have been reviewed before the data has been collected based on the introduction, method, and proposed statistical analysis plan, and published regardless of whether the results were statistically significant or not. One consequence of no longer selectively publishing significant studies is that many effect sizes will turn out to be smaller than researchers thought. For example, in the 100 replication studies performed in the Reproducibility Project: Psychology, observed effect sizes in replication studies were on average half the size of those observed in the original studies (Open Science Collaboration, 2015).

To not just *report* but *interpret* an effect size, nothing is gained by the common practice of finding the corresponding verbal label of ‘small’, ‘medium’, or ‘large’. Instead, researchers who want to argue that an effect is meaningful need to provide empirical and falsifiable arguments for such a claim (Anvari et al., 2021; Primbs et al., 2022). One approach to argue that effect sizes are meaningful is by explicitly specifying a **smallest effect size of interest**, for example based on a cost-benefit analysis. Alternatively, researchers can interpret effect sizes relative to other effects in the literature (Baguley, 2009; Funder & Ozer, 2019).

## 6.6 Correlations and Variance Explained

The  $r$  family effect sizes are based on the proportion of variance that is explained by group membership (e.g., a correlation of  $r = 0.5$  indicates 25% of the variance ( $r^2$ ) is explained by the difference between groups). You might remember that  $r$  is used to refer to a correlation. The correlation of two continuous variables can range from 0 (completely unrelated) to 1 (perfect positive relationship) or -1 (perfect negative relationship). To get a better feel of correlations, play the online game [guess the correlation](#) where you will see a scatterplot, and have to guess the correlation between the variables (see Figure 6.6).

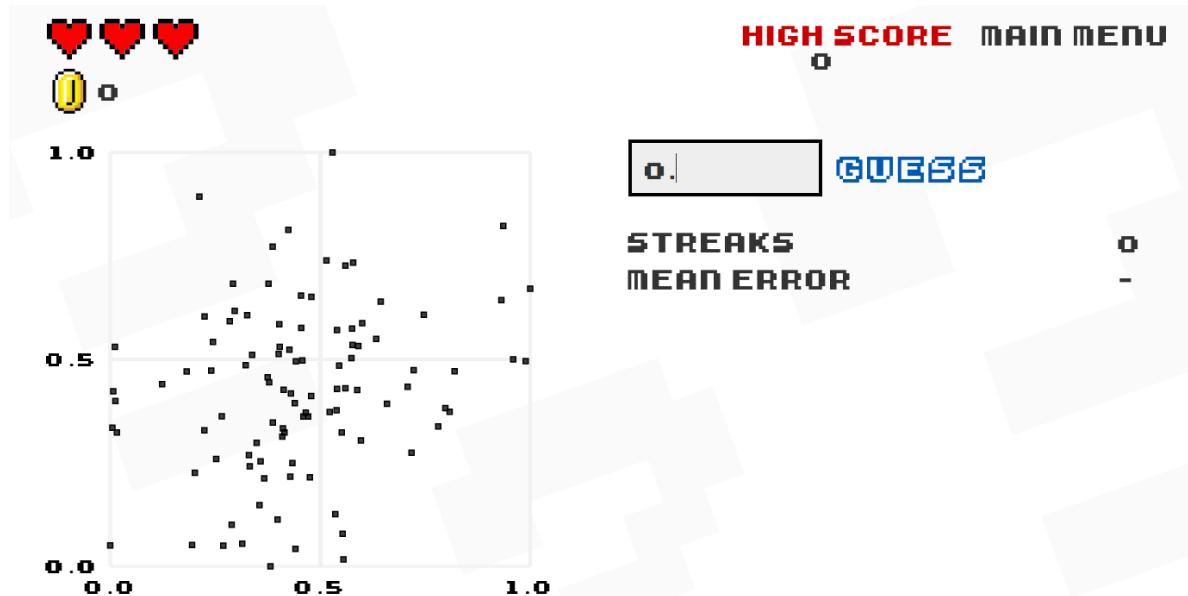


Figure 6.6: Screenshot from Guess the Correlation game (the correct answer is  $r = 0.24$ ).

The  $r$  family effect sizes are calculated from the sum of squares (the difference between individual observations and the mean for the group, squared and summed) for the effect, divided by the sums of squares for other factors in the design. Earlier, I mentioned that the median effect size in psychology is  $d_s = 0.43$ . However, the authors actually report their results as a correlation,  $r = 0.21$ . We can convert Cohen's  $d$  into  $r$  (but take care that this only applies to  $d_s$ , not  $d_z$ ):

$$r = \frac{d_s}{\sqrt{d_s^2 + \frac{N^2 - 2N}{n_1 \times n_2}}}$$

$N$  is the total sample size of both groups, whereas  $n_1$  and  $n_2$  are the sample sizes of the individual groups you are comparing (it is common to use capital  $N$  for the total sample size,

and lowercase  $n$  for sample sizes per group). You can go to <http://rpsychologist.com/d3/correlation/> to look at a good visualization of the proportion of variance that is explained by group membership, and the relationship between  $r$  and  $r^2$ . The amount of variance explained is often quite a small number, and we see in Figure 6.7 that a correlation of 0.21 (the median from the meta-meta-analysis by Richard and colleagues) we see the proportion of variance explained is only 4.4%. Funder and Ozer (2019) warn against misinterpreting small values for the variance explained as an indication that the effect is not meaningful (and they even consider the practice of squaring the correlation to be “actively misleading”).

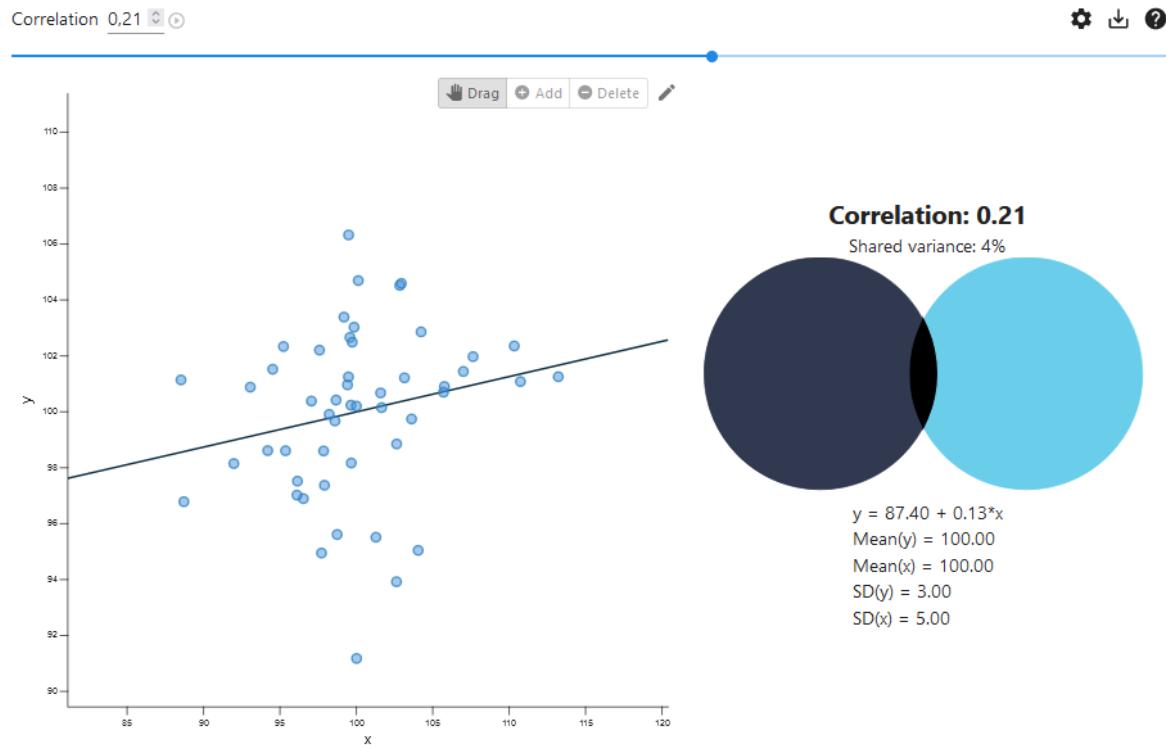


Figure 6.7: Screenshot from correlation effect size visualization by Kristoffer Magnusson for  $r = 0.21$ .

As we have seen before, it can be useful to interpret effect sizes to identify effects that are practically insignificant, or those that are implausibly large. Let's take a look at a study that examines the number of suicides as a function of the amount of country music played on the radio. You can find the paper [here](#). It won an Ig Nobel prize for studies that first make you laugh, and then think, although in this case, the the study should not make you think about

country music, but about the importance of interpreting effect sizes.

The authors predicted the following:

We contend that the themes found in country music foster a suicidal mood among people already at risk of suicide and that it is thereby associated with a high suicide rate.

Then they collected data:

Our sample is comprised of 49 large metropolitan areas for which data on music were available. Exposure to country music is measured as the proportion of radio airtime devoted to country music. Suicide data were extracted from the annual Mortality Tapes, obtained from the Inter-University Consortium for Political and Social Research (ICPSR) at the University of Michigan. The dependent variable is the number of suicides per 100,000 population.

And they concluded:

A significant zero-order correlation was found between white suicide rates and country music ( $r = .54$ ,  $p < .05$ ). The greater the airtime given to country music, the greater the white suicide rate.

We can again compare the size of this effect with other known effects in psychology. In the database by Richard and colleagues, there are very few effects this large, but some examples are: that leaders are most effective if they have charisma ( $r = 0.54$ ), good leader–subordinate relations promote subordinate satisfaction ( $r = 0.53$ ), and people can recognize emotions across cultures ( $r = 0.53$ ). These effects are all large and obvious, which should raise some doubts about whether the relationship between listening to country music and suicides can be of the same size. Is country music really that bad? If we search the literature, we find that [other researchers were not able to reproduce the analysis of the original authors](#). It is possible that the results are either spurious, or a Type 1 error.

Eta squared, written  $\eta^2$  (part of the  $r$  family of effect sizes, and an extension of  $r$  that can be used for more than two sets of observations) measures the proportion of the variation in  $Y$  that is associated with membership of the different groups defined by  $X$ , or the sum of squares of the effect divided by the total sum of squares:

$$\eta^2 = \frac{SS_{\text{effect}}}{SS_{\text{total}}}$$

An  $\eta^2$  of .13 means that 13% of the total variance can be accounted for by group membership. Although  $\eta^2$  is an efficient way to compare the sizes of effects within a study (given that every effect is interpreted in relation to the total variance, all  $\eta^2$  from a single study sum to 100%), eta squared cannot easily be compared between studies, because the total variability in a

study ( $SS_{total}$ ) depends on the design of a study, and increases when additional variables are manipulated (e.g., when independent variables are added). Keppel (1991) has recommended partial eta squared ( $\eta_p^2$ ) to improve the comparability of effect sizes between studies.  $\eta_p^2$  expresses the sum of squares of the effect in relation to the sum of squares of the effect plus the sum of squares of the error associated with the effect. Partial eta squared is calculated as:

$$\eta_p^2 = \frac{SS_{\text{effect}}}{SS_{\text{effect}} + SS_{\text{error}}}$$

For designs with fixed factors (manipulated factors, or factors that exhaust all levels of the independent variable, such as alive vs. dead), but not for designs with measured factors or covariates, partial eta squared can be computed from the  $F$ -value and its degrees of freedom (Cohen, 1988):

$$\eta_p^2 = \frac{F \times df_{\text{effect}}}{F \times df_{\text{effect}} + df_{\text{error}}}$$

For example, for an  $F(1, 38) = 7.21$ ,  $\eta_p^2 = 7.21 \times 1 / (7.21 + 1 + 38) = 0.16$ .

Eta squared can be transformed into Cohen's  $d$ :

$$d = 2 \times f \text{ where } f^2 = \eta^2 / (1 - \eta^2)$$

## 6.7 Correcting for Bias

Population effect sizes are almost always estimated on the basis of samples, and as a measure of the population effect size estimate based on sample averages, Cohen's  $d$  slightly overestimates the true population effect. When Cohen's  $d$  refers to the population, the Greek letter  $\delta$  is typically used. Therefore, corrections for bias are used (even though these corrections do not always lead to a completely unbiased effect size estimate). In the  $d$  family of effect sizes, the correction for bias in the population effect size estimate of Cohen's  $d$  is known as Hedges'  $g$  (although different people use different names –  $d_{unbiased}$  is also used). This correction for bias is only noticeable in small sample sizes, but since we often use software to calculate effect sizes anyway, it makes sense to always report Hedges'  $g$  instead of Cohen's  $d$  (Thompson, 2007).

As with Cohen's  $d$ ,  $\eta^2$  is a biased estimate of the true effect size in the population. Two less biased effect size estimates have been proposed, namely epsilon squared  $\varepsilon^2$  and omega squared  $\omega^2$ . For all practical purposes, these two effect sizes correct for bias equally well (Albers & Lakens, 2018; Okada, 2013), and should be preferred above  $\eta^2$ . Partial epsilon squared ( $\varepsilon_p^2$ ) and partial omega squared ( $\omega_p^2$ ) can be calculated based on the  $F$ -value and degrees of freedom.

$$\omega_p^2 = \frac{F - 1}{F + \frac{\text{df}_{\text{error}} + 1}{\text{df}_{\text{effect}}}}$$

$$\varepsilon_p^2 = \frac{F - 1}{F + \frac{\text{df}_{\text{error}}}{\text{df}_{\text{effect}}}}$$

The partial effect sizes  $\eta_p^2$ ,  $\varepsilon_p^2$  and  $\omega_p^2$  cannot be generalized across different designs. For this reason, generalized eta-squared ( $\eta_G^2$ ) and generalized omega-squared ( $\omega_G^2$ ) have been proposed (Olejnik & Algina, 2003), although they are not very popular. In part, this might be because summarizing the effect size in an ANOVA design with a single index has limitations, and perhaps it makes more sense to describe the pattern of results, as we will see in the section below.

## 6.8 Effect Sizes for Interactions

The effect size used for power analyses for ANOVA designs is Cohen's  $f$ . For two independent groups, Cohen's  $f = 0.5 * \text{Cohen's } d$ . For more than two groups, Cohen's  $f$  can be converted into eta-squared and back with  $f = \frac{\eta^2}{(1-\eta^2)}$  or  $\eta^2 = \frac{f^2}{(1+f^2)}$ . When predicting interaction effects in ANOVA designs, planning the study based on an expected effect size such as  $\eta_p^2$  or Cohen's  $f$  might not be the most intuitive approach.

Let's start with the effect size for a simple two group comparison, and assume we have observed a mean difference of 1, and a standard deviation of 2. This means that the standardized effect size is  $d = 0.5$ . An independent  $t$ -test is mathematically identical to an  $F$ -test with two groups. For an  $F$ -test, the effect size used for power analyses is Cohen's  $f$ , which is calculated based on the standard deviation of the population means divided by the population standard deviation (which we know for our measure is 2), or:

$$f = \frac{\sigma_m}{\sigma} \tag{6.1}$$

where for equal sample sizes

$$\sigma_m = \sqrt{\frac{\sum_{i=1}^k (m_i - m)^2}{k}}. \tag{6.2}$$

In this formula  $m$  is the grand mean,  $k$  is the number of means, and  $m_i$  is the mean in each group. The formula above might look a bit daunting, but calculating Cohen's  $f$  is not that difficult for two groups.

If we take the means of 0 and 1, and a standard deviation of 2, the grand mean (the  $m$  in the formula above) is  $(0 + 1)/2 = 0.5$ . The formula says we should subtract this grand mean from the mean of each group, square this value, and sum them. So we have  $(0 - 0.5)^2$  and  $(1$

$- 0.5)^2$ , which are both 0.25. We sum these values ( $0.25 + 0.25 = 0.5$ ), divide them by the number of groups ( $0.5/2 = 0.25$ ), and take the square root, we find that  $\sigma_m = 0.5$ . We can now calculate Cohen's  $f$  (using  $\sigma = 2$  for our measure):

$$f = \frac{\sigma_m}{\sigma} = \frac{0.5}{2} = 0.25 \quad (6.3)$$

We confirm that for two groups Cohen's  $f$  is half as large as Cohen's  $d$ .

Now we have the basis to look at interaction effects. Different patterns of means in an ANOVA can have the same Cohen's  $f$ . There are two types of interactions, as visualized below in Figure 6.8. In an **ordinal interaction**, the mean of one group ("B1") is always higher than the mean for the other group ("B2"). **Disordinal interactions** are also known as 'cross-over' interactions, and occur when the group with the larger mean changes between conditions. The difference is important, since the disordinal interaction in Figure 6.8 has a larger effect size than the ordinal interaction.

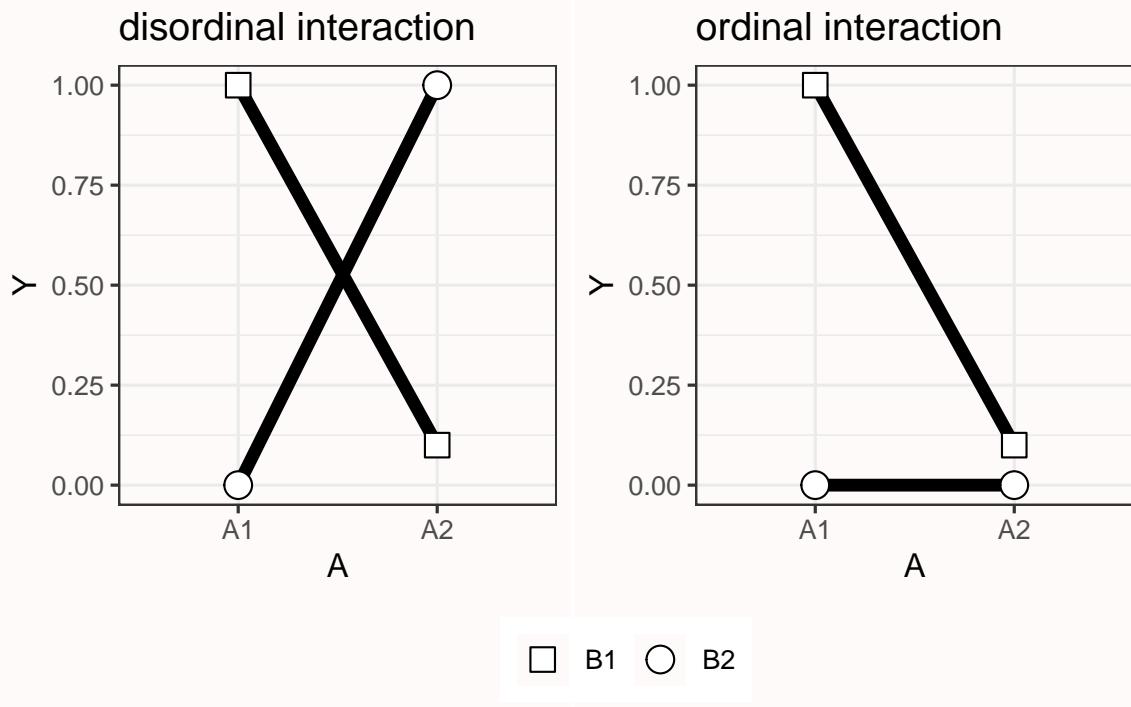


Figure 6.8: Schematic illustration of a disordinal (or cross-over) and ordinal interaction.

Mathematically the interaction effect is computed as the cell mean minus the sum of the grand mean, the marginal mean in each condition of one factor minus the grand mean, and

the marginal mean in each condition for the other factor minus grand mean (Maxwell & Delaney, 2004).

Let's consider two cases, one where we have a perfect disordinal interaction (the means of 0 and 1 flip around in the other condition, and are 1 and 0) or an ordinal interaction (the effect is present in one condition, with means 0 and 1, but disappears in the other condition, with means 0 and 0; see Figure 6.9).

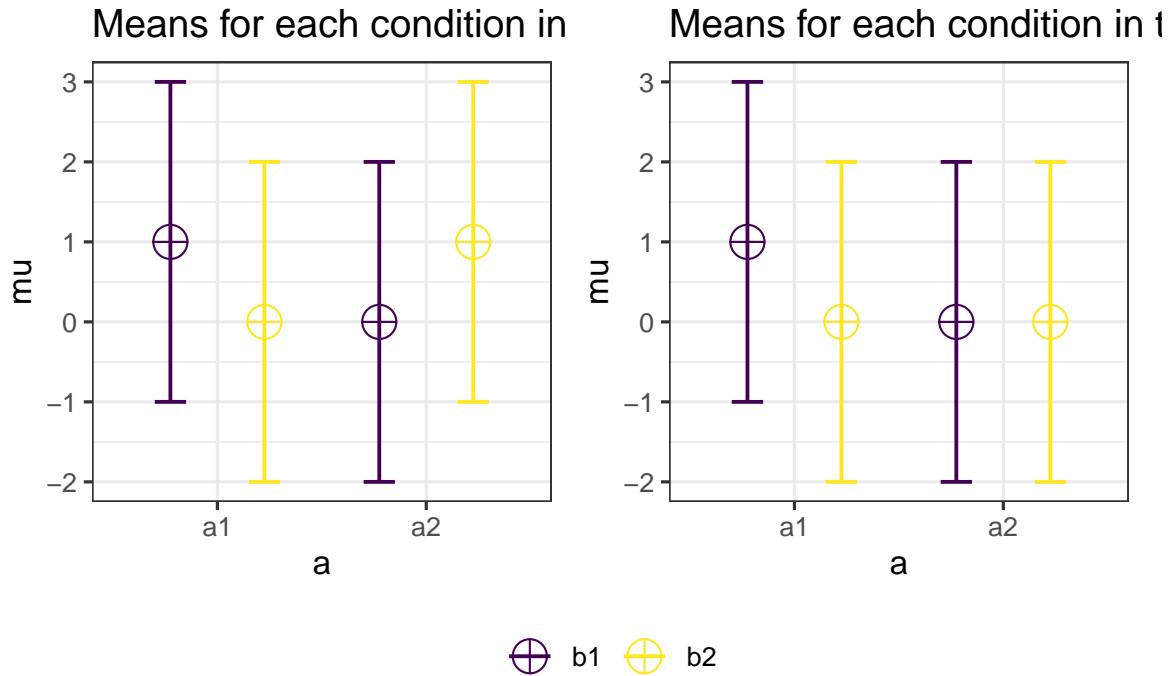


Figure 6.9: Disordinal (or cross-over) and ordinal interaction with means of 0 and 1,  $n = 50$  per group, and an  $SD$  of 2.

We can calculate the interaction effect as follows (we will go through the steps in some detail). First, let's look at the disordinal interaction. The grand mean is  $(1 + 0 + 0 + 1) / 4 = 0.5$ .

We can compute the marginal means for A1, A2, B1, and B2, which is simply averaging per row and column, which gets us for the A1 row  $(1+0)/2=0.5$ . For this perfect disordinal interaction, all marginal means are 0.5. This means there are no main effects. There is no main effect of factor A (because the marginal means for A1 and A2 are both exactly 0.5), nor is there a main effect of B.

We can also calculate the interaction effect. For each cell we take the value in the cell (e.g., for a1b1 this is 1) and compute the difference between the cell mean and the additive effect of the two factors as:

$1 \cdot 2013$  (the grand mean of  $0.5 + (\text{the marginal mean of } a_1 \text{ minus the grand mean, or } 0.5 - 0.5 = 0) + (\text{the marginal mean of } b_1 \text{ minus the grand mean, or } 0.5 \cdot 2013 \cdot 0.5 = 0)$ ). Thus, for each cell we get:

$$a_1 b_1: 1 \cdot 2013 (0.5 + (0.5 \cdot 2013 \cdot 0.5) + (0.5 \cdot 2013 \cdot 0.5)) = 0.5$$

$$a_1 b_2: 0 \cdot 2013 (0.5 + (0.5 \cdot 2013 \cdot 0.5) + (0.5 \cdot 2013 \cdot 0.5)) = -0.5$$

$$a_2 b_1: 0 - (0.5 + (0.5 \cdot 2013 \cdot 0.5) + (0.5 \cdot 2013 \cdot 0.5)) = -0.5$$

$$a_2 b_2: 1 - (0.5 + (0.5 \cdot 2013 \cdot 0.5) + (0.5 \cdot 2013 \cdot 0.5)) = 0.5$$

$$\text{Cohen's } f \text{ is then } f = \sqrt{\frac{0.5^2 + (-0.5^2) + (-0.5^2) + 0.5^2}{4}} = 0.25$$

For the ordinal interaction the grand mean is  $(1 + 0 + 0 + 0) / 4$ , or 0.25. The marginal means are  $a_1: 0.5$ ,  $a_2: 0$ ,  $b_1: 0.5$ , and  $b_2: 0$ .

Completing the calculation for all four cells for the ordinal interaction gives:

$$a_1 b_1: 1 \cdot 2013 (0.25 + (0.5 \cdot 2013 \cdot 0.25) + (0.5 \cdot 2013 \cdot 0.25)) = 0.25$$

$$a_1 b_2: 0 \cdot 2013 (0.25 + (0.5 \cdot 2013 \cdot 0.25) + (0.0 \cdot 2013 \cdot 0.25)) = -0.25$$

$$a_2 b_1: 0 \cdot 2013 (0.25 + (0.0 \cdot 2013 \cdot 0.25) + (0.5 \cdot 2013 \cdot 0.25)) = -0.25$$

$$a_2 b_2: 0 \cdot 2013 (0.25 + (0.0 \cdot 2013 \cdot 0.25) + (0.0 \cdot 2013 \cdot 0.25)) = 0.25$$

$$\text{Cohen's } f \text{ is then } f = \sqrt{\frac{0.25^2 + (-0.25^2) + (-0.25^2) + 0.25^2}{4}} = 0.125.$$

We see the effect size of the cross-over interaction ( $f = 0.25$ ) is twice as large as the effect size of the ordinal interaction ( $f = 0.125$ ). This should make sense if we think about the interaction as a test of contrasts. In the disordinal interaction we are comparing cells  $a_1 b_1$  and  $a_2 b_2$  against  $a_1 b_2$  and  $a_2 b_1$ , or  $(1+1)/2$  vs.  $(0+0)/2$ . Thus, if we see this as a  $t$ -test for a contrast, we see the mean difference is 1. For the ordinal interaction, we have  $(1+0)/2$  vs.  $(0+0)/2$ , so the mean difference is halved, namely 0.5. This obviously matters for the statistical power we will have when we examine interaction effects in our experiments.

Just stating that you expect a ‘medium’ Cohen’s  $f$  effect size for an interaction effect in your power analysis is not the best approach. Instead, start by thinking about the pattern of means and standard deviations (and for within factors, the correlation between dependent variables) and then compute the effect size from the data pattern. If you prefer not to do so by hand, you can use [Superpower](#) (Daniël Lakens & Caldwell, 2021). This also holds for more complex designs, such as multilevel models. In these cases, it is often the case that power analyses are easier to perform with simulation-based approaches, than based on plugging a single effect size in to power analysis software (DeBruine & Barr, 2021).

## 6.9 Why Effect Sizes Selected for Significance are Inflated

Another way to think about this is through the concept of a **truncated distribution**. If effect sizes are only reported if the  $p$ -value is statistically significant, then we only have access to effect sizes that are larger than some minimal value (S. F. Anderson et al., 2017; Taylor & Muller, 1996). In Figure 6.10 only effects larger than  $d = 0.4$  can be significant, so all effect sizes below this threshold are censored, and only effect sizes in the gray part of the distribution will be available to researchers. Without the lower part of the effect size distribution effect sizes will on average be inflated.

Estimates based on samples from the population will show variability. The larger the sample, the closer our estimates will be to the true population values, as explained in the next chapter on [confidence intervals](#). Sometimes we will observe larger estimates than the population value, and sometimes we will observe smaller values. As long as we have an unbiased collection of effect size estimates, combining effect sizes estimates through a meta-analysis can increase the accuracy of the estimate. Regrettably, the scientific literature is often biased. It is specifically common that statistically significant studies are published (e.g., studies with  $p$  values smaller than 0.05) while studies with  $p$  values larger than 0.05 remain unpublished (Franco et al., 2014; Sterling, 1959). Instead of having access to all effect sizes, anyone reading the literature only has access to effects that passed a significance filter. This will introduce systematic bias in our effect size estimates.

To explain how selection for significance introduces bias, it is useful to understand the concept of a truncated or censored distribution. If we want to measure the average length of people in The Netherlands we would collect a representative sample of individuals, measure how tall they are, and compute the average score. If we collect sufficient data the estimate will be close to the true value in the population. However, if we collect data from participants who are on a theme park ride where people need to be at least 150 centimeters tall to enter, the mean we compute is based on a truncated distribution where only individuals taller than 150 cm are included. Smaller individuals are missing. Imagine we have measured the height of two individuals in the theme park ride, and they are 164 and 184 cm tall. Their average height is  $(164+184)/2 = 174$  cm. Outside the entrance of the theme park ride is one individual who is 144 cm tall. Had we measured this individual as well, our estimate of the average length would be  $(144+164+184)/3 = 164$  cm. Removing low values from a distribution will lead to overestimation of the true value. Removing high values would lead to underestimation of the true value.

The scientific literature suffers from publication bias. Non-significant test results – based on whether a  $p$  value is smaller than 0.05 or not – are often less likely to be published. When an effect size estimate is 0 the  $p$  value is 1. The further removed effect sizes are from 0, the smaller the  $p$  value. All else equal (e.g., studies have the same sample size, and measures have the same distribution and variability) if results are selected for statistical significance (e.g.,  $p < .05$ ) they are also selected for larger effect sizes. As small effect sizes will be observed with their corresponding probabilities, their absence will inflate effect size estimates. Every study in

the scientific literature provides its own estimate of the true effect size, just as every individual provides its own estimate of the average height of people in a country. When these estimates are combined – as happens in [meta-analyses](#) in the scientific literature – the meta-analytic effect size estimate will be biased (or systematically different from the true population value) whenever the distribution is truncated. To achieve unbiased estimates of population values when combining individual studies in the scientific literature in meta-analyses researchers need access to the complete distribution of values – or all studies that are performed, regardless of whether they yielded a  $p$  value above or below 0.05.

In Figure 6.10 we see a distribution centered at an effect size of Cohen's  $d = 0.5$  for a two-sided  $t$ -test with 50 observations in each independent condition. Given an alpha level of 0.05 in this test only effect sizes larger than  $d = 0.4$  will be statistically significant (i.e., all observed effect sizes in the grey area). The threshold for which observed effect sizes will be statistically significant is determined by the sample size and the alpha level (and not influenced by the true effect size). The white area under the curve illustrates Type 2 errors – non-significant results that will be observed if the alternative hypothesis is true. If researchers only have access to the effect sizes estimates in the grey area – a truncated distribution where non-significant results are removed – a weighted average effect size from only these studies will be upwardly biased.

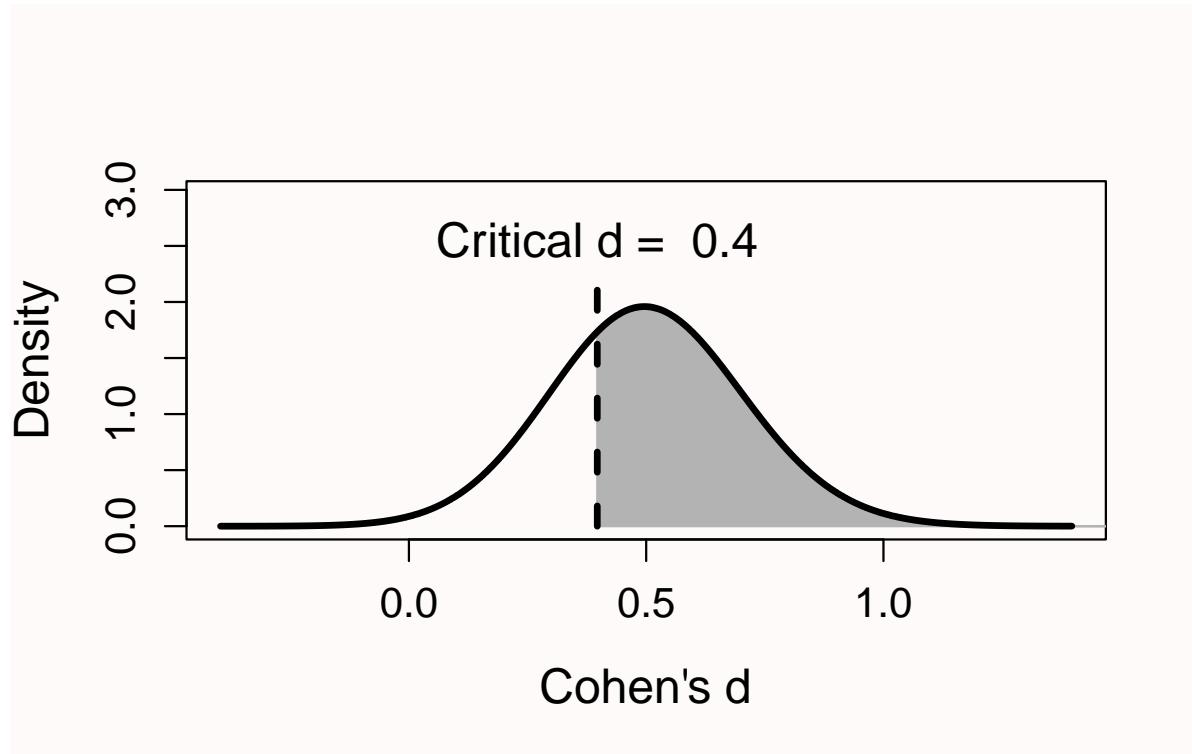


Figure 6.10: Range of effect sizes that will be statistically significant in an independent  $t$ -test with 50 participants per group and a true effect size of  $d = 0.5$ .

If researchers only have access to the effect sizes estimates in the grey area – a truncated distribution where non-significant results are removed – a weighted average effect size from only these studies will be upwardly biased. We can see this in the two forest plots visualizing meta-analyses in Figure 6.11. In the top meta-analysis all 5 studies are included, even though study C and D yield non-significant results (as can be seen from the fact that the 95% CI overlaps with 0). The estimated effect size based on all 5 studies is  $d = 0.4$ . In the bottom meta-analysis the two non-significant studies are removed - as would happen when there is publication bias. Without these two studies the estimated effect size in the meta-analysis,  $d = 0.5$ , is inflated. The extent to which meta-analyses are inflated depends on the true effect size and the sample size of the studies.

The inflation will be greater the larger the part of the distribution is truncated, and the closer the true population effect size is to 0. In our example about the height of individuals the inflation would be greater had we truncated the distribution by removing everyone smaller than 170 cm instead of 150 cm. If the true average height of individuals was 194 cm, removing the few people that are expected to be smaller than 150 (based on the assumption of normally distributed data) would have less of an effect on how much our estimate is inflated than when the true average height was 150 cm, in which case we would remove 50% of individuals. In statistical tests where results are selected for significance at a 5% alpha level more data will be removed if the true effect size is smaller, but also when the sample size is smaller. If the sample size is smaller, statistical power is lower, and more of the values in the distribution (those closest to 0) will be non-significant.

Any single estimate of a population value will vary around the true population value. The effect size estimate from a single study can be smaller than the true effect size, even if studies have been selected for significance. For example, it is possible that the true effect size is 0.5, you have observed an effect size of 0.45, but only effect sizes smaller than 0.4 are truncated when selecting studies based on statistical significance (as in the figure above). At the same time, this single effect size estimate of 0.45 is inflated. What inflates the effect size is the long-run procedure used to generate the value. In the long run effect sizes estimates based on a procedure where estimates are selected for significance will be upwardly biased. This means that a single observed effect size of  $d = 0.45$  will be inflated if it is generated based on a procedure where all non-significant effects are truncated, but it will be unbiased if it is generated based on a distribution where all observed effect sizes are reported, regardless of whether they are significant or not. This also means that a single researcher can not guarantee that the effect sizes they contribute to a literature will contribute to an unbiased effect sizes estimate: There needs to be a system in place where all researchers report all observed effect sizes to prevent bias. An alternative is to not have to rely on other researchers, and collect sufficient data in a single study to have a highly accurate effect size estimate. Multi-lab replication studies are an example of such an approach, where dozens of researchers collect a large number (up to thousands) of observations.

The most extreme consequence of the inflation of effect size estimates occurs when the true effect size in the population is 0, but due to selection of statistically significant results, only

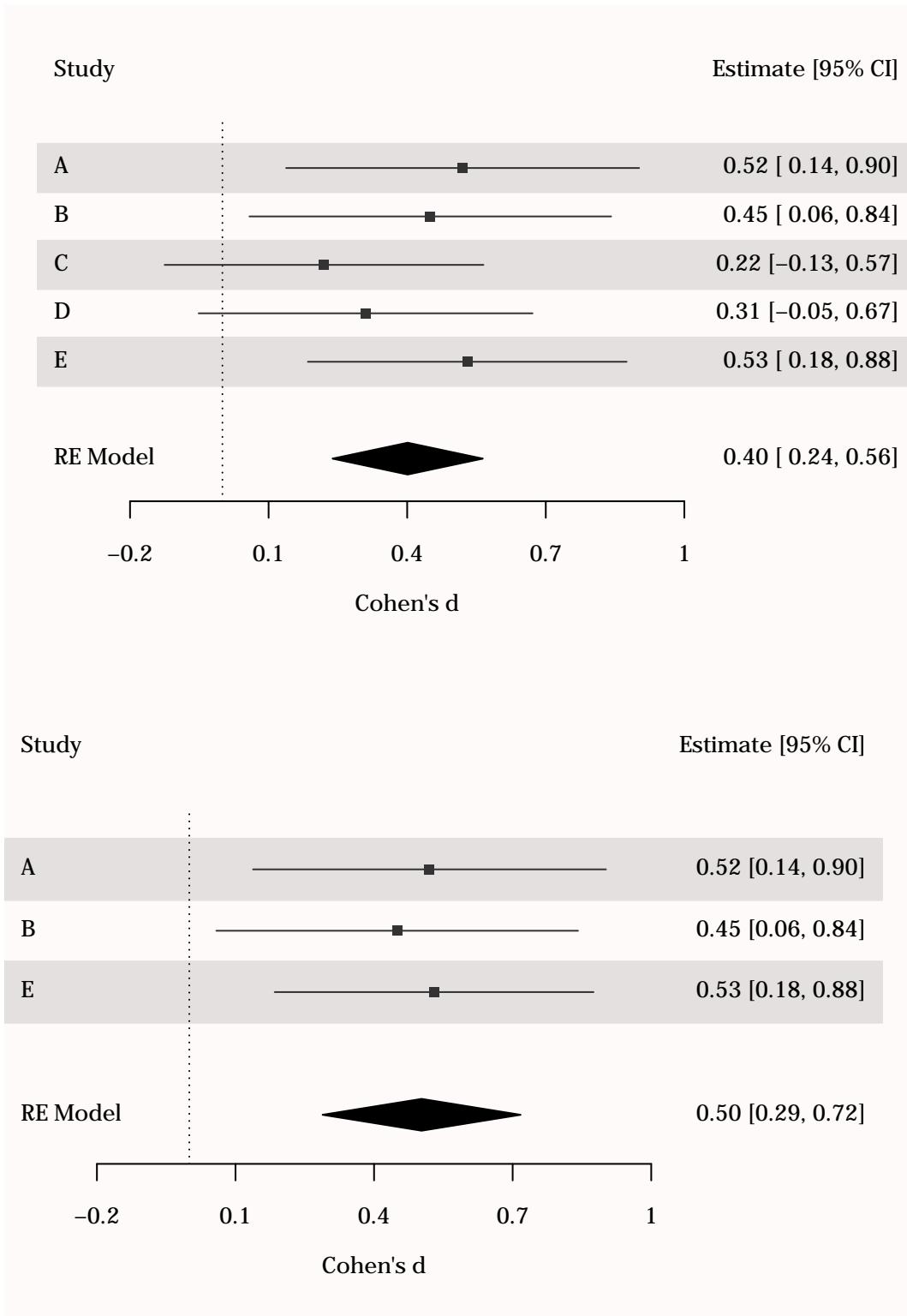


Figure 6.11: Two forest plots, the top with all studies reported, the bottom only reporting statistically significant effects.

significant effects in the expected direction are published. Note that if all significant results are published (and not only effect sizes in the expected direction) 2.5% of Type 1 error rates will be in the positive direction, and 2.5% will be in the negative direction, and the average effect size would be actually be 0. Thus, as long as the true effect size is exactly 0, and all Type 1 errors are published, the effect size estimate would be unbiased. In practice, we see scientists often do not simply publish all results, but only statistically significant results in the desired direction. An example of this is the literature on ego-depletion, where hundreds of studies were published, most showing statistically significant effects, but unbiased large scale replication studies revealed effect sizes of 0 (Hagger et al., 2016; Vohs et al., 2021).

What can be done about the problem of biased effect sizes estimates if we mainly have access to the studies that passed a significance filter? Statisticians have developed approaches to adjust biased effect size estimates by taking a truncated distribution into account (Taylor & Muller, 1996). This approach has recently been implemented in R (S. F. Anderson et al., 2017). Implementing this approach in practice is difficult, because we never know for sure if an effect size estimate is biased, and if it is biased, how much bias there is. Furthermore, selection based on significance is only one form of bias, whereas researchers who selectively report significant results may engage in additional problematic research practices, such as selectively reporting results, which are not accounted for in the adjustment. Nevertheless, it can be used as a more conservative approach to estimate effect sizes in a biased literature. Other researchers have referred to this problem as a Type M error (Gelman & Carlin, 2014) and have suggested that researchers always report the average inflation factor of effect sizes. I do not believe this approach is useful. The Type M error is not an error, but a bias in estimation, and it is more informative to compute the adjusted estimate based on a truncated distribution as proposed by Taylor and Muller (1996), than to compute the average inflation for a specific study design. If effects are on average inflated by a factor of 1.3 (the Type M error) it does not mean that the observed effect size is inflated by this factor, and the truncated effect sizes estimator by Taylor and Muller will provide researchers with an actual estimate based on their observed effect size. Type M errors might have a function in education, but they are not useful for scientists.

Of course the real solution to bias in effect size estimates due to significance filters that lead to truncated or censored distributions is to stop selectively reporting results. Designing highly informative studies that have high power to both reject the null, as a smallest effect size of interest in an equivalence test, is a good starting point. Publishing research as Registered Reports is even better. Eventually, if we do not solve this problem ourselves, it is likely that we will face external regulatory actions that force us to include all studies that have received ethical review board approval to a public registry, and update the registration with the effect size estimate, as is done for clinical trials.

## 6.10 The Minimal Statistically Detectable Effect

Given any alpha level and sample size it is possible to directly compute the **minimal statistically detectable effect**, or the **critical effect size**, which is the smallest effect size that, if observed, would be statistically significant given a specified alpha level and sample size (J. Cook et al., 2014). As explained in the previous section, if researchers selectively have access to only significant results, all effect sizes should be larger than the minimal statistically detectable effect, and the average effect size estimate will be upwardly inflated. For any critical  $t$  value (e.g.,  $t = 1.96$  for  $\alpha = 0.05$ , for large sample sizes) we can compute a critical mean difference (Phillips et al., 2001), or a critical standardized effect size. For a two-sided independent  $t$  test the critical mean difference is:

$$M_{crit} = t_{crit} \sqrt{\frac{sd_1^2}{n_1} + \frac{sd_2^2}{n_2}}$$

and the corresponding critical standardized mean difference is:

$$d_{crit} = t_{crit} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

G\*Power provides the critical test statistic (such as the critical  $t$  value) when performing a power analysis. For example, Figure 6.12 shows that for a correlation based on a two-sided test, with  $\alpha = 0.05$ , and  $N = 30$ , only effects larger than  $r = 0.361$  or smaller than  $r = -0.361$  can be statistically significant. This reveals that when the sample size is relatively small, the observed effect needs to be quite substantial to be statistically significant.

It is important to realize that due to random variation each study has a probability to yield effects larger than the critical effect size, even if the true effect size is small (or even when the true effect size is 0, in which case each significant effect is a Type I error). At the same time, researchers often do not want to perform an experiment where effects they are interested in can not even become statistically significant, which is why it can be useful to compute the minimal statistically significant effect as part of a [sample size justification](#).

## 6.11 Test Yourself

**Q1:** One of the largest effect sizes in the meta-meta analysis by Richard and colleagues from 2003 is that people are likely to perform an action if they feel positively about the action and believe it is common. Such an effect is (with all due respect to all of the researchers who contributed to this meta-analysis) somewhat trivial. Even so, the correlation was  $r = .66$ ,

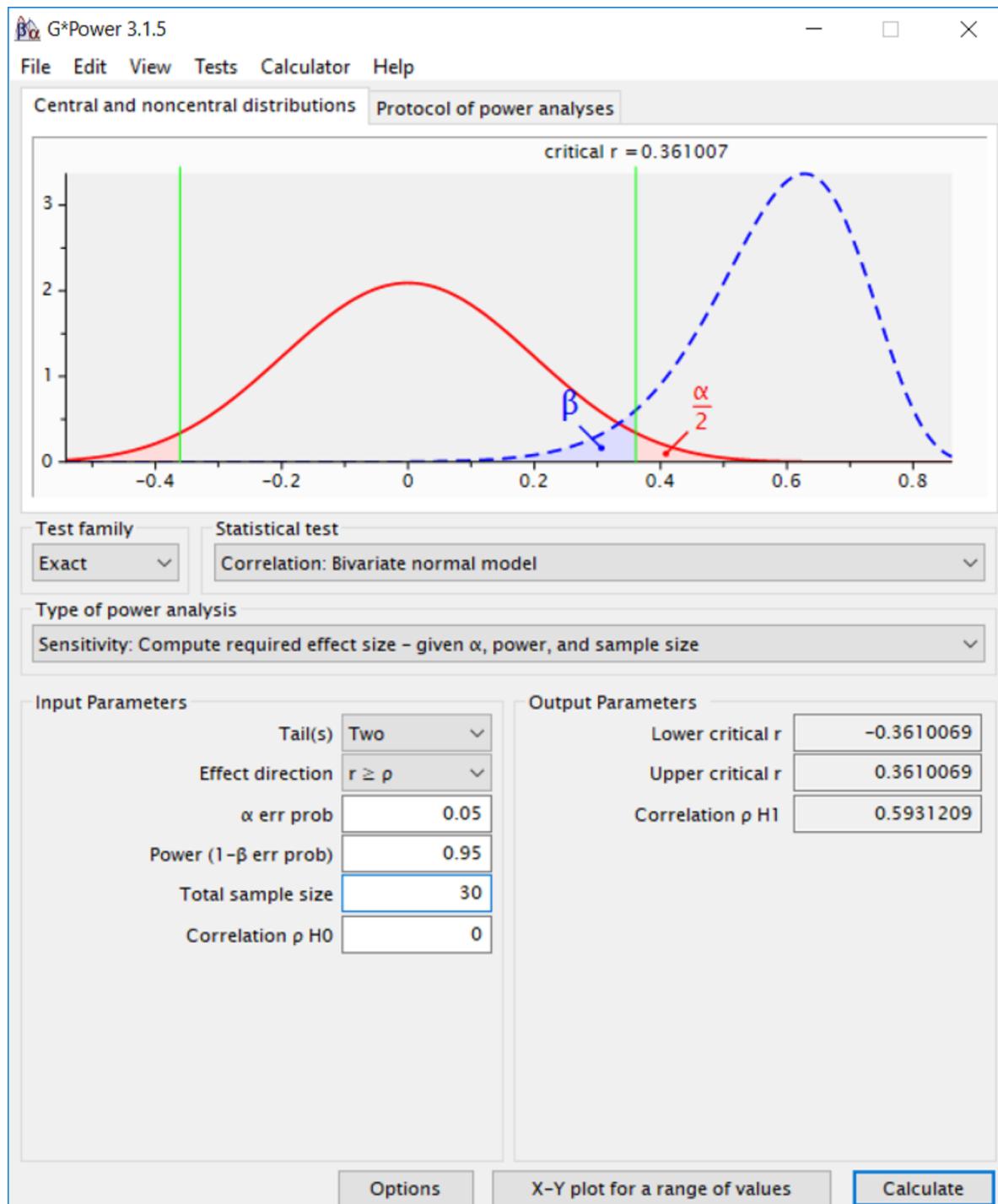


Figure 6.12: The critical correlation of a test based on a total sample size of 30 and  $\alpha = 0.05$  calculated in G\*Power.

which equals a Cohen's  $d$  of 1.76. What, according to the online app at <https://rpsychologist.com/cohend/>, is the probability of superiority for an effect of this size?

- (A) 70.5%
- (B) 88.1%
- (C) 89.3%
- (D) 92.1%

**Q2:** Cohen's  $d$  is to \_\_\_\_\_ as eta-squared is to \_\_\_\_\_

- (A)  $r$ ; epsilon-squared
- (B) Hedges'  $g$ ; omega-squared
- (C) Cohen's  $d_s$ ; generalized eta-squared

**Q3:** A correlation of  $r = 1.2$  is:

- (A) Impossible
- (B) Implausibly large for an effect size in the social sciences
- (C) In line with the median effect size in psychology

**Q4:** Let's assume the difference between two means we observe is 1, and the pooled standard deviation is also 1. If we simulate a large number of studies with those values, what, on average, happens to the  $t$ -value and Cohen's  $d$ , as a function of the sample size in these simulations?

- (A) Given the mean difference and standard deviation, as the sample size becomes bigger, the  $t$ -value become larger, and Cohen's  $d$  becomes larger.
- (B) Given the mean difference and standard deviation, as the sample size becomes bigger, the  $t$ -value gets closer to the true value, and Cohen's  $d$  becomes larger.
- (C) Given the mean difference and standard deviation, as the sample size becomes bigger, the  $t$ -value become larger, and Cohen's  $d$  gets closer to the true value.

- (D) Given the mean difference and standard deviation, as the sample size becomes bigger, the  $t$ -value gets closer to the true value, and Cohen's  $d$  gets closer to the true value.

**Q5:** Go to <http://rpsychologist.com/d3/correlation/> to look at a good visualization of the proportion of variance that is explained by group membership, and the relationship between  $r$  and  $r^2$ . Look at the scatterplot and the shared variance for an effect size of  $r = .21$  (Richard et al., 2003). Given that  $r = 0.21$  was their estimate of the median effect size in psychological research (not corrected for bias), how much variance in the data do variables in psychology on average explain?

- (A) 2%
- (B) 21%
- (C) 4%
- (D) 44%

**Q6:** By default, the sample size for the online correlation visualization linked to above is 50. Click on the cogwheel to access the settings, change the sample size to 500, and click the button 'New Sample'. What happens?

- (A) The proportion of explained variance is 5 times as large.
- (B) The proportion of explained variance is 5 times as small.
- (C) The proportion of explained variance is 52 times as large.
- (D) The proportion of explained variance stays the same.

**Q7:** In an old paper you find a statistical result reported as  $t(36) = 2.14$ ,  $p < 0.05$  for an independent  $t$ -test without a reported effect size. Using the online MOTE app <https://doomlab.shinyapps.io/mote/> (choose "Independent t -t" from the Mean Differences dropdown menu) or the MOTE R function `d.ind.t.t`, what is the Cohen's  $d$  effect size for this effect, given 38 participants (e.g., 19 in each group, leading to  $N - 2 = 36$  degrees of freedom) and an alpha level of 0.05?

- (A)  $d = 0.38$

- (B)  $d = 0.41$
- (C)  $d = 0.71$
- (D)  $d = 0.75$

**Q8:** In an old paper you find a statistical result from a 2x3 between-subjects ANOVA reported as  $F(2, 122) = 4.13, p < 0.05$ , without a reported effect size. Using the online MOTE app <https://doomlab.shinyapps.io/mote/> (choose Eta – F from the Variance Overlap dropdown menu) or the MOTE R function `eta.F`, what is the effect size expressed as partial eta-squared?

- (A)  $\eta_p^2 = 0.06$
- (B)  $\eta_p^2 = 1.00$
- (C)  $\eta_p^2 = 0.032$
- (D)  $\eta_p^2 = 0.049$

**Q9:** You realize that computing omega-squared corrects for some of the bias in eta-squared. For the old paper with  $F(2, 122) = 4.13, p < 0.05$ , and using the online MOTE app <https://doomlab.shinyapps.io/mote/> (choose Omega – F from the Variance Overlap dropdown menu) or the MOTE R function `omega.F`, what is the effect size in partial omega-squared? HINT: The total sample size is the  $df_{error} + k$ , where  $k$  is the number of groups (which is 6 for the 2x3 ANOVA).

- (A)  $\eta_p^2 = 0.05$
- (B)  $\eta_p^2 = 0.75$
- (C)  $\eta_p^2 = 0.032$
- (D)  $\eta_p^2 = 0.024$

**Q10:** Several times in this chapter the effect size Cohen's  $d$  was converted to  $r$ , or vice versa. We can use the `effectsize` R package (that can also be used to compute effect sizes when you analyze your data in R) to convert the median  $r = 0.21$  observed in Richard and colleagues' meta-meta-analysis to  $d$ : `effectsize::r_to_d(0.21)` which (assuming equal sample sizes per condition) yields  $d = 0.43$  (the conversion assumes equal sample sizes in each group). Which Cohen's  $d$  corresponds to a  $r = 0.1$ ?

- (A)  $d = 0.05$
- (B)  $d = 0.10$
- (C)  $d = 0.20$
- (D)  $d = 0.30$

**Q11:** It can be useful to convert effect sizes to  $r$  when performing a meta-analysis where not all effect sizes that are included are based on mean differences. Using the `d_to_r()` function in the `effectsize` package, what does a  $d = 0.8$  correspond to (again assuming equal sample sizes per condition)?

- (A)  $r = 0.30$
- (B)  $r = 0.37$
- (C)  $r = 0.50$
- (D)  $r = 0.57$

**Q12:** From questions 10 and 11 you might have noticed something peculiar. The benchmarks typically used for ‘small’, ‘medium’, and ‘large’ effects for Cohen’s  $d$  are  $d = 0.2$ ,  $d = 0.5$ , and  $d = 0.8$ , and for a correlation are  $r = 0.1$ ,  $r = 0.3$ , and  $r = 0.5$ . Using the `d_to_r()` function in the `effectsize` package, check to see whether the benchmark for a ‘large’ effect size correspond between  $d$  and  $r$ .

As McGrath & Meyer (2006) write: “Many users of Cohen’s (1988) benchmarks seem unaware that those for the correlation coefficient and  $d$  are not strictly equivalent, because Cohen’s generally cited benchmarks for the correlation were intended for the infrequently used biserial correlation rather than for the point biserial.”

Download the paper by McGrath and Meyer, 2006 (you can find links to the pdf [here](#)), and on page 390, right column, read which solution the authors prefer.

- (A) Think carefully about the limitations of using benchmarks.
- (B) Just stop using these silly benchmarks.
- (C) The benchmarks for  $d$  would need to be changed to 0.20, 0.67, and 1.15
- (D) The benchmarks for correlations  $r$  would need to be changed to .10, .24, and .37

### **6.11.1 Open Questions**

1. What is the difference between standardized and unstandardized effect sizes?
2. Give a definition of an ‘effect size’.
3. What are some of the main uses of effect sizes?
4. How can effect sizes improve statistical inferences, beyond looking at the  $p$ -value?
5. What is the effect size  $r$ , and which values can it take?
6. What is the effect size  $d$ , and which values can it take?
7. What are the unbiased effect sizes that correspond to  $d$  and eta-squared called?
8. Give an example when small effects are meaningless, and when they are not.
9. Researchers often use Cohen’s (1988) benchmarks to interpret effect sizes. Why is this not best practice?
10. What is the difference between ordinal and disordinal interaction effects? And if the means across different conditions are either 0 or 1 on a scale, which type of interaction will have a larger effect size?

# 7 Confidence Intervals

When we report point estimates, we should acknowledge and quantify the uncertainty in these estimates. Confidence intervals provide a way to quantify the precision of an estimate. By reporting an estimate with a confidence interval, results are reported within a range of values that contain the true value of the parameter with a desired percentage. For example, when we report an effect size estimate with a 95% confidence interval, the expectation is that the interval is wide enough such that 95% of the time the range of values around the estimate contains the true parameter value (if all test assumptions are met).

## 7.1 Population vs. Sample

In statistics, we differentiate between the population and the sample. The population is everyone you are interested in, such as all people in the world, elderly who are depressed, or people who buy innovative products. Your sample is everyone you were able to measure from the population you are interested in. We similarly distinguish between a parameter and a statistic. A parameter is a characteristic of the population, while a statistic is a characteristic of a sample. Sometimes, you have data about your entire population. For example, we have measured the height of all the people who have ever walked on the moon. We can calculate the average height of these twelve individuals, and so we know the true parameter. We do not need inferential statistics. However, we do not know the average height of all people who have ever walked on the earth. Therefore, we need to estimate this parameter, using a statistic based on a sample. Although it is rare that a study includes the entire population, it is not impossible, as illustrated in Figure 7.1.

When the entire population is measured there is no need to perform a hypothesis test. After all, there is no population to generalize to (although it is possible to argue we are still making an inference, even when the entire population is observed, because we have observed a *metaphorical population* from one of many possible worlds, see D. Spiegelhalter (2019)). When data from the entire population has been collected, the population effect size is known and there is no confidence interval to compute. If the total population size is known, but not measured completely, then the confidence interval width should shrink to zero the closer a study gets to measuring the entire population. This is known as the finite population correction factor for the variance of the estimator (Kish, 1965). The variance of a sample mean is

## Methods

In this registry-based study of all children in Norway ( $n = 1\,354\,393$ ) aged 5–17 years from 2008 to 2016, we examined whether parental income was associated with childhood diagnoses of mental disorders identified through national registries from primary healthcare, hospitalizations and specialist outpatient services.

Figure 7.1: Example of a registry-based study in which the entire population was included in the study. From <https://doi.org/10.1093/ije/dyab066>

$\sigma^2/n$ , which for finite populations is multiplied by the finite population correction factor of the standard error:

$$FPC = \sqrt{\frac{(N-n)}{(N-1)}}$$

where  $N$  is the size of the population, and  $n$  is the size of the sample. When  $N$  is much larger than  $n$ , the correction factor will be close to 1 (and therefore this correction is typically ignored when populations are very large, even when populations are finite), and will not have a noticeable effect on the variance. When the total population is measured the correction factor is 0, such that the variance becomes 0 as well. For example, when the total population consists of 100 top athletes, and data is collected from a sample of 35 athletes, the finite population correction is  $\sqrt{(100 - 35)/(100 - 1)} = 0.81$ . The **superb** R package can compute population corrected confidence intervals (Cousineau & Chiasson, 2019).

## 7.2 What is a Confidence Interval?

Confidence intervals are a statement about the percentage of confidence intervals that contain the true parameter value. This behavior of confidence intervals is nicely visualized on this website by Kristoffer Magnusson: <http://rpsychologist.com/d3/CI/>. In Figure 7.2 we see blue dots that represent means from a sample, and that fall around a red dashed vertical line, which represents the true value of the parameter in the population. Due to variation in the sample, the estimates do not all fall on the red dashed line. The horizontal lines around the blue dots are the confidence intervals. By default, the visualization shows 95% confidence intervals. Most of the lines are black (which means the confidence interval overlaps with the orange dashed line indicating the true population value), but some are red (indicating they do not capture the true population value). In the long run, 95% of the horizontal bars will be black, and 5% will be red.

## 95% confidence intervals

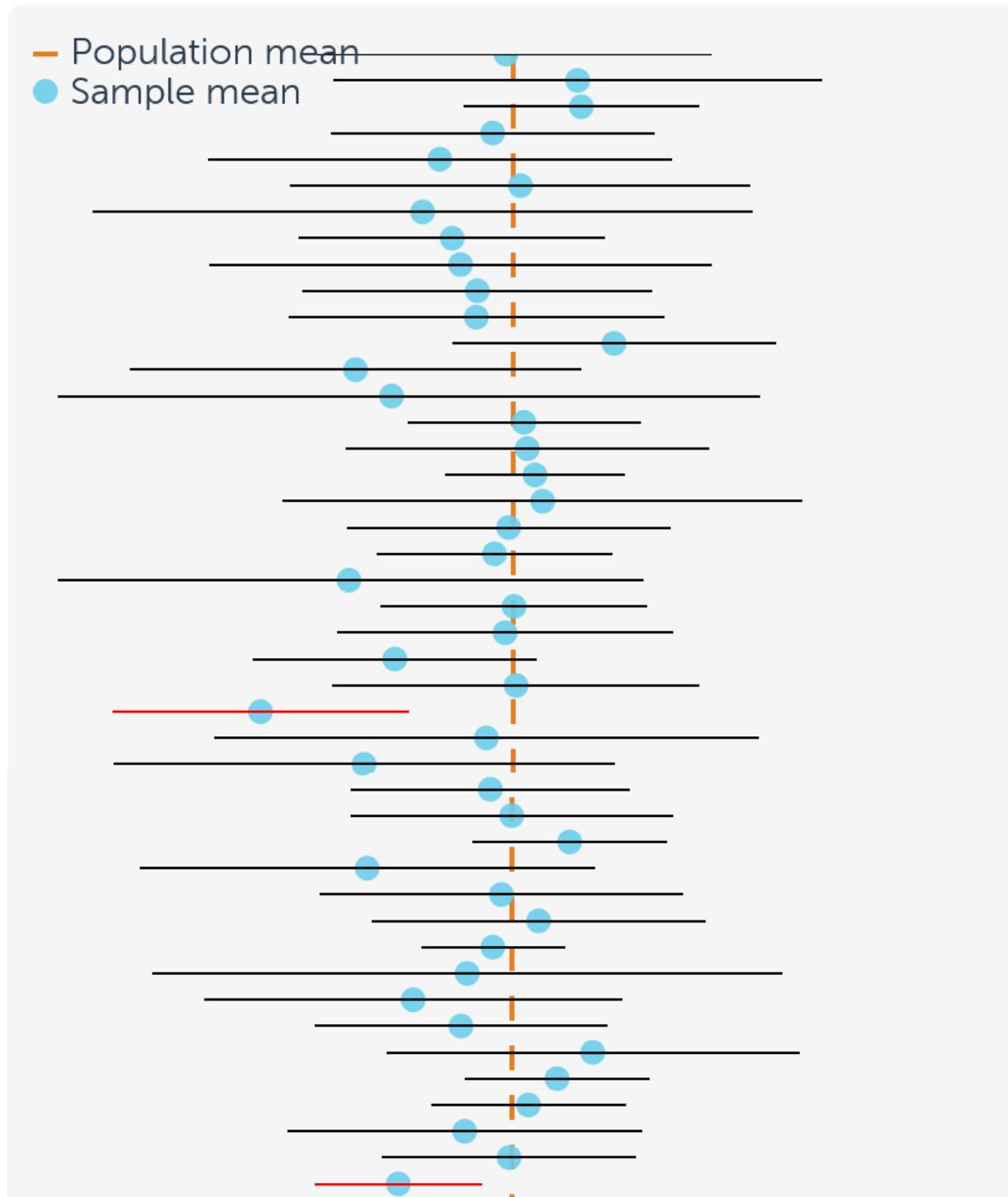


Figure 7.2: Series of simulated point estimates and confidence intervals.

We can now see what is meant by the sentence “Confidence intervals are a statement about the percentage of confidence intervals that contain the true parameter value”. In the long run, for 95% of the samples, the orange dashed line (the population parameter) is contained within the 95% confidence interval around the sample mean, and in 5% of the confidence intervals this is not true. As we will see when we turn to the formula for confidence intervals, the width of a confidence interval depends on the sample size and the standard deviation. The larger the sample size, the smaller the confidence intervals. In Figure 7.3 we see how the confidence interval around the mean becomes more narrow as the sample size increases (assuming a standard deviation of 1). If we imagine confidence interval lines around each observed mean with a width equal to the width of the difference between the upper and lower confidence interval around 1, we see that in the long run 95% of the observed means would have a 95% confidence interval that contains the true population parameter (a mean of 1), but that approximately 5% of the means are more extreme and do not have a confidence interval that contains the population parameter of 1.

### Confidence Interval and Observed Means as the Sample Size Increases

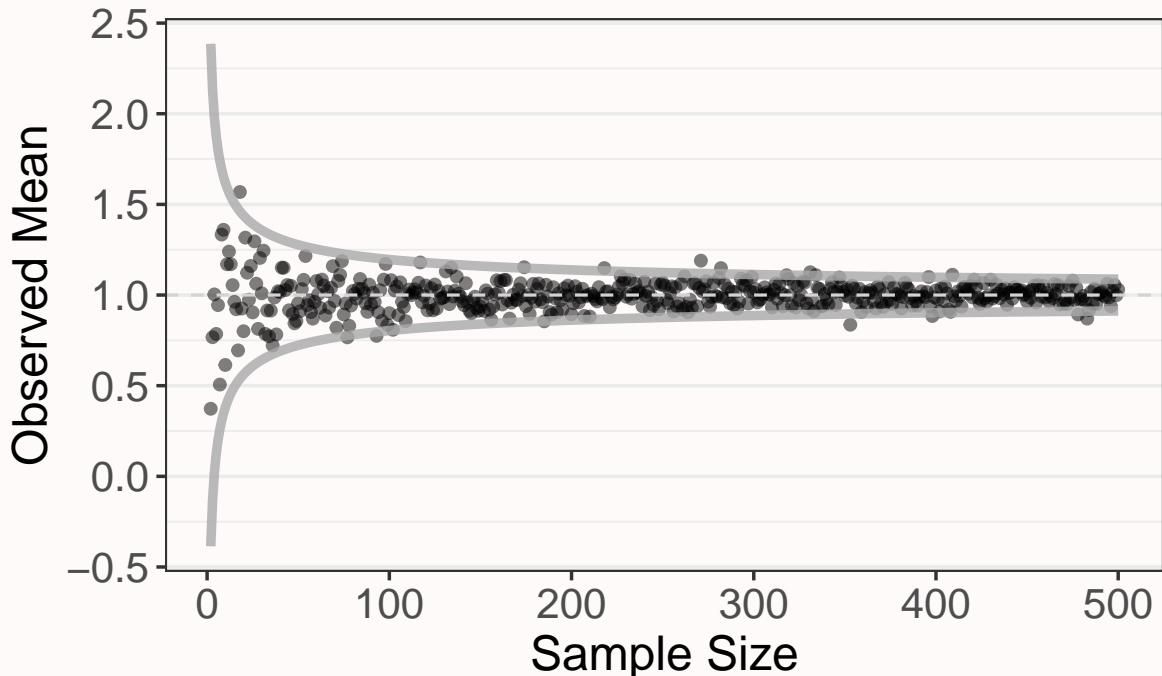


Figure 7.3: The observed mean for 500 samples drawn from a population with a true mean of 1 as the sample size increases to 500 observations. Grey lines indicate the upper and lower 95% confidence interval around the mean.

### 7.3 Interpreting a single confidence interval

Whenever we compute or encounter a single confidence interval it is important to realize that someone else performing exactly the same experiment would, purely due to random variation, have observed a different confidence interval, effect size, and  $p$ -value. Because of this random variation a single confidence interval is difficult to interpret. Misinterpretations are common. For example, Cumming (2014) writes “We can be 95% confident that our interval includes  $\mu$  and can think of the lower and upper limits as likely lower and upper bounds for  $\mu$ .” Both these statements are incorrect (Richard D. Morey et al., 2016). It is incorrect to claim we can be 95% confident that our interval includes the true population mean, because if we study whether our friend can predict whether a coin comes up heads or tails in 100 flips, and they correctly predict the coin flip in 61 out of 100 flips with a 95% confidence interval from 0.507 to 0.706, it is perfectly reasonable to use some Bayesian reasoning and assume (with more than the remaining 5% confidence) it was just a fluke, and the true success rate when guessing the outcome of coin flips is 50%. It is also incorrect to believe the lower and upper limits are likely lower and upper bounds for  $\mu$ , as anyone else performing the same experiment would have observed a different confidence interval, with a different upper and lower bound, when analyzing a single sample drawn from the same population. If a lot of data has been collected (say thousands of observations) this problem practically disappears, because the remaining uncertainty is too small to matter.

One useful way to think of a confidence interval is as an indication of the resolution with which an effect is estimated. If the resolution is low, it is difficult to get a clear picture, but if the resolution is extremely high, the picture is clear enough for all practical use cases. If we have estimated an effect in a very narrow range, say a  $M = 0.52$ , 95% CI [0.49; 0.55], and we feel warranted to assume that no one cares about differences less than 0.05 on the measure, a confidence interval communicates that the data have been estimated with sufficient precision. Similarly, if the sample size is small, and the confidence is very wide, say a  $M = 0.52$ , 95% CI [0.09; 0.95], and we feel warranted to assume that differences of almost 1 on the measure matter for situations in which the estimate would be used, the confidence interval communicates that the effect size estimate is not precise enough. This evaluation of the resolution of the estimate can be useful, and is missing if only a  $p$ -value or effect size are reported. For this reason, it is recommended to report confidence intervals around estimates. Confidence intervals are often reported within brackets, but an interesting alternative (especially for tables) is to use subscripts:  $_{0.09}^{0.52}_{0.95}$  (Louis & Zeger, 2009).

It is tempting to use a Bayesian interpretation of a single confidence interval, where one would say “I believe there is a 95% probability that this interval contains the true population parameter”. A Bayesian interpretation has lost frequentist error control which means that depending on the prior this belief might be misguided in much more than 5% of the studies. This is not something a Bayesian worries about, as the focus of their inferences is not on limiting errors in the long run, but on quantifying beliefs. A frequentist can not make probability claims for single observations. After the data has been collected, a frequentist can only state that the

current confidence interval either contains the true population parameter, or it does not. In the long run,  $\alpha\%$  of the confidence intervals will not include the true population parameter, and this single confidence interval could be one of these flukes. Even though a frequentist and Bayesian confidence interval can be identical under certain priors (Albers et al., 2018), the different definitions of probability lead to different interpretations of a single confidence interval. A frequentist can easily interpret a confidence *procedure*, but it is not so easy to interpret a single confidence interval (Richard D. Morey et al., 2016). This should not surprise us, because it is difficult to interpret any single study (which is why we need to perform replication studies). When confidence intervals are interpreted as a long-run procedure, they are directly related to *p*-values.

## 7.4 The relation between confidence intervals and *p*-values

There is a direct relationship between the CI around an effect size and statistical significance of a null-hypothesis significance test. For example, if an effect is statistically significant ( $p < 0.05$ ) in a two-sided independent *t*-test with an alpha of .05, the 95% CI for the mean difference between the two groups will not include zero. Confidence intervals are sometimes said to be more informative than *p*-values, because they not only provide information about whether an effect is statistically significant (i.e., when the confidence interval does not overlap with the value representing the null hypothesis), but also communicate the precision of the effect size estimate. This is true, but as mentioned in the chapter on *p*-values it is still recommended to add exact *p*-values, which facilitates the re-use of results for secondary analyses (Appelbaum et al., 2018), and allows other researchers to compare the *p*-value to an alpha level they would have preferred to use (Lehmann & Romano, 2005).

In order to maintain the direct relationship between a confidence interval and a *p*-value it is necessary to adjust the confidence interval level whenever the alpha level is adjusted. For example, if an alpha level of 5% is corrected for three comparisons to  $0.05/3 = 0.0167$ , the corresponding confidence interval would be a  $1 - 0.0167 = 0.9833$  confidence interval. Similarly, if a *p*-value is computed for a one-sided *t*-test, there is only an upper or lower limit of the interval, and the other end of the interval ranges to  $-\infty$  or  $\infty$ .

To maintain a direct relationship between an *F*-test and its confidence interval, a 90% CI for effect sizes from an *F*-test should be provided. The reason for this is explained by [Karl Wuensch](#). Where Cohen's *d* can take both positive and negative values,  $r^2$  or  $\eta^2$  are squared, and can therefore only take positive values. This is related to the fact that *F*-tests (as commonly used in ANOVA) are one-sided. If you calculate a 95% CI, you can get situations where the confidence interval includes 0, but the test reveals a statistical difference with a  $p < .05$  (for a more mathematical explanation, see Steiger (2004)). This means that a 95% CI around Cohen's *d* in an independent *t*-test equals a 90% CI around  $\eta^2$  for exactly the same test performed as an ANOVA. As a final detail, because eta-squared cannot be smaller than zero, the lower bound for the confidence interval cannot be smaller than 0. This means that a confidence interval for

an effect that is not statistically different from 0 has to start at 0. You report such a CI as 90% CI [.00; .XX] where the XX is the upper limit of the CI.

Confidence intervals are often used in forest plots that communicate the results from a meta-analysis. In the plot below, we see 4 rows. Each row shows the effect size estimate from one study (in Hedges'  $g$ ). For example, study 1 yielded an effect size estimate of 0.53, with a confidence interval around the effect size from 0.12 to 0.94. The horizontal black line, similarly to the visualization we played around with before, is the width of the confidence interval. When it does not touch the effect size 0 (indicated by a black vertical dotted line) the effect is statistically significant.

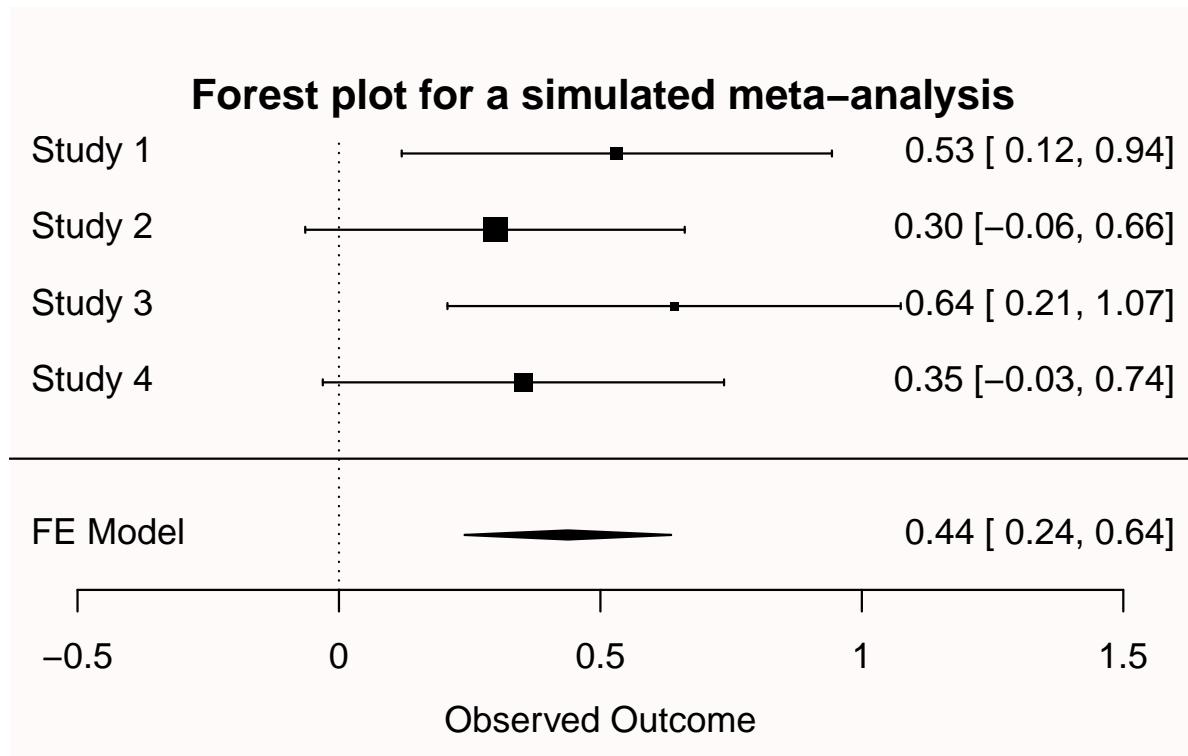


Figure 7.4: Meta-analysis of 4 studies.

We can see, based on the fact that the confidence intervals do not overlap with 0, that studies 1 and 3 were statistically significant. The diamond shape named the FE model (Fixed Effect model) is the meta-analytic effect size. Instead of using a black horizontal line, the upper limit and lower limit of the confidence interval are indicated by the left and right points of the diamond, and the center of the diamond is the meta-analytic effect size estimate. A meta-analysis calculates the effect size by combining and weighing all studies. The confidence interval for a meta-analytic effect size estimate is always narrower than that for a single study, because of the combined sample size of all studies included in the meta-analysis.

In the preceding section, we focused on examining whether the confidence interval overlapped with 0. This is a confidence interval approach to a null-hypothesis significance test. Even though we are not computing a  $p$ -value, we can directly see from the confidence interval whether  $p < \alpha$ . The confidence interval approach to hypothesis testing makes it quite intuitive to think about performing tests against non-zero null hypotheses (Bauer & Kieser, 1996). For example, we could test whether we can reject an effect of 0.5 by examining if the 95% confidence interval does not overlap with 0.5. We can test whether an effect is *smaller* than 0.5 by examining if the 95% confidence interval falls completely *below* 0.5. We will see that this leads to a logical extension of null-hypothesis testing where, instead of testing to reject an effect of 0, we can test whether we can reject other effects of interest in **range predictions** and **equivalence tests**.

## 7.5 The Standard Error and 95% Confidence Intervals

To calculate a confidence interval, we need the standard error. The standard error (SE) estimates the variability between sample means that would be obtained after taking several measurements from the same population. It is easy to confuse it with the standard deviation, which is the degree to which individuals within the sample differ from the sample mean. Formally, statisticians distinguish between  $\sigma$  and  $\hat{\sigma}$ , where the hat means the value is estimated from a sample, and the lack of a hat means it is the population value – but I'll leave out the hat, even when I'll mostly talk about estimated values based on a sample in the formulas below. Mathematically (where  $\sigma$  is the standard deviation),

$$\text{Standard Error (SE)} = \sigma / \sqrt{n}$$

The standard error of the sample will tend to zero with increasing sample size, because the estimate of the population mean will become more and more accurate. The standard deviation of the sample will become more and more similar to the population standard deviation as the sample size increases, but it will not become smaller. Where the standard deviation is a statistic that is descriptive of your sample, the standard error describes bounds on a random sampling process.

The standard error is used to construct confidence intervals (CI) around sample estimates, such as the mean, or differences between means, or whatever statistics you might be interested in. To calculate a confidence interval around a mean (indicated by the Greek letter mu:  $\mu$ ), we use the  $t$  distribution with the corresponding degrees of freedom ( $df$  : in a one-sample  $t$ -test, the degrees of freedom are  $n-1$ ):

$$\mu \pm t_{df, 1-(\alpha/2)} SE$$

With a 95% confidence interval, the  $\alpha = 0.05$ , and thus the critical  $t$ -value for the degrees of freedom for  $1 - \alpha / 2$ , or the 0.975th quantile is calculated. Remember that a  $t$ -distribution has slightly thicker tails than a  $Z$ -distribution. Where the 0.975th quantile for a  $Z$ -distribution is 1.96, the value for a  $t$ -distribution with for example  $df = 19$  is 2.093. This value is multiplied by the standard error, and added (for the upper limit of the confidence interval) or subtracted (for the lower limit of the confidence interval) from the mean.

## 7.6 Overlapping Confidence Intervals

Confidence intervals are often used in plots. In Figure 7.5 below, three estimates are visualized (the dots), surrounded by three lines (the 95% confidence intervals). The left two dots (X and Y) represent the *means* of the independent groups X and Y on a scale from 0 to 8 (see the axis from 0-8 on the left side of the plot). The dotted lines between the two confidence intervals visualize the overlap between the confidence intervals around the means. The two confidence intervals around the means of X and Y are commonly shown in a figure in a scientific article. The third dot, slightly larger, is the *mean difference* between X and Y, and the slightly thicker line visualizes the confidence interval of this mean difference. The difference score is expressed using the axis on the right (from -3 to 5). In the plot below, the mean of group X is 3, the mean of group Y is 5.6, and the difference is 2.6. The plot is based on 50 observations per group, and the confidence interval around the mean difference ranges from 0.49 to 4.68, which is quite wide.

As mentioned earlier, when a 95% confidence interval does not contain 0, the effect is statistically different from 0. In Figure 7.5 above, the mean difference and the 95% confidence interval around it are indicated by the ‘difference’ label. As the 95% confidence interval does not contain 0, the  $t$ -test is significant at an alpha of 0.05. The  $p$ -value is indicated in the plot as 0.016. Even though the two means differ statistically significantly from each other, the confidence interval around each mean overlap. One might intuitively believe that an effect is only statistically significant if the confidence interval around the individual means do not overlap, but this is not true. The significance test is related to the confidence interval around the mean difference.

## 7.7 Prediction Intervals

Even though 95% of confidence intervals will contain the true parameter in the long run, a 95% confidence interval will not contain 95% of future individual observations (or 95% of future means; this will be discussed in the next section). Sometimes, researchers want to predict the interval within which a single value will fall. This is called the prediction interval. It is always much wider than a confidence interval. The reason is that individual observations can vary

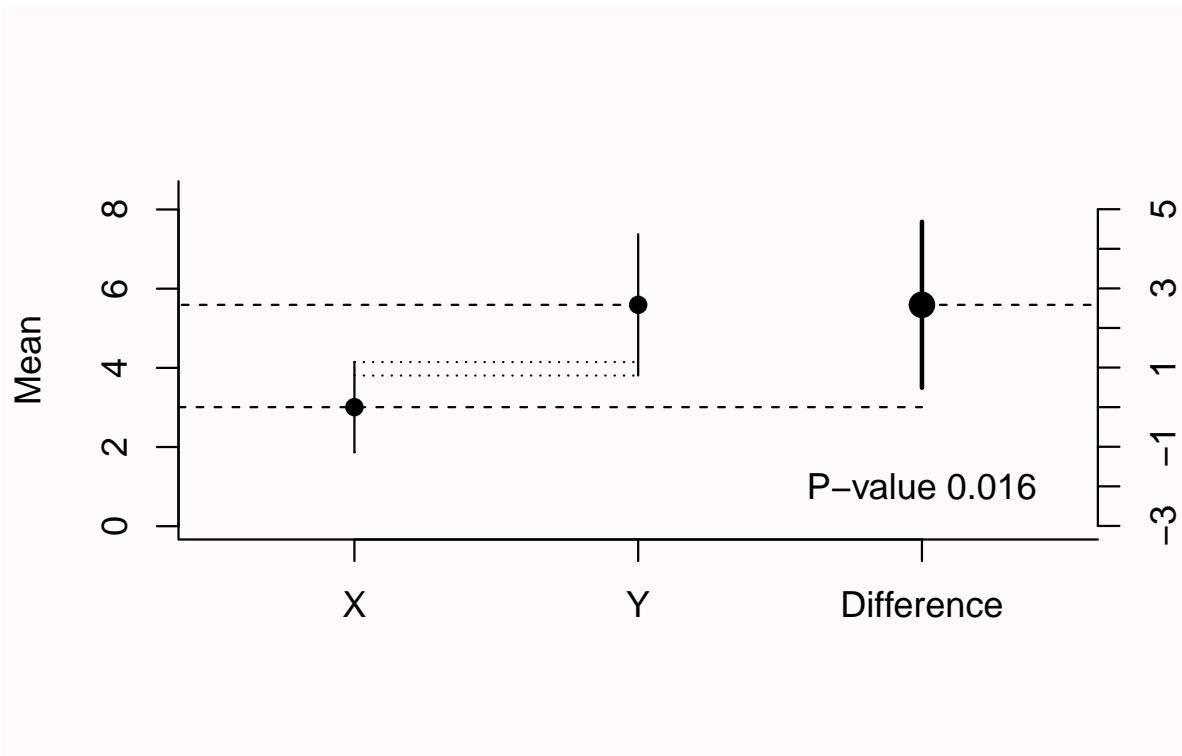


Figure 7.5: Means and 95% confidence intervals of two independent groups and the mean difference between the two groups and its 95% confidence interval.

substantially, but means of future samples (which fall within a normal confidence interval 95% of the time) will vary much less.

In Figure 7.6, the orange background illustrates the 95% confidence interval around the mean, and the yellow background illustrates the 95% prediction interval (PI).

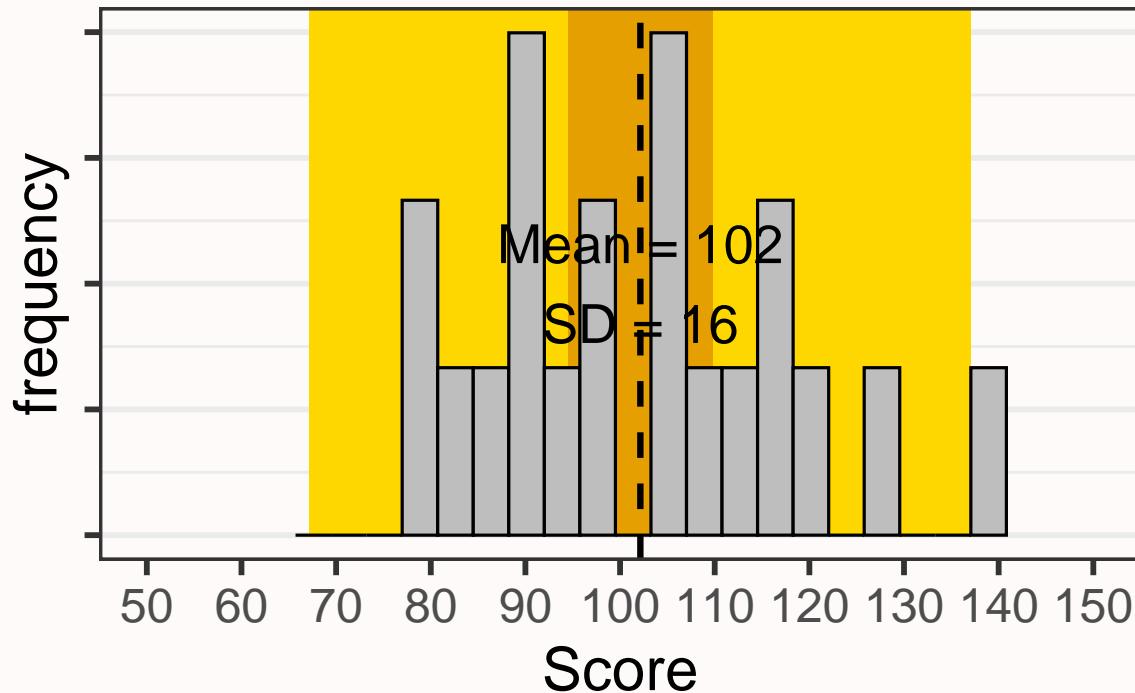


Figure 7.6: A comparison of a 95% confidence interval (orange) and 95% prediction interval (yellow).

To calculate the prediction interval, we need a slightly different formula for the standard error than that which was used for the confidence interval, namely:

$$\text{Standard Error (SE)} = \sigma / \sqrt{(1 + 1/n)}$$

When we rewrite the formula used for the confidence interval to  $\sigma / \sqrt{(1/N)}$ , we see that the difference between a confidence interval and the prediction interval is in the “1+” which always leads to wider intervals. Prediction intervals are **wider**, because they are constructed so that they will contain **a single future value** 95% of the time, instead of the **mean**. The fact that prediction intervals are wide is a good reminder that it is difficult to predict what will happen for any single individual.

## 7.8 Capture Percentages

It can be difficult to understand why a 95% confidence interval does not provide us with the interval where 95% of future means will fall. The percentage of means that falls within a single confidence interval is called the **capture percentage**. A capture percentage is not something we would ever use to make inferences about data, but it is useful to learn about capture percentages to prevent misinterpreting confidence intervals. In Figure 7.7 we see two randomly simulated studies with the same sample size from the same population. The true effect size in both studies is 0, and we see that the 95% confidence intervals for both studies contain the true population value of 0. However, the two confidence intervals cover quite different ranges of effect sizes, with the confidence interval in Study 1 ranging from -0.07 to 0.48, and the confidence interval in Study 2 ranging from -0.50 to 0.06. It cannot be true that in the future, we should expect 95% of the effect sizes to fall between -0.07 to 0.48 **and** 95% of the effect sizes to fall between -0.50 to 0.06.

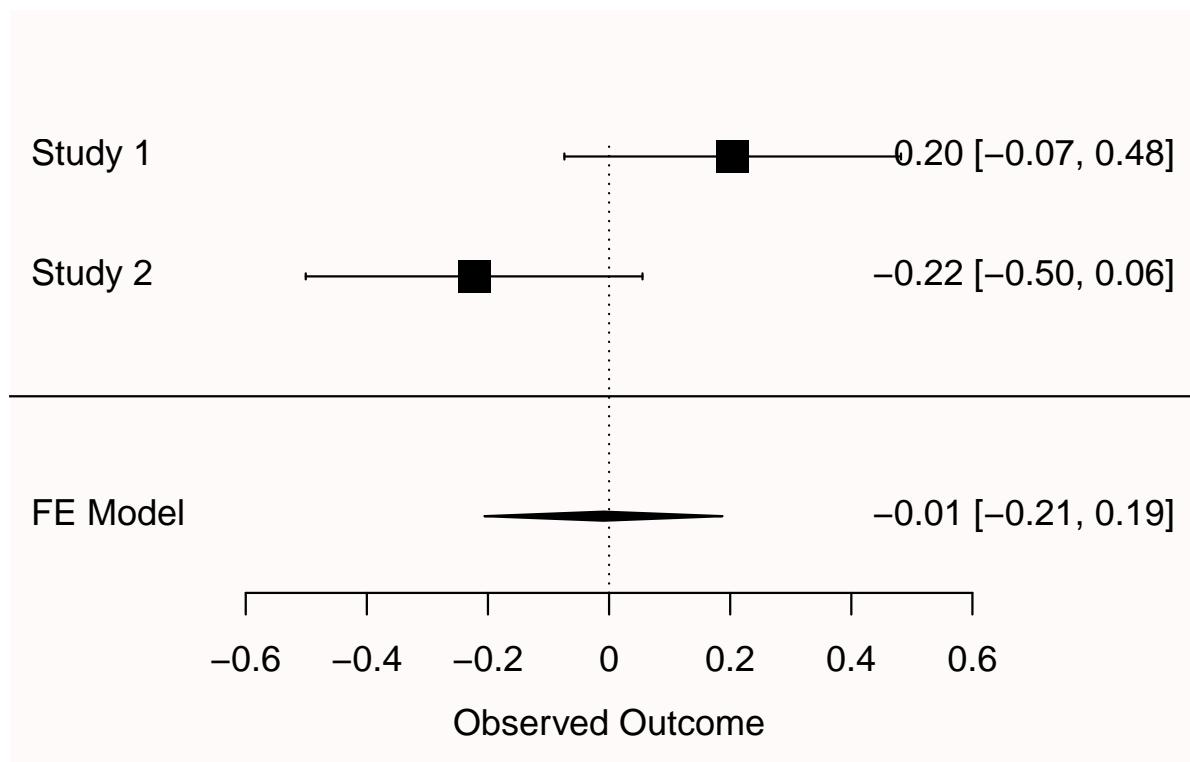


Figure 7.7: Meta-analysis of two simulated studies from the same population.

The only situation in which a 95% confidence interval happens to also be a 95% capture percentage is when the observed effect size in a sample happens to be exactly the same as the true population parameter. In Figure 7.7, that means we would need to observe an effect of exactly 0. However, you can't know whether the observed effect size happens to be exactly

the same as the population effect size. When a sample estimate is not identical to the true population value (which is almost always the case) less than 95% of future effect sizes will fall within the CI from your current sample. As we have observed two studies with the observed effect sizes a bit removed from the true effect size, we will find effect size estimates in future studies that fall outside the observed 95% confidence interval quite often. So, the percentage of future means that fall within a single confidence interval depends upon which single confidence interval you happened to observe. Based on simulation studies it is possible to show that on average, in the long run, a 95% CI has an 83.4% capture probability (Cumming & Maillardet, 2006).

## 7.9 Calculating Confidence Intervals around Standard Deviations.

If we calculate a standard deviation (SD) from a sample, this value is an estimate of the true value in the population. In small samples, this estimate can be quite far off the population value. But due to the law of large numbers, as our sample size increases, we will be measuring the standard deviation more accurately. Since the sample standard deviation is an estimate with uncertainty, we can calculate a 95% confidence interval around it. For some reason, this is rarely done in practice, maybe because researchers are often more interested in means, and less interested in standard deviations. But standard deviations are an interesting property of our measures, and one could even make theoretical predictions about an increase or decrease in standard deviations between conditions. Currently, researchers rarely theorize about variation in standard deviations, and perhaps because of this, basically never compute confidence intervals around the standard deviations they report.

Keeping the uncertainty of standard deviations in mind can be important. When researchers perform an a-priori power analysis based on an effect size of interest expressed on a raw scale, they need accurate estimates of the standard deviation. Sometimes researchers will use pilot data to get an estimate of the standard deviation. Since the estimate of the population standard deviation based on a pilot study has some uncertainty (as pilot studies usually have a relatively small sample size), the a-priori power analysis will inherit this uncertainty (see the ‘Test Yourself’ questions below). To circumvent this, use validated or existing measures for which accurate estimates of the standard deviation in your population of interest are available. And keep in mind that all estimates from a sample have uncertainty.

## 7.10 Computing Confidence Intervals around Effect Sizes

In 1994, Cohen (1994) reflected on the reason confidence intervals were rarely reported: “I suspect that the main reason they are not reported is that they are so embarrassingly large!” This might be, but another reason might have been that statistical software rarely provided confidence intervals around effect sizes in the time when Cohen wrote his article. It has become

increasingly easy to report confidence intervals with the popularity of free software packages in R, even though these packages might not provide solutions for all statistical tests yet. The [Journal Article Reporting Standards](#) recommend to report “effect-size estimates and confidence intervals on estimates that correspond to each inferential test conducted, when possible”.

One easy solution to calculating effect sizes and confidence intervals is [MOTE](#) made by Dr. Erin Buchanan and her lab. The website comes with a full collection of tutorials, comparisons with other software packages, and demonstration videos giving accessible overviews of how to compute effect sizes and confidence intervals for a wide range of tests based on summary statistics. This means that whichever software you use to perform statistical tests, you can enter sample sizes and means, standard deviations, or test statistics to compute effect sizes and their confidence intervals. For example, the video below gives an overview of how to compute a confidence interval around Cohen’s  $d$  for an independent  $t$ -test.

MOTE is also available as an R package (Buchanan et al., 2017). Although many solutions exists to compute Cohen’s  $d$ , MOTE sets itself apart by allowing researchers to compute effect sizes and confidence intervals for many additional effect sizes, such as (partial) omega squared for between subjects ANOVA ( $\omega^2$  and  $\omega_p^2$ ), generalized omega squared for ANOVA ( $\omega_G^2$ ), epsilon squared for ANOVA ( $\varepsilon^2$ ) and (partial) generalized eta squared for ANOVA ( $\eta_G^2$ ).

```
MOTE:::d.ind.t(m1 = 1.7, m2 = 2.1, sd1 = 1.01, sd2 = 0.96, n1 = 77, n2 = 78, a = .05)$estimate
```

```
[1] "$d_s$ = -0.41, 95\\% CI [-0.72, -0.09]"
```

MBESS is another R package that has a range of options to compute effect sizes and their confidence intervals (Kelley, 2007). The code below reproduces the example for MOTE above.

```
MBESS::smd(Mean.1 = 1.7, Mean.2 = 2.1, s.1 = 1.01, s.2 = 0.96, n.1 = 77, n.2 = 78)
```

```
[1] -0.406028
```

If you feel comfortable analyzing your data in R, the [effectsize](#) package offers a complete set of convenient solutions to compute effect sizes and confidence intervals (Ben-Shachar et al., 2020).

```
set.seed(33)
x <- rnorm(n = 20, mean = 0, sd = 2.5) # create sample from normal distribution
y <- rnorm(n = 200, mean = 1.5, sd = 3.5) # create sample from normal distribution
effectsize::cohens_d(x, y)
```

Cohens_d	CI	CI_low	CI_high
-0.443983	0.95	-0.9050135	0.0180575

I am personally impressed by the way the **effectsize** package incorporates the state of the art (although I might be a bit biased). For example, after our recommendation to, by default, use Welch's *t*-test instead of students *t*-test (Delacre et al., 2017), and based on a recent simulation study recommended to report Hedges'  $g_s^*$  as the effect size for Welch's *t*-test (Delacre et al., 2021), the **effectsize** package was the first to incorporate it.

```
effectsize::cohens_d(x, y, pooled_sd = FALSE)
```

Cohens_d	CI	CI_low	CI_high
-0.5328286	0.95	-0.8972774	-0.1613137

Free statistical software **jamovi** and **JASP** are strong alternatives to SPSS that (unlike SPSS) allows users to compute Cohen's *d* and the confidence interval for both independent and dependent *t*-tests.

For **jamovi**, the ESCI module allows users to compute effect sizes and confidence intervals, and is accompanied by educational material that focuses more on estimation and less on testing (Cumming & Calin-Jageman, 2016).

**JASP** offers a wide range of frequentist and Bayesian analyses, and in addition to Cohen's *d* also allows users to compute omega squared  $\omega^2$ , the less biased version of  $\eta^2$  (Albers & Lakens, 2018; Okada, 2013).

## 7.11 Test Yourself

**Q1:** Go to the online app by Kristoffer Magnusson: <http://rpsychologist.com/d3/CI/>. You might want more confidence intervals to contain the true population parameter than 95%. Drag the 'Slide me' button to the far right, and you will see the simulation for 99% confidence intervals. Which statement is true?

- (A) The confidence intervals are larger, and the sample means fall closer to the true mean.
- (B) The confidence intervals are smaller, and the sample means fall closer to the true mean.

### Standardized Mean Difference

$d_{\text{unbiased}} = -0.62$  95% CI [-1.23, -0.06]

Note that the standardized effect size is  $d_{\text{unbiased}}$  because the denominator used was SDpooled which had a value of 0.83

The standardized effect size has been corrected for bias.

The bias-corrected version of Cohen's d is sometimes also (confusingly) called Hedges' g.

### Decision Making

t-table

t	df	p
-2.22	48.0	0.031

### Descriptives Plot

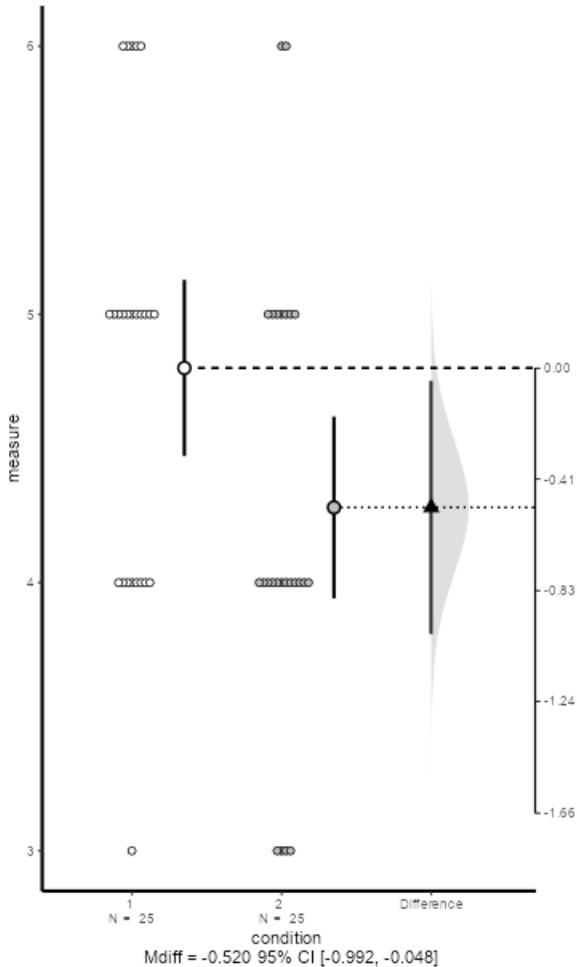


Figure 7.8: Output from ESCI module in jamovi.

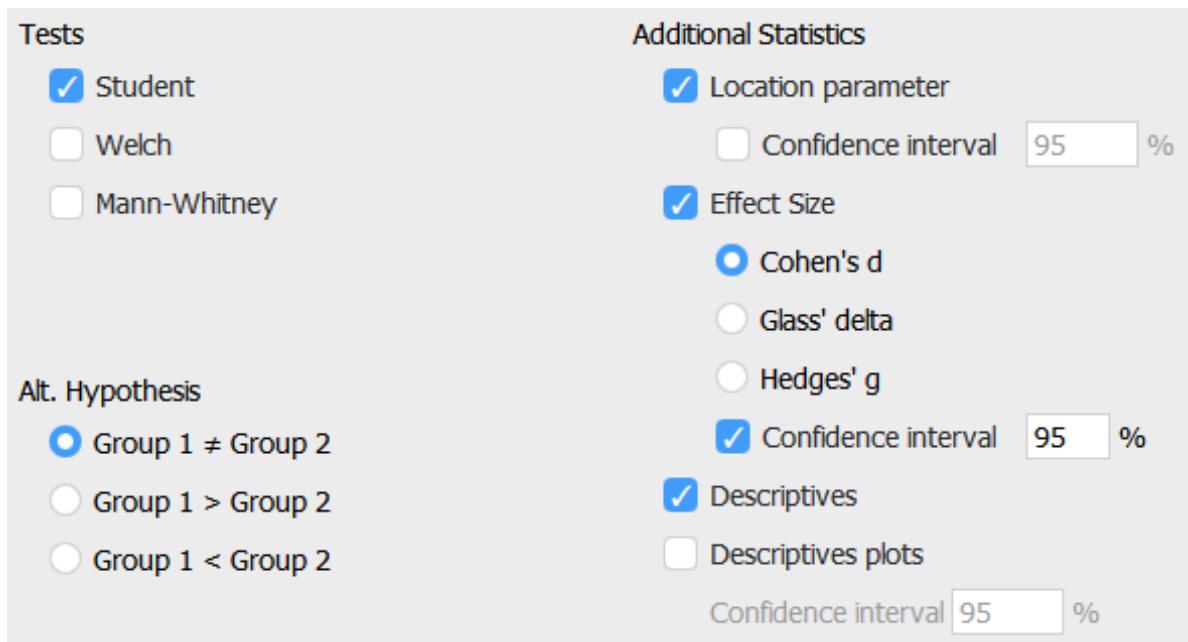


Figure 7.9: JASP menu option allows you to select Cohen's  $d$  and a CI around it.

## Independent Samples T-Test ▾

### Independent Samples T-Test

	t	df	p	Mean Difference	SE Difference	Cohen's d	95% CI for Cohen's d	
							Lower	Upper
dv	-0.212	298.000	0.832	-0.189	0.892	-0.024	-0.251	0.202

Note. Student's t-test.

Figure 7.10: JASP output returns Cohen's  $d$  and the confidence interval around it.

- (C) The confidence intervals are larger, and the sample means fall as close to the true mean as for a 95% confidence interval.
- (D) The confidence intervals are smaller, and the sample means fall as close to the true mean as for a 95% confidence interval.

**Q2:** As we could see from the formulas for confidence intervals, sample means and their confidence intervals depend on the sample size. We can change the sample size in the online app (see the setting underneath the visualization). By default, the sample size is set to 5. Change the sample size to 50 (you can type it in). Which statement is true?

- (A) The larger the sample size, the larger the confidence intervals. The sample size does not influence how the sample means vary around the true population mean.
- (B) The larger the sample size, the smaller the confidence intervals. The sample size does not influence how the sample means vary around the true population mean.
- (C) The larger the sample size, the larger the confidence intervals, and the closer the sample means are to the true population mean.
- (D) The larger the sample size, the smaller the confidence intervals, and the closer the sample means are to the true population mean.

**Q3:** In the forest plot below, we see the effect size (indicated by the square) and the confidence interval of the effect size (indicated by the line around the effect). Which of the studies 1 to 4 in the forest plot below were statistically significant?

- (A) Studies 1, 2, 3, and 4
- (B) Only study 3
- (C) None of the four studies
- (D) Studies 1, 2 and 4

**Q4:** The light black diamond in the bottom row is the fixed effects meta-analytic effect size estimate. Instead of using a black horizontal line, the upper limit and lower limit of the confidence interval are indicated by the left and right points of the diamond. The center of the diamond is the meta-analytic effect size estimate. A meta-analysis calculates the effect size by combining and weighing all studies. Which statement is true?

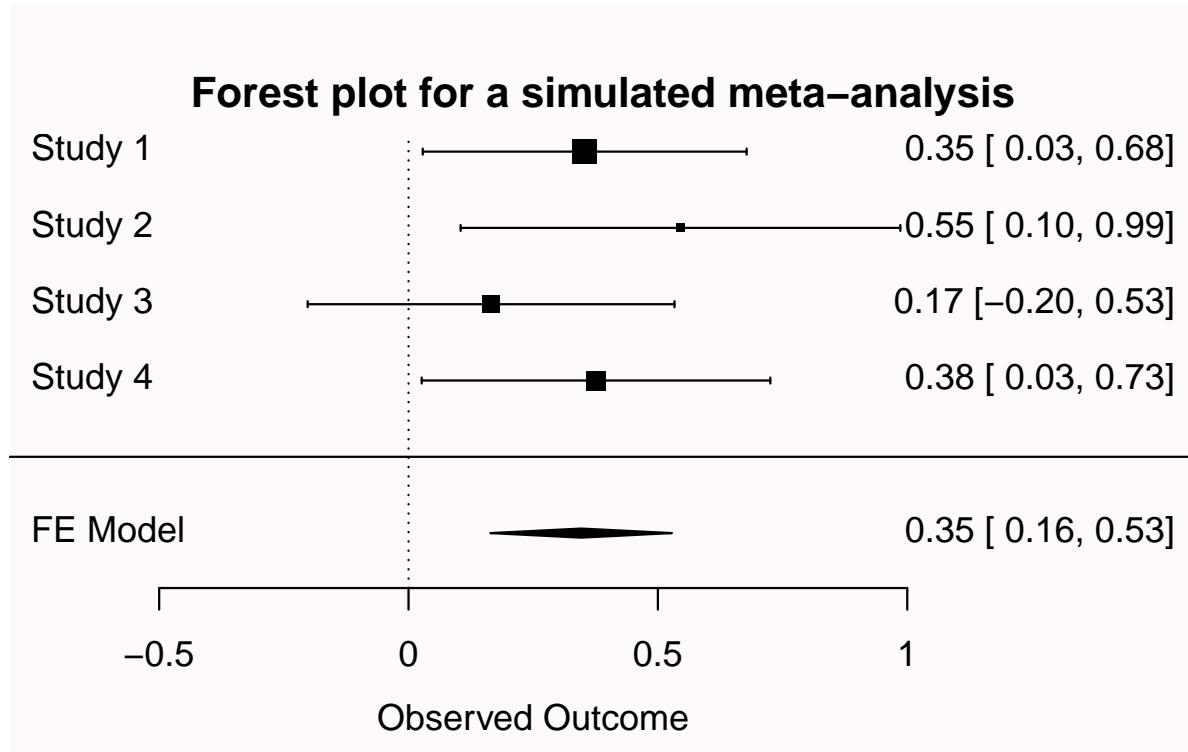


Figure 7.11: Meta-analysis of 4 studies.

- (A) The confidence interval for a fixed effect meta-analytic effect size estimate is always wider than that for a single study, because of the additional variation between studies.
- (B) The confidence interval for a fixed effect meta-analytic effect size estimate is always more narrow than that for a single study, because of the combined sample size of all studies included in the meta-analysis.
- (C) The confidence interval for a fixed effect meta-analytic effect size estimate does not become wider or more narrow compared to the confidence interval of a single study, it just becomes closer to the true population parameter.

**Q5:** Let's assume a researcher calculates a mean of 7.5, and a standard deviation of 6.3, in a sample of 20 people. The critical value for a  $t$ -distribution with  $df = 19$  is 2.093. Calculate the upper limit of the confidence interval around the mean using the formula below. Is it:

$$\mu \pm t_{df,1-(\alpha/2)}SE$$

- (A) 1.40
- (B) 2.95
- (C) 8.91
- (D) 10.45

Copy the code below into R and run the code. It will generate plots like the one in Figure 7.5. Run the entire script as often as you want (notice the variability in the  $p$ -values due to the relatively low power in the test!), to answer the following question. The  $p$ -value in the plot will tell you if the difference is statistically significant, and what the  $p$ -value is. Run the simulation until you find a  $p$ -value close to  $p = 0.05$ .

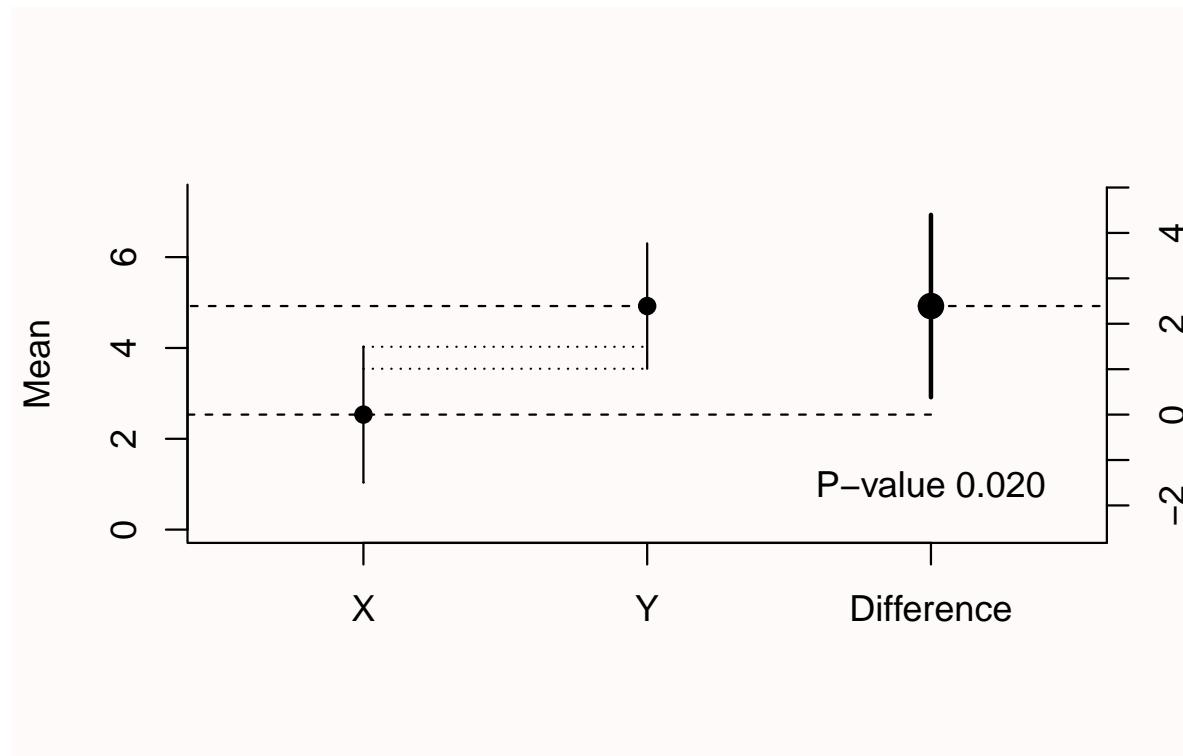
```
x <- rnorm(n = 50, mean = 3, sd = 5) # get sample group 1
y <- rnorm(n = 50, mean = 5, sd = 5) # get sample group 2

d <- data.frame(
  labels = c("X", "Y", "Difference"),
  mean = c(mean(x), mean(y), mean(y) - mean(x)),
  lower = c(t.test(x)[[4]][1], t.test(y)[[4]][1], t.test(y, x)[[4]][1]),
  upper = c(t.test(x)[[4]][2], t.test(y)[[4]][2], t.test(y, x)[[4]][2])
)
```

```

par(bg = backgroundcolor)
plot(NA, xlim = c(.5, 3.5), ylim = c(0, max(d$upper[1:2] + 1)), bty = "l",
      xaxt = "n", xlab = "", ylab = "Mean")
points(d$mean[1:2], pch = 19)
segments(0, d$mean[1], 3, d$mean[1], lty = 2)
segments(2, d$mean[2], -1, d$mean[2], lty = 2)
axis(1, 1:3, d$labels)
segments(1:2, d$lower[1:2], 1:2, d$upper[1:2])
axis(4, seq((d$mean[1] - 3), (d$mean[1] + 5), by = 1), seq(-3, 5, by = 1))
points(3, d$mean[1] + d$mean[3], pch = 19, cex = 1.5)
segments(3, d$mean[1] + d$lower[3], 3, d$mean[1] + d$upper[3], lwd = 2)
segments(3, d$mean[2], 5, d$mean[2], lty = 2)
mtext("Difference", side = 4, at = d$mean[1], line = 3)
segments(1:1, d$upper[1:1], 1:2, d$upper[1:1], lty = 3)
segments(1:1, d$lower[1:2], 1:2, d$lower[1:2], lty = 3)
text(3, 1, paste("P-value", sprintf("%.3f", t.test(x, y)$p.value)))

```



**Q6:** How much do two 95% confidence intervals around individual means from independent

groups overlap when the mean difference between the two means is only just statistically significant ( $p = 0.05$  at an alpha of 0.05)?

- (A) When the 95% confidence interval around one mean does not contain the mean of the other group, the groups always differ significantly from each other.
- (B) When the 95% confidence interval around one mean does not overlap with the 95% confidence interval of the mean of the other group, the groups always differ significantly from each other.
- (C) When the overlap between the two confidence intervals around each mean overlap a little bit (the upper bound of the CI overlaps with the lower quarter of the confidence interval around the other mean) the groups differ significantly from each other at approximately  $p = 0.05$ .
- (D) There is no relationship between the overlap of the 95% confidence intervals around two independent means, and the  $p$ -value for the difference between these groups.

Note that this visual overlap rule can only be used when the comparison is made between independent groups, not between dependent groups! The 95% confidence interval around effect sizes is therefore typically more easily interpretable in relation to the significance of a test.

Let's experience this through simulation. The simulation in the R script below generates a large number of additional samples, after the initial one that was plotted. The simulation returns the number of CI that contains the mean (which should be 95% in the long run). The simulation also returns the % of means from future studies that fall within the 95% of the original study, or the capture percentage. It differs from (and is often lower, but sometimes higher, than) the confidence interval.

```
library(ggplot2)

n <- 20 # set sample size
nsims <- 100000 # set number of simulations

x <- rnorm(n = n, mean = 100, sd = 15) # create sample from normal distribution

# 95% Confidence Interval
ciu <- mean(x) + qt(0.975, df = n - 1) * sd(x) * sqrt(1 / n)
cil <- mean(x) - qt(0.975, df = n - 1) * sd(x) * sqrt(1 / n)

# 95% Prediction Interval
piu <- mean(x) + qt(0.975, df = n - 1) * sd(x) * sqrt(1 + 1 / n)
```

```

pil <- mean(x) - qt(0.975, df = n - 1) * sd(x) * sqrt(1 + 1 / n)

ggplot(as.data.frame(x), aes(x)) + # plot data
  geom_rect(aes(xmin = pil, xmax = piu, ymin = 0, ymax = Inf),
            fill = "gold") + # draw yellow PI area
  geom_rect(aes(xmin = cil, xmax = ciu, ymin = 0, ymax = Inf),
            fill = "#E69F00") + # draw orange CI area
  geom_histogram(colour = "black", fill = "grey", aes(y = after_stat(density)), bins = 20) +
  xlab("Score") +
  ylab("frequency") +
  theme_bw(base_size = 20) +
  theme(panel.grid.major.x = element_blank(), axis.text.y = element_blank(),
        panel.grid.minor.x = element_blank()) +
  geom_vline(xintercept = mean(x), linetype = "dashed", linewidth = 1) +
  coord_cartesian(xlim = c(50, 150)) +
  scale_x_continuous(breaks = c(seq(50, 150, 10))) +
  annotate("text", x = mean(x), y = 0.02, label = paste(
    "Mean = ", round(mean(x)), "\n",
    "SD = ", round(sd(x)), sep = ""), size = 6.5)

# Simulate Confidence Intervals
ciu_sim <- numeric(nsims)
cil_sim <- numeric(nsims)
mean_sim <- numeric(nsims)

for (i in 1:nsims) { # for each simulated experiment
  x <- rnorm(n = n, mean = 100, sd = 15) # create sample from normal distribution
  ciu_sim[i] <- mean(x) + qt(0.975, df = n - 1) * sd(x) * sqrt(1 / n)
  cil_sim[i] <- mean(x) - qt(0.975, df = n - 1) * sd(x) * sqrt(1 / n)
  mean_sim[i] <- mean(x) # store means of each sample
}

# Save only those simulations where the true value was inside the 95% CI
ciu_sim <- ciu_sim[ciu_sim < 100]
cil_sim <- cil_sim[cil_sim > 100]

# Calculate how many times the observed mean fell within the 95% CI of the original study
mean_sim <- mean_sim[mean_sim > cil & mean_sim < ciu]

cat((100 * (1 - (length(ciu_sim) / nsims + length(cil_sim) / nsims))), 
    "% of the 95% confidence intervals contained the true mean")
cat("The capture percentage for the plotted study, or the % of values within"

```

```
the observed confidence interval from", cil, "to", ciu,
"is:", 100 * length(mean_sim) / nsims, "%")
```

**Q7:** Run the simulations multiple times. Look at the output you will get in the R console. For example: “95.077 % of the 95% confidence intervals contained the true mean” and “The capture percentage for the plotted study, or the % of values within the observed confidence interval from 88.17208 to 103.1506 is: 82.377 %”. While running the simulations multiple times, look at the confidence interval around the sample mean, and relate this to the capture percentage. Run the simulation until you have seen a range of means closer and further away from the true mean in the simulation (100). Which statement is true?

- (A) The farther the sample mean is from the true population mean, the lower the capture percentage.
- (B) The farther the sample mean is from the true population mean, the higher the capture percentage.

**Q8:** Simulations in R are randomly generated, but you can make a specific simulation reproducible by setting the seed of the random generation process. Copy-paste “set.seed(1000)” to the first line of the R script, and run the simulation. The sample mean should be 94. What is the capture percentage? (Don’t forget to remove the set.seed command if you want to generate more random simulations!).

- (A) 95%
- (B) 42.1%
- (C) 84.3%
- (D) 89.2%

Capture percentages are rarely directly used to make statistical inferences. The main reason we discuss them here is really to prevent the common misunderstanding that 95% of future means fall within a single confidence interval: Capture percentages clearly show that is not true. Prediction intervals are also rarely used in psychology, but are more common in data science.

**Q9** So far we have looked at confidence intervals around means, but we can also compute confidence intervals around standard deviations. If you run lines the first lines of the code below, you will see that with an alpha level of 0.05, 100 observations, and a true standard deviation of 1, the 95% CI around the standard deviation is [0.88; 1.16]. Change the assumed

population standard deviation from 1 to 2 (`st_dev <- 2`). Keep all other settings the same. What is the 95% CI around the standard deviation of 2 with 100 observations?

```
alpha_level <- 0.05 # set alpha level
n <- 100 # set number of observations
st_dev <- 1 # set true standard deviation
effect <- 0.5 # set effect size (raw mean difference)

# calculate lower and upper critical values c_l and c_u
c_l <- sqrt((n - 1)/qchisq(alpha_level/2, n - 1, lower.tail = FALSE))
c_u <- sqrt((n - 1)/qchisq(alpha_level/2, n - 1, lower.tail = TRUE))

# calculate lower and upper confidence interval for sd
st_dev * c_l
st_dev * c_u

# d based on lower bound of the 95CI around the SD
effect/(st_dev * c_l)
# d based on upper bound of the 95CI around the SD
effect/(st_dev * c_u)

pwr::pwr.t.test(d = effect/(st_dev * c_l), power = 0.9, sig.level = 0.05)
pwr::pwr.t.test(d = effect/(st_dev * c_u), power = 0.9, sig.level = 0.05)

# Power analysis for true standard deviation for comparison
pwr::pwr.t.test(d = effect/st_dev, power = 0.9, sig.level = 0.05)
```

- (A) 95% CI [1.38; 3.65]
- (B) 95% CI [1.76; 2.32]
- (C) 95% CI [1.82; 2.22]
- (D) 95% CI [1.84; 2.20]

**Q10:** Change the assumed population standard deviation back from 2 to 1. Lower the sample size from 100 to 20 (`n <- 20`). This will inform us about the width of the confidence interval for a standard deviation when we run a pilot study with 20 observations. Keep all other settings the same. What is the 95% CI around the standard deviation of 1 with 20 observations?

- (A) 95% CI [0.91; 1.11]

- (B) 95% CI [0.82; 1.28]
- (C) 95% CI [0.76; 1.46]
- (D) 95% CI [1.52; 2.92]

**Q11:** If we want the 95% CI around the standard deviation of 1 to be at most 0.05 away from the assumed population standard deviation, how large should our number of observations be? Note that this means we want the 95% CI to fall within 0.95 and 1.05. But notice from the calculations above that the distribution of the sample standard deviations is not symmetrical. Standard deviations can't be smaller than 0 (because they are the square rooted variance). So in practice the question is: What is the **smallest** number of observations for the upper 95% CI to be smaller than 1.05? Replace n with each of the values in the answer options.

- (A)  $n = 489$
- (B)  $n = 498$
- (C)  $n = 849$
- (D)  $n = 948$

Let's explore what the consequences of an inaccurate estimate of the population standard deviation are on a-priori power analyses. Let's imagine we want to perform an a-priori power analysis for a smallest effect size of interest of half a scale point (on a scale from 1-5) on a measure that has an (unknown) true population standard deviation of 1.2.

**Q12:** Change the number of observations to 50. Change the assumed population standard deviation to 1.2. Keep the effect as 0.5. The 95% confidence interval for the standard deviation based on a sample of 50 observation ranges from 1.002 to 1.495. To perform an a-priori power analysis we need to calculate Cohen's d, which is the difference divided by the standard deviation. In our example, we want to at least observe a difference of 0.5. What is Cohen's d (effect/SD) for the lower bound of the 95% confidence interval (where SD = 1.002) or the upper bound (where SD = 1.495)?

- (A)  $d = 0.33$  and  $d = 0.50$
- (B)  $d = 0.40$  and  $d = 0.60$
- (C)  $d = 0.43$  and  $d = 0.57$

- (D)  $d = 0.29$  and  $d = 0.55$

If we draw a sample of 50 observations we can happen to observe a value that, due to random variation, is much smaller or much larger than the true population value. We can examine the effect this has on the number of observations that we think will be required when we perform an a-priori power analysis.

**Q13:** An a-priori power analysis is performed that uses the estimate of Cohen's d based on the lower 95% CI of the standard deviation. Which statement is true?

- (A) Because the lower bound of the 95% CI is **smaller** than the true population SD, Cohen's d is **smaller**, and the a-priori power analysis will yield a sample size that is **smaller** than the sample size we really need.
- (B) Because the lower bound of the 95% CI is **smaller** than the true population SD, Cohen's d is **larger**, and the a-priori power analysis will yield a sample size that is **larger** than the sample size we really need.
- (C) Because the lower bound of the 95% CI is **smaller** than the true population SD, Cohen's d is **smaller**, and the a-priori power analysis will yield a sample size that is **larger** than the sample size we really need.
- (D) Because the lower bound of the 95% CI is **smaller** than the true population SD, Cohen's d is **larger**, and the a-priori power analysis will yield a sample size that is **smaller** than the sample size we really need.

**Q14:** Let's check if our answer on the previous question was correct. We still have an alpha level of 0.05,  $n = 50$ , a standard deviation of 1.2, and an effect of interest of 0.5. Run the power analyses using the `pwr` package. The first power analysis uses Cohen's d based on the lower bound of the 95% confidence interval. The second power analysis uses the upper bound of the 95% confidence interval. (There is also a third power analysis based on the (in real-life situations unknown) true standard deviation, just for comparison). Which statement is true (note that the sample size for a power analysis is rounded up, as we can't collect a partial observation)?

- (A) The sample size per group is 68 when calculating the effect size based on the lower bound on the 95% CI around the standard deviation, and 86 when using the upper bound of the 95% CI around the standard deviation.
- (B) The sample size per group is 68 when calculating the effect size based on the lower bound on the 95% CI around the standard deviation, and 123 when using the

upper bound of the 95% CI around the standard deviation.

- (C) The sample size per group is 86 when calculating the effect size based on the lower bound on the 95% CI around the standard deviation, and 123 when using the upper bound of the 95% CI around the standard deviation.
- (D) The sample size per group is 86 when calculating the effect size based on the lower bound on the 95% CI around the standard deviation, and 189 when using the upper bound of the 95% CI around the standard deviation.

### **7.11.1 Open Questions**

1. What is the definition of a confidence interval?
2. How is a confidence interval related to statistical significance?
3. What happens to a confidence interval when the sample size increases?
4. What is the difference between a confidence interval and a capture percentage?
5. What is a prediction interval?
6. If you have data from the entire population, do you need to calculate a confidence interval?
7. What are confidence intervals a statement about?
8. What does it mean to say that after you have collected the data, the confidence interval either contains the true parameter, or it doesn't?
9. What is the difference, all else equal, between estimates from small vs. large samples?
10. Why do researchers rarely (if ever) compute confidence intervals around standard deviations? What would be a situation where it could be interesting to report confidence intervals around standard deviations?

# 8 Sample Size Justification

You can listen to an audio recording of this chapter [here](#). You can download a translation of this chapter in Chinese [here](#)

Scientists perform empirical studies to collect data that helps to answer a research question. The more data that is collected, the more informative the study will be with respect to its inferential goals. A sample size justification should consider how informative the data will be given an inferential goal, such as estimating an effect size, or testing a hypothesis. Even though a sample size justification is sometimes requested in manuscript submission guidelines, when submitting a grant to a funder, or submitting a proposal to an ethical review board, the number of observations is often simply *stated*, but not *justified*. This makes it difficult to evaluate how informative a study will be. To prevent such concerns from emerging when it is too late (e.g., after a non-significant hypothesis test has been observed), researchers should carefully justify their sample size before data is collected. In this chapter, which is largely identical to Daniël Lakens (2022a), we will explore in detail how to justify your sample size.

Table 8.1: Overview of possible justifications for the sample size in a study.

Type of justification	When is this justification applicable?
Measure entire population	A researcher can specify the entire population, it is finite, and it is possible to measure (almost) every entity in the population.
Resource constraints	Limited resources are the primary reason for the choice of the sample size a researcher can collect.
Accuracy	The research question focusses on the size of a parameter, and a researcher collects sufficient data to have an estimate with a desired level of accuracy.
A-priori power analysis	The research question aims to test whether certain effect sizes can be statistically rejected with a desired statistical power.

Type of justification	When is this justification applicable?
Heuristics	A researcher decides upon the sample size based on a heuristic, general rule or norm that is described in the literature, or communicated orally.
No justification	A researcher has no reason to choose a specific sample size, or does not have a clearly specified inferential goal and wants to communicate this honestly.

## 8.1 Six Approaches to Justify Sample Sizes

Researchers often find it difficult to justify their sample size (i.e., a number of participants, observations, or any combination thereof). In this review article six possible approaches are discussed that can be used to justify the sample size in a quantitative study (see Table 8.1). This is not an exhaustive overview, but it includes the most common and applicable approaches for single studies. The topic of power analysis for meta-analyses is outside the scope of this chapter, but see Hedges & Pigott (2001) and Valentine et al. (2010). The first justification is that data from (almost) the entire population has been collected. The second justification centers on resource constraints, which are almost always present, but rarely explicitly evaluated. The third and fourth justifications are based on a desired statistical power or a desired accuracy. The fifth justification relies on heuristics, and finally, researchers can choose a sample size without any justification. Each of these justifications can be stronger or weaker depending on which conclusions researchers want to draw from the data they plan to collect.

All of these approaches to the justification of sample sizes, even the ‘no justification’ approach, give others insight into the reasons that led to the decision for a sample size in a study. It should not be surprising that the ‘heuristics’ and ‘no justification’ approaches are often unlikely to impress peers. However, it is important to note that the value of the information that is collected depends on the extent to which the final sample size allows a researcher to achieve their inferential goals, and not on the sample size justification that is chosen.

The extent to which these approaches make other researchers judge the data that is collected as *informative* depends on the details of the question a researcher aimed to answer and the parameters they chose when determining the sample size for their study. For example, a badly performed a-priori power analysis can quickly lead to a study with very low informational value. These six justifications are not mutually exclusive, and multiple approaches can be considered when designing a study.

## 8.2 Six Ways to Evaluate Which Effect Sizes are Interesting

The informativeness of the data that is collected depends on the inferential goals a researcher has, or in some cases, the inferential goals scientific peers will have. A shared feature of the different inferential goals considered in this review article is the question which effect sizes a researcher considers meaningful to distinguish. This implies that researchers need to evaluate which effect sizes they consider interesting. These evaluations rely on a combination of statistical properties and domain knowledge. In Table 8.2 six possibly useful considerations are provided. This is not intended to be an exhaustive overview, but it presents common and useful approaches that can be applied in practice. Not all evaluations are equally relevant for all types of sample size justifications. The online Shiny app accompanying Daniël Lakens (2022a) provides researchers with an interactive form that guides researchers through the considerations for a sample size justification. These considerations often rely on the same information (e.g., effect sizes, the number of observations, the standard deviation, etc.) so these six considerations should be seen as a set of complementary approaches that can be used to evaluate which effect sizes are of interest.

To start, researchers should consider what their smallest effect size of interest is. Second, although only relevant when performing a hypothesis test, researchers should consider which effect sizes could be statistically significant given a choice of an alpha level and sample size. Third, it is important to consider the (range of) effect sizes that are expected. This requires a careful consideration of the source of this expectation and the presence of possible biases in these expectations. Fourth, it is useful to consider the width of the confidence interval around possible values of the effect size in the population, and whether we can expect this confidence interval to reject effects we considered a-priori plausible. Fifth, it is worth evaluating the power of the test across a wide range of possible effect sizes in a sensitivity power analysis. Sixth, a researcher can consider the effect size distribution of related studies in the literature.

Table 8.2: Overview of possible ways to evaluate which effect sizes are interesting.

Type of evaluation	Which question should a researcher ask?
Smallest effect size of interest	What is the smallest effect size that is considered theoretically or practically interesting?
The minimal statistically detectable effect	Given the test and sample size, what is the critical effect size that can be statistically significant?
Expected effect size	Which effect size is expected based on theoretical predictions or previous research?
Width of confidence interval	Which effect sizes are excluded based on the expected width of the confidence interval around the effect size?

Type of evaluation	Which question should a researcher ask?
Sensitivity power analysis	Across a range of possible effect sizes, which effects does a design have sufficient power to detect when performing a hypothesis test?
Distribution of effect sizes in a research area	What is the empirical range of effect sizes in a specific research area and which effects are <i>a priori</i> unlikely to be observed?

### 8.3 The Value of Information

Since all scientists are faced with resource limitations, they need to balance the cost of collecting each additional datapoint against the increase in information that datapoint provides. This is referred to as the *value of information* (Eckermann et al., 2010). Calculating the value of information is notoriously difficult (Detsky, 1990). Researchers need to specify the cost of collecting data, and weigh the costs of data collection against the increase in utility that having access to the data provides. From a value of information perspective not every data point that can be collected is equally valuable (J. Halpern et al., 2001; Wilson, 2015). Whenever additional observations do not change inferences in a meaningful way, the costs of data collection can outweigh the benefits.

The value of additional information will in most cases be a non-monotonic function, especially when it depends on multiple inferential goals. A researcher might be interested in comparing an effect against a previously observed large effect in the literature, a theoretically predicted medium effect, and the smallest effect that would be practically relevant. In such a situation the expected value of sampling information will lead to different optimal sample sizes for each inferential goal. It could be valuable to collect informative data about a large effect, with additional data having less (or even a negative) marginal utility, up to a point where the data becomes increasingly informative about a medium effect size, with the value of sampling additional information decreasing once more until the study becomes increasingly informative about the presence or absence of a smallest effect of interest.

Because of the difficulty of quantifying the value of information, scientists typically use less formal approaches to justify the amount of data they set out to collect in a study. Even though the cost-benefit analysis is not always made explicit in reported sample size justifications, the value of information perspective is almost always implicitly the underlying framework that sample size justifications are based on. Throughout the subsequent discussion of sample size justifications, the importance of considering the value of information given inferential goals will repeatedly be highlighted.

## 8.4 Measuring (Almost) the Entire Population

In some instances it might be possible to collect data from (almost) the entire population under investigation. For example, researchers might use census data, are able to collect data from all employees at a firm or study a small population of top athletes. Whenever it is possible to measure the entire population, the sample size justification becomes straightforward: the researcher used all the data that is available.

## 8.5 Resource Constraints

A common reason for the number of observations in a study is that resource constraints limit the amount of data that can be collected at a reasonable cost (Lenth, 2001). In practice, sample sizes are always limited by the resources that are available. Researchers practically always have resource limitations, and therefore even when resource constraints are not the primary justification for the sample size in a study, it is always a secondary justification.

Despite the omnipresence of resource limitations, the topic often receives little attention in texts on experimental design (for an example of an exception, see Metin Bulus & Dong (2021)). This might make it feel like acknowledging resource constraints is not appropriate, but the opposite is true: Because resource limitations always play a role, a responsible scientist carefully evaluates resource constraints when designing a study. Resource constraint justifications are based on a trade-off between the costs of data collection, and the value of having access to the information the data provides. Even if researchers do not explicitly quantify this trade-off, it is revealed in their actions. For example, researchers rarely spend all the resources they have on a single study. Given resource constraints, researchers are confronted with an optimization problem of how to spend resources across multiple research questions.

Time and money are two resource limitations all scientists face. A PhD student has a certain time to complete a PhD thesis, and is typically expected to complete multiple research lines in this time. In addition to time limitations, researchers have limited financial resources that often directly influence how much data can be collected. A third limitation in some research lines is that there might simply be a very small number of individuals from whom data can be collected, such as when studying patients with a rare disease. A resource constraint justification puts limited resources at the center of the justification for the sample size that will be collected, and *starts* with the resources a scientist has available. These resources are translated into an expected number of observations ( $N$ ) that a researcher expects they will be able to collect with an amount of money in a given time. The challenge is to evaluate whether collecting  $N$  observations is worthwhile. How do we decide if a study will be informative, and when should we conclude that data collection is *not* worthwhile?

When evaluating whether resource constraints make data collection uninformative, researchers need to explicitly consider which inferential goals they have when collecting data (Parker &

Berman, 2003). Having data always provides more knowledge about the research question than not having data, so in an absolute sense, all data that is collected has value. However, it is possible that the benefits of collecting the data are outweighed by the costs of data collection.

It is most straightforward to evaluate whether data collection has value when we know for certain that someone will make a decision, with or without data. In such situations any additional data will reduce the error rates of a well-calibrated decision process, even if only ever so slightly. For example, without data we will not perform better than a coin flip if we guess which of two conditions has a higher true mean score on a measure. With some data, we can perform better than a coin flip by picking the condition that has the highest mean. With a small amount of data we would still very likely make a mistake, but the error rate is smaller than without any data. In these cases, the value of information might be positive, as long as the reduction in error rates is more beneficial than the cost of data collection.

Another way in which a small dataset can be valuable is if its existence eventually makes it possible to perform a meta-analysis (Maxwell & Kelley, 2011). This argument in favor of collecting a small dataset requires 1) that researchers share the data in a way that a future meta-analyst can find it, and 2) that there is a decent probability that someone will perform a high-quality meta-analysis that will include this data in the future (S. D. Halpern et al., 2002). The uncertainty about whether there will ever be such a meta-analysis should be weighed against the costs of data collection.

One way to increase the probability of a future meta-analysis is if researchers commit to performing this meta-analysis themselves, by combining several studies they have performed into a small-scale meta-analysis (Cumming, 2014). For example, a researcher might plan to repeat a study for the next 12 years in a class they teach, with the expectation that after 12 years a meta-analysis of 12 studies would be sufficient to draw informative inferences (but see ter Schure & Grünwald (2019)). If it is not plausible that a researcher will collect all the required data by themselves, they can attempt to set up a collaboration where fellow researchers in their field commit to collecting similar data with identical measures. If it is not likely that sufficient data will emerge over time to reach the inferential goals, there might be no value in collecting the data.

Even if a researcher believes it is worth collecting data because a future meta-analysis will be performed, they will most likely perform a statistical test on the data. To make sure their expectations about the results of such a test are well-calibrated, it is important to consider which effect sizes are of interest, and to perform a sensitivity power analysis to evaluate the probability of a Type II error for effects of interest. From the six ways to evaluate which effect sizes are interesting that will be discussed in the second part of this review, it is useful to consider the smallest effect size that can be statistically significant, the expected width of the confidence interval around the effect size, and effects that can be expected in a specific research area, and to evaluate the power for these effect sizes in a sensitivity power analysis. If a decision or claim is made, a compromise power analysis is worthwhile to consider when deciding upon the error rates while planning the study. When reporting a resource constraints sample size justification

it is recommended to address the five considerations in Table 8.3). Addressing these points explicitly facilitates evaluating if the data is worthwhile to collect. To make it easier to address all relevant points explicitly, an interactive form to implement the recommendations in this chapter can be found at [https://shiny.ieis.tue.nl/sample\\_size\\_justification/](https://shiny.ieis.tue.nl/sample_size_justification/).

Table 8.3: Overview of recommendations when reporting a sample size justification based on resource constraints.

What to address	How to address it?
Will a future meta-analysis be performed?	Consider the plausibility that sufficient highly similar studies will be performed in the future to make a meta-analysis possible.
Will a decision or claim be made regardless of the amount of data that is available?	If a decision is made then any data that is collected will reduce error rates. Consider using a compromise power analysis to determine Type I and Type II error rates. Are the costs worth the reduction in errors?
What is the critical effect size?	Report and interpret the critical effect size, with a focus on whether expected effect sizes could yield significant results. If not indicate the interpretation of the data will not be based on p values.
What is the width of the confidence interval?	Report and interpret the width of the confidence interval. What will an estimate with this much uncertainty be useful for? If the null hypothesis is true, would rejecting effects outside of the confidence interval be worthwhile (ignoring how a design might have low power to actually test against these values)?
Which effect sizes will a design have decent power to detect?	Report a sensitivity power analysis, and report the effect sizes that can be detected across a range of desired power levels (e.g., 80%, 90%, and 95%) or plot a sensitivity analysis.

## 8.6 A-priori Power Analysis

When designing a study where the goal is to test whether a statistically significant effect is present, researchers often want to make sure their sample size is large enough to prevent erroneous conclusions for a range of effect sizes they care about. In this approach to justifying

a sample size, the value of information is to collect observations up to the point that the probability of an erroneous inference is, in the long run, not larger than a desired value. If a researcher performs a hypothesis test, there are four possible outcomes:

1. A false positive (or Type I error), determined by the  $\alpha$  level. A test yields a significant result, even though the null hypothesis is true.
2. A false negative (or Type II error), determined by  $\beta$ , or  $1 - \text{power}$ . A test yields a non-significant result, even though the alternative hypothesis is true.
3. A true negative, determined by  $1 - \alpha$ . A test yields a non-significant result when the null hypothesis is true.
4. A true positive, determined by  $1 - \beta$ . A test yields a significant result when the alternative hypothesis is true.

Given a specified effect size, alpha level, and power, an a-priori power analysis can be used to calculate the number of observations required to achieve the desired error rates, given the effect size. Power analyses can be performed based on standardized effect sizes or effect sizes expressed on the original scale. It is important to know the standard deviation of the effect (see the ‘Know Your Measure’ section) but I find it slightly more convenient to talk about standardized effects in the context of sample size justifications. Figure 8.1 illustrates how the statistical power increases as the number of observations (per group) increases in an independent  $t$  test with a two-sided alpha level of 0.05. If we are interested in detecting an effect of  $d = 0.5$ , a sample size of 90 per condition would give us more than 90% power. Statistical power can be computed to determine the number of participants, or the number of items (Westfall et al., 2014) but can also be performed for single case studies (Ferron & Onghena, 1996; McIntosh & Rittmo, 2021).

Although it is common to set the Type I error rate to 5% and aim for 80% power, error rates should be justified (Daniël Lakens, Adolfi, et al., 2018). As explained in the section on compromise power analysis, the default recommendation to aim for 80% power lacks a solid justification. In general, the lower the error rates (and thus the higher the power), the more informative a study will be, but the more resources are required. Researchers should carefully weigh the costs of increasing the sample size against the benefits of lower error rates, which would probably make studies designed to achieve a power of 90% or 95% more common for articles reporting a single study. An additional consideration is whether the researcher plans to publish an article consisting of a set of replication and extension studies, in which case the probability of observing multiple Type I errors will be very low, but the probability of observing mixed results even when there is a true effect increases (Daniël Lakens & Etz, 2017), which would also be a reason to aim for studies with low Type II error rates, perhaps even by slightly increasing the alpha level for each individual study.

Figure 8.2 visualizes two distributions. The left distribution (dashed line) is centered at 0. This is a model for the null hypothesis. If the null hypothesis is true a statistically significant result will be observed if the effect size is extreme enough (in a two-sided test either in the positive or negative direction), but any significant result would be a Type I error (the dark

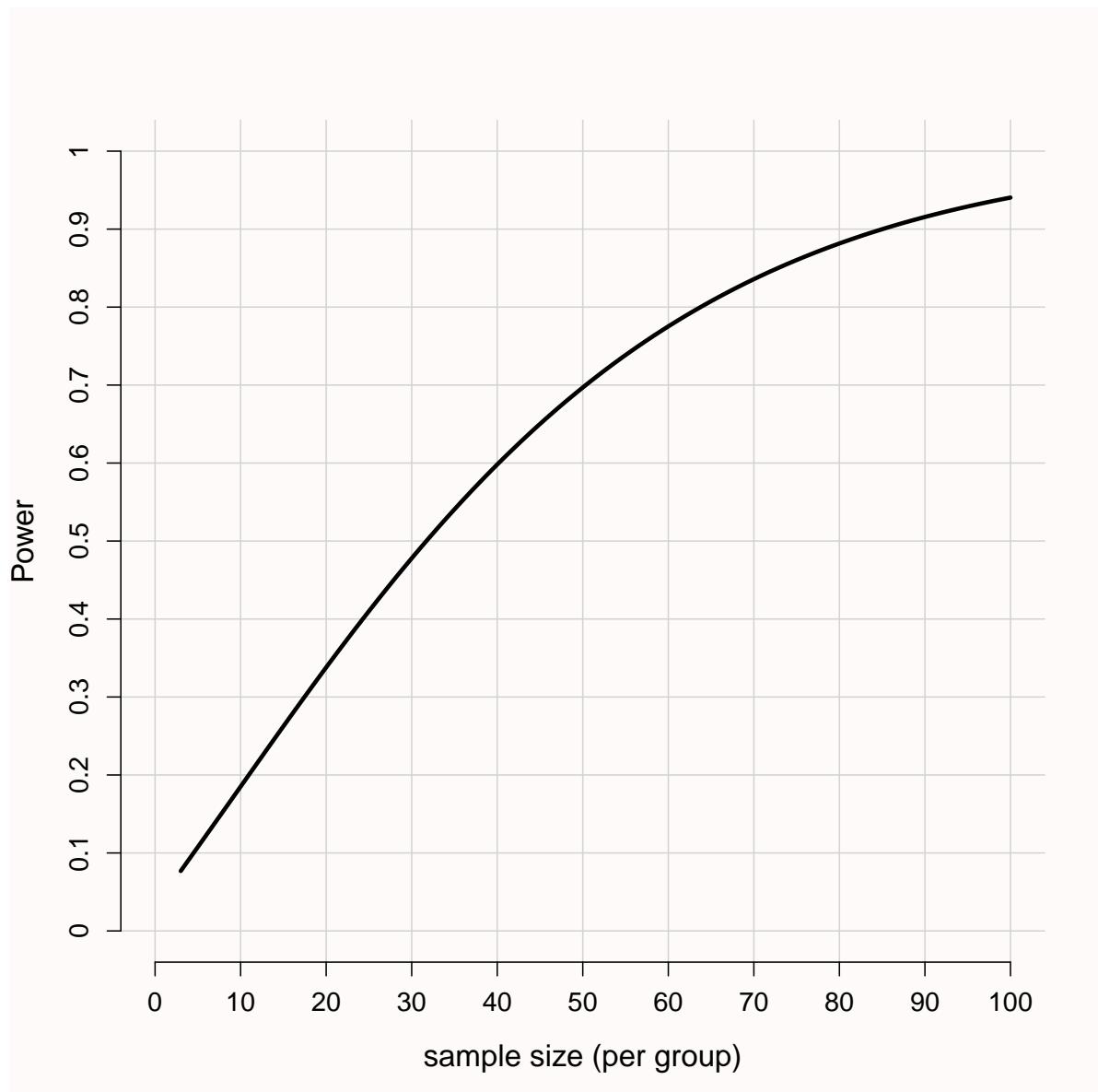


Figure 8.1: Power curve for an independent  $t$  test with an effect of  $d = 0.5$  and  $\alpha = 0.05$  as a function of the sample size.

grey areas under the curve). If there is no true effect, formally statistical power for a null hypothesis significance test is undefined. Any significant effects observed if the null hypothesis is true are Type I errors, or false positives, which occur at the chosen alpha level. The right distribution (solid line) is centered on an effect of  $d = 0.5$ . This is the specified model for the alternative hypothesis in this study, illustrating the expectation of an effect of  $d = 0.5$  if the alternative hypothesis is true. Even though there is a true effect, studies will not always find a statistically significant result. This happens when, due to random variation, the observed effect size is too close to 0 to be statistically significant. Such results are false negatives (the light grey area under the curve on the right). To increase power, we can collect a larger sample size. As the sample size increases, the distributions become more narrow, reducing the probability of a Type II error. These figures can be reproduced and adapted in an [online shiny app](#).

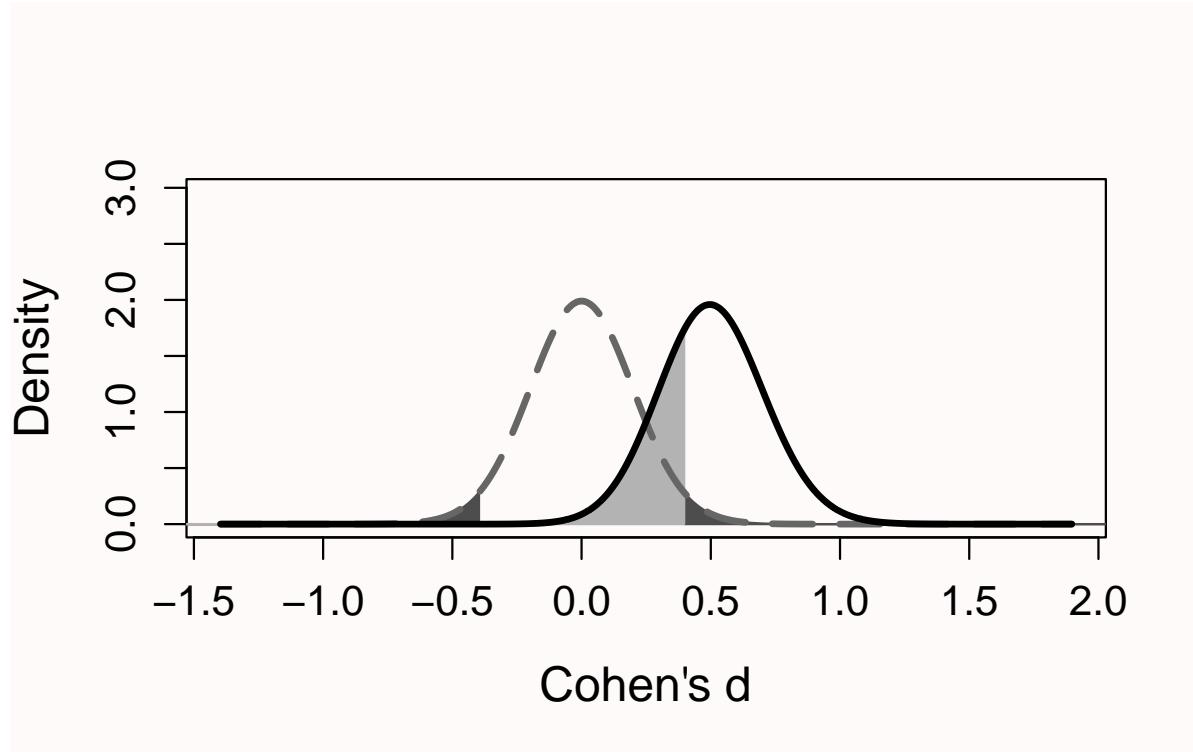


Figure 8.2: Null ( $d = 0$ , grey dashed line) and alternative ( $d = 0.5$ , solid black line) hypothesis, with  $\alpha = 0.05$  and  $n = 80$  per group.

It is important to highlight that the goal of an a-priori power analysis is *not* to achieve sufficient power for the true effect size. The true effect size is unknown. The goal of an a-priori power analysis is to achieve sufficient power, given a specific *assumption* of the effect size a researcher wants to detect. Just like a Type I error rate is the maximum probability of making a Type I error conditional on the assumption that the null hypothesis is true, an a-priori power analysis

is computed under the assumption of a specific effect size. It is unknown if this assumption is correct. All a researcher can do is to make sure their assumptions are well justified. Statistical inferences based on a test where the Type II error rate is controlled are conditional on the assumption of a specific effect size. They allow the inference that, assuming the true effect size is at least as large as that used in the a-priori power analysis, the maximum Type II error rate in a study is not larger than a desired value.

This point is perhaps best illustrated if we consider a study where an a-priori power analysis is performed both for a test of the *presence* of an effect, as for a test of the *absence* of an effect. When designing a study, it essential to consider the possibility that there is no effect (e.g., a mean difference of zero). An a-priori power analysis can be performed both for a null hypothesis significance test, as for a test of the absence of a meaningful effect, such as an equivalence test that can statistically provide support for the null hypothesis by rejecting the presence of effects that are large enough to matter (Daniël Lakens, 2017; Meyners, 2012; J. L. Rogers et al., 1993). When multiple primary tests will be performed based on the same sample, each analysis requires a dedicated sample size justification. If possible, a sample size is collected that guarantees that all tests are informative, which means that the collected sample size is based on the largest sample size returned by any of the a-priori power analyses.

For example, if the goal of a study is to detect or reject an effect size of  $d = 0.4$  with 90% power, and the alpha level is set to 0.05 for a two-sided independent  $t$  test, a researcher would need to collect 133 participants in each condition for an informative null hypothesis test, and 136 participants in each condition for an informative equivalence test. Therefore, the researcher should aim to collect 272 (that is, 136 participants in each condition) participants in total for an informative result for both tests that are planned. This does not guarantee a study has sufficient power for the true effect size (which can never be known), but it guarantees the study has sufficient power given an assumption of the effect a researcher is interested in detecting or rejecting. Therefore, an a-priori power analysis is useful, as long as a researcher can justify the effect sizes they are interested in.

If researchers correct the alpha level when testing multiple hypotheses, the a-priori power analysis should be based on this corrected alpha level. For example, if four tests are performed, an overall Type I error rate of 5% is desired, and a Bonferroni correction is used, the a-priori power analysis should be based on a corrected alpha level of .0125.

An a-priori power analysis can be performed analytically or by performing computer simulations. Analytic solutions are faster but less flexible. A common challenge researchers face when attempting to perform power analyses for more complex or uncommon tests is that available software does not offer analytic solutions. In these cases simulations can provide a flexible solution to perform power analyses for any test (Morris et al., 2019). The following code is an example of a power analysis in R based on 10000 simulations for a one-sample  $t$  test against zero for a sample size of 20, assuming a true effect of  $d = 0.5$ . All simulations consist of first randomly generating data based on assumptions of the data generating mechanism (e.g., a normal distribution with a mean of 0.5 and a standard deviation of 1), followed by a

test performed on the data. By computing the percentage of significant results, power can be computed for any design.

```
p <- numeric(10000) # to store p-values
for (i in 1:10000) { # simulate 10k tests
  x <- rnorm(n = 20, mean = 0.5, sd = 1)
  p[i] <- t.test(x)$p.value # store p-value
}
sum(p < 0.05) / 10000 # Compute power
```

There is a wide range of tools available to perform power analyses. Whichever tool a researcher decides to use, it will take time to learn how to use the software correctly to perform a meaningful a-priori power analysis. Resources to educate psychologists about power analysis consist of book-length treatments (Aberson, 2019; Cohen, 1988; Julious, 2004; Murphy et al., 2014), general introductions (Baguley, 2004; Brysbaert, 2019; Faul et al., 2007; Maxwell et al., 2008; Perugini et al., 2018), and an increasing number of applied tutorials for specific tests (Brysbaert & Stevens, 2018; DeBruine & Barr, 2021; P. Green & MacLeod, 2016; Kruschke, 2013; Daniël Lakens & Caldwell, 2021; Schoemann et al., 2017; Westfall et al., 2014). It is important to be trained in the basics of power analysis, and it can be extremely beneficial to learn how to perform simulation-based power analyses. At the same time, it is often recommended to enlist the help of an expert, especially when a researcher lacks experience with a power analysis for a specific test.

When reporting an a-priori power analysis, make sure that the power analysis is completely reproducible. If power analyses are performed in R it is possible to share the analysis script and information about the version of the package. In many software packages it is possible to export the power analysis that is performed as a PDF file. For example, in G\*Power analyses can be exported under the ‘protocol of power analysis’ tab. If the software package provides no way to export the analysis, add a screenshot of the power analysis to the supplementary files.

The reproducible report needs to be accompanied by justifications for the choices that were made with respect to the values used in the power analysis. If the effect size used in the power analysis is based on previous research, the factors presented in Table 8.5 (if the effect size is based on a meta-analysis) or Table 8.6 (if the effect size is based on a single study) should be discussed. If an effect size estimate is based on the existing literature, provide a full citation, and preferably a direct quote from the article where the effect size estimate is reported. If the effect size is based on a smallest effect size of interest, this value should not just be stated, but justified (e.g., based on theoretical predictions or practical implications, see Daniël Lakens, Scheel, et al. (2018)). For an overview of all aspects that should be reported when describing an a-priori power analysis, see Table 8.4.

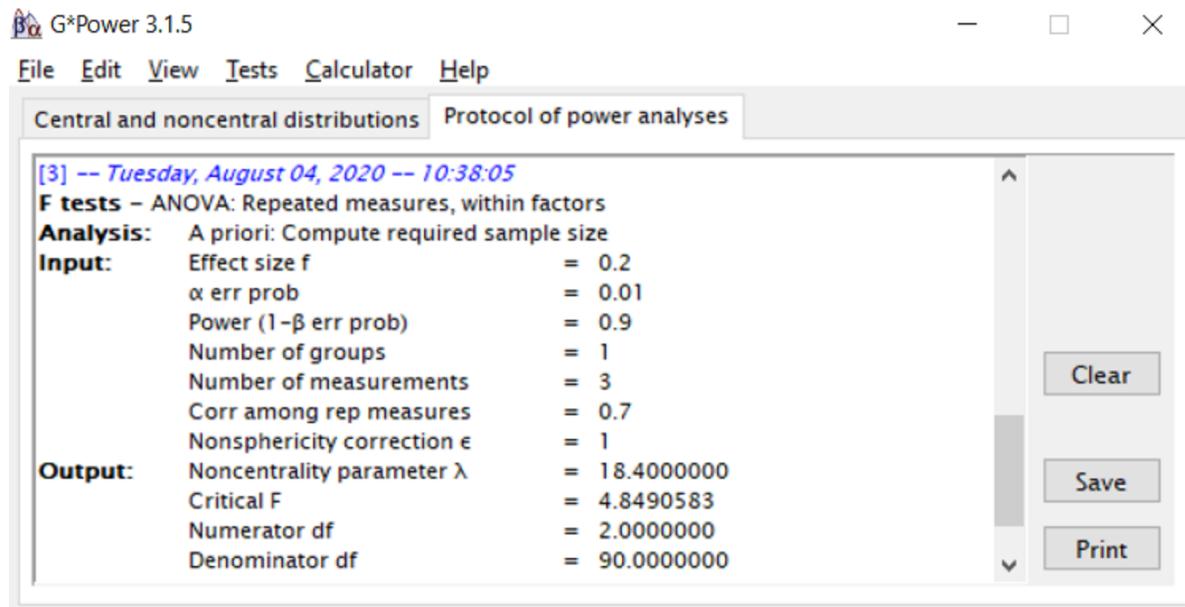


Figure 8.3: All details about the power analysis that is performed can be exported in G\*Power.

Table 8.4: Overview of recommendations when reporting an a-priori power analysis.

What to take into account?	How to take it into account?
List all primary analyses that are planned.	Specify all planned primary analyses that test hypotheses for which Type I and Type II error rates should be controlled.
Specify the alpha level for each analysis.	List and justify the Type I error rate for each analysis. Make sure to correct for multiple comparisons where needed.
What is the desired power?	List and justify the desired power (or Type II error rate) for each analysis.
For each power analysis, specify the effect size metric, the effect size, and the justification for powering for this effect size.	Report the effect size metric (e.g., Cohen's d, Cohen's f), the effect size (e.g., 0.3), and the justification for the effect size, and whether it is based on a smallest effect size of interest, a meta-analytic effect size estimate, the estimate of a single previous study, or some other source.
Consider the possibility that the null hypothesis is true.	Perform a power analysis for the test that is planned to examine the absence of a meaningful effect (e.g., power for an equivalence test).

What to take into account?	How to take it into account?
Make sure the power analysis is reproducible.	Include the code used to run the power analysis, or print a report containing the details about the power analyses that has been performed.

## 8.7 Planning for Precision

Some researchers have suggested to justify sample sizes based on a desired level of precision of the estimate (Cumming & Calin-Jageman, 2016; Kruschke, 2018; Maxwell et al., 2008). The goal when justifying a sample size based on precision is to collect data to achieve a desired width of the confidence interval around a parameter estimate. The width of the confidence interval around the parameter estimate depends on the standard deviation and the number of observations. The only aspect a researcher needs to justify for a sample size justification based on accuracy is the desired width of the confidence interval with respect to their inferential goal, and their assumption about the population standard deviation of the measure.

If a researcher has determined the desired accuracy, and has a good estimate of the true standard deviation of the measure, it is straightforward to calculate the sample size needed for a desired level of accuracy. For example, when measuring the IQ of a group of individuals a researcher might desire to estimate the IQ score within an error range of 2 IQ points for 95% of the observed means, in the long run. The required sample size to achieve this desired level of accuracy (assuming normally distributed data) can be computed by:

$$N = \left( \frac{z \cdot sd}{error} \right)^2$$

where  $N$  is the number of observations,  $z$  is the critical value related to the desired confidence interval,  $sd$  is the standard deviation of IQ scores in the population, and  $error$  is the width of the confidence interval within which the mean should fall, with the desired error rate. In this example,  $(1.96 \times 15 / 2)^2 = 216.1$  observations. If a researcher desires 95% of the means to fall within a 2 IQ point range around the true population mean, 217 observations should be collected. If a desired accuracy for a non-zero mean difference is computed, accuracy is based on a non-central  $t$ -distribution. For these calculations, an expected effect size estimate needs to be provided, but it has relatively little influence on the required sample size (Maxwell et al., 2008). It is also possible to incorporate uncertainty about the observed effect size in the sample size calculation, known as *assurance* (Kelley & Rausch, 2006). The MBESS package in R provides functions to compute sample sizes for a wide range of tests (Kelley, 2007).

What is less straightforward is to justify how a desired level of accuracy is related to inferential goals. There is no literature that helps researchers to choose a desired width of the confidence

interval. Morey (2020) convincingly argues that most practical use-cases of planning for precision involve an inferential goal of distinguishing an observed effect from other effect sizes (for a Bayesian perspective, see Kruschke (2018)). For example, a researcher might expect an effect size of  $r = 0.4$  and would treat observed correlations that differ more than 0.2 (i.e.,  $0.2 < r < 0.6$ ) differently, in that effects of  $r = 0.6$  or larger are considered too large to be caused by the assumed underlying mechanism (Hilgard, 2021), while effects smaller than  $r = 0.2$  are considered too small to support the theoretical prediction. If the goal is indeed to get an effect size estimate that is precise enough so that two effects can be differentiated with high probability, the inferential goal is actually a hypothesis test, which requires designing a study with sufficient power to reject effects (e.g., testing a range prediction of correlations between 0.2 and 0.6).

If researchers do not want to test a hypothesis, for example because they prefer an estimation approach over a testing approach, then in the absence of clear guidelines that help researchers to justify a desired level of precision, one solution might be to rely on a generally accepted norm of precision. This norm could be based on ideas about a certain resolution below which measurements in a research area no longer lead to noticeably different inferences. Just as researchers normatively use an alpha level of 0.05, they could plan studies to achieve a desired confidence interval width around the observed effect that is determined by a norm. Future work is needed to help researchers choose a confidence interval width when planning for accuracy (see also the section on which confidence interval to use in Bayesian tests of [range predictions](#)).

## 8.8 Heuristics

When a researcher uses a heuristic, they are not able to justify their sample size themselves, but they trust in a sample size recommended by some authority. When I started as a PhD student in 2005 it was common to collect 15 participants in each between subject condition. When asked why this was a common practice, no one was really sure, but people trusted that there was a justification somewhere in the literature. Now, I realize there was no justification for the heuristics we used. As Berkeley (1735) already observed: “Men learn the elements of science from others: And every learner hath a deference more or less to authority, especially the young learners, few of that kind caring to dwell long upon principles, but inclining rather to take them upon trust: And things early admitted by repetition become familiar: And this familiarity at length passeth for evidence.”

Some papers provide researchers with simple rules of thumb about the sample size that should be collected. Such papers clearly fill a need, and are cited a lot, even when the advice in these articles is flawed. For example, Wilson VanVoorhis & Morgan (2007) translate an absolute *minimum* of 50+8 observations for regression analyses suggested by a rule of thumb examined in S. B. Green (1991) into the recommendation to collect ~50 observations. Green actually concludes in his article that “In summary, no specific minimum number of subjects or minimum ratio of subjects-to-predictors was supported”. He does discuss how a general rule of thumb

of  $N = 50 + 8$  provided an accurate minimum number of observations for the ‘typical’ study in the social sciences because these have a ‘medium’ effect size, as Green claims by citing Cohen (1988). Cohen actually didn’t claim that the typical study in the social sciences has a ‘medium’ effect size, and instead said (1988, p. 13): “Many effects sought in personality, social, and clinical-psychological research are likely to be small effects as here defined”. We see how a string of mis-citations eventually leads to a misleading rule of thumb.

Rules of thumb seem to primarily emerge due to mis-citations and/or overly simplistic recommendations. Simonsohn, Nelson, and Simmons (2011) recommended that “Authors must collect at least 20 observations per cell”. A later recommendation by the same authors presented at a conference suggested to use  $n > 50$ , unless you study large effects (Simmons et al., 2013). Regrettably, this advice is now often mis-cited as a justification to collect no more than 50 observations per condition without considering the expected effect size. If authors justify a specific sample size (e.g.,  $n = 50$ ) based on a general recommendation in another paper, either they are mis-citing the paper, or the paper they are citing is flawed.

Another common heuristic is to collect the same number of observations as were collected in a previous study. This strategy is not recommended in scientific disciplines with widespread publication bias, and/or where novel and surprising findings from largely exploratory single studies are published. Using the same sample size as a previous study is only a valid approach if the sample size justification in the previous study also applies to the current study. Instead of stating that you intend to collect the same sample size as an earlier study, repeat the sample size justification, and update it in light of any new information (such as the effect size in the earlier study, see Table 8.6).

Peer reviewers and editors should carefully scrutinize rules of thumb sample size justifications, because they can make it seem like a study has high informational value for an inferential goal even when the study will yield uninformative results. Whenever one encounters a sample size justification based on a heuristic, ask yourself: ‘Why is this heuristic used?’ It is important to know what the logic behind a heuristic is to determine whether the heuristic is valid for a specific situation. In most cases, heuristics are based on weak logic, and not widely applicable. That said, it might be possible that fields develop valid heuristics for sample size justifications. For example, it is possible that a research area reaches widespread agreement that effects smaller than  $d = 0.3$  are too small to be of interest, and all studies in a field use sequential designs (see below) that have 90% power to detect a  $d = 0.3$ . Alternatively, it is possible that a field agrees that data should be collected with a desired level of accuracy, irrespective of the true effect size. In these cases, valid heuristics would exist based on generally agreed goals of data collection. For example, Simonsohn (2015) suggests to design replication studies that have 2.5 times as large sample sizes as the original study, as this provides 80% power for an equivalence test against an equivalence bound set to the effect the original study had 33% power to detect, assuming the true effect size is 0. As original authors typically do not specify which effect size would falsify their hypothesis, the heuristic underlying this ‘small telescopes’ approach is a good starting point for a replication study with the inferential goal to reject the presence of an effect as large as was described in an earlier publication. It is the

responsibility of researchers to gain the knowledge to distinguish valid heuristics from mindless heuristics, and to be able to evaluate whether a heuristic will yield an informative result given the inferential goal of the researchers in a specific study, or not.

## 8.9 No Justification

It might sound like a *contradictio in terminis*, but it is useful to distinguish a final category where researchers explicitly state they do not have a justification for their sample size. Perhaps the resources were available to collect more data, but they were not used. A researcher could have performed a power analysis, or planned for precision, but they did not. In those cases, instead of pretending there was a justification for the sample size, honesty requires you to state there is no sample size justification. This is not necessarily bad. It is still possible to discuss the smallest effect size of interest, the minimal statistically detectable effect, the width of the confidence interval around the effect size, and to plot a sensitivity power analysis, in relation to the sample size that was collected. If a researcher truly had no specific inferential goals when collecting the data, such an evaluation can perhaps be performed based on reasonable inferential goals peers would have when they learn about the existence of the collected data.

Do not try to spin a story where it looks like a study was highly informative when it was not. Instead, transparently evaluate how informative the study was given effect sizes that were of interest, and make sure that the conclusions follow from the data. The lack of a sample size justification might not be problematic, but it might mean that a study was not informative for most effect sizes of interest, which makes it especially difficult to interpret non-significant effects, or estimates with large uncertainty.

## 8.10 What is Your Inferential Goal?

The inferential goal of data collection is often in some way related to the size of an effect. Therefore, to design an informative study, researchers will want to think about which effect sizes are interesting. First, it is useful to consider three effect sizes when determining the sample size. The first is the smallest effect size a researcher is interested in, the second is the smallest effect size that can be statistically significant (only in studies where a significance test will be performed), and the third is the effect size that is expected. Beyond considering these three effect sizes, it can be useful to evaluate ranges of effect sizes. This can be done by computing the width of the expected confidence interval around an effect size of interest (for example, an effect size of zero), and examine which effects could be rejected. Similarly, it can be useful to plot a sensitivity curve and evaluate the range of effect sizes the design has decent power to detect, as well as to consider the range of effects for which the design has low power. Finally, there are situations where it is useful to consider a range of effects that is likely to be observed in a specific research area.

## 8.11 What is the Smallest Effect Size of Interest?

The strongest possible sample size justification is based on an explicit statement of the smallest effect size that is considered interesting. The smallest effect size of interest can be based on theoretical predictions or practical considerations. For a review of approaches that can be used to determine the smallest effect size of interest in randomized controlled trials, see J. Cook et al. (2014) and Keefe et al. (2013), for reviews of different methods to determine a smallest effect size of interest, see King (2011) and Copay et al. (2007), and for a discussion focused on psychological research, see Daniël Lakens, Scheel, et al. (2018).

It can be challenging to determine the smallest effect size of interest whenever theories are not very developed, or when the research question is far removed from practical applications, but it is still worth thinking about which effects would be too small to matter. A first step forward is to discuss which effect sizes are considered meaningful in a specific research line with your peers. Researchers will differ in the effect sizes they consider large enough to be worthwhile (Murphy et al., 2014). Just as not every scientist will find every research question interesting enough to study, not every scientist will consider the same effect sizes interesting enough to study, and different stakeholders will differ in which effect sizes are considered meaningful (Kelley & Preacher, 2012).

Even though it might be challenging, there are important benefits of being able to specify the smallest effect size of interest. The population effect size is always uncertain (indeed, estimating this is typically one of the goals of the study), and therefore whenever a study is powered for an expected effect size, there is considerable uncertainty about whether the statistical power is high enough to detect the true effect in the population. However, if the smallest effect size of interest can be specified and agreed upon after careful deliberation, it becomes possible to design a study that has sufficient power (given the inferential goal to detect or reject the smallest effect size of interest with a certain error rate). Put differently, although the smallest effect of interest may be subjective (one researcher might find effect sizes smaller than  $d = 0.3$  meaningless, while another researcher might still be interested in effects smaller than  $d = 0.1$ ), and there might be uncertainty about the parameters required to specify the smallest effect size of interest (e.g., when performing a cost-benefit analysis), once researchers determine the smallest effect size of interest, a study can be designed with a known Type II error rate to detect or reject this value. For this reason an a-priori power based on a smallest effect size of interest is generally preferred, whenever researchers are able to specify one (Aberson, 2019; Albers & Lakens, 2018; G. W. Brown, 1983; Cascio & Zedeck, 1983; Dienes, 2014; Lenth, 2001).

## 8.12 The Minimal Statistically Detectable Effect

The [minimal statistically detectable effect](#) is the smallest effect size that, if observed, would yield a statistically significant  $p$ -value (J. Cook et al., 2014). In Figure 8.4, the distribution of

Cohen's  $d$  is plotted for 15 participants per group when the true effect size is either  $d = 0$  or  $d = 0.5$ . This figure is similar to Figure 8.2, with the addition that the critical  $d$  is indicated. We see that with such a small number of observations in each group only observed effects larger than  $d = 0.75$  will be statistically significant. Whether such effect sizes are interesting, and can realistically be expected, should be carefully considered and justified.

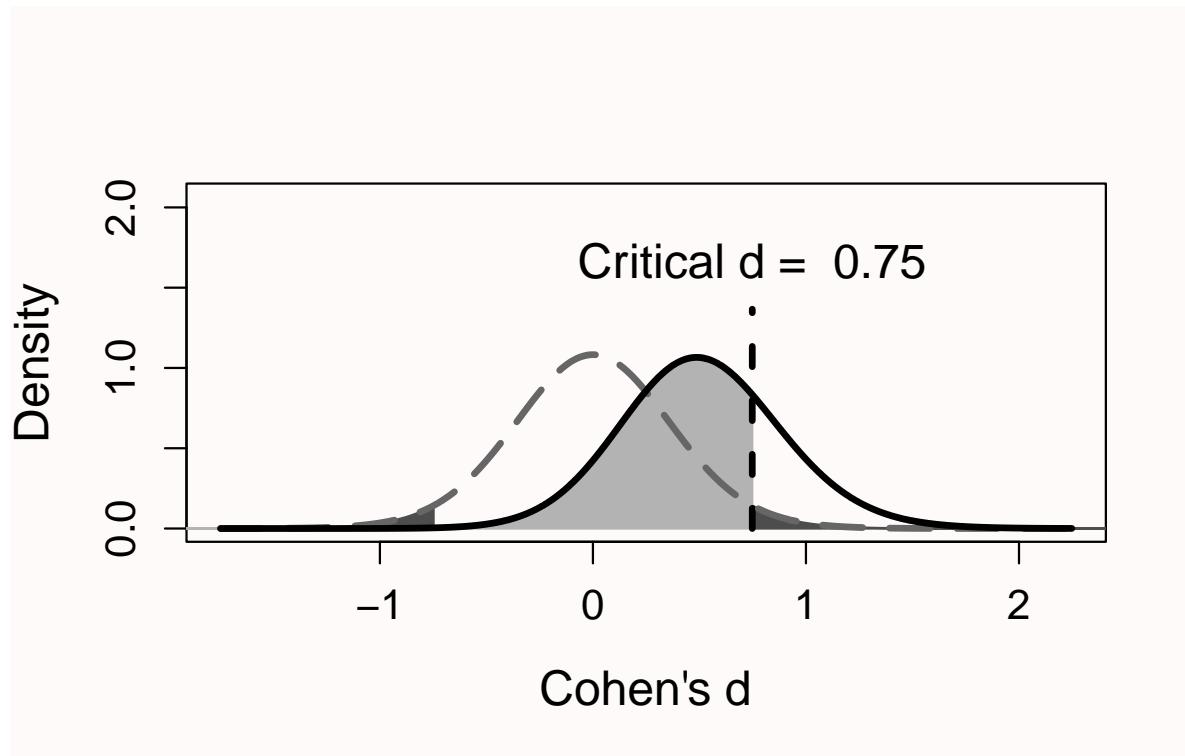


Figure 8.4: Critical effect size for an independent  $t$  test with  $n = 15$  per group and  $\alpha = 0.05$ .

Computing a minimal statistically detectable effect is useful for a study where no a-priori power analysis is performed, both for studies in the published literature that do not report a sample size justification (Daniël Lakens, Scheel, et al., 2018), as for researchers who rely on heuristics for their sample size justification.

It can be informative to ask yourself whether the critical effect size for a study design is within the range of effect sizes that can realistically be expected. If not, then whenever a significant effect is observed in a published study, either the effect size is surprisingly larger than expected, or more likely, it is an upwardly biased effect size estimate. In the latter case, given publication bias, published studies will lead to biased effect size estimates. If it is still possible to increase the sample size, for example by ignoring rules of thumb and instead performing an a-priori power analysis, then do so. If it is not possible to increase the sample size, for example due to resource constraints, then reflecting on the minimal statistically detectable effect should make it clear that an analysis of the data should not focus on  $p$  values, but on the effect size and

the confidence interval (see Table 8.3).

It is also useful to compute the minimal statistically detectable effect if an ‘optimistic’ power analysis is performed. For example, if you believe a best case scenario for the true effect size is  $d = 0.57$  and use this optimistic expectation in an a-priori power analysis, effects smaller than  $d = 0.4$  will not be statistically significant when you collect 50 observations in a two independent group design. If your worst case scenario for the alternative hypothesis is a true effect size of  $d = 0.35$  your design would not allow you to declare a significant effect if effect size estimates close to the worst case scenario are observed. Taking into account the minimal statistically detectable effect size should make you reflect on whether a hypothesis test will yield an informative answer, and whether your current approach to sample size justification (e.g., the use of rules of thumb, or letting resource constraints determine the sample size) leads to an informative study, or not.

## 8.13 What is the Expected Effect Size?

Although the true population effect size is always unknown, there are situations where researchers have a reasonable expectation of the effect size in a study, and want to use this expected effect size in an a-priori power analysis. Even if expectations for the observed effect size are largely a guess, it is always useful to explicitly consider which effect sizes are expected. A researcher can justify a sample size based on the effect size they expect, even if such a study would not be very informative with respect to the smallest effect size of interest. In such cases a study is informative for one inferential goal (testing whether the expected effect size is present or absent), but not highly informative for the second goal (testing whether the smallest effect size of interest is present or absent).

There are typically three sources for expectations about the population effect size: a meta-analysis, a previous study, or a theoretical model. It is tempting for researchers to be overly optimistic about the expected effect size in an a-priori power analysis, as higher effect size estimates yield lower sample sizes, but being too optimistic increases the probability of observing a false negative result. When reviewing a sample size justification based on an a-priori power analysis, it is important to critically evaluate the justification for the expected effect size used in power analyses.

## 8.14 Using an Estimate from a Meta-Analysis

In a perfect world effect size estimates from a meta-analysis would provide researchers with the most accurate information about which effect size they could expect. Due to widespread publication bias in science, effect size estimates from meta-analyses are regrettably not always accurate. They can be biased, sometimes substantially so. Furthermore, meta-analyses typically have considerable heterogeneity, which means that the meta-analytic effect size estimate

differs for subsets of studies that make up the meta-analysis. So, although it might seem useful to use a meta-analytic effect size estimate of the effect you are studying in your power analysis, you need to take great care before doing so.

If a researcher wants to enter a meta-analytic effect size estimate in an a-priori power analysis, they need to consider three things (see Table 8.5). First, the studies included in the meta-analysis should be similar enough to the study they are performing that it is reasonable to expect a similar effect size. In essence, this requires evaluating the generalizability of the effect size estimate to the new study. It is important to carefully consider differences between the meta-analyzed studies and the planned study, with respect to the manipulation, the measure, the population, and any other relevant variables.

Second, researchers should check whether the effect sizes reported in the meta-analysis are homogeneous. If there is substantial heterogeneity in the meta-analytic effect sizes, it means not all included studies can be expected to have the same true effect size estimate. A meta-analytic estimate should be used based on the subset of studies that most closely represent the planned study. Note that heterogeneity remains a possibility (even direct replication studies can show heterogeneity when unmeasured variables moderate the effect size in each sample (Kenny & Judd, 2019; Olsson-Collentine et al., 2020)), so the main goal of selecting similar studies is to use existing data to increase the probability that your expectation is accurate, without guaranteeing it will be.

Third, the meta-analytic effect size estimate should not be biased. Check if the bias detection tests that are reported in the meta-analysis are state-of-the-art, or perform multiple bias detection tests yourself (Carter et al., 2019), and consider bias corrected effect size estimates (even though these estimates might still be biased, and do not necessarily reflect the true population effect size).

Table 8.5: Overview of recommendations when justifying the use of a meta-analytic effect size estimate for a power analysis.

What to take into account	How to take it into account?
Are the studies in the meta-analysis similar?	Are the studies in the meta-analyses very similar in design, measures, and the population to the study you are planning? Evaluate the generalizability of the effect size estimate to your study.
Are the studies in the meta-analysis homogeneous?	Is there heterogeneity in the meta-analysis? If so, use the meta-analytic effect size estimate of the most relevant homogenous subsample.

What to take into account	How to take it into account?
Is the effect size estimate unbiased?	Did the original study report bias detection tests, and was there bias? If so, it might be wise to use a more conservative effect size estimate, based on bias correction techniques while acknowledging these corrected effect size estimates might not represent the true meta-analytic effect size estimate.

## 8.15 Using an Estimate from a Previous Study

If a meta-analysis is not available, researchers often rely on an effect size from a previous study in an a-priori power analysis. The first issue that requires careful attention is whether the two studies are sufficiently similar. Just as when using an effect size estimate from a meta-analysis, researchers should consider if there are differences between the studies in terms of the population, the design, the manipulations, the measures, or other factors that should lead one to expect a different effect size. For example, intra-individual reaction time variability increases with age, and therefore a study performed on older participants should expect a smaller standardized effect size than a study performed on younger participants. If an earlier study used a very strong manipulation, and you plan to use a more subtle manipulation, a smaller effect size should be expected. Finally, effect sizes do not generalize to studies with different designs. For example, the effect size for a comparison between two groups is most often not similar to the effect size for an interaction in a follow-up study where a second factor is added to the original design (Daniël Lakens & Caldwell, 2021).

Even if a study is sufficiently similar, statisticians have warned against using effect size estimates from small pilot studies in power analyses. Leon, Davis, and Kraemer (2011) write:

Contrary to tradition, a pilot study does not provide a meaningful effect size estimate for planning subsequent studies due to the imprecision inherent in data from small samples.

The two main reasons researchers should be careful when using effect sizes from studies in the published literature in power analyses is that effect size estimates from studies can differ from the true population effect size due to random variation, and that publication bias inflates effect sizes. Figure 8.5 shows the distribution for  $\eta_p^2$  for a study with three conditions with 25 participants in each condition when the null hypothesis is true (dotted grey curve) and when there is a ‘medium’ true effect of  $\eta_p^2 = 0.0588$  [solid black curve; Richardson (2011)]. As in Figure 8.4 the critical effect size is indicated, which shows observed effects smaller than  $\eta_p^2 = 0.08$  will not be significant with the given sample size. If the null hypothesis is true, effects larger than  $\eta_p^2 = 0.08$  will be a Type I error (the dark grey area), and when the alternative

hypothesis is true effects smaller than  $\eta_p^2 = 0.08$  will be a Type II error (light grey area). It is clear all significant effects are larger than the true effect size ( $\eta_p^2 = 0.0588$ ), so power analyses based on a significant finding (e.g., because only significant results are published in the literature) will be based on an overestimate of the true effect size, introducing bias.

But even if we had access to all effect sizes (e.g., from pilot studies you have performed yourself) due to random variation the observed effect size will sometimes be quite small. Figure 8.5 shows it is quite likely to observe an effect of  $\eta_p^2 = 0.01$  in a small pilot study, even when the true effect size is 0.0588. Entering an effect size estimate of  $\eta_p^2 = 0.01$  in an a-priori power analysis would suggest a total sample size of 957 observations to achieve 80% power in a follow-up study. If researchers only follow up on pilot studies when they observe an effect size in the pilot study that, when entered into a power analysis, yields a sample size that is feasible to collect for the follow-up study, these effect size estimates will be upwardly biased, and power in the follow-up study will be systematically lower than desired (Albers & Lakens, 2018).

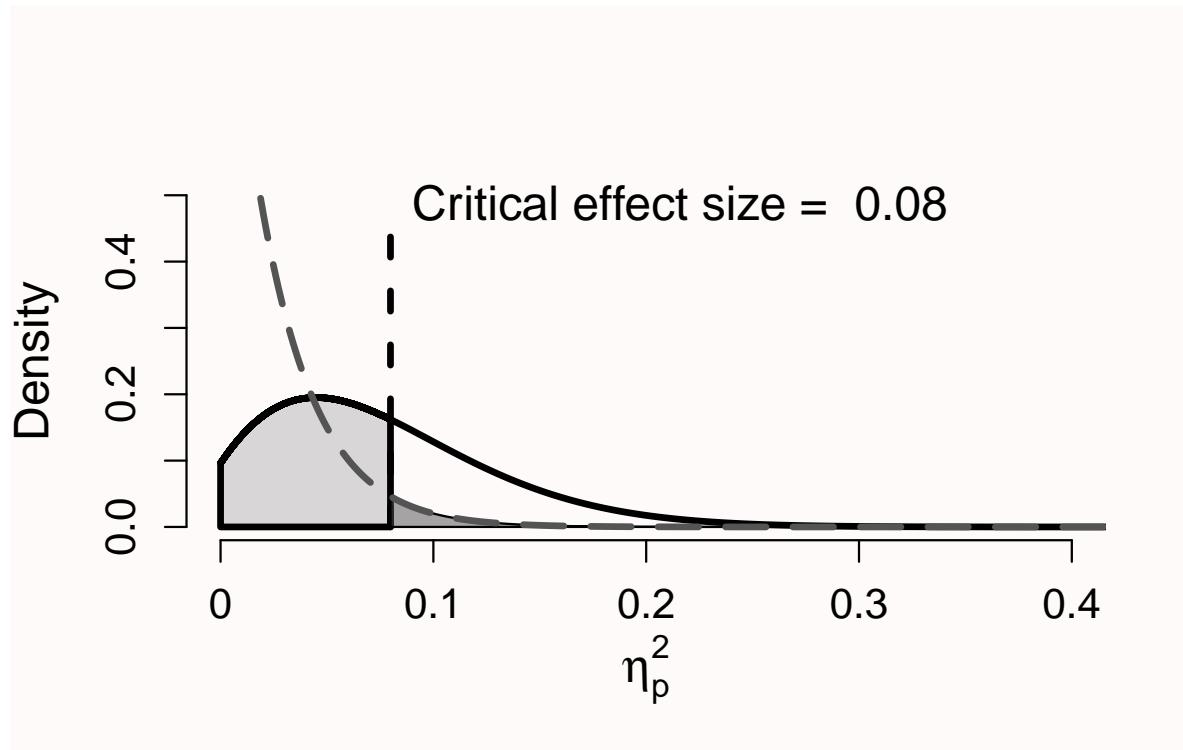


Figure 8.5: Distribution of partial eta squared under the null hypothesis (dotted grey curve) and a medium true effect of 0.0588 (solid black curve) for 3 groups with 25 observations.

In essence, the problem with using small studies to estimate the effect size that will be entered into an a-priori power analysis is that due to publication bias or follow-up bias the effect sizes researchers end up using for their power analysis do not come from a full  $F$  distribution, but

from what is known as a *truncated F* distribution (Taylor & Muller, 1996). For example, imagine if there is extreme publication bias in the situation illustrated in Figure 8.5. The only studies that would be accessible to researchers would come from the part of the distribution where  $\eta_p^2 > 0.08$ , and the test result would be statistically significant. It is possible to compute an effect size estimate that, based on certain assumptions, corrects for bias. For example, imagine we observe a result in the literature for a One-Way ANOVA with 3 conditions, reported as  $F(2, 42) = 0.017$ ,  $\eta_p^2 = 0.176$ . If we would take this effect size at face value and enter it as our effect size estimate in an a-priori power analysis, the suggested sample size to achieve 80% power would suggest we need to collect 17 observations in each condition.

However, if we assume bias is present, we can use the BUCSS R package (S. F. Anderson et al., 2017) to perform a power analysis that attempts to correct for bias. In the example above, a power analysis that takes bias into account (under a specific model of publication bias, based on a truncated *F* distribution where only significant results are published) suggests collecting 73 participants in each condition instead. It is possible that the bias corrected estimate of the non-centrality parameter used to compute power is zero, in which case it is not possible to correct for bias using this method. As an alternative to formally modeling a correction for publication bias whenever researchers assume an effect size estimate is biased, researchers can simply use a more conservative effect size estimate, for example by computing power based on the lower limit of a 60% two-sided confidence interval around the effect size estimate, which Perugini et al. (2014) refer to as *safeguard power*. Both these approaches lead to a more conservative power analysis, but not necessarily a more accurate power analysis. It is simply not possible to perform an accurate power analysis on the basis of an effect size estimate from a study that might be biased and/or had a small sample size (Teare et al., 2014). If it is not possible to specify a smallest effect size of interest, and there is great uncertainty about which effect size to expect, it might be more efficient to perform a study with a sequential design (discussed below).

To summarize, an effect size from a previous study in an a-priori power analysis can be used if three conditions are met (see Table 8.6). First, the previous study is sufficiently similar to the planned study. Second, there was a low risk of bias (e.g., the effect size estimate comes from a Registered Report, or from an analysis for which results would not have impacted the likelihood of publication). Third, the sample size is large enough to yield a relatively accurate effect size estimate, based on the width of a 95% CI around the observed effect size estimate. There is always uncertainty around the effect size estimate, and entering the upper and lower limit of the 95% CI around the effect size estimate might be informative about the consequences of the uncertainty in the effect size estimate for an a-priori power analysis.

Table 8.6: Overview of recommendations when justifying the use of an effect size estimate from a single study.

What to take into account	How to take it into account?
Is the study sufficiently similar?	Consider if there are differences between the studies in terms of the population, the design, the manipulations, the measures, or other factors that should lead one to expect a different effect size.
Is there a risk of bias?	Evaluate the possibility that if the effect size estimate had been smaller you would not have used it (or it would not have been published). Examine the difference when entering the reported, and a bias corrected, effect size estimate in a power analysis.
How large is the uncertainty?	Studies with a small number of observations have large uncertainty. Consider the possibility of using a more conservative effect size estimate to reduce the possibility of an underpowered study for the true effect size (such as a safeguard power analysis).

## 8.16 Using an Estimate from a Theoretical Model

When your theoretical model is sufficiently specific such that you can build a computational model, and you have knowledge about key parameters in your model that are relevant for the data you plan to collect, it is possible to estimate an effect size based on the effect size estimate derived from a computational model. For example, if one had strong ideas about the weights for each feature stimuli share and differ on, it could be possible to compute predicted similarity judgments for pairs of stimuli based on Tversky's contrast model (Tversky, 1977), and estimate the predicted effect size for differences between experimental conditions. Although computational models that make point predictions are relatively rare, whenever they are available, they provide a strong justification of the effect size a researcher expects.

## 8.17 Compute the Width of the Confidence Interval around the Effect Size

If a researcher can estimate the standard deviation of the observations that will be collected, it is possible to compute an a-priori estimate of the width of the 95% confidence interval around

an effect size (Kelley, 2007). Confidence intervals represent a range around an estimate that is wide enough so that in the long run the true population parameter will fall inside the confidence intervals  $100 - \alpha$  percent of the time. In any single study the true population effect either falls in the confidence interval, or it doesn't, but in the long run one can *act* as if the confidence interval includes the true population effect size (while keeping the error rate in mind). Cumming (2013) calls the difference between the observed effect size and the upper bound of the 95% confidence interval (or the lower bound of the 95% confidence interval) the margin of error.

If we compute the 95% CI for an effect size of  $d = 0$  based on the  $t$  statistic and sample size (Smithson, 2003), we see that with 15 observations in each condition of an independent  $t$  test the 95% CI ranges from  $d = -0.716$  to  $d = 0.716$ . Confidence intervals around effect sizes can be computed using the MOTE Shiny app: <https://www.aggieerin.com/shiny-server/>. The margin of error is half the width of the 95% CI, 0.716. A Bayesian estimator who uses an uninformative prior would compute a credible interval with the same (or a very similar) upper and lower bound (Albers et al., 2018; Kruschke, 2011), and might conclude that after collecting the data they would be left with a range of plausible values for the population effect that is too large to be informative. Regardless of the statistical philosophy you plan to rely on when analyzing the data, the evaluation of what we can conclude based on the width of our interval tells us that with 15 observation per group we will not learn a lot.

One useful way of interpreting the width of the confidence interval is based on the effects you would be able to reject if the true effect size is 0. In other words, if there is no effect, which effects would you have been able to reject given the collected data, and which effect sizes would not be rejected, if there was no effect? Effect sizes in the range of  $d = 0.7$  are findings such as “People become aggressive when they are provoked”, “People prefer their own group to other groups”, and “Romantic partners resemble one another in physical attractiveness” (Richard et al., 2003). The width of the confidence interval tells you that you can only reject the presence of effects that are so large, if they existed, you would probably already have noticed them. If it is true that most effects that you study are realistically much smaller than  $d = 0.7$ , there is a good possibility that we do not learn anything we didn't already know by performing a study with  $n = 15$ . Even without data, in most research lines we would not consider certain large effects plausible (although the effect sizes that are plausible differ between fields, as discussed below). On the other hand, in large samples where researchers can for example reject the presence of effects larger than  $d = 0.2$ , if the null hypothesis was true, this analysis of the width of the confidence interval would suggest that peers in many research lines would likely consider the study to be informative.

We see that the margin of error is almost, but not exactly, the same as the minimal statistically detectable effect ( $d = 0.748$ ). The small variation is due to the fact that the 95% confidence interval is calculated based on the  $t$  distribution. If the true effect size is not zero, the confidence interval is calculated based on the non-central  $t$  distribution, and the 95% CI is asymmetric. Figure 8.6 visualizes three  $t$  distributions, one symmetric at 0, and two asymmetric distributions with a noncentrality parameter (the normalized difference between the means) of 2 and 3.

The asymmetry is most clearly visible in very small samples (the distributions in the plot have 5 degrees of freedom) but remains noticeable in larger samples when calculating confidence intervals and statistical power. For example, for a true effect size of  $d = 0.5$  observed with 15 observations per group would yield  $d_s = 0.50$ , 95% CI [-0.23, 1.22]. If we compute the 95% CI around the critical effect size, we would get  $d_s = 0.75$ , 95% CI [0.00, 1.48]. We see the 95% CI ranges from exactly 0 to 1.484, in line with the relation between a confidence interval and a  $p$  value, where the 95% CI excludes zero if the test is statistically significant. As noted before, the different approaches recommended here to evaluate how informative a study is are often based on the same information.

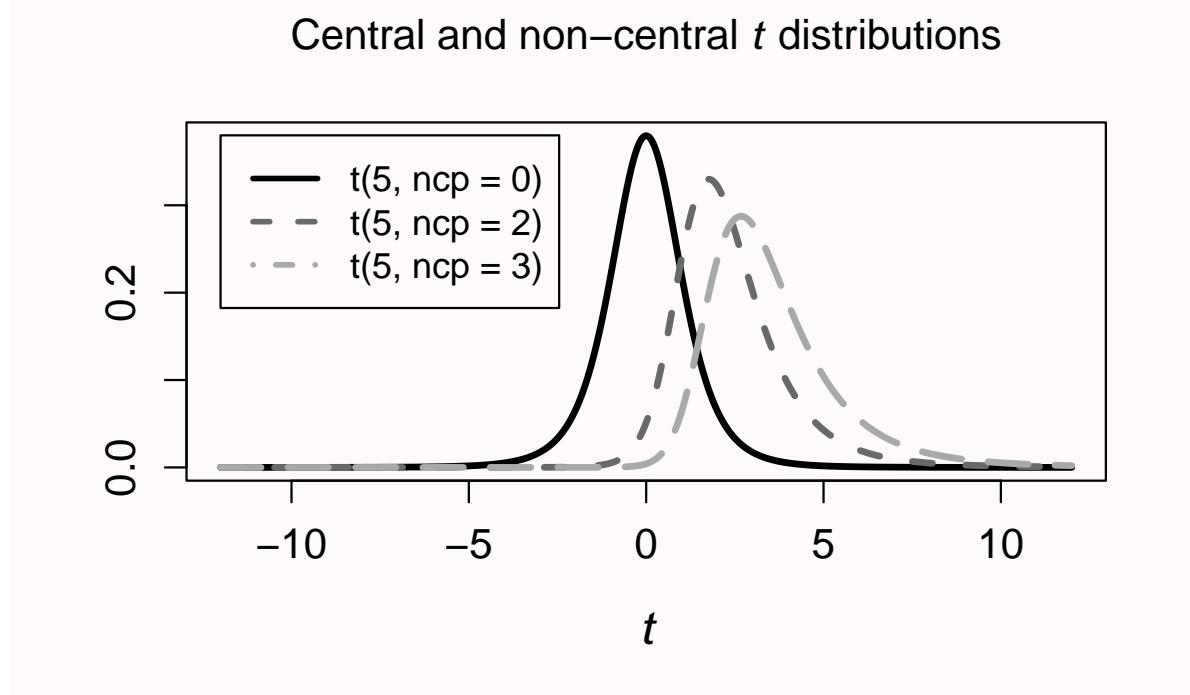


Figure 8.6: Central (black dashed line) and 2 non-central (dark grey and light grey dashed lines)  $t$  distributions; ncp = noncentrality parameter.

## 8.18 Plot a Sensitivity Power Analysis

A sensitivity power analysis fixes the sample size, desired power, and alpha level, and answers the question which effect size a study could detect with a desired power. A sensitivity power analysis is therefore performed when the sample size is already known. Sometimes data has already been collected to answer a different research question, or the data is retrieved from an

existing database, and you want to perform a sensitivity power analysis for a new statistical analysis. Other times, you might not have carefully considered the sample size when you initially collected the data, and want to reflect on the statistical power of the study for (ranges of) effect sizes of interest when analyzing the results. Finally, it is possible that the sample size will be collected in the future, but you know that due to resource constraints the maximum sample size you can collect is limited, and you want to reflect on whether the study has sufficient power for effects that you consider plausible and interesting (such as the smallest effect size of interest, or the effect size that is expected).

Assume a researcher plans to perform a study where 30 observations will be collected in total, 15 in each between participant condition. Figure 8.7 shows how to perform a sensitivity power analysis in G\*Power for a study where we have decided to use an alpha level of 5%, and desire 90% power. The sensitivity power analysis reveals the designed study has 90% power to detect effects of at least  $d = 1.23$ . Perhaps a researcher believes that a desired power of 90% is quite high, and is of the opinion that it would still be interesting to perform a study if the statistical power was lower. It can then be useful to plot a sensitivity curve across a range of smaller effect sizes.

The two dimensions of interest in a sensitivity power analysis are the effect sizes, and the power to observe a significant effect assuming a specific effect size. Fixing the sample size, these two dimensions can be plotted against each other to create a sensitivity curve. For example, a sensitivity curve can be plotted in G\*Power by clicking the ‘X-Y plot for a range of values’ button, as illustrated in Figure 8.8. Researchers can examine which power they would have for an a-priori plausible range of effect sizes, or they can examine which effect sizes would provide reasonable levels of power. In simulation-based approaches to power analysis, sensitivity curves can be created by performing the power analysis for a range of possible effect sizes. Even if 50% power is deemed acceptable (in which case deciding to act as if the null hypothesis is true after a non-significant result is a relatively noisy decision procedure), Figure 8.8 shows a study design where power is extremely low for a large range of effect sizes that are reasonable to expect in most fields. Thus, a sensitivity power analysis provides an additional approach to evaluate how informative the planned study is, and can inform researchers that a specific design is unlikely to yield a significant effect for a range of effects that one might realistically expect.

If the number of observations per group had been larger, the evaluation might have been more positive. We might not have had any specific effect size in mind, but if we had collected 150 observations per group, a sensitivity analysis could have shown that power was sufficient for a range of effects we believe is most interesting to examine, and we would still have approximately 50% power for quite small effects. For a sensitivity analysis to be meaningful, the sensitivity curve should be compared against a smallest effect size of interest, or a range of effect sizes that are expected. A sensitivity power analysis has no clear cut-offs to examine (Bacchetti, 2010). Instead, the idea is to make a holistic trade-off between different effect sizes one might observe or care about, and their associated statistical power.

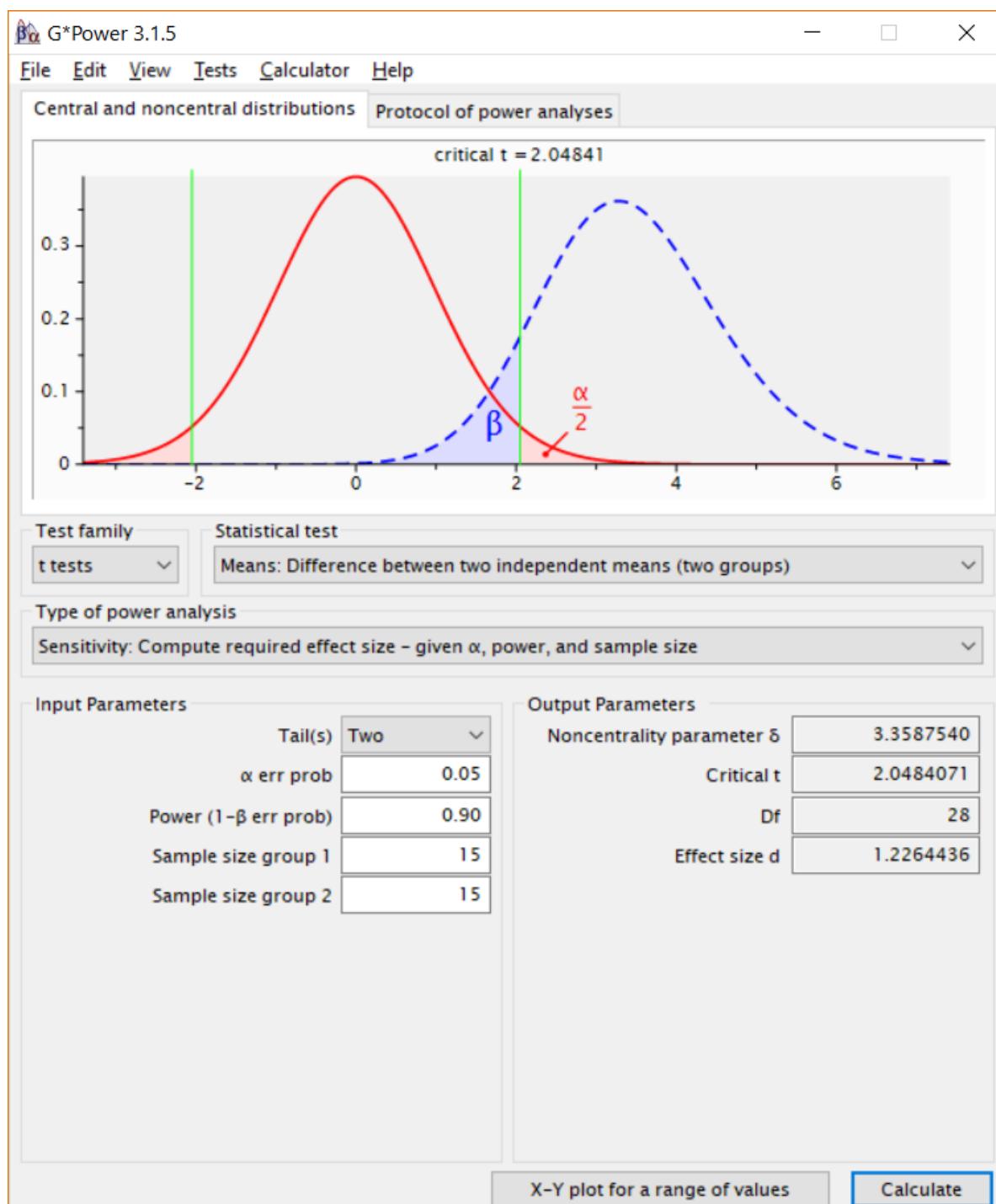


Figure 8.7: Sensitivity power analysis in G\*Power software.

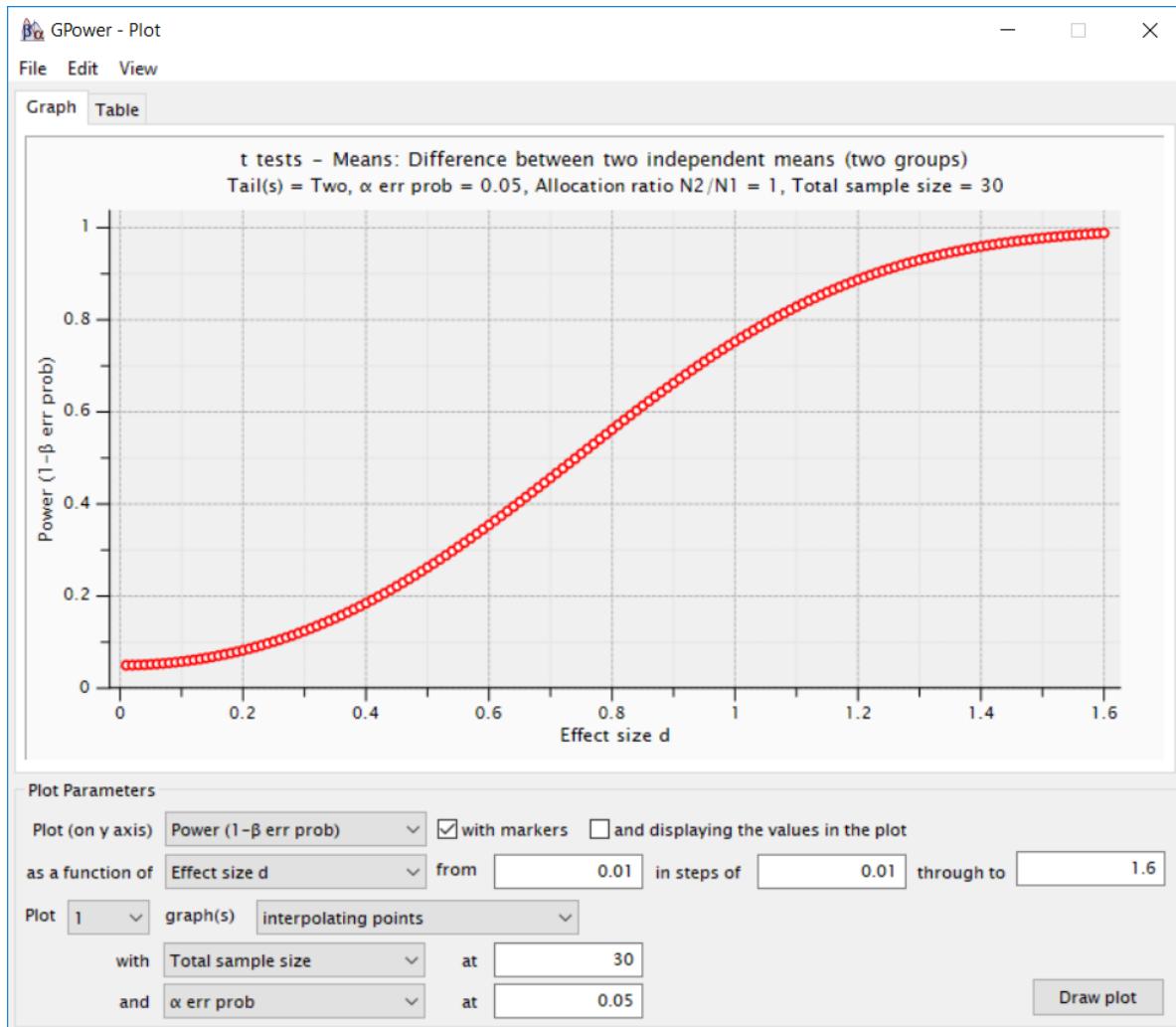


Figure 8.8: Plot of the effect size against the desired power when  $n = 15$  per group and alpha = 0.05.

## 8.19 The Distribution of Effect Sizes in a Research Area

In my personal experience the most commonly entered effect size estimate in an a-priori power analysis for an independent  $t$  test is Cohen's benchmark for a 'medium' effect size, because of what is known as the *default effect*. When you open G\*Power, a 'medium' effect is the default option for an a-priori power analysis. Cohen's benchmarks for small, medium, and large effects should not be used in an a-priori power analysis (J. Cook et al., 2014; Correll et al., 2020), and Cohen regretted having proposed these benchmarks (Funder & Ozer, 2019). The large variety in research topics means that any 'default' or 'heuristic' that is used to compute statistical power is not just unlikely to correspond to your actual situation, but it is also likely to lead to a sample size that is substantially misaligned with the question you are trying to answer with the collected data.

Some researchers have wondered what a better default would be, if researchers have no other basis to decide upon an effect size for an a-priori power analysis. Brysbaert (2019) recommends  $d = 0.4$  as a default in psychology, which is the average observed in replication projects and several meta-analyses. It is impossible to know if this average effect size is realistic, but it is clear there is huge heterogeneity across fields and research questions. Any average effect size will often deviate substantially from the effect size that should be expected in a planned study. Some researchers have suggested to change Cohen's benchmarks based on the distribution of effect sizes in a specific field (Bosco et al., 2015; Funder & Ozer, 2019; Hill et al., 2008; Kraft, 2020; Lovakov & Agadullina, 2021). As always, when effect size estimates are based on the published literature, one needs to evaluate the possibility that the effect size estimates are inflated due to publication bias. Due to the large variation in effect sizes within a specific research area, there is little use in choosing a large, medium, or small effect size benchmark based on the empirical distribution of effect sizes in a field to perform a power analysis.

Having some knowledge about the distribution of effect sizes in the literature can be useful when interpreting the confidence interval around an effect size. If in a specific research area almost no effects are larger than the value you could reject in an equivalence test (e.g., if the observed effect size is 0, the design would only reject effects larger than for example  $d = 0.7$ ), then it is a-priori unlikely that collecting the data would tell you something you didn't already know.

It is more difficult to defend the use of a specific effect size derived from an empirical distribution of effect sizes as a justification for the effect size used in an a-priori power analysis. One might argue that the use of an effect size benchmark based on the distribution of effects in the literature will outperform a wild guess, but this is not a strong enough argument to form the basis of a sample size justification. There is a point where researchers need to admit they are not ready to perform an a-priori power analysis due to a lack of clear expectations (Scheel, Tiokhin, et al., 2021). Alternative sample size justifications, such as a justification of the sample size based on resource constraints, perhaps in combination with a sequential study design, might be more in line with the actual inferential goals of a study.

## **8.20 Additional Considerations When Designing an Informative Study**

So far, the focus has been on justifying the sample size for quantitative studies. There are a number of related topics that can be useful to design an informative study. First, in addition to a-priori or prospective power analysis and sensitivity power analysis, it is important to discuss compromise power analysis (which is useful) and post-hoc or retrospective power analysis (which is not useful, e.g., Zumbo & Hubley (1998), Lenth (2007)). When sample sizes are justified based on an a-priori power analysis it can be very efficient to collect data in sequential designs where data collection is continued or terminated based on interim analyses of the data. Furthermore, it is worthwhile to consider ways to increase the power of a test without increasing the sample size. An additional point of attention is to have a good understanding of your dependent variable, especially its standard deviation. Finally, sample size justification is just as important in qualitative studies, and although there has been much less work on sample size justification in this domain, some proposals exist that researchers can use to design an informative study. Each of these topics is discussed in turn.

## **8.21 Compromise Power Analysis**

In a compromise power analysis the sample size and the effect are fixed, and the error rates of the test are calculated, based on a desired ratio between the Type I and Type II error rate. A compromise power analysis is useful both when a very large number of observations will be collected, as when only a small number of observations can be collected.

In the first situation a researcher might be fortunate enough to be able to collect so many observations that the statistical power for a test is very high for all effect sizes that are deemed interesting. For example, imagine a researcher has access to 2000 employees who are all required to answer questions during a yearly evaluation in a company where they are testing an intervention that should reduce subjectively reported stress levels. You are quite confident that an effect smaller than  $d = 0.2$  is not large enough to be subjectively noticeable for individuals (Jaeschke et al., 1989). With an alpha level of 0.05 the researcher would have a statistical power of 0.994, or a Type II error rate of 0.006. This means that for the smallest effect size of interest of  $d = 0.2$  the researcher is 8.3 times more likely to make a Type I error than a Type II error.

Although the original idea of designing studies that control Type I and Type II error rates was that researchers would need to justify their error rates (Neyman & Pearson, 1933), a common heuristic is to set the Type I error rate to 0.05 and the Type II error rate to 0.20, meaning that a Type I error is 4 times as unlikely as a Type II error. This default use of 80% power (or a Type II error rate/ $\beta$  of 0.20) is based on a personal preference of Cohen (1988), who writes:

It is proposed here as a convention that, when the investigator has no other basis for setting the desired power value, the value .80 be used. This means that  $\beta$  is set at .20. This arbitrary but reasonable value is offered for several reasons (Cohen, 1965, pp. 98-99). The chief among them takes into consideration the implicit convention for  $\alpha$  of .05. The  $\beta$  of .20 is chosen with the idea that the general relative seriousness of these two kinds of errors is of the order of .20/.05, i.e., that Type I errors are of the order of four times as serious as Type II errors. This .80 desired power convention is offered with the hope that it will be ignored whenever an investigator can find a basis in his substantive concerns in his specific research investigation to choose a value ad hoc.

We see that conventions are built on conventions: the norm to aim for 80% power is built on the norm to set the alpha level at 5%. What we should take away from Cohen is not that we should aim for 80% power, but that we should justify our error rates based on the relative seriousness of each error. This is where compromise power analysis comes in. If you share Cohen's belief that a Type I error is 4 times as serious as a Type II error, and building on our earlier study on 2000 employees, it makes sense to adjust the Type I error rate when the Type II error rate is low for all effect sizes of interest (Cascio & Zedeck, 1983). Indeed, Erdfelder et al. (1996) created the G\*Power software in part to give researchers a tool to perform compromise power analysis.

Figure 8.9 illustrates how a compromise power analysis is performed in G\*Power when a Type I error is deemed to be equally costly as a Type II error (that is, when the  $\beta/\alpha$  ratio = 1), which for a study with 1000 observations per condition would lead to a Type I error and a Type II error of 0.0179. As Faul, Erdfelder, Lang, and Buchner (2007) write:

Of course, compromise power analyses can easily result in unconventional significance levels greater than  $\alpha = .05$  (in the case of small samples or effect sizes) or less than  $\alpha = .001$  (in the case of large samples or effect sizes). However, we believe that the benefit of balanced Type I and Type II error risks often offsets the costs of violating significance level conventions.

This brings us to the second situation where a compromise power analysis can be useful, which is when we know the statistical power in our study is low. Although it is highly undesirable to make decisions when error rates are high, if one finds oneself in a situation where a decision must be made based on little information, Winer (1962) writes:

The frequent use of the .05 and .01 levels of significance is a matter of convention having little scientific or logical basis. When the power of tests is likely to be low under these levels of significance, and when Type I and Type II errors are of approximately equal importance, the .30 and .20 levels of significance may be more appropriate than the .05 and .01 levels.

For example, if we plan to perform a two-sided  $t$  test, can feasibly collect at most 50 observations in each independent group, and expect a population effect size of 0.5, we would have 70%

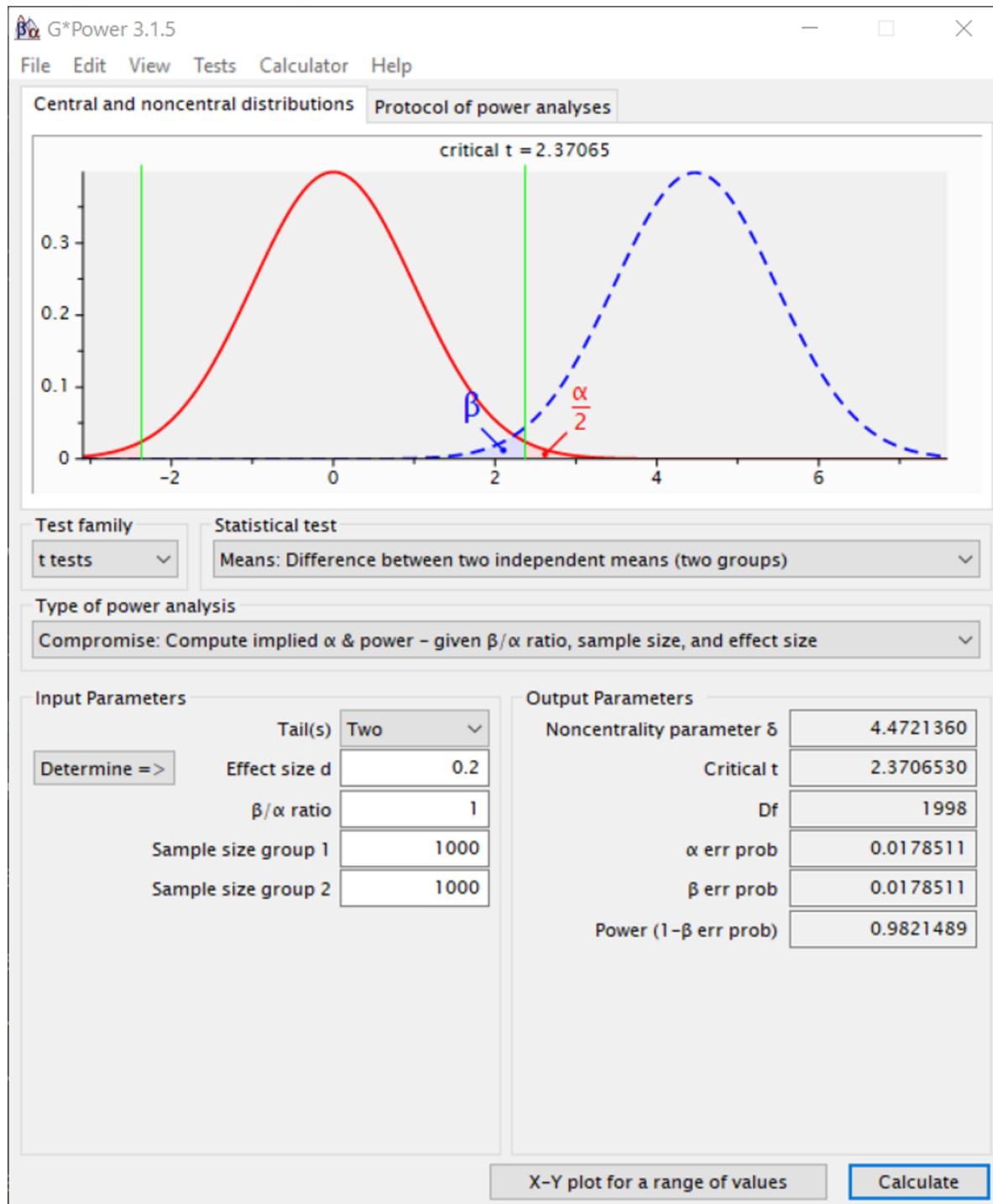


Figure 8.9: Compromise power analysis in G\*Power.

power if we set our alpha level to 0.05. Alternatively, using compromise power analysis, we can choose to weigh both types of error equally ( $\beta/\alpha$  ratio = 1) by setting both the alpha level and Type II error rate to 0.149. Doing so, we would have 85.10% power to detect the expected population effect size of  $d = 0.5$  instead. The choice of  $\alpha$  and  $\beta$  in a compromise power analysis can be extended to take prior probabilities of the null and alternative hypothesis into account (Maier & Lakens, 2022; Miller & Ulrich, 2019; Murphy et al., 2014).

A compromise power analysis requires a researcher to specify the sample size. This sample size itself requires a justification, so a compromise power analysis will typically be performed together with a resource constraint justification for a sample size. It is especially important to perform a compromise power analysis if your resource constraint justification is strongly based on the need to make a decision, in which case a researcher should think carefully about the Type I and Type II error rates stakeholders are willing to accept. However, a compromise power analysis also makes sense if the sample size is very large, but a researcher did not have the freedom to set the sample size. This might happen if, for example, data collection is part of a larger international study and the sample size is based on other research questions. In designs where the Type II error rate is very small (and power is very high) some statisticians have also recommended to lower the alpha level to prevent Lindley's paradox, a situation where a significant effect ( $p < \alpha$ ) is evidence for the null hypothesis (Good, 1992; Jeffreys, 1939). Lowering the alpha level as a function of the statistical power of the test can prevent this paradox, providing another argument for a compromise power analysis when sample sizes are large (Maier & Lakens, 2022). Finally, a compromise power analysis needs a justification for the effect size, either based on a smallest effect size of interest or an effect size that is expected. Table 8.7 lists three aspects that should be discussed alongside a reported compromise power analysis.

Table 8.7: Overview of recommendations when justifying error rates based on a compromise power analysis.

What to take into account	How to take it into account?
What is the justification for the sample size?	Specify why a specific sample size is collected (e.g., based on resource constraints or other factors that determined the sample size).
What is the justification for the effect size?	Is the effect size based on a smallest effect size of interest or an expected effect size?
What is the desired ratio of Type I vs Type II error rates?	Weigh the relative costs of a Type I and a Type II error by carefully evaluating the consequences of each type of error.

## 8.22 What to do if Your Editor Asks for Post-hoc Power?

Post-hoc, retrospective, or observed power is used to describe the statistical power of the test that is computed assuming the effect size that has been estimated from the collected data is the true effect size (Lenth, 2007; Zumbo & Hubley, 1998). Post-hoc power is therefore not performed before looking at the data, based on effect sizes that are deemed interesting, as in an a-priori power analysis, and it is unlike a sensitivity power analysis where a range of interesting effect sizes is evaluated. Because a post-hoc or retrospective power analysis is based on the effect size observed in the data that has been collected, it does not add any information beyond the reported  $p$  value, but it presents the same information in a different way. Despite this fact, editors and reviewers often ask authors to perform post-hoc power analysis to interpret non-significant results. This is not a sensible request, and whenever it is made, you should not comply with it. Instead, you should perform a sensitivity power analysis, and discuss the power for the smallest effect size of interest and a realistic range of expected effect sizes.

Post-hoc power is directly related to the  $p$  value of the statistical test (Hoenig & Heisey, 2001). For a  $z$  test where the  $p$  value is exactly 0.05, post-hoc power is always 50%. The reason for this relationship is that when a  $p$  value is observed that equals the alpha level of the test (e.g., 0.05), the observed  $z$  score of the test is exactly equal to the critical value of the test (e.g.,  $z = 1.96$  in a two-sided test with a 5% alpha level). Whenever the alternative hypothesis is centered on the critical value half the values we expect to observe if this alternative hypothesis is true fall below the critical value, and half fall above the critical value. Therefore, a test where we observed a  $p$  value identical to the alpha level will have exactly 50% power in a post-hoc power analysis, as the analysis assumes the observed effect size is true.

For other statistical tests, where the alternative distribution is not symmetric (such as for the  $t$  test, where the alternative hypothesis follows a non-central  $t$  distribution, see Figure 8.6), a  $p = 0.05$  does not directly translate to an observed power of 50%, but by plotting post-hoc power against the observed  $p$  value we see that the two statistics are always directly related. As Figure 8.10 shows, if the  $p$  value is non-significant (i.e., larger than 0.05) the observed power will be less than approximately 50% in a  $t$  test. Lenth (2007) explains how observed power is also completely determined by the observed  $p$  value for  $F$  tests, although the statement that a non-significant  $p$  value implies a power less than 50% no longer holds.

When editors or reviewers ask researchers to report post-hoc power analyses they would like to be able to distinguish between true negatives (concluding there is no effect, when there is no effect) and false negatives (a Type II error, concluding there is no effect, when there actually is an effect). Since reporting post-hoc power is just a different way of reporting the  $p$  value, reporting the post-hoc power will not provide an answer to the question editors are asking (Hoenig & Heisey, 2001; Lenth, 2007; Schulz & Grimes, 2005; Yuan & Maxwell, 2005). To be able to draw conclusions about the absence of a meaningful effect, one should perform an [equivalence test](#), and design a study with high power to reject the smallest effect size of

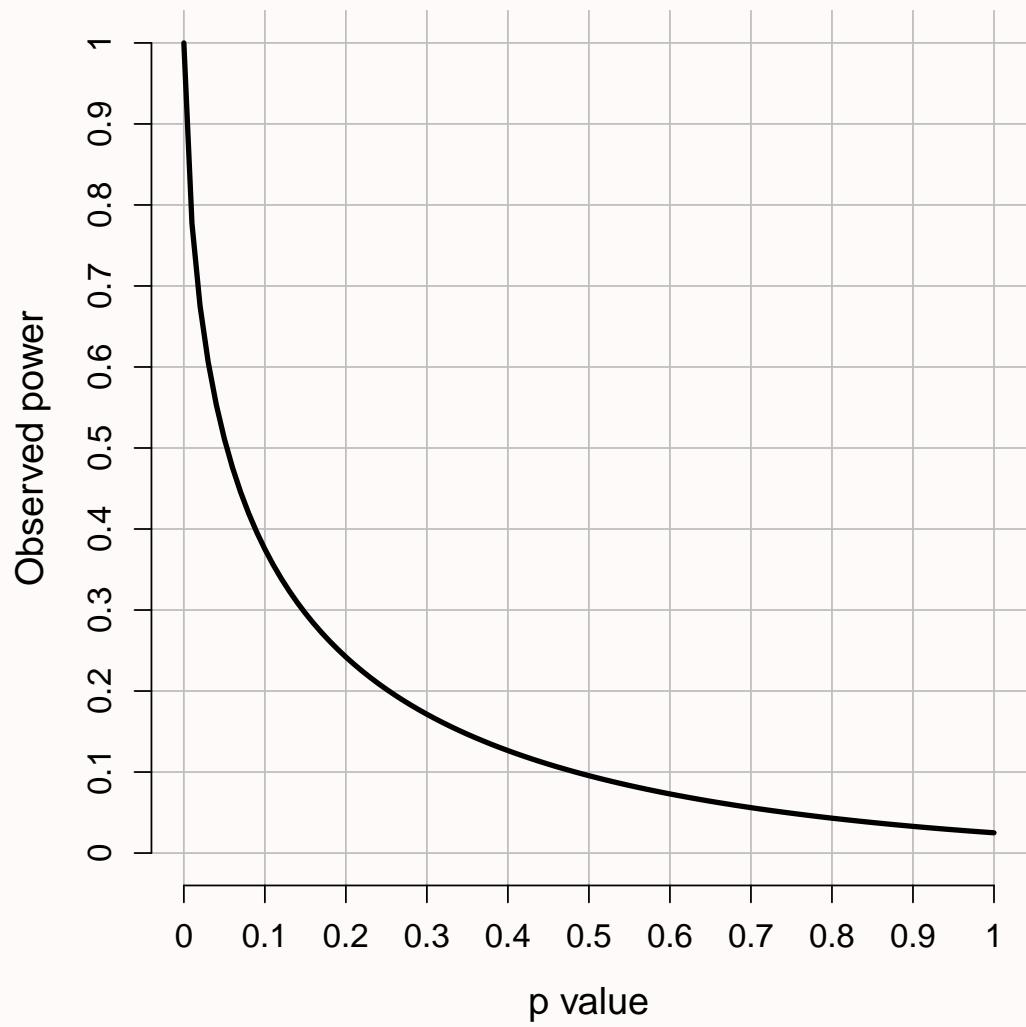


Figure 8.10: Relationship between  $p$  values and power for an independent  $t$  test with  $\alpha = 0.05$  and  $n = 10$ .

interest. Alternatively, if no smallest effect size of interest was specified when designing the study, researchers can report a sensitivity power analysis.

## 8.23 Sequential Analyses

Whenever the sample size is justified based on an a-priori power analysis it can be very efficient to collect data in a sequential design. Sequential designs control error rates across multiple looks at the data (e.g., after 50, 100, and 150 observations have been collected) and can reduce the average expected sample size that is collected compared to a fixed design where data is only analyzed after the maximum sample size is collected (Proschan et al., 2006; Wassmer & Brannath, 2016). Sequential designs have a long history (Dodge & Romig, 1929), and exist in many variations, such as the Sequential Probability Ratio Test (Wald, 1945), combining independent statistical tests (Westberg, 1985), group sequential designs (Jennison & Turnbull, 2000), sequential Bayes factors (Schönbrodt et al., 2017), and safe testing (Grünwald et al., 2019). Of these approaches, the Sequential Probability Ratio Test is most efficient if data can be analyzed after every observation (Schnuerch & Erdfelder, 2020). Group sequential designs, where data is collected in batches, provide more flexibility in data collection, error control, and corrections for effect size estimates (Wassmer & Brannath, 2016). Safe tests provide optimal flexibility if there are dependencies between observations (ter Schure & Grünwald, 2019).

Sequential designs are especially useful when there is considerable uncertainty about the effect size, or when it is plausible that the true effect size is larger than the smallest effect size of interest the study is designed to detect (Daniël Lakens, 2014). In such situations data collection has the possibility to terminate early if the effect size is larger than the smallest effect size of interest, but data collection can continue to the maximum sample size if needed. Sequential designs can prevent waste when testing hypotheses, both by stopping early when the null hypothesis can be rejected, as by stopping early if the presence of a smallest effect size of interest can be rejected (i.e., stopping for futility). Group sequential designs are currently the most widely used approach to sequential analyses, and can be planned and analyzed using `rpact` or `gsDesign`. Shiny apps are available for both `rpact` and `gsDesign`.

## 8.24 Increasing Power Without Increasing the Sample Size

The most straightforward approach to increase the informational value of studies is to increase the sample size. Because resources are often limited, it is also worthwhile to explore different approaches to increasing the power of a test without increasing the sample size. The first option is to use directional (one-sided) tests where relevant, instead of two-sided tests. Researchers often make directional predictions, such as ‘we predict X is larger than Y’. The statistical test that logically follows from this prediction is a directional (or one-sided)  $t$  test. A directional

test moves the Type I error rate to one side of the tail of the distribution, which lowers the critical value, and therefore requires less observations to achieve the same statistical power.

Although there is some discussion about when directional tests are appropriate, they are perfectly defensible from a Neyman-Pearson perspective on hypothesis testing (Cho & Abe, 2013), which makes a (preregistered) directional test a straightforward approach to both increase the power of a test, as the riskiness of the prediction. However, there might be situations where you do not want to ask a directional question. Sometimes, especially in research with applied consequences, it might be important to examine if a null effect can be rejected, even if the effect is in the opposite direction as predicted. For example, when you are evaluating a recently introduced educational intervention, and you predict the intervention will increase the performance of students, you might want to explore the possibility that students perform worse, to be able to recommend abandoning the new intervention. In such cases it is also possible to distribute the error rate in a ‘lop-sided’ manner, for example assigning a stricter error rate to effects in the negative than in the positive direction (Rice & Gaines, 1994).

Another approach to increase the power without increasing the sample size, is to increase the alpha level of the test, as explained in the section on compromise power analysis. Obviously, this comes at an increased probability of making a Type I error. The risk of making either type of error should be carefully weighed, which typically requires taking into account the prior probability that the null hypothesis is true (Cascio & Zedeck, 1983; Miller & Ulrich, 2019; Mudge et al., 2012; Murphy et al., 2014). If you *have* to make a decision, or want to make a claim, and the data you can feasibly collect is limited, increasing the alpha level is justified, either based on a compromise power analysis, or based on a cost-benefit analysis (Baguley, 2004; Field et al., 2004).

Another widely recommended approach to increase the power of a study is use a within participant design where possible. In almost all cases where a researcher is interested in detecting a difference between groups, a within participant design will require collecting less participants than a between participant design. The reason for the decrease in the sample size is explained by the equation below from Maxwell et al. (2017). The number of participants needed in a two group within-participants design (NW) relative to the number of participants needed in a two group between-participants design (NB), assuming normal distributions, is:

$$NW = \frac{NB(1 - \rho)}{2}$$

The required number of participants is divided by two because in a within-participants design with two conditions every participant provides two data points. The extent to which this reduces the sample size compared to a between-participants design also depends on the correlation between the dependent variables (e.g., the correlation between the measure collected in a control task and an experimental task), as indicated by the  $(1 - \rho)$  part of the equation. If the correlation is 0, a within-participants design simply needs half as many participants as a between participant design (e.g., 64 instead 128 participants). The higher the correlation,

the larger the relative benefit of within-participants designs, and whenever the correlation is negative (up to -1) the relative benefit disappears.

In Figure 8.11 we see two normally distributed scores with a mean difference of 6, where the standard deviation of each mean is 15, and the correlation between the measurements is 0. The standard deviation of the difference score is  $\sqrt{2}$  times as large as the standard deviation in each measurement, and indeed,  $15 \times \sqrt{2} = 21.21$ , which is rounded to 21. This situation where the correlation between measurements is zero equals the situation in an independent *t*-test, where the correlation between measurements is not taken into account.

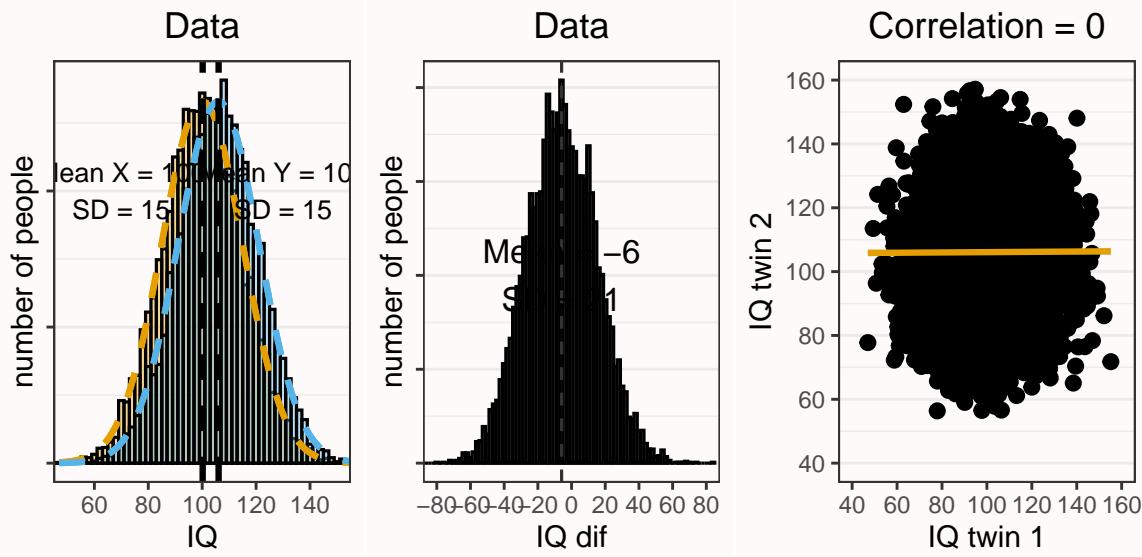


Figure 8.11: Distributions of two dependent groups with means 100 and 106 and a standard deviation of 15, distribution of the differences, and correlation of 0.

In Figure 8.12 we can see what happens when the two variables are correlated, for example with  $r = 0.7$ . Nothing has changed when we plot the means. The correlation between measurements is now strongly positive, and the important difference is in the standard deviation of the difference scores, which is 11 instead of 21 in the uncorrelated example. Because the standardized effect size is the difference divided by the standard deviation, the effect size (Cohen's  $d_z$  in within designs) is larger in this test than in the uncorrelated test.

Coordinate system already present. Adding new coordinate system, which will replace the existing one.

The correlation between dependent variables is an important aspect of within designs. I recommend explicitly reporting the correlation between dependent variables in within designs

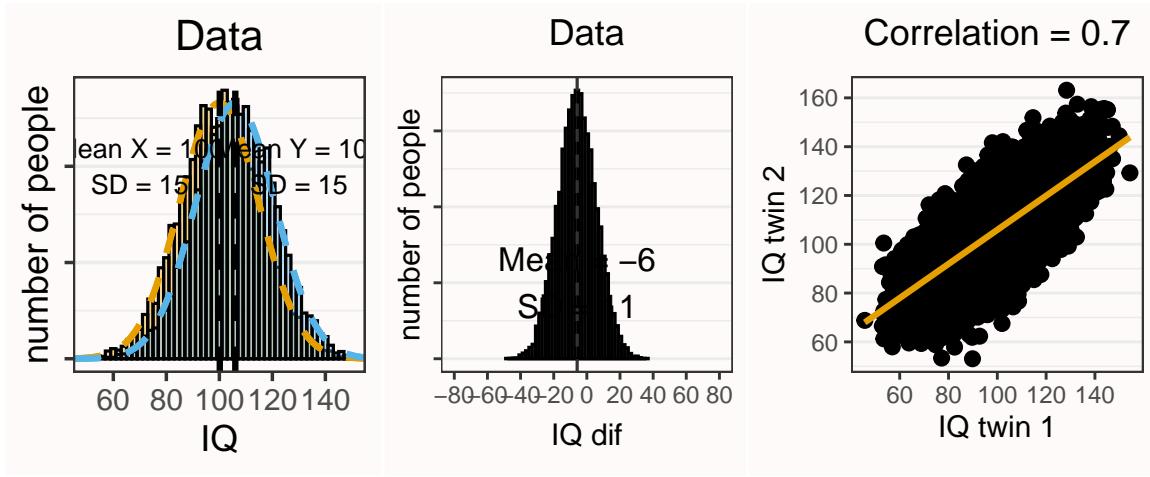


Figure 8.12: Distributions of two independent groups with means 100 and 106 and a standard deviation of 15, distribution of the differences, and correlation of 0.7.

(e.g., participants responded significantly slower ( $M = 390$ ,  $SD = 44$ ) when they used their feet than when they used their hands ( $M = 371$ ,  $SD = 44$ ,  $r = .953$ ),  $t(17) = 5.98$ ,  $p < 0.001$ , Hedges'  $g = 0.43$ ,  $M_{diff} = 19$ , 95% CI [12; 26]). Since most dependent variables in within designs in psychology are positively correlated, within designs will increase the power you can achieve given the sample size you have available. Use within-designs when possible, but weigh the benefits of higher power against the downsides of order effects or carryover effects that might be problematic in a within-subject design (Maxwell et al., 2017).

You can use this [Shiny app](#) to play around with different means, standard deviations, and correlations, and see the effect of the distribution of the difference scores.

In general, the smaller the variation, the larger the standardized effect size (because we are dividing the raw effect by a smaller standard deviation) and thus the higher the power given the same number of observations. Some additional recommendations are provided in the literature (Allison et al., 1997; Bausell & Li, 2002; Hallahan & Rosenthal, 1996), such as:

1. Use better ways to screen participants for studies where participants need to be screened before participation.
2. Assign participants unequally to conditions (if data in the control condition is much cheaper to collect than data in the experimental condition, for example).
3. Use reliable measures that have low error variance (Williams et al., 1995).
4. Smart use of preregistered covariates (Meyvis & Van Osselaer, 2018).

It is important to consider if these ways to reduce the variation in the data do not come at too large a cost for external validity. For example, in an *intention-to-treat* analysis in randomized controlled trials participants who do not comply with the protocol are maintained in the analysis such that the effect size from the study accurately represents the effect of implementing the intervention in the population, and not the effect of the intervention only on those people who perfectly follow the protocol (Gupta, 2011). Similar trade-offs between reducing the variance and external validity exist in other research areas.

## 8.25 Know Your Measure

Although it is convenient to talk about standardized effect sizes, it is generally preferable if researchers can interpret effects in the raw (unstandardized) scores, and have knowledge about the standard deviation of their measures (Baguley, 2009; Lenth, 2001). To make it possible for a research community to have realistic expectations about the standard deviation of measures they collect, it is beneficial if researchers within a research area use the same validated measures. This provides a reliable knowledge base that makes it easier to plan for a desired accuracy, and to use a smallest effect size of interest on the unstandardized scale in an a-priori power analysis.

In addition to knowledge about the standard deviation it is important to have knowledge about the correlations between dependent variables (for example because Cohen's  $d_z$  for a dependent  $t$  test relies on the correlation between means). The more complex the model, the more aspects of the data-generating process need to be known to make predictions. For example, in hierarchical models researchers need knowledge about variance components to be able to perform a power analysis (DeBruine & Barr, 2021; Westfall et al., 2014). Finally, it is important to know the reliability of your measure (Parsons et al., 2019), especially when relying on an effect size from a published study that used a measure with different reliability, or when the same measure is used in different populations, in which case it is possible that measurement reliability differs between populations. With the increasing availability of open data, it will hopefully become easier to estimate these parameters using data from earlier studies.

If we calculate a standard deviation from a sample, this value is an estimate of the true value in the population. In small samples, our estimate can be quite far off, while due to the law of large numbers, as our sample size increases, we will be measuring the standard deviation more accurately. Since the sample standard deviation is an estimate with uncertainty, we can calculate a confidence interval around the estimate (Smithson, 2003), and design pilot studies that will yield a sufficiently reliable estimate of the standard deviation. The confidence interval for the variance  $\sigma^2$  is provided in the following formula, and the confidence for the standard deviation is the square root of these limits:

$$(N - 1)s^2/\chi_{N-1:\alpha/2}^2, (N - 1)s^2/\chi_{N-1:1-\alpha/2}^2$$

Whenever there is uncertainty about parameters, researchers can use sequential designs to perform an *internal pilot study* (Wittes & Brittain, 1990). The idea behind an internal pilot study is that researchers specify a tentative sample size for the study, perform an interim analysis, use the data from the internal pilot study to update parameters such as the variance of the measure, and finally update the final sample size that will be collected. As long as interim looks at the data are blinded (e.g., information about the conditions is not taken into account) the sample size can be adjusted based on an updated estimate of the variance without any practical consequences for the Type I error rate (Friede & Kieser, 2006; Proschan, 2005). Therefore, if researchers are interested in designing an informative study where the Type I and Type II error rates are controlled, but they lack information about the standard deviation, an internal pilot study might be an attractive approach to consider (M. Chang, 2016).

## 8.26 Conventions as meta-heuristics

Even when a researcher might not use a heuristic to directly determine the sample size in a study, there is an indirect way in which heuristics play a role in sample size justifications. Sample size justifications based on inferential goals such as a power analysis, accuracy, or a decision all require researchers to choose values for a desired Type I and Type II error rate, a desired accuracy, or a smallest effect size of interest. Although it is sometimes possible to justify these values as described above (e.g., based on a cost-benefit analysis), a solid justification of these values might require dedicated research lines. Performing such research lines will not always be possible, and these studies might themselves not be worth the costs (e.g., it might require less resources to perform a study with an alpha level that most peers would consider conservatively low, than to collect all the data that would be required to determine the alpha level based on a cost-benefit analysis). In these situations, researchers might use values based on a convention.

When it comes to a desired width of a confidence interval, a desired power, or any other input values required to perform a sample size computation, it is important to transparently report the use of a heuristic or convention (for example by using the accompanying online Shiny app). A convention such as the use of a 5% Type 1 error rate and 80% power practically functions as a lower threshold of the minimum informational value peers are expected to accept *without* any justification (whereas *with* a justification, higher error rates can also be deemed acceptable by peers). It is important to realize that none of these values are set in stone. Journals are free to specify that they desire a higher informational value in their author guidelines (e.g., Nature Human Behavior requires Registered Reports to be designed to achieve 95% statistical power, and my own department has required staff to submit ERB proposals where, whenever possible, the study was designed to achieve 90% power). Researchers who choose to design studies with a higher informational value than a conventional minimum should receive credit for doing so.

In the past some fields have changed conventions, such as the 5 sigma threshold now used in physics to declare a discovery instead of a 5% Type I error rate. In other fields such attempts have been unsuccessful (e.g., Johnson (2013)). Improved conventions should be context dependent, and it seems sensible to establish them through consensus meetings (Mullan & Jacoby, 1985). Consensus meetings are common in medical research, and have been used to decide upon a smallest effect size of interest (for an example, see Fried et al. (1993)). In many research areas current conventions can be improved. For example, it seems peculiar to have a default alpha level of 5% both for single studies and for meta-analyses, and one could imagine a future where the default alpha level in meta-analyses is much lower than 5%. Hopefully, making the lack of an adequate justification for certain input values in specific situations more transparent will motivate fields to start a discussion about how to improve current conventions. The online Shiny app links to good examples of justifications where possible, and will continue to be updated as better justifications are developed in the future.

## 8.27 Sample Size Justification in Qualitative Research

A value of information perspective to sample size justification also applies to qualitative research. A sample size justification in qualitative research should be based on the consideration that the cost of collecting data from additional participants does not yield new information that is valuable enough given the inferential goals. One widely used application of this idea is known as *saturation* and is indicated by the observation that new data replicates earlier observations, without adding new information (Morse, 1995). For example, let's imagine we ask people why they have a pet. Interviews might reveal reasons that are grouped into categories, but after interviewing 20 people, no new categories emerge, at which point saturation has been reached. Alternative philosophies to qualitative research exist, and not all value planning for saturation. Regrettably, principled approaches to justify sample sizes have not been developed for these alternative philosophies (Marshall et al., 2013).

When sampling, the goal is often not to pick a representative sample, but a sample that contains a sufficiently diverse number of subjects such that saturation is reached efficiently. Fugard and Potts (2015) show how to move towards a more informed justification for the sample size in qualitative research based on 1) the number of codes that exist in the population (e.g., the number of reasons people have pets), 2) the probability a code can be observed in a single information source (e.g., the probability that someone you interview will mention each possible reason for having a pet), and 3) the number of times you want to observe each code. They provide an R formula based on binomial probabilities to compute a required sample size to reach a desired probability of observing codes.

A more advanced approach is used in Rijnsoever (2017), which also explores the importance of different sampling strategies. In general, purposefully sampling information from sources you expect will yield novel information is much more efficient than random sampling, but this also requires a good overview of the expected codes, and the sub-populations in which

each code can be observed. Sometimes, it is possible to identify information sources that, when interviewed, would at least yield one new code (e.g., based on informal communication before an interview). A good sample size justification in qualitative research is based on 1) an identification of the populations, including any sub-populations, 2) an estimate of the number of codes in the (sub-)population, 3) the probability a code is encountered in an information source, and 4) the sampling strategy that is used.

## 8.28 Discussion

Providing a coherent sample size justification is an essential step in designing an informative study. There are multiple approaches to justifying the sample size in a study, depending on the goal of the data collection, the resources that are available, and the statistical approach that is used to analyze the data. An overarching principle in all these approaches is that researchers consider the value of the information they collect in relation to their inferential goals.

The process of justifying a sample size when designing a study should sometimes lead to the conclusion that it is not worthwhile to collect the data, because the study does not have sufficient informational value to justify the costs. There will be cases where it is unlikely there will ever be enough data to perform a meta-analysis (for example because of a lack of general interest in the topic), the information will not be used to make a decision or claim, and the statistical tests do not allow you to test a hypothesis with reasonable error rates or to estimate an effect size with sufficient accuracy. If there is no good justification to collect the maximum number of observations that one can feasibly collect, performing the study anyway is a waste of time and/or money (G. W. Brown, 1983; Button et al., 2013; S. D. Halpern et al., 2002).

The awareness that sample sizes in past studies were often too small to meet any realistic inferential goals is growing among psychologists (Button et al., 2013; Fraley & Vazire, 2014; Lindsay, 2015; Sedlmeier & Gigerenzer, 1989). As an increasing number of journals start to require sample size justifications, some researchers will realize they need to collect larger samples than they were used to. This means researchers will need to request more money for participant payment in grant proposals, or that researchers will need to increasingly collaborate (Moshontz et al., 2018). If you believe your research question is important enough to be answered, but you are not able to answer the question with your current resources, one approach to consider is to organize a research collaboration with peers, and pursue an answer to this question collectively.

## 8.29 Test Yourself

**Q1:** A student has at most 2 months to collect data. They need to pay participants for their participation, and their budget is limited to 250 euro. They decide to collect all the

participants they can in the amount of time, and with the money they have available. What type of sample size justification is this?

- (A) Collecting the entire population
- (B) A resource justification
- (C) A heuristic
- (D) No justification

**Q2:** What is the goal of an a-priori power analysis?

- (A) Achieve a desired statistical power for the true effect size, and controlling the Type 1 error rate.
- (B) Achieve a desired statistical power for an effect size of interest, and controlling the Type 1 error rate.
- (C) Achieve a desired statistical power for the true effect size, and controlling the Type 2 error rate.
- (D) Achieve a desired statistical power for an effect size of interest, and controlling the Type 2 error rate.

**Q3:** A researcher already knows the sample size they will be able to collect. Given this sample size, they choose to compute equal Type 1 and Type 2 error rates for an effect size of interest. This is known as:

- (A) An a-priori power analysis
- (B) A sensitivity power analysis
- (C) A post-hoc power analysis
- (D) A compromise power analysis

**Q4:** Looking at the formula in the section ‘Increasing Power Without Increasing the Sample Size’. which two factors contribute to the fact that within subject designs can have much more power, with the same number of participants, than between subject designs?

- (A) The fact a participant contributes multiple observations, and the fact that effect sizes within individuals are typically larger than effect sizes between individuals.
- (B) The fact a participant contributes multiple observations, and the effect of the correlation between measurements.
- (C) The fact that order effects increase the effect size, and the effect of the correlation between measurements.
- (D) The fact that order effects increase the effect size, and the fact that effect sizes within individuals are typically larger than effect sizes between individuals.

**Q5:** Which factors determine the minimal statistically detectable effect?

- (A) The power of the study
- (B) The true effect size in the study
- (C) The sample size and alpha level
- (D) The observed effect size in the sample

**Q6:** All else equal, if you want to perform a study that has the highest possible informational value, which approach to specifying the effect size of interest is the best choice?

- (A) Specify a smallest effect size of interest.
- (B) Compute the minimal statistically detectable effect.
- (C) Use an effect size estimate from a meta-analysis.
- (D) Perform a sensitivity power analysis.

**Q7:** In an a-priori power analysis based on an empirical estimate of the literature, which 2 issues are important to consider, both when using an estimate from a meta-analysis, as from a single study?

- (A) Evaluate the risk of bias in the estimate, and evaluate the uncertainty in the effect size estimate.

- (B) Evaluate the heterogeneity underlying the effect size estimate, and evaluate the similarity of the study/studies the estimate is based on with the study you plan to perform.
- (C) Evaluate the risk of bias in the estimate, and evaluate the similarity of the study/studies the estimate is based on with the study you plan to perform.
- (D) Evaluate the heterogeneity underlying the effect size estimate, and evaluate the uncertainty in the effect size estimate.

**Q8:** Imagine a researcher did not justify their sample size before performing the study, and had no justification for the sample size they choose. After submitting their scientific article to a journal reviewers ask for a justification of the sample size. Of course, honesty requires the authors to write down there was no justification, but how can they still evaluate the informational value of the study for effect sizes of interest?

- (A) Perform an a-priori power analysis
- (B) Perform a compromise power analysis
- (C) Perform a sensitivity power analysis
- (D) Perform a post-hoc or retrospective power analysis

**Q9:** Why can it be useful to consider the effect size distribution of findings in a specific research area when evaluating the informational value of the study you are planning?

- (A) If your study can only reject effects that are so large that they are very unlikely to be observed in a specific research area, collecting the data will not teach us anything we do not already know.
- (B) You can use this information to design a study that has high power for the smallest effect size that is observed in a specific literature, which will lead to a study with high informational value.
- (C) You can use this information to design a study that has high power to detect the median effect size in this literature, which will lead to a study with high informational value.
- (D) You can use this information to design a study that has high power to reject the median effect size in this literature, which will lead to a study with high informational value.

informational value.

**Q10:** Why is it nonsensical to ask researchers to perform a post-hoc or retrospective power analysis, where the observed effect size and the collected sample size is used to calculate the statistical power of a test, when a non-significant finding is observed?

- (A) Post-hoc power analyses are always based on assumptions, and therefore, when the assumptions are wrong, the post-hoc power analysis will not be informative.
- (B) Due to the relationship between post-hoc power and a  $p$ -value, whenever an effect is non-significant, post-hoc power will be low, so the post-hoc power analysis does not provide any useful additional information.
- (C) A post-hoc power analysis is mathematically identical to a sensitivity power analysis for a specific effect size estimate, and it is better to plot power for a range of effect sizes, than for a specific value.
- (D) The question is whether a non-significant effect is a true negative, or a false negative, and the risk of these errors should be controlled in advance through an a-priori power analysis, not after the data is collected through a post-hoc power analysis.

**Q11:** Researchers should not perform a post-hoc power analysis. There are two solutions, one that can be implemented when designing a study, and one when interpreting a non-significant result after the data is in. Which solution can be implemented when the data is in?

- (A) Evaluate the accuracy of the effect size estimate, or perform a sensitivity power analysis.
- (B) Plan a study to have high power for an equivalence test, or perform a sensitivity power analysis.
- (C) Evaluate the accuracy of the effect size estimate, or perform a compromise power analysis.
- (D) Plan a study to have high power for an equivalence test, or perform a compromise power analysis.

**Q12:** What is a way/are ways to increase the statistical power of a test, without increasing the sample size?

- (A) Perform a one-sided test instead of a two-sided test.
- (B) Increase the alpha-level of the test.
- (C) Use measures with higher (compared to lower) error variance.
- (D) All of the other answer options are correct.

### 8.29.1 Open Questions

1. Why are resource constraints, if not the primary justification, always a secondary justification (if it is not possible to collect data from the entire population)?
2. What is the goal of an a-priori power analysis, and why is the goal not to achieve a desired Type 2 error rate for the true effect size?
3. Which factors determine the Minimal Statistically Detectable Effect, and why can it be useful to compute it for a study you are planning to perform?
4. What is a benefit of planning for precision, given that the effect size is typically unknown (and might even be 0). Which aspect of the decisions that need to be made when planning for precision is most difficult to justify?
5. What is a problem of using heuristics as the basis of a sample size justification?
6. It seems counter-intuitive to have a ‘no justification’ category in a chapter on sample size justification, but why is it important to explicitly state there was no justification?
7. From all effect sizes that might be related to the inferential goal in a study, which of the 6 categories in Table 8.2 is the best approach (if it can be specified)?
8. Why can’t you simply take an effect size estimate from a meta-analysis as the basis of an a-priori power analysis for a related study?
9. Why can’t you simply take an effect size estimate from a single study as the basis of an a-priori power analysis for a related study?
10. What is the goal in a compromise power analysis?
11. Why is ‘post-hoc’ or ‘retrospective’ power not a useful way to draw inferences about non-significant results? BEST
12. When would you perform a sensitivity power analysis?
13. How can the statistical power of a study be increased, without increasing the sample size?

14. Why can it be beneficial to use a within-design compared to a between-design (where possible)?

# 9 Equivalence Testing and Interval Hypotheses

Most scientific studies are designed to test the prediction that an effect or a difference exists. Does a new intervention work? Is there a relationship between two variables? These studies are commonly analyzed with a null hypothesis significance test. When a statistically significant  $p$ -value is observed, the null hypothesis can be rejected, and researchers can claim that the intervention works, or that there is a relationship between two variables, with a maximum error rate. But if the  $p$ -value is not statistically significant, researchers very often draw a logically incorrect conclusion: They conclude there is no effect based on  $p > 0.05$ .

Open a result section of an article you are writing, or the result section of an article you have recently read. Search for “ $p > 0.05$ ”, and look carefully at what you or the scientists concluded (in the results section, but also check which claim they make in the discussion section). If you see the conclusion that there was ‘no effect’ or there was ‘no association between variables’, you have found an example where researchers forgot that *absence of evidence is not evidence of absence* (Altman & Bland, 1995). A non-significant result in itself only tells us that we cannot reject the null hypothesis. It is tempting to ask after  $p > 0.05$  ‘so, is the true effect zero?’ But the  $p$ -value from a null hypothesis significance test cannot answer that question. (remember the concept of (mu) discussed in the chapter on *p values*: the answer is neither yes nor no, but we should ‘unask’ the question).

There should be many situations where researchers are interested in examining whether a meaningful effect is absent. For example, it can be important to show two groups do not differ on factors that might be a confound in the experimental design (e.g., examining whether a manipulation intended to increase fatigue did not affect the mood of the participants, by showing that positive and negative affect did not differ between the groups). Researchers might want to know if two interventions work equally well, especially when the newer intervention costs less or requires less effort (e.g., is online therapy just as efficient as in person therapy?). And other times we might be interested to demonstrate the absence of an effect because a theoretical model predicts there is no effect, or because we believe a previously published study was a false positive, and we expect to show the absence of an effect in a replication study (Dienes, 2014). And yet, when you ask researchers if they have ever designed a study where the goal was to show that there was no effect, for example by predicting that there would be no difference between two conditions, many people say they have never designed a study where their main prediction was that the effect size was 0. Researchers almost always predict there is a difference. One reason might be that many researchers would not even know how to statistically support a prediction of an effect size of 0, because they were not trained in the use of equivalence testing.

It is never possible to show an effect is *exactly* 0. Even if you collected data from every person in the world, the effect in any single study will randomly vary around the true effect size of 0 - you might end up with a mean difference that is very close to, but not exactly, zero, in any finite sample. Hodges & Lehmann (1954) were the first to discuss the statistical problem of testing whether two populations have the same mean. They suggest (p. 264) to: “test that their means do not differ by more than an amount specified to represent the smallest difference of practical interest”. Nunnally (1960) similarly proposed a ‘fixed-increment’ hypothesis where researchers compare an observed effect against a range of values that is deemed too small to be meaningful. Defining a range of values considered practically equivalent to the absence of an effect is known as an **equivalence range** (Bauer & Kieser, 1996) or a **region of practical equivalence** (Kruschke, 2013). The equivalence range should be specified in advance, and requires careful consideration of the smallest effect size of interest.

Although researchers have repeatedly attempted to introduce tests against an equivalence range in the social sciences (Cribbie et al., 2004; Hoenig & Heisey, 2001; Levine et al., 2008; Quertemont, 2011; J. L. Rogers et al., 1993), this statistical approach has only recently become popular. During the replication crisis, researchers searched for tools to interpret null results when performing replication studies. Researchers wanted to be able to publish informative null results when replicating findings in the literature that they suspected were false positives. One notable example were the studies on pre-cognition by Daryl Bem, which ostensibly showed that participants were able to predict the future (Bem, 2011). Equivalence tests were proposed as a statistical approach to answer the question whether an observed effect is small enough to conclude that a previous study could not be replicated (S. F. Anderson & Maxwell, 2016; Daniël Lakens, 2017; Simonsohn, 2015). Researchers specify a smallest effect size of interest (for example an effect of 0.5, so for a two-sided test any value outside a range from -0.5 to 0.5) and test whether effects more extreme than this range can be rejected. If so, they can reject the presence of effects that are deemed large enough to be meaningful.

One can distinguish a **nil null hypothesis**, where the null hypothesis is an effect of 0, from a **non-nil null hypothesis**, where the null hypothesis is any other effect than 0, for example effects more extreme than the smallest effect size of interest (Nickerson, 2000). As Nickerson writes:

The distinction is an important one, especially relative to the controversy regarding the merits or shortcomings of NHST inasmuch as criticisms that may be valid when applied to nil hypothesis testing are not necessarily valid when directed at null hypothesis testing in the more general sense.

Equivalence tests are a specific implementation of **interval hypothesis tests**, where instead of testing against a null hypothesis of no effect (that is, an effect size of 0; **nil null hypothesis**), an effect is tested against a null hypothesis that represents a range of non-zero effect sizes (**non-nil null hypothesis**). Indeed, one of the most widely suggested improvements that mitigates the most important limitations of null hypothesis significance testing is to replace the nil null hypothesis with the test of a range prediction (by specifying a non-nil null hypothesis) in

an interval hypothesis test (Daniël Lakens, 2021). To illustrate the difference, Panel A in Figure 9.1 visualizes the results that are predicted in a two-sided null hypothesis test with a nil hypothesis, where the test examines whether an effect of 0 can be rejected. Panel B shows an interval hypothesis where an effect between 0.5 and 2.5 is predicted, where the non-nil null hypothesis consists of values smaller than 0.5 or larger than 2.5, and the interval hypothesis test examines whether values in these ranges can be rejected. Panel C illustrates an equivalence test, which is basically identical to an interval hypothesis test, but the predicted effects are located in a range around 0, and contain effects that are deemed too small to be meaningful.

When an equivalence test is reversed, a researcher designs a study to reject effects less extreme than a smallest effect size of interest (see Panel D in Figure 9.1), it is called a **minimum effect test** (Murphy & Myors, 1999). A researcher might not just be interested in rejecting an effect of 0 (as in a null hypothesis significance test) but in rejecting a range of effects that are too small to be meaningful. All else equal, a study designed to have high power for a minimum effect requires more observations than if the goal had been to reject an effect of zero. As the confidence interval needs to reject a value that is closer to the observed effect size (e.g., 0.1 instead of 0) it needs to be more narrow, which requires more observations.

One benefit of a minimum effect test compared to a null hypothesis test is that there is no distinction between statistical significance and practical significance. As the test value is chosen to represent the minimum effect of interest, whenever it is rejected, the effect is both statistically and practically significant (Murphy et al., 2014). Another benefit of minimum effect tests is that, especially in correlational studies in the social sciences, variables are often connected through causal structures that result in real but theoretically uninteresting nonzero correlations between variables, which has been labeled the ‘crud factor’ (Meehl, 1990a; Orben & Lakens, 2020). Because an effect of zero is unlikely to be true in large correlational datasets, rejecting a nil null hypothesis is not a severe test. Even if the hypothesis is incorrect, it is likely that an effect of 0 will be rejected due to ‘crud’. For this reason, some researchers have suggested to test against a minimum effect of  $r = 0.1$ , as correlations below this threshold are quite common due to theoretically irrelevant correlations between variables (C. J. Ferguson & Heene, 2021).

Figure 9.1 illustrates two-sided tests, but it is often more intuitive and logical to perform one-sided tests. In that case, a minimum effect test would, for example, aim to reject effects smaller than 0.1, and an equivalence test would aim to reject effects larger than for example 0.1. Instead of specifying an upper and lower bound of a range, it is sufficient to specify a single value for one-sided tests. A final variation of a one-sided non-nil null hypothesis test is known as a test for **non-inferiority**, which examines if an effect is larger than the lower bound of an equivalence range. Such a test is for example performed when a novel intervention should not be noticeably worse than an existing intervention, but it can be a tiny bit worse. For example, if a difference between a novel and existing intervention is not smaller than -0.1, and effects smaller than -0.1 can be rejected, one can conclude an effect is non-inferior (Mazzolari et al., 2022; Schumi & Wittes, 2011). We see that extending nil null hypothesis tests to non-nil null hypotheses allow researchers to ask questions that might be more interesting.

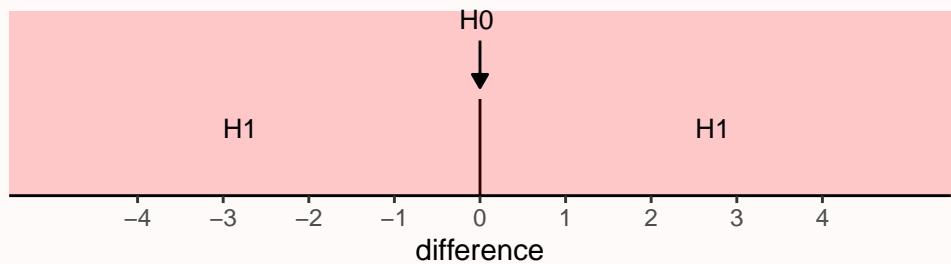
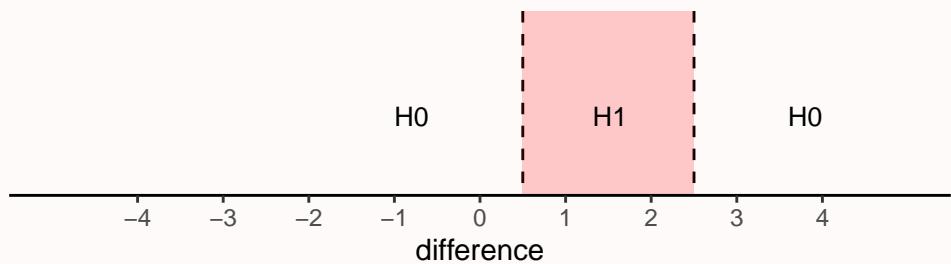
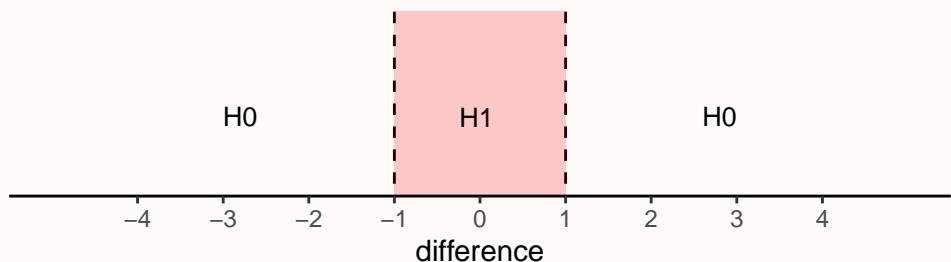
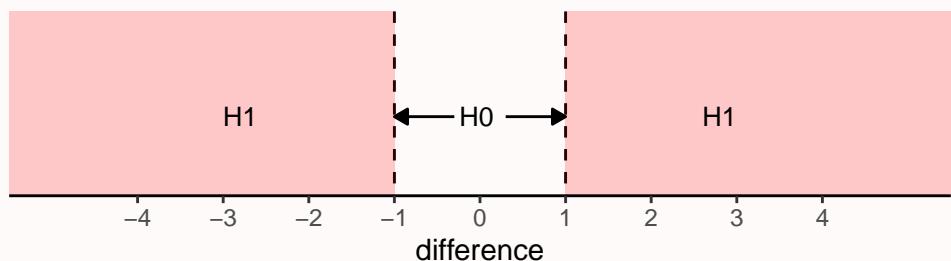
**A: Two-sided NHST****B: Interval Hypothesis Test****C: Equivalence Test****D: Minimum Effect Test**

Figure 9.1: Two-sided null hypothesis test (A), interval hypothesis test (B), equivalence test (C) and minimum effect test (D).

## 9.1 Equivalence tests

Equivalence tests were first developed in pharmaceutical sciences (Hauck & Anderson, 1984; Westlake, 1972) and later formalized as the **two one-sided tests (TOST)** approach to equivalence testing (Schuirmann, 1987; Seaman & Serlin, 1998; Wellek, 2010). The TOST procedure entails performing two one-sided tests to examine whether the observed data is surprisingly larger than a lower equivalence boundary ( $\Delta_L$ ), or surprisingly smaller than an upper equivalence boundary ( $\Delta_U$ ):

$$t_L = \frac{\bar{M}_1 - \bar{M}_2 - \Delta_L}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

and

$$t_U = \frac{\bar{M}_1 - \bar{M}_2 - \Delta_U}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where  $M$  indicates the means of each sample,  $n$  is the sample size, and  $\sigma$  is the pooled standard deviation:

$$\sigma = \sqrt{\frac{(n_1 - 1) s^2_1 + (n_2 - 1) s^2_2}{n_1 + n_2 - 2}}$$

If both one-sided tests are significant, we can reject the presence of effects large enough to be meaningful. The formulas are highly similar to the normal formula for the  $t$ -statistic. The difference between a NHST  $t$ -test and the TOST procedure is that the lower equivalence boundary  $\Delta_L$  and the upper equivalence boundary  $\Delta_U$  are subtracted from the mean difference between groups (in a normal  $t$ -test, we compare the mean difference against 0, and thus the delta drops out of the formula because it is 0).

To perform an equivalence test, you don't need to learn any new statistical tests, as it is just the well-known  $t$ -test against a different value than 0. It is somewhat surprising that the use of *ttests* to perform equivalence tests is not taught alongside their use in null hypothesis significance tests, as there is some indication that this could prevent common misunderstandings of  $p$ -values (Parkhurst, 2001). Let's look at an example of an equivalence test using the TOST procedure.

In a study where researchers are manipulating fatigue by asking participants to carry heavy boxes around, the researchers want to ensure the manipulation does not inadvertently alter participants' moods. The researchers assess positive and negative emotions in both conditions, and want to claim there are no differences in positive mood. Let's assume that positive mood

in the experimental fatigue condition ( $m_1 = 4.55$ ,  $sd_1 = 1.05$ ,  $n_1 = 15$ ) did not differ from the mood in the control condition ( $m_2 = 4.87$ ,  $sd_2 = 1.11$ ,  $n_2 = 15$ ). The researchers conclude: “Mood did not differ between conditions,  $t = -0.81$ ,  $p = .42$ ”. Of course, mood did differ between conditions, as  $4.55 - 4.87 = -0.32$ . The claim is that there was no *meaningful* difference in mood, but to make such a claim in a correct manner, we first need to specify which difference in mood is large enough to be meaningful. For now, let’s assume the researcher consider any effect less extreme half a scale point too small to be meaningful. We now test if the observed mean difference of -0.32 is small enough such that we can reject the presence of effects that are large enough to matter.

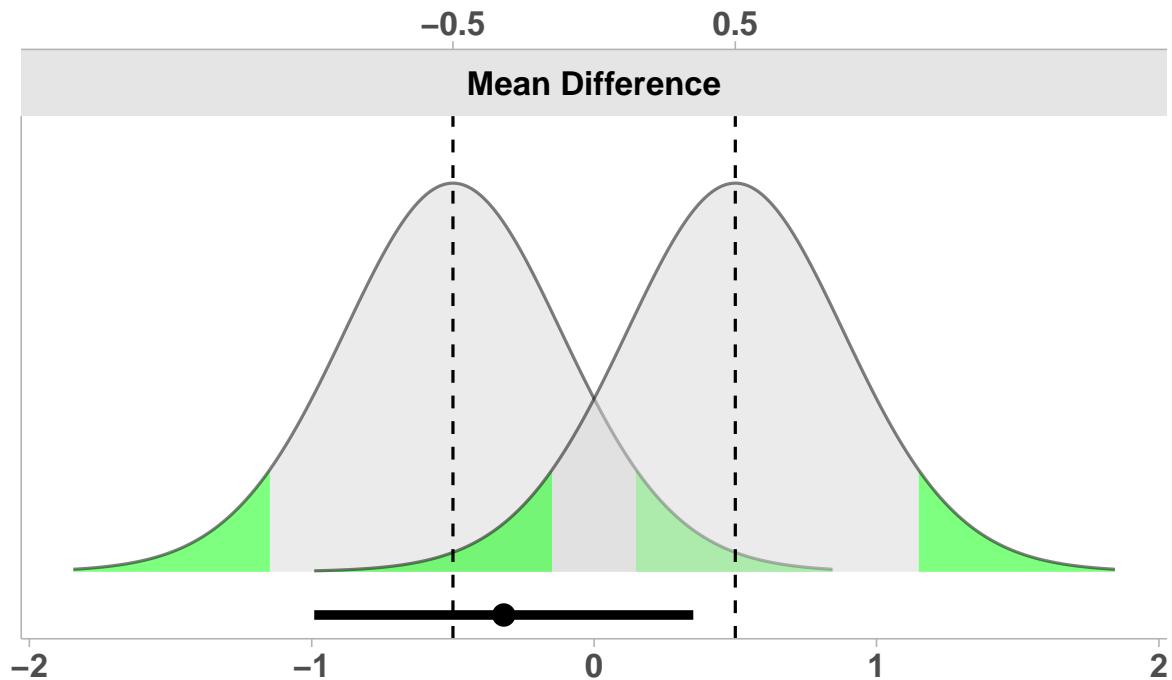
The TOSTER package (originally created by myself but recently redesigned by [Aaron Caldwell](#)) can be used to plot two  $t$ -distributions and their critical regions indicating when we can reject the presence of effects smaller than -0.5 and larger than 0.5. It can take some time to get used to the idea that we are rejecting values more extreme than the equivalence bounds. Try to consistently ask in any hypothesis test: Which values can the test reject? In a nil null hypothesis test, we can reject an effect of 0, and in the equivalence test in the Figure below, we can reject values lower than -0.5 and higher than 0.5. In Figure 9.2 we see two  $t$ -distributions centered on the upper and lower bound of the specified equivalence range (-0.5 and 0.5).

```
res <- TOSTER::tsum_TOST(m1 = 4.55, m2 = 4.87, sd1 = 1.05, sd2 = 1.11,
                           n1 = 15, n2 = 15, low_eqbound = -0.5, high_eqbound = 0.5)

plot(res, type = "tnull")
```

Below the two curves we see a line that represents the confidence interval ranging from -0.99 to 0.35, and a dot on the line that indicates the observed mean difference of -0.32. Let’s first look at the left curve. We see the green highlighted area in the tails that highlights which observed mean differences would be extreme enough to statistically reject an effect of -0.5. Our observed mean difference of -0.32 lies very close to -0.5, and if we look at the left distribution, the mean is not far enough away from -0.5 to fall in the green area that indicates when observed differences would be statistically significant. We can also perform the equivalence test using the TOSTER package, and look at the results.

```
TOSTER::tsum_TOST(m1 = 4.55,
                   m2 = 4.87,
                   sd1 = 1.05,
                   sd2 = 1.11,
                   n1 = 15,
                   n2 = 15,
                   low_eqbound = -0.5,
                   high_eqbound = 0.5)
```



Note: green indicates rejection region for null equivalence and MET hypotheses

Figure 9.2: The mean difference and its confidence interval plotted below the *t*-distributions used to perform the two-one-sided tests against -0.5 and 0.5.

### Welch Two Sample t-test

```
The equivalence test was non-significant, t(27.91) = 0.456, p = 3.26e-01
The null hypothesis test was non-significant, t(27.91) = -0.811, p = 4.24e-01
NHST: don't reject null significance hypothesis that the effect is equal to zero
TOST: don't reject null equivalence hypothesis
```

### TOST Results

	t	df	p.value
t-test	-0.8111	27.91	0.424
TOST Lower	0.4563	27.91	0.326
TOST Upper	-2.0785	27.91	0.023

### Effect Sizes

	Estimate	SE	C.I.	Conf.	Level
Raw	-0.3200	0.3945	[-0.9912, 0.3512]		0.9
Hedges's g(av)	-0.2881	0.3799	[-0.8733, 0.3021]		0.9

Note: SMD confidence intervals are an approximation. See vignette("SMD\_calcs").

In the line ‘t-test’ the output shows the traditional nil null hypothesis significance test (which we already knew was not statistically significant:  $t = 0.46$ ,  $p = 0.42$ ). Just like the default  $t$ -test in R, the tsum\_TOST function will by default calculate Welch’s  $t$ -test (instead of Student’s  $t$ -test), which is a better default (Delacre et al., 2017), but you can request Student’s  $t$ -test by adding var.equal = TRUE as an argument to the function.

We also see a test indicated by TOST Lower. This is the first one-sided test examining if we can reject effects lower than -0.5. From the test result, we see this is not the case:  $t = 0.46$ ,  $p = 0.33$ . This is an ordinary  $t$ -test, just against an effect of -0.5. Because we cannot reject differences more extreme than -0.5, it is possible that a difference we consider meaningful (e.g., a difference of -0.60) is present. When we look at the one-sided test against the upper bound of the equivalence range (0.5) we see that we can statistically reject the presence of mood effects larger than 0.5, as in the line TOST Upper we see  $t = -2.08$ ,  $p = 0.02$ . Our final conclusion is therefore that, even though we can reject effects more extreme than 0.5 based on the observed mean difference of -0.32, we cannot reject effects more extreme than -0.5. Therefore, we cannot completely reject the presence of meaningful mood effects. As the data does not allow us to claim the effect is different from 0, nor that the effect is, if anything, too small to matter (based on an equivalence range from -0.5 to 0.5), the data are **inconclusive**. We cannot distinguish between a Type 2 error (there is an effect, but in this study we just did not detect it) or a true negative (there really is no effect large enough to matter).

Note that because we fail to reject the one-sided test against the lower equivalence bound, the possibility remains that there is a true effect size that is large enough to be considered meaningful. This statement is true, even when the effect size we have observed (-0.32) is

closer to zero than to the equivalence bound of -0.5. One might think the observed effect size needs to be more extreme (i.e.,  $< -0.5$  or  $> 0.5$ ) than the equivalence bound to maintain the possibility that there is an effect that is large enough to be considered meaningful. But that is not required. The 90% CI indicates that some values below -0.5 cannot be rejected. As we can expect that 90% of confidence intervals in the long run capture the true population parameter, it is perfectly possible that the true effect size is more extreme than -0.5. And, the effect might even be more extreme than the values captured by this confidence interval, as 10% of the time, the computed confidence interval is expected to not contain the true effect size. Therefore, when we fail to reject the smallest effect size of interest, we retain the possibility that an effect of interest exists. If we can reject the nil null hypothesis, but fail to reject values more extreme than the equivalence bounds, then we can claim there is an effect, and it might be large enough to be meaningful.

One way to reduce the probability of an inconclusive effect is to collect sufficient data. Let's imagine the researchers had not collected 15 participants in each condition, but 200 participants. They otherwise observe exactly the same data. As explained in the chapter on [confidence intervals](#), as the sample size increases, the confidence interval becomes more narrow. For a TOST equivalence test to be able to reject both the upper and lower bound of the equivalence range, the confidence interval needs to fall completely within the equivalence range. In Figure 9.3 we see the same result as in Figure 9.2, but now if we had collected 200 observations. Because of the larger sample size, the confidence is more narrow than when we collected 15 participants. We see that the 90% confidence interval around the observed mean difference now excludes both the upper and lower equivalence bound. This means that we can now reject effects outside of the equivalence range (even though barely, with a  $p = 0.048$  as the one-sided test against the lower equivalence bound is only just statistically significant).

```
result <- TOSTER::tsum_TOST(m1 = 4.55, m2 = 4.87, sd1 = 1.05, sd2 = 1.11, n1 = 200, n2 = 200)

# plot the result
plot(result, type = "tnull", estimates = "raw")

# print the result
result
```

Welch Two Sample t-test

The equivalence test was significant,  $t(396.78) = 1.666$ ,  $p = 4.82e-02$   
 The null hypothesis test was significant,  $t(396.78) = -2.962$ ,  $p = 3.24e-03$   
 NHST: reject null significance hypothesis that the effect is equal to zero  
 TOST: reject null equivalence hypothesis

TOST Results

```

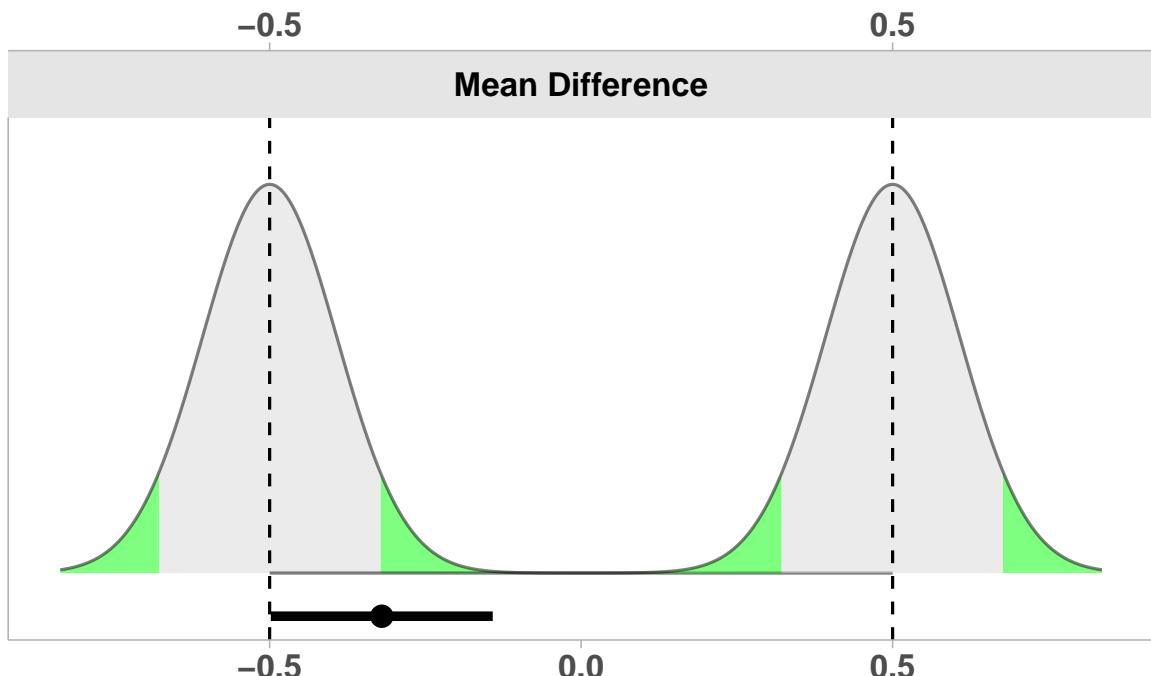
t      df p.value
t-test    -2.962 396.8  0.003
TOST Lower 1.666 396.8  0.048
TOST Upper -7.590 396.8 < 0.001

```

#### Effect Sizes

	Estimate	SE	C.I. Conf.	Level
Raw	-0.3200	0.1080	[-0.4981, -0.1419]	0.9
Hedges's g(av)	-0.2956	0.1008	[-0.4605, -0.1304]	0.9

Note: SMD confidence intervals are an approximation. See vignette("SMD\_calcs").



Note: green indicates rejection region for null equivalence and MET hypotheses

Figure 9.3: The mean difference and its confidence interval for an equivalence test with an equivalence range of -0.5 and 0.5.

In Figure 9.4 we see the same results, but now visualized as a confidence density plot (Schweder & Hjort, 2016), which is a graphical summary of the distribution of confidence. A confidence density plot allows you to see which effects can be rejected with difference confidence interval widths. We see the bounds of the green area (corresponding to a 90% confidence interval) fall inside the equivalence bounds. Thus, the equivalence test is statistically significant, and we can statistically reject the presence of effects outside the equivalence range. We can also see that the 95% confidence interval excludes 0, and therefore, a traditional null

hypothesis significance test is also statistically significant.

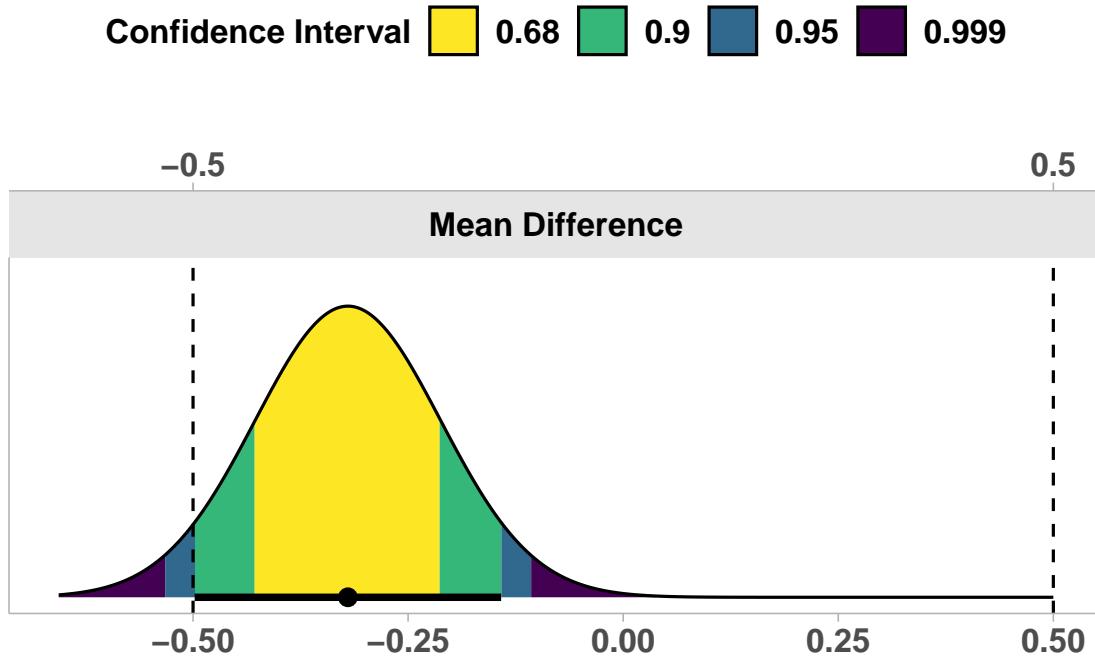


Figure 9.4: The mean difference and its confidence interval for an equivalence test with an equivalence range of -0.5 and 0.5.

In other words, both the null hypothesis test and the equivalence test have yielded significant results. This means we can claim that the observed effect is statistically different from zero, and that the effect is statistically smaller than effects we deemed large enough to matter when we specified the equivalence range from -0.5 to 0.5. This illustrates how combining equivalence tests and nil null hypothesis tests can prevent us from mistaking statistically significant effects for practically significant effects. In this case, with 200 participants, we can reject an effect of 0, but the effect, if any, is not large enough to be meaningful.

## 9.2 Reporting Equivalence Tests

It is common practice to only report the test yielding the higher  $p$ -value of the two one-sided tests when reporting an equivalence test. Because both one-sided tests need to be statistically significant to reject the null hypothesis in an equivalence test (i.e., the presence of effects large enough to matter), when the larger of the two hypothesis tests rejects the equivalence bound, so does the other test. Unlike in null hypothesis significance tests it is not common

to report standardized effect sizes for equivalence tests, but there can be situations where researchers might want to discuss how far the effect is removed from the equivalence bounds on the raw scale. Prevent the erroneous interpretation to claim there is ‘no effect’, that an effect is ‘absent’, that the true effect size is ‘zero’, or vague verbal descriptions, such as that two groups yielded ‘similar’ or ‘comparable’ data. A significant equivalence test rejects effects more extreme than the equivalence bounds. Smaller true effects have not been rejected, and thus it remains possible that there is a true effect. Because a TOST procedure is a frequentist test based on a *p*-value, all other [misconceptions of \*p\*-values](#) should be prevented as well.

When summarizing the main result of an equivalence test, for example in an abstract, always report the equivalence range that the data is tested against. Reading ‘based on an equivalence test we concluded the absence of a meaningful effect’ means something very different if the equivalence bounds were  $d = -0.9$  to  $0.9$  than when the bounds were  $d = -0.2$  to  $d = 0.2$ . So instead, write ‘based on an equivalence test with an equivalence range of  $d = -0.2$  to  $0.2$ , we conclude the absence of an effect we deemed meaningful’. Of course, whether peers agree you have correctly concluded the absence of a meaningful effect depends on whether they agree with your justification for a smallest effect of interest! A more neutral conclusion would be a statement such as: ‘based on an equivalence test, we rejected the presence of effects more extreme than  $-0.2$  to  $0.2$ , so we can act (with an error rate of alpha) as if the effect, if any, is less extreme than our equivalence range’. Here, you do not use value-laden terms such as ‘meaningful’. If both a null hypothesis test and an equivalence test are non-significant, the finding is best described as ‘inconclusive’: There is not enough data to reject the null, or the smallest effect size of interest. If both the null hypothesis test and the equivalence test are statistically significant, you can claim there is an effect, but at the same time claim the effect is too small to be of interest (given your justification for the equivalence range).

Equivalence bounds can be specified in raw effect sizes, or in standardized mean differences. It is better to specify the equivalence bounds in terms of raw effect sizes. Setting them in terms of Cohen’s *d* leads to bias in the statistical test, as the observed standard deviation has to be used to translate the specified Cohen’s *d* into a raw effect size for the equivalence test (and when you set equivalence bounds in standardized mean differences, TOSTER will warn: “Warning: setting bound type to SMD produces biased results!”). The bias is in practice not too problematic in any single equivalence test, and being able to specify the equivalence bounds in standardized mean differences lowers the threshold to perform an equivalence test when they do not know the standard deviation of their measure. But as equivalence testing becomes more popular, and fields establish smallest effect sizes of interest, they should do so in raw effect size differences, not in standardized effect size differences.

### 9.3 Minimum Effect Tests

If a researcher has specified a smallest effect size of interest, and is interested in testing whether the effect in the population is larger than this smallest effect of interest, a minimum effect test

can be performed. As with any hypothesis test, we can reject the smallest effect of interest whenever the confidence interval around the observed effect does not overlap with it. In the case of a minimum effect test, however, the confidence interval should be fall completely beyond the smallest effect size of interest. For example, let's assume a researcher performs a minimum effect test with 200 observations per condition against a smallest effect size of interest of a mean difference of 0.5.

#### Welch Two Sample t-test

The minimal effect test was significant,  $t(396.78) = 12.588$ ,  $p = 4.71e-04$

The null hypothesis test was significant,  $t(396.78) = 7.960$ ,  $p = 1.83e-14$

NHST: reject null significance hypothesis that the effect is equal to zero

TOST: reject null MET hypothesis

#### TOST Results

	t	df	p.value
t-test	7.960	396.8	< 0.001
TOST Lower	12.588	396.8	1
TOST Upper	3.332	396.8	< 0.001

#### Effect Sizes

	Estimate	SE	C.I. Conf. Level
Raw	0.8600	0.1080	[0.6819, 1.0381]
Hedges's g(av)	0.7945	0.1041	[0.6234, 0.9646]

Note: SMD confidence intervals are an approximation. See vignette("SMD\_calcs").

Below the two curves we again see a line that represents the confidence interval ranging from 0.68 to 1.04, and a dot on the line that indicates the observed mean difference of 0.86. The entire confidence interval lies well above the minimum effect of 0.5, and we can therefore not just reject the nil null hypothesis, but also effects smaller than the minimum effect of interest. Therefore, we can claim that the effect is large enough to be not just statistically significant, but also practically significant (as long as we have justified our smallest effect size of interest well). Because we have performed a two-sided minimum effect test, the minimum effect test would also have been significant if the confidence interval had been completely on the opposite side of -0.5.

Earlier we discussed how combining traditional NHST and an equivalence test could lead to more informative results. It is also possible to combine a minimum effect test and an equivalence test. One might even say that such a combination is the most informative test of a prediction whenever a smallest effect size of interest can be specified. In principle, this is true. As long as we are able to collect enough data, we will always get an informative and straightforward answer when we combine a minimum effect test with an equivalence test:

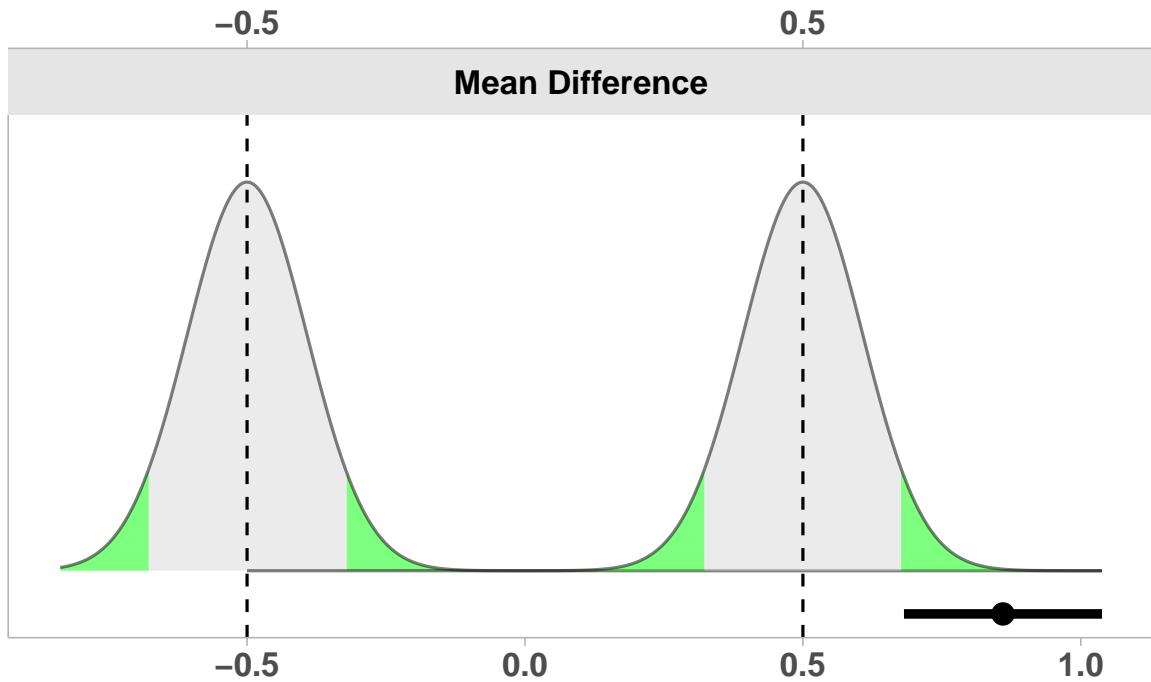


Figure 9.5: The mean difference and its confidence interval plotted below the  $t$ -distributions used to perform the two-one-sided tests against  $-0.5$  and  $0.5$  when performing a minimum effect test.

Either we can reject all effects that are too small to be of interest, or we can reject all effects that are large enough to be of interest. As we will see below in the section on power analysis for interval hypotheses, whenever the true effect size is close to the smallest effect size of interest, a large amount of observations will need to be collected. And if the true effect size happens to be identical to the smallest effect size of interest, neither the minimum effect test nor the equivalence test can be correctly rejected (and any significant test would be a Type 1 error). If a researcher can collect sufficient data (so that the test has high statistical power), and is relatively confident that the true effect size will be larger or smaller than the smallest effect of interest, then the combination of a minimum effect test and an equivalence test can be attractive as such a hypothesis test is likely to yield an informative answer to the research question.

## 9.4 Power Analysis for Interval Hypothesis Tests

When designing a study it is a sensible strategy to always plan for both the presence and the absence of an effect. Several scientific journals require a sample size justification for Registered Reports where the statistical power to reject the null hypothesis is high, but where the study is also capable of demonstrating the absence of an effect, for example by also performing a power analysis for an equivalence test. As we saw in the chapter on [error control](#) and [likelihoods](#) null results are to be expected, and if you only think about the possibility of observing a null effect when the data has been collected, it is often too late.

The statistical power for interval hypotheses depend on the alpha level, the sample size, the smallest effect of interest you decide to test against, and the true effect size. For an equivalence test, it is common to perform a power analysis assuming the true effect size is 0, but this might not always be realistic. The closer the expected effect size is to the smallest effect size of interest, the larger the sample size needed to reach a desired power. Don't be tempted to assume a true effect size of 0, if you have good reason to expect a small but non-zero true effect size. The sample size that the power analysis indicates you need to collect might be smaller, but in reality you also have a higher probability of an inconclusive result. Earlier versions of TOSTER only enabled researchers to perform power analyses for equivalence tests assuming a true effect size of 0, but a new power function by Aaron Caldwell allows users to specify `delta`, the expected effect size.

Assume a researchers desired to achieve 90% power for an equivalence test with an equivalence range from -0.5 to 0.5, with an alpha level of 0.05, and assuming a population effect size of 0. A power analysis for an equivalence test can be performed to examine the required sample size.

```
TOSTER::power_t_TOST(power = 0.9, delta = 0,
                      alpha = 0.05, type = "two.sample",
                      low_eqbound = -0.5, high_eqbound = 0.5)
```

```
Two-sample TOST power calculation
```

```
power = 0.9
beta = 0.1
alpha = 0.05
n = 87.26261
delta = 0
sd = 1
bounds = -0.5, 0.5
```

NOTE: n is number in \*each\* group

We see that the required sample size is 88 participants in each condition for the independent  $t$ -test. Let's compare this power analysis to a situation where the researcher expects a true effect of  $d = 0.1$ , instead of a true effect of 0. To be able to reliably reject effects larger than 0.5, we will need a larger sample size, just as how we need a larger sample size for a null hypothesis test powered to detect  $d = 0.4$  than a null hypothesis test powered to detect  $d = 0.5$ .

```
TOSTER::power_t_TOST(power = 0.9, delta = 0.1,
                      alpha = 0.05, type = "two.sample",
                      low_eqbound = -0.5, high_eqbound = 0.5)
```

```
Two-sample TOST power calculation
```

```
power = 0.9
beta = 0.1
alpha = 0.05
n = 108.9187
delta = 0.1
sd = 1
bounds = -0.5, 0.5
```

NOTE: n is number in \*each\* group

We see the sample size has now increased to 109 participants in each condition. As mentioned before, it is not necessary to perform a two-sided equivalence test. It is also possible to perform a one-sided equivalence test. An example of a situation where such a directional test is appropriate is a replication study. If a previous study observed an effect of  $d = 0.48$ , and

you perform a replication study, you might decide to consider any effect smaller than  $d = 0.2$  a failure to replicate - including any effect in the opposite direction, such as an effect of  $d = -0.3$ . Although most software for equivalence tests requires you to specify an upper and lower bound for an equivalence range, you can mimic a one-sided test by setting the equivalence bound in the direction you want to ignore to a low value so that the one-sided test against this value will always be statistically significant. This can also be used to perform a power analysis for a minimum effect test, where one bound is the minimum effect of interest, and the other bound is set to an extreme value on the other side of the expected effect size.

In the power analysis for an equivalence test example below, the lower bound is set to -5 (it should be set low enough such that lowering it even further has no noticeable effect). We see that the new power function in the TOSTER package takes the directional prediction into account, and just as with directional predictions in a nil null hypothesis test, a directional prediction in an equivalence test is more efficient, and only 70 observations are needed to achieve 90% power.

```
# New TOSTER power functions allows power for expected non-zero effect.
TOSTER::power_t_TOST(power = 0.9, delta = 0,
                      alpha = 0.05, type = "two.sample",
                      low_eqbound = -5, high_eqbound = 0.5)
```

Two-sample TOST power calculation

```
power = 0.9
beta = 0.1
alpha = 0.05
n = 69.19784
delta = 0
sd = 1
bounds = -5.0, 0.5
```

NOTE: n is number in \*each\* group

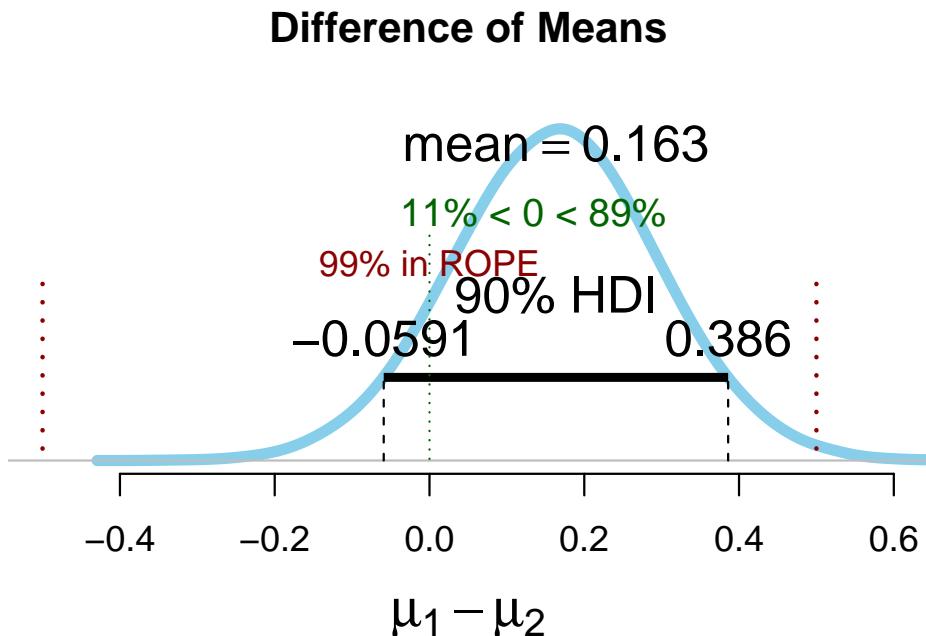
Statistical software offers options for power analyses for some statistical tests, but not for all tests. Just as with power analysis for a nil null hypothesis test, it can be necessary to use a simulation-based approach to power analysis.

## 9.5 The Bayesian ROPE procedure

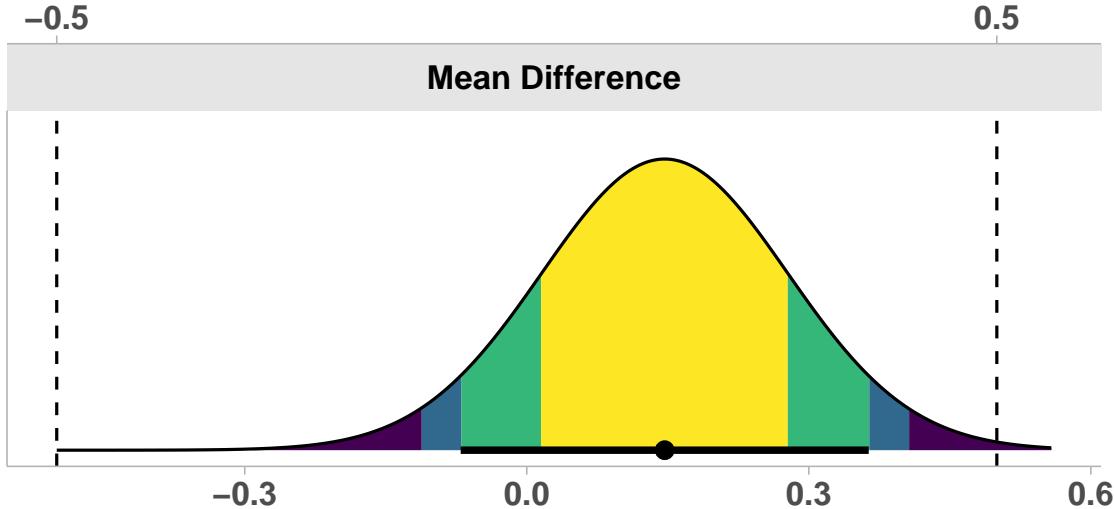
In Bayesian estimation, one way to argue for the absence of a meaningful effect is the **region of practical equivalence** (ROPE) procedure (Kruschke, 2013), which is “somewhat analogous

to frequentist equivalence testing” (Kruschke & Liddell, 2017). In the ROPE procedure, an equivalence range is specified, just as in equivalence testing, but the Bayesian highest density interval based on a posterior distribution (as explained in the chapter on [Bayesian statistics](#)) is used instead of the confidence interval.

If the prior used by Kruschke was perfectly uniform, and the ROPE procedure and an equivalence test used the same confidence interval (e.g., 90%), the two tests would yield identical results. There would only be philosophical differences in how the numbers are interpreted. The BEST package in R that can be used to perform the ROPE procedure by default uses a ‘broad’ prior, and therefore results of the ROPE procedure and an equivalence test are not exactly the same, but they are very close. One might even argue the two tests are ‘practically equivalent’. In the R code below, random normally distributed data for two conditions is generated (with means of 0 and a standard deviation of 1) and the ROPE procedure and a TOST equivalence test are performed.



**Confidence Interval**  0.68  0.9  0.95  0.999



The 90% HDI ranges from -0.06 to 0.39, with an estimated mean based on the prior and the data of 0.164. The HDI falls completely between the upper and the lower bound of the equivalence range, and therefore values more extreme than -0.5 or 0.5 are deemed implausible. The 95% CI ranges from -0.07 to 0.36 with an observed mean difference of 0.15. We see that the numbers are not identical, because in Bayesian estimation the observed values are combined with a prior, and the mean estimate is not purely based on the data. But the results are very similar, and will in most cases lead to similar inferences. The BEST R package also enables researchers to perform simulation based power analyses, which take a long time but, when using a broad prior, yield a result that is basically identical to the sample size from a power analysis for an equivalence test. The biggest benefit of ROPE over TOST is that it allows you to incorporate prior information. If you have reliable prior information, ROPE can use this information, which is especially useful if you don't have a lot of data. If you use informed priors, check the robustness of the posterior against reasonable changes in the prior in sensitivity analyses.

## 9.6 Which interval width should be used?

Because the TOST procedure is based on two one-sided tests, a 90% confidence interval is used when the one-sided tests are performed at an alpha level of 5%. Because both the test against

the upper bound and the test against the lower bound needs to be statistically significant to declare equivalence (which as explained in the chapter on error control is an intersection-union approach to multiple testing) it is not necessary to correct for the fact that two tests are performed. If the alpha level is adjusted for multiple comparisons, or if the alpha level is justified instead of relying on the default 5% level (or both), the corresponding confidence interval should be used, where  $CI = 100 - (2 * \alpha)$ . Thus, the width of the confidence interval is directly related to the choice for the alpha level, as we are making decisions to reject the smallest effect size of interest, or not, based on whether the confidence interval excluded the effect that is tested against.

When using a Highest Density Interval from a Bayesian perspective, such as the ROPE procedure, the choice for a width of a confidence interval does not follow logically from a desired error rate, or any other principle. Kruschke (2014) writes: “How should we define ‘reasonably credible’? One way is by saying that any points within the 95% HDI are reasonably credible.” McElreath (2016) has recommended the use of 67%, 89%, and 97%, because “No reason. They are prime numbers, which makes them easy to remember.”. Both these suggestions lack a solid justification. As Gosset (or Student), observed (1904):

Results are only valuable when the amount by which they probably differ from the truth is so small as to be insignificant for the purposes of the experiment. What the odds selected should be depends-

1. On the degree of accuracy which the nature of the experiment allows, and
2. On the importance of the issues at stake.

There are only two principled solutions. First, if a highest density interval width is used to make claims, these claims will be made with certain error rates, and researchers should quantify the risk of erroneous claims by computing frequentist error rates. This would make the ROPE procedure a Bayesian/Frequentist compromise procedure, where the computation of a posterior distribution allows for Bayesian interpretations of which parameters values are believed to be most probable, while decisions based on whether or not the HDI falls within an equivalence range have a formally controlled error rate. Note that when using an informative prior, an HDI does not match a CI, and the error rate when using an HDI can only be derived through simulations. The second solution is to not make any claims, present the full posterior distribution, and let readers draw their own conclusions.

## 9.7 Setting the Smallest Effect Size of Interest

To perform an equivalence test we need to specify which observed values are too small to be meaningful. We can never say that an effect is exactly zero, but we can examine whether observed effects are too small to be theoretically or practically interesting. This requires that we specify the **smallest effect size of interest** (SESOI). The same concept goes by many names, such as a minimal important difference, or clinically significant difference (King, 2011).

Take a moment to think about what the smallest effect size is that you would still consider theoretically or practically meaningful for the next study you are designing. It might be difficult to determine what the smallest effect size is that you would consider interesting, and the question what the smallest effect size of interest is might be something you have never really thought about to begin with. However, determining your smallest effect size of interest has important practical benefits. First, if researchers in a field are able to specify which effects would be too small to matter, it becomes very straightforward to power a study for the effects that are meaningful. The second benefit of specifying the smallest effect size of interest is that it makes your study falsifiable. Having your predictions falsified by someone else might not feel that great for you personally, but it is quite useful for science as a whole (Popper, 2002). After all, if there is no way a prediction can be wrong, why would anyone be impressed if the prediction is right?

To start thinking about which effect sizes matter, ask yourself whether *any* effect in the predicted direction is actually support for the alternative hypothesis. For example, would an effect size of a Cohen's  $d$  of 10 be support for your hypothesis? In psychology, it should be rare that a theory predicts such a huge effect, and if you observed a  $d = 10$ , you would probably check for either a computation error, or a confound in the study. On the other end of the scale, would an effect of  $d = 0.001$  be in line with the theoretically proposed mechanism? Such an effect is incredibly small, and is well below what an individual would notice, as it would fall below the **just noticeable difference** given perceptual and cognitive limitations. Therefore, a  $d = 0.001$  would in most cases lead researchers to conclude "Well, this is really too small to be something that my theory has predicted, and such a small effect is practically equivalent to the absence of an effect." However, when we make a directional prediction, we say that these types of effects are all part of our alternative hypothesis. Even though many researchers would agree such tiny effects are too small to matter, they still officially support for our alternative hypothesis if we have a directional prediction with a nil null hypothesis. Furthermore, researchers rarely have the resources to statistically reject the presence of effects this small, so the claim that such effects would still support a theoretical prediction makes the theory **practically unfalsifiable**: A researcher could simply respond to any replication study showing a non-significant small effect (e.g.,  $d = 0.05$ ) by saying: "That does not falsify my prediction. I suppose the effect is just a bit smaller than  $d = 0.05$ ", without ever having to admit the prediction is falsified. This is problematic, because if we do not have a process of replication and falsification, a scientific discipline risks a slide towards the unfalsifiable (C. J. Ferguson & Heene, 2012). So whenever possible, when you design an experiment or you have a theory and a theoretical prediction, carefully think about, and clearly state, what the smallest effect size of interest is.

## 9.8 Specifying a SESOI based on theory

One example of a theoretically predicted smallest effect size of interest can be found in the study by Burriss et al. (2015), who examined whether women displayed increased redness in

the face during the fertile phase of their ovulatory cycle. The hypothesis was that a slightly redder skin signals greater attractiveness and physical health, and that sending this signal to men yields an evolutionary advantage. This hypothesis presupposes that men can detect the increase in redness with the naked eye. Burris et al. collected data from 22 women and showed that the redness of their facial skin indeed increased during their fertile period. However, this increase was not large enough for men to detect with the naked eye, so the hypothesis was falsified. Because the just-noticeable difference in redness of the skin can be measured, it was possible to establish a theoretically motivated SESOI. A theoretically motivated smallest effect size of interest can be derived from just-noticeable differences, which provide a lower bound on effect sizes that can influence individuals, or based on computational models, which can provide a lower bound on parameters in the model that will still be able to explain observed findings in the empirical literature.

## 9.9 Anchor based methods to set a SESOI

Building on the idea of a just-noticeable difference, psychologists are often interested in effects that are large enough to be noticed by single individuals. One procedure to estimate what constitutes a meaningful change on an individual level is the anchor-based method (Jaeschke et al., 1989; King, 2011; Norman et al., 2004). Measurements are collected at two time points (e.g., a quality of life measure before and after treatment). At the second time point, an independent measure (the anchor) is used to determine if individuals show no change compared to time point 1, or if they have improved, or worsened. Often, the patient is directly asked to answer the anchor question, and indicate if they subjectively feel the same, better, or worse at time point 2 compared to time point 1. Button et al. (2015) used an anchor-based method to estimate that a minimal clinically important difference on the Beck Depression Inventory corresponded to a 17.5% reduction in scores from baseline.

Anvari and Lakens (2021) applied the anchor-based method to examine a smallest effect of interest as measured by the widely used Positive and Negative Affect Scale (PANAS). Participants completed the 20 item PANAS at two time points several days apart (using a Likert scale going from 1 = “very slightly or not at all”, to 5 = “extremely”). At the second time point they were also asked to indicate if their affect had changed a little, a lot, or not at all. When people indicated their affect had changed “a little”, the average change in Likert units was 0.26 scale points for positive affect and 0.28 scale points for negative affect. Thus, an intervention to improve people’s affective state that should lead to what individuals subjectively consider at least a little improvement might set the SESOI at 0.3 units on the PANAS.

## **9.10 Specifying a SESOI based on a cost-benefit analysis**

Another principled approach to justify a smallest effect size of interest is to perform a cost-benefit analysis. Research shows that cognitive training may improve mental abilities in older adults which might benefit older drivers (Ball et al., 2002). Based on these findings, Viamonte, Ball, and Kilgore (2006) performed a cost-benefit analysis and concluded that based on the cost of the intervention (\$247.50), the probability of an accident for drivers older than 75 ( $p = 0.0710$ ), and the cost of an accident (\$22,000), performing the intervention on all drivers aged 75 or older was more efficient than not intervening or only intervening after a screening test. Furthermore, sensitivity analyses revealed that intervening for all drivers would remain beneficial as long as the reduction in collision risk is 25%. Therefore, a 25% reduction in the probability of elderly above 75 getting into a car accident could be set as the smallest effect size of interest.

For another example, economists have examined the value of a statistical life, based on willingness to pay to reduce the risk of death, at \$1.5 - \$2.5 million (in the year 2000, in western countries, see Mrozek & Taylor (2002)). Building on this work, Abelson (2003) calculated the willingness to pay to prevent acute health issues such as eye irritation at about \$40-\$50 per day. A researcher may be examining a psychological intervention that reduces the amount of times people touch their face close to their eyes, thereby reducing eye irritations caused by bacteria. If the intervention costs \$20 per year to administer, it therefore should reduce the average number of days with eye irritation in the population by at least 0.5 days for the intervention to be worth the cost. A cost-benefit analysis can also be based on the resources required to empirically study a very small effect when weighed against the value this knowledge would have for the scientific community.

## **9.11 Specifying the SESOI using the small telescopes approach**

Ideally, researchers who publish empirical claims would always specify which observations would falsify their claim. Regrettably, this is not yet common practice. This is particularly problematic when a researcher performs a close replication of earlier work. Because it is never possible to prove an effect is exactly zero, and the original authors seldom specify which range of effect sizes would falsify their hypotheses, it has proven to be very difficult to interpret the outcome of a replication study (S. F. Anderson & Maxwell, 2016). When does the new data contradict the original finding?

Consider a study in which you want to test the idea of the wisdom of crowds. You ask 20 people to estimate the number of coins in a jar, expecting the average to be very close to the true value. The research question is whether the people can on average correctly guess the number of coins, which is 500. The observed mean guess by 20 people is 550, with a standard deviation of 100. The observed difference from the true value is statistically significant,  $t(19)=2.37$ ,  $p = 0.0375$ , with a Cohen's  $d$  of 0.5. Can it really be that the group average is so far off? Is there

no Wisdom of Crowds? Was there something special about the coins you used that make it especially difficult to guess their number? Or was it just a fluke? You set out to perform a close replication of this study.

You want your study to be informative, regardless of whether there is an effect or not. This means you need to design a replication study that will allow you to draw an informative conclusion, regardless of whether the alternative hypothesis is true (the crowd will not estimate the true number of coins accurately) or whether the null hypothesis is true (the crowd will guess 500 coins, and the original study was a fluke). But since the original researcher did not specify a smallest effect size of interest, when would a replication study allow you to conclude the original study is contradicted by the new data? Observing a mean of exactly 500 would perhaps be considered by some to be quite convincing, but due to random variation you will (almost) never find a mean score of exactly 500. A non-significant result can't be interpreted as the absence of an effect, because your study might have too small a sample size to detect meaningful effects, and the result might be a Type 2 error. So how can we move forward and define an effect size that is meaningful? How can you design a study that has the ability to falsify a previous finding?

Uri Simonsohn (2015) defines a small effect as “one that would give 33% power to the original study”. In other words, the effect size that would give the original study odds of 2:1 *against* observing a statistically significant result if there was an effect. The idea is that if the original study had 33% power, the probability of observing a significant effect, if there was a true effect, is too low to reliably distinguish signal from noise (or situations where there is a true effect from situations where there is no true effect). Simonsohn (2015, p. 561) calls this the **small telescopes approach**, and writes: “Imagine an astronomer claiming to have found a new planet with a telescope. Another astronomer tries to replicate the discovery using a larger telescope and finds nothing. Although this does not prove that the planet does not exist, it does nevertheless contradict the original findings, because planets that are observable with the smaller telescope should also be observable with the larger one.”

Although this approach to setting a smallest effect size of interest (SESOI) is arbitrary (why not 30% power, or 35%) it suffices for practical purposes (and you are free to choose a power level you think is too low). The nice thing about this definition of a SESOI is that if you know the sample size of the original study, you can always calculate the effect size that study had 33% power to detect. You can thus always use this approach to set a smallest effect size of interest. If you fail to find support for an effect size the original study has 33% power to detect, it does not mean there is no true effect, and not even that the effect is too small to be of any theoretical or practical interest. But using the small telescopes approach is a good first step, since it will get the conversation started about which effects are meaningful and allows researchers who want to replicate a study to specify when they would consider the original claim falsified.

With the small telescopes approach, the SESOI is based only on the sample size in the original study. A smallest effect size of interest is set only for effects in the same direction. All effects smaller than this effect (including large effects in the opposite direction) are interpreted as a

failure to replicate the original results. We see that the small telescopes approach is a **one-sided equivalence test**, where only the upper bound is specified, and the smallest effect size of interest is determined based on the sample size of the original study. The test examines if we can reject effects as large or larger than the effect the original study has 33% power to detect. It is a simple one-sided test, not against 0, but against a SESOI.

For example, consider our study above in which 20 guessers tried to estimate the number of coins. The results were analyzed with a two-sided one-sample  $t$ -test, using an alpha level of 0.05. To determine the effect size that this study had 33% power for, we can perform a sensitivity analysis. In a sensitivity analysis we compute the required effect size given the alpha, sample size, and desired statistical power. Note that Simonsohn uses a two-sided test in his power analyses, which we will follow here – if the original study reported a pre-registered directional prediction, the power analysis should be based on a one-sided test. In this case, the alpha level is 0.05, the total sample size is 20, and the desired power is 33%. We compute the effect size that gives us 33% power and see that it is a Cohen's  $d$  of 0.358. This means we can set our smallest effect size of interest for the replication study to  $d = 0.358$ . If we can reject effects as large or larger than  $d = 0.358$ , we can conclude that the effect is smaller than anything the original study had 33% power for. The screenshot below illustrates the correct settings in G\*Power, and the code in R is:

```
library("pwr")

pwr::pwr.t.test(
  n = 20,
  sig.level = 0.05,
  power = 0.33,
  type = "one.sample",
  alternative = "two.sided"
)
```

One-sample t test power calculation

```
n = 20
d = 0.3577466
sig.level = 0.05
power = 0.33
alternative = two.sided
```

Determining the SESOI based on the effect size the original study had 33% power to detect has an additional convenient property. Imagine the true effect size is actually 0, and you perform a statistical test to see if the data is statistically smaller than the SESOI based on the small

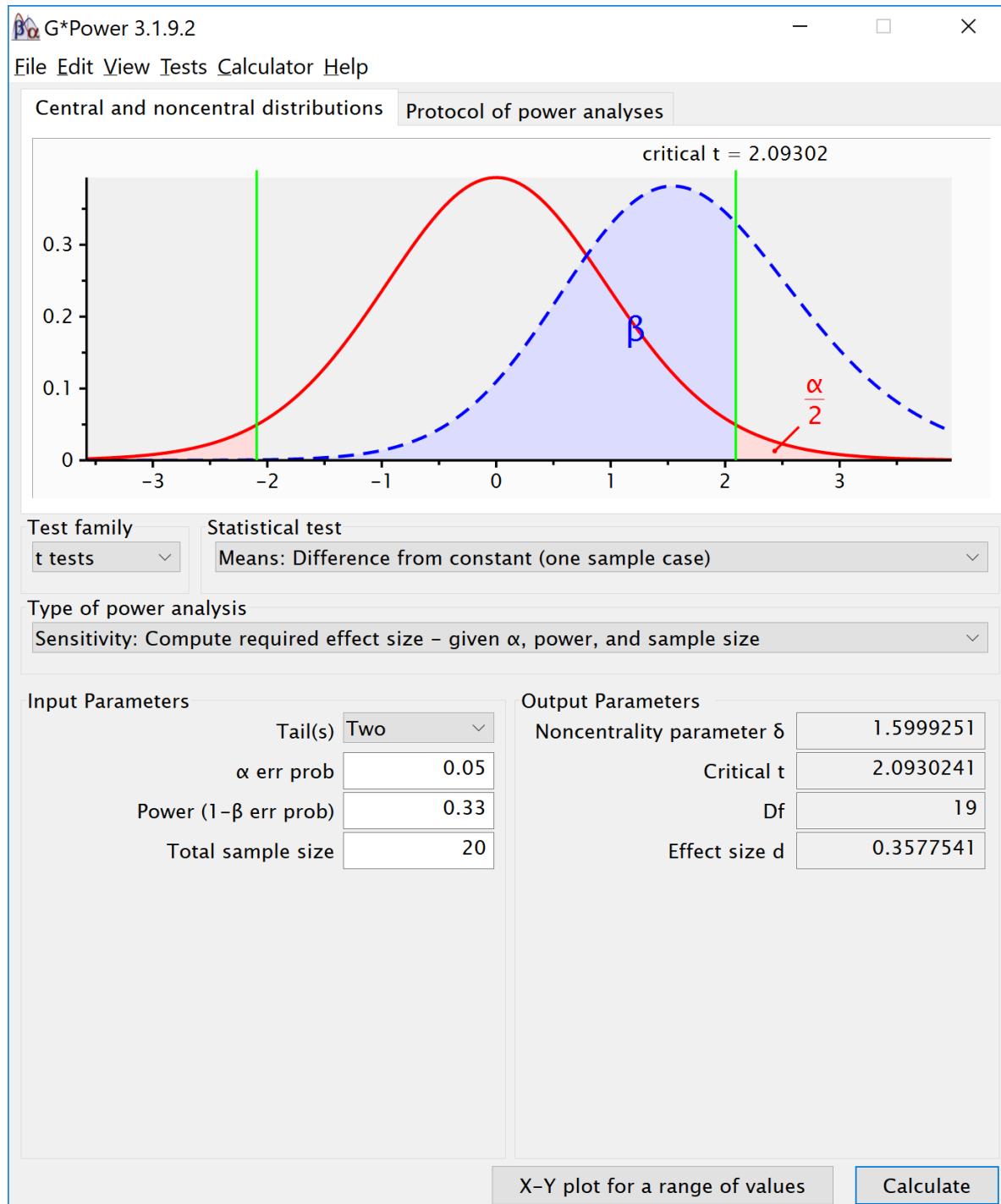


Figure 9.6: Screenshot illustrating a sensitivity power analysis in G\*Power to compute the effect size an original study had 33% power to detect.

telescopes approach (which is called an inferiority test). If you increase the sample size by 2.5 times, you will have approximately 80% power for this one-sided equivalence test, assuming the true effect size is exactly 0 (e.g.,  $d = 0$ ). People who do a replication study can follow the small telescopes recommendations, and very easily determine both the smallest effect size of interest, and the sample size needed to design an informative replication study, assuming the true effect size is 0 (but see the section above for a-priori power analyses where you want to test for equivalence, but do not expect a true effect size of 0).

The figure below, from Simonsohn (2015) illustrates the small telescopes approach using a real-life example. The original study by Zhong and Liljenquist (2006) had a tiny sample size of 30 participants in each condition and observed an effect size of  $d = 0.53$ , which was barely statistically different from zero. Given a sample size of 30 per condition, the study had 33% power to detect effects larger than  $d = 0.401$ . This “small effect” is indicated by the green dashed line. In R, the smallest effect size of interest is calculated using:

```
pwr::pwr.t.test(
  n = 30,
  sig.level = 0.05,
  power = 1/3,
  type = "two.sample",
  alternative = "two.sided"
)
```

Two-sample t test power calculation

```
n = 30
d = 0.401303
sig.level = 0.05
power = 0.3333333
alternative = two.sided
```

NOTE: n is number in \*each\* group

Note that 33% power is a rounded value, and the calculation uses 1/3 (or 0.333333...).

We can see that the first replication by Gámez and colleagues also had a relatively small sample size ( $N = 47$ , compared to  $N = 60$  in the original study), and was not designed to yield informative results when interpreted with a small telescopes approach. The confidence interval is very wide and includes the null effect ( $d = 0$ ) and the smallest effect size of interest ( $d = 0.401$ ). Thus, this study is inconclusive. We can't reject the null, but we can also not reject effect sizes of 0.401 or larger that are still considered to be in line with the original result. The second replication has a much larger sample size, and tells us that we can't reject the null,

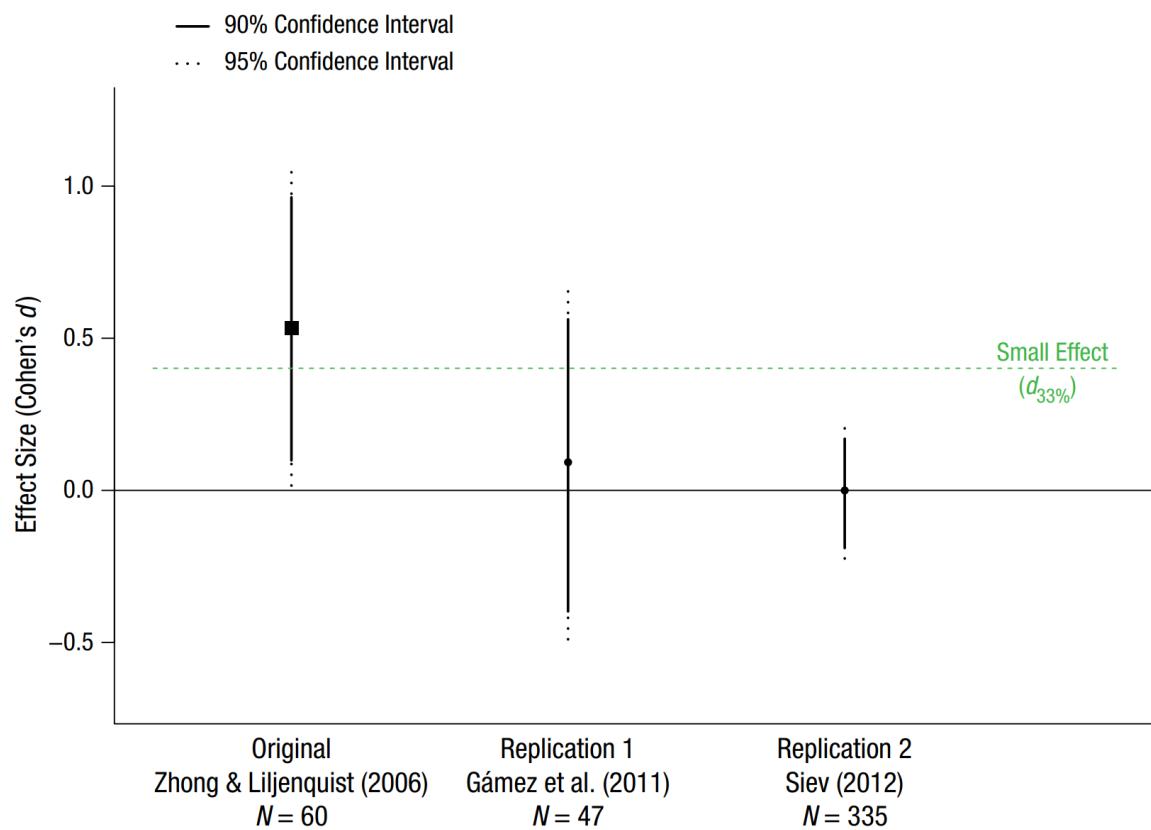


Figure 9.7: Example used in Simonsohn (2015) of an original study and two replication studies.

but we can reject the smallest effect size of interest, suggesting that the effect is smaller than what is considered an interesting effect based on the small telescopes approach.

Although the *small telescope* recommendations are easy to use, one should take care not to turn any statistical procedure into a heuristic. In our example above with the 20 referees, a Cohen's  $d$  of 0.358 would be used as a smallest effect size of interest, and a sample size of 50 would be collected (2.5 times the original 20), but if someone would make the effort to perform a replication study, it would be relatively easy to collect a larger sample size. Alternatively, had the original study been extremely large, it would have had high power for effects that might not be practically significant, and we would not want to collect 2.5 times as many observations in a replication study. Indeed, as Simonsohn writes: "whether we need 2.5 times the original sample size or not depends on the question we wish to answer. If we are interested in testing whether the effect size is smaller than d33%, then, yes, we need about 2.5 times the original sample size no matter how big that original sample was. When samples are very large, however, that may not be the question of interest." Always think about the question you want to ask, and design the study so that it provides an informative answer for a question of interest. Do not automatically follow a 2.5 times  $n$  heuristic, and always reflect on whether the use of a suggested procedure is appropriate in your situation.

## 9.12 Setting the Smallest Effect Size of Interest to the Minimal Statistically Detectable Effect

Given a sample size and alpha level, every test has a minimal statistically detectable effect. For example, given a test with 86 participants in each group, and an alpha level of 5%, only  $t$ -tests which yield a  $t = 1.974$  will be statistically significant. In other words,  $t = 1.974$  is the **critical  $t$ -value**. Given a sample size and alpha level, the critical  $t$ -value can be transformed into a **critical  $d$ -value**. As visualized in Figure 9.8, with  $n = 50$  in each group and an alpha level of 5% the critical  $d$ -value is 0.4. This means that only effects larger than 0.4 will yield a  $p < 0.05$ . The critical  $d$ -value is influenced by the sample size per group, and the alpha level, but does not depend on the true effect size.

It is possible to observe a statistically significant test result if the true effect size is *smaller* than the critical effect size. Due to random variation, it is possible to observe a larger value in a *sample* than is the true value in the population. This is the reason the statistical power of a test is never 0 in a null hypothesis significance test. As illustrated in Figure 9.9, even if the true effect size is smaller than the critical value (i.e., if the true effect size is 0.2) we see from the distribution that we can expect some *observed effect sizes* to be larger than 0.4 when the *true population effect size* is  $d = 0.2$  – if we compute the statistical power for this test, it turns out we can expect 16.77% of the *observed effect sizes* will be larger than 0.4, in the long run. That is not a lot, but it is something. This is also the reason why publication bias combined with underpowered research is problematic: It leads to a large **overestimation of**

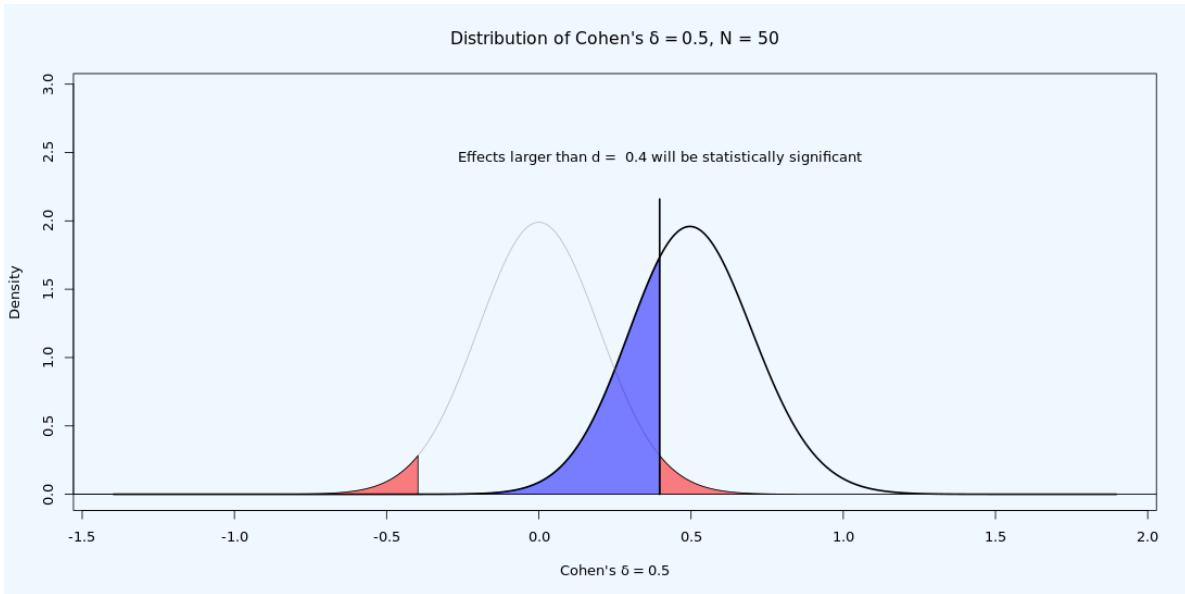


Figure 9.8: Null and alternative distribution with Type 1 and Type 2 error indicating the smallest effect size that will be statistically significant with  $n = 50$  per condition.

**the true effect size** when only observed effect sizes from statistically significant findings in underpowered studies end up in the scientific literature.

We can use the minimal statistically detectable effect to set the SESOI for replication studies. If you attempt to replicate a study, one justifiable option when choosing the smallest effect size of interest (SESOI) is to use the smallest observed effect size that could have been statistically significant in the study you are replicating. In other words, you decide that effects that could not have yielded a  $p$ -value less than  $\alpha\%$  in an original study will not be considered meaningful in the replication study. The assumption here is that the original authors were interested in observing a significant effect, and thus were not interested in observed effect sizes that could not have yielded a significant result. It might be likely that the original authors did not consider which effect sizes their study had good statistical power to detect, or that they were interested in smaller effects but gambled on observing an especially large effect in the sample purely as a result of random variation. Even then, when building on earlier research that does not specify a SESOI, a justifiable starting point might be to set the SESOI to the smallest effect size that, when observed in the original study, **could have been statistically significant**. Not all researchers might agree with this (e.g., the original authors might say they actually cared just as much about an effect of  $d = 0.001$ ). However, as we try to change the field from the current situation where no one specifies what would falsify their hypothesis, or what their smallest effect size of interest is, this approach is one way to get started. In practice, as explained in the section on [post-hoc power](#), due to the relation between  $p = 0.05$  and 50% power for the observed effect size, this justification for a SESOI will mean that the

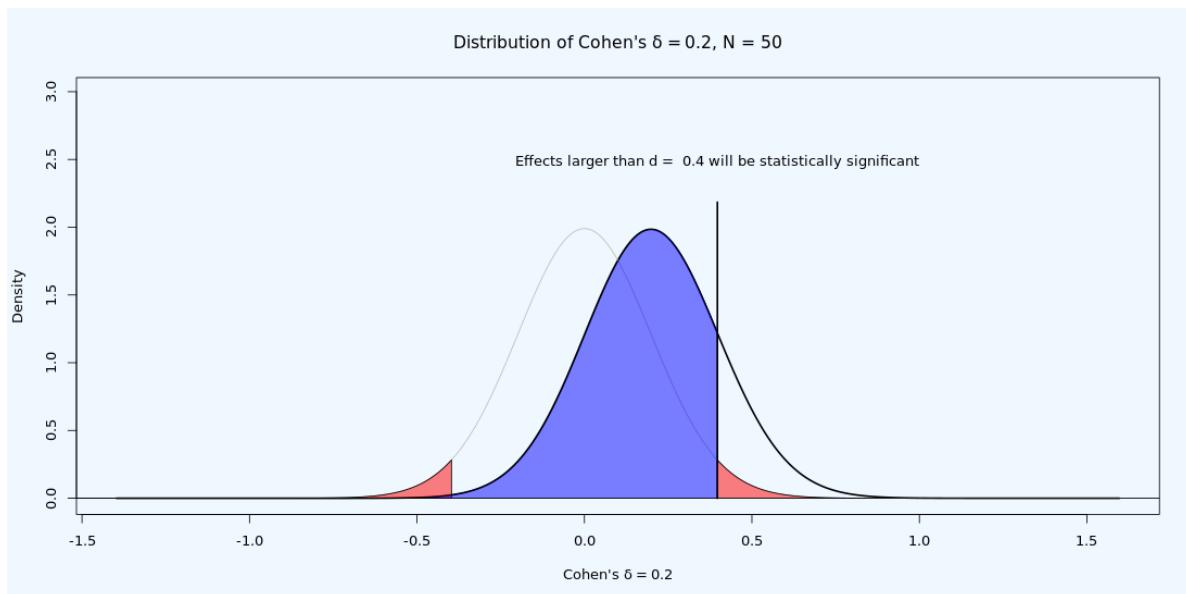


Figure 9.9: Null and alternative distribution with Type 1 and Type 2 error indicating the smallest effect size that will be statistically significant with  $n = 50$  per condition.

SESOI is set to the effect size the original study had 50% power to detect for an independent  $t$ -test. This approach is in some ways similar to the small telescopes approach by Simonsohn (2015), except that it will lead to a somewhat larger SESOI.

Setting a smallest effect size of interest for a replication study is a bit like a tennis match. Original authors serve and hit the ball across the net, saying ‘look, something is going on’. The approach to set the SESOI to the effect size that could have been significant in the original study is a return volley which allows you to say ‘there does not seem to be anything large enough that could have been significant in your own original study’ after performing a well-designed replication study with high statistical power to reject the SESOI. This is never the end of the match – the original authors can attempt to return the ball with a more specific statement about effects their theory predicts, and demonstrate such a smaller effect size is present. But the ball is back in their court, and if they want to continue to claim there is an effect, they will have to support their claim by new data.

Beyond replication studies, the minimal statistically detectable effect can also be computed based on the sample sizes that are typically used in a research field. For example, imagine a line of research in which a hypothesis has almost always been tested by performing a one-sample  $t$ -test, and where the sample sizes that are collected are always smaller than 100 observations. A one-sample  $t$ -test on 100 observations, using an alpha of .05 (two sided), has 80% power to detect an effect of  $d = 0.28$  (as can be calculated in a sensitivity power analysis). In a new study, concluding that one can reliably reject the presence of effects more extreme than  $d = 0.28$  suggests that sample sizes of 100 might not be enough to detect effects in such

research lines. Rejecting the presence of effects more extreme than  $d = 0.28$  does not test a theoretical prediction, but it contributes to the literature by answering a **resource question**. It suggests that future studies in this research line will need to change the design of their studies by substantially increasing the sample size. Setting the smallest effect size of interest based on this approach does not answer any theoretical question (after all, the SESOI is not based on any theoretical prediction). But this approach to specifying a smallest effect size of interest can make a useful contribution to the literature by informing peers that the effect is not large enough so that it can be reliably studied given the sample sizes researchers are typically collecting. It does not mean that the effect is not interesting, but it might be an indication that the field will need to coordinate data collection in the future, because the effect is too small to reliable study with the sample sizes that have been collected in the past.

## 9.13 Test Yourself

### 9.13.1 Questions about equivalence tests

**Q1:** When the 90% CI around a mean difference falls just within the equivalence range from -0.4 to 0.4, we can reject the smallest effect size of interest. Based on your knowledge about confidence intervals, when the equivalence range is changed to -0.3 – 0.3, what is needed for the equivalence test to be significant (assuming the effect size estimate and standard deviation remains the same)?

- (A) A larger effect size.
- (B) A lower alpha level.
- (C) A larger sample size.
- (D) Lower statistical power.

**Q2:** Why is it incorrect to conclude that there is no effect, when an equivalence test is statistically significant?

- (A) An equivalence test is a statement about the data, not about the presence or absence of an effect.
- (B) The result of an equivalence test could be a Type 1 error, and therefore, one should conclude that there is no effect, or a Type 1 error has been observed.

- (C) An equivalence test rejects values as large or larger than the smallest effect size of interest, so the possibility that there is a small non-zero effect cannot be rejected.
- (D) We conclude there is no effect when the equivalence test is non-significant, not when the equivalence test is significant.

**Q3:** Researchers are interested in showing that students who use an online textbook perform just as well as students who use a paper textbook. If so, they can recommend teachers to allow students to choose their preferred medium, but if there is a benefit, they would recommend the medium that leads to better student performance. They randomly assign students to use an online textbook or a paper textbook, and compare their grades on the exam for the course (from the worst possible grade, 1, to the best possible grade, 10). They find that the both groups of students perform similarly, with for the paper textbook condition  $m = 7.35$ ,  $sd = 1.15$ ,  $n = 50$ , and the online textbook  $m = 7.13$ ,  $sd = 1.21$ ,  $n = 50$ ). Let's assume we consider any effect as large or larger than half a grade point (0.5) worthwhile, but any difference smaller than 0.5 too small to matter, and the alpha level is set at 0.05. What would the authors conclude? Copy the code below into R, replacing all zeroes with the correct numbers. Type `?tsum_TOST` for help with the function.

```

result <- TOSTER::tsum_TOST(
  m1 = 0.00,
  sd1 = 0.00,
  n1 = 0,
  m2 = 0.00,
  sd2 = 0.00,
  n2 = 0,
  low_eqbound = -0.0,
  high_eqbound = 0.0,
  eqbound_type = "raw",
  alpha = 0.05
)

# print the result
result

# plot the result
plot(result, type = "tnull", estimates = "raw")

```

- (A) We can **reject** an effect size of zero, and we can **reject** the presence of effects as large or larger than the smallest effect size of interest.

- (B) We can **not reject** an effect size of zero, and we can **reject** the presence of effects as large or larger than the smallest effect size of interest.
- (C) We can **reject** an effect size of zero, and we can **not reject** the presence of effects as large or larger than the smallest effect size of interest.
- (D) We can **not reject** an effect size of zero, and we can **not reject** the presence of effects as large or larger than the smallest effect size of interest.

**Q4:** If we increase the sample size in question Q3 to 150 participants in each condition, and assuming the observed means and standard deviations would be exactly the same, which conclusion would we draw?

- (A) We can **reject** an effect size of zero, and we can **reject** the presence of effects as large or larger than the smallest effect size of interest.
- (B) We can **not reject** an effect size of zero, and we can **reject** the presence of effects as large or larger than the smallest effect size of interest.
- (C) We can **reject** an effect size of zero, and we can **not reject** the presence of effects as large or larger than the smallest effect size of interest.
- (D) We can **not reject** an effect size of zero, and we can **not reject** the presence of effects as large or larger than the smallest effect size of interest.

**Q5:** If we increase the sample size in question Q3 to 500 participants in each condition, and assuming the observed means and standard deviations would be exactly the same, which conclusion would we draw?

- (A) We can **reject** an effect size of zero, and we can **reject** the presence of effects as large or larger than the smallest effect size of interest.
- (B) We can **not reject** an effect size of zero, and we can **reject** the presence of effects as large or larger than the smallest effect size of interest.
- (C) We can **reject** an effect size of zero, and we can **not reject** the presence of effects as large or larger than the smallest effect size of interest.
- (D) We can **not reject** an effect size of zero, and we can **not reject** the presence of effects as large or larger than the smallest effect size of interest.

Sometimes the result of a test is **inconclusive**, as both the null hypothesis test, and the equivalence test, of not statistically significant. The only solution in such a case is to collect additional data. Sometimes both the null hypothesis test and the equivalence test are statistically significant, in which case the effect is **statistically different from zero, but practically insignificant** (based on the justification for the SESOI).

**Q6:** We might wonder what the statistical power was for the test in Q3, assuming there was no true difference between the two groups (so a true effect size of 0). Using the new and improved `power_t_TOST` function in the TOSTER R package, we can compute the power using a sensitivity power analysis (i.e., entering the sample size per group of 50, the assumed true effect size of 0, the equivalence bounds, and the alpha level. Note that because the equivalence bounds were specified on a raw scale in Q3, we will also need to specify an estimate for the true standard deviation in the population. Let's assume this true standard deviation is 1.2. Round the answer to two digits after the decimal. Type `?power_t_TOST` for help with the function. What was the power in Q3?

```
TOSTER::power_t_TOST(  
  n = 50,  
  delta = 0.0,  
  sd = 1.2,  
  low_eqbound = -0.0,  
  high_eqbound = 0.0,  
  alpha = 0.05,  
  type = "two.sample"  
)
```

- (A) 0.00
- (B) 0.05
- (C) 0.33
- (D) 0.40

**Q7:** Assume we would only have had 15 participants in each group in Q3, instead of 50. What would be the statistical power of the test with this smaller sample size (keeping all other settings as in Q6)? Round the answer to 2 digits.

- (A) 0.00
- (B) 0.05

- (C) 0.33
- (D) 0.40

**Q8:** You might remember from discussions on statistical power for a null hypothesis significance test that the statistical power is never smaller than 5% (if the true effect size is 0, power is formally undefined, but we will observe at least 5% Type 1 errors, and the power increases when introducing a true effect). In a two-sided equivalence tests, power can be lower than the alpha level. Why?

- (A) Because in an equivalence test the Type 1 error rate is not bounded at 5%.
- (B) Because in an equivalence test the null hypothesis and alternative hypothesis are reversed, and therefore the Type 2 error rate does not have a lower bound (just as the Type 1 error rate in NHST has no lower bound).
- (C) Because the confidence interval needs to fall between the lower and upper bound of the equivalence interval, and with small sample sizes, this probability can be close to zero (because the confidence interval is very wide).
- (D) Because the equivalence test is based on a confidence interval, and not on a *p*-value, and therefore power is not limited by the alpha level.

**Q9:** A well designed study has high power to detect an effect of interest, but also to reject the smallest effect size of interest. Perform an a-priori power analysis for the situation described in Q3. Which sample size in **each group** needs to be collected to achieve a desired statistical power of 90% (or 0.9), assuming the true effect size is 0, and we still assume the true standard deviation is 1.2? Use the code below, and round up the sample size (as we cannot collect a partial observation).

```
TOSTER::power_t_TOST(
  power = 0.00,
  delta = 0.0,
  sd = 0.0,
  low_eqbound = -0.0,
  high_eqbound = 0.0,
  alpha = 0.05,
  type = "two.sample"
)
```

- (A) 100

- (B) 126
- (C) 200
- (D) 252

**Q10:** Assume that when performing the power analysis for Q9 we did not expect the true effect size to be 0, but we actually expected a mean difference of 0.1 grade point. Which sample size in **each group** would we need to collect for the equivalence test, now that we expect a true effect size of 0.1? Change the variable `delta` in `power_t_TOST` to answer this question.

- (A) 117
- (B) 157
- (C) 314
- (D) 3118

**Q11:** Change the equivalence range to -0.1 and 0.1 for Q9 (and set the expected effect size of `delta` to 0). To be able to reject effects outside such a very narrow equivalence range, you'll need a large sample size. With an alpha of 0.05, and a desired power of 0.9 (or 90%), how many participants would you need in **each group**?

- (A) 117
- (B) 157
- (C) 314
- (D) 3118

You can see it takes a very large sample size to have high power to reliably reject very small effects. This should not be surprising. After all, it also requires a very large sample size to *detect* small effects! This is why we typically leave it to a future meta-analysis to detect, or reject, the presence of small effects.

**Q12:** You can do equivalence tests for all tests. The TOSTER package has functions for *t*-tests, correlations, differences between proportions, and meta-analyses. If the test you want to perform is not included in any software, remember that you can just use a 90% confidence interval, and test whether you can reject the smallest effect size of interest. Let's perform an

equivalence test for a meta-analysis. Hyde, Lindberg, Linn, Ellis, and Williams (2008) report that effect sizes for gender differences in mathematics tests across the 7 million students in the US represent trivial differences, where a trivial difference is specified as an effect size smaller than  $d = 0.1$ . The table with Cohen's d and se is reproduced below:

Grades	d + se
Grade 2	0.06 +/- 0.003
Grade 3	0.04 +/- 0.002
Grade 4	-0.01 +/- 0.002
Grade 5	-0.01 +/- 0.002
Grade 6	-0.01 +/- 0.002
Grade 7	-0.02 +/- 0.002
Grade 8	-0.02 +/- 0.002
Grade 9	-0.01 +/- 0.003
Grade 10	0.04 +/- 0.003
Grade 11	0.06 +/- 0.003

For grade 2, when we perform an equivalence test with boundaries of  $d = -0.1$  and  $d = 0.1$ , using an alpha of 0.01, which conclusion can we draw? Use the TOSTER function TOSTmeta, and enter the alpha, effect size (ES), standard error (se), and equivalence bounds.

```
TOSTER::TOSTmeta(
  ES = 0.00,
  se = 0.000,
  low_eqbound_d = -0.0,
  high_eqbound_d = 0.0,
  alpha = 0.05
)
```

- (A) We can **reject** an effect size of zero, and we can **reject** the presence of effects as large or larger than the smallest effect size of interest.
- (B) We can **not reject** an effect size of zero, and we can **reject** the presence of effects as large or larger than the smallest effect size of interest.
- (C) We can **reject** an effect size of zero, and we can **not reject** the presence of effects as large or larger than the smallest effect size of interest.
- (D) We can **not reject** an effect size of zero, and we can **not reject** the presence of effects as large or larger than the smallest effect size of interest.

### 9.13.2 Questions about the small telescopes approach

**Q13:** What is the smallest effect size of interest based on the small telescopes approach, when the original study collected 20 participants in each condition of an independent  $t$ -test, with an **alpha level of 0.05**. Note that for this answer, it happens to depend on whether you enter the power as 0.33 or 1/3 (or 0.333). You can use the code below, which relies on the **pwr** package.

```
pwr::pwr.t.test(  
  n = 0,  
  sig.level = 0.00,  
  power = 0,  
  type = "two.sample",  
  alternative = "two.sided"  
)
```

- (A)  $d = 0.25$  (setting power to 0.33) or 0.26 (setting power to 1/3)
- (B)  $d = 0.33$  (setting power to 0.33) or 0.34 (setting power to 1/3)
- (C)  $d = 0.49$  (setting power to 0.33) or 0.50 (setting power to 1/3)
- (D)  $d = 0.71$  (setting power to 0.33) or 0.72 (setting power to 1/3)

**Q14:** Let's assume you are trying to replicate a previous result based on a correlation in a two-sided test. The study had 150 participants. Calculate the SESOI using a small telescopes justification for a replication of this study that will use an alpha level of 0.05. Note that for this answer, it happens to depend on whether you enter the power as 0.33 or 1/3 (or 0.333). You can use the code below.

```
pwr::pwr.r.test(  
  n = 0,  
  sig.level = 0,  
  power = 0,  
  alternative = "two.sided")
```

- (A)  $r = 0.124$  (setting power to 0.33) or 0.125 (setting power to 1/3)
- (B)  $r = 0.224$  (setting power to 0.33) or 0.225 (setting power to 1/3)
- (C)  $r = 0.226$  (setting power to 0.33) or 0.227 (setting power to 1/3)

- (D)  $r = 0.402$  (setting power to 0.33) or  $0.403$  (setting power to 1/3)

**Q15:** In the age of big data researchers often have access to large databases, and can run correlations on samples of thousands of observations. Let's assume the original study in the previous question did not have 150 observations, but 15000 observations. We still use an alpha level of 0.05. Note that for this answer, it happens to depend on whether you enter the power as 0.33 or 1/3 (or 0.333). What is the SESOI based on the small telescopes approach?

- (A)  $r = 0.0124$  (setting power to 0.33) or  $0.0125$  (setting power to 1/3)
- (B)  $r = 0.0224$  (setting power to 0.33) or  $0.0225$  (setting power to 1/3)
- (C)  $r = 0.0226$  (setting power to 0.33) or  $0.0227$  (setting power to 1/3)
- (D)  $r = 0.0402$  (setting power to 0.33) or  $0.0403$  (setting power to 1/3)

Is this effect likely to be practically or theoretically significant? Probably not. This would be a situation where the small telescopes approach is not a very useful procedure to determine a smallest effect size of interest.

**Q16:** Using the small telescopes approach, you set the SESOI in a replication study to  $d = 0.35$ , and set the alpha level to 0.05. After collecting the data in a well-powered replication study that was as close to the original study as practically possible, you find no significant effect, and you can reject effects as large or larger than  $d = 0.35$ . What is the correct interpretation of this result?

- (A) There is no effect.
- (B) We can statistically reject (using an alpha of 0.05) effects anyone would find theoretically meaningful.
- (C) We can statistically reject (using an alpha of 0.05) effects anyone would find practically relevant.
- (D) We can statistically reject (using an alpha of 0.05) effects the original study had 33% power to detect.

### **9.13.3 Questions about specifying the SESOI as the Minimal Statistically Detectable Effect**

**Q17:** Open the online Shiny app that can be used to compute the minimal statistically detectable effect for two independent groups: [https://shiny.ieis.tue.nl/d\\_p\\_power/](https://shiny.ieis.tue.nl/d_p_power/). Three sliders influence what the figure looks like: The sample size per condition, the true effect size, and the alpha level. Which statement is true?

- (A) The critical  $d$ -value is influenced by the sample size per group, the true effect size, but **not** by the alpha level.
- (B) The critical  $d$ -value is influenced by the sample size per group, the alpha level, but **not** by the true effect size.
- (C) The critical  $d$ -value is influenced by the alpha level, the true effect size, but **not** by the sample size per group.
- (D) The critical  $d$ -value is influenced by the sample size per group, the alpha level, and by the true effect size.

**Q18:** Imagine researchers performed a study with 18 participants in each condition, and performed a  $t$ -test using an alpha level of 0.01. Using the Shiny app, what is the smallest effect size that could have been statistically significant in this study?

- (A)  $d = 0.47$
- (B)  $d = 0.56$
- (C)  $d = 0.91$
- (D)  $d = 1$

**Q19:** You expect the true effect size in your next study to be  $d = 0.5$ , and you plan to use an alpha level of 0.05. You collect 30 participants in each group for an independent  $t$ -test. Which statement is true?

- (A) You have low power for all possible effect sizes.
- (B) Observed effect sizes of  $d = 0.5$  will never be statistically significant.
- (C) Observed effect sizes of  $d = 0.5$  will never be statistically significant.

- (D) Observed effect sizes of  $d = 0.5$  will be statistically significant.

The example we have used so far was based on performing an independent  $t$ -test, but the idea can be generalized. A shiny app for an  $F$ -test is available here: [https://shiny.ieis.tue.nl/f\\_p\\_power/](https://shiny.ieis.tue.nl/f_p_power/). The effect size associated to the power of an  $F$ -test is partial eta squared ( $\eta_p^2$ ), which for a One-Way ANOVA (visualized in the Shiny app) equals eta-squared.

The distribution for eta-squared looks slightly different from the distribution of Cohen's  $d$ , primarily because an  $F$ -test is a one-directional test (and because of this, eta-squared values are all positive, while Cohen's  $d$  can be positive or negative). The light grey line plots the expected distribution of eta-squared when the null is true, with the red area under the curve indicating Type 1 errors, and the black line plots the expected distribution of eta-squared when the true effect size is  $\eta_p^2 = 0.059$ . The blue area indicates the expected effect sizes smaller than the critical value of 0.04, which will not be statistically significant, and thus will be Type 2 errors.

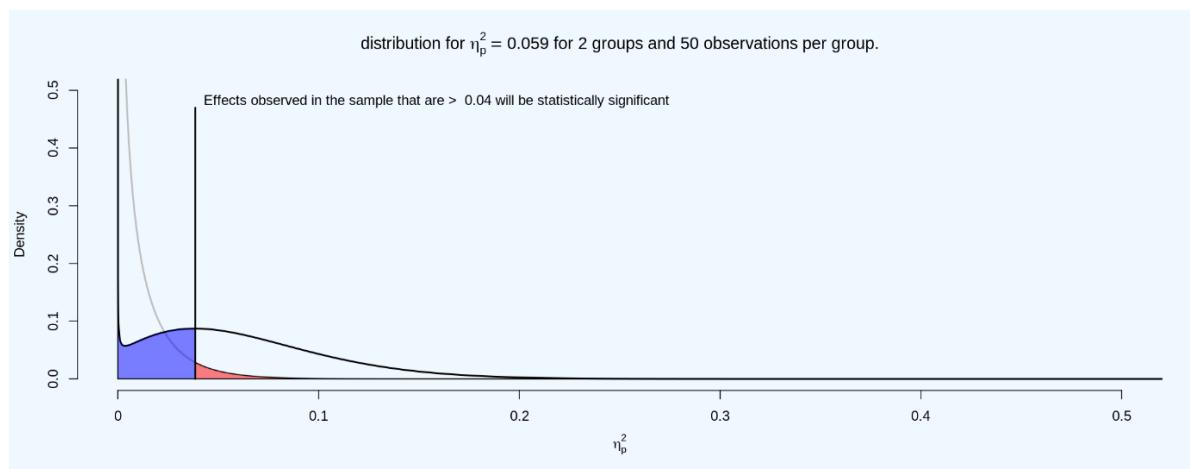


Figure 9.10: Illustration of the critical  $F$ -value for two groups, 50 observations per group, and an alpha level of 0.05.

**Q20:** Set the number of participants (per condition) to 14, and the number of groups to 3. Using the Shiny app at [https://shiny.ieis.tue.nl/f\\_p\\_power/](https://shiny.ieis.tue.nl/f_p_power/) which effect sizes (expressed in partial eta-squared, as indicated on the vertical axis) can be statistically significant with  $n = 14$  per group, and 3 groups?

- (A) Only effects larger than 0.11
- (B) Only effects larger than 0.13
- (C) Only effects larger than 0.14

- (D) Only effects larger than 0.16

Every sample size and alpha level implies a minimal statistically detectable effect that can be statistically significant in your study. Looking at which observed effects you can detect is a useful way to make sure you could actually detect the smallest effect size you are interested in.

**Q21:** Using the minimal statistically detectable effect, you set the SESOI in a replication study to  $d = 0.35$ , and set the alpha level to 0.05. After collecting the data in a well-powered replication study that was as close to the original study as practically possible, you find no significant effect, and you can reject effects as large or larger than  $d = 0.35$ . What is the correct interpretation of this result?

- (A) There is no effect.
- (B) We can statistically reject (using an alpha of 0.05) effects anyone would find theoretically meaningful.
- (C) We can statistically reject (using an alpha of 0.05) effects anyone would find practically relevant.
- (D) We can statistically reject (using an alpha of 0.05) effects that could have been statistically significant in the original study.

#### 9.13.4 Open Questions

1. What is meant with the statement ‘Absence of evidence is not evidence of absence’?
2. What is the goal of an equivalence test?
3. What is the difference between a nil null hypothesis and a non-nil null hypothesis?
4. What is a minimal effect test?
5. What conclusion can we draw if a null-hypothesis significance test and equivalence test are performed for the same data, and neither test is statistically significant?
6. When designing equivalence tests to have a desired statistical power, why do you need a larger sample size, the narrower the equivalence range is?
7. While for a null hypothesis significance test one always has some probability to observe a statistically significant result, it is possible to perform an equivalence test with 0% power. When would this happen?

8. Why is it incorrect to say there is ‘no effect’ when the equivalence test is statistically significant?
9. Specify one way in which the Bayesian ROPE procedure and an equivalence test are similar, and specify one way in which they are different.
10. What is the anchor based method to specify a smallest effect size of interest?
11. What is a cost-benefit approach to specify a smallest effect size of interest?
12. How can researchers use theoretical predictions to specify a smallest effect size of interest?
13. What is the idea behind the ‘small telescopes’ approach to equivalence testing?

# 10 Sequential Analysis

Repeatedly analyzing incoming data while data collection is in progress has many advantages. Researchers can stop the data collection at an interim analysis when they can reject the null hypothesis or the smallest effect size of interest, even if they would be willing to collect more data if needed, or if the results show there is an unexpected problem with the study (e.g., participants misunderstand the instructions or questions). One could easily argue that psychological researchers have an ethical obligation to repeatedly analyze accumulating data, given that continuing data collection whenever the desired level of confidence is reached, or whenever it is sufficiently clear that the expected effects are not present, is a waste of the time of participants and the money provided by taxpayers. In addition to this ethical argument, designing studies that make use of sequential analyses can be more efficient than when data is only analyzed a single time, when the maximum sample size a researcher is willing to collect has been reached.

Sequential analyses should not be confused with **optional stopping**, which was discussed in the chapter on error control. In optional stopping, researchers use an unadjusted alpha level (e.g., 5%) to repeatedly analyze the data as it comes in, which can substantially inflate the Type 1 error rate. The critical difference with **sequential analysis** is that the Type 1 error rate is controlled. By lowering the alpha level at each interim analysis, the overall Type I error rate can be controlled, much like a Bonferroni correction is used to prevent inflation of the Type 1 error rate for multiple comparisons. Indeed, the Bonferroni correction is a valid (but conservative) approach to control the error rate in sequential analyses (Wassmer & Brannath, 2016).

In sequential analysis a researcher designs a study such that they are able to perform **interim analyses**, say when 25%, 50%, and 75% of the data is collected. At each interim analysis a test is performed at a corrected alpha level, so that over all planned analyses the desired Type 1 error rate is maintained. Sequential analyses are commonly used in medical trials, where quickly discovering an effective treatment can be a matter of life and death. If at an interim analysis, researchers decide that a new drug is effective, in turn they may well want to terminate the trial and give the working drug to patients in the control condition to improve their health, or even save their lives. For example, the safety and efficacy of the Pfizer–BioNTech COVID-19 vaccine used an experimental design where they planned to analyze the data 5 times, and controlled the overall Type 1 error rate by lowering the alpha level for each **interim analysis**.

**Table 6. Interim Analysis Plan and Boundaries for Efficacy and Futility**

Analysis	Number of Cases	Success Criteria <sup>a</sup>	Futility Boundary
		VE Point Estimate (Case Split)	VE Point Estimate (Case Split)
IA1	32	76.9% (6:26)	11.8% (15:17)
IA2	62	68.1% (15:47)	27.8% (26:36)
IA3	92	62.7% (25:67)	38.6% (35:57)
IA4	120	58.8% (35:85)	N/A
Final	164	52.3% (53:111)	

Abbreviations: IA = interim analysis; N/A = not applicable; VE = vaccine efficacy.

Note: Case split = vaccine : placebo.

a. Interim efficacy claim:  $P(VE > 30\% | \text{data}) > 0.995$ ; success at the final analysis:  $P(VE > 30\% | \text{data}) > 0.986$ .

Figure 10.1: Screenshot of the planned interim analyses examining the safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine.

The use of sequential analyses is only slowly becoming more popular in many scientific disciplines, but sequential analysis techniques have a long history. As early as 1929, Dodge and Romig realized that analyzing the data sequentially was more efficient than doing so once (Dodge & Romig, 1929). Wald (1945), who popularized the idea of sequential tests of hypotheses in 1945, performed his work during the second world war. He was only allowed to publish his findings after the war had ended, as he explains in a historical note:

Because of the substantial savings in the expected number of observations effected by the sequential probability ratio test, and because of the simplicity of this test procedure in practical applications, the National Defense Research Committee considered these developments sufficiently useful for the war effort to make it desirable to keep the results out of the reach of the enemy, at least for a certain period of time. The author was, therefore, requested to submit his findings in a restricted report which was dated September, 1943.

Sequential analyses are well-established procedures, and have been developed in great detail over the last decades (Jennison & Turnbull, 2000; Proschan et al., 2006; Wassmer & Brannath, 2016). Here, we will explain the basics of how to control error rates in group sequential analyses, perform a-priori power analysis and compare when sequential designs will be more or less efficient than fixed designs. Before we discuss these topics, it is useful to clarify some terminology. A **look** (also called **stage**) means analyzing all the data collected up to a specific point; that is, you look after 50, 100, and 150 observations, and analyze all the data that has been collected up to that point. After 50 and 100 observations we perform an **interim analysis**, and after 150 observations we perform the **final analysis**, after which we always stop. Not all looks have to occur in practice. If the analysis reveals a statistically significant result at look 1, data collection can be terminated. We can stop because we reject  $H_0$  (e.g., in a null hypothesis significance test), or because we reject  $H_1$  (e.g., in an equivalence test). We

can also stop for **curtailment** or for **futility**: It is either impossible, or very unlikely for the final analysis to yield  $p < \alpha$ . The **overall alpha level** in a sequential design differs from the alpha level at each look. For example, if we want an overall Type I error rate of 5% for a two-sided test with 3 looks, the alpha level for each look could be 0.0221 (if we decide to use the correction for the alpha level proposed by Pocock (1977)). In this chapter we will focus on group sequential designs, where data is collected in multiple groups, but other sequential approaches exist, as explained in the chapter on sample size justification.

## 10.1 Choosing alpha levels for sequential analyses.

If one would analyze the data at multiple looks without correcting the alpha level, the Type 1 error rate would inflate (Armitage et al., 1969). As Armitage and colleagues show, with equally spaced looks, the alpha level inflates to 0.142 after 5 looks, 0.374 after 100 looks, and 0.530 after 1000 looks. Looking at the data twice is conceptually similar to deciding if a result is significant if one of two dependent variables shows a statistically significant effect. However, an important difference is that in the case of sequential analyses the multiple tests are not independent, but dependent. A test at look 2 combines the old data collected at look 1 with the new data at look 2. This means the Type 1 error rate inflates less quickly compared to independent tests, and we will see below this enables more efficient and flexible solutions to controlling error rates.

When controlling the Type 1 error rate in sequential analyses, a decision needs to be made about how to spend the alpha level across all looks at the data. For example, when a researcher plans a study with one interim look and one final look at the data, boundary critical Z-values need to be set for the first look (at  $n$  out of  $N$  observations) and the second look (at  $N$  observations). These two critical values,  $c_1$  and  $c_2$  (for the first and the second analysis) need to be chosen such that the overall probability ( $\Pr$ ) that the null hypothesis is rejected – when in the first analysis the observed Z-score is larger than the critical value for the first look,  $Z_n > c_1$ , and (if we did not reject the hypothesis in the first analysis, so  $Z_n < c_1$ , and we continue data collection) when in the second analysis the observed Z-score is larger than the critical value for the second look,  $Z_N > c_2$  – equals the desired overall alpha level when the null hypothesis is true. In formal terms, for a directional test:

$$\Pr\{Z_n \geq c_1\} + \Pr\{Z_n < c_1, Z_N \geq c_2\} = \alpha$$

With more than one interim analysis, additional critical values have to be determined following the same rationale. If you combine multiple looks at the data with multiple comparisons, you would correct the alpha level twice, once for multiple comparisons, and then for multiple looks. Because the alpha level is corrected, it does not matter which statistical test you perform at each look, all that matters is that the  $p$ -value is compared to the corrected alpha level. The

corrections discussed below are valid for any design where the data is normally distributed, and where each group of observations is independent of the previous group.

## 10.2 The Pocock correction

The first decision researchers need to make is how they want to correct the Type I error rate across looks. Four common approaches are the Pocock correction, the O'Brien-Fleming correction, the Haybittle & Peto correction, and the Wang and Tsiatis approach. Users are also free to specify their own preferred way to spend the alpha level across looks.

The Pocock correction is the simplest way to correct the alpha level for multiple looks. Conceptually, it is very similar to the Bonferroni correction. The Pocock correction has been created such that the alpha level is identical for each look at the data, resulting in constant critical values (expressed as  $z$  values)  $u_k = c$  to reject the null hypothesis,  $H_0$ , at look  $k$ . The following code uses the package `rpact` to design a study for a sequential analysis:

```
library(rpact)
design <- getDesignGroupSequential(
  kMax = 2,
  typeOfDesign = "P",
  sided = 2,
  alpha = 0.05,
  beta = 0.1
)
print(summary(design))

## Sequential analysis with a maximum of 2 looks (group sequential design)
```

Pocock design, two-sided overall significance level 5%, power 90%,  
undefined endpoint, inflation factor 1.1001, ASN H1 0.7759, ASN H01 1.0094,  
ASN HO 1.0839.

Stage	1	2
Information rate	50%	100%
Efficacy boundary (z-value scale)	2.178	2.178
Stage levels (one-sided)	0.0147	0.0147
Cumulative alpha spent	0.0294	0.0500
Overall power	0.5893	0.9000

The output tells us we have designed a study with 2 looks (one interim, one final) using the Pocock spending function. The last line returns one-sided alpha levels. The `rpact` package focuses on Confirmatory Adaptive Clinical Trial Design and Analysis. In clinical trials, researchers mostly test directional predictions, and thus, the default setting is to perform a one-sided test. In clinical trials it is common to use a 0.025 significance level for one-sided tests, but in many other fields, 0.05 is a more common default. We can get the two-sided alpha levels by multiplying the one-sided alpha levels by two:

```
design$stageLevels * 2
```

```
[1] 0.02938579 0.02938579
```

We can check the output against the [Wikipedia page for the Pocock correction](#) where we indeed see that with 2 looks at the data the alpha level for each look is 0.0294. The Pocock correction is slightly more efficient than using a Bonferroni correction (in which case the alpha levels would be 0.025), because of the dependency in the data (at the second look, the data analyzed at the first look is again part of the analysis).

`rpact` makes it easy to plot the boundaries (based on the critical values) for each look. Looks are plotted as a function of the ‘Information Rate’, which is the percentage of the total data that has been collected at a look. In Figure 10.2 there are two equally spaced looks, so when 50% of the data has been collected (Information Rate 0.5) and when 100% of the data has been collected (Information Rate 1). We see the critical values (solid black lines) are larger than the 1.96 we would use for a fixed design with a 5% alpha level, namely  $Z = 2.178$  (black dashed line). Whenever we observe a test statistic that is more extreme than these critical values at the first or second look, we can reject the null hypothesis.

The analysis can also be performed in the `rpact` [shiny app](#) which also allows users to create all plots through simple menu options, and download a complete report of the analyses (e.g., for a preregistration document).

## 10.3 Comparing Spending Functions

We can visualize the corrections for different types of designs for each of 3 looks (2 interim looks and one final look) in the same plot (see Figure 10.4). The plot below shows the Pocock, O’Brien-Fleming, Haybittle-Peto, and Wang-Tsiatis correction with  $\Delta = 0.25$ . We see that researchers can choose different approaches to spend their alpha level across looks. Researchers can choose to spend their alpha conservatively (keeping most of the alpha for the last look), or more liberally (spending more alpha at the earlier looks, which increases the probability of stopping early for many true effect sizes).

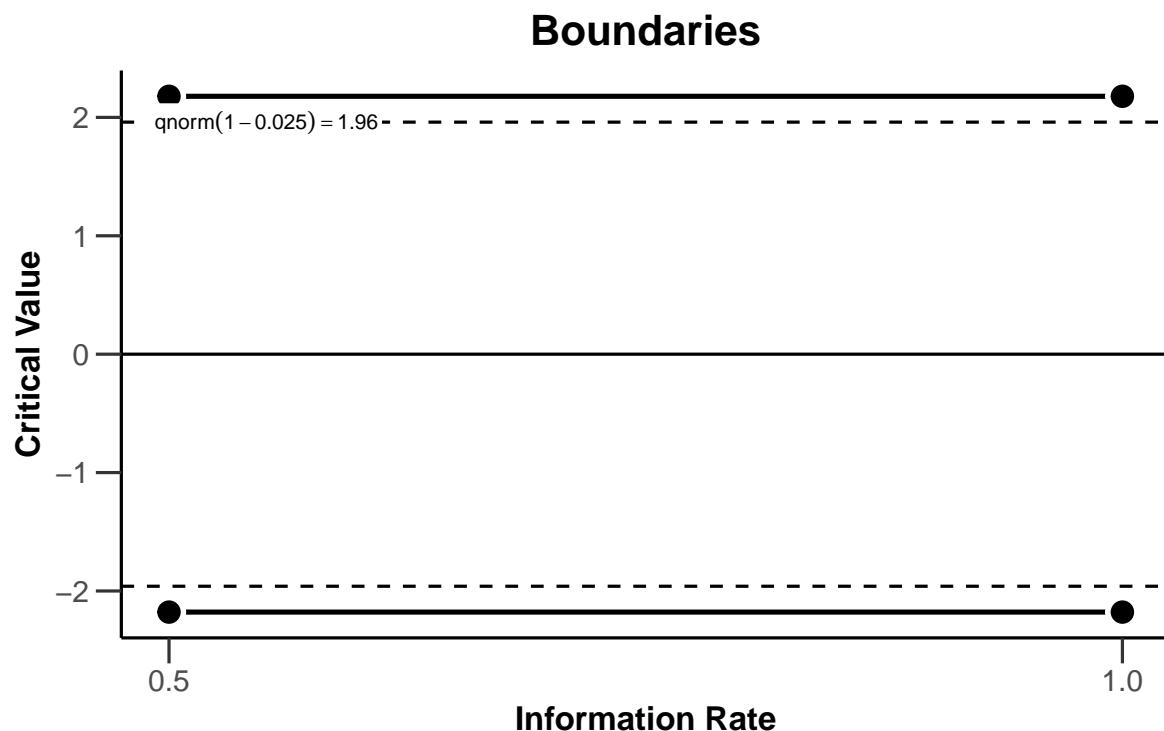


Figure 10.2: Plot of critical boundaries at each look for a 2 look design with a Pocock correction.

The screenshot shows the rPACT Shiny app interface. At the top, there is a header with the PACT logo, 'Trial Design', and a 'Help' dropdown. The main area is divided into two sections: 'Design' (selected) and 'Endpoint'. The 'Design' section contains the following parameters:

- Design:** Group Sequential (radio button selected), Inverse Normal, Fisher.
- Maximum number of stages:** A slider set to 2 (labeled 'kMax = 2') with a range from 1 to 10.
- Test:** Two-sided power checked, Two-sided selected.
- Significance level:** alpha = 0.05.
- Type II error rate:** beta = 0.2.
- Power:** Power = 0.8.
- Type of design:** Pocock (P).

A note at the bottom of the design section states: "Double click on a cell to edit it; hit Ctrl+Enter to finish editing, or Esc to cancel".

The 'R Command' section displays the following R code:

```
design <- getDesignGroupSequential(typeOfDesign = "P", informationRates = c(0.5, 1),
alpha = 0.05, twoSidedPower = TRUE, sided = 2)
summary(design)
```

The 'R Output' section shows the results of the R command:

```
Sequential analysis with a maximum of 2 looks (group sequential design)

Stage          1      2
Information rate 50%  100%
Efficacy boundary (z-value scale) 2.178 2.178
Cumulative alpha spent   0.0294 0.0500
Two-sided local significance level 0.0294 0.0294
```

Figure 10.3: Screenshot of rPACT Shiny app.

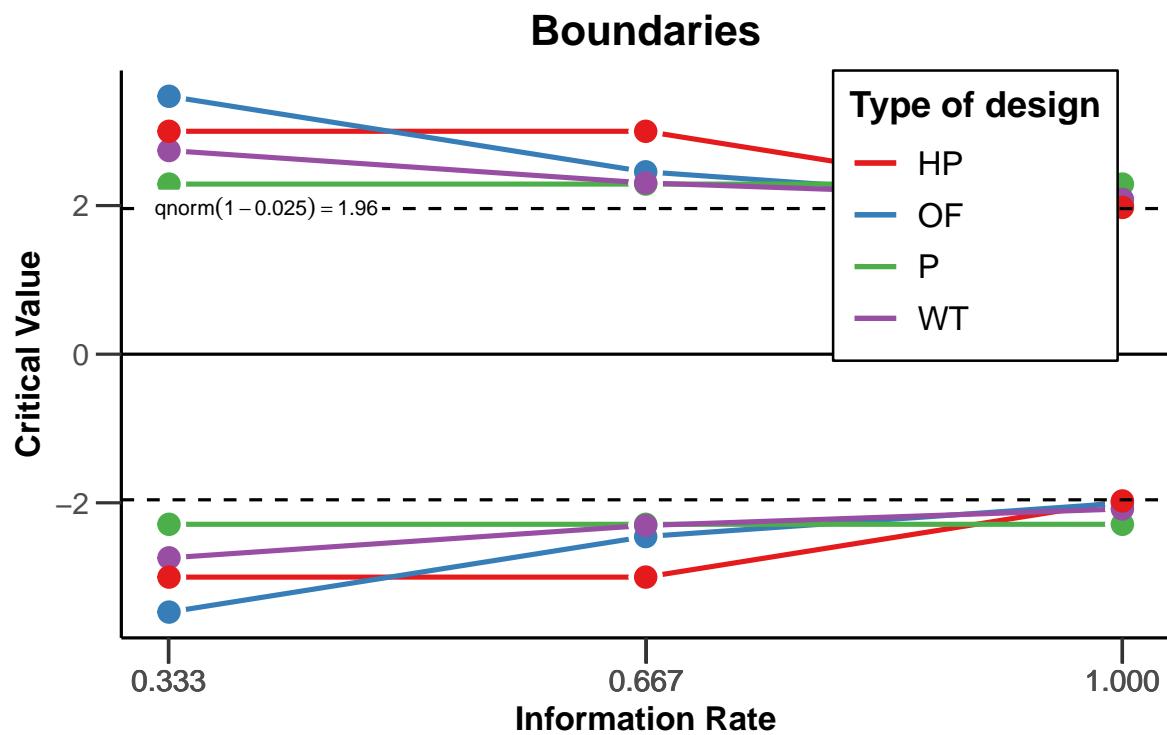


Figure 10.4: Four different spending functions for 3 looks: O'Brien-Fleming (OF), Pocock (P), Haybittle-Peto (HP), Wang-Tsiatis (WT).

We can see that the O'Brien and Fleming correction is much more conservative at the first look but is close to the uncorrected critical value of 1.96 at the last look (the black dashed line - for two-sided tests all critical values are mirrored in the negative direction): 3.471, 2.454, and 2.004. The Pocock correction has the same critical value at each look (2.289, 2.289, and 2.289). The Haybittle and Peto correction has the same critical value at each look but the last (3, 3, and 1.975). With the Wang and Tsiatis correction, critical values decrease with each look (2.741, 2.305, and 2.083).

Being conservative during early looks is sensible if you mainly want to monitor the results for unexpected developments. A Pocock correction is more useful when there is substantial uncertainty both in whether an effect is present and how large the effect size is, as it gives a higher probability of stopping the experiment early if the effects are large. Because the statistical power of a test depends on the alpha level, lowering the alpha level at the final look means that the statistical power is lower compared to a fixed design, and that to achieve a desired power, the sample size of a study needs to be increased to maintain the same statistical power at the last look. This increase in sample size can be compensated by stopping data collection early, in which case a sequential design is more efficient than a fixed design. Because the alpha at the last look for O'Brien-Fleming or Haybittle-Peto designs are very similar to the statistical power for a fixed design with only one look, the required sample size is also very similar. The Pocock correction requires a larger increase in the maximum sample size to achieve the desired power compared to a fixed design.

Corrected alpha levels can be computed to many digits, but this quickly reaches a level of precision that is meaningless in real life. The observed Type I error rate for all tests you will do in your lifetime is not noticeably different if you set the alpha level at 0.0194, 0.019, or 0.02 (see the concept of '[significant digits](#)'). Even as we calculate and use alpha thresholds up to many digits in sequential tests, the messiness of most research makes these alpha levels have [false precision](#). Keep this in mind when interpreting your data.

## 10.4 Alpha spending functions

The approaches to specify the shape of decision boundaries across looks discussed so far have an important limitation (Proschan et al., 2006). They require a pre-specified number of looks (e.g., 4), and the sample size for the interim looks need to be pre-specified as well (e.g., after 25%, 50%, 75%, and 100% of observations). It is logically not always feasible to stop the data collection exactly at 25% of the planned total sample size. An important contribution to the sequential testing literature was made by Lan and DeMets (1983) who introduced the alpha spending approach to correct the alpha level. In this approach the cumulative Type I error rate spent across the looks is pre-specified through a function (the *alpha spending function*) to control the overall significance level  $\alpha$  at the end of the study.

The main benefit of these alpha spending functions is that error rates at interim analyses can be controlled, while neither the number nor the timing of the looks needs to be specified in

advance. This makes alpha spending approaches much more flexible than earlier approaches to controlling the Type 1 error in group sequential designs. When using an alpha spending function it is important that the decision to perform an interim analysis is not based on collected data, as this can still increase the Type I error rate. As long as this assumption is met, it is possible to update the alpha levels at each look during a study.

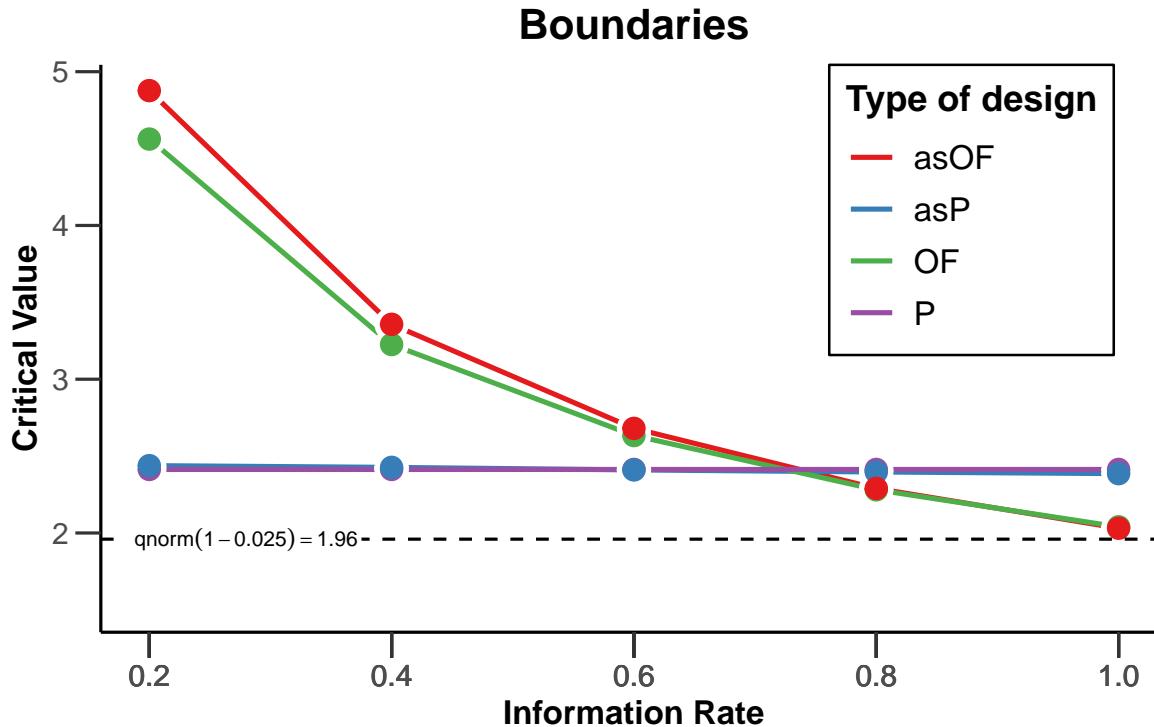


Figure 10.5: Comparison of Pocock (P) and O'Brien-Fleming correction (OF), Pocock-like (asP) and O'Brien-Fleming like (asOF) alpha spending functions, for 5 looks.

## 10.5 Updating boundaries during a study

Although alpha spending functions control the Type I error rate even when there are deviations from the pre-planned number of looks, or their timing, this does require recalculating the boundaries used in the statistical test based on the amount of information that has been observed. Let us assume a researcher designs a study with three equally spaced looks at the data (two interim looks, one final look), using a Pocock-type alpha spending function, where results will be analyzed in a two-sided  $t$ -test with an overall desired Type I error rate of 0.05, and a desired power of 0.9 for a Cohen's  $d$  of 0.5. An a-priori power analysis (which we will explain in the next section) shows that we achieve the desired power in our sequential

design if we plan to look after 65.4, 130.9, and 196.3 observations in each condition. Since we cannot collect partial participants, we should round these numbers up, and because we have 2 independent groups, we will collect 66 observations for look 1 (33 in each condition), 132 at the second look (66 in each condition) and 198 at the third look (99 in each condition). The code below computes the alpha levels at each look (or stage) for a two-sided test:

```
design <- getDesignGroupSequential(kMax = 3,
                                    typeOfDesign = "asP",
                                    sided = 2,
                                    alpha = 0.05,
                                    beta = 0.1)
design$stageLevels * 2
```

```
[1] 0.02264162 0.02173822 0.02167941
```

Now imagine that due to logistical issues, we do not manage to analyze the data until we have collected data from 76 observations (38 in each condition) instead of the planned 66 observations. Such logistical issues are common in practice and one of the main reasons alpha spending functions for group sequential designs were developed. Our first look at the data does not occur at the planned time of collecting 33.3% of the total sample, but at  $76/198 = 38.4\%$  of the planned sample. We can recalculate the alpha level we should use for each look at the data, based on the current look, and planned future looks. Instead of using the alpha levels 0.0226, 0.0217, and 0.0217 at the three respective looks (as calculated above, and note how in the Pocock-like alpha spending function the alpha levels are almost, but not exactly, the same at each look, unlike the Pocock correction where they are identical at each look). We can adjust the information rates by explicitly specifying them using `informationRates` in the code below. The first look now occurs at 76/198 of the planned sample. The second look is still planned to occur at 2/3 of the sample, and the final look at the planned maximum sample size.

```
design <- getDesignGroupSequential(
  typeOfDesign = "asP",
  informationRates = c(76/198, 2/3, 1),
  alpha = 0.05,
  sided = 2)
design$stageLevels * 2
```

```
[1] 0.02532710 0.02043978 0.02164755
```

The updated alpha levels are 0.0253 for the current look, 0.0204 for the second look, and 0.0216 for the final look. The alpha level we will use for the first look is therefore not 0.0226

(as originally planned) but the slightly higher 0.0253. The second look will now use a slightly lower alpha of 0.0204 instead of 0.0217. The differences are small, but the fact that there is a formal method to control the alpha level that provides the flexibility to look at different times than originally planned is extremely useful.

It is also possible to correct the alpha level if the final look at the data changes, for example because you are not able to collect the intended sample size, or because due to unforeseen circumstances you collect more data than planned. This is nowadays increasingly common as people preregister their studies, or publish using Registered Reports. Sometimes they end up with slightly more data than planned, which raises the question is they should analyze the planned sample size, or all the data. Analyzing all the collected data prevents wasting responses from participants, and uses all the information available, but it increases the flexibility in the data analysis (as researchers can now choose to analyze either the data from the planned sample, or all the data they have collected). Alpha spending functions solve this conundrum by allowing researchers to analyze all the data, while updating the alpha level that is used to control the overall alpha level.

If more data is collected than was planned, we can no longer use the alpha spending function that was chosen (i.e., the Pocock spending function), and instead have to provide a **user-defined alpha spending function** by updating the timing and alpha spending function to reflect the data collection as it actually occurred up to the final look. Assuming the second look in our earlier example occurred as originally planned at 2/3 of the data we planned to collect, but the last look occurred at 206 participants instead of 198, we can compute an updated alpha level for the last look. Given the current total sample size, we need to recompute the alpha levels for the earlier looks, which now occurred at  $76/206 = 0.369$ ,  $132/206 = 0.641$ , and for the last look at  $206/206 = 1$ .

The first and second look occurred with the adjusted alpha levels we computed after the first adjustment (alpha levels of 0.0253 and 0.0204). We have already spent part of our total alpha at the first two looks. We can look at the ‘Cumulative alpha spent’ in the results from the design we specified above, and see how much of our Type I error rate we spent so far:

```
design$alphaSpent
```

```
[1] 0.02532710 0.03816913 0.05000000
```

We see that we have spent 0.0253 after look 1, and 0.0382 after look 2. We also know we want to spend the remainder of our Type I error rate at the last look, for a total of 0.05.

Our actual alpha spending function is no longer captured by the Pocock spending function after collecting more data than planned, so instead we specify a user defined spending function. We can perform these calculations using the code below by specifying the **userAlphaSpending** information, after choosing the **asUser** design:

```

design <- getDesignGroupSequential(
  typeOfDesign = "asUser",
  informationRates = c(72/206, 132/206, 1),
  alpha = 0.05,
  sided = 2,
  userAlphaSpending = c(0.0253, 0.0382, 0.05)
)
design$stageLevels * 2

```

```
[1] 0.02530000 0.01987072 0.02075796
```

The alpha levels for looks in the past do not correspond with the alpha levels we used, but the final alpha level (0.0208) gives the alpha level we should use for our final analysis based on a sample size that is larger than what we planned to collect. The difference with the alpha level we would have used if we collected the planned sample size is really small (0.0216 vs. 0.0208), in part because we did not miss the planned sample size by a lot. Such small differences in alpha levels will not really be noticeable in practice, but it is very useful that there is a formally correct solution to deal with collecting more data than planned, while controlling the Type 1 error rate. If you use sequential designs, you can use these corrections whenever you overshoot the sample size you planned to collect in a preregistration.

## 10.6 Sample Size for Sequential Designs

At the final look, sequential designs require somewhat more participants than a fixed design, depending on how much the alpha level at this look is lowered due to the correction for multiple comparisons. That said, due to early stopping, sequential designs will on average require less participants. Let's first examine how many participants we would need in a fixed design, where we only analyze our data once. We have an alpha level of 0.05, and a Type 2 (beta) error of 0.1 - in other words, the desired power is 90%. We will perform one test, and assuming a normal distribution our critical Z-score would be 1.96, for an alpha level of 5%.

```

design <- getDesignGroupSequential(
  kMax = 1,
  typeOfDesign = "P",
  sided = 2,
  alpha = 0.05,
  beta = 0.1
)
power_res <- getSampleSizeMeans(

```

```

design = design,
groups = 2,
alternative = 0.5,
stDev = 1,
allocationRatioPlanned = 1,
normalApproximation = FALSE)

print(power_res)

## Design plan parameters and output for means

### Design parameters

* *Critical values*: 1.960
* *Two-sided power*: FALSE
* *Significance level*: 0.0500
* *Type II error rate*: 0.1000
* *Test*: two-sided

### User defined parameters

* *Alternatives*: 0.5

### Default parameters

* *Mean ratio*: FALSE
* *Theta H0*: 0
* *Normal approximation*: FALSE
* *Standard deviation*: 1
* *Treatment groups*: 2
* *Planned allocation ratio*: 1

### Sample size and output

* *Number of subjects fixed*: 170.1
* *Number of subjects fixed (1)*: 85
* *Number of subjects fixed (2)*: 85
* *Lower critical values (treatment effect scale)*: -0.303
* *Upper critical values (treatment effect scale)*: 0.303
* *Local one-sided significance levels*: 0.0500

### Legend

```

```
* *(i)*: values of treatment arm i
```

We see that we need 85 participants in each group, (or 86, since the sample size is actually 85.03 and the required number of observations is rounded up, and so we need 172 participants in total. Other power analysis software, such as G\*Power, should yield the same required sample size. We can now examine our design above with 2 looks and a Pocock-like alpha spending function for a 2 sided test with an alpha of 0.05. We will look 2 times, and expect a true effect of  $d = 0.5$  (which we enter by specifying an alternative of 0.5, and a stDev of 1).

```
seq_design <- getDesignGroupSequential(
  kMax = 2,
  typeOfDesign = "asP",
  sided = 2,
  alpha = 0.05,
  beta = 0.1
)

# Compute the sample size we need
power_res_seq <- getSampleSizeMeans(
  design = seq_design,
  groups = 2,
  alternative = 0.5,
  stDev = 1,
  allocationRatioPlanned = 1,
  normalApproximation = FALSE)

print(power_res_seq)

## Design plan parameters and output for means

### Design parameters

* *Information rates*: 0.500, 1.000
* *Critical values*: 2.157, 2.201
* *Futility bounds (binding)*: -Inf
* *Cumulative alpha spending*: 0.03101, 0.05000
* *Local one-sided significance levels*: 0.01550, 0.01387
* *Two-sided power*: FALSE
* *Significance level*: 0.0500
* *Type II error rate*: 0.1000
* *Test*: two-sided
```

```

### User defined parameters

* *Alternatives*: 0.5

### Default parameters

* *Mean ratio*: FALSE
* *Theta H0*: 0
* *Normal approximation*: FALSE
* *Standard deviation*: 1
* *Treatment groups*: 2
* *Planned allocation ratio*: 1

### Sample size and output

* *Maximum number of subjects*: 188.9
* *Maximum number of subjects (1)*: 94.5
* *Maximum number of subjects (2)*: 94.5
* *Number of subjects [1]*: 94.5
* *Number of subjects [2]*: 188.9
* *Reject per stage [1]*: 0.6022
* *Reject per stage [2]*: 0.2978
* *Early stop*: 0.6022
* *Expected number of subjects under H0*: 186
* *Expected number of subjects under H0/H1*: 172.7
* *Expected number of subjects under H1*: 132.1
* *Lower critical values (treatment effect scale) [1]*: -0.451
* *Lower critical values (treatment effect scale) [2]*: -0.323
* *Upper critical values (treatment effect scale) [1]*: 0.451
* *Upper critical values (treatment effect scale) [2]*: 0.323
* *Local one-sided significance levels [1]*: 0.03101
* *Local one-sided significance levels [2]*: 0.02774

### Legend

* *(i)*: values of treatment arm i
* *[k]*: values at stage k

```

The sample size per condition at the first look is 47.24, and at the second look it is 94.47, which means we are now collecting 190 instead of 172 participants. This is a consequence of lowering our alpha level at each look (from 0.05 to 0.028). To compensate for the lower alpha level, we need to increase the sample size of the study to achieve the same power.

However, the maximum sample size is not the expected sample size for this design, because of the possibility that we can stop data collection at an earlier look in the sequential design. In the long run, if  $d = 0.5$ , and we use an Pocock-like alpha spending function, and ignoring upward rounding because we can only collect a complete number of observations, we will sometimes collect 96 participants and stop after the first look, and the remaining time continue to 190 participants. As we see in the rows ‘Reject per stage’ the data collection is expected to stop after the first look in 0.6 of the studies because we have observed a significant result. The remainder of the time ( $1 - 0.6$ ) = 0.4.

This means that, assuming there is a true effect of  $d = 0.5$ , the *expected* sample size on average is the probability of stopping at each look, multiplied by the number of observations we collect at each look, so  $0.6 * 96 + 0.3 * 190 = 133.39$ . The `rpact` package returns 132.06 under “Expected number of subjects under  $H_1$ ” - the small difference is due to the fact that `rpact` does not round the number of observations up, although it should). So, assuming the true effect is  $d = 0.5$ , in any single study we might need to collect slightly more data than in a fixed design (where we would collect 172), but on average we will need to collect less observations in a sequential design.

Because power is a curve, and the true effect size is unknown, it is useful to plot power across a range of possible effect sizes, so that we can explore the expected sample size, in the long run, if we use a sequential design, for different true effect sizes.

```
# Use getPowerMeans and set max N to 190 based on analysis above
sample_res <- getPowerMeans(
  design = seq_design,
  groups = 2,
  alternative = seq(0, 1, 0.01),
  stDev = 1,
  allocationRatioPlanned = 1,
  maxNumberOfSubjects = 190,
  normalApproximation = FALSE)

plot(sample_res, type = 6)
```

## Expected Sample Size and Power / Early Stop

$N_{\max}=190$ , standard deviation=1, H0: mean difference=0

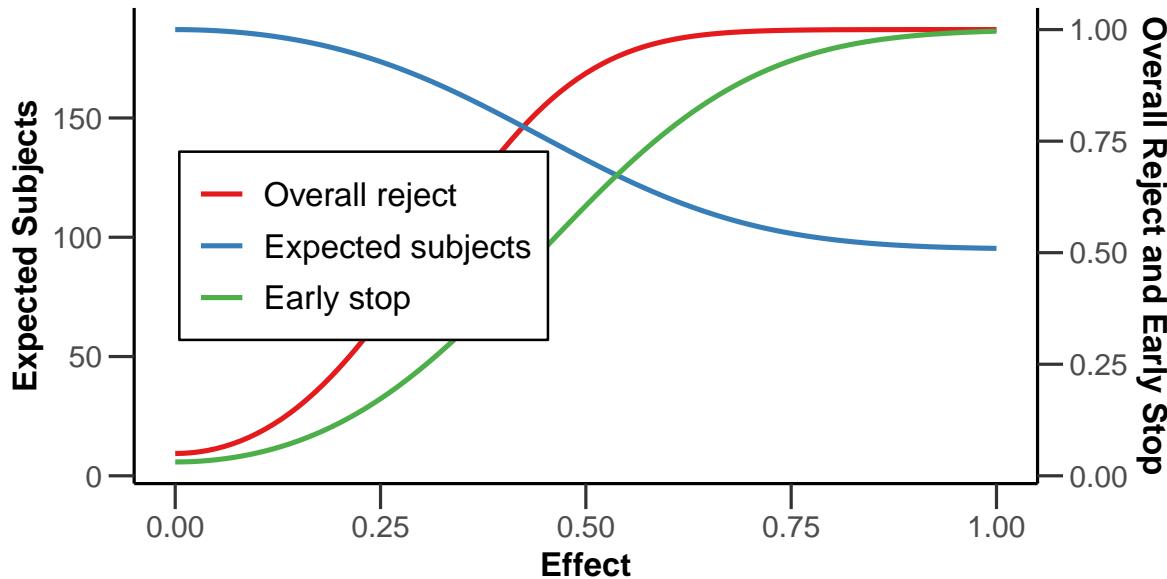


Figure 10.6: Power curve for a sequential design with 2 looks.

The blue line in Figure 10.6 indicates the expected number of observations we need to collect. Not surprisingly, when the true effect size is 0, we will almost always continue data collection to the end. We will only stop if we observe a Type 1 error, which is rare, and thus the expected number of observations is very close to the maximum sample size we are willing to collect. On the other side of the graph we see the scenario for when the true effect size is  $d = 1$ . With such a large effect size, we will have high power at our first look, and we will almost always be able to stop at the first look. The red line indicates the power at the final look, and the green line indicates the probability of stopping early.

The Pocock correction leads to a substantially lower alpha level at the last look, which requires an increase in sample size to compensate. As we saw before, the O'Brien-Fleming spending function does not require such a severe reduction in the alpha level at the last look. As the power analysis below shows, with 2 looks, this design would not need an increase in sample size at all in practice.

```
seq_design <- getDesignGroupSequential(
  kMax = 2,
  typeOfDesign = "asOF",
  sided = 2,
```

```

alpha = 0.05,
beta = 0.1
)

# Compute the sample size we need
power_res_seq <- getSampleSizeMeans(
  design = seq_design,
  groups = 2,
  alternative = 0.5,
  stDev = 1,
  allocationRatioPlanned = 1,
  normalApproximation = FALSE)

print(summary(power_res_seq))

```

## Sample size calculation for a continuous endpoint

Sequential analysis with a maximum of 2 looks (group sequential design), overall significance level 5% (two-sided).

The results were calculated for a two-sample t-test, H0: mu(1) - mu(2) = 0, H1: effect = 0.5, standard deviation = 1, power 90%.

Stage	1	2
Information rate	50%	100%
Efficacy boundary (z-value scale)	2.963	1.969
Overall power	0.2525	0.9000
Number of subjects	85.3	170.6
Expected number of subjects under H1	149.1	
Cumulative alpha spent	0.0031	0.0500
Two-sided local significance level	0.0031	0.0490
Lower efficacy boundary (t)	-0.661	-0.304
Upper efficacy boundary (t)	0.661	0.304
Exit probability for efficacy (under H0)	0.0031	
Exit probability for efficacy (under H1)	0.2525	

Legend:

\* \*(t)\*: treatment effect scale

This design meets the desired power when we collect 172 participants - exactly as many as

when we would *not* look at the data once. We basically get a free look at the data, with the expected number of participants (assuming  $d = 0.5$ ) dropping to 149.1. Increasing the number of looks to 4 comes at only a very small required increase in the number of participants to maintain the same statistical power, but further decreases the expected sample size. Especially for a conservative a-priori power analysis, or when performing an a-priori power analysis for a smallest effect size of interest, and there is a decent probability that the true effect size is larger, using sequential analysis is a very attractive option.

## 10.7 Stopping for futility

So far, the sequential designs we have discussed would only stop at an interim analysis if we can reject  $H_0$ . A well-designed study also takes into account the possibility that there is no effect, as we discussed in the chapter on [equivalence testing](#). In the sequential analysis literature, stopping to reject the presence of the smallest effect size of interest is called **stopping for futility**. In the most extreme case, it could be impossible after an interim analysis that the final analysis will yield a statistically significant result. To illustrate this in a hypothetical scenario, imagine that after collecting 182 out of 192 observations, the observed mean difference between two independent conditions is 0.1, while the study was designed with the idea that the smallest effect deemed worthwhile is a mean difference of 0.5. If the primary dependent variable is measured on a 7 point Likert scale, it might be that even if every of the remaining 5 participants in the control condition answers 1, and every of the remaining participants in the experimental condition answers 7, the effect size after 192 observations will not yield  $p < \alpha$ . If the goal of your study was to detect whether there was an effect of at least a mean difference of 0.5, at this point a researcher knows that goal will not be reached. Stopping a study at an interim analysis because the final result cannot yield a significant effect is called *non-stochastic curtailment*.

In less extreme but more common situations, it might still be possible for the study to observe a significant effect, but the probability might be very small. The probability of finding a significant result, given the data that have been observed up to an interim analysis, is called **conditional power**. Performing the conditional power analysis on the effect size that was originally expected might be too optimistic, but it is also undesirable to use the observed effect size, which typically has quite some uncertainty. One proposal is to update the expected effect size based on the observed data. If a Bayesian updating procedure is used, this is called **predictive power** (D. J. Spiegelhalter et al., 1986). It is possible to use **adaptive designs** that allow researchers to increase the final number of observations based on an interim analysis without inflating the Type 1 error rate (see Wassmer & Brannath (2016)).

Alternatively, if the observed effect size is smaller than expected, one might want to stop for futility. As an illustration of a simple stopping rule for futility, imagine a researcher who will stop for futility whenever the observed effect size is either zero, or in the opposite direction as was predicted. In Figure 10.7 the red line indicates critical values to declare a significant

effect. In essence, this means that if the observed  $z$ -score for the interim test is either 0 or negative, data collection will be terminated. This can be specified by adding `futilityBounds = c(0, 0)` to the specification of the sequential design. One can choose in advance to stop whenever the criteria to stop for futility have been met, (i.e., a binding futility rule), but it is typically recommended to allow the possibility to continue data collection (i.e., a non-binding futility rule, specified by setting `bindingFutility = FALSE`).

```
design <- getDesignGroupSequential(
  sided = 1,
  alpha = 0.05,
  beta = 0.1,
  typeOfDesign = "asP",
  futilityBounds = c(0, 0),
  bindingFutility = FALSE
)
```

In Figure 10.7 we see a sequential design where data collection is stopped to reject  $H_0$  when the observed  $z$ -score is larger than the values indicated by the red line, computed based on a Pocock-like alpha spending function (as in Figure 10.4). In addition, data collection will stop when at an interim analysis a  $z$ -score lower than or equal to 0 is observed, as indicated by the blue line. If the data collection is not stopped at a look, it will continue to the next look. At the last look, the red and blue lines meet, because we will either reject  $H_0$  at the critical value, or fail to reject  $H_0$ .

Manually specifying the futility bounds is not ideal, as we risk stopping data collection because we fail to reject  $H_0$ , when there is a high probability of a Type 2 error. It is better to set the futility bounds by directly controlling the Type 2 error across looks at the data. Just as we are willing to distribute our Type I error rate across interim analyses, we can distribute our Type II error rate across looks, and decide to stop for futility when we fail to reject the effect size of interest with a desired Type 2 error rate.

When a study is designed such that the null hypothesis significance test has 90% power to detect an effect of  $d = 0.5$ , 10% of the time  $H_0$  will not be rejected when it should. In these 10% of cases where we make a Type 2 error, the conclusion will be that an effect of 0.5 is not present, when in reality, there is an effect of  $d = 0.5$  (or larger). In an equivalence against a smallest effect size of interest of  $d = 0.5$ , the conclusion that an effect of 0.5 or larger is not present, when in reality there is an effect of  $d = 0.5$  (or larger), is called a Type 1 error: We incorrectly conclude the effect is practically equivalent to zero. Therefore, what is a Type 2 error in NHST when  $H_0$  is  $d = 0$  and  $H_1 = d = 0.5$  is a Type 1 error in an equivalence test where  $H_0$  is  $d = 0.5$  and  $H_1$  is  $d = 0$  (Jennison & Turnbull, 2000). Controlling the Type 2 error in a sequential design can therefore be seen as controlling the Type 1 error for an equivalence test against the effect size the study is powered for. If we design a study to have a 5% Type 1 error rate and equally low Type 2 error rate (e.g., 5%, or 95% power), the study is an informative test for the presence or the absence of an effect of interest.

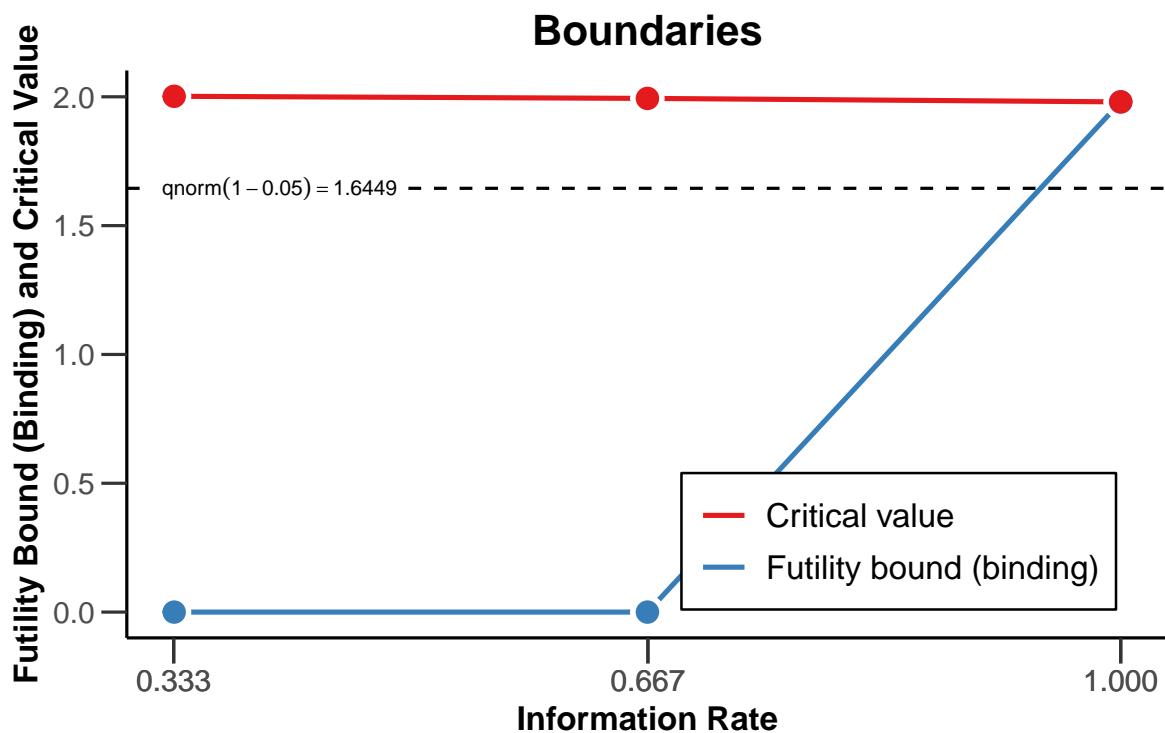


Figure 10.7: Pocock-type boundaries for 3 looks to stop when rejecting  $H_0$  (red line) or to stop for futility (blue line) when the observed effect is in the opposite direction.

If the true effect size is (close to) 0, sequential designs that stop for futility are more efficient than designs that do not stop for futility. Adding futility bounds based on beta-spending functions reduces power, which needs to be compensated by increasing the sample size, but this can be compensated by the fact that studies can stop earlier for futility, which can make designs more efficient. When specifying a smallest effect size of interest is not possible, researchers might not want to incorporate stopping for futility into the study design. To control the Type 2 error rate across looks, a **beta-spending function** needs to be chosen, such as a Pocock type beta spending function, an O'Brien-Fleming type beta spending function, or a user defined beta spending function. For example, a Pocock-like beta-spending function is added through *typeBetaSpending* = "bsP". The beta-spending function does not need to be the same as the alpha-spending function. In **ract** beta-spending functions can only be chosen for directional (one-sided) tests. After all, you can consider an effect in both directions support for your hypothesis, and an effect in the opposite direction as a reason to reject the alternative hypothesis.

```
design <- getDesignGroupSequential(  
    kMax = 2,  
    typeOfDesign = "asP",  
    sided = 1,  
    alpha = 0.05,  
    beta = 0.1,  
    typeBetaSpending = "bsP",  
    bindingFutility = FALSE  
)  
  
plot(design)
```

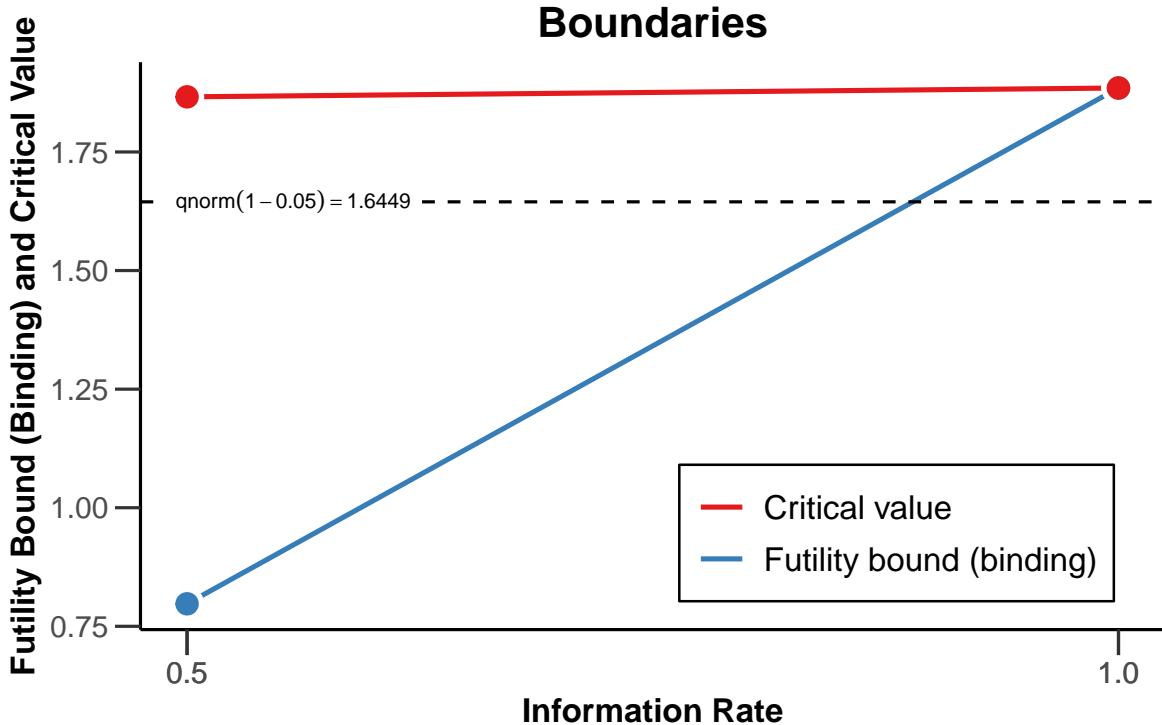


Figure 10.8: Pocock-type boundaries for 3 looks to stop when rejecting  $H_0$  (red line) or to stop for futility (blue line) based on a Pocock-type beta-spending function.

With a beta-spending function, the expected number of subjects under  $H_1$  will increase, so if the alternative hypothesis is true, designing a study to be able to stop for futility comes at a cost. However, it is possible that  $H_0$  is true, and when it is, stopping for futility reduces the expected sample size. In Figure 10.9 you can see that the probability of stopping (the green line) is now also high when the true effect size is 0, as we will now stop for futility, and if we do, the expected sample size (the blue line) is lower compared to Figure 10.6. It is important to design studies that have a high informational value to reject the presence of a meaningful effect at the final analysis, but whether stopping for futility early is an option you want to build into a study is a choice that requires considering the probability that the null hypothesis is true and a (perhaps small) increase in the sample size.

## 10.8 Reporting the results of a sequential analysis

Group sequential designs have been developed to efficiently test hypotheses using the Neyman-Pearson approach for statistical inference, where the goal is to decide how to act, while controlling error rates in the long run. Group sequential designs do not have the goal to quantify the

## Expected Sample Size and Power / Early Stop

$N_{\max}=190$ , standard deviation=1,  $H_0$ : mean difference=0

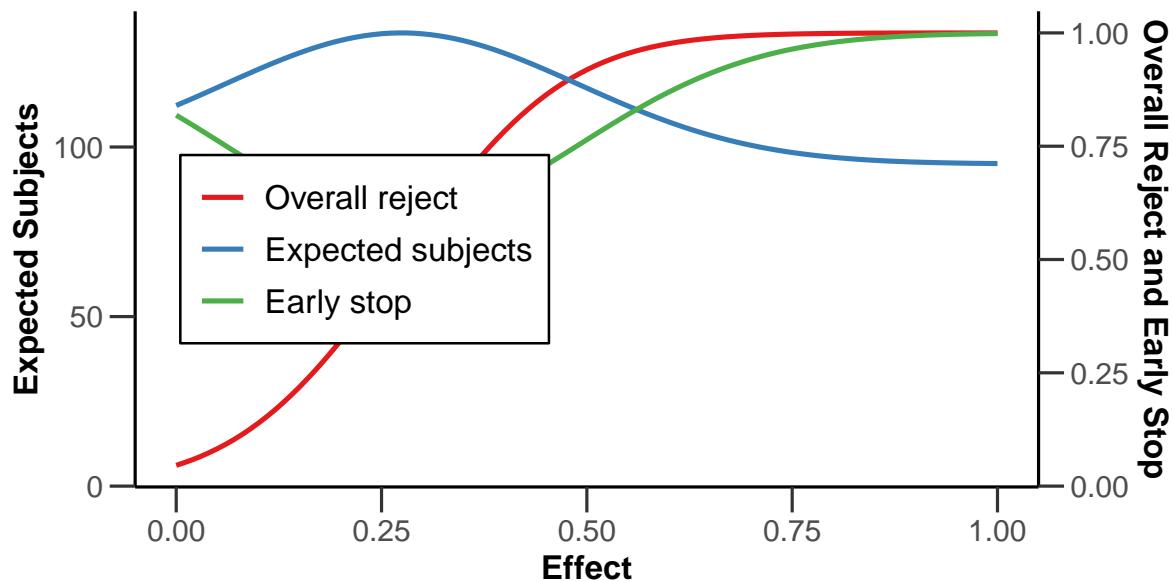


Figure 10.9: Power curve for a sequential design with 2 looks with stopping for futility.

strength of evidence, or provide accurate estimates of the effect size (Proschan et al., 2006). Nevertheless, after having reached a conclusion about whether a hypothesis can be rejected or not, researchers will often want to also interpret the effect size estimate when reporting results.

A challenge when interpreting the observed effect size in sequential designs is that whenever a study is stopped early when  $H_0$  is rejected, there is a risk that the data analysis was stopped because, due to random variation, a large effect size was observed at the time of the interim analysis. This means that the observed effect size at these interim analyses over-estimates the true effect size. As Schönbrodt et al. (2017) show, a meta-analysis of studies that used sequential designs will yield an accurate effect size, because studies that stop early have smaller sample sizes, and are weighted less, which is compensated by the smaller effect size estimates in those sequential studies that reach the final look, and are weighted more because of their larger sample size. However, researchers might want to interpret effect sizes from single studies before a meta-analysis can be performed, and in this case, reporting an adjusted effect size estimate can be useful. Although sequential analysis software only allows one to compute adjusted effect size estimates for certain statistical tests, we recommend reporting both the adjusted effect size where possible, and to always also report the unadjusted effect size estimate for future meta-analyses.

A similar issue is at play when reporting  $p$  values and confidence intervals. When a sequential design is used, the distribution of a  $p$  value that does not account for the sequential nature of the design is no longer uniform when  $H_0$  is true. A  $p$  value is the probability of observing a result *at least as extreme* as the result that was observed, given that  $H_0$  is true. It is no longer straightforward to determine what ‘at least as extreme’ means a sequential design (T. D. Cook, 2002). The most widely recommended procedure to determine what “at least as extreme” means is to order the outcomes of a series of sequential analyses in terms of the look at which the study was stopped, where earlier stopping is more extreme than later stopping, and where studies with higher  $z$  values are more extreme, when different studies are stopped at the same time (Proschan et al., 2006). This is referred to as *stagewise ordering*, which treats rejections at earlier looks as stronger evidence against  $H_0$  than rejections later in the study (Wassmer & Brannath, 2016). Given the direct relationship between a  $p$  value and a confidence interval, confidence intervals for sequential designs have also been developed.

Reporting adjusted  $p$  values and confidence intervals, however, might be criticized. After a sequential design, a correct interpretation from a Neyman-Pearson framework is to conclude that  $H_0$  is rejected, the alternative hypothesis is rejected, or that the results are inconclusive. The reason that adjusted  $p$  values are reported after sequential designs is to allow readers to interpret them as a measure of evidence. Dupont (1983) provides good arguments to doubt that adjusted  $p$  values provide a valid measure of the strength of evidence. Furthermore, a strict interpretation of the Neyman-Pearson approach to statistical inferences also provides an argument against interpreting  $p$  values as measures of evidence (Daniël Lakens, 2022b). Therefore, it is recommended, if researchers are interested in communicating the evidence in the data for  $H_0$  relative to the alternative hypothesis, to report likelihoods or Bayes factors,

which can always be reported and interpreted after the data collection has been completed. Reporting the unadjusted  $p$ -value in relation to the alpha level communicates the basis to reject hypotheses, although it might be important for researchers performing a meta-analysis based on  $p$ -values (e.g., a  $p$ -curve or  $z$ -curve analysis, as explained in the chapter on bias detection) that these are sequential  $p$ -values. Adjusted confidence intervals are useful tools to evaluate the observed effect estimate relative to its variability at an interim or the final look at the data. Note that the adjusted parameter estimates are only available in statistical software for a few commonly used designs in pharmaceutical trials, such as comparisons of mean differences between groups, or survival analysis.

Below, we see the same sequential design we started with, with 2 looks and a Pocock-type alpha spending function. After completing the study with the planned sample size of 95 participants per condition (where we collect 48 participants at look 1, and the remaining 47 at look 2), we can now enter the observed data using the function `getDataset`. The means and standard deviations are entered for each stage, so at the second look, only the data from the second 95 participants in each condition are used to compute the means (1.51 and 1.01) and standard deviations (1.03 and 0.96).

```
design <- getDesignGroupSequential(
  kMax = 2,
  typeOfDesign = "asP",
  sided = 2,
  alpha = 0.05,
  beta = 0.1
)

dataMeans <- getDataset(
  n1 = c(48, 47),
  n2 = c(48, 47),
  means1 = c(1.12, 1.51), # for directional test, means 1 > means 2
  means2 = c(1.03, 1.01),
  stDevs1 = c(0.98, 1.03),
  stDevs2 = c(1.06, 0.96)
)

res <- getAnalysisResults(
  design,
  equalVariances = TRUE,
  dataInput = dataMeans
)

print(summary(res))
```

```

[PROGRESS] Stage results calculated [0.0458 secs]
[PROGRESS] Conditional power calculated [0.0329 secs]
[PROGRESS] Conditional rejection probabilities (CRP) calculated [0.0013 secs]
[PROGRESS] Repeated confidence interval of stage 1 calculated [0.6484 secs]
[PROGRESS] Repeated confidence interval of stage 2 calculated [0.824 secs]
[PROGRESS] Repeated confidence interval calculated [1.47 secs]
[PROGRESS] Repeated p-values of stage 1 calculated [0.2493 secs]
[PROGRESS] Repeated p-values of stage 2 calculated [0.2523 secs]
[PROGRESS] Repeated p-values calculated [0.5027 secs]
[PROGRESS] Final p-value calculated [0.0012 secs]
[PROGRESS] Final confidence interval calculated [0.0671 secs]

```

## Analysis results for a continuous endpoint

Sequential analysis with 2 looks (group sequential design).

The results were calculated using a two-sample t-test (two-sided, alpha = 0.05), equal variances option.

H0:  $\mu(1) - \mu(2) = 0$  against H1:  $\mu(1) - \mu(2) \neq 0$ .

Stage		1	2
Fixed weight		0.5	1
Efficacy boundary (z-value scale)		2.157	2.201
Cumulative alpha spent		0.0310	0.0500
Stage level		0.0155	0.0139
Cumulative effect size		0.090	0.293
Cumulative (pooled) standard deviation		1.021	1.013
Overall test statistic		0.432	1.993
Overall p-value		0.3334	0.0238
Test action	continue		accept
Conditional rejection probability		0.0073	
95% repeated confidence interval		[-0.366; 0.546]	[-0.033; 0.619]
Repeated p-value		>0.5	0.0819
Final p-value			0.0666
Final confidence interval			[-0.020; 0.573]
Median unbiased estimate			0.281

Imagine we have performed a study planned to have at most 2 equally spaced looks at the data, where we perform a two-sided test with an alpha of 0.05, and we use a Pocock type alpha spending function, and we observe mean differences between the two conditions at the last look. Based on a Pocock-like alpha spending function with two equally spaced looks the

alpha level for a two-sided  $t$ -test is 0.003051, and 0.0490. We can thus reject  $H_0$  after look 2. But we would also like to report an effect size, and adjusted  $p$  values and confidence intervals.

The results show that the action after look 1 was to continue data collection, and that we could reject  $H_0$  at the second look. The unadjusted mean difference is provided in the row “Overall effect size” and at the final look this was 0.293. The adjusted mean difference is provided in the row “Median unbiased estimate” and is lower, and the adjusted confidence interval is in the row “Final confidence interval”, giving the result 0.281, 95% CI [-0.02, 0.573].

The unadjusted  $p$  values for a one-sided test are reported in the row “Overall  $p$ -value”. The actual  $p$  values for our two-sided test would be twice as large, so 0.6668, 0.0477. The adjusted  $p$ -value at the final look is provided in the row “Final  $p$ -value” and it is 0.06662.

## 10.9 Test Yourself

**Q1:** Sequential analyses can increase the efficiency of the studies you perform. Which statement is true for a sequential design in which researchers only stop if  $H_0$  can be rejected (and did not specify a rule to stop for futility)?

- (A) Sequential analyses will reduce the sample size of every study you will perform.
- (B) Sequential analyses will on average reduce the sample size of studies you will perform.
- (C) Sequential analyses will on average reduce the sample size of studies you will perform, as long as there is a true effect (when a rule to stop for futility has not been specified).
- (D) Sequential analyses will on average require the same sample size as fixed designs, but offer more flexibility.

**Q2:** What is the difference between sequential analysis and optional stopping?

- (A) The only difference is that a sequential analysis is transparently reporting, while optional stopping is typically not disclosed in a paper.
- (B) In sequential analysis the Type 1 error rate is controlled, while in optional stopping the Type 1 error rate is inflated.
- (C) In optional stopping data collection is only terminated when a significant result has been observed, while in sequential analysis data collection can also stop when

the absence of a meaningful effect has been established.

- (D) In sequential analysis it is not possible to design a study where you analyze the data after every participant, while you can do this in optional stopping.

**Q3:** What is the defining feature of the Pocock correction?

- (A) It uses a very conservative alpha level for early looks, and the alpha level at the last look is close to the unadjusted alpha level in a fixed design.
- (B) It uses the same alpha level at each look (or almost the same alpha level at each look, when using an Pocock-like alpha spending function).
- (C) It uses a critical value of 3 at each interim analysis, and spends the remaining Type 1 error rate at the last look.
- (D) It has a parameter that can be chosen such that the Type 1 error rate is spent more conservatively or more liberally in early interim analyses.

**Q4:** A benefit of the O'Brien-Fleming correction is that the alpha level at the last look is close to the alpha level. Why is this a benefit?

- (A) It means that the sample size based on an a-priori power analysis (which depends on the alpha level) is close to the sample size in a fixed design, while allowing additional looks at the data.
- (B) It means the Type 1 error rate is inflated only a little bit, compared to a fixed design.
- (C) It means the Type 1 error rate is only a bit more conservative, compared to a fixed design.
- (D) It means that the sample size based on an a-priori power analysis (which depends on the alpha level) is always identical to the sample size in a fixed design, while allowing additional looks at the data.

**Q5:** A researcher uses a sequential design for a study with 5 looks at the data, with a desired overall alpha level of 0.05 for a two-sided test, and chooses a **Pocock correction**. After continuing data collect to the third look, the researcher observes a *p*-value of 0.011. Which statement is true? Note: remember that `ract` returns one-sided alpha levels. You can use the following code by replacing 0 and specifying the `typeOfDesign`:

```

design <- rpact::getDesignGroupSequential(
  kMax = 0,
  typeOfDesign = "",
  sided = 0,
  alpha = 0.0
)
print(summary(design))

```

- (A) The researcher can reject the null hypothesis and can terminate data collection.
- (B) The researcher fails to reject the null hypothesis and needs to continue the data collection.

**Q6:** A researcher uses a sequential design for a study with 5 looks at the data, with a desired overall alpha level of 0.05, and chooses an **O'Brien-Fleming correction**. After continuing data collect to the third look, the researcher observes a  $p$ -value of 0.011. Which statement is true (you can use the same code as for Q5)?

- (A) The researcher can reject the null hypothesis and can terminate data collection.
- (B) The researcher fails to reject the null hypothesis and needs to continue the data collection.

**Q7:** For the design in Q5 (using the Pocock correction), what is the sample size required to achieve 80% power (the default – you can change the default by specifying a different value than `beta = 0.2` in the `getDesignGroupSequential` function) for an effect size of  $d = 0.5$  (which equals a mean difference of 0.5 with a standard deviation of 1). You can use the code below.

```

design <- rpact::getDesignGroupSequential(
  kMax = 5,
  typeOfDesign = "OF",
  sided = 2,
  alpha = 0.05
)

power_res <- rpact::getSampleSizeMeans(
  design = design,
  groups = 2,
  alternative = 0.5,
  stDev = 1,

```

```

allocationRatioPlanned = 1,
normalApproximation = FALSE)

print(power_res)

```

- (A) 64 (32 in each independent group)
- (B) 128 (64 in each independent group)
- (C) 154 (77 in each independent group)
- (D) 158 (79 in each independent group)

**Q8:** For the design in the previous question, what is the sample size required to achieve 80% power for an effect size of  $d = 0.5$  for a fixed design with only one look instead of 5? First update the design (by changing the `kMax` to 1) and then re-run the code provided with the previous question.

- (A) 64 (32 in each independent group)
- (B) 128 (64 in each independent group)
- (C) 154 (77 in each independent group)
- (D) 158 (79 in each independent group)

We see the sample size increases quite a bit because of the choice for the Pocock correction, and the number of looks (5, which lead to a low alpha level at the final look). The ratio of the maximum sample size for a sequential design and the sample size for a fixed design is known as the **inflation factor**, which is independent of the effect size. Although a-priori power analyses have not been programmed for all types of tests, the inflation factor can be used to compute the increased number of observations that is required relative to a fixed design for any test. Researchers can perform an a-priori power analysis for a fixed design with any tool they would normally use, and multiply the total number of observations with the inflation factor to determine the required sample size for a sequential design. The inflation factor can be retrieved using the `getDesignCharacteristics` function.

**Q9:** First, re-run the code to create a sequential design with 5 looks and a Pocock correction at the data used in Q7. Then, run the code below, and find the inflation factor. What is the inflation factor, or the required increase in the sample size for a sequential design with 5 looks using the Pocock correction, compared to a fixed design? Note that `rpact` does not round

up the number of observations for per group to whole numbers when calculating the inflation factor.

```
rpact::getDesignCharacteristics(design)
```

- (A) The inflation factor is 1
- (B) The inflation factor is 1.0284
- (C) The inflation factor is 1.2286
- (D) The inflation factor is 1.2536

**Q10:** We see the inflation factor is quite large, and there is a certain probability that we will have to collect more observations than using a fixed design. Re-run the code for Q7 (for the Pocock design with 5 looks). We see that on average, if there is a true effect of 0.5, we will be more efficient than in a fixed design. What is the expected number of subjects under  $H_1$ , as provided by `rpact`?

- (A) 101.9
- (B) 104.3
- (C) 125.3
- (D) 152.8

We see the sequential design will on average be more efficient than a fixed design, but the decision about the trade-off between the specific sequential design used, and whether the possible benefit is worth the risk of collecting additional data, must be made on a case-by-case basis.

**Q11:** First, change the code to create a sequential design with 5 looks at the data used in Q7 from the Pocock correction to the OF (O'Brien-Fleming) correction. Then, run the code in question 9 again, and find the inflation factor for this design. What is the inflation factor?

- (A) The inflation factor is 1
- (B) The inflation factor is 1.0284
- (C) The inflation factor is 1.2286

- (D) The inflation factor is 1.2536

**Q12:** It is also possible to stop for futility (or to reject the presence of a specific effect of interest). Researchers should decide between binding and non-binding beta-spending functions, but they do not need to decide between binding and non-binding alpha spending functions. If a researcher observed a statistically significant result at an interim analysis, but decides not to stop the data collection, but continue the data collection (for example to get a more precise effect size estimate) what are the consequences?

- (A) The Type 1 error rate will inflate, and the Type 2 error rate will inflate.
- (B) The Type 1 error rate will inflate, and the Type 2 error rate will not inflate.
- (C) The Type 1 error rate will not inflate, and the Type 2 error rate will inflate. v
- (D) The Type 1 error rate will not inflate, and the Type 2 error rate will not inflate.

**Q13:** In the plot below you see the  $t$ -score boundaries for a sequential design to stop to reject  $H_0$  (the red line) and to reject  $H_1$  (the blue line). At the second interim look, you perform a test, and observe a  $t$ -value of 2. Which decision would you make?

- (A) You can reject  $H_0$  and stop data collection.
- (B) You can reject  $H_1$  and stop data collection.
- (C) You reject both  $H_0$  and  $H_1$  and stop data collection.
- (D) You fail to reject both  $H_0$  and  $H_1$  and continue data collection.

### 10.9.1 Open Questions

1. What is the difference between sequential analysis and optional stopping?
2. What is a possible benefit of using a sequential design over a fixed design?
3. What does it mean to stop data collection for futility?
4. What is the difference in the philosophy of how the alpha is spent across looks between the Pocock and O'Brien-Fleming approaches?
5. What is the benefit of the fact that the alpha level at the final look when using an O'Brien-Fleming correction is close to the uncorrected alpha level?

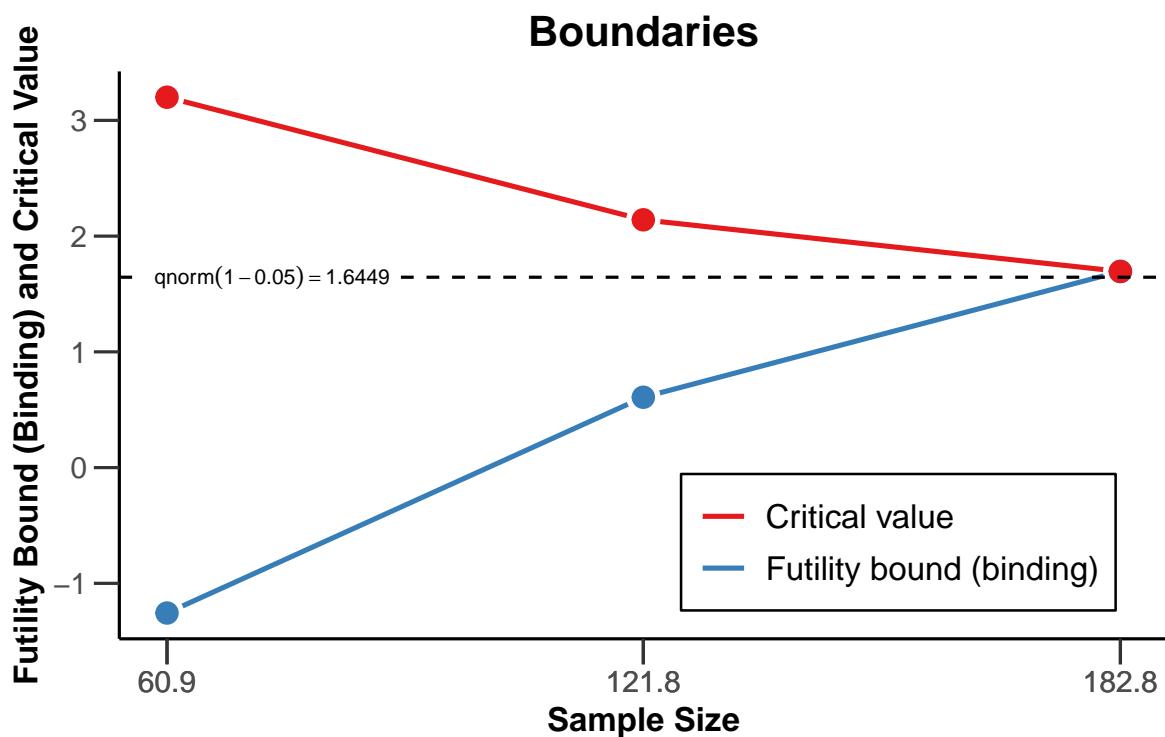


Figure 10.10: Example of O'Brien-Fleming-type boundaries for 3 looks to stop when rejecting  $H_0$  (red line) or to stop for futility (blue line) with a 5% Type 1 and Type 2 error.

6. What is the difference between the Pocock and O'Brien-Fleming correction, and the corresponding Pocock and O'Brien-Fleming alpha spending functions developed by Lan and DeMets?
7. How can it be that even though the maximum sample size for a sequential design is slightly larger than the sample size for a fixed design, sequential designs can still be more efficient?
8. When does incorporating a stopping rule for futility increase the efficiency of a sequential design?
9. On average, what is the effect of stopping early in a sequential design on the effect size estimate? What is an argument to not correct the effect size estimate when reporting it?

# 11 Meta-analysis

Every single study is just a data-point in a future meta-analysis. If you draw small samples from a population, the mean and standard deviation in the sample can differ considerably from the mean and standard deviation in the population. There is great variability in small samples. Parameter estimates from small samples are very imprecise, and therefore the 95% confidence intervals around effect sizes are very wide. Indeed, this led Cohen (1994) to write “I suspect that the main reason [confidence intervals](#) are not reported is that they are so embarrassingly large!” If we want a more precise estimate of our parameter of interest, such as the mean difference or correlation in the population, we need either run extremely large single studies, or alternatively, combine data from several studies by performing a meta-analysis. The most common approach to combine studies is to perform a meta-analysis of effect size estimates.

You can perform a meta-analysis for a set of studies in a single article you plan to publish (often called an internal meta-analysis), or you can search the literature for multiple studies reported in as many different articles as possible, and perform a meta-analysis on all studies others have published. An excellent introduction to meta-analyses is provided in the book by Borenstein (2009). There is commercial software you can use to perform meta-analyses, but I highly recommend against using such software. Almost all commercial software packages lack transparency, and do not allow you to share your analysis code and data with other researchers. In this chapter, we will be using R to perform a meta-analysis of effect sizes, using the `metafor` package by Viechtbauer (2010). An important benefit of using `metafor` is that your meta-analysis can be made completely reproducible. If you plan to perform a narrative review, it is relatively little additional effort to also code the effect sizes and sample size, and perform an effect size meta-analysis, and to code the statistical tests and  $p$ -values, to perform a  $p$ -curve or  $z$ -curve analysis (which will be discussed in the next chapter on [bias detection](#)).

## 11.1 Random Variation

People find it difficult to think about random variation. Our mind is more strongly geared towards recognizing patterns than randomness. In this section, the goal is to learn what random variation looks like, and how the number of observations collected determines the amount of variation.

Intelligence tests have been designed such that the mean Intelligence Quotient of the entire population of adults is 100, with a standard deviation of 15. This will not be true for every sample we draw from the population. Let's get a feel for what the IQ scores from a sample look like. Which IQ scores will people in our sample have?

We will start by manually calculating the mean and standard deviation of a random sample of 10 individuals. Their IQ scores are: 91.15, 86.52, 75.64, 115.72, 95.83, 105.44, 87.10, 100.81, 82.63, and 106.22. If we sum these 10 scores and divide them by 10, we get the mean of our sample: 94.71. We can also calculate the standard deviation from our sample. First, we subtract the overall mean (94.71) from each individual IQ score. Then, we square these differences and then sum these squared differences (giving 1374.79). We divide this sum of the squared difference by the sample size minus 1 ( $10-1=9$ ), and finally take the square root of this value, which gives the standard deviation: 12.36. Copy the code below, remove the `set.seed(3190)` line (which makes the code reproducible but creates the same data as in the plot below each time) and run it to randomly simulate 10 IQ scores and plot them.

```
library(ggplot2)
set.seed(3190) # set seed for reproducibility
n <- 10 # set sample size
x <- rnorm(n = n, mean = 100, sd = 15) # simulate data

# plot data adding normal distribution and annotations
ggplot(as.data.frame(x), aes(x)) +
  geom_histogram(colour = "black", fill = "grey", aes(y = after_stat(density)), binwidth = 2,
                 stat_function(fun = dnorm, args = c(mean = 100, sd = 15), linewidth = 1, color = "red", lty = 2))
  xlab("IQ") +
  ylab("number of people") +
  theme_bw(base_size = 20) +
  geom_vline(xintercept = mean(x), colour = "gray20", linetype = "dashed") +
  coord_cartesian(xlim = c(50, 150)) +
  scale_x_continuous(breaks = seq(50, 150, 10)) +
  annotate("text", x = mean(x), y = 0.02, label = paste("Mean = ", round(mean(x)), "\n", "SD = ", round(sd(x), 2), "\n", "N = ", n))
  theme(plot.background = element_rect(fill = "#fffffa")) +
  theme(panel.background = element_rect(fill = "#fffffa"))
```

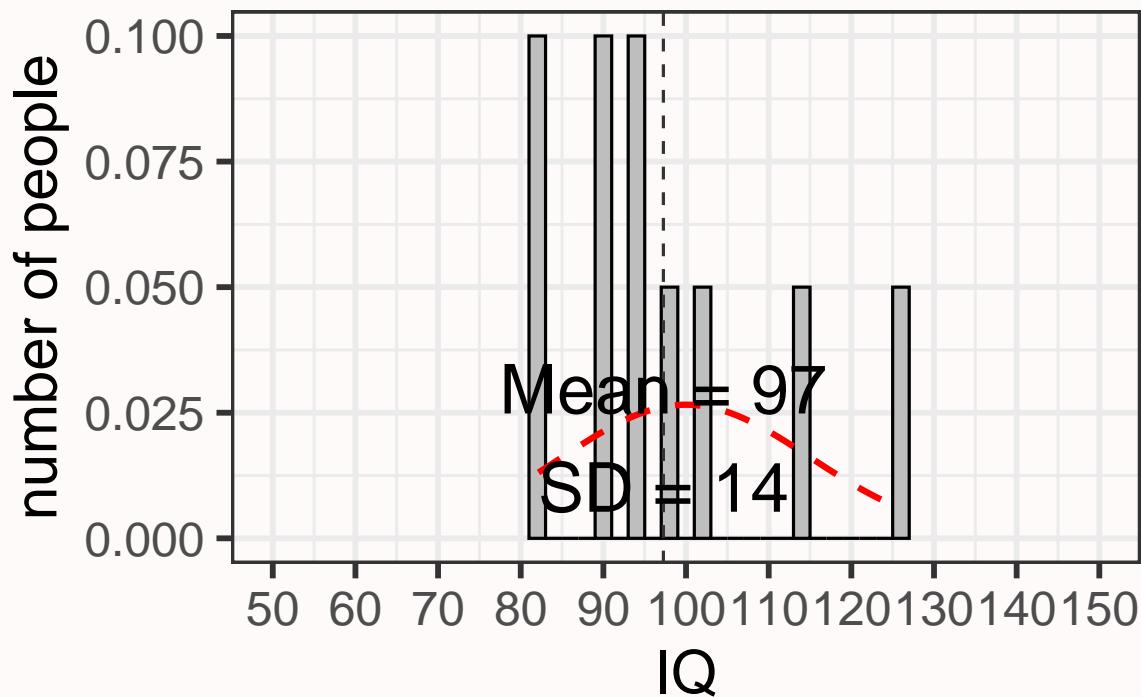
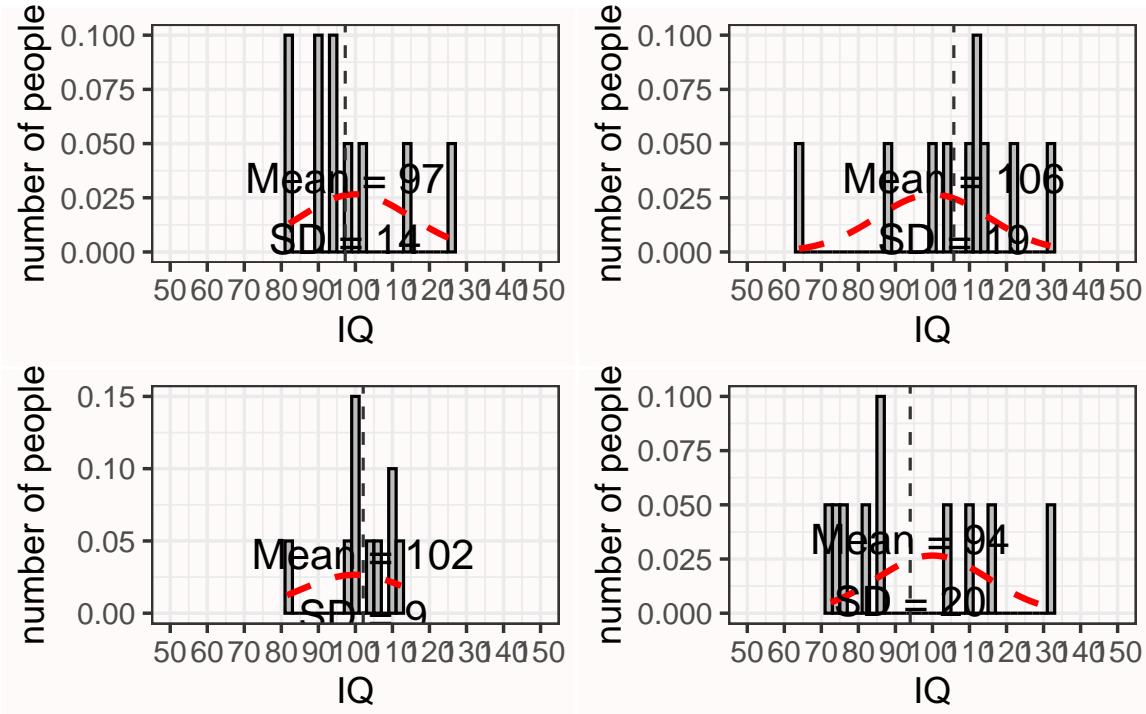


Figure 11.1: Simulation of 10 random datapoints with mean = 100 and sd = 15 in the population.

The plot above provides one example of a randomly simulated dataset of 10 points drawn from a normal distribution with a mean of 100 and a standard deviation of 15. The grey bars indicate the frequency with which each IQ score was observed. The red dotted line illustrates the normal distribution based on the mean and sd of the population. Both the observed mean (97; thin vertical dashed line), as well as the observed standard deviation (14), differ from the true population values. If we simulate 4 additional datasets, we see both the mean and the standard deviation vary.



Imagine we did not yet know what the mean IQ was in our population (where  $M = 100$ ), or the standard deviation (where  $SD = 15$ ), and that we would only have access to one dataset. Our estimate might be rather far off. This type of variation is to be expected in small samples of 10 participants, given the true standard deviation. The variability in the mean is determined by the standard deviation of the measurement. In real life, the standard deviation can be reduced by for example using multiple and reliable measurements (which is why an IQ test has not just one question, but many different questions). But we can also make sure our sample mean is closer to the population mean by increasing the sample size.

A new simulated sample with 100 participants is plotted below. We are slowly seeing what is known as the **normal distribution** (and the frequency scores start to resemble the red dotted line illustrating the normal distribution of the population). This is the well-known bell shaped curve that represents the distribution of many variables in scientific research (although some other types of distributions are quite common as well). The mean and standard deviation are much closer to the true mean and standard deviation, and this is true for most of the simulated samples if you set  $n < 100$  in the code above and run additional simulations.

If we simulate a really large sample of 1000 observations, we will see the benefits of collecting a large sample size in terms of accuracy of the measurement. Not every simulated study of 1000 people will yield the true mean and standard deviation, but it will happen quite often.

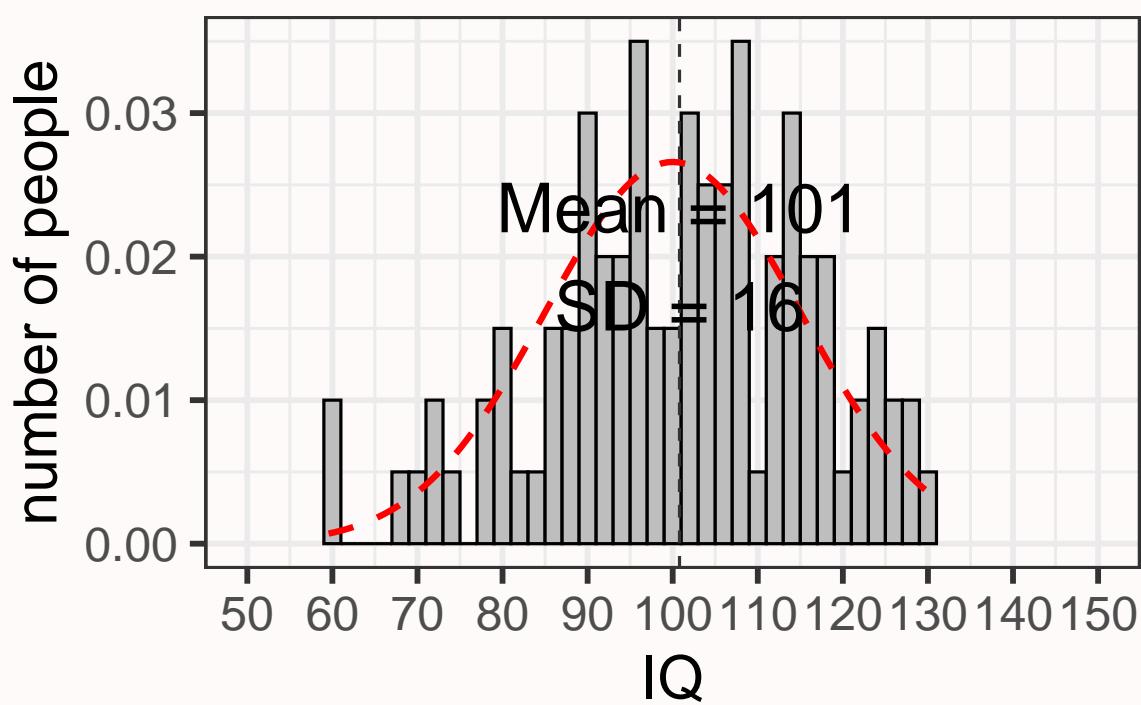


Figure 11.2: 100 random datapoints with mean = 100 and sd = 15 in the population.

And note how although the distribution is very close to a normal distribution, even with 1000 people it is not perfect.

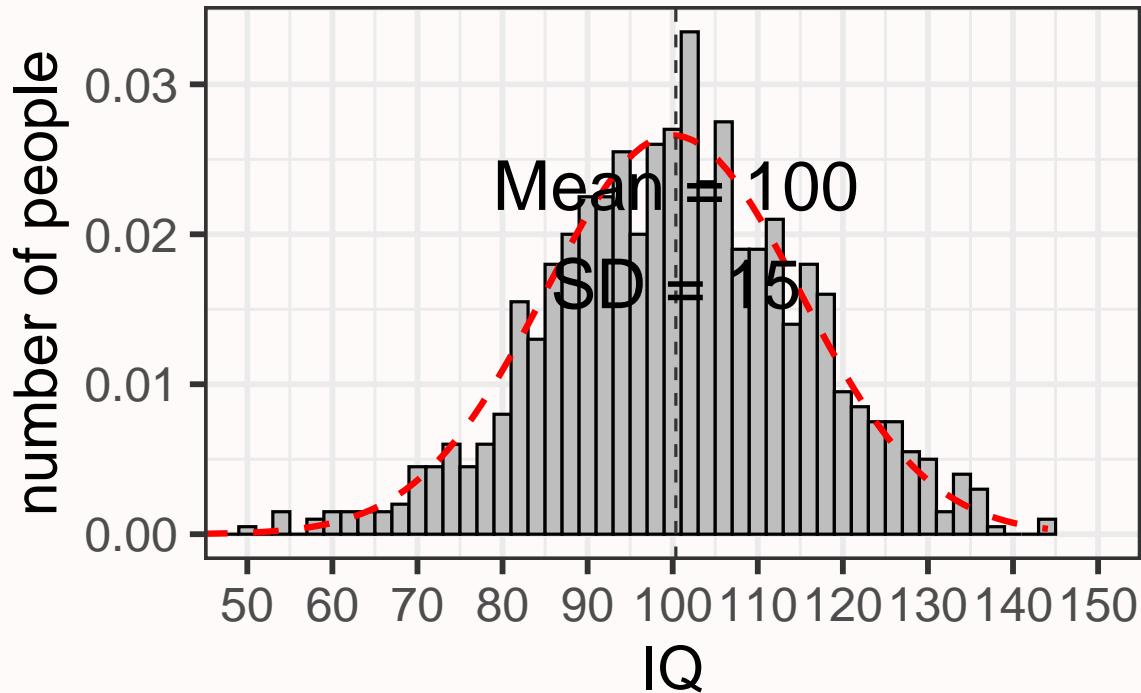


Figure 11.3: 1000 random datapoints with mean = 100 and sd = 15 in the population.

So far, we have simulated only a single group of observations, but it is also informative to examine the variation we will observe when we compare the means in two independent groups. Assume we have a new IQ training program that will increase people's IQ score by 6 points. People in condition 1 are in the control condition – they do not get IQ training. People in condition 2 get IQ training. Let's simulate 10 people in each group, assuming mean IQ in the control condition is 100 and in the experimental group is 106 (the SD is still 15 in each group).

The two groups differ in how close they are to their true means, and as a consequence, the difference between groups varies as well. Note that this difference is the main variable in statistical analyses when comparing two groups in for example a *t*-test. In this specific simulation, we got quite extreme results, with a score of 96 (when the population mean is 100) and a score of 111 (when the population mean is 106). So in this sample, due to random variation, we calculate an effect size estimate that is quite a bit larger than the true effect size. Let's simulate 4 additional datasets to see the variation.

We see that there is quite some variation, up to the point that in one simulation the sample

## Data

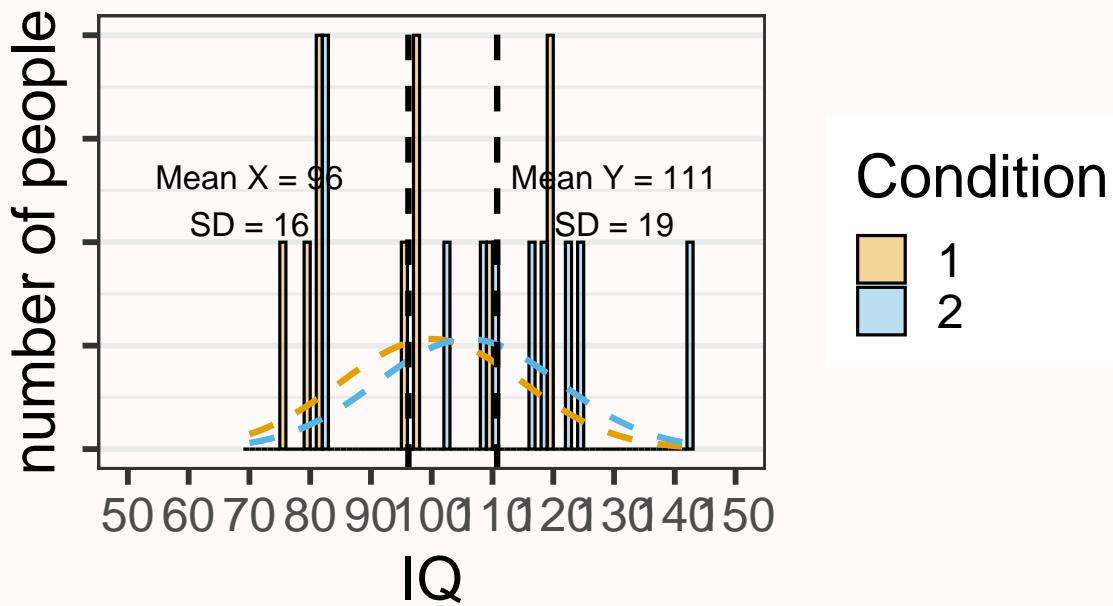


Figure 11.4: Simulation of 10 observations in two independent groups.

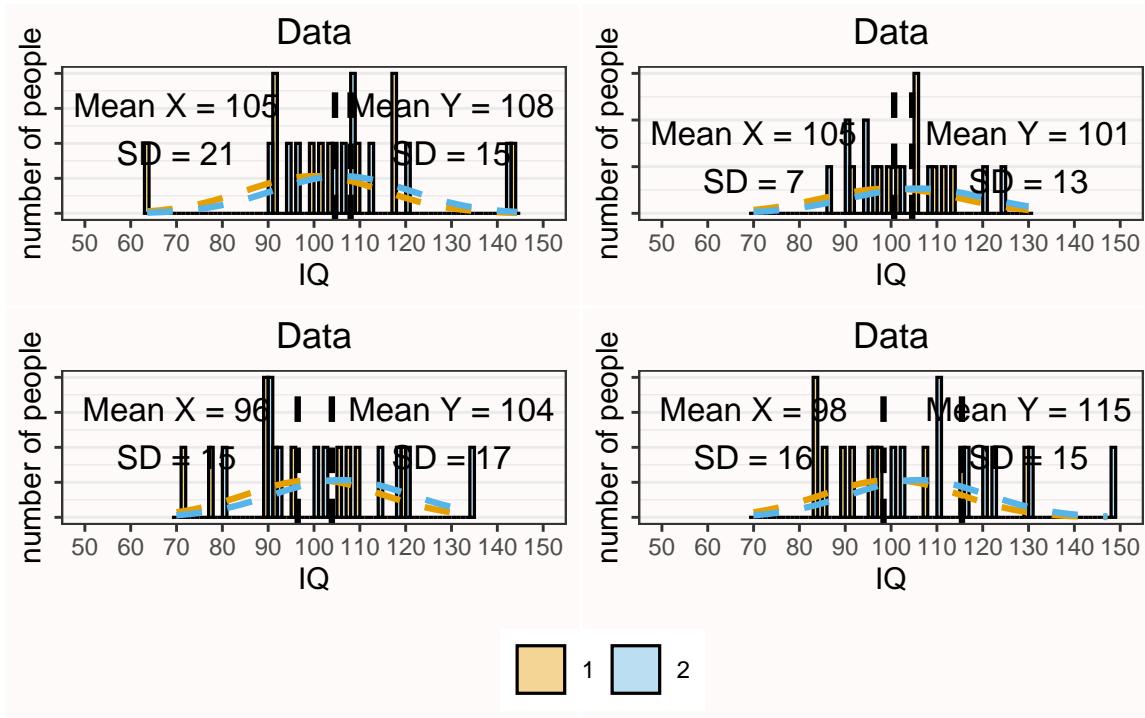


Figure 11.5: Four simulated samples of independent groups.

means are in the opposite direction of the population means. Again, increasing the sample size will mean that, in the long run, the sample means will get closer to the population means, and that we are more accurately estimating the difference between conditions. With 250 observations in each group, a randomly simulated set of observations for the two groups might look like Figure 11.6. Note that this difference might not look impressive. However, the difference would pass a significance test (an independent  $t$ -test) with a very low alpha level.

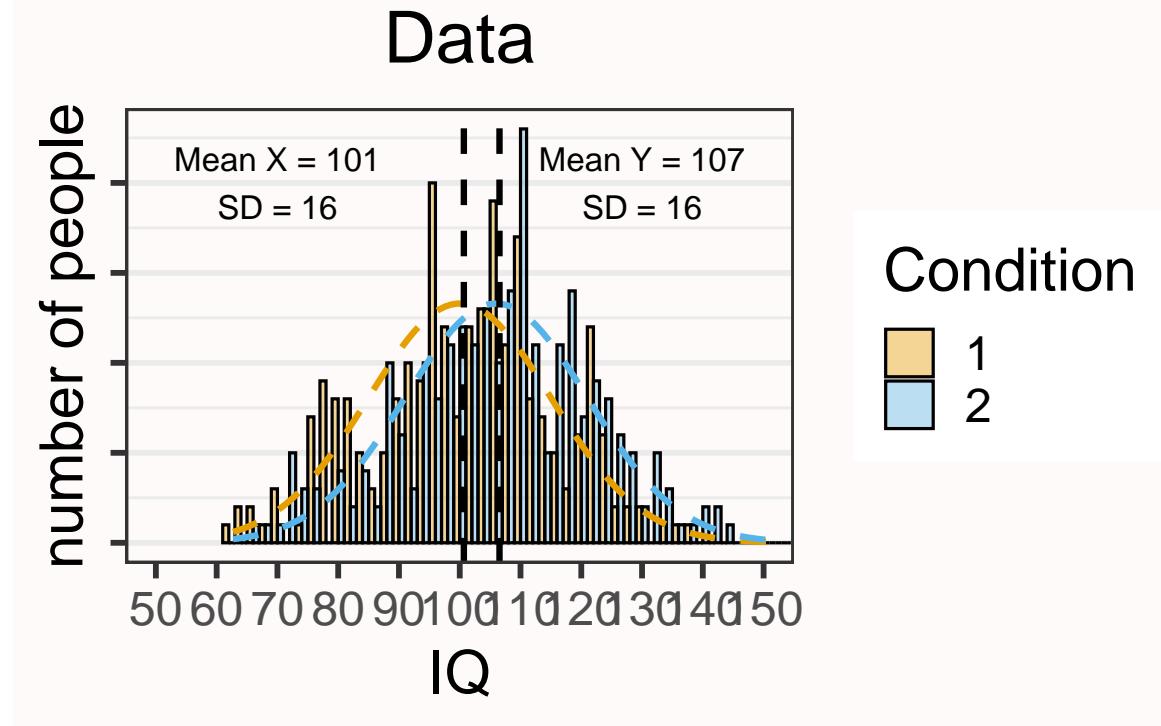


Figure 11.6: Simulated sample of 250 independent observations.

The variation in the estimate of the mean decreases as the sample size increases. The larger the sample size, the more precise the estimate of the mean becomes. The **standard deviation of the sample** ( $\sigma_x$ ) of single IQ scores is 15, irrespective of the sample size, and the larger the sample size, the more accurately we can measure the true standard deviation. But the **standard deviation of the sampling distribution of the sample mean** ( $\sigma_{\bar{x}}$ ) decreases, as the sample size increases, and is referred to as the **standard error (SE)**. The estimated standard deviation of the sample mean, or the standard error, calculated based on the observed standard deviation of the sample ( $\sigma_x$ ) is:

$$SE = \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$$

Based on this formula, and assuming an observed standard deviation of the sample of 15, the

standard error of the mean is 4.74 for a sample size of 10, and 0.95 for a sample size of 250. Because estimates with a lower standard error are more precise, the effect size estimates in a meta-analysis are weighed based on the standard error, with the more precise estimates getting more weight.

So far we have seen random variation in means, but correlations will show similar variation as a function of the sample size. We will continue with our example of measuring IQ scores, but now we search for fraternal (so not identical) twins, and measure their IQ. Estimates from the literature suggest the true correlation of IQ scores between fraternal twins is around  $r = 0.55$ . We find 30 fraternal twins, measure their IQ scores, and plot the relation between the IQ of both individuals. In this simulation, we assume all twins have a mean IQ of 100 with a standard deviation of 15.

The correlation is calculated based on the IQ scores of one fraternal twin (x) and the IQ scores of the other fraternal twin (y) for each pair of twins, and the total number of pairs (N). In the numerator of the formula, the number of pairs is multiplied by the sum of the product of x and y, and from this value the sum of x multiplied by the sum of y is subtracted. In the denominator, the square root is taken from the number of pairs multiplied by the sum of x squared, from which the sum of x, which is then squared, is subtracted, and multiplied by the same calculation but now for y.

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

When we randomly simulate observations for 30 twins, we get the following result.

On the x-axis, we see the IQ score of one twin, and on the y-axis we see the IQ score of the second twin, for each pair. The black dotted diagonal line illustrates the true correlation (0.55), while the yellow line shows the observed correlation (in this case,  $r = 0.43$ ). The slope of the yellow line is determined by the observed correlation, but the position of the line is influenced by the mean IQ scores in both groups (in this simulation, the mean on the y-axis is 105, somewhat above 100, and the mean on the x-axis is 102, also slightly above 100. The blue area is the 95% confidence interval around the observed correlation. As we saw in the chapter on confidence intervals, 95% of the time (in the long run) the blue area will contain the true correlation (the dotted black line). As in the examples based on means, increasing the sample size to 300 narrows the confidence interval considerably, and will mean that most of the time the correlation in the sample is much closer to the correlation in the population. As the sample size increases, the estimate of the correlation becomes more precise, following the formula of the standard error of a correlation:

$$SE_{r_{xy}} = \frac{1 - r_{xy}^2}{\sqrt{(n - 2)}}$$

**Correlation = 0.43**

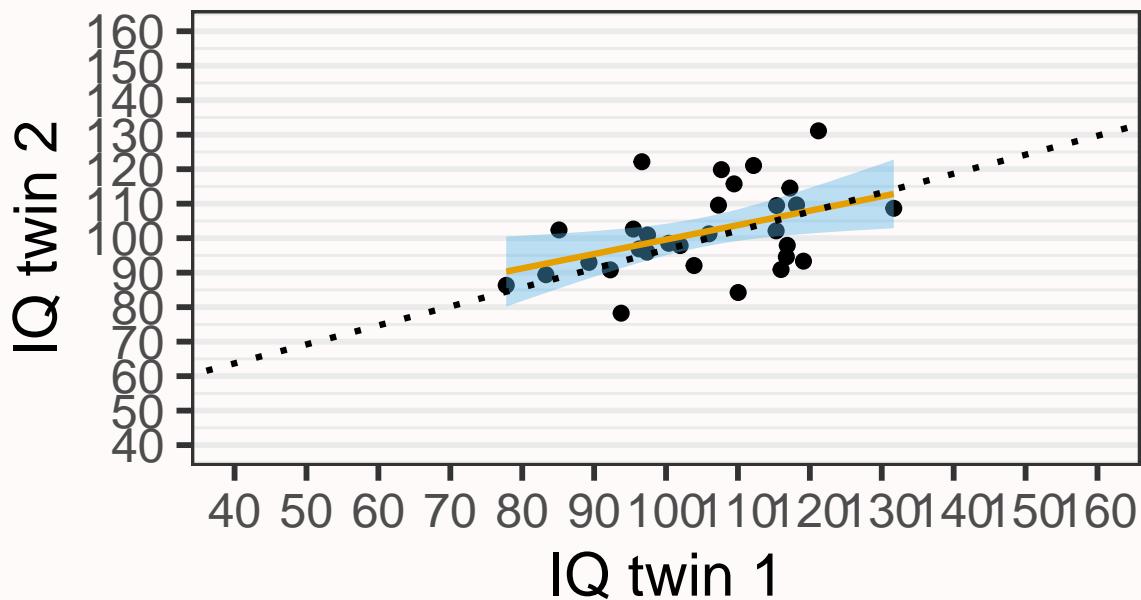


Figure 11.7: Correlation based on 30 pairs.

**Correlation = 0.52**

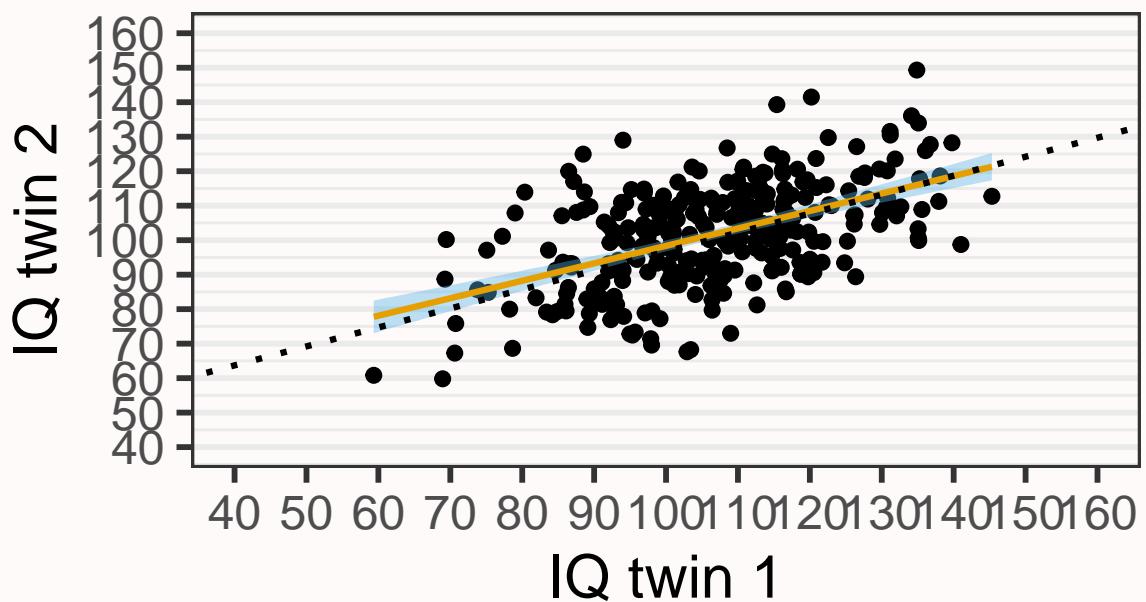


Figure 11.8: Correlation based on 300 pairs.

Because estimates of means, standard deviations, or correlations based on small samples have relatively large uncertainty, it is preferable to collect larger samples. However, this is not always possible, and often the goal of a study is not to provide an accurate estimate, but to test a hypothesis. A study often requires less observations to achieve sufficient power for a hypothesis test, than are required to be able to accurately estimate a parameter (Maxwell et al., 2008). Therefore, scientists often rely on meta-analyses, where data from multiple studies are combined, to provide accurate estimates.

## 11.2 A single study meta-analysis

Let's first begin with something you will hardly ever do in real life: a meta-analysis of a single study. This is a little silly, because a simple *t*-test or correlation will tell you the same thing – but it is educational to compare a *t*-test with a meta-analysis of a single study, before we look at how multiple studies are combined into a meta-analysis.

A difference between an independent *t*-test and a meta-analysis is that a *t*-test is performed on the raw data, while a meta-analysis is typically performed on the effect size(s) of individual studies. The `metafor` R package contains a very useful function called `escalc` that can be used to calculate effect sizes, their variances, and confidence intervals around effect size estimates. So let's start by calculating the effect size to enter into our meta-analysis. As explained in the chapter on [effect sizes](#) the two main effect sizes used for meta-analyses of continuous variables are the standardized mean difference (*d*) or the correlation (*r*), although it is of course also possible to perform meta-analyses on dichotomous variables (we will see an example below). The code below will calculate the **standardized mean difference** (SMD) from two independent groups from **means** (specified by *m1i* and *m2i*), **standard deviations** (*sd1i* and *sd2i*), and the number of observations in each group (*n1i* and *n2i*). By default, `metafor` computes the effect size '**Hedges' g**' which is the unbiased version of Cohen's *d* (see the section on [Cohen's d](#) in the chapter on Effect Sizes).

```
library(metafor)
g <- escalc(measure = "SMD",
             n1i = 50, # sample size in Group 1
             m1i = 5.6, # observed mean in Group 1
             sd1i = 1.2, # observed standard deviation in Group 1
             n2i = 50, # sample size in Group 2
             m2i = 4.9, # observed mean in Group 2
             sd2i = 1.3) # observed standard deviation in Group 2
g
```

yi	vi
0.5552575	0.0415416

The output gives you Hedge's  $g$  (under the `yi` column, which always returns the effect size, in this case the standardized mean difference) and the variance of the effect size estimate (under `vi`). As explained in Borenstein (2009) formula 4.18 to 4.24 the standardized mean difference Hedges'  $g$  is calculated by dividing the difference between means by the pooled standard deviation, multiplied by a correction factor,  $J$ :

$$J = (1 - 3/(4df - 1))$$

$$g = J \times \left( \frac{\bar{X}_1 - \bar{X}_2}{S_{\text{within}}} \right)$$

and a very good approximation of the variance of Hedges'  $g$  is provided by:

$$Vg = J^2 \times \left( \frac{n_1 + n_2}{n_1 n_2} + \frac{g^2}{2(n_1 + n_2)} \right)$$

The variance of the standardized mean difference depends only on the sample size (`n1` and `n2`) and the value of the standardized mean difference itself. **To perform the required calculations for a meta-analysis, you need the effect sizes and their variance.** This means that if you have coded the effect sizes and the sample sizes (per group) from studies in the literature, you have the information you need to perform a meta-analysis. You do not need to manually calculate the effect size and its variance using the two formula above – the `escalc` function does this for you. We can now easily perform a single study meta-analysis using the `rma` function in the `metafor` package:

```
meta_res <- rma(yi, vi, data = g)
meta_res
```

```
Random-Effects Model (k = 1; tau^2 estimator: REML)

tau^2 (estimated amount of total heterogeneity): 0
tau (square root of estimated tau^2 value):      0
I^2 (total heterogeneity / total variability):   0.00%
H^2 (total variability / sampling variability):  1.00
```

```
Test for Heterogeneity:  
Q(df = 0) = 0.0000, p-val = 1.0000
```

#### Model Results:

```
estimate      se     zval    pval   ci.lb   ci.ub  
0.5553  0.2038  2.7243  0.0064  0.1558  0.9547 **  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Under ‘Model Results’ we find the effect size Hedges’  $g$  (0.56) and the standard error (0.2), the  $Z$ -test statistic testing the mean difference against the null-hypothesis (2.72), and the 95% confidence interval [ci.lb = 0.16; ci.ub = 0.95] around the effect size (the interval width can be specified using the ‘level =’ option). We also see the  $p$ -value for the test of the meta-analytic effect size against 0. In this case we can reject the null-hypothesis ( $p = 0.006$ ).

In a meta-analysis, a  $Z$ -test is used to examine whether the null-hypothesis can be rejected. This assumes a normally distributed random effect size model. Normally, you would analyze data from a single study with two groups using a  $t$ -test, which not surprisingly uses a  $t$ -distribution. I don’t know why statistical computations sometimes care a lot about a small amount of bias (the difference between the effect size  $d$  and  $g$ , for example) and sometimes not (the difference between  $Z$  and  $t$ ), but meta-analysts seem happy with  $Z$ -scores (in fact, with large enough sample sizes (which is commonly true in a meta-analysis) the difference between a  $Z$ -test and  $t$ -test is tiny). If we directly compare a single-study meta-analysis based on a  $Z$ -test with a  $t$ -test, we will see some tiny differences in the results.

As explained in the chapter on [effect sizes](#) we can directly calculate the effect size Hedges’  $g$  (and it’s 95% confidence interval) using MOTE (Buchanan et al., 2017). The MOTE package uses the  $t$ -distribution when calculating confidence intervals around the effect size (and we can see this makes only a tiny difference compared to using the  $Z$ -distribution in a meta-analysis with 50 observations in each group).

The  $t$ -value is 2.835, and the  $p$ -value is 0.006. The results are very similar to those computed when performing a meta-analysis, with  $g = 0.55$ , 95% CI[0.16; 0.94], where the effect size and the upper bound for the confidence interval differ only 0.01 after rounding.

It is now common to visualize the results of a meta-analysis using a forest plot. According to H. M. Cooper et al. (2009) the first forest plot was published in 1978 (Freiman et al., 1978), with the goal to visualize a large set of studies that had concluded the absence of an effect based on non-significant results in small studies (see Figure 11.9). By plotting the width of the confidence interval for each study, it becomes possible to see that even though the studies do not reject an effect size of 0, and thus were all non-significant, many studies also did not reject the presence of a meaningful favorable treatment effect. To make large studies more

noticeable in a forest plot, later versions added a square to indicate the estimated effect size, where the size of the square was proportional to the weight that will be assigned to the study when computing the combined effect.

In Figure 11.10 we see a modern version of a forest plot, with the effect size for Study 1 marked by the black square at 0.56, and the confidence interval visualized by lines extending to 0.16 on the left and 0.95 on the right. The numbers printed on the right-hand side of the forest plot provide the exact values for the effect size estimate and the lower and upper bound of the confidence interval. On the lower half of the forest plot, we see a stretched-out diamond, in a row labeled ‘RE Model’, for ‘Random Effects model’. The diamond summarizes the meta-analytic effect size estimate, being centered on that effect size estimate with the left and right endpoints at the 95% confidence interval of the estimate. Because we only have a single study, the meta-analytic effect size estimate is the same as the effect size estimate for our single study.

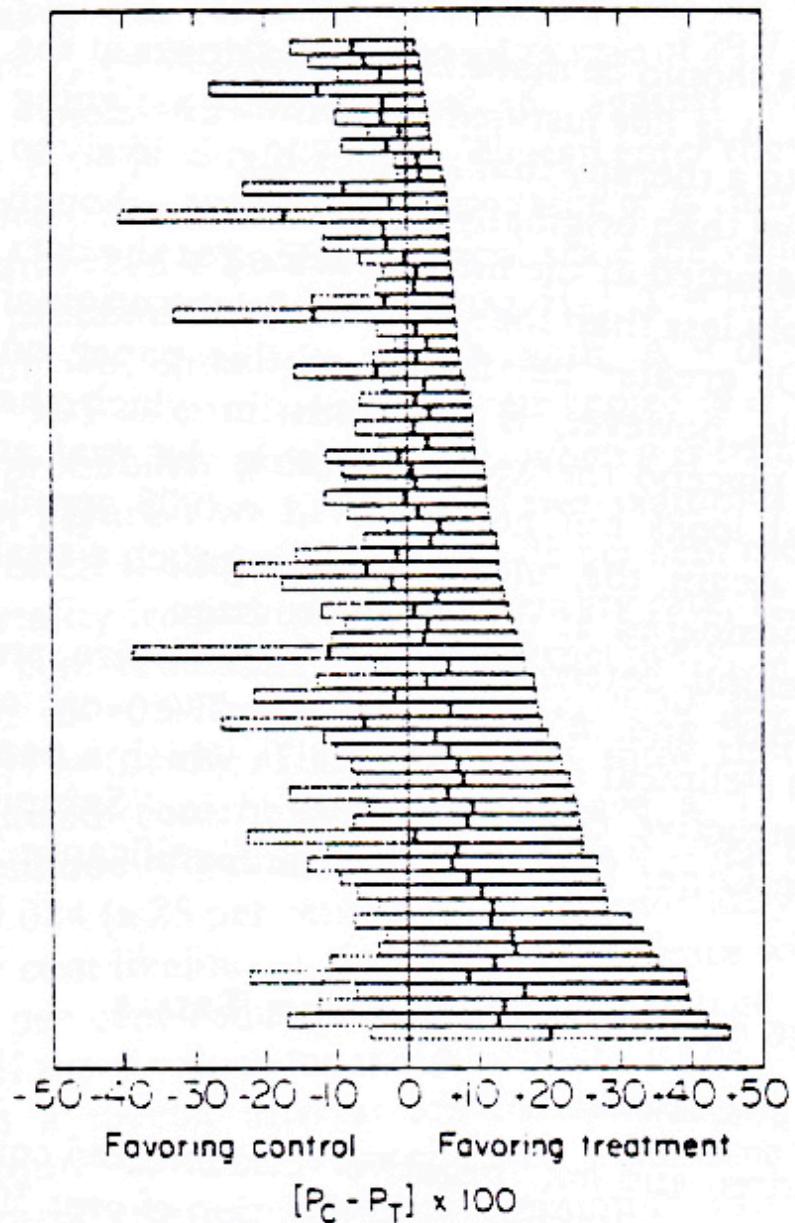
## 11.3 Simulating meta-analyses of mean standardized differences

Meta-analyses get a bit more exciting when we are using them to analyze results from multiple studies. When multiple studies are combined in a meta-analysis, effect size estimates are not simply averaged, but they are **weighed** by the **precision** of the effect size estimate, which is determined by standard error, which is in turn determined by the sample size of the study. Thus, the larger the sample size of an individual study, the more weight it gets in the meta-analysis, meaning that it has more influence on the meta-analytic effect size estimate.

One intuitive way to learn about meta-analyses is to simulate studies and meta-analyze them. The code below simulates 12 studies. There is a true effect in the simulated studies, as the difference in means in the population is 0.4 (and given the standard deviation of 1, Cohen’s  $d = 0.4$  as well). The studies vary in their sample size between 30 observations and 100 observations per condition. The meta-analysis is performed, and a forest plot is created.

```
set.seed(94)
nSims <- 12 # number of simulated studies
m1 <- 0.4 # population mean Group 1
sd1 <- 1 # standard deviation Group 1
m2 <- 0 # population mean Group 2
sd2 <- 1 # standard deviation Group 1
metadata <- data.frame(yi = numeric(0), vi = numeric(0)) # create dataframe

for (i in 1:nSims) { # for each simulated study
  n <- sample(30:100, 1) # pick a sample size per group
  x <- rnorm(n = n, mean = m1, sd = sd1)
  y <- rnorm(n = n, mean = m2, sd = sd2)
```



**Figure 2. Ninety per Cent Confidence Limits for the True Percentage Difference for the 71 Trials.**

**The vertical bar at the center of each interval indicates the observed value,  $\hat{P}_c - \hat{P}_t$ , for each trial.**

Figure 11.9: First version of a forest plot by Freiman and colleagues, 1978 (image from <https://www.jameslindlibrary.org/freiman-ja-chalmers-tc-smith-h-kuebler-rr-1978/>). 365

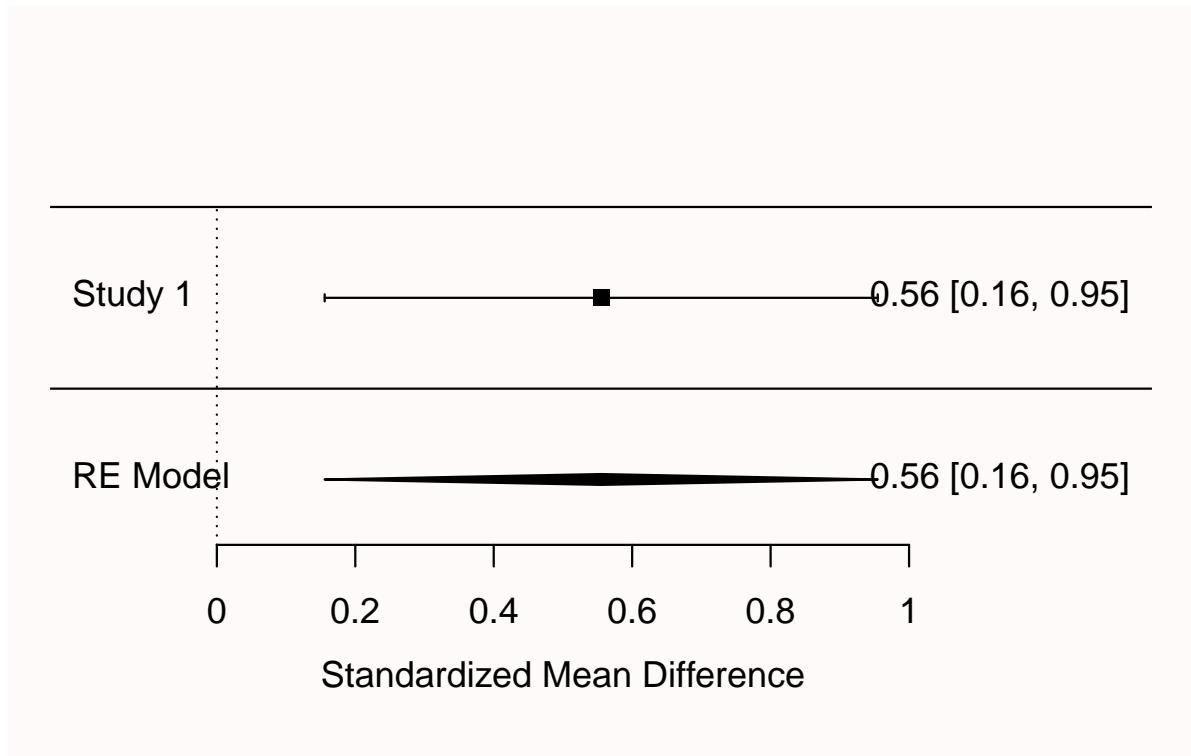


Figure 11.10: Forest plot for a single study.

```

metadata[i,1:2] <- metafor::escalc(n1i = n, n2i = n, m1i = mean(x),
                                    m2i = mean(y), sd1i = sd(x), sd2i = sd(y), measure = "SMD")
}
result <- metafor::rma(yi, vi, data = metadata, method = "FE")
par(bg = "#fffffa")
metafor::forest(result)

```

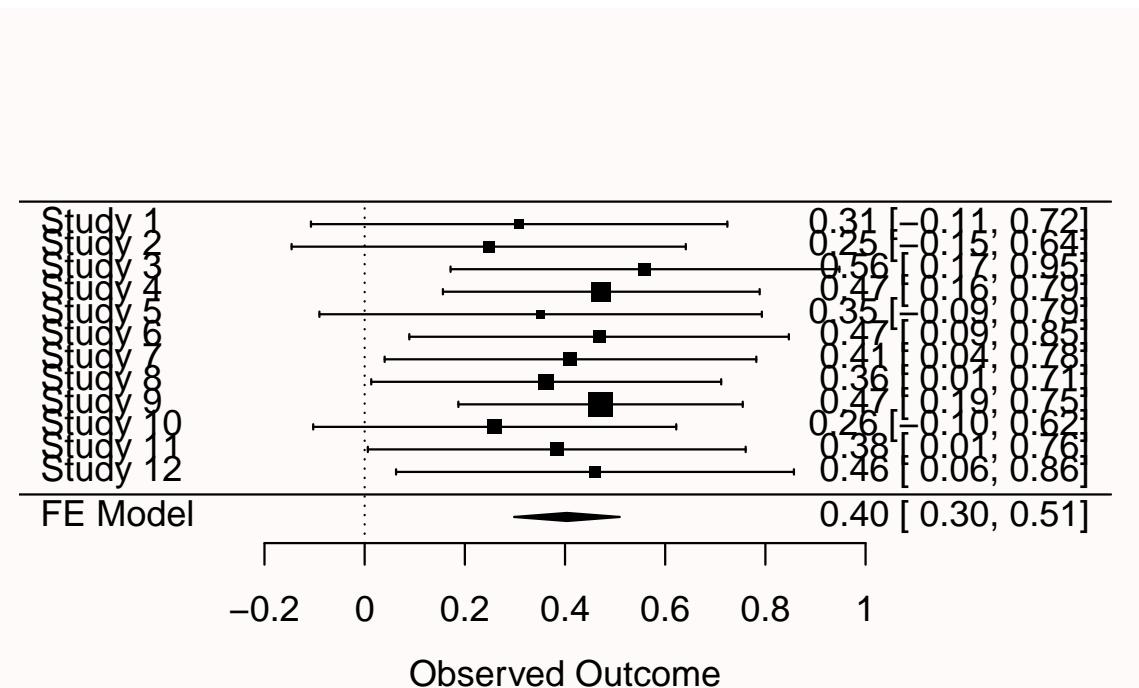


Figure 11.11: Forest plot for 12 simulated studies.

We see 12 rows, one for each study, each with their own effect size and confidence interval. If you look closely, you can see the squares that indicate the effect size estimate for each study differ in size. The larger the sample size, the bigger the square. Study 5 had a relatively small sample size, which can be seen by both the small square and the relatively wide confidence interval. Study 9 had a larger sample size, and thus a slightly larger square and narrower confidence interval. At the bottom of the graph we find the meta-analytic effect size and its confidence interval, both visualized by a diamond and numerically. The model is referred to as an **FE Model**, or **Fixed Effect (FE) model**. The alternative approach is an RE Model, or **Random Effects (RE) model** (the difference is discussed below).

You might notice that the first two studies in the meta-analysis were not statistically significant.

Take a moment to think for yourself if you would have continued this research line, after not finding an effect twice in a row. If you feel like it, run the code above several times (remove the set.seed argued used to make the simulation reproducible first, or you will get the same result each time) and see how often this happens with a population effect size and range of sample sizes in this simulation. As should be clear from discussion of mixed results in the chapter on likelihoods, it is important to think meta-analytically. In the long run, there will be situations where you will find one or two non-significant results early in a research line, even when there is a true effect.

Let's also look at the statistical results of the meta-analysis, which is a bit more interesting now that we have 12 studies:

```
Fixed-Effects Model (k = 12)

I^2 (total heterogeneity / total variability):  0.00%
H^2 (total variability / sampling variability):  0.25

Test for Heterogeneity:
Q(df = 11) = 2.7368, p-val = 0.9938

Model Results:

estimate      se     zval    pval   ci.lb   ci.ub
 0.4038  0.0538  7.5015 <.0001  0.2983  0.5093  ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see a test for **heterogeneity**, a topic we will return to [below](#). We see the model results, which in this specific simulation yielded a meta-analytic effect size estimate of 0.4. The confidence interval around the effect size estimate [0.3 ; 0.51] is much narrower than we saw before for a single study. This is because the 12 studies we simulated together have quite a large sample size, and the larger the sample size, the smaller the standard error, and thus the narrower the confidence interval is. The meta-analytic effect size estimate is statistically different from 0 ( $p < 0.0001$ ) so we can reject the null hypothesis even if we use a stringent alpha level, such as 0.001. Note that, as discussed in the chapter on sample size justification and in line with the section on justifying error rates in the chapter on error control, it seems sensible to use a much lower alpha level than 5% in meta-analyses. It is possible to set the alpha level in `metafor`, e.g. using `level = 0.999` (for an alpha level of 0.001), but this adjusts all confidence intervals, including those of the individual studies, which will mostly have used an alpha level of 0.05, so it is easier to just manually check if the test is significant at your chosen alpha level (e.g., 0.001).

## 11.4 Fixed Effect vs Random Effects

There are two possible models when performing a meta-analysis. One model, known as a fixed effect model, assumes there is one effect size that generates the data in all studies in the meta-analysis. This model assumes there is no variation between individual studies – all have exactly the same true effect size. The author of the `metafor` package we used in this chapter prefers to use the term [equal-effects model](#) instead of fixed effect model. The perfect example of this is the simulations we have done so far. We specified a single true effect in the population, and generated random samples from this population effect.

Alternatively, one can use a model where the true effect differs in some way in each individual study. We don't have a single true effect in the population, but a range of **randomly distributed** true effect sizes (hence the 'random effects' model). Studies differ in some way from each other (or some sets of studies differ from other sets), and their true effect sizes differ as well. Note the difference between a fixed effect model, and a random effects model, in that the plural 'effects' is used only in the latter. Borenstein et al (2009) state there are two reasons to use a fixed effect model: When all studies are functionally equivalent, and when your goal is *not* to generalize to other populations. This makes the random effects model generally the better choice, although some people have raised the concern that random-effects models give more weight to smaller studies, which can be more biased. By default, `metafor` will use a random effects model. We used the `method="FE"` command to explicitly ask for a fixed effect model. In the meta-analyses that we will simulate in the rest of this chapter, we will leave out this command and simulate random effects meta-analyses, as this is the better choice in many real life meta-analyses.

## 11.5 Simulating meta-analyses for dichotomous outcomes

Although meta-analyses on mean differences are very common, a meta-analysis can be performed on different effect sizes measures. To show a slightly less common example, let's simulate a meta-analysis based on odds ratios. Sometimes the main outcome in an experiment is a dichotomous variable, such as the success or failure on a task. In such study designs we can calculate risk ratios, odds ratios, or risk differences as the effect size measure. Risk differences are sometimes judged easiest to interpret, but odds ratios are most often used for a meta-analysis because they have attractive statistical properties. An **odds ratio** is a ratio of two odds. To illustrate how an odds ratio is calculated, it is useful to consider the four possible outcomes in a 2 x 2 table of outcomes:

	Success	Failure	N
Experimental	A	B	n1
Control	C	D	n2

The odds ratio is calculated as:

$$OR = \frac{AD}{BC}$$

The meta-analysis is performed on log transformed odds ratios (because log transformed odds ratios are symmetric around 1, see Borenstein et al., 2009), and thus the log of the odds ratio is used, which has a variance which is approximated by:

$$\text{Var}(\log OR) = \frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}$$

Let's assume that we train students in using a spaced learning strategy (they work through a textbook every week instead of cramming the week before the exam). Without such training, 70 out of 100 students succeed in passing the course after the first exam, but with this training, 80 out of 100 students pass.

	Success	Failure	N
Experimental	80	20	100
Control	70	30	100

The odds of passing in the experimental group is 80/20, or 4, while odds in the control condition are 70/30, or 2.333. The ratio of these two odds is then:  $4/2.333 = 1.714$ , or:

$$OR = \frac{80 \times 30}{20 \times 70} = 1.714$$

We can simulate studies with dichotomous outcomes, where we set the percentage of successes and failures in the experimental and control condition. In the script below, by default the percentage of success in the experimental condition is 70%, and in the control condition it is 50%.

```
library(metafor)
set.seed(5333)
nSims <- 12 # Number of simulated experiments

pr1 <- 0.7 # Set percentage of successes in Group 1
pr2 <- 0.5 # Set percentage of successes in Group 2

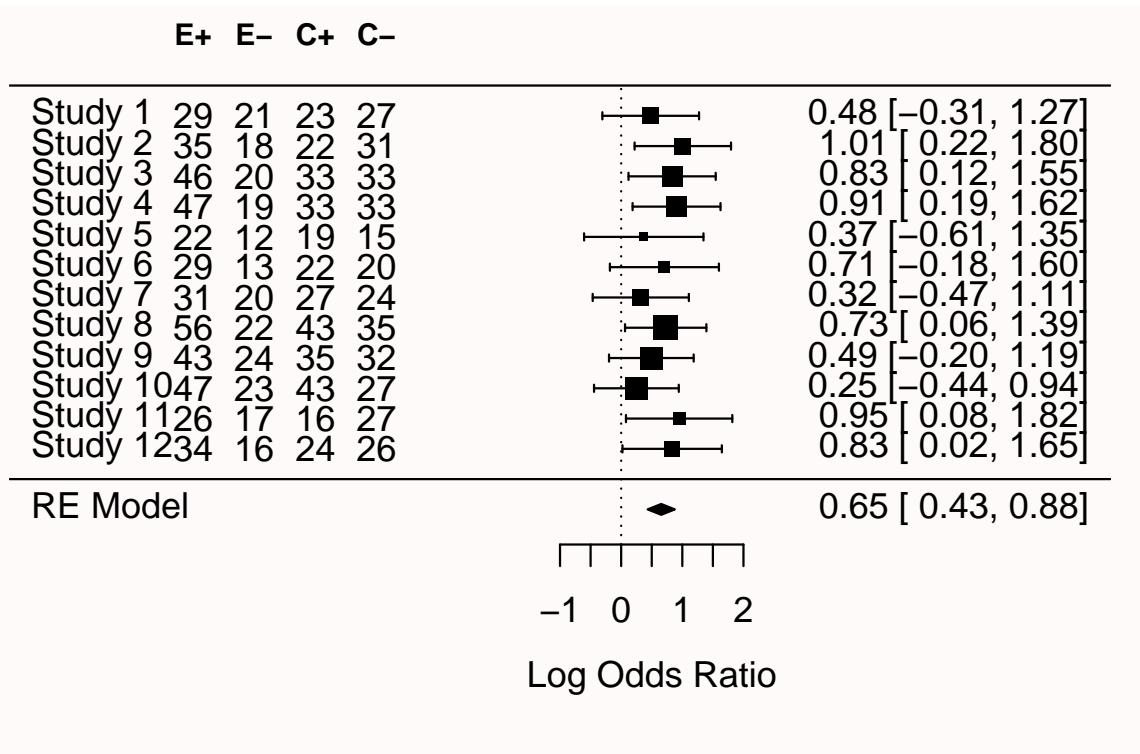
ai <- numeric(nSims) # set up empty vector for successes Group 1
bi <- numeric(nSims) # set up empty vector for failures Group 1
ci <- numeric(nSims) # set up empty vector for successes Group 2
di <- numeric(nSims) # set up empty vector for failures Group 2
```

```

for (i in 1:nSims) { # for each simulated experiment
  n <- sample(30:80, 1)
  x <- rbinom(n, 1, pr1) # participants (1 = success, 0 = failure)
  y <- rbinom(n, 1, pr2) # participants (1 = success, 0 = failure)
  ai[i] <- sum(x == 1) # Successes Group 1
  bi[i] <- sum(x == 0) # Failures Group 1
  ci[i] <- sum(y == 1) # Successes Group 2
  di[i] <- sum(y == 0) # Failures Group 2
}

# Combine data into dataframe
metadata <- cbind(ai, bi, ci, di)
# Create escalc object from metadata dataframe
metadata <- escalc(measure = "OR",
                    ai = ai, bi = bi, ci = ci, di = di,
                    data = metadata)
# Perform Meta-analysis
result <- rma(yi, vi, data = metadata)
# Create forest plot. Using ilab and ilab.xpos arguments to add counts
par(mar=c(5, 4, 0, 2))
par(bg = "#ffffafa")
forest(result,
       ilab = cbind(metadata$ai, metadata$bi, metadata$ci, metadata$di),
       xlim = c(-10, 8),
       ilab.xpos = c(-7, -6, -5, -4))
text(c(-7, -6, -5, -4), 14.7, c("E+", "E-", "C+", "C-"), font = 2, cex = .8)

```



The forest plot presents the studies and four columns of data after the study label, which contain the number of successes and failures in the experimental groups (E+ and E-), and the number of successes and failures in the control group (C+ and C-). Imagine we study the percentage of people who get a job within 6 months after a job training program, compared to a control condition. In Study 1, which had 50 participants in each condition, 29 people in the job training condition got a job within 6 months, and 21 did not get a job. In the control condition, 23 people got a job, but 27 did not. The effect size estimate for the random effects model is 0.65. Feel free to play around with the script, adjusting the number of studies, or the sample sizes in each study, to examine the effect it has on the meta-analytic effect size estimate.

We can also get the meta-analytic test results by printing the test output. We see that there was no heterogeneity in this meta-analysis. This is true (we simulated identical studies), but is highly unlikely to ever happen in real life where variation in effect sizes between studies included in a meta-analysis is a much more realistic scenario.

```
# Print result meta-analysis
result
```

```

Random-Effects Model (k = 12; tau^2 estimator: REML)

tau^2 (estimated amount of total heterogeneity): 0 (SE = 0.0645)
tau (square root of estimated tau^2 value):      0
I^2 (total heterogeneity / total variability):   0.00%
H^2 (total variability / sampling variability):  1.00

Test for Heterogeneity:
Q(df = 11) = 4.8886, p-val = 0.9364

Model Results:

estimate      se     zval    pval   ci.lb   ci.ub
0.6548  0.1132  5.7824  <.0001  0.4328  0.8767  ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

## 11.6 Heterogeneity

Although researchers often primarily use meta-analysis to compute a meta-analytic effect size estimate, and test whether this effect is statistically different from zero, **an arguably much more important use of meta-analyses is to explain variation between (sets of) studies.** This variation among (sets of) studies is referred to as **heterogeneity**. One goal of meta-analyses is not just to code effect sizes and estimate the meta-analytic effect size, but to code factors in studies that can explain heterogeneity, and examine which of these factors account for heterogeneity. This can help in theory evaluation or theory development. Tests have been developed to examine whether the studies included in a meta-analysis vary more than would be expected if the underlying true effect size in all studies was the same, and measures have been developed to quantify this variation.

If all studies have the same true population effect size, the only source of variation is random error. If there are real differences between (sets of) studies, there are two sources of variation, namely random variation from study to study *and* real differences in effect sizes in (sets of) studies.

A classical measure of heterogeneity is Cochran's *Q* statistic, which is the weighted sum of the squared differences between effect size estimates in each study, and the meta-analytic effect size estimate. The *Q* statistic can be used to test whether the absence of heterogeneity can be statistically rejected (by comparing it to the expected amount of variation, which is the degrees of freedom, *df*, or the number of studies -1, see Borenstein et al., 2009), but it can

have low power if the number of studies in the meta-analysis is small (Huedo-Medina et al., 2006).

On theoretical grounds one might argue that some heterogeneity will always happen in a meta-analysis, and therefore it is more interesting to quantify the extent to which there is heterogeneity. The  $I^2$  index aims to quantify statistical heterogeneity. It is calculated as follows:

$$I^2 = \frac{(Q - k - 1)}{Q} \times 100\%$$

where  $k$  is the number of studies (and  $k - 1$  is the degrees of freedom).  $I^2$  ranges from 0 to 100 and can be interpreted as the percentage of the total variability in a set of effect sizes that is due to heterogeneity. When  $I^2 = 0$  all variability in the effect size estimates can be explained by within-study error, and when  $I^2 = 50$  half of the total variability can be explained by true heterogeneity.  $I^2$  values of 25%, 50%, and 75% can be interpreted as low, medium, and high heterogeneity. Finally, in a random effects meta-analysis,  $\tau^2$  estimates the variance of the true effects, and  $\tau$  is the estimated standard deviation, as expressed on the same scale as the effect size. A benefit of  $\tau^2$  is that it does not depend on the precision, as  $I^2$  does, which tends to 100% if the studies included in the meta-analysis are very large (Rücker et al., 2008), but a downside is that  $\tau^2$  is more difficult to interpret (Harrer et al., 2021).

The script below simulates a similar meta-analysis to the example for standardized means above, but with a small variation. Of the 30 studies, 15 are generated based on a true mean difference of 0.2, while the other 15 studies are based on a true effect size of 0.5. Thus, in this set of studies the true effect size varies, and there is true heterogeneity. We use the `confint` function in the `metafor` package to report  $I^2$  and  $\tau^2$ , and their confidence intervals, and we see the test for heterogeneity is statistically significant. In other words, we would conclude that the meta-analytic effect size is statistically different from 0, but we would also conclude that there is unexplained variability in effect sizes in this set of studies, and the effect size is not the same for all studies in the meta-analysis.

```
library(metafor)
set.seed(3)
nSims <- 30 # Number of simulated experiments (need to be divided by 2)
metadata <- data.frame(yi = numeric(0), vi = numeric(0)) # create dataframe
true_es <- numeric(nSims) # set up empty vector for true effect sizes
study <- numeric(nSims) # set up empty vector for study numbers
group <- numeric(nSims) # set up empty vector for group

m1 <- 0.5 # population mean Group 1
sd1 <- 1 # standard deviation Group 1
m2 <- 0 # population mean Group 2
sd2 <- 1 # standard deviation Group 1
```

```

for (i in 1:(nSims/2)) {
  n <- sample(30:100, 1)
  x <- rnorm(n = n, mean = m1, sd = sd1) # simulate random normally distributed data
  y <- rnorm(n = n, mean = m2, sd = sd2) # simulate random normally distributed data
  metadata[i,1:2] <- metafor::escalc(n1i = n, n2i = n, m1i = mean(x), m2i = mean(y), sd1i = s
  true_es[i] <- paste("Study", i, "Effect = 0.5") # true effect size
  group[i] <- paste("Manipulation A") # Group label
  study[i] <- paste("Study",i) # Study
}

m1 <- 0.2 # population mean Group 1
sd1 <- 1 # standard deviation Group 1
m2 <- 0 # population mean Group 2
sd2 <- 1 # standard deviation Group 1

for (i in (nSims/2+1):nSims) { # for third quarter of each simulated study
  n <- sample(30:100, 1)
  x <- rnorm(n = n, mean = m1, sd = sd1) # simulate random normally distributed data
  y <- rnorm(n = n, mean = m2, sd = sd2) # simulate random normally distributed data
  metadata[i,1:2] <- metafor::escalc(n1i = n, n2i = n, m1i = mean(x), m2i = mean(y), sd1i = s
  true_es[i] <- paste("Study", i, "Effect = 0.2") # true effect size
  group[i] <- paste("Manipulation B") # Group label
  study[i] <- paste("Study",i) # Study
}

# Combine data into dataframe
metadata <- cbind.data.frame(metadata, true_es, study, group)
# Shuffle rows to make it difficult to see which effect size comes from which group
metadata <- metadata[sample(nrow(metadata)),]

# Perform Meta-analysis
result <- rma(yi, vi, data = metadata, slab = paste(true_es))
# Print result meta-analysis
result
confint(result) # Get confidence interval for indices of heterogeneity

```

Random-Effects Model (k = 30; tau<sup>2</sup> estimator: REML)

tau<sup>2</sup> (estimated amount of total heterogeneity): 0.0274 (SE = 0.0160)  
 tau (square root of estimated tau<sup>2</sup> value): 0.1655  
 I<sup>2</sup> (total heterogeneity / total variability): 45.43%

```
H^2 (total variability / sampling variability): 1.83
```

```
Test for Heterogeneity:
```

```
Q(df = 29) = 52.3032, p-val = 0.0050
```

```
Model Results:
```

estimate	se	zval	pval	ci.lb	ci.ub
0.3578	0.0453	7.8909	<.0001	0.2689	0.4467

\*\*\*

```
---
```

```
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

	estimate	ci.lb	ci.ub
tau^2	0.0274	0.0044	0.0717
tau	0.1655	0.0666	0.2677
I^2(%)	45.4253	11.8823	68.5417
H^2	1.8324	1.1348	3.1788

The forest plot in Figure 11.12 shows there is more variation around the meta-analytic effect size estimate than expected purely based on random variation. The plot has 2 vertical lines, one at 0.2, and one at 0.5. Of course, in a real meta-analysis we would not know what the true effect sizes of subsets are, but these lines help us to see that one subset of the effects varies randomly around 0.2, and a second subset varies randomly around 0.5.

Based on the test for heterogeneity, we can reject the null hypothesis that there is no heterogeneity in the meta-analysis. Tests for heterogeneity themselves have Type 1 and Type 2 error rates, and with a small number of studies (such as in our example,  $n = 12$ ) tests for heterogeneity can have low power. If you remove the set.seed command and run the code multiple times, you will see that the test for heterogeneity will often not be significant, even though there is true heterogeneity in the simulation. In large meta-analyses, power can be so high that the test always yields a  $p$ -value small enough to reject the null hypothesis, but then it is important to look at the  $I^2$  estimate.

Recently there has been considerable attention to the possibility that effect sizes within research lines have substantial heterogeneity (Bryan et al., 2021). Large heterogeneity can impact the power of studies, and therefore has consequences for how studies are planned (Kenny & Judd, 2019). Although heterogeneity seems to be low in direct replication studies (Olsson-Collentine et al., 2020) it is high in most meta-analyses, which has been argued to reflect a lack of understanding of the effects in those research lines (Linden & Hönekopp, 2021).

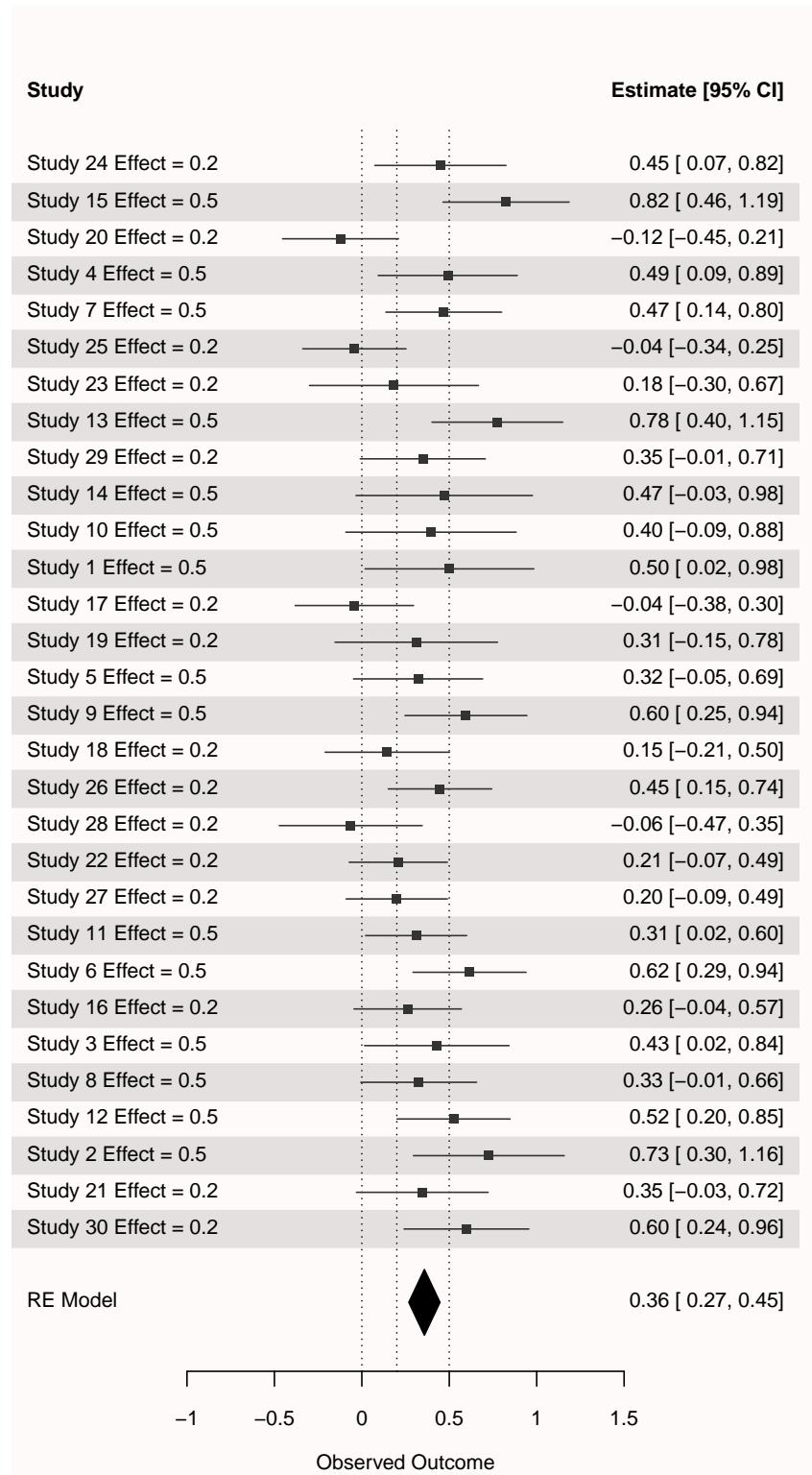


Figure 11.12: Forest plot of 30 studies with 2 subsets of studies, one set with a true effect size of 0.2, and one set with a true effect size of 0.5.

## 11.7 Exploring heterogeneity through subgroup analyses

Let's imagine that while coding our meta-analysis we noticed that in the total set of studies two different manipulations were used. Manipulation A was stronger, and therefore we would theoretically expect a stronger effect, while manipulation B was more subtle. We have coded which manipulation was used in each of the studies included in our meta-analysis. This allows us to explore whether the heterogeneity observed above can be explained by the type of manipulation. In other words, we are testing if the effect size is moderated by the type of manipulation. This is a sub-group analysis. It is conceptually very similar to an ANOVA where we test whether there are differences between groups. We see the test of moderators yields a statistically significant result, which means we can reject the null hypothesis that the two groups do not differ in their effect size. Furthermore, there is no statistically significant amount of heterogeneity left after splitting up the studies in these two groups.

```
# Based on https://www.metafor-project.org/doku.php/tips:comp_two_independent_estimates

# We take original dataset and run a meta-regression using the "mods" argument
# This pools the estimates of tau, but this is often a good approach
rma(yi, vi, mods = ~ group, data = metadata, digits = 3)
```

```
Mixed-Effects Model (k = 30; tau^2 estimator: REML)

tau^2 (estimated amount of residual heterogeneity):      0.006 (SE = 0.010)
tau (square root of estimated tau^2 value):             0.081
I^2 (residual heterogeneity / unaccounted variability): 16.42%
H^2 (unaccounted variability / sampling variability):   1.20
R^2 (amount of heterogeneity accounted for):            76.32%

Test for Residual Heterogeneity:
QE(df = 28) = 31.550, p-val = 0.293

Test of Moderators (coefficient 2):
QM(df = 1) = 17.231, p-val < .001

Model Results:

           estimate      se    zval   pval    ci.lb    ci.ub
intrcpt       0.514  0.053  9.643  <.001    0.410    0.618  ***
groupManipulation B -0.303  0.073 -4.151  <.001   -0.446   -0.160  ***
---
***
```

```
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

If we plot the two subgroups, we see in Figure 11.13 that effect sizes vary randomly across the true effect sizes. Again, in real meta-analyses these true effect sizes would not be known.

## 11.8 Strengths and weaknesses of meta-analysis

The conclusions from meta-analyses have been debated from the very first meta-analyses that were performed. It's ironic that, as far as I can find, the 'garbage in-garbage out' criticism of meta-analysis originates from Eysenck (1978) because although it is valid criticism, Eysenck himself published literal garbage, as he was found [guilty of scientific misconduct](#), which has led to a large number of [retractions](#) and expressions of concern. Eysenck wrote about a meta-analysis that yielded results he did not like:

The most surprising feature of Smith and Glass's (1977) exercise in mega-silliness is their advocacy of low standards of judgment. More, they advocate and practice the abandonment of critical judgments of any kind. A mass of reports—good, bad, and indifferent—are fed into the computer in the hope that people will cease caring about the quality of the material on which the conclusions are based. If their abandonment of scholarship were to be taken seriously, a daunting but improbable likelihood, it would mark the beginning of a passage into the dark age of scientific psychology. The notion that one can distill scientific knowledge from a compilation of studies mostly of poor design, relying on subjective, unvalidated, and certainly unreliable clinical judgments, and dissimilar with respect to nearly all the vital parameters, dies hard. This article, it is to be hoped, is the final death rattle of such hopes. "Garbage in—garbage out" is a well-known axiom of computer specialists; it applies here with equal force.

The problem of 'garbage in, garbage out' remains one of the most common, and difficult to deal with, criticisms of meta-analysis. It is true that a meta-analysis cannot turn low quality data into a good effect size estimate, or highly heterogeneous effect sizes into a useful estimate of an effect size that generalizes to all studies included in the meta-analysis. The decision of which studies to include in a meta-analysis is a difficult one, and often leads to disagreements in the conclusions of meta-analyses performed on the same set of studies (C. J. Ferguson, 2014; Goodyear-Smith et al., 2012). Finally, meta-analyses can be biased, in the same way individual studies are biased, which is a topic explored in more detail in the chapter on [bias detection](#).

A strength of meta-analysis is that combining highly similar studies into a single analysis increases the statistical power of the test, as well as the accuracy of the effect size estimate. Whenever it is not possible, or it is efficient, to perform studies with a large number of observations in each study, an unbiased meta-analysis can provide better statistical inferences.

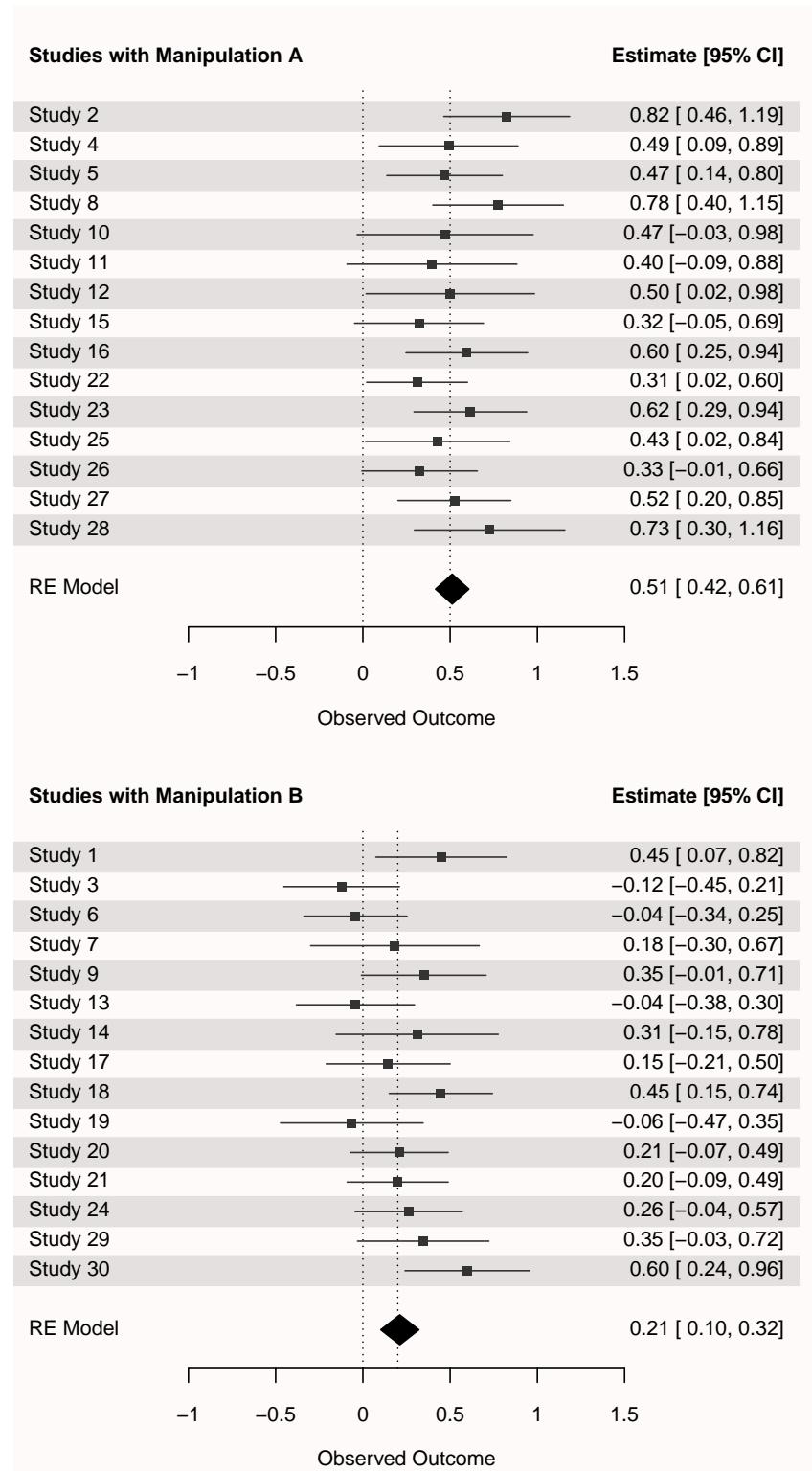


Figure 11.13: Two forest plots for the 2 different subgroups with Manipulation A or B.

Furthermore, including an **internal meta-analysis** in a multi-study paper (when all studies are sufficiently similar) can be a way to reduce the file-drawer problem, by allowing researchers to publish mixed results. At the same time, researchers have raised the concern that if researchers selectively report studies when they perform an internal meta-analysis they simply increase the flexibility in the data analysis, and are more likely to erroneously claim support for their hypothesis (Vosgerau et al., 2019). Researchers should publish all well-designed studies they have performed in a research line, and if the studies are similar and unbiased, a meta-analysis will improve inferences. At the same time, the result of a meta-analysis may be biased, and should not be interpreted as the final answer. For this reason, an analysis of the heterogeneity of effect size estimates, and the use of statistical techniques to detect bias, are an essential part of any meta-analysis.

## **11.9 Which results should you report to be included in a future meta-analysis?**

It would be a useful educational exercise for any researcher who publishes quantitative studies to code a dozen studies for a meta-analysis. A notorious problem when performing a meta-analysis is that researchers do not report all the results a meta-analyst needs in order to include the study in their meta-analysis. Sometimes the original researcher can be contacted and the missing information can be provided, but as every single study is just a data-point in a future meta-analysis, it is best to report all the required results to be included in a future meta-analysis.

The single best approach to guarantee that all required information for a future meta-analysis is available to meta-analysts is to share the (anonymized) data and analysis code with the manuscript. This will enable meta-analysts to compute any statistical information they need. Access to individual observations allows meta-analysts to perform analyses on subgroups, and makes it possible to perform more advanced statistical tests (Stewart & Tierney, 2002). Finally, access to the raw data, instead of only access to the summary statistics, makes it easier to find flawed individual studies that should not be included in a meta-analysis (Lawrence et al., 2021). As open data becomes the norm, efforts to standardize measures and develop specifications for datasets will facilitate the availability of raw data as input for meta-analyses. This will also facilitate the re-use of data, and allow researchers to perform meta-analyses unrelated to the main research question. If you want to share the raw data you will collect, make sure you address this in your [informed consent form](#).

When summarizing data in a scientific article, report the number of observations associated with each statistical test. Most articles will mention the total sample size, but if some observations are removed while cleaning the data, also report the final number of observations included in a test. When a statistical test is based on multiple conditions (e.g., a *t*-test), report the sample size in each independent group. If this information is missing, meta-analysts will often have to assume that the total number of observations is distributed equally across conditions,

which is not always correct. Report full test results for significant and non-significant results (e.g., never write  $F < 1$ ,  $ns$ ). Write out the full test result, including an effect size estimate, regardless of the  $p$ -value, as non-significant results are especially important to be included in meta-analyses. When reporting effect sizes, report how they were computed (e.g., when reporting standardized mean differences, did you compute Cohen's  $d$  or Hedges'  $g$ ). Report exact  $p$ -values for each test, or the full test statistics that can be used to recompute the  $p$ -value. Report means and standard deviations for each group of observations, and for within-subject designs, report the correlation between dependent variables (which is currently almost always never reported, but is needed to compute Cohen's  $d_{av}$  and perform simulation based power analyses based on the predicted data pattern). It might be useful to use a table to summarize all statistical tests if many tests are reported, but the raw data cannot be shared (for example in the supplemental material).

## 11.10 Improving the reproducibility of meta-analyses

Although meta-analyses do not provide definitive conclusions, they are typically interpreted as state-of-the-art empirical knowledge about a specific effect or research area. Large-scale meta-analyses often accumulate a massive number of citations and influence future research and theory development. It is therefore essential that published meta-analyses are of the highest possible quality.

At the same time, the conclusions from meta-analyses are often open for debate and are subject to change as new data becomes available. We recently proposed practical recommendations to increase the reproducibility of meta-analyses to facilitate quality control, improve reporting guidelines, allow researchers to re-analyze meta-analyses based on alternative inclusion criteria, and future-proof meta-analyses by making sure the collected meta-analytic data is shared so that continuously accumulating meta-analyses can be performed, and so that novel statistical techniques can be applied on the collected data as they become available (Daniël Lakens et al., 2016). The need for the improvement in reproducibility of meta-analysis is clear - a recent review of 150 meta-analyses in Psychological Bulletin revealed that only 1 meta-analysis shared the statistical code (Polanin et al., 2020). This is unacceptable in the current day and age. In addition to inspecting how well your meta-analysis adheres to the [JARS Quantitative Meta-Analysis Reporting Standards](#), following the recommendations summarized in Table 11.4 should substantially improve the state-of-the-art in meta-analyses.

Table 11.4: Six practical recommendations to improve the quality and reproducibility of meta-analyses.

What?	How?
-------	------

Facilitate cumulative science	Disclose all meta-analytic data (effect sizes, sample sizes for each condition, test statistics and degrees of freedom, means, standard deviations, and correlations between dependent observations) for each data point. Quote relevant text from studies that describe the meta-analytic data to prevent confusion, such as when one effect size is selected from a large number of tests reported in a study. When analyzing subgroups, include quotes from the original study that underlie this classification, and specify any subjective decisions.
Facilitate quality control	Specify which effect size calculations are used and which assumptions are made for missing data (e.g., assuming equal sample sizes in each condition, imputed values for unreported effect sizes), if necessary for each effect size extracted from the literature. Specify who extracted and coded the data, knowing it is preferable that two researchers independently extract effect sizes from the literature.
Use reporting guidelines	A minimal requirement when reporting meta-analyses is to adhere to one of the reporting standards (e.g., PRISMA). The reporting guidelines ask authors of meta-analyses to report essential information that should be made available either in the main text of the article, or by providing a completed checklist as supplementary material during review and after publication.
Preregister	Whenever possible, pre-register the meta-analysis research protocol (e.g., using PROSPERO) to distinguish between confirmatory and exploratory analyses. Perform a prospective meta-analysis where possible.
Facilitate reproducibility	Allow others to re-analyze the data to examine how sensitive the results are to subjective choices such as inclusion criteria. Always include a link to data files that can be directly analyzed with statistical software, either by providing completely reproducible scripts containing both the data and the reported analyses in free software (e.g., R), or at the very minimum a spreadsheet that contains all meta-analytic data that can easily analyzed in any statistical program.
Recruit expertise	Consider consulting a librarian before you start the literature search, and a statistician before coding the effect sizes, for advice on how to make the literature search and effect size calculations reproducible.

For another open educational resource on meta-analysis in R, see [Doing Meta-Analysis in R](#).

## 11.11 Test Yourself

**Q1:** What is true about the standard deviation of the sample, and the standard deviation of the mean (or the standard error)?

- (A) As the sample size increases, the standard deviation of the sample becomes smaller, and the standard deviation of the mean (or standard error) becomes smaller.
- (B) As the sample size increases, the standard deviation of the sample becomes more accurate, and the standard deviation of the mean (or standard error) becomes smaller.
- (C) As the sample size increases, the standard deviation of the sample becomes more smaller, and the standard deviation of the mean (or standard error) becomes more accurate.
- (D) As the sample size increases, the standard deviation of the sample becomes more accurate, and the standard deviation of the mean (or standard error) becomes more accurate.

**Q2:** If we would perform a meta-analysis by just averaging all observed effect sizes, an effect size of  $d = 0.7$  from a small study with 20 observations would influence the meta-analytic effect size estimate just as much as a  $d = 0.3$  from a study with 2000 observations. How is the meta-analytic effect size estimate computed instead?

- (A) Effect sizes estimates from small studies undergo a small study correction before being included.
- (B) Effect size estimates from small studies are ignored when computing a meta-analytic effect size estimate.
- (C) Effect sizes are weighed based on the precision of their estimate, determined by the standard error.
- (D) Effect sizes are weighed based on how close they are to the meta-analytic effect size estimate, with studies further removed receiving less weight.

**Q3:** The size of the squares indicating effect sizes in a forest plot vary as a function of the:

- (A) Power of the study
- (B) Size of the effect
- (C) Sample size
- (D) Type 1 error rate

**Q4:** One can compute a ‘fixed effect model’ or a ‘random effects model’ when performing a meta-analysis on studies in the scientific literature. Which statement is true?

- (A) It is generally recommended to compute a **fixed effect** model, mainly because not all studies included in a meta-analysis will be functionally similar.
- (B) It is generally recommended to compute a **random effects** model, mainly because not all studies included in a meta-analysis will be functionally similar.
- (C) It is generally recommended to compute a **fixed effect** model, as this reduces the heterogeneity in the set of studies.
- (D) It is generally recommended to compute a **random effects** model, as this reduces the heterogeneity in the set of studies.

**Q5:** When there is no heterogeneity in the effect size estimates included in a meta-analysis, a fixed effect and random effects model will yield similar conclusions. If there is variability in the effect size estimates, the two models can yield different results. Below, we see two forest plots based on the same 5 simulated studies. The top plot is based on a random effects meta-analysis, the bottom plot based on a fixed effect meta-analysis. A random effects meta-analysis incorporates uncertainty about the variability of effect size estimates into the final meta-analytic estimate. How does this translate into a difference between the two plots?

- (A) There is no difference in the meta-analytic effect size estimate between the plots, as each effect size estimate from the 5 studies is identical.
- (B) The effect size in the random effects model is identical to the estimate from the fixed effect model, but the confidence interval is larger.
- (C) The effect size in the random effects model is identical to the estimate from the fixed effect model, but the confidence interval is smaller.
- (D) The effect size in the random effects model is larger than the estimate from the fixed effect model, as it incorporates additional uncertainty about bias in the

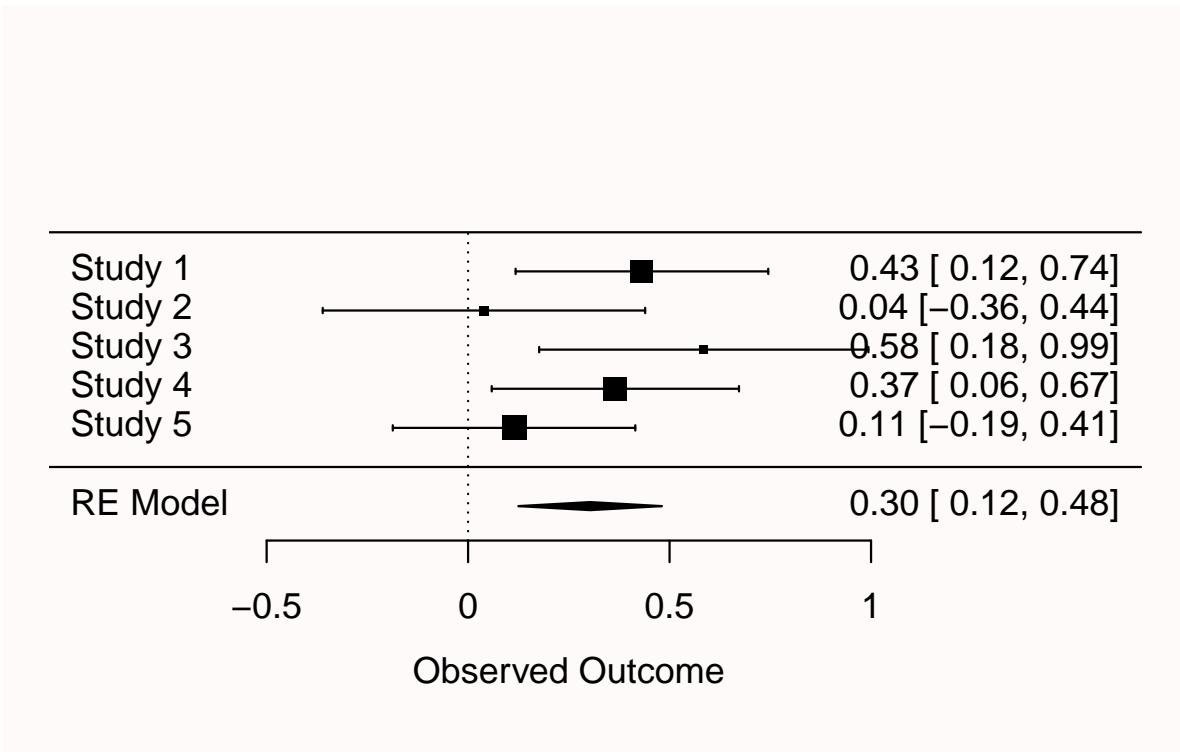


Figure 11.14: Simulated studies under a random effects model.

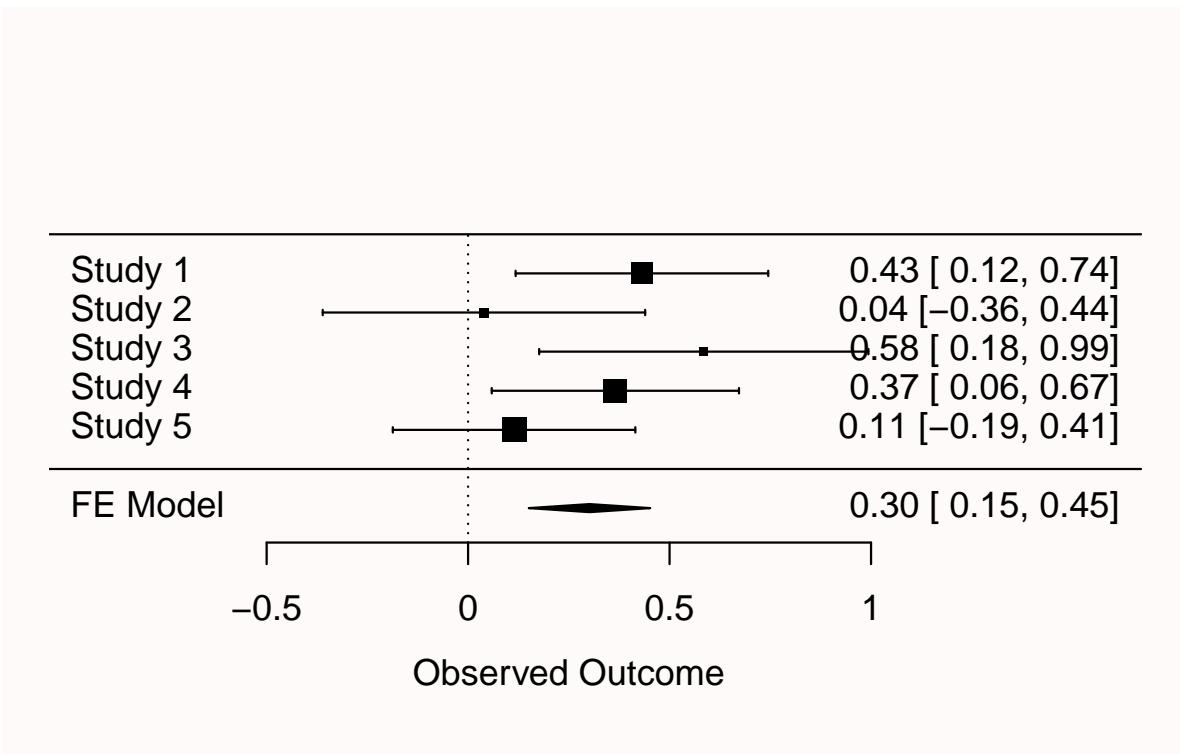


Figure 11.15: Simulated studies under a fixed effect model.

estimate.

**Q6:** Which statement is true about the two measures of heterogeneity discussed above, Cochran's  $Q$  and  $I^2$ ?

- (A) Cochran's  $Q$  relies on a hypothesis testing approach to detecting heterogeneity, and with few studies, it can have low power.  $I^2$  relies on an estimation approach, and with few studies, it can have large uncertainty.
- (B) Cochran's  $Q$  relies on an estimation approach, and with few studies, it can have large uncertainty.  $I^2$  relies on a hypothesis testing approach to detecting heterogeneity, and with few studies, it can have low power.

**Q7:** Researchers who perform very similar studies in a research line can combine all studies (whether they all yield statistically significant results, or not) into an internal meta-analysis, combining the effect sizes into a meta-analytic estimate. What is a strength of this approach, and what is a risk?

- (A) A strength is an internal meta-analysis can reduce the Type 1 error rate when multiple studies have been performed, each with their own 5% alpha level, but a weakness is that by selectively including studies in an internal meta-analysis, researcher have additional flexibility to  $p$ -hack.
- (B) A strength is an internal meta-analysis can reduce the Type 1 error rate when multiple studies have been performed, each with their own 5% alpha level, but a weakness is that the effect size estimate might be biased compared to the estimates from the single studies, especially when there is heterogeneity.
- (C) A strength is an internal meta-analysis can prevent publication bias by providing a way to report all results (including non-significant results), but a weakness is that by selectively including studies in an internal meta-analysis, researcher have additional flexibility to  $p$ -hack.
- (D) A strength is an internal meta-analysis can prevent publication bias by providing a way to report all results (including non-significant results), but a weakness is that the effect size estimate might be biased compared to the estimates from the single studies, especially when there is heterogeneity.

**Q8:** What is the best way to guarantee the statistical results in a meta-analysis are computationally reproducible? Choose the best answer.

- (A) Use open source software, such as `metafor` for R, share the analysis data, and share the analysis code, so that anyone can run the code on the data.
- (B) Use commercial software, such as 'Comprehensive Meta-Analysis', that although it does not allow you to export the data or the analysis code, allows you to share a picture of a forest plot.

### **11.11.1 Open Questions**

1. What is the difference between the standard deviation and the standard error, and what happens to each as the sample size increases?
2. What is an internal meta-analysis?
3. If you analyze only a single study, a *t*-test and a meta-analysis differ slightly in the results. Why?
4. What are the black squares in a forest plot, and what are the horizontal lines through each black square?
5. Effect sizes in a meta-analysis are not simply averaged. Why not, and how are they combined instead?
6. What is the difference between a fixed effect and random effects meta-analysis?
7. What is heterogeneity in a meta-analysis, and why is it interesting?
8. What is the problem of 'garbage in, garbage out'?

## 12 Bias detection

Diagoras, who is called the atheist, being at Samothrace, one of his friends showed him several pictures of people who had endured very dangerous storms; “See,” says he, “you who deny a providence, how many have been saved by their prayers to the Gods.” “Ay,” says Diagoras, “I see those who were saved, but where are those painted who were shipwrecked?” *The Tusculanae Disputationes, Cicero, 45 BC.*

Bias can be introduced throughout the research process. It is useful to prevent this or to detect it. Some researchers recommend a skeptical attitude towards any claim you read in the scientific literature. For example, the philosopher of science Deborah Mayo (2018) writes: “Confronted with the statistical news flash of the day, your first question is: Are the results due to selective reporting, cherry picking, or any number of other similar ruses?”. You might not make yourself very popular if this is the first question you ask a speaker at the next scientific conference you are attending, but at the same time it would be naïve to ignore the fact that researchers more or less intentionally introduce bias into their claims.

At the most extreme end of practices that introduce bias into scientific research is **research misconduct**: Making up data or results, or changing or omitting data or results such that the research isn’t accurately represented in the research record. For example, [Andrew Wakefield](#) authored a fraudulent paper in 1998 that claimed a link between the measles, mumps, and rubella (MMR) vaccine and autism. It was retracted in 2010, but only after it caused damage to trust in vaccines among some parts of the general population. Another example from psychology concerned a study by [James Vicary](#) on subliminal priming. He claimed to have found that by flashing ‘EAT POPCORN’ and ‘DRINK COCA-COLA’ subliminally during a movie screen in a cinema, the sales of popcorn and Coca-Cola had increased with 57.5 and 18.1 percent, respectively. However, it was later found that Vicary most likely committed scientific fraud, as there was no evidence that the study was ever performed (S. Rogers, 1992–1993). The website Retraction Watch maintains a [database](#) that tracks reasons why scientific papers are retracted, including data fabrication. It is unknown how often data fabrication occurs in practice, but as discussed in the chapter on [research integrity](#), we should expect that at least a small percentage of scientists have fabricated, falsified or modified data or results at least once.

A different category of mistakes are statistical reporting errors, which range from reporting incorrect degrees of freedom, to reporting  $p = 0.056$  as  $p < 0.05$  (Nuijten et al., 2015). Although we should do our best to prevent errors, everyone makes them, and data and code sharing become more common, it will become easier to detect errors in the work of other researchers.



Figure 12.1: Scene in The Dropout about the company Theranos that falsely claimed to have devices that could perform blood tests on very small amounts of blood. In the scene, two whistleblowers confront their bosses when they are pressured to remove datapoints that do not show the desired results.

As Dorothy Bishop (2018) writes: “As open science becomes increasingly the norm, we will find that everyone is fallible. The reputations of scientists will depend not on whether there are flaws in their research, but on how they respond when those flaws are noted.”

[Statcheck](#) is software that automatically extracts statistics from articles and recomputes their *p*-values, as long as statistics are reported following guidelines from the American Psychological Association (APA). It checks if the reported statistics are internally consistent: Given the test statistics and degrees of freedom, is the reported *p*-value accurate? If it is, that makes it less likely that you have made a mistake (although it does not prevent coherent mistakes!) and if it is not, you should check if all the information in your statistical test is accurate. Statcheck is not perfect, and it will make Type 1 errors where it flags something as an error when it actually is not, but it is an easy to use tool to check your articles before you submit them for publication, and early meta-scientific work suggests it can reduce statistical reporting errors (Nuijten & Wicherts, 2023).

Some inconsistencies in data are less easy to automatically detect, but can be identified manually. For example, N. J. L. Brown & Heathers (2017) show that many papers report means that are not possible given the sample size (known as the [GRIM test](#)). For example, Matti Heino noticed in a [blog post](#) that three of the reported means in the table in a classic study by Festinger and Carlsmith are mathematically impossible. With 20 observations per condition, and a scale from -5 to 5, all means should end in a multiple of 1/20, or 0.05. The three means

ending in X.X8 or X.X2 are not consistent with the reported sample size and scale. Of course, such inconsistencies can be due to failing to report that there was missing data for some of the questions, but the GRIM test has also been used to uncover [scientific misconduct](#).

#### AVERAGE RATINGS ON INTERVIEW QUESTIONS FOR EACH CONDITION

Question on Interview	Experimental Condition		
	Control (N = 20)	One Dollar (N = 20)	Twenty Dollars (N = 20)
How enjoyable tasks were (rated from -5 to +5)	-.45	+1.35	-.05
How much they learned (rated from 0 to 10)	3.08	2.80	3.15
Scientific importance (rated from 0 to 10)	5.60	6.45	5.18
Participate in similar exp. (rated from -5 to +5)	-.62	+1.20	-.25

Figure 12.2: Screenshot of the table reporting the main results from Festinger and Carlsmith, 1959.

## 12.1 Publication bias

Publication bias is one of the biggest challenges that science faces. **Publication bias** is the practice of selectively submitting and publishing scientific research, often based on whether or not the results are ‘statistically significant’ or not. The scientific literature is dominated by these statistically significant results. In 1959 Sterling (1959) counted how many hypothesis tests in 4 journals in psychology yielded significant results, and found that 286 out of 294 articles examined (i.e., 97%) rejected the null hypothesis with a statistically significant result. A replication by Bozarth and Roberts (1972) yielded an estimate of 94%.

At the same time, we know that many studies researchers perform do not yield significant results. When scientists only have access to significant results, but not to all results, they are lacking a complete overview of the evidence for a hypothesis (Smart, 1964). In extreme cases, selective reporting can lead to a situation where there are hundreds of statistically significant results in the published literature, but no true effect because there are even more

non-significant studies that are not shared. This is known as the **file-drawer problem**, when non-significant results are hidden away in file-drawers (or nowadays, folders on your computer) and not available to the scientific community. Such bias is most likely to impact the main hypothesis test a researcher reports, as whether or not they have an interesting story to tell often depends on the result of this one test. Bias is less likely to impact papers without a main hypothesis that needs to be significant to for the narrative in the manuscript, for example when researchers just describe effect sizes across a large correlation table of tests that have been performed. Every scientist should work towards solving publication bias, because it is extremely difficult to learn what is likely to be true as long as scientists do not share all their results. Greenwald (1975) even considers selectively reporting only significant results an ethical violation.

**First, it is a truly gross ethical violation for a researcher to suppress reporting of difficult-to-explain or embarrassing data in order to present a neat and attractive package to a journal editor. Second, it is to be hoped that journal editors will base publication decisions on criteria of importance and methodological soundness, uninfluenced by whether a result supports or rejects a null hypothesis.**

Figure 12.3: Screenshot of the final sentences of Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. Psychological Bulletin, 82(1), 1–20.

Publication bias can only be fixed by making all your research results available to fellow scientists, irrespective of the  $p$ -value of the main hypothesis test. Registered Reports are one way to combat publication bias, as this type of scientific article is reviewed based on the introduction, method, and statistical analysis plan, before the data is collected (Chambers & Tzavella, 2022; Nosek & Lakens, 2014). After peer review by experts in the field, who might suggest improvements to the design and analysis, the article can get an ‘in principle acceptance’, which means that as long as the research plan is followed, the article will be published, regardless of the results. This should facilitate the publication of null results, and as shown in Figure 12.4, an analysis of the first published Registered Reports in psychology revealed that 31 out of 71 (44%) articles observed positive results, compared to 146 out of 152 (96%) of comparable standard scientific articles published during the same time period

(Scheel, Schijen, et al., 2021).

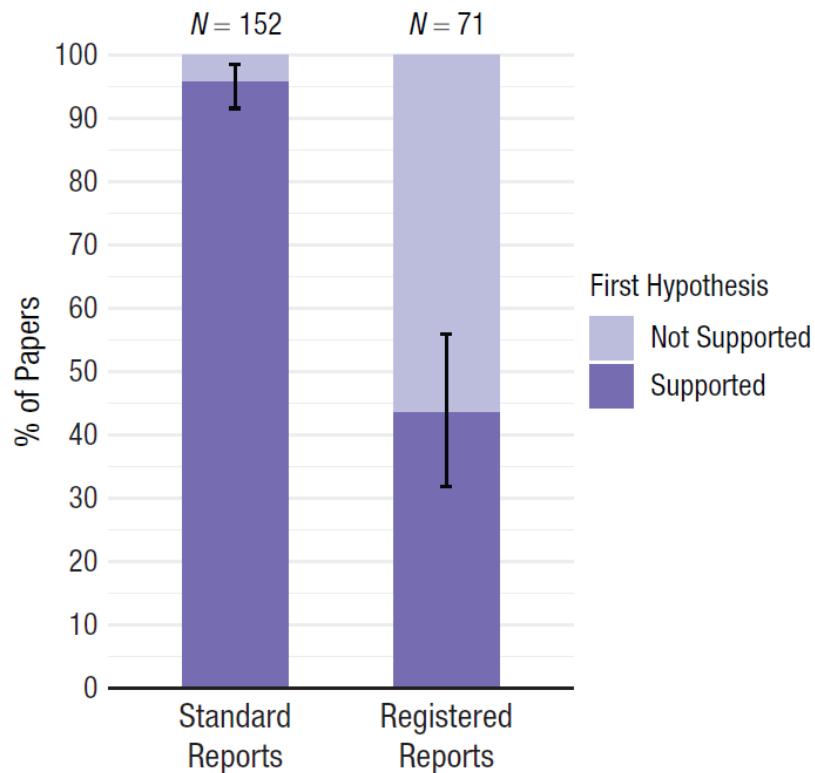


Figure 12.4: Positive result rates for standard reports and Registered Reports. Error bars indicate 95% confidence intervals around the observed positive result rate.

In the past, Registered Reports did not exist, and scientists did not share all results (Franco et al., 2014; Greenwald, 1975; Sterling, 1959), and as a consequence, we have to try to detect the extent to which publication bias impacts our ability to accurately evaluate the literature. Meta-analyses should always carefully examine the impact of publication bias on the meta-analytic effect size estimate - even though only an estimated 57% of meta-analyses published in Psychological Bulletin between 1990 to 2017 report that they assessed publication bias (Polanin et al., 2020). In more recent meta-analyses published in educational research, 82% used bias detection tests, but the methods used were typically far from the state-of-the-art (Ropovik et al., 2021). Several techniques to detect publication bias have been developed, and this continues to be a very active field of research. All techniques are based on specific assumptions, which you should consider before applying a test (Carter et al., 2019). There is no silver bullet: None of these techniques can fix publication bias. None of them can tell you with certainty what the true meta-analytic effect size is corrected for publication bias. The best these methods can do is detect publication bias caused by specific mechanisms, under

specific conditions. Publication bias can be detected, but it cannot be corrected.

In the chapter on [likelihoods](#) we saw how mixed results are to be expected, and can be strong evidence for the alternative hypothesis. It is not only the case that mixed results should be expected, but exclusively observing statistically significant results, especially when the statistical power is low, is very surprising. With the commonly used lower limit for statistical power of 80%, we can expect a non-significant result in one out of five studies when there is a true effect. Some researchers have pointed out that *not* finding mixed results can be very unlikely (or ‘too good to be true’) in a set of studies (Francis, 2014; Schimmack, 2012). We don’t have a very good feeling for what real patterns of studies look like, because we are continuously exposed to a scientific literature that does not reflect reality. Almost all multiple study papers in the scientific literature present only statistically significant results, even though this is unlikely.

The [online Shiny app we used to compute binomial likelihoods](#) displays, if you scroll to the bottom of the page, binomial probabilities to find multiple significant findings given a specific assumption about the power of the tests. Francis (2014) used these binomial likelihoods to calculate the test of excessive significance (Ioannidis & Trikalinos, 2007) for 44 articles published in the journal Psychological Science between 2009 and 2012 that contained four studies or more. He found that for 36 of these articles, the likelihood of observing four significant results, given the average power computed based on the observed effect sizes, was less than 10%. Given his choice of an alpha level of 0.10, this binomial probability is a hypothesis test, and allows the claims (at a 10% alpha level) that whenever the binomial probability of the number of statistically significant results is lower than 10%, the data is surprising, and we can reject the hypothesis that this is an unbiased set of studies. In other words, it is unlikely that this many significant results would be observed, suggesting that publication bias or other selection effects have played a role in these articles.

One of these 44 articles had been co-authored by myself (Jostmann et al., 2009). At this time, I knew little about statistical power and publication bias, and being accused of improper scientific conduct was stressful. And yet, the accusations were correct - we had selectively reported results, and selectively reported analyses that worked. Having received virtually no training on this topic, we educated ourselves, and uploaded an unpublished study to the website psychfiledrawer.org (which no longer exists) to share our filedrawer. Some years later, we assisted when Many Labs 3 included one of the studies we had published in the set of studies they were replicating (Ebersole et al., 2016), and when a null result was observed, we wrote “We have had to conclude that there is actually no reliable evidence for the effect” (Jostmann et al., 2016). I hope this educational materials prevents others from making a fool of themselves as we did.

## 12.2 Bias detection in meta-analysis

New methods to detect publication bias are continuously developed, and old methods become outdated (even though you can still see them appear in meta-analyses). One outdated method is known as **fail-safe N**. The idea was to calculate the number of non-significant results one would need to have in file-drawers before an observed meta-analytic effect size estimate would no longer be statistically different from 0. It is [no longer recommended](#), and Becker (2005) writes “Given the other approaches that now exist for dealing with publication bias, the failsafe N should be abandoned in favor of other, more informative analyses”. Currently, the only use fail-safe N has is as a tool to identify meta-analyses that are not state-of-the-art.

Before we can explain a second method (Trim-and-Fill), it’s useful to explain a common way to visualize meta-analyses, known as a **funnel plot**. In a funnel plot, the x-axis is used to plot the effect size of each study, and the y-axis is used to plot the ‘precision’ of each effect size (typically, the standard error of each effect size estimate). The larger the number of observations in a study, the more precise the effect size estimate, the smaller the standard error, and thus the higher up in the funnel plot the study will be. An infinitely precise study (with a standard error of 0) would be at the top of y-axis.

The script below simulates meta-analyses for `nsims` studies, and stores all the results needed to examine bias detection. In the first section of the script, statistically significant results in the desired direction are simulated, and in the second part null results are generated. The script generates a percentage of significant results as indicated by `pub.bias` - when set to 1, all results are significant. In the code below, `pub.bias` is set to 0.05. Because there is no true effect in the simulation (`m1` and `m2` are equal, so there is no difference between the groups), the only significant results that should be expected are the 5% false positives. Finally, the meta-analysis is performed, the results are printed, and a funnel plot is created.

```
library(metafor)
library(truncnorm)

nsims <- 100 # number of simulated experiments
pub.bias <- 0.05 # set percentage of significant results in the literature

m1 <- 0 # too large effects will make non-significant results extremely rare
sd1 <- 1
m2 <- 0
sd2 <- 1
metadata.sig <- data.frame(m1 = NA, m2 = NA, sd1 = NA, sd2 = NA,
                           n1 = NA, n2 = NA, pvalues = NA, pcurve = NA)
metadata.nonsig <- data.frame(m1 = NA, m2 = NA, sd1 = NA, sd2 = NA,
                           n1 = NA, n2 = NA, pvalues = NA, pcurve = NA)
```

```

# simulate significant effects in the expected direction
if(pub.bias > 0){
  for (i in 1:nsims*pub.bias) { # for each simulated experiment
    p <- 1 # reset p to 1
    n <- round(truncnorm::rtruncnorm(1, 20, 1000, 100)) # n based on truncated normal
    while (p > 0.025) { # continue simulating as long as p is not significant
      x <- rnorm(n = n, mean = m1, sd = sd1)
      y <- rnorm(n = n, mean = m2, sd = sd2)
      p <- t.test(x, y, alternative = "greater", var.equal = TRUE)$p.value
    }
    metadata.sig[i, 1] <- mean(x)
    metadata.sig[i, 2] <- mean(y)
    metadata.sig[i, 3] <- sd(x)
    metadata.sig[i, 4] <- sd(y)
    metadata.sig[i, 5] <- n
    metadata.sig[i, 6] <- n
    out <- t.test(x, y, var.equal = TRUE)
    metadata.sig[i, 7] <- out$p.value
    metadata.sig[i, 8] <- paste0("t(", out$parameter, ")=", out$statistic)
  }
}

# simulate non-significant effects (two-sided)
if(pub.bias < 1){
  for (i in 1:nsims*(1-pub.bias)) { # for each simulated experiment
    p <- 0 # reset p to 1
    n <- round(truncnorm::rtruncnorm(1, 20, 1000, 100))
    while (p < 0.05) { # continue simulating as long as p is significant
      x <- rnorm(n = n, mean = m1, sd = sd1) # produce simulated participants
      y <- rnorm(n = n, mean = m2, sd = sd2) # produce simulated participants
      p <- t.test(x, y, var.equal = TRUE)$p.value
    }
    metadata.nonsig[i, 1] <- mean(x)
    metadata.nonsig[i, 2] <- mean(y)
    metadata.nonsig[i, 3] <- sd(x)
    metadata.nonsig[i, 4] <- sd(y)
    metadata.nonsig[i, 5] <- n
    metadata.nonsig[i, 6] <- n
    out <- t.test(x, y, var.equal = TRUE)
    metadata.nonsig[i, 7] <- out$p.value
    metadata.nonsig[i, 8] <- paste0("t(", out$parameter, ")=", out$statistic)
  }
}

```

```

# Combine significant and non-significant effects
metadata <- rbind(metadata.nonsig, metadata.sig)

# Use escalc to compute effect sizes
metadata <- escalc(n1i = n1, n2i = n2, m1i = m1, m2i = m2, sd1i = sd1,
  sd2i = sd2, measure = "SMD", data = metadata[complete.cases(metadata),])
# add se for PET-PEESE analysis
metadata$sei <- sqrt(metadata$vi)

#Perform meta-analysis
result <- metafor::rma(yi, vi, data = metadata)
result

# Print a Funnel Plot
metafor::funnel(result, level = 0.95, refline = 0)
abline(v = result$b[1], lty = "dashed") # vertical line at meta-analytic ES
points(x = result$b[1], y = 0, cex = 1.5, pch = 17) # add point

```

Let's start by looking at what unbiased research looks like, by running the code, keeping pub.bias at 0.05, such that only 5% Type 1 errors enter the scientific literature.

```

Random-Effects Model (k = 100; tau^2 estimator: REML)

tau^2 (estimated amount of total heterogeneity): 0.0000 (SE = 0.0018)
tau (square root of estimated tau^2 value):      0.0006
I^2 (total heterogeneity / total variability):   0.00%
H^2 (total variability / sampling variability):  1.00

Test for Heterogeneity:
Q(df = 99) = 91.7310, p-val = 0.6851

Model Results:

estimate      se      zval     pval    ci.lb    ci.ub
-0.0021  0.0121  -0.1775  0.8591  -0.0258  0.0215

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

When we examine the results of the meta-analysis we see there are 100 studies in the meta-analysis ( $k = 100$ ), and there is no statistically significant heterogeneity ( $p = 0.69$ , which

is not too surprising, as we programmed the simulation to have a true effect size of 0, and there is no heterogeneity in effect sizes). We also get the results for the meta-analysis. The meta-analytic estimate is  $d = -0.002$ , which is very close to 0 (as it should be, because the true effect size is indeed 0). The standard error around this estimate is 0.012. With 100 studies, we have a very accurate estimate of the true effect size. The  $Z$ -value for the test against  $d = 0$  is -0.177, and the  $p$ -value for this test is 0.86. We cannot reject the hypothesis that the true effect size is 0. The CI around the effect size estimate (-0.026, 0.021) includes 0.

If we examine the funnel plot in Figure 12.5 we see each study represented as a dot. The larger the sample size, the higher up in the plot, and the smaller the sample size, the lower in the plot. The white pyramid represents the area within which a study is not statistically significant, because the observed effect size (x-axis) is not far enough removed from 0 such that the confidence interval around the observed effect size would exclude 0. The lower the standard error, the more narrow the confidence interval, and the smaller the effect sizes need to be in order to be statistically significant. At the same time, the smaller the standard error, the closer the effect size will be to the true effect size, so the less likely we will see effects far away from 0. We should expect 95% of the effect size estimates to fall within the funnel, if it is centered on the true effect size. We see only a few studies (five, to be exact) fall outside the white pyramid on the right side of the plot. These are the 5% significant results that we programmed in the simulation. Note that all 5 of these studies are false positives, as there is no true effect. If there was a true effect (you can re-run the simulation and set  $d$  to 0.5 by changing `m1 <- 0` in the simulation to `m1 <- 0.5`) the pyramid cloud of points would move to the right, and be centered on 0.5 instead of 0.

We can now compare the unbiased meta-analysis above with a biased meta-analysis. We can simulate a situation with extreme publication bias. Building on the estimate by Scheel, Schijen, et al. (2021), let's assume 96% of the studies show positive results. We set `pub.bias <- 0.96` in the code. We keep both means at 0, so there still is not real effect, but we will end up with mainly Type 1 errors in the predicted direction in the final set of studies. After simulating biased results, we can perform the meta-analysis to see if the statistical inference based on the meta-analysis is misleading.

```
Random-Effects Model (k = 100; tau^2 estimator: REML)

tau^2 (estimated amount of total heterogeneity): 0 (SE = 0.0019)
tau (square root of estimated tau^2 value):      0
I^2 (total heterogeneity / total variability):   0.00%
H^2 (total variability / sampling variability):  1.00

Test for Heterogeneity:
Q(df = 99) = 77.6540, p-val = 0.9445
```

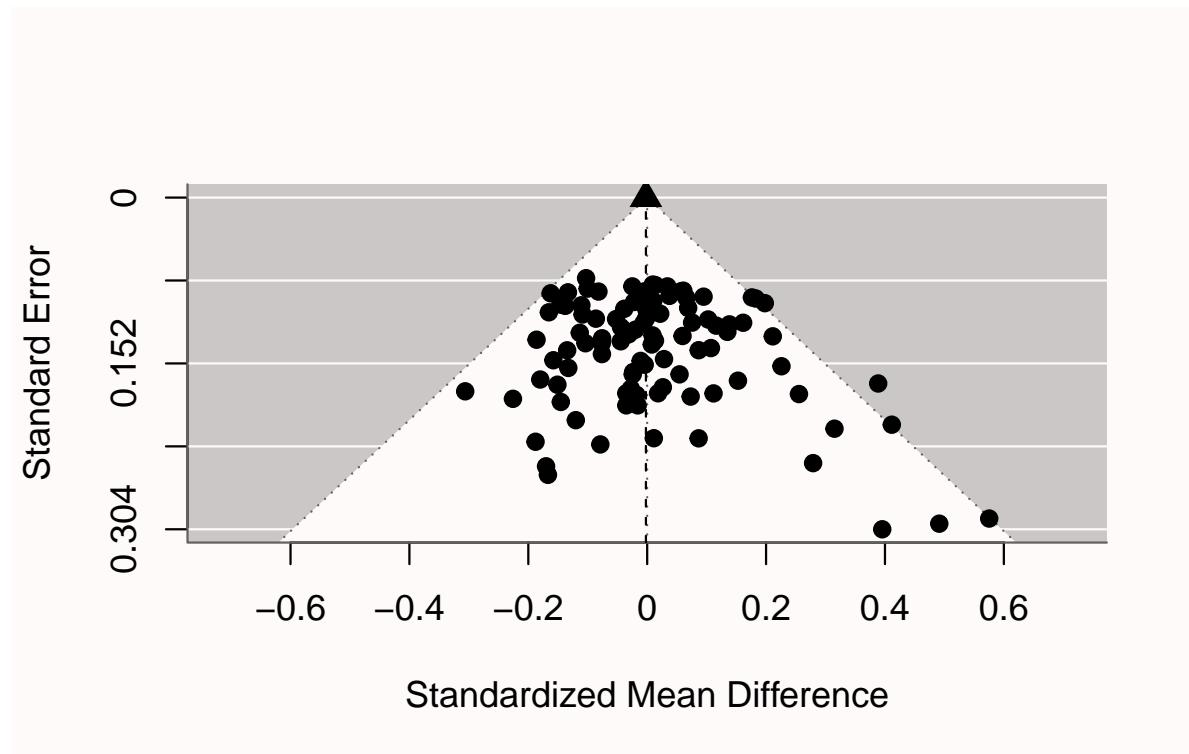


Figure 12.5: Funnel plot of unbiased null results.

Model Results:

```
estimate      se     zval    pval   ci.lb   ci.ub
 0.2701  0.0125  21.6075 <.0001  0.2456  0.2946 ***

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The biased nature of the set of studies we have analyzed becomes clear if we examine the funnel plot in fig-funnel2. The pattern is quite peculiar. We see four unbiased null results, as we programmed into the simulation, but the remainder of the 96 studies are statistically significant, even though the null is true. We see most studies fall just on the edge of the white pyramid. Because  $p$ -values are uniformly distributed under the null, the Type 1 errors we observe often have  $p$ -values in the range of 0.02 to 0.05, unlike what we would expect if there was a true effect. These just significant  $p$ -values fall just outside of the white pyramid. The larger the study, the smaller the effect size that is significant. The fact that the effect sizes do not vary around a single true effect size (e.g.,  $d = 0$  or  $d = 0.5$ ), but rather effect sizes become smaller with larger sample sizes (or smaller standard errors), is a strong indicator of bias. The vertical dotted line and black triangle at the top of the plot illustrate the observed (upwardly biased) meta-analytic effect size estimate.

One might wonder if such extreme bias ever really emerges in scientific research. It does. In Figure 12.7 we see a funnel plot by Carter & McCullough (2014) who examined bias in 198 published studies testing the ‘ego-depletion’ effect, the idea that self-control relies on a limited resource. Using bias detecting techniques such as PET-PEESE meta-regression they concluded based on the unbiased effect size estimated by PET ( $d = -0.1$ , which did not differ statistically from 0) that the true unbiased effect size might be  $d = 0$ , even though the meta-analytic effect size estimate without correcting for bias was  $d = 0.62$ . Do you notice any similarities to the extremely biased meta-analysis we simulated above? You might not be surprised that, even though before 2015 researchers thought there was a large and reliable literature demonstrating ego-depletion effects, a Registered Replication report yielded a non-significant effect size estimate (Hagger et al., 2016), and even when the original researchers tried to replicate their own work, they failed to observe a significant effect of ego-depletion (Vohs et al., 2021). Imagine the huge amount of wasted time, effort, and money on a literature that was completely based on bias in scientific research. Obviously, such research waste has ethical implications, and researchers need to take responsibility for preventing such waste in the future.

We can also see signs of bias in the forest plot for a meta-analysis. In Figure 12.8, two forest plots are plotted side by side. The left forest plot is based on unbiased data, the right forest plot is based on biased data. The forest plots are a bit big with 100 studies, but we see that in the left forest plot, the effects randomly vary around 0 as they should. On the right, beyond the first four studies, all confidence intervals magically just exclude an effect of 0.

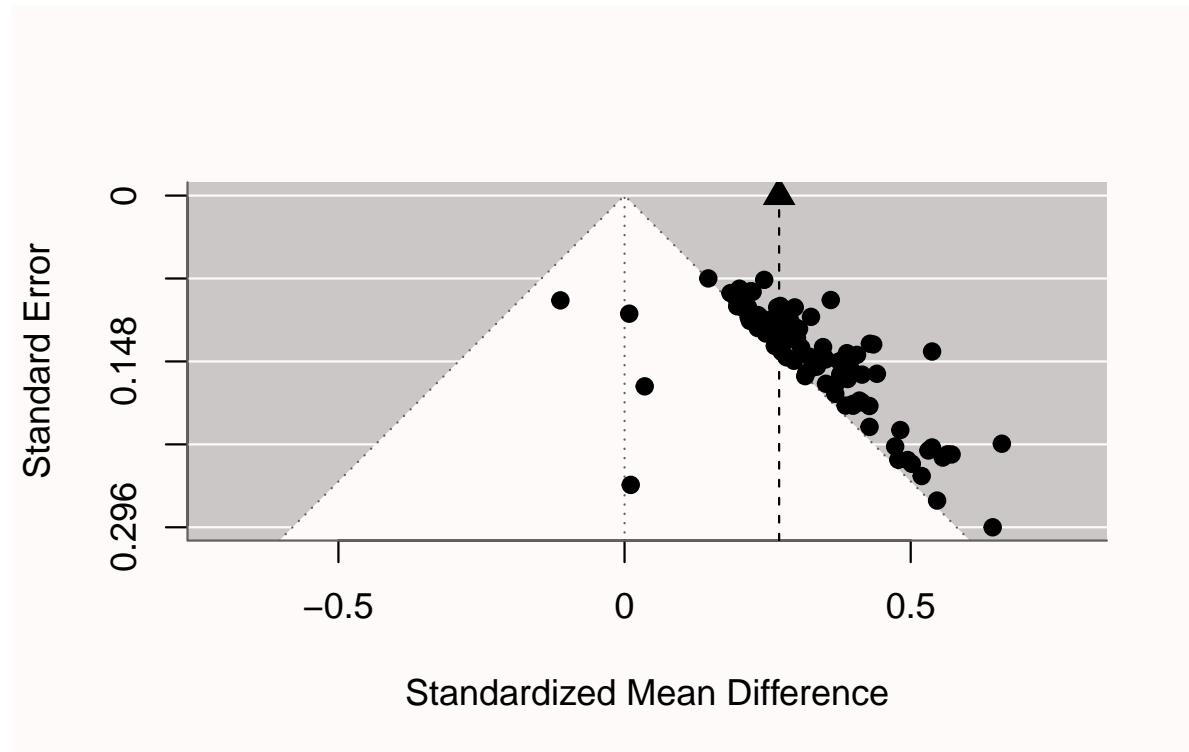


Figure 12.6: Funnel plot of biased null results with mostly significant results.

**FE  $d = 0.62$  PET  $d = -0.1$  PEESE  $d = 0.25$**

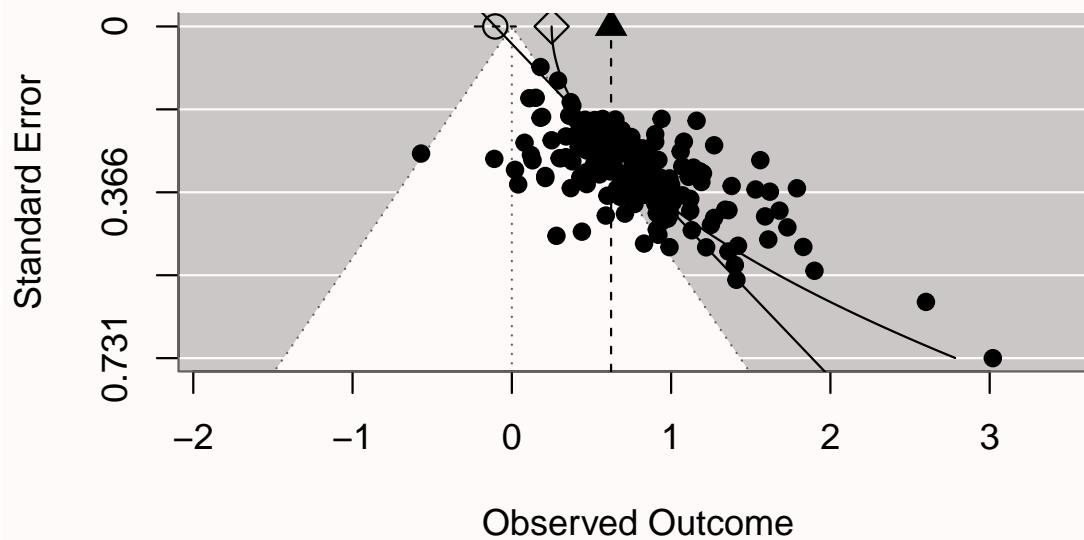


Figure 12.7: Funnel plot from Carter and McCullough (2014) visualizing bias in 198 published tests of the ego-depletion effect, including PET-PEESE bias-corrected effect size estimates.

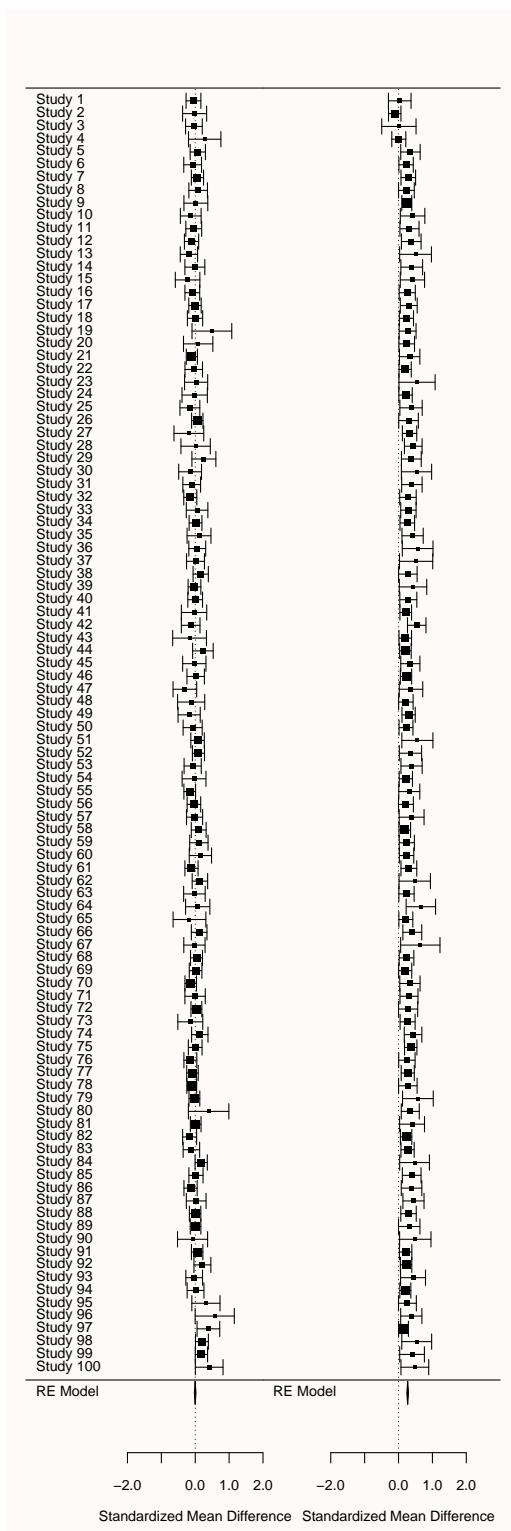


Figure 12.8: Forest plot of unbiased meta-analysis (left) and biased meta-analyses (right).

When there is publication bias because researchers only publish statistically significant results ( $p < \alpha$ ), and you calculate the effect size in a meta-analysis, the meta-analytic effect size estimate is **higher** when there is publication bias (where researchers publish only effects with  $p < \alpha$ ) compared to when there is no publication bias. This is because publication bias filters out the smaller (non-significant) effect sizes, which are then not included in the computation of the meta-analytic effect size. This leads to a meta-analytic effect size estimate that is larger than the true population effect size. With strong publication bias, we know the meta-analytic effect size is inflated, but we don't know by how much. The true effect size could just be a bit smaller, but the true effect size could also be 0, such as in the case of the ego-depletion literature.

## 12.3 Trim and Fill

Trim and fill is a technique that aims to augment a dataset by adding hypothetical ‘missing’ studies (that may be in the ‘file-drawer’). The procedure starts by removing (‘trimming’) small studies that bias the meta-analytic effect size, then estimates the true effect size, and ends with ‘filling’ in a funnel plot with studies that are assumed to be missing due to publication bias. In the Figure 12.9, you can see the same funnel plot as above, but now with added hypothetical studies (the unfilled circles which represent ‘imputed’ studies). If you look closely, you’ll see these points each have a mirror image on the opposite side of the meta-analytic effect size estimate (this is clearest in the lower half of the funnel plot). If we examine the result of the meta-analysis that includes these imputed studies, we see that trim and fill successfully alerts us to the fact that the meta-analysis is biased (if not, it would not add imputed studies) but it fails miserably in correcting the effect size estimate. In the funnel plot, we see the original (biased) effect size estimate indicated by the triangle, and the meta-analytic effect size estimate adjusted with the trim-and-fill method (indicated by the black circle). We see the meta-analytic effect size estimate is a bit lower, but given that the true effect size in the simulation was 0, the adjustment is clearly not sufficient.

Trim-and-fill is not very good under many realistic publication bias scenarios. The method is criticized for its reliance on the strong assumption of symmetry in the funnel plot. When publication bias is based on the  $p$ -value of the study (arguably the most important source of publication bias in many fields) the trim-and-fill method does not perform well enough to yield a corrected meta-analytic effect size estimate that is close to the true effect size (Peters et al., 2007; Terrin et al., 2003). When the assumptions are met, it can be used as a **sensitivity analysis**. Researchers should not report the trim-and-fill corrected effect size estimate as a realistic estimate of the unbiased effect size. If other bias-detection tests (like  $p$ -curve or  $z$ -curve discussed below) have already indicated the presence of bias, the trim-and-fill procedure might not provide additional insights.

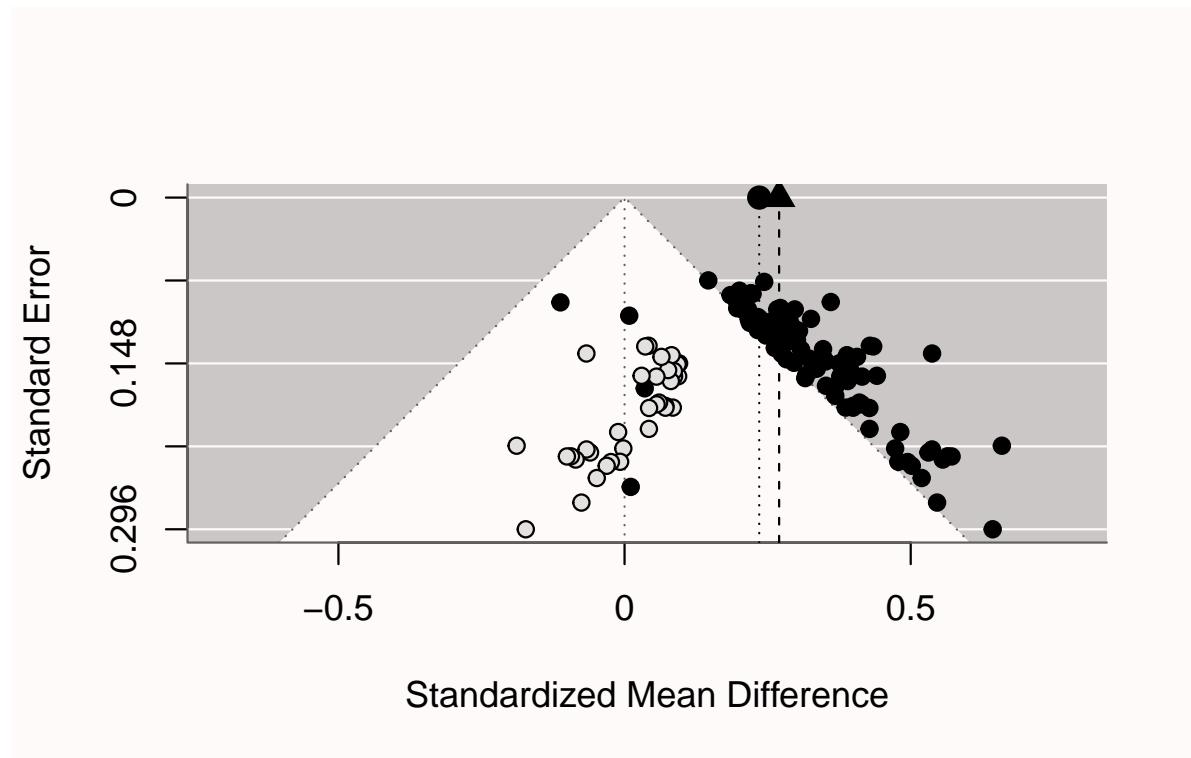


Figure 12.9: Funnel plot with assumed missing effects added through trim-and-fill.

## 12.4 PET-PEESE

A novel class of solutions to publication bias is **meta-regression**. Instead of plotting a line through individual data-points, in meta-regression a line is plotted through data points that each represent a study. As with normal regression, the more data meta-regression is based on, the more precise the estimate is and, therefore, the more studies in a meta-analysis, the better meta-regression will work in practice. If the number of studies is small, all bias detection tests lose power, and this is something that one should keep in mind when using meta-regression. Furthermore, regression requires sufficient variation in the data, which in the case of meta-regression means a wide range of sample sizes (recommendations indicate meta-regression performs well if studies have a range from 15 to 200 participants in each group – which is not typical for most research areas in psychology). Meta-regression techniques try to estimate the population effect size if precision was perfect (so when the standard error = 0).

One meta-regression technique is known as PET-PEESE (Stanley et al., 2017; Stanley & Doucouliagos, 2014). It consists of a ‘precision-effect-test’ (PET) which can be used in a Neyman-Pearson hypothesis testing framework to test whether the meta-regression estimate can reject an effect size of 0 based on the 95% CI around the PET estimate at the intercept  $SE = 0$ . Note that when the confidence interval is very wide due to a small number of observations, this test might have low power, and have an a-priori low probability of rejecting the null effect. The estimated effect size for PET is calculated with:  $d = \beta_0 + \beta_1 SE_i + u_i$  where  $d$  is the estimated effect size,  $SE$  is the standard error, and the equation is estimated using weighted least squares (WLS), with  $1/SE^2_i$  as the weights. The PET estimate underestimates the effect size when there is a true effect. Therefore, the PET-PEESE procedure recommends first using PET to test whether the null can be rejected, and if so, then the ‘precision-effect estimate with standard error’ (PEESE) should be used to estimate the meta-analytic effect size. In PEESE, the standard error (used in PET) is replaced by the variance (i.e., the standard error squared), which Stanley & Doucouliagos (2014) find reduces the bias of the estimated meta-regression intercept.

PET-PEESE has limitations, as all bias detection techniques have. The biggest limitations are that it does not work well when there are few studies, all the studies in a meta-analysis have small sample sizes, or when there is large heterogeneity in the meta-analysis (Stanley et al., 2017). When these situations apply (and they will in practice), PET-PEESE might not be a good approach. Furthermore, there are some situations where there might be a correlation between sample size and precision, which in practice will often be linked to heterogeneity in the effect sizes included in a meta-analysis. For example, if true effects are different across studies, and people perform power analyses with accurate information about the expected true effect size, large effect sizes in a meta-analysis will have small sample sizes, and small effects will have large sample sizes. Meta-regression is, like normal regression, a way to test for an association, but you need to think about the causal mechanism behind the association.

Let’s explore how PET-PEESE meta-regression attempts to give us an unbiased effect size estimate, under specific assumptions of how publication bias is caused. In we once again see

the funnel plot, now complemented with 2 additional lines through the plots. The vertical line at  $d = 0.27$  is the meta-analytic effect size estimate, which is upwardly biased because we are averaging over statistically significant studies only. There are 2 additional lines, which are the meta-regression lines for PET-PEESE based on the formulas detailed previously. The straight diagonal line gives us the PET estimate at a SE of 0 (an infinite sample, at the top of the plot), indicated by the circle. The dotted line around this PET estimate is the 95% confidence interval for the estimate. In this case, the 95% CI contains 0, which means that based on the PET estimate of  $d = 0.02$ , we cannot reject a meta-analytic effect size of 0. Note that even with 100 studies, the 95% CI is quite wide. Meta-regression is, just like normal regression, only as accurate as the data we have. This is one limitation of PET-PEESE meta-regression: With small numbers of studies in the meta-analysis, it has low accuracy. If we had been able to reject the null based on the PET estimate, we would then have used the PEESE estimate (indicated by the diamond shape) of  $d = 0.17$  for the meta-analytic effect size, corrected for bias (while never knowing whether the model underlying the PEESE estimate corresponded to the true bias generating mechanisms in the meta-analysis, and thus if the meta-analytic estimate was accurate).

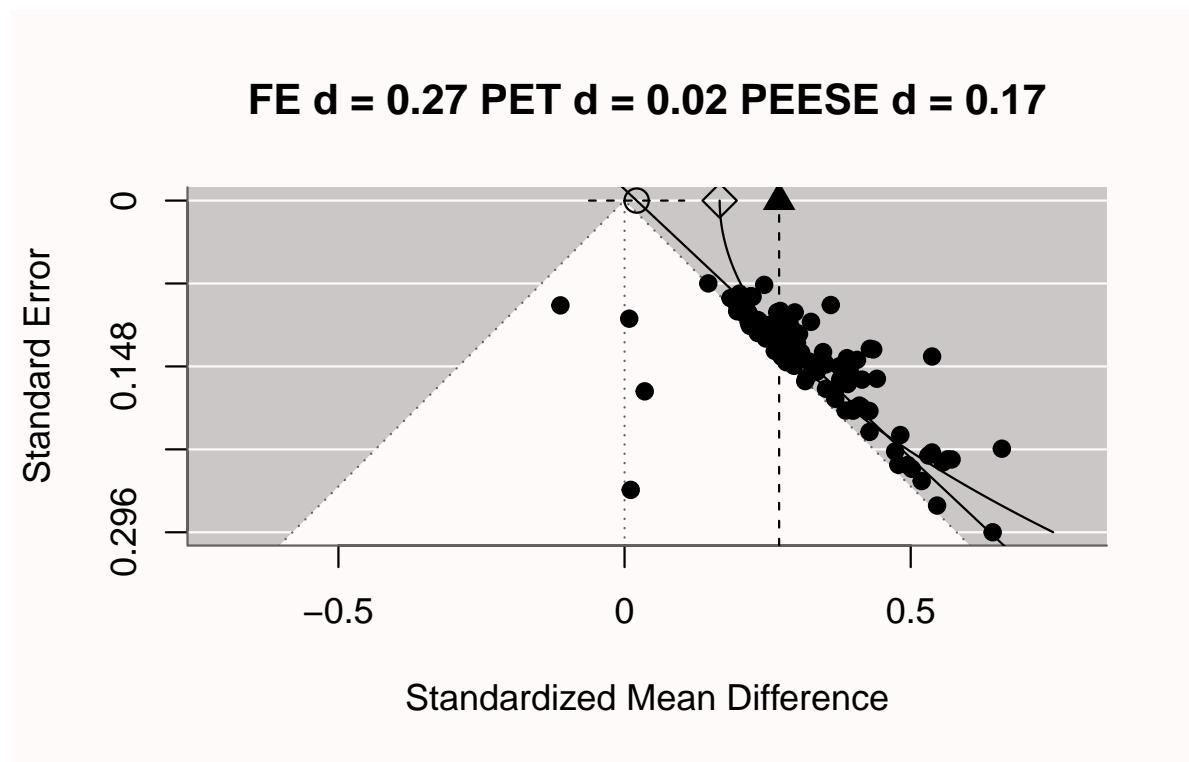


Figure 12.10: Funnel plot with PETPEESE regression lines.

## 12.5 *P*-value meta-analysis

In addition to a meta-analysis of effect sizes, it is possible to perform a meta-analysis of *p*-values. The first of such approaches is known as the **Fisher's combined probability test**, and more recent bias detection tests such as *p*-curve analysis (Simonsohn et al., 2014) and *p*-uniform\* (Aert & Assen, 2018) build on this idea. These two techniques are an example of selection model approaches to test and adjust for meta-analysis (Iyengar & Greenhouse, 1988), where a *model about the data generating process* of the effect sizes is combined with a *selection model* of how publication bias impacts which effect sizes become part of the scientific literature. An example of a data generating process would be that results of studies are generated by statistical tests where all test assumptions are met, and the studies have some average power. A selection model might be that all studies are published, as long as they are statistically significant at an alpha level of 0.05.

*P*-curve analysis uses exactly this selection model. It assumes all significant results are published, and examines whether the data generating process mirrors what would be expected if the studies have a certain power, or whether the data generating process mirrors the pattern expected if the null hypothesis is true. As discussed in the section on [which \*p\*-values you can expect](#), for continuously distributed test statistics (e.g., *t*-values, *F*-values, *Z*-scores) we should observe uniformly distributed *p*-values when the null hypothesis is true, and more small significant *p*-values (e.g., 0.01) than large significant *p*-values (e.g., 0.04) when the alternative hypothesis is true. *P*-curve analysis performs two tests. In the first test, *p*-curve analysis examines whether the *p*-value distribution is flatter than what would be expected if the studies you analyze had 33% power. This value is somewhat arbitrary (and can be adjusted), but the idea is to reject at the smallest level of statistical power that would lead to useful insights about the presence of effects. If the average power in the set of studies is less than 33%, there might be an effect, but the studies are not designed well enough to learn about it by performing statistical tests. If we can reject the presence of a pattern of *p*-values that has at least 33% power, this suggests the distribution looks more like one expected when the null hypothesis is true. That is, we would doubt there is an effect in the set of studies included in the meta-analysis, *even though all individual studies were statistically significant*.

The second test examines whether the *p*-value distribution is sufficiently right-skewed (more small significant *p*-values than large significant *p*-values), such that the pattern suggests we can reject a uniform *p*-value distribution. If we can reject a uniform *p*-value distribution, this suggests the studies might have examined a true effect and had at least some power. If the second test is significant, we would act as if the set of studies examines some true effect, even though there might be publication bias. As an example, let's consider Figure 3 from Simonsohn and colleagues (2014). The authors compared 20 papers in the Journal of Personality and Social Psychology that used a covariate in the analysis, and 20 studies that did not use a covariate. The authors suspected that researchers might add a covariate in their analyses to try to find a *p*-value smaller than 0.05, when the first analysis they tried did not yield a significant effect.

**Figure 3. P-curves for JPSP studies suspected to have been *p*-hacked (A) and not *p*-hacked (B).**

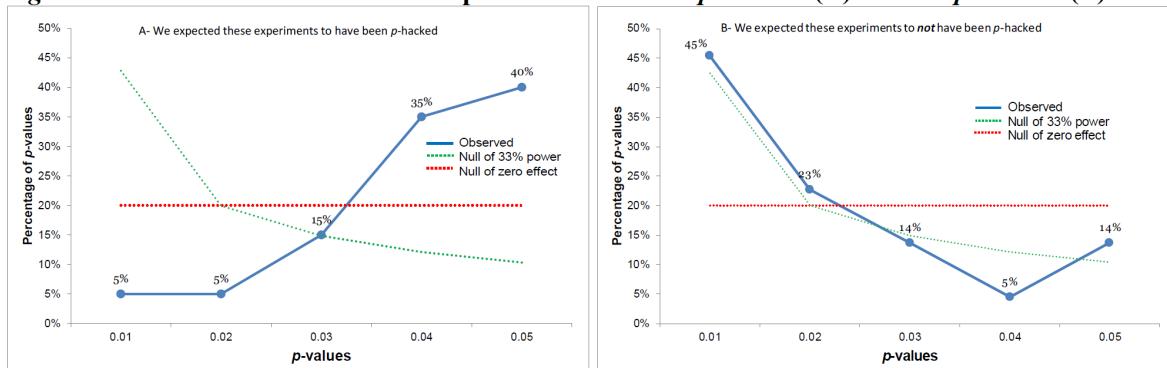


Figure 12.11: Figure 3 from Simonsohn et al (2014) showing a *p*-curve with and without bias.

The *p*-curve distribution of the observed *p*-values is represented by five points in the blue line. *P*-curve analysis is performed *only* on statistically significant results, based on the assumption that these are always published, and thus that this part of the *p*-value distribution contains all studies that were performed. The 5 points illustrate the percentage of *p*-values between 0 and 0.01, 0.01 and 0.02, 0.02 and 0.03, 0.03 and 0.04, and 0.04 and 0.05. In the figure on the right, you see a relatively normal right-skewed *p*-value distribution, with more low than high *p*-values. The *p*-curve analysis shows that the blue line in the right figure is more right-skewed than the uniform red line (where the red line is the uniform *p*-value distribution expected if there was no effect). Simonsohn and colleagues summarize this pattern as an indication that the set of studies has ‘evidential value’, but this terminology is somewhat misleading. The formally correct interpretation is that we can reject a *p*-value distribution as expected when the null hypothesis was true in all studies included in the *p*-curve analysis. Rejecting a uniform *p*-value distribution does not automatically mean there is evidence for the theorized effect (e.g., the pattern could be caused by a mix of null effects and a small subset of studies that show an effect due to a methodological confound).

In the left figure we see the opposite pattern, with mainly high *p*-values around 0.05, and almost no *p*-values around 0.01. Because the blue line is significantly flatter than the green line, the *p*-curve analysis suggests this set of studies is the result of selection bias and was not generated by a set of sufficiently powered studies. *P*-curve analysis is a useful tool. But it is important to correctly interpret what a *p*-curve analysis can tell you. A right-skewed *p*-curve does not prove that there is no bias, or that the theoretical hypothesis is true. A flat *p*-curve does not prove that the theory is incorrect, but it does show that the studies that were meta-analyzed look more like the pattern that would be expected if the null hypothesis was true, and there was selection bias.

The script stores all the test statistics for the 100 simulated *t*-tests that are included in the meta-analysis. The first few rows look like:

```
t(136)=0.208132209831132
t(456)=-1.20115958535433
t(58)=0.0422284763301259
t(358)=0.0775200850900646
t(188)=2.43353676652346
```

Print all test results with `cat(metadata$pcurve, sep = "\n")`, and go to the online *p*-curve app at <http://www.p-curve.com/app4/>. Paste all the test results, and click the ‘Make the *p*-curve’ button. Note that the *p*-curve app will only yield a result when there are *p*-values smaller than 0.05 - if all test statistics yield a *p* > 0.05, the *p*-curve cannot be computed, as these tests are ignored.

The distribution of *p*-values clearly looks like it comes from a uniform distribution (as it indeed does), and the statistical test indicates we can reject a *p*-value distribution as steep or steeper as would be generated by a set of studies with 33% power, *p* < 0.0001. The app also provides an estimate of the average power of the tests that generated the observed *p*-value distribution, 5%, which is indeed correct. Therefore, we can conclude these studies, even though many effects are statistically significant, are more in line with selective reporting of Type 1 errors, than with a *p*-value distribution that should be expected if there was a true effect that was studied with sufficient statistical power. The theory might still be true, but the set of studies we have analyzed here do not provide support for the theory.

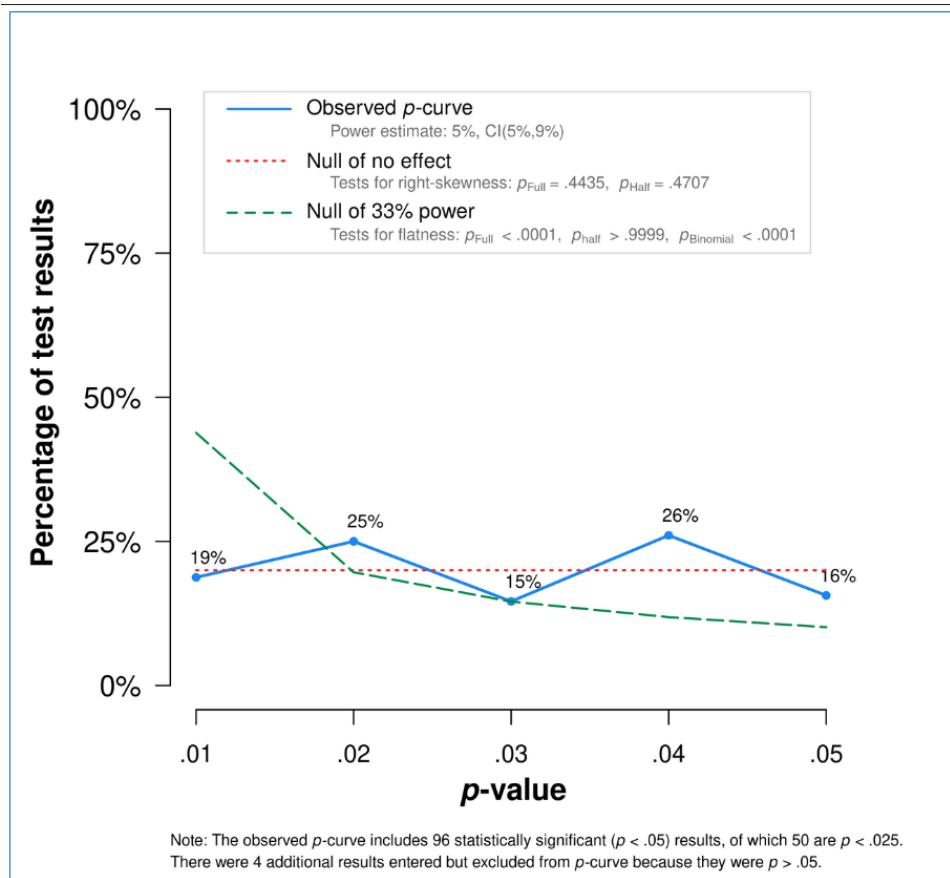
A similar meta-analytic technique is *p*-uniform\*. This technique is similar to *p*-curve analysis and selection bias models, but it uses the results both from significant and non-significant studies, and can be used to estimate a bias-adjusted meta-analytic effect size estimate. The technique uses a random-effects model to estimate the effect sizes for each study, and weighs them based on a selection model that assumes significant results are more likely to be published than non-significant results. Below, we see the output of the *p*-uniform\* which estimates the bias-corrected effect size to be *d* = 0.0126. This effect size is not statistically different from 0, *p* = 0.3857, and therefore this bias detection technique correctly indicates that even though all effects were statistically significant, the set of studies does not provide a good reason to reject a meta-analytic effect size estimate of 0.

```
puniform::puniform(m1i = metadata$m1, m2i = metadata$m2, n1i = metadata$n1,
  n2i = metadata$n2, sd1i = metadata$sd1, sd2i = metadata$sd2, side = "right")
```

Method: P

Effect size estimation p-uniform

est	ci.lb	ci.ub	L.0	pval	ksig
0.0126	-0.0811	0.0887	-0.2904	0.3857	96



	<b>Binomial Test</b> (Share of results $p < .025$ )	<b>Continuous Test</b> (Aggregate with Stouffer Method)	
		<b>Full p-curve</b> ( $p's < .05$ )	<b>Half p-curve</b> ( $p's < .025$ )
1) Studies contain evidential value. <i>(Right skew)</i>	$p=.3798$	$Z=-0.14$ , $p=.4435$	$Z=-0.07$ , $p=.4707$
2) Studies' evidential value, if any, is inadequate. <i>(Flatter than 33% power)</i>	$p=.0001$	$Z=-5.38$ , $p<.0001$	$Z=6$ , $p>.9999$
<b>Statistical Power</b>			
Estimate: 5%			
90% Confidence interval: (5%, 9%)			

Power of tests included in p-curve  
*(correcting for selective reporting)*

Figure 12.12: Result of the p-curve analysis of the biased studies.

====

Publication bias test p-uniform

L.pb	pval
7.9976	<.001

====

Fixed-effect meta-analysis

est.fe	se.fe	zval.fe	pval.fe	ci.lb.fe	ci.ub.fe	Qstat	Qpval
0.2701	0.0125	21.6025	<.001	0.2456	0.2946	77.6031	0.945

An alternative technique that also meta-analyzes the  $p$ -values from individual studies is a  $z$ -curve analysis, which is a meta-analysis of observed power ((Bartoš & Schimmack, 2020; Brunner & Schimmack, 2020); for an example, see (Sotola, 2022)). Like a traditional meta-analysis,  $z$ -curve analysis transforms observed test results ( $p$ -values) into  $z$ -scores. In an unbiased literature where the null hypothesis is true, we should observe approximately  $\alpha\%$  significant results. If the null is true, the distribution of  $z$ -scores is centered on 0.  $Z$ -curve analysis computes absolute  $z$ -values, and therefore  $\alpha\%$  of  $z$ -scores should be larger than the critical value (1.96 for a 5% alpha level). In Figure 12.13  $z$ -scores for 1000 studies are plotted, with a true effect size of 0, where exactly 5% of the observed results are statistically significant.

If there is a true effect, the distribution of  $z$ -scores shifts away from 0, as a function of the statistical power of the test. The higher the power, the further to the right the distribution of  $z$ -scores will be located. For example, when examining an effect with 66% power, an unbiased distribution of  $z$ -scores, computed from observed  $p$ -values, looks like the distribution in Figure 12.14.

In any meta-analysis the studies that are included will differ in their statistical power, and their true effect size (due to heterogeneity).  $Z$ -curve analysis uses mixtures of normal distributions centered at means 0 to 6 to fit a model of the underlying effect sizes that best represents the observed results in the included studies (for the technical details, see Bartoš & Schimmack (2020)). The  $z$ -curve then aims to estimate the average power of the set of studies, and then calculates the *observed discovery rate* (ODR: the percentage of significant results, or the observed power), the *expected discovery rate* (EDR: the proportion of the area under the curve on the right side of the significance criterion) and the expected replication rate (ERR: the expected proportion of successfully replicated significant studies from all significant studies). The  $z$ -curve is able to correct for selection bias for positive results (under specific assumptions), and can estimate the EDR and ERR using only the significant  $p$ -values.

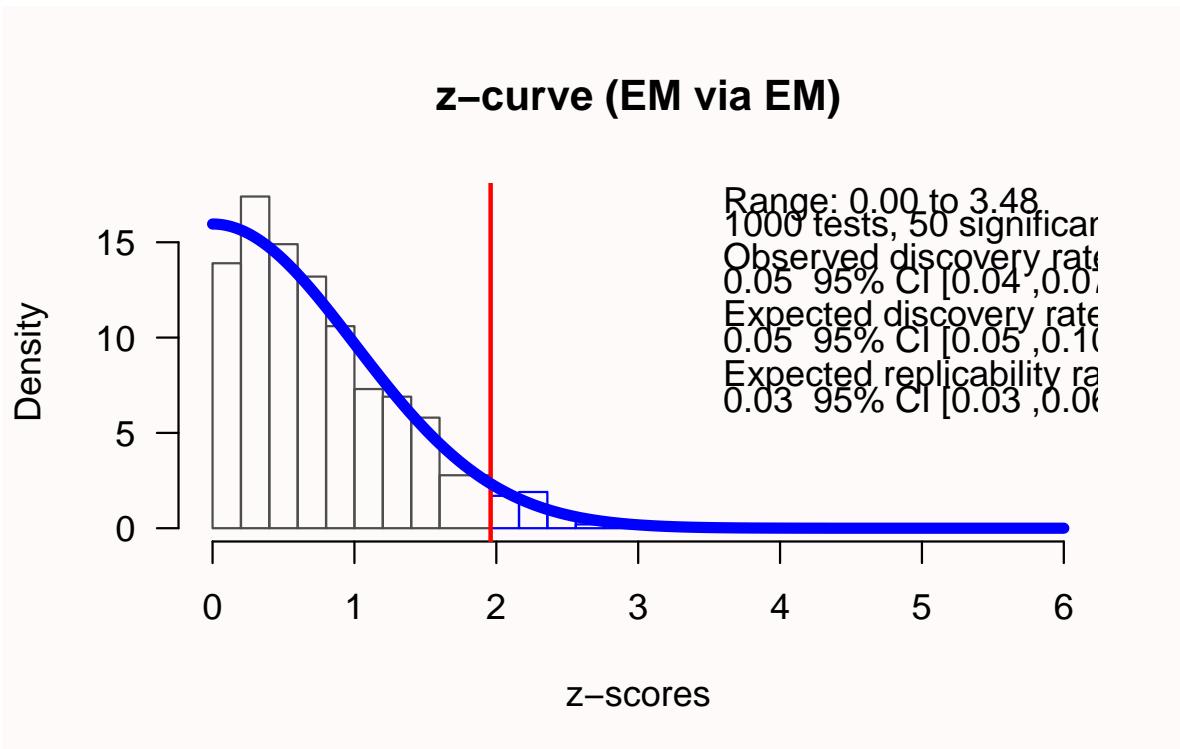


Figure 12.13: Z-curve analysis for 1000 studies with a true effect size of 0 without publication bias.

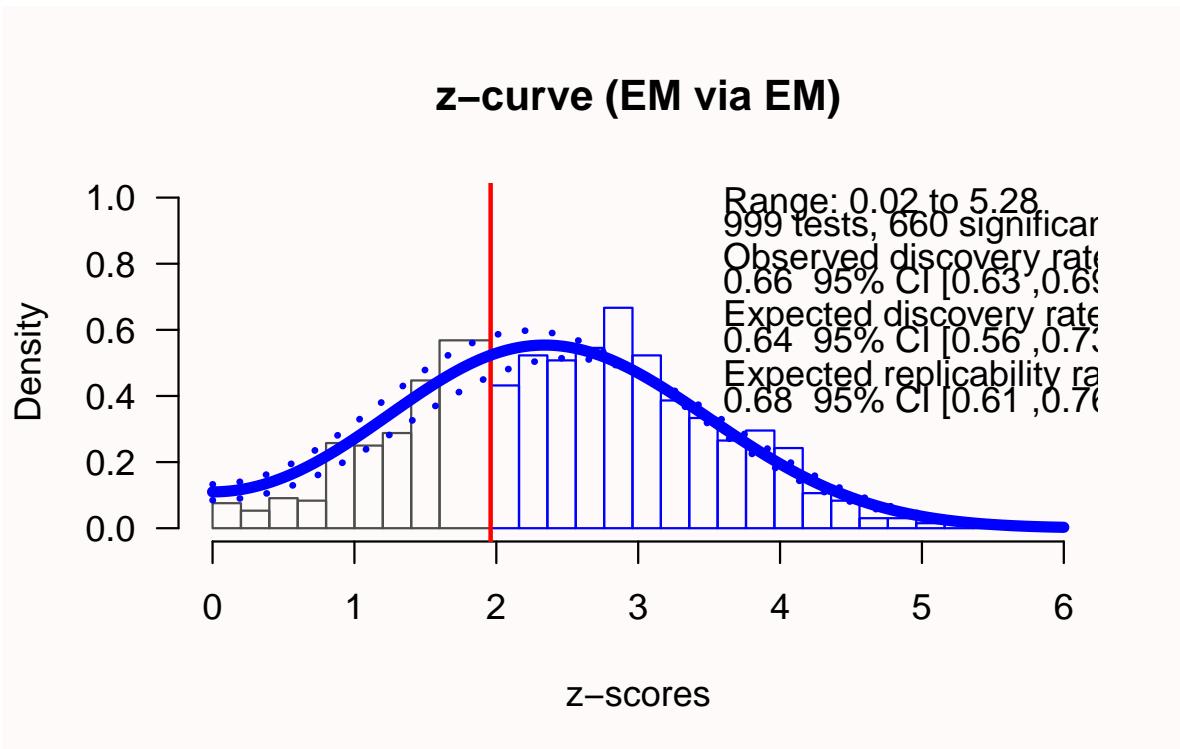
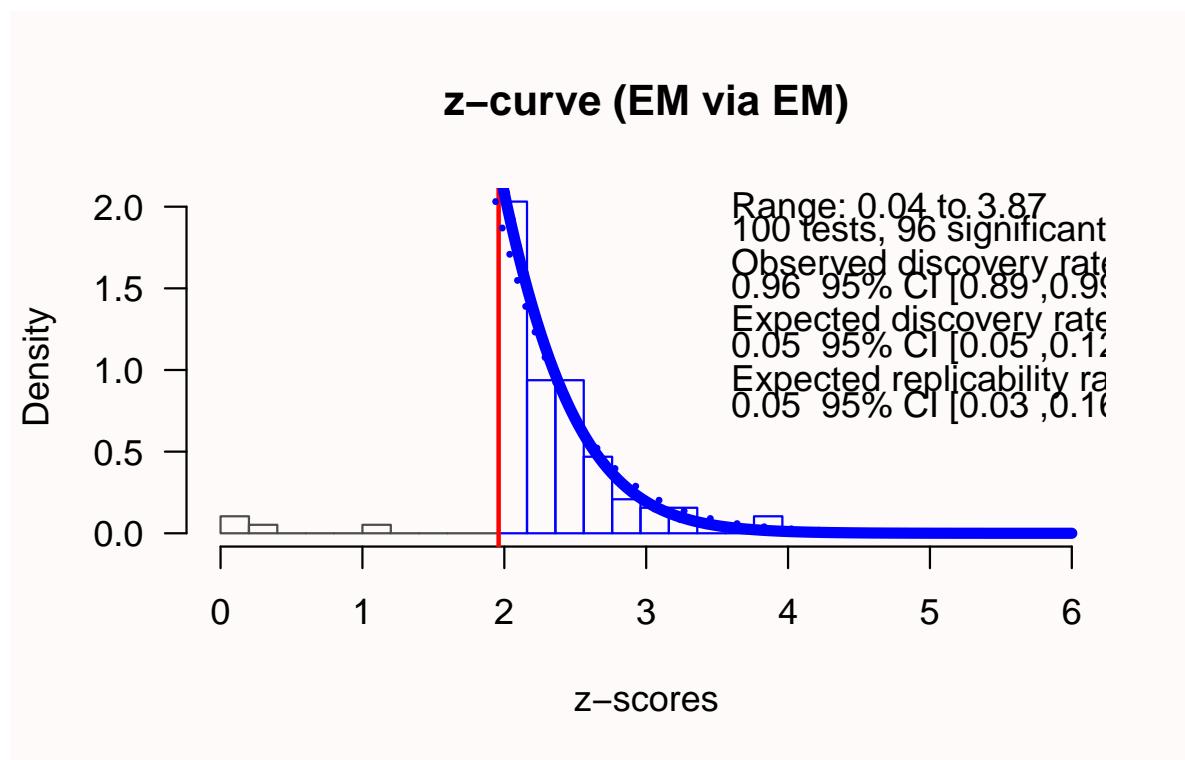


Figure 12.14: Z-curve analysis for 1000 studies with a true effect size of  $d = 0.37$  and  $n = 100$  per condition in an independent  $t$ -test without publication bias.

To examine the presence of bias, it is preferable to submit non-significant and significant  $p$ -values to a  $z$ -curve analysis, even if only the significant  $p$ -values are used to produce estimates. Publication bias can then be examined by comparing the ODR to the EDR. If the percentage of significant results in the set of studies (ODR) is much higher than the expected discovery rate (EDR), this is a sign of bias. If we analyze the same set of biased studies as we used to illustrate the bias detection techniques discussed above,  $z$ -curve analysis should be able to indicate the presence of bias. We can perform the  $z$ -curve with the following code:

```
z_res <- zcurve::zcurve(p = metadata$pvalues, method = "EM", bootstrap = 1000)
summary(z_res, all = TRUE)
plot(z_res, annotation = TRUE, CI = TRUE)
```



```
Call:
zcurve::zcurve(p = metadata$pvalues, method = "EM", bootstrap = 1000)

model: EM via EM

      Estimate   l.CI    u.CI
ERR       0.052  0.025  0.160
```

EDR	0.053	0.050	0.119
Soric FDR	0.947	0.389	1.000
File Drawer R	17.987	7.399	19.000
Expected N	1823	806	1920
Missing N	1723	706	1820

Model converged in 38 + 205 iterations

Fitted using 96 p-values. 100 supplied, 96 significant (ODR = 0.96, 95% CI [0.89, 0.99]).  
 $Q = -6.69$ , 95% CI [-23.63, 11.25]

We see that the distribution of  $z$ -scores looks peculiar. Most expected  $z$ -scores between 0 and 1.96 are missing. 96 out of 100 studies were significant, which makes the observed discovery rate (ODR), or observed power (across all these studies with different sample sizes) 0.96, 95% CI[0.89; 0.99]. The expected discovery rate (EDR) is only 0.053, which differs statistically from the observed discovery rate, as indicated by the fact that the confidence interval of the EDR does not overlap with the ODR of 0.96. This means there is clear indication of selection bias based on the  $z$ -curve analysis. The expected replicability rate for these studies is only 0.052, which is in line with the expectation that we will only observe 5% Type 1 errors, as there was no true effect in this simulation. Thus, even though we only entered significant  $p$ -values,  $z$ -curve analysis correctly suggests that we should not expect these results to replicate at a higher frequency than the Type 1 error rate.

## 12.6 Conclusion

Publication bias is a big problem in science. It is present in almost all meta-analyses performed on the primary hypothesis test in scientific articles, because these articles are much more likely to be submitted and accepted for publication if the primary hypothesis test is statistically significant. Meta-analytic effect size estimates that are not adjusted for bias will almost always overestimate the true effect size, and bias-adjusted effect sizes might still be misleading. Having messed up the scientific literature through publication bias, there is no way for us to know whether we are computing accurate meta-analytic effect sizes estimates from the literature. Publication bias inflates the effect size estimate to an unknown extent, and there have already been several cases where the true effect size turned out to be zero. The publication bias tests in this chapter might not provide certainty about the unbiased effect size, but they can function as a red flag to indicate when bias is present, and provide adjusted estimates that, if the underlying model of publication bias is correct, might well be closer to the truth.

There is a lot of activity in the literature on tests for publication bias. There are many different tests, and you need to carefully check the assumptions of each test before applying it. Most tests don't work well when there is large heterogeneity, and heterogeneity is quite likely. A meta-analysis should always examine whether there is publication bias, preferably using

multiple publication bias tests, and therefore it is useful to not just code effect sizes, but also code test statistics or  $p$ -values. None of the bias detection techniques discussed in this chapter will be a silver bullet, but they will be better than naively interpreting the uncorrected effect size estimate from the meta-analysis.

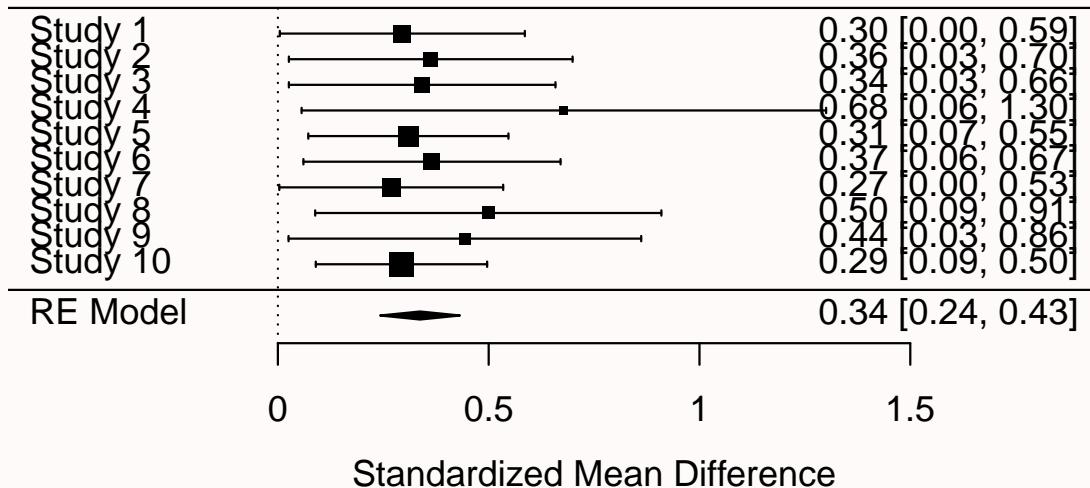
For another open educational resource on tests for publication bias, see [Doing Meta-Analysis in R](#).

## 12.7 Test Yourself

**Q1:** What happens when there is publication bias because researchers only publish statistically significant results ( $p < \alpha$ ), and you calculate the effect size in a meta-analysis?

- (A) The meta-analytic effect size estimate is **identical** whether there is publication bias (where researchers publish only effects with  $p < \alpha$ ) or no publication bias.
- (B) The meta-analytic effect size estimate is **closer to the true effect size** when there is publication bias (where researchers publish only effects with  $p < \alpha$ ) compared to when there is no publication bias.
- (C) The meta-analytic effect size estimate is **inflated** when there is publication bias (where researchers publish only effects with  $p < \alpha$ ) compared to when there is no publication bias.
- (D) The meta-analytic effect size estimate is **lower** when there is publication bias (where researchers publish only effects with  $p < \alpha$ ) compared to when there is no publication bias.

**Q2:** The forest plot in the figure below looks quite peculiar. What do you notice?



- (A) All effect sizes are quite similar, suggesting large sample sizes and highly accurate effect size measures.
- (B) The studies look as if they were designed based on perfect a-priori power analyses, all yielding just significant results.
- (C) The studies have confidence intervals that only just fail to include 0, suggesting most studies are only just statistically significant. This suggests publication bias.
- (D) All effects are in the same direction, which suggests that one-sided tests have been performed, even though these might not have been preregistered.

**Q3:** Which statement is true?

- (A) With extreme publication bias, all individual studies in a literature can be significant, but the standard errors are so large that the meta-analytic effect size estimate is not significantly different from 0.
- (B) With extreme publication bias, all individual studies in a literature can be significant, but the meta-analytic effect size estimate will be severely inflated,

giving the impression there is overwhelming support for  $H_1$  when actually the true effect size is either small, or even 0.

- (C) With extreme publication bias, all individual studies are significant, but meta-analytic effect size estimates are automatically corrected for publication bias in most statistical packages, and the meta-analytic effect size estimate is therefore quite reliable.
- (D) Regardless of whether there is publication bias, the meta-analytic effect size estimate is severely biased, and it should never be considered a reliable estimate of the population.

**Q4:** Which statement is true based on the plot below, visualizing a PET-PEESE meta-regression?

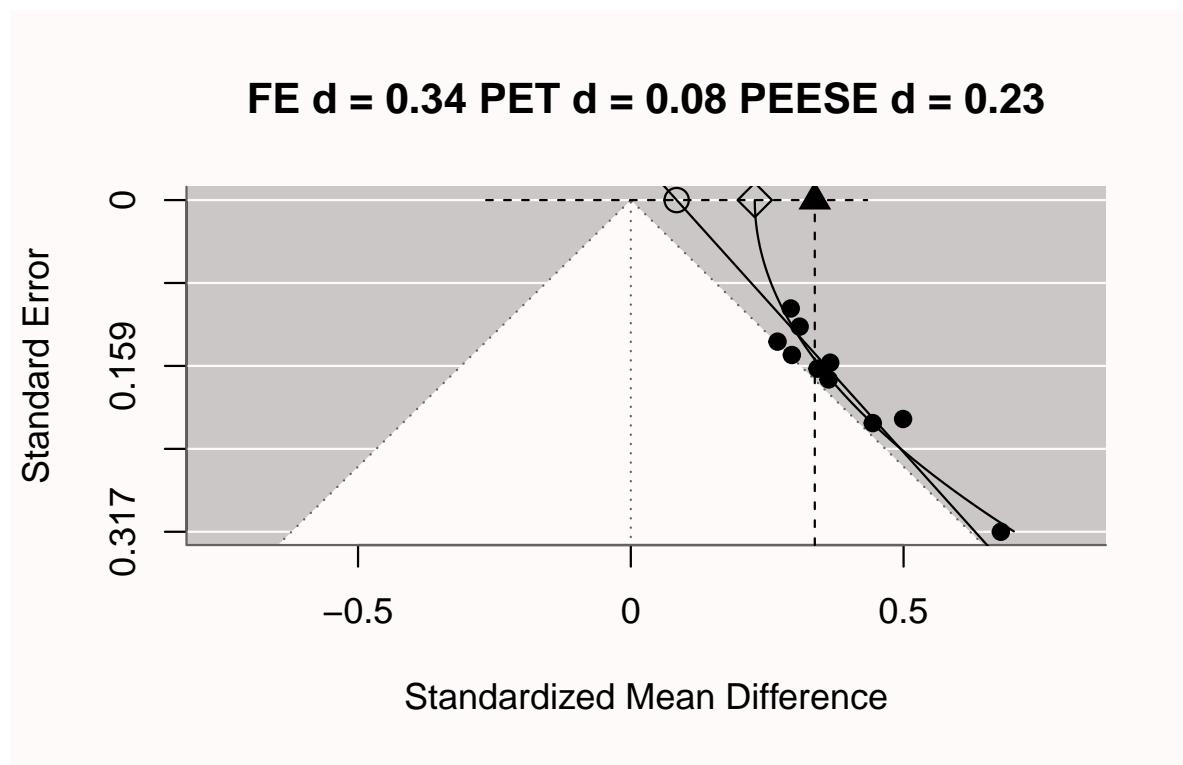


Figure 12.15: Funnel plot with PETPEESE regression lines for the same studies as in Q2.

- (A) Using PET-PEESE meta-regression we can show that the true effect size is  $d = 0$  (based on the PET estimate).

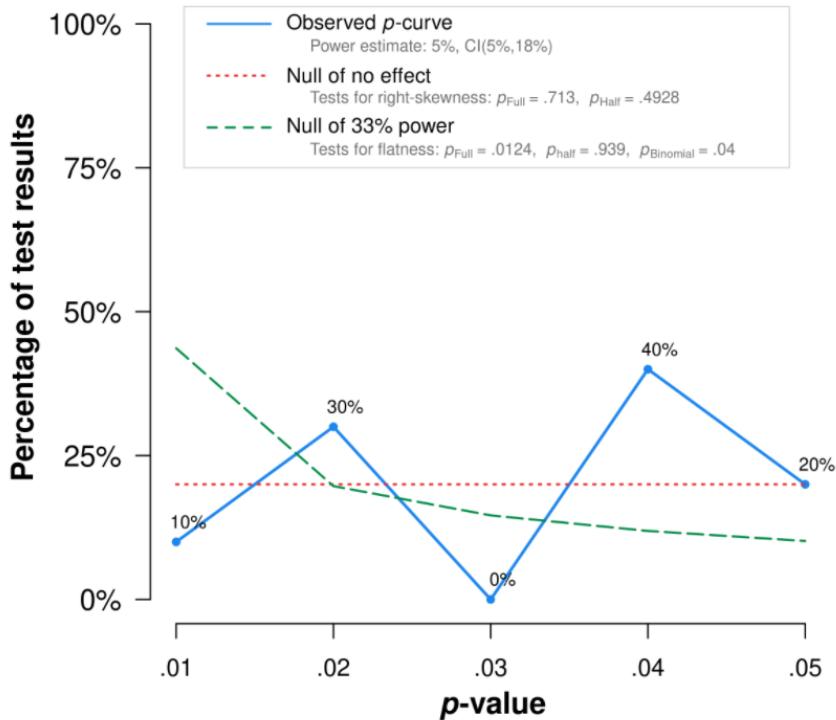
- (B) Using PET-PEESE meta-regression we can show that the true effect size is  $d = r \text{ round}(PEESE$b[1], 2)$  (based on the PEESE estimate).
- (C) Using PET-PEESE meta-regression we can show that the true effect size is  $d = r \text{ round(result.biased$b, 2)}$  (based on the normal meta-analytic effect size estimate).
- (D) The small sample size (10 studies) means PET has very low power to reject the null, and therefore it is not a reliable indicator of bias - but there might be reason to worry.

**Q5:** Take a look at the figure and output table of the *p*-curve app below, which gives the results for the studies in Q2. Which interpretation of the output is correct?

- (A) Based on the continuous Stouffer's test for the full *p*-curve, we cannot reject a *p*-value distribution expected under  $H_0$ , and we can reject a *p*-value distribution as expected if  $H_1$  is true and studies had 33% power.
- (B) Based on the continuous Stouffer's test for the full *p*-curve, we can conclude the observed *p*-value distribution is not skewed enough to be interpreted as the presence of a true effect size, therefore the theory used to deduce these studies is incorrect.
- (C) Based on the continuous Stouffer's test for the full *p*-curve, we can conclude the observed *p*-value distribution is skewed enough to be interpreted in line with a *p*-value distribution as expected if  $H_1$  is true and studies had 33% power.
- (D) Based on the continuous Stouffer's test for the full *p*-curve, we can conclude the observed *p*-value distribution is flatter than we would expect if the studies had 33% power, and therefore, we can conclude these studies are based on fabricated data.

**Q6:** The true effect size in the studies simulated in Q2 is 0 - there is no true effect. Which statement about the *z*-curve analysis below is true?

- (A) The expected discovery rate and the expected replicability rate are both statistically significant, and therefore we can expect the observed effects to successfully replicate in future studies.
- (B) Despite the fact that the average observed power (the observed discovery rate) is 100%, *z*-curve correctly predicts the expected replicability rate (which is 5%, as only Type 1 errors will be statistically significant).



Note: The observed p-curve includes 10 statistically significant ( $p < .05$ ) results, of which 4 are  $p < .025$ .  
There were no non-significant results entered.

	Binomial Test (Share of results $p < .025$ )		Continuous Test (Aggregate with Stouffer Method)	
	Full p-curve ( $p < .05$ )	Half p-curve ( $p < .025$ )	Full p-curve ( $p < .05$ )	Half p-curve ( $p < .025$ )
1) Studies contain evidential value. <i>(Right skew)</i>	$p=.8281$		$Z=0.56, p=.713$	$Z=-0.02, p=.4928$
2) Studies' evidential value, if any, is inadequate. <i>(Flatter than 33% power)</i>	$p=.04$		$Z=-2.24, p=.0124$	$Z=1.55, p=.939$
<b>Statistical Power</b>				
Power of tests included in p-curve <i>(correcting for selective reporting)</i>				
Estimate: 5% 90% Confidence interval: (5%, 18%)				

Figure 12.16: Result of the p-curve analysis of the biased studies in Q2.

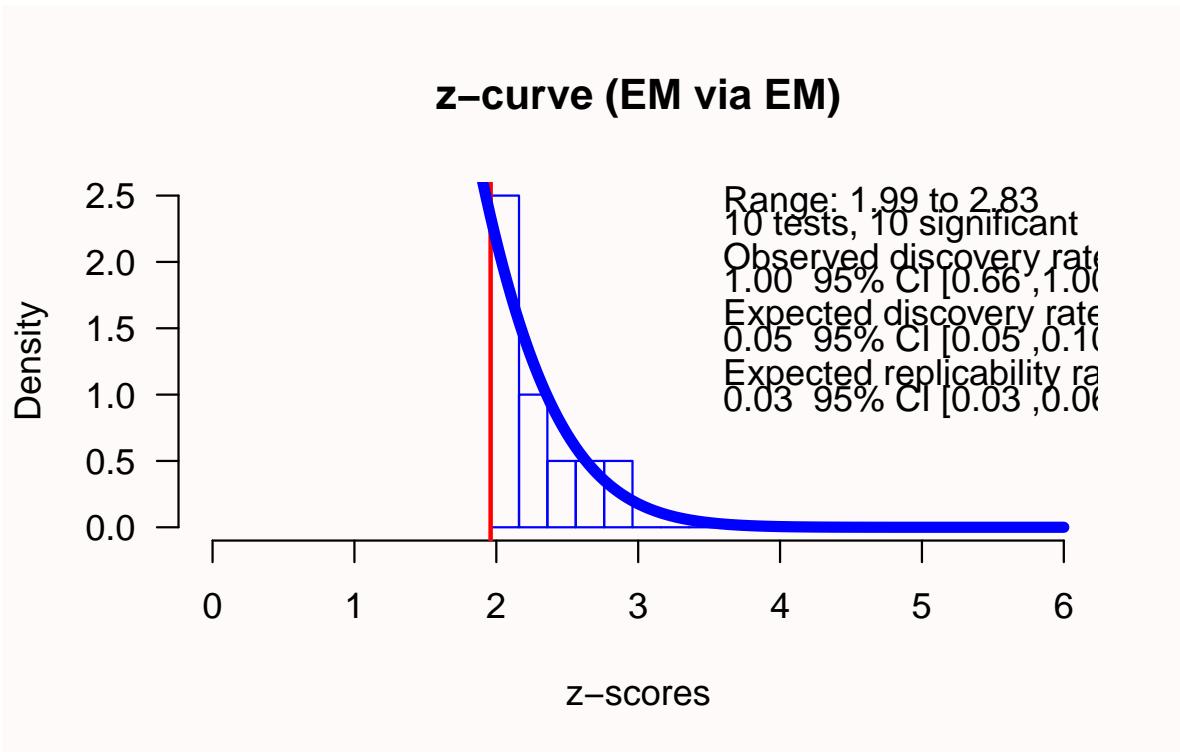


Figure 12.17: Result of the z-curve analysis of the biased studies in Q2.

- (C)  $Z$ -curve is not able to find an indication of bias, as the expected discovery rate and the expected replicability rate do not differ from each other statistically.
- (D) Although the observed discovery rate is 1 (indicating an observed power of 100%) the confidence interval ranges from 0.66 to 1, which indicates that the studies could have a lower but more realistic power, and the fact that 100% of the results were significant could have happened by chance.

**Q7:** We did not yet perform a trim and fill analysis, and given the analyses above (e.g., the  $z$ -curve analysis), which statement is true?

- (A) The trim-and-fill method would most likely not indicate any missing studies to 'fill'.
- (B) The trim-and-fill method has known low power to detect bias, and would contradict the  $z$ -curve or  $p$ -curve analysis reported above.
- (C) The trim-and-fill analysis would indicate bias, but so did the  $p$ -curve and  $z$ -curve analysis, and the adjusted effect size estimate by trim-and-fill does not adequately correct for bias, so the analysis would not add anything.
- (D) The trim-and-fill method provides a reliable estimate of the true effect size, which is not provided by any of the other methods discussed so far, and therefore it should be reported alongside other bias detection tests.

**Q8:** Publication bias is defined as the practice of selectively submitting and publishing scientific research. Throughout this chapter, we have focused on selectively submitting *significant* results. Can you think of a research line or a research question where researchers might prefer to selectively publish *non-significant* results?

### 12.7.1 Open Questions

1. What is the idea behind the GRIM test?
2. What is the definition of 'publication bias'?
3. What is the file-drawer problem?
4. In a funnel plot, what is true for studies that fall inside the funnel (when it is centered on 0)?
5. What is true for the trim-and-fill approach with respect to its ability to detect and correct effect size estimates?

6. When using the PET-PEESE approach, what is important to consider when the meta-analysis has a small number of studies?
7. What conclusions can we draw from the 2 tests that are reported in a p-curve analysis?

# 13 Preregistration and Transparency

For as long as data has been used to support scientific claims, people have tried to selectively present data in line with what they wish to be true. An example of a scientist who did this is Daryl Bem, a parapsychologist who studies whether people have extra-sensory perception that allows them to predict the future. By using selective reporting, and publishing 9 studies in a top journal claiming that people could predict the future, Bem kick-started the replication crisis in psychology back in 2011. In Figure Figure 13.1 you can see the results and discussion from a study he performed (Bem, 2011). In this study, participants pressed a left or right button to predict whether a picture was hidden behind a left or right curtain. At the moment they made the decision, not even the computer had randomly determined where this picture would appear, so any performance better than average would be very surprising.

## Results and Discussion

Across all 100 sessions, participants correctly identified the future position of the erotic pictures significantly more frequently than the 50% hit rate expected by chance: 53.1%,  $t(99) = 2.51$ ,  $p = .01$ ,  $d = 0.25$ .<sup>3</sup> In contrast, their hit rate on the nonerotic pictures did not differ significantly from chance: 49.8%,  $t(99) = -0.15$ ,  $p = .56$ . This was true across all types of nonerotic pictures: neutral pictures, 49.6%; negative pictures, 51.3%; positive pictures, 49.4%; and romantic but nonerotic pictures, 50.2%.

Figure 13.1: Screenshot from the Results and Discussion section of Bem, 2011.

It is clear there are 5 tests against guessing average (for erotic, neutral, negative, positive, and ‘romantic but non-erotic’ pictures). A Bonferroni correction would lead us to use an alpha level of 0.01 (an alpha of 0.05/5 tests) and the main result, that participants guessed the future position of erotic pictures above guessing average, with a  $p$ -value of 0.013, would not have allowed Bem to reject the null hypothesis, given a pre-specified alpha level corrected for multiple comparisons.

Which of the five categories (erotic, neutral, negative, positive, and romantic but non-erotic pictures) would you have predicted people would perform better than guessing average at, if we had evolved the ability to predict the future? Do you think Bem actually predicted an effect for the erotic pictures only, before he had seen the data? You might not trust that Bem predicted an effect only for this specific group of stimuli, and that he was ‘cooking’ - making multitudes of observations, and selecting the significant result, only to **HARK** - hypothesize after the results are known (Kerr, 1998) in his introduction of the study. Do you think other researchers should simply trust that you predicted a reported outcome, if you performed a study with multiple conditions, and you found an effect in only one condition? Or should they be skeptical, and doubt they can take the claims in Bem’s paper at face value?

### 13.1 Preregistration of the Statistical Analysis Plan

In the past, researchers have proposed solutions to prevent bias in the literature due to inflated Type 1 error rates as a result of selective reporting. For example, Bakan (1966) discussed the problematic aspects of choosing whether or not to perform a directional hypothesis test after looking at the data. If a researcher chooses to perform a directional hypothesis test only when the two-sided hypothesis test yields a  $p$ -value between 0.05 and 0.10 (i.e., when a test yields  $p = 0.08$ , the researcher decides after seeing the result that a one-sided test was also warranted, and reports the  $p$ -value as 0.04, one-sided) then in practice the Type 1 error rate is doubled (i.e., is 0.10 instead of 0.05). Bakan (p. 431) writes:

How should this be handled? Should there be some central registry in which one registers one’s decision to run a one- or two-tailed test before collecting the data? Should one, as one eminent psychologist once suggested to me, send oneself a letter so that the postmark would prove that one had pre-decided to run a one-tailed test?

De Groot (1969) already pointed out the importance to “work out in advance the investigative procedure (or experimental design) on paper to the fullest possible extent” which should include “a statement of the confirmation criteria, including formulation of null hypotheses, if any, choice of statistical test(s), significance level and resulting confirmation intervals” and “for each of the details mentioned, a brief note on their rationale, i.e., a justification of the investigator’s particular choices.”

The rise of the internet has made it possible to create online [registries](#) that allow researchers to specify their study design, the sample plan, and statistical analysis plan before the data is collected. A time-stamp, and sometimes even a dedicated Digital Object Identifier (DOI) transparently communicates to peers that the research question and analysis plan were specified before looking at the data. Some tools go even further, such as [OpenSafely](#) which logs all analyses that are performed, and all changes to the analysis code. This is important, because you can’t *test* a hypothesis on the data that is used to generate it. If you come up with a hypothesis by looking at data, the hypothesis might be true, but nothing has been done to

severely test the hypothesis yet. When exploring data, you can perform a hypothesis test, but you cannot *test* a hypothesis.

In some fields, such as medicine, it is now required to register certain studies, such as clinical trials. For example, the [International Committee of Journal Editors](#) writes:

the ICMJE requires, and recommends, that all medical journal editors require registration of clinical trials in a public trials registry at or before the time of first patient enrollment as a condition of consideration for publication.

The use of **study registries** has been promoted by the Food and Drug Administration (FDA) since 1997. In these registries a description of the study and contact information was provided with the main goal to make it easier for the public to take part in clinical trials. From 2000 onwards registries have increasingly been used to prevent bias, and regulations have become increasingly strict in terms of reporting both the primary outcome of studies before data collection, as well as updating the registry with the results after data collection is complete, although these rules are not always followed (Goldacre et al., 2018).

The requirement to register the primary outcome of interest on [ClinicalTrials.gov](#) was correlated with a substantial drop in the number of studies that observed statistically significant results, which could indicate that removing flexibility in how data was analyzed prevented false positive results from being reported. Kaplan and Irvin (2015) analyzed the results of randomized controlled trials evaluating drugs or dietary supplements for the treatment or prevention of cardiovascular disease. They observed how 17 of 30 studies (57%) published before the requirement to register studies on ClinicalTrials.gov yielded statistically significant results, while only 2 out of 25 (8%) studies published after 2000 observed statistically significant results. Of course, correlation is not causation, so we can not conclude there is a causal effect. But if you go to the doctor when you are sick, and the doctor tells you that luckily there are two cures, one proven effective in a study published in 1996, and one from a study published in 2004, which cure would you pick?

When implemented perfectly, study registries allow the scientific community to know about the planned analyses before data collection, and the main result of the planned hypotheses. However, these results might not necessarily end up in the published literature, and this is especially a risk for studies where predictions are not confirmed. One step beyond study registration is a novel publication format known as **Registered Reports**. Journals that publish Registered evaluate studies based on the introduction, method, and statistical analyses, but not on the results (Chambers & Tzavella, 2022; Nosek & Lakens, 2014). The idea to review studies before data collection is not new, and has proposed repeatedly during the last half century (Wiseman et al., 2019). As discussed in the section on [publication bias](#), Registered Reports have a substantially larger probability of reporting findings that do not support the hypotheses compared to the traditional scientific literature (Scheel, Schijen, et al., 2021).

The benefits and downsides of publishing research as a Registered Report are still being examined, but one benefit is that researchers get feedback from expert reviewers at a time where

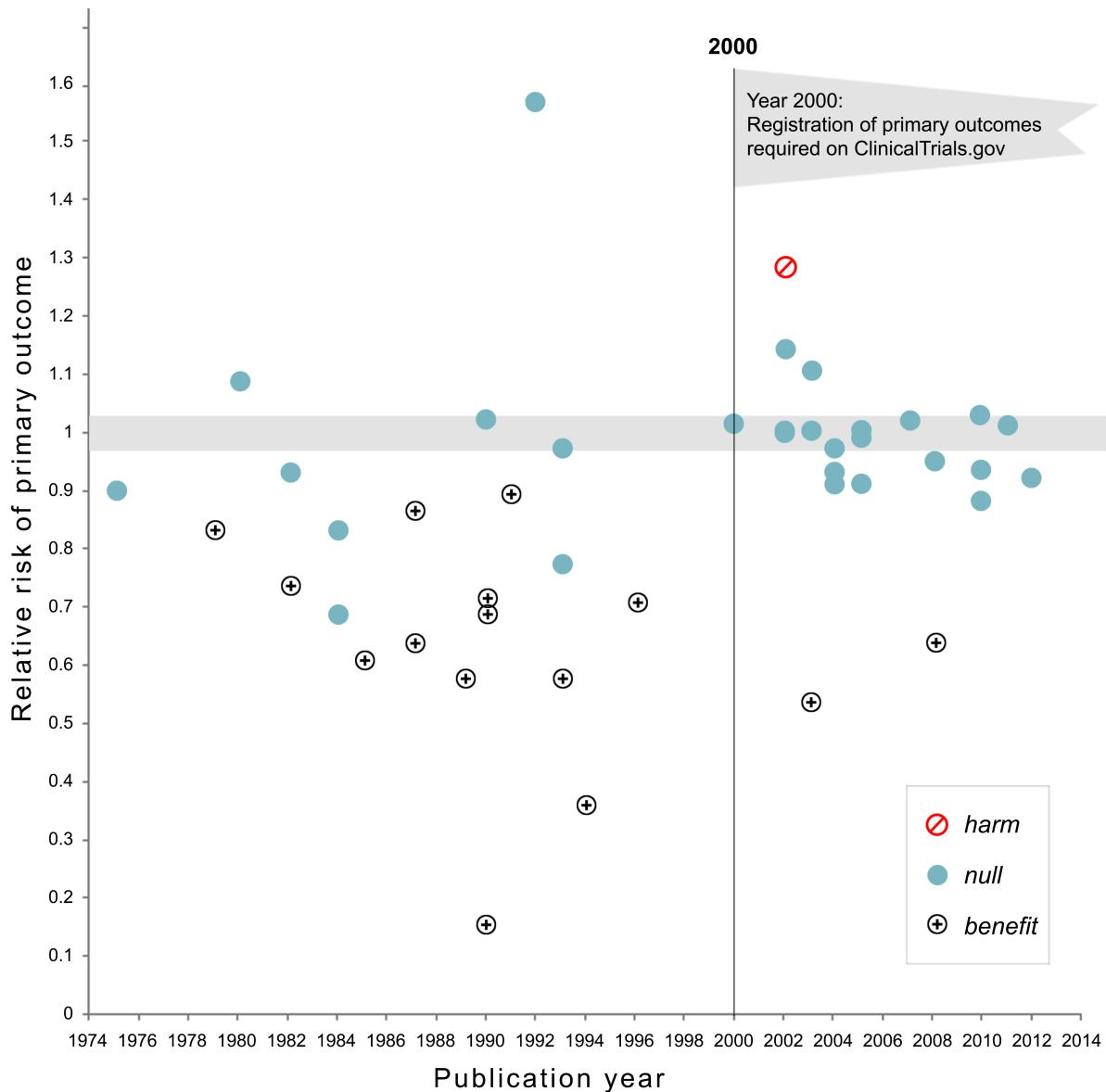


Figure 13.2: Figure from Kaplan and Irvin (2015) showing the substantial drop in statistically significant results after the registration of primary outcomes was required on ClinicalTrials.gov.

they can still improve the study, instead of after data has already been collected. Moving the process of criticism on a study from after the study has been performed (as in traditional peer review) to before the study has been performed (through Registered Reports, or by implementing ‘Red Teams’ Daniël Lakens (2020) or methodological review boards Daniël Lakens (2023) at universities) is an idea worth exploring, and could make the process of scientific criticism more collaborative, as reviewers can help to improve a study, instead of deciding whether or not any flaws in a manuscript are too consequential to recommend publication.

## 13.2 The value of preregistration

Preregistrations are time-stamped documents that describe the analyses researchers plan to perform, while transparently communicating that the analyses have not been selected based on information in the data that determines the result of the planned analyses. Preregistration has the primary goal to allow others to transparently evaluate the capacity of a test to falsify a prediction, or how *severely* a hypothesis has been tested (Daniël Lakens, 2019). The severity of a test is determined by how likely a prediction is proven wrong when it is wrong, and proven right when it is right. During the research process researchers can make decisions that increase the probability their prediction will be statistically supported, even when it is wrong. For example, when Daryl Bem decided which of the 5 sets of stimuli to focus on in his results section, the choice to only focus on erotic stimuli (versus the mean of all stimuli, or the stimuli in another condition, such as negative pictures) was only justified by the fact that the *p*-value ended up being statistically significant. The opposite can also happen, where researchers desire to obtain a *non-significant* test result, and where researchers make decisions that lead to a higher likelihood of not corroborating a prediction (e.g., by reducing the statistical power), even if the prediction was correct. The goal of preregistration is to prevent researchers from non-transparently reducing the capacity of the test to falsify a prediction by allowing readers of their work to see how they planned to test their prediction before they had access to the data, and evaluate whether any changes from their original plan reduce how severely they tested their prediction.

Preregistration adds value for people who, based on their philosophy of science, increase their trust in claims that are supported by severe tests and predictive successes. Preregistration itself does not make a study better or worse compared to a non-preregistered study (Daniël Lakens, 2019). Instead, it merely allows researchers to transparently evaluate the **severity** of a test. The severity of a test is in theory unrelated to whether a study is preregistered. However, in practice, whenever reward structures in science introduce researcher bias, preregistration is likely to increase the severity of tests. Preregistration would not add value when the correct analysis approach was completely clear to researchers and their readers, for example because the theory is so well specified that there is only one rational analysis plan. In most scientific fields, however, theories rarely completely constrain the test of predictions. Despite this, it is important to recognize that there are cases where deviating from a preregistered analysis plan will lead to a *more severely* tested hypothesis. For example, if your preregistration failed to

consider that some participants would be too drunk to respond meaningfully on your task, or when you forgot to specify what you would do if data was not normally distributed, then changing the original analysis plan would be seen by most fellow researchers as a more severe test, and not as a way to increase the probability you would find a statistically significant effect. If you transparently list all deviations from your analysis plan, and provide strong justifications for these deviations, the readers can draw their own conclusions about how severely a hypothesis has been tested.

Preregistration is a tool, and researchers who use it should do so because they have a goal that preregistration facilitates. If the use of a tool is detached from a philosophy of science it risks becoming a heuristic. Researchers should not choose to preregister because it has become a new norm, but they should preregister because they can justify based on their philosophy of science how preregistration supports their goals. There are many types of research for which preregistration is not necessary. Although it is always good to be as transparent as possible when doing research, from a philosophy of science perspective, the unique value of preregistration is limited to research which aims to severely test predictions. Outside of this type of research, transparency (for example by sharing data, materials, and a lab notebook detailing decisions that were made) can be valuable to allow researchers to evaluate results in more detail. In addition to the primary goal of preregistration to allow others to evaluate how severely a prediction has been tested, researchers have reported secondary benefits of preregistering, such as feeling the preregistration improved their experimental design, analysis plan, and theoretical predictions (Sarafoglou et al., 2022). Although it is not necessary to publicly preregister to reap these benefits, a public preregistration can motivate researchers to more carefully think about their study in advance. This use of preregistration was already pointed out by Bakan (1967):

Some years ago I developed a lecture of advice to graduate students on the conduct of research. My intent was hardly cynical. Yet this lecture was consistently interpreted by my students as cynical, a reaction that helped me to understand the very poor state of psychological research. The major point of the lecture, repeatedly made in various presentations of “good experimental design,” was that the manner in which data were to be analyzed and interpreted should be thought out carefully before the data were collected. Ideally, I argued, one should be able to write up the sections defining the problem, reviewing the literature, and explaining the methods used, exactly as they would appear in the final report. One could then proceed to block out the tables that one would report, and to write two or three different versions of the discussion section - without collecting any data whatsoever! Indeed, I argued, it was a good exercise to fill in the tables with some “made up” data to make sure that the data one eventually collected could be used to defend the assertions one would eventually make.

### **13.3 How to preregister**

The more detail a preregistration document has, the easier it is for others to transparently evaluate the severity of the tests that are performed. Because it is difficult to come up with all aspects that one should include, researchers have created websites to guide researchers through this process (e.g., <https://aspredicted.org/>), including submission guidelines, and templates (van 't Veer & Giner-Sorolla, 2016). The template by Van 't Veer and Giner-Sorolla is an excellent start, and is intended as a place to begin for people who have no experience preregistering their research. Another useful paper by Wicherts et al. (2016) provides a checklist of aspects to consider when planning, executing, analyzing, and reporting research.

Although these checklists were useful to introduce scientists to the idea of preregistration, it is important to raise the bar to the level we need to have high quality preregistrations that actually fulfill their goal to allow peer to evaluate the severity of a test. The first step towards this is for authors to follow reporting guidelines in their field. In psychology, this means following the Journal Article Reporting Standards (JARS) (Appelbaum et al., 2018). The reporting guidelines encompass more suggestions than needed for a preregistration document, but I would recommend using JARS both for your preregistration document and when writing up the final report, as it is a very well-thought through set of recommendations. Taking JARS into account when planning or reporting your research is likely to improve your research.

The Journal Article Reporting Standards inform you about information that needs to be present on the title page, the abstract of your paper, the introduction, the method section, the results section, and the discussion. For example, JARS states that you should add an Author Note on the title page that includes “Registration information if the study has been registered”. The method and result sections receive a lot of attention in JARS, and these two sections also deserve a lot of attention in a preregistration. Remember that a severe test has a high probability of finding a predicted effect if the prediction is correct, and a high probability of not finding a predicted effect if the prediction is incorrect. Practices that inflate the Type 1 error rate increase the possibility of finding a predicted effect if a prediction is actually wrong. Low power, unreliable measures, a flawed procedure, or a bad design increase the possibility of not finding an effect when the prediction was actually correct. Incorrect analyses risk answering a question that is unrelated to the prediction researchers set out to test (sometimes referred to as a [Type 3 error](#)). As we see, JARS aims to address these threats to the severity of a test by asking authors to provide detailed information in their methods and results sections.

### **13.4 Journal Article Reporting Standards**

Although in the following I will focus on quantitative experimental studies with random assignment to conditions (you can download the JARS table [here](#)), JARS includes tables for

**TABLE 1 | Checklist for different types of researcher degrees of freedom in the planning, executing, analyzing, and reporting of psychological studies.**

Code	Related	Type of degrees of freedom
Hypothesizing		
T1	R6	Conducting explorative research without any hypothesis
T2		Studying a vague hypothesis that fails to specify the direction of the effect
Design		
D1	A8	Creating multiple manipulated independent variables and conditions
D2	A10	Measuring additional variables that can later be selected as covariates, independent variables, mediators, or moderators
D3	A5	Measuring the same dependent variable in several alternative ways
D4	A7	Measuring additional constructs that could potentially act as primary outcomes
D5	A12	Measuring additional variables that enable later exclusion of participants from the analyses (e.g., awareness or manipulation checks)
D6		Failing to conduct a well-founded power analysis
D7	C4	Failing to specify the sampling plan and allowing for running (multiple) small studies
Collection		
C1		Failing to randomly assign participants to conditions
C2		Insufficient blinding of participants and/or experimenters
C3		Correcting, coding, or discarding data during data collection in a non-blinded manner
C4	D7	Determining the data collection stopping rule on the basis of desired results or intermediate significance testing
Analyses		
A1		Choosing between different options of dealing with incomplete or missing data on <i>ad hoc</i> grounds
A2		Specifying pre-processing of data (e.g., cleaning, normalization, smoothing, motion correction) in an <i>ad hoc</i> manner
A3		Deciding how to deal with violations of statistical assumptions in an <i>ad hoc</i> manner
A4		Deciding on how to deal with outliers in an <i>ad hoc</i> manner
A5	D3	Selecting the dependent variable out of several alternative measures of the same construct
A6		Trying out different ways to score the chosen primary dependent variable
A7	D4	Selecting another construct as the primary outcome
A8	D1	Selecting independent variables out of a set of manipulated independent variables
A9	D1	Operationalizing manipulated independent variables in different ways (e.g., by discarding or combining levels of factors)
A10	D2	Choosing to include different measured variables as covariates, independent variables, mediators, or moderators
A11		Operationalizing non-manipulated independent variables in different ways
A12	D5	Using alternative inclusion and exclusion criteria for selecting participants in analyses
A13		Choosing between different statistical models
A14		Choosing the estimation method, software package, and computation of SEs
A15		Choosing inference criteria (e.g., Bayes factors, alpha level, sidedness of the test, corrections for multiple testing)
Reporting		
R1		Failing to assure reproducibility (verifying the data collection and data analysis)
R2		Failing to enable replication (re-running of the study)
R3		Failing to mention, misrepresenting, or misidentifying the study preregistration
R4		Failing to report so-called "failed studies" that were originally deemed relevant to the research question
R5		Misreporting results and <i>p</i> -values
R6	T1	Presenting exploratory analyses as confirmatory (HARKing)

Figure 13.3: Screenshot of Table 1 in Wicherts et al., 2016, which depicts the checklist for preregistrations.

experiments without randomization, clinical trials, longitudinal designs, meta-analyses, and replication studies. The following items in the JARS table are relevant for a preregistration:

1. *Describe the unit of randomization and the procedure used to generate the random assignment sequence, including details of any restriction (e.g., blocking, stratification).*
2. *Report inclusion and exclusion criteria, including any restrictions based on demographic characteristics.*

This prevents flexibility concerning the participants that will be included in the final analysis.

3. *Describe procedures for selecting participants, including*

- *Sampling method if a systematic sampling plan was implemented*
- *Percentage of the sample approached that actually participated*

You might often not know what percentage of the sample you approach will participate, and getting this information might require some pilot data, as you might not be able to reach the desired final sample size (see below) with the sampling plan.

4. *Describe the sample size, power, and precision, including*

- *Intended sample size*
- *Determination of sample size, including*
  - *Power analysis, or methods used to determine precision of parameter estimates*
  - *Explanation of any interim analyses and stopping rules employed*

Clearly stating the intended sample size prevents practices such as optional stopping, which inflate the Type 1 error rate. Be aware (or if not, JARS will remind you) that you might end up with an achieved sample size that differs from the intended sample size, and consider possible reasons why you might not manage to collect the intended sample size. A sample size needs to be justified, as do the assumptions in a power analysis (e.g., is the expected effect size realistic, or is the smallest effect size of interest indeed of interest to others?). If you used sequential analyses, specify how you controlled the Type 1 error rate while analyzing the data repeatedly as it came in.

Because there is a range of possible sample size justifications, I recommend using the online Shiny app that accompanies the [sample size justification](#) chapter. The Shiny app can be found [here](#). The Shiny app guides you through four steps.

First, researchers should specify the population they are sampling from. To describe the sample, researchers can simply follow the JARS guidelines, such as the Quantitative Design Reporting Standards:

Report major demographic characteristics (e.g., age, sex, ethnicity, socioeconomic status) and important topic-specific characteristics (e.g., achievement level in studies of educational interventions). In the case of animal research, report the genus, species, and strain number or other specific identification, such as the name and location of the supplier and the stock designation. Give the number of animals and the animals' sex, age, weight, physiological condition, genetic modification status, genotype, health-immune status, drug or test naivete, and previous procedures to which the animal may have been subjected. Report inclusion and exclusion criteria, including any restrictions based on demographic characteristics.

Researchers should also indicate if they can collect data from the entire sample (in which case the sample size justification is completed), and if not, which resource limitations they have (e.g., the time and money they have available for data collection).

In the second step, researchers should consider which **effects of interest** they can specify. Ideally, they are able to determine a smallest effect size of interest, but other approaches can also be used. In the third step researchers specify an **inferential goal** such as test a hypothesis, or measure an effect with accuracy. Finally, researchers should specify the total sample size (based on the number of participants and the number of observations per participant) and explain the informational value of the study (e.g., why is the sample size large enough to yield an informative answer to the research question?). After filling out the relevant fields in the Shiny app, researchers can download a PDF file that contains a complete sample size justification.

##### *5. Describe planned data diagnostics, including*

- *Criteria for post-data collection exclusion of participants, if any*
- *Criteria for deciding when to infer missing data and methods used for imputation of missing data*
- *Defining and processing of statistical outliers*
- *Analyses of data distributions*
- *Data transformations to be used, if any*

After collecting the data, the first step is to examine the data quality, and test assumptions for the planned analytic methods. It is common to exclude data from participants who did not follow instructions, and these decision procedures should be prespecified. At each preregistration you will discover additional unforeseen consequences that will be added to these sections. If data is missing, you might not want to remove a participant entirely, but use a method to impute missing data. Because outliers can have an undue influence on the results, you might want to preregister ways to mitigate the impact of outliers. For practical recommendations on how to classify, detect, and manage outliers, see (Leys et al., 2019). If you are planning to perform statistical tests that have assumptions (e.g., the assumption of normality for Welch's *t*-test) you need to preregister how you will decide whether these assumptions are met, and if not, what you will do.

6. *Describe the analytic strategy for inferential statistics and protection against experiment-wise error for*

- Primary hypotheses
- Secondary hypotheses
- Exploratory hypotheses

The difference between these three levels of hypotheses is not adequately explained in the JARS material, but H. Cooper (2020) explains the distinction a *bit* more, although it remains quite vague. The way I would distinguish these three categories is as follows. First, a study is designed to answer a **primary hypothesis**. The Type 1 and Type 2 error rates for this primary hypothesis are as low as the researcher can afford to make them. **Secondary hypotheses** are questions that a researcher considers interesting when planning the study, but that are not the main goal of the study. Secondary hypotheses might concern additional variables that are collected, or even sub-group analyses that are deemed interesting from the outset. For these hypotheses, the Type 1 error rate is still controlled at a level the researchers consider justifiable. However, the Type 2 error rate is not controlled for secondary analyses. The effect that is expected on additional variables might be much smaller than the effect for the primary hypothesis, or analyses on subgroups will have smaller sample sizes. Therefore, the study will yield an informative answer if a significant effect is observed, but a non-significant effect can not be interpreted because the study lacked power (both to reject the null hypothesis, or for an equivalence test). By labeling a question as a secondary hypothesis, a researcher specifies in advance that non-significant effects will not lead to clear conclusions.

Finally, there is a left-over category of analyses that are performed in an article. I would refer to this category as **exploratory results**, not exploratory hypotheses, because a researcher might not have hypothesized these analyses at all, but comes up with these tests during data analysis. JARS requires researchers to report such results ‘in terms of both substantive findings and error rates that may be uncontrolled’. An exploratory result might be deemed impressive by readers, or not, depending on their prior belief, but it has not been severely tested. All findings need to be independently replicated if we want to be able to build on them - but all else equal, this requirement is more imminent for exploratory results.

## 13.5 Deviating from a Preregistration

Although some researchers manage to report a study that was performed exactly in line with their preregistration, many researchers deviate from their preregistration when they perform their study and analyze their data (Akker et al., 2023). Common reasons for deviations are collected sample sizes that do not match the preregistered sample size, excluding data from the analysis for reasons not prespecified, performing a different statistical test than preregistered, or implementing changes to the analysis plan due to errors during the data collection. A deviation from a preregistration would occur when researchers preregistered to analyze all

data, but after inspection of the data, decide to exclude a subset of the observations, and subsequently use the results of the analysis based on a subset of the data as the basis of their claim, while the analysis they originally planned is ignored.

The goal of a statistical hypothesis test in an error statistical philosophy is to make valid claims that are severely tested (Mayo & Spanos, 2011). One justifiable reason to deviate from a preregistered statistical analysis plan is to increase the validity of the scientific claim – even if doing so comes at the expense of the severity of the test. Validity refers to “the approximate truth of an inference” (Shadish et al., 2001). When researchers can make a convincing argument that the preregistered analysis plan leads to a statistical test with low validity, a less severe but more valid test of the hypothesis might lead to a claim that has more [verisimilitude](#), or truth-likeness (Niiniluoto, 1998).

Both validity, which is a property of the inference, and severity, which is a property of the test are continuous dimensions. A statistical test can be more or less severe, and the inference can be more or less valid. It is important to note that in practice a claim based on a hypothesis test that contains a deviation from a preregistration will be more severely tested than a claim based on a non-preregistered test. Such deviations should not just be reported, but the consequences of the deviation should also be evaluated. Table 1 provides four examples of tests with lower or higher severity and lower or higher validity.

	<b>Lower validity</b>	<b>Higher validity</b>
<b>Lower severity</b>	Selectively reporting one out of five variables that measure a construct of interest because only this test yields $p < .05$ .	Deviating from a preregistration to exclude observations not caused by processes related to the research question.
<b>Higher severity</b>	Following a preregistered analysis of all data even though 15% of respondents did not follow the instructions.	Following a preregistered statistical analysis plan with high construct and statistical validity.

Figure 13.4: Examples of reporting practices that lead to tests with higher or lower severity and claims with higher or lower validity.

There are different reasons to deviate from a preregistration. Lakens (2023) distinguishes: 1) unforeseen events, 2) errors in the preregistration, 3) missing information, 4) violations of untested assumptions, and 5) falsification of auxiliary hypotheses. Some deviations have no impact on the test severity, while others decrease the severity substantially, even if they

are often still more severe than non-preregistered tests. Under some circumstances deviating from a preregistration can increase the severity of a test. It can also be justified to deviate from a preregistration if doing so increases the validity of the inference. For every deviation clearly specify when, where, and why a deviation from a preregistration occurred, followed by an evaluation of the impact of the deviation on the severity of the test (and where relevant, the validity of the inference). Forms to report deviations are available online (e.g., <https://osf.io/6fk87> and <https://osf.io/yrvcg>).

## 13.6 What Does a Formalized Analytic Strategy Look Like?

A hypothesis test is a methodological procedure to evaluate a prediction that can be described on a **conceptual level** (e.g., “Learning how to preregister improves your research”), an **operationalized level** (e.g., “Researchers who have read this text will control their alpha level more carefully, and they will more precisely specify what would corroborate or falsify their prediction in a preregistration document”), and a **statistical level** (e.g., “An independent *t*-test comparing coded preregistration documents written by people who read this text will show a statistically lower number of ways in which the hypothesis could be tested, which implies more careful Type 1 error control, compared to people who did not read this text”). In a pre-registration document, the goal should be to specify the hypothesis in detail at the statistical level. Furthermore, each statistical hypothesis should be clearly linked to the conceptual and operationalized level. In some studies people perform multiple tests, and it is often not clear which pattern of results would falsify the researchers’ predictions. Currently, preregistrations differ widely in how detailed they are, and not all preregistration have sufficient detail to treat them as confirmatory tests of predictions (Waldron & Allen, 2022).

Preregistration is a relatively new practice for most researchers. It should not be surprising that there is often quite some room for improvement in the way researchers preregister. It is not sufficient to preregister - the goal is to preregister well enough so that others can evaluate the severity with which you tested your hypothesis. How do we do this? First, it is important to acknowledge that it is difficult to describe a hypothesis verbally. Just like we use notation to describe statistics because it removes ambiguity, verbal descriptions of hypotheses rarely sufficiently constrain potential flexibility in the data analysis.

For example, in the verbal description of a statistical hypothesis in the previous paragraph (which read “An independent *t*-test comparing coded preregistration documents written by people who read this text will show a statistically lower number of ways in which the hypothesis could be tested, which implies more careful Type 1 error control, compared to people who did not read this text”) it is not clear what alpha level I plan to use for the *t*-test, or whether I will perform Student’s *t*-test or Welch’s *t*-test. Researchers often implicitly treat a  $p > 0.05$  as falsifying a prediction, but this is a common misconception of ***p*-values**, and a hypothesis is often better falsified using a statistical test that can reject the presence of predicted outcomes,

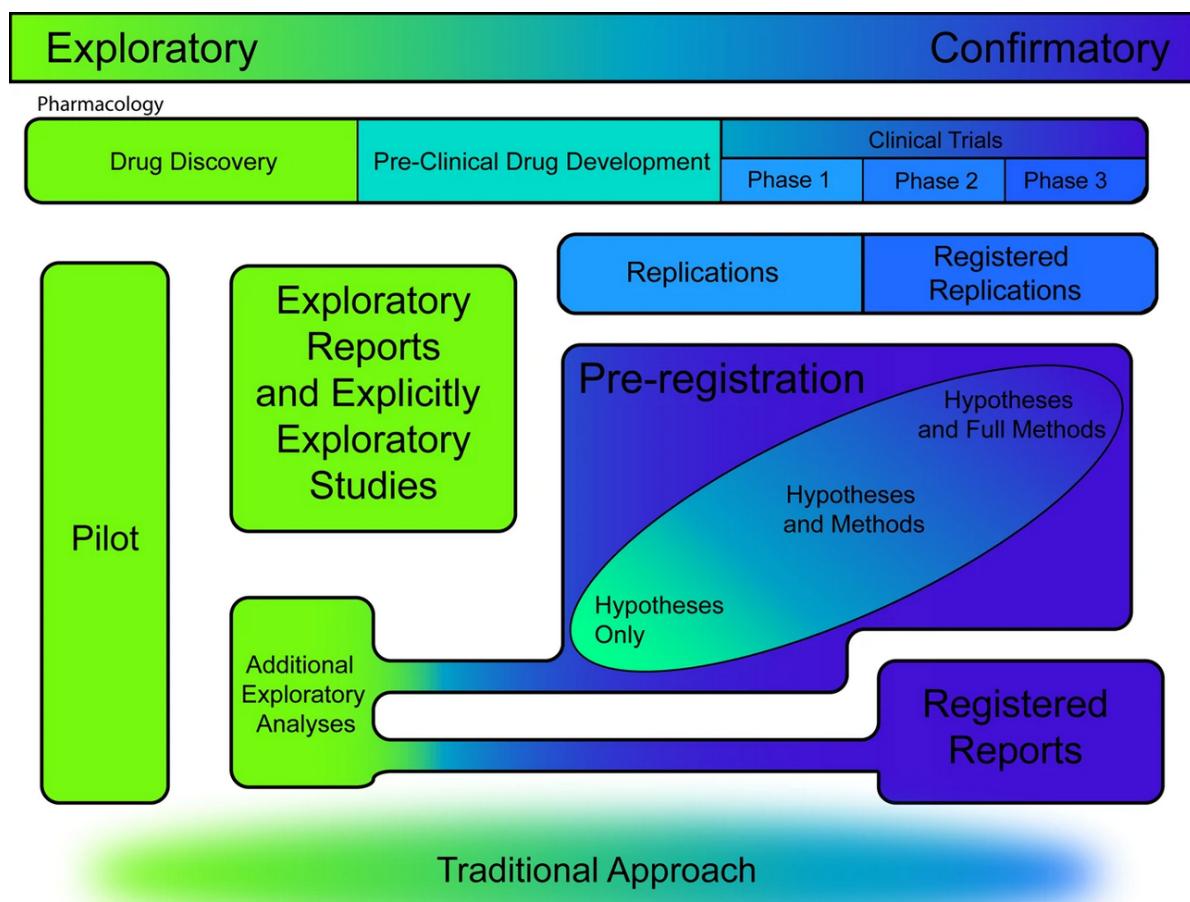


Figure 13.5: Different study types plotted on a dimension from fully exploratory to fully confirmatory (from Waldron & Allen, 2022).

such as an [equivalence test](#). Specifying explicitly how you will evaluate a hypothesis makes it clear how your prediction will be proven wrong.

In Daniël Lakens & DeBruine (2020) we discuss how a good way to remove ambiguity in a hypothesis test described in a preregistration document is to make sure it is [machine readable](#). Machines are notoriously bad at dealing with ambiguous descriptions, so if the hypothesis is understandable for a machine, it will be clearly specified. A *hypothesis* is tested in an *analysis* that takes *data* as input and returns *test results*. Some of these test results will be compared to *criteria*, used in the *evaluation* of the test result. For example, imagine a *hypothesis* predicts that the mean in one group will be higher than the mean in another group. The *data* is *analyzed* with Welch's *t*-test, and if the *resulting p-value* is smaller than a specified *criterion alpha* (e.g., 0.01) the prediction is *evaluated* as being *corroborated*. Our prediction is *falsified* if we can reject effects deemed large enough to matter in an equivalence test, and the result is *inconclusive* otherwise. In a clear preregistration of a hypothesis test, all these components (the analysis, the way results will be compared to criteria, and how results will be evaluated in terms of corroborating or falsifying a prediction) will be clearly specified.

The most transparent way to specify the statistical hypothesis is in **analysis code**. The gold standard for a preregistration is to create a simulated dataset that looks like the data you plan to collect, and write an analysis script that can be run on the dataset you plan to collect. Simulating data might sound difficult, but there are [great packages](#) for this in R, and an increasing number of tutorials. Since you will need to perform the analyses anyway, doing so before you collect the data helps you to carefully think through your experiment. By preregistering the analysis code, you make sure all steps in the data analysis are clear, including assumption checks, exclusion of outliers, and the exact analysis you plan to run (including any parameters that need to be specified for the test). For some examples, see <https://osf.io/un3zx>, <https://osf.io/c4t28>, and section 25 of <https://osf.io/gjsft/>.

In addition to sharing the analysis code, you will need to specify how you will **evaluate** the test result when the analysis code is run on the data you will collect. This is often not made explicit in preregistrations, but it is an essential part of a hypothesis test, especially when there are multiple primary hypotheses, such as in our prediction that “Researchers who have read this text will become better at controlling their alpha level *and* more clearly specify what would corroborate or falsify their prediction”. If our hypothesis really predicts that both of these outcomes should occur, then the evaluation of our hypothesis should specify that the prediction is falsified if only one of these two effects occurs.

## 13.7 Are you ready to preregister a hypothesis test?

It should often happen that the theory you use to make predictions is not strong enough to lead to falsifiable hypotheses. Especially early in research lines there are too many uncertainties about which analyses we will run, or which effects would be too small to matter. At these stages in the research process, it is more common to have a cyclical approach where researchers

do an experiment to see what happens, use the insights to reformulate their theory, and design another experiment. The philosopher of science Van Fraassen summarizes this in his statement: “experimentation is the continuation of theory construction by other means.” During this process, we often need to examine whether certain assumption we make hold. This often requires tests of **auxiliary hypotheses** concerning the measures and manipulations we use (Uygun Tunç & Tunç, 2022).

As you prepare a preregistration document, you might be faced with many uncertainties that you don’t exactly know how to address. It might be that this is a sign that you are not yet ready to preregister a prediction. When testing hypotheses, corroborating a prediction should be impressive, and falsifying a prediction should be consequential for the theory you are testing. If you make arbitrary choices as you write down your predictions, the test might be neither impressive nor consequential. Sometimes you just want to collect data to describe, examine the relationship between variables, or explore boundary conditions, without testing anything. If that’s the case, don’t feel forced into a hypothesis testing straight-jacket (Scheel, Tiokhin, et al., 2021). Of course, such a study also does not allow you to make any claims that have been severely tested, but that should not be the goal of every study, especially in new research lines.

## 13.8 Test Yourself

In this assignment we will go through the steps to complete a high-quality preregistration. This assignment will continue in the next chapter, where we will focus on a computationally reproducible analysis pipeline and implementing open science practices such as sharing data and code. **Open science** is a set of practices for reproducibility, transparency, sharing, and collaboration based on openness of data and tools that allows others to reuse and scrutinize research. You might want to complete this assignment for a real research project you are involved in. If you are not involved in any real research projects, you can perform a simple study analyzing publicly accessible data just for this assignment. I will illustrate this using a hypothesis that can be answered based on movie ratings on the [Internet Movie Database \(IMDB\)](#). You can come up with any hypothesis you want based on another data source (but don’t spend too much time on data collection, as that is not the goal of this assignment).

To organize your preregistration, you can follow templates others have created, which you can find at <https://osf.io/zab38/wiki/home/>. The default OSF preregistration template is good for hypothesis testing studies. Keep the JARS reporting guidelines in mind while writing your preregistration.

One of my favorite movies is Fight Club. It stars Brad Pitt and Edward Norton. At a *conceptual level* my hypothesis is that Brad Pitt and Edward Norton are both great actors, and because they are both great actors, the movies they play in are equally good. At an *operationalized level* my hypothesis is that on average movies that star Brad Pitt and Edward

Norton will receive the same rating in the Internet Movie Database. IMDB provides both an IMDB rating, and the metascore (provided by metacritic.com).

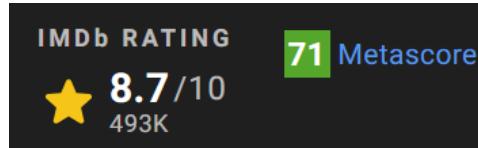


Figure 13.6: Screenshot of a IMDB and metacritic rating.

I will operationalize movie ratings as the IMDB scores, and movies that star Brad Pitt or Edward Norton as all movies they have appeared in, according to the following two search commands on IMDB:

For Brad Pitt [http://www.imdb.com/filmosearch?role=nm0000093&explore=title\\_type&mode=detail&page=1](http://www.imdb.com/filmosearch?role=nm0000093&explore=title_type&mode=detail&page=1)

For Edward Norton: [http://www.imdb.com/filmosearch?role=nm0001570&explore=title\\_type&mode=detail&page=1](http://www.imdb.com/filmosearch?role=nm0001570&explore=title_type&mode=detail&page=1)

**Q1:** Write down your hypothesis on a *conceptual level*. In some research areas you might be able to capture this hypothesis in a formal quantitative model, but more often you will describe this hypothesis verbally. Try to be as precise as possible, even though all verbal descriptions of hypotheses will inherently be limited. It is also useful to discuss whether the data has already been collected, or not, and explain whether your hypothesis is influenced by knowledge about the data (in a typical preregistration, data will not have been collected).

**Q2:** Write down your hypothesis on an *operationalized level*. All variables in the study (i.e., independent and/or dependent variables) should be clearly specified.

The next steps are to specify the hypothesis on a statistical level. Sometimes resources limit the statistical questions one can answer. Whenever this is the case, it is useful to first perform a sample size justification. If this is not the case, it is more useful to first specify the statistical question and then calculate the required sample size. As I already know that the number of movies Brad Pitt and Edward Norton appear in is limited, I will first proceed with a sample size justification.

**Q3:** Justify your sample size. I will use my own Shiny app to go through the steps of a [sample size justification](#).

1.1: Describe the population you are sampling from.

The population consists of all movies that Brad Pitt and Edward Norton starred in (up to March 2023) since the start of their career, as indexed by the internet movie database ([www.imdb.com](http://www.imdb.com)). The total number of observations is limited by the movies Brad Pitt and Edward Norton have appeared in to date, which is 62 and 39, respectively.

1.2: Can you collect data from the entire population?

yes.

The total number of observations is limited by the movies Brad Pitt and Edward Norton have appeared in too date, which is 62 and 39, respectively.

## 2. Which Effect Sizes are of Interest?

The smallest effect size of interest is always a matter of discussion among peers. In this case, I personally believe a difference in movie ratings that is less than 0.5 on a 10 point scale (as used on the IMDB) is sufficiently small to support my prediction that movies with Brad Pitt and Edward Norton are equally good. In other words, if the raw difference is larger than -0.5 and smaller than 0.5, I will conclude the two sets of movies are rated equally well.

The minimal statistically detectable effect given the number of movies (62 and 39) can be computed by examining the effect we have 50% power for in an independent *t*-test, for example as computed by G\*Power. Before we can compute the minimal statistically detectable effect, we need to specify our alpha level. We know our sample size is limited, and statistical power will be an issue. At the same time, we need to make a decision based on the available data, as it will take many years before we have a larger sample size. In such a context where the sample size is fixed, and a decision must be made, it is sensible to balance the Type 1 and Type 2 error in a compromise power analysis. Below, we determine an alpha level of 0.15 is a defensible decision. That means we have a rather high probability of incorrectly concluding the two groups of movies have the same rating, but we do so to reduce the probability of incorrectly concluding the two movies do not have the same rating, when they actually have.

– Sunday, March 05, 2023 – 12:19:19

t tests - Means: Difference between two independent means (two groups)

Analysis: Sensitivity: Compute required effect size

Input: Tail(s) = Two

```
$\alpha$ err prob = 0.15  
Power (1-$\beta$ err prob) = 0.5  
Sample size group 1 = 62  
Sample size group 2 = 39
```

Output: Noncentrality parameter  $\delta$  = 1.4420104

Critical t = 1.4507883

Df = 99

Effect size d = 0.2947141

The Minimal Statistically Detectable Effect is thus  $d = 0.295$ .

A sensitivity analysis shows that given the sample size we can collect ( $n = 62$  and  $39$ ), and the smallest effect size of interest (half a scale point), and assuming a standard deviation of movie ratings of  $sd = 0.9$  (an estimate based on pilot data), statistical power is 91% if we assume the difference in movie ratings is exactly 0. It is possible that movie ratings actually differ slightly, and the difference is not exactly 0. If we assume the true difference in movie ratings is 0.1 power is still 86%. This decent statistical power is a direct consequence of increasing the alpha level. With a 5% Type 1 error rate, power would be 64% (assuming a true difference between movies of 0.1), and the combined error rate would be  $((100 - 64) + 5) = 41\%$ , but by increasing the alpha level to 15% the combined error rate has dropped to  $((100 - 86) + 15) = 29\%$ , which means the probability of making an error, assuming  $H_0$  and  $H_1$  are equally likely to be true, has been reduced.

```
TOSTER::power_t_TOST(  
  n = c(62, 39),  
  delta = 0.1,  
  sd = 0.9,  
  low_eqbound = -0.5,  
  high_eqbound = 0.5,  
  alpha = 0.05,  
  type = "two.sample"  
)
```

We can conclude that we will have decent power for our planned test, given our smallest effect size of interest and our high alpha level.

### 3. Inferential goal

Our inferential goal is to conduct a statistical test while controlling our error rates, and therefore, we plan to make a decision. We use the sensitivity analysis above to justify our error rates (but we could also have used a compromise power analysis by more formally minimizing the combined error rate, Maier & Lakens (2022)).

### 4. Informational Value of the Study

Finally, we evaluate the informational value of our study. First, we are using all available data, and we have tried to reduce the combined Type 1 and Type 2 error rate by somewhat balancing the two error rates. Our goal is to make a decision based on the available data. Our decision has a relatively high probability to be wrong, but the value of our study is that it allows us to make a decision as well as possible, given the available data. Therefore, if anyone else wants to know the answer to the question if movies starring Brad Pitt and Edward Norton are indeed equally good, our results will give the best possible answer we currently have available, even if there is substantial remaining uncertainty after the study.

**Q4:** Write down your hypothesis on a *statistical level* and specify the code for this analysis. Be as specific as possible. Look through the JARS recommendations above, and the checklist by Wicherts et al. (2016), to make sure you did not miss any details you need to specify (e.g., how will you pre-process the data, which software version will you use, etc.).

We can now specify the test we plan to perform on a statistical level. We do not expect any missing data or outliers, and will analyze all movie ratings in an equivalence test with equivalence bounds of -0.5 and 0.5, and alpha level of 0.15, and as group sizes are unequal Welch's *t*-test (which does not assume equal variances) will be performed Delacre et al. (2017). The following analysis code (which I will run in R version 4.2.0, and TOSTER package version 0.4.1) assumes the data will be stored in the `imdb_ratings` dataframe with a column for the movie ratings for Brad Pitt (with the column name `brad_pitt_score`) and a column for the movie ratings for Edward Norton (with the column name `edward_norton_score`)

```
TOSTER::t_TOST(  
  x = imdb_ratings$brad_pitt_score,  
  y = imdb_ratings$edward_norton_score,  
  low_eqbound = -0.5,  
  high_eqbound = 0.5,  
  eqbound_type = "raw",  
  alpha = 0.15,  
  var.equal = FALSE  
)
```

Finally, we need to specify our criteria and how we will evaluate the results. The `t_TOST` function will also perform a null-hypothesis significance test, which is convenient, because we can now consider our hypothesis supported if the equivalence test is significant at  $p < 0.15$ , or when the 70% confidence interval falls completely within the equivalence bounds. We can consider our hypothesis falsified if the null hypothesis significance test is significant at  $p < 0.15$ . If neither of the two tests is significant, our results are inconclusive. If both tests are significant, our hypothesis is also falsified, as there is an effect, but it is too small to matter. So more formally, our hypothesis is corroborated if the TOST  $p < 0.015$  & NHST  $p > 0.015$ , it is falsified if NHST  $p < 0.015$ , and inconclusive otherwise.

We have now completed a preregistration of a very simple study. In real studies, the preregistration process is often not this easy. A sample size justification often requires some [knowledge about the variables in your analysis](#), power analyses become more uncertain the more complex the analysis, and it might be challenging to make decisions about data pre-processing and how to deal with outliers. If you struggle with your preregistration, you might simply not be [ready to preregister](#). You might not be in the *tightening phase* or your research, or the equivalent of a phase 3 trial in clinical trials, and need to perform more descriptive research before you can perform a [real test](#) of your hypothesis.

### **13.8.1 Practical Aspects of an Online Preregistration**

Over the last years multiple online services have been created that allow you to preregister your hypothesis plan. I will discuss three solutions: ZPID, OSF, and AsPredicted. In decreasing order, these three services differ in how high they set the bar for researchers to preregister on their platform. ZPID is specific for psychological science, the OSF is accessible for anyone, and AsPredicted requires an email address from an academic institution to preregister.

If you preregister, your preregistration will be archived. This will cost time (at ZPID due to manual checks) and money (because of long-term data storage). You should only preregister real studies (even though AsPredicted allows class assignments as preregistrations). For an assignment it suffices to store a PDF file that contains all information related to your preregistration, without actually creating a time-stamped version in one of these databases.

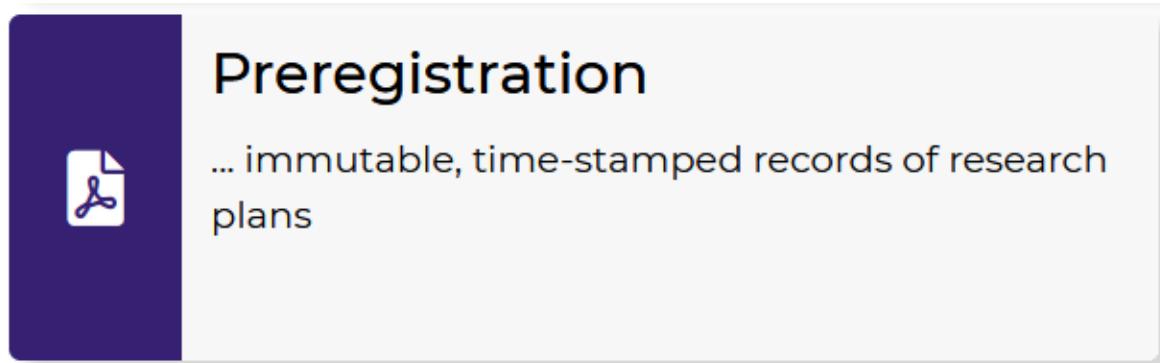
### **13.8.2 Pre-registering on PsychArchives by ZPID**

Go to <https://psa.psycharchives.org>. Log in with your ORCID if you have one or can make one (which should be available to most people working or studying at a research institution), or make a dedicated Leibniz Psychology account.

Click ‘Start a new submission’

**Start new submission**

and scroll down to the preregistration option and click it.

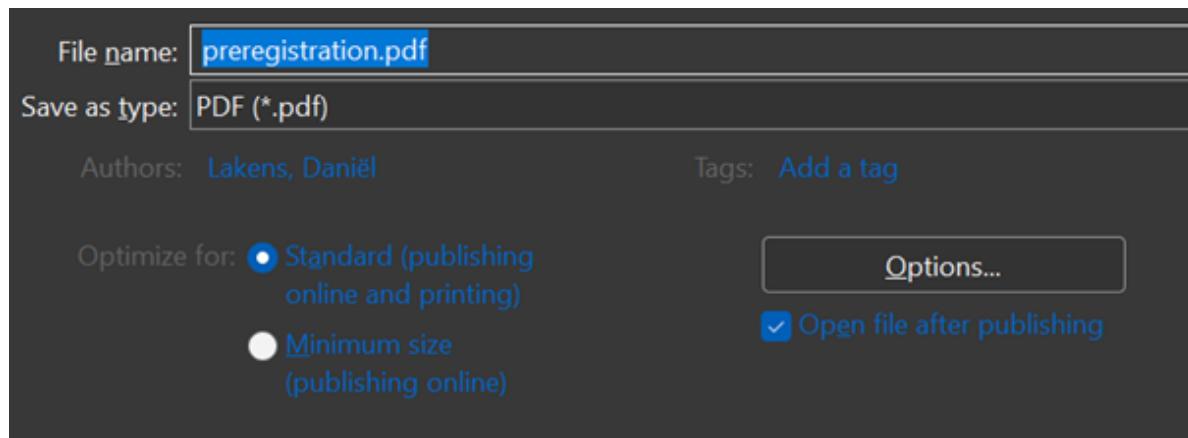


You will see a recommendation to submit your preregistration in a PDF/A file, which is a PDF which does not allow certain restrictions that hinder long-term archiving. This already reveals how PsychArchives will ask you to meet certain standards that they deem best practices, that you might not think about yourself. This is a good thing!

In Microsoft word, you can save a file as a PDF/A compliant pdf file by choosing ‘File’> ‘Save As’, choose PDF from the dropdown menu, click on ‘More options...’ below the dropdown menu:



Click the ‘Options...’ button:



And check the box: PDF/A compliant

## PDF options



**PDF/A compliant**



**Optimize for image quality**



**Bitmap text when fonts may not be embedded**



**Encrypt the document with a password**

**OK**

**Cancel**

When opening a PDF/A compliant PDF, some PDF readers will show a warning message:



This file claims compliance with the PDF/A standard and has been opened read-only to prevent modification.

Click ‘Next’. You can now upload a preregistration document. PsychArchives will motivate you to add descriptions and meta-data to files.

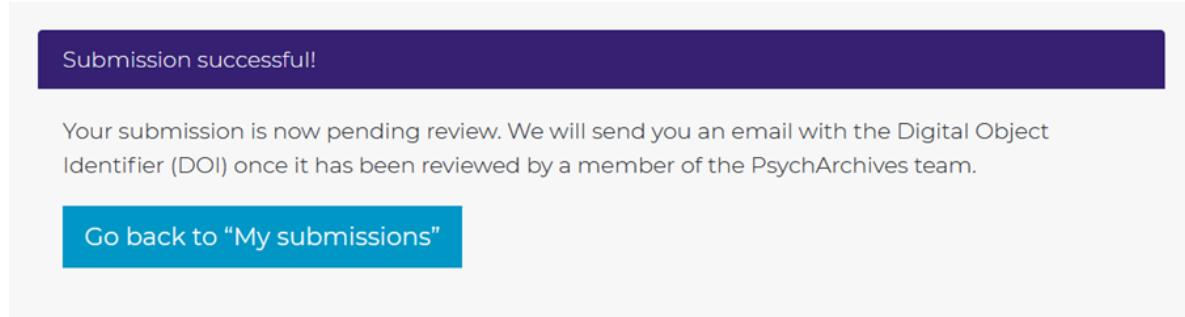
Filename	Delete	Download	Description
preregistration.pdf			Preregistration Document

You then choose a sharing level (level 0 means the file is for public use, level 1 is for scientific use only) and a license. Click ‘Save and Next’.

Now we see why PsychArchives is setting a higher bar than some other services, because we need to specify meta-data for our file. This meta-data will make the files we share (such as a preregistration, the paper based on this preregistration, and the data and code used in the paper) more easily findable, which is important if the scientific community wants to reap the benefits of open science in the future. This preregistration is not yet linked to any other files, but in a real research project, future submissions to PsychArchives would be linked to this

preregistration. Adding good descriptions about files also helps others to find them in the future.

When you submit your preregistration, you will see that your preregistration will be manually checked. Staff at PsychArchives will check whether your submission adheres to all guidelines. If not, you will be told what to improve, and have to resubmit the file. Note that this check is about the uploaded files and meta-data – staff does not check the quality of your preregistration!



You can find the PsychArchives preregistration at the following persistent identifier link:  
<https://doi.org/10.23668/psycharchives.12575>

### 13.8.3 Pre-registering on the Open Science Framework

Go to [www.osf.io](http://www.osf.io), create an account or log in, and click the button ‘Create new project’. Enter a title. If you are from Europe, and you want to adhere to the GDPR privacy regulations, make sure to select ‘Germany’ as the storage location. Click ‘Create’, and click ‘Go to project’.

## Create new project

X

### Title

Are Brad Pitt and Edward Norton Equally Good Actors?

### Storage location

Germany - Frankfurt



► More

Cancel

Create

Add a description, so that others will understand what this project is about, and a license, so that others know how they can re-use any material you are sharing.

You can choose to make your project public now, or to make it public later. It is common for researchers to make the project public upon publication, but if you are not worried about others re-using your content (and any ideas in it) you can make the project open immediately. To do so, click the 'Make public' button in the top right of the project.

To preregister a study, click the 'Registrations' button in the top bar, and click 'new registration'.

**Reminder:** You should not use the OSF to preregister just to practice how to preregister. The OSF registries website is intended to be a searchable database of all official scientific registrations and preregistrations. When you complete a registration, it will remain in [OSF Registries](#) for ever, and there is no way to remove it. This costs money and reduces the usefulness of the database. Only register real studies.

For hypothesis testing research the default OSF preregistration template will provide a useful structure to your preregistration. It will also allow you to upload supplementary files (such as the html file detailing our sample size justification that we created above).

The screenshot shows the OSF (Open Science Framework) interface. At the top, there's a navigation bar with links for 'My Projects', 'Search', 'Support', and a user profile for 'Rebecca Rosenblatt'. Below the navigation bar, there's a sub-navigation bar with 'Templates of OSF Registration Forms', 'Files' (which is selected), 'Wiki', 'Analytics', 'Registrations', and 'Forks'. The main content area displays a Microsoft Word document titled 'AsPredicted registration.docx (Version: 1)'. The document contains a list of questions for pre-registration:

**AsPredicted registration:**

1. Have any data been collected for this study already? (optional)
  - Yes, at least some data have been collected for this study already
  - No, no data have been collected for this study yet
2. What's the main question being asked or hypothesis being tested in this study? (optional)
3. Describe the key dependent variable(s) specifying how they will be measured. (optional)
4. How many and which conditions will participants be assigned to? (optional)
5. Specify exactly which analyses you will conduct to examine the main question/hypothesis. (optional)
6. Any secondary analyses? (optional)
7. How many observations will be collected or what will determine the sample size? No need to justify decision, but be precise about exactly how the number will be determined. (optional)

You will be asked a range of relevant questions to complete. For a completed preregistration related to the study comparing movie ratings of movies starring Brad Pitt and Edward Norton, see: <https://doi.org/10.17605/OSF.IO/RJYCP>. Note that all pre-registrations on the OSF will become public after four years. This is the only platform where you will not be able to keep your preregistrations private indefinitely.

To share your preregistration during the review process, the OSF allows you to create a 'view-only' link. As long as you did not enter identifying information in any of the files in your project, peer reviewers will be able to see your preregistration, without revealing your identity: <https://help.osf.io/article/201-create-a-view-only-link-for-a-project>.

### 13.8.4 Pre-registering on AsPredicted

AsPredicted offers a preregistration service with a focus on simplicity. The website will not allow preregistrations that are too long. This can typically be accomplished by removing all justifications for choices from a preregistration document. This makes life easier for people who need to read the preregistration. AsPredicted focuses on distinguishing exploratory from confirmatory research, but because not all deviations from an analysis plan reduce the severity of a test, I think it is important to make sure the word limit does not limit the ability of peers to evaluate whether changes to the original analysis plan increase or reduce the severity of a test. If you feel the word limit constrains you when writing your preregistration, you can use the AsPredicted template on the OSF to preregister.

Go to <https://aspredicted.org/> and create a new AsPredicted pre-registration by clicking the ‘create’ button. Fill in your name, e-mail, and institution.

## Creating a New Pre-Registration

I am just trying things out. (Check the box and the submission will self destruct within 24 hours) 

**Participating Authors (Up to 5)**  
More than 5 authors?

Order	First	Last	email	Affiliation
1	Daniel	Lakens	D.Lakens@tue.nl	Eindhoven University of Technology

Scroll down, and answer questions 1 to 11. At 2) paste your answer to Q1, at 3) paste your answer to Q2, at 4) explain how many groups you will compare (e.g., Edward Norton vs. Brad Pitt), at 5) and 6) enter the answer at Q4, and at 7) enter the answer from Q3. Answer the remaining questions. Please indicate at 10) that you are using AsPredicted for a ‘Class project or assignment’ if you want to complete an actual preregistration.

Preview your pre-registration, and submit the preregistration. If you have added co-authors, they need to approve the submission. AsPredicted gives you the option to make an anonymous PDF file for peer review purposes, or you can just make the preregistration public.

To share this pre-registration you need to make a .pdf. If you are submitting for peer-review you probably want to first make an anonymous .pdf, and once the paper is accepted you make a public .pdf.  
If you click below you will see more information about this process, and can still change your mind.

[Make Anonymous PDF](#) [Make Public](#)

To share your preregistration during the review process, AsPredicted allows you to download an anonymous PDF. You can see the preregistration corresponding to the research question above at: <https://aspredicted.org/nx35m.pdf>

# 14 Computational Reproducibility

Technology has greatly improved how scientists work. The internet has made it easy to share information – including data, materials, and code – and new software and online platforms have emerged to facilitate the workflow of scientists (Spellman, 2015). One important goal of a scientific workflow is to make sure that the final work you publish is computationally reproducible. **Computational reproducibility** means that when you use the **same data** as in the published article, you can reproduce the **same results**. In other words, if the authors of the published article send you their data and code, you should be able to get the exact same numbers as they report in their article. Current research on the computational reproducibility of scientific articles suggests it is often not possible to run the original code on the data to reproduce results (Crüwell et al., 2023; Hardwicke et al., 2018; Obels et al., 2020; Stodden et al., 2018). Sometimes the code will simply not run on the data, or not all analyses are part of the code.

However, computational reproducibility is important, both for other scholars to be able to verify your results, and to build on your results. We could consider computational reproducibility a minimum standard of your own workflow. However, meeting this standard requires training. When I was a PhD, we often had a problem known as ‘data rot’. When I submitted an article for publication, and received the reviews after several months, I could not always easily reproduce my own analyses. For example, I might not have stored how I dealt with outliers, and could not exactly reproduce the original results. Sometimes, ‘data rot’ had eaten away at either my data or my analysis code, and it no longer worked.

Obviously, there is no such thing as ‘data rot’. The problem was I did not use a reproducible workflow. In this chapter, we will learn what a computationally reproducible workflow looks like, and how you can share computationally reproducible results with your published paper. The goal of applying a computationally reproducible workflow to your projects is to allow someone else (or yourself, one year from now) to take your data, run your code, and get exactly the same results as you reported in your work.

Although there are multiple ways to achieve a fully reproducible workflow, in this chapter I aim to introduce you to what I believe might be one emerging standard in psychological research. Through an example, you will learn to work with a version control system (such as GitHub, which integrates nicely with the Open Science Framework) as you are programming in R, which stores previous versions of files. You will then learn how to write a completely reproducible data analysis script(including figures), that you can save as an HTML file or a PDF file, using RMarkdown. Finally, we will take a look at Code Ocean, a novel online platform that allows

you to share computationally reproducible code online, making it extremely easy for others to run (small variations of) your code. While you will not learn how to become an experienced programmer by the end of this chapter, you will see what a fully reproducible workflow would look like, and get some initial experience with tools you will most likely want to explore more in the future.

Getting software and code to work on your system might be a challenge, and regrettably, I can't offer ICT support. Differences between Windows, Linux, and Apple operating systems means you might need to search the internet for solutions to problems you run into – this is very normal, and even experienced programmers do this all the time. If you get stuck, you can check what you did against what the code should look like by visiting the public versions of part of this example:

GitHub repository: [https://github.com/Lakens/reproducibility\\_assignment](https://github.com/Lakens/reproducibility_assignment)

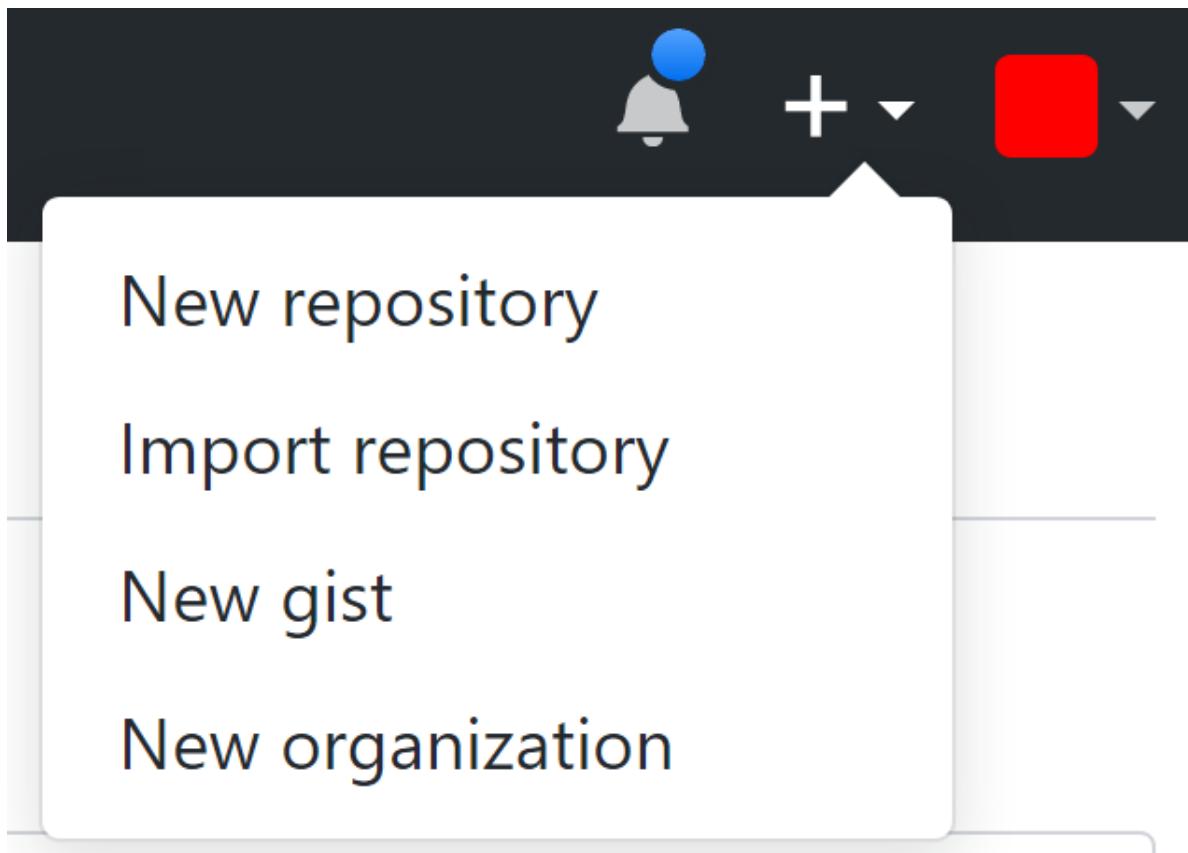
OSF project: <https://osf.io/jky8s/>

Code Ocean container: <https://codeocean.com/capsule/2529779/tree/v1>

## 14.1 Step 1: Setting up a GitHub repository

In this assignment we will use GitHub, but an open source alternative is to use [GitLab](#). If you haven't created a GitHub account before, do so now. Go to <https://github.com/> and create an account. Git is a version control system for tracking changes in computer files and coordinating work on those files among multiple people. Version control allows you to track changes to files and revert back to previous versions if needed. GitHub and GitLab are web-based hosting services that make it easier to use version control with Git. We will be using GitHub because it is what I am most familiar with, and it integrates with slightly more tools, but feel free to use GitLab instead.

If you have an account, you can create a new repository. A **repository** is a collection of folders and files that make up your project. In the top-right of the GitHub page, click the + symbol, and select 'New repository' from the dropdown menu.

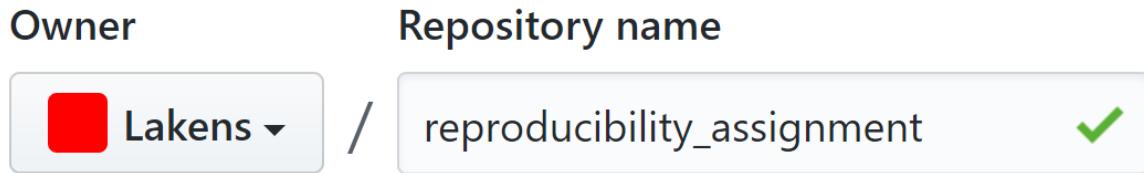


The first thing to do is name your repository. When it comes to naming folders and files, it is important to follow **best practices for file naming**:

- Keep names short, but clear. `data_analysis_project` is easier to understand for others than `dat_an_prjct`
- Do not use spaces. Options include:
  - Underscore: `this_is_a_file.R` (this is my personal favorite)
  - Camelcase: `ThisIsAFile.R`
  - Dashes: `this-is-a-file.R`
  - No spaces: `thisisafile.R`
- If you want to number multiple sequential files, do not use `1_start`, `2_end`, but use leading zeroes whenever you might number more than 10 files, so for example `01`, `02`, etc., or `001`, `002`, etc.
- Do not use special characters such as `$#&*{}`: in file names.

- If you want to use date information, use the YYYYMMDD format.

Let's name our repository: reproducibility\_assignment



You can add a short description (e.g., ‘This is an assignment to practice an open and reproducible data analysis workflow’). If you are an academic or student, you can get an academic account, which gives some extra options, such as keeping repositories private: <https://education.github.com/pack>

Click the checkbox before ‘Initialize this repository with a README’. A readme file is a useful way to provide a more detailed description of your project, that will be visible when people visit your GitHub project page. It can also contain instructions on how to reproduce analyses, such as which files to run in which order, and any changes to files that need to be made as the files are run.

You are also asked whether you want to add a **license**. Adding a license is a way to easily communicate to others how they can use the data, code, and materials that you will share in your GitHub repository. Note that not making a choice about a license is also a choice: if you do not add a license your work is under exclusive copyright by default, which means others can't easily re-use it. You can [learn more about licenses](#), but for now, a simple choice is the MIT license, which puts only very limited restrictions on reuse, but more restrictive licenses also exist. You can select the choice of license (such as the MIT license) from the dropdown menu. It lets people do anything they want with your code as long as they provide attribution back to you and don't hold you liable. There are also [creative commons licenses](#) that you can use when you are sharing something else than software, such as research materials (for example, this educational material is shared under a [CC-BY-NC-SA 4.0](#) license).

We are now ready to create the repository. Click

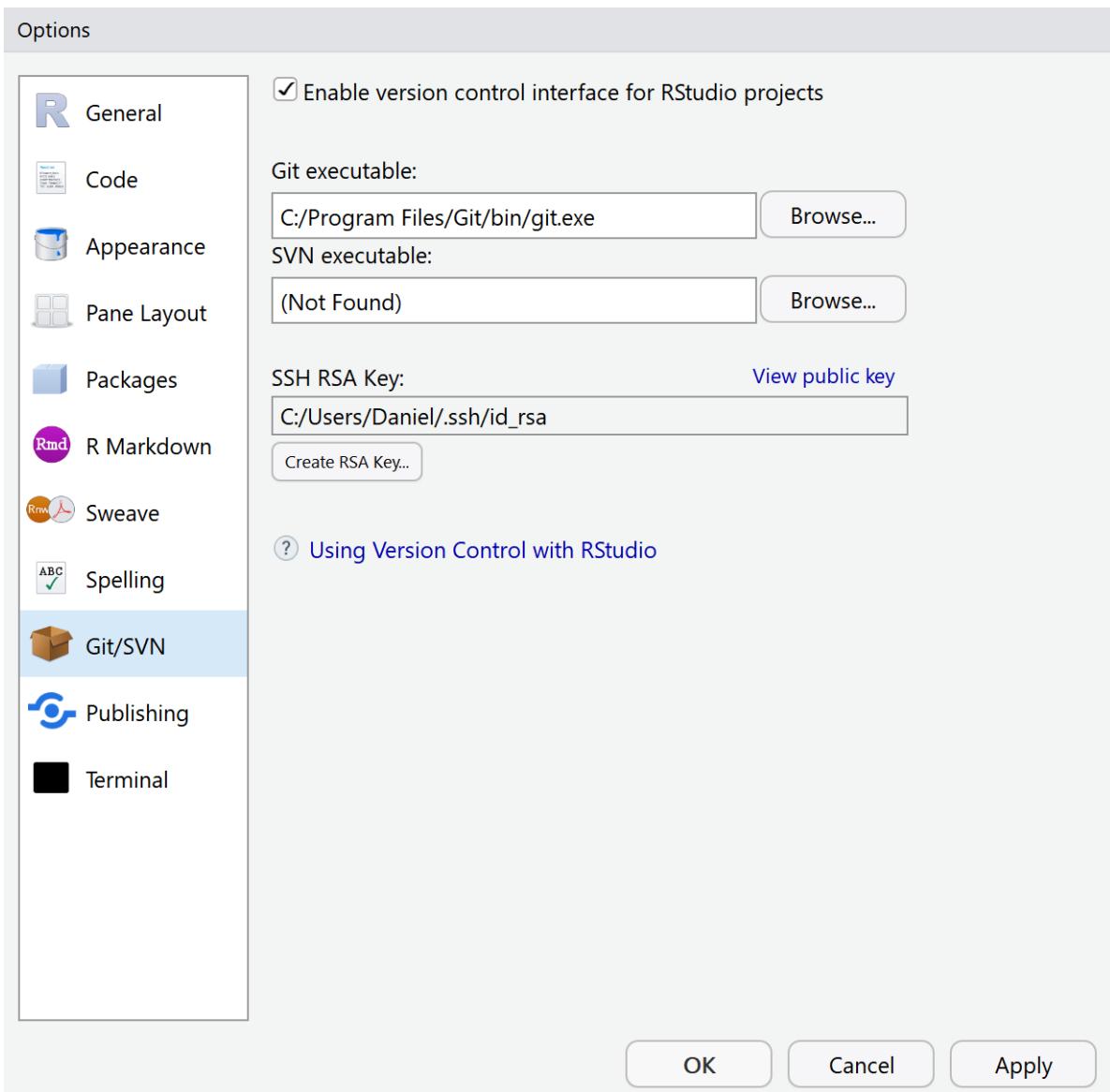
**Create repository**

It might feel unintuitive, but it is important to remember that you are not expected to directly interact with your new GitHub repository through the GitHub website. The repository page will give you information about the contents of the repository, and the history of the files in the repository, but it is not particularly easy to add files or download files directly through the website. The idea is that you use other software to interact with your GitHub repository.

## 14.2 Step 2: Cloning your GitHub repository into RStudio

R Studio can communicate with GitHub. To allow RStudio to work together with GitHub, you first need to set up the system. A detailed explanation for different operating systems is provided [here](#). First, download Git: <https://git-scm.com/downloads> for your operating system, and install it (you can accept all defaults during the installation process). If you haven't done so already, download and install R: <https://cran.r-project.org/>, and download and install the free version of R Studio (scroll down for the installers): <https://www.rstudio.com/products/rstudio/download/>.

In R Studio, go to Tools > Global Options, and select the Git/SVN menu option.



Check if the Git executable (“git.exe”) has been found automatically. If not, you will need to click the ‘Browse...’ button and find it manually. It will always be in the location where you installed Git.

Click the ‘Create RSA Key...’ button. A window will appear:

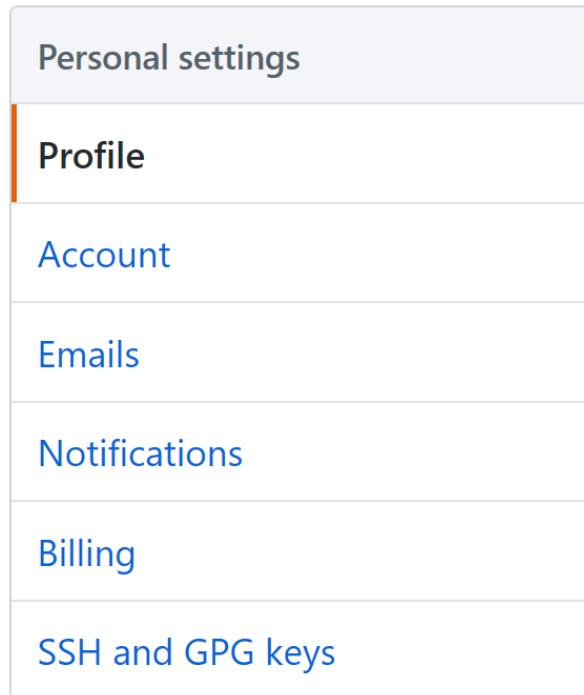
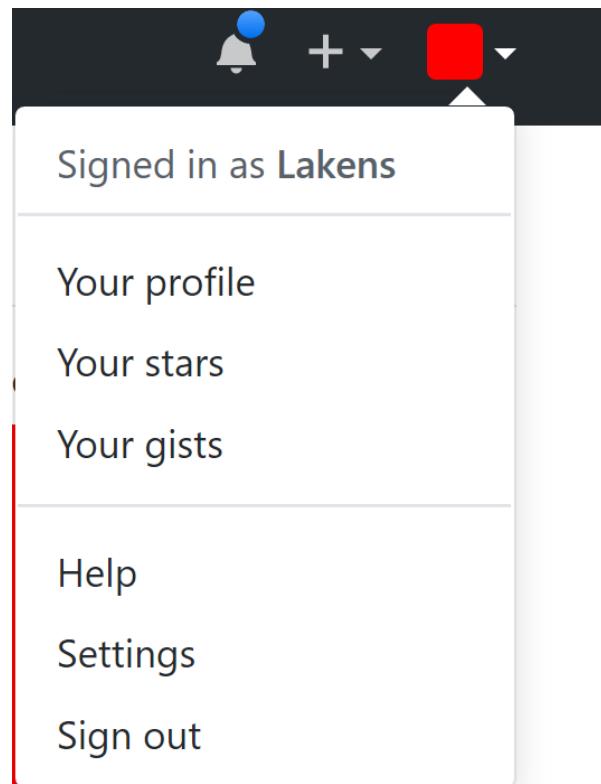
### Create RSA Key

```
Generating public/private rsa key pair.  
Your identification has been saved in C:/Users/Daniel/.ssh/id_rsa.  
Your public key has been saved in C:/Users/Daniel/.ssh/id_rsa.pub.  
The key fingerprint is:  
SHA256:7P/9jPIVOFgI2dxE8HqbVV[REDACTED] daniel@Lakens_Laptop  
The key's randomart image is:  
+---[RSA 2048]---+  
| .+=o.+ |  
| ..oo.o.*|  
| . + *+|  
| . =.o*+|  
| S o =+oB|  
| . o =o+|  
| . . o. o|  
| . E. ==|  
| ...o*B*|  
+---[SHA256]---+
```

[Close](#)

You can close the window. Still under the RStudio options, click the blue hyperlink ‘View public key’. A window will appear, telling you that you can use CTRL+C to copy the key. Do so.

Go to GitHub, and go to settings and then select the option SSH and GPG keys:

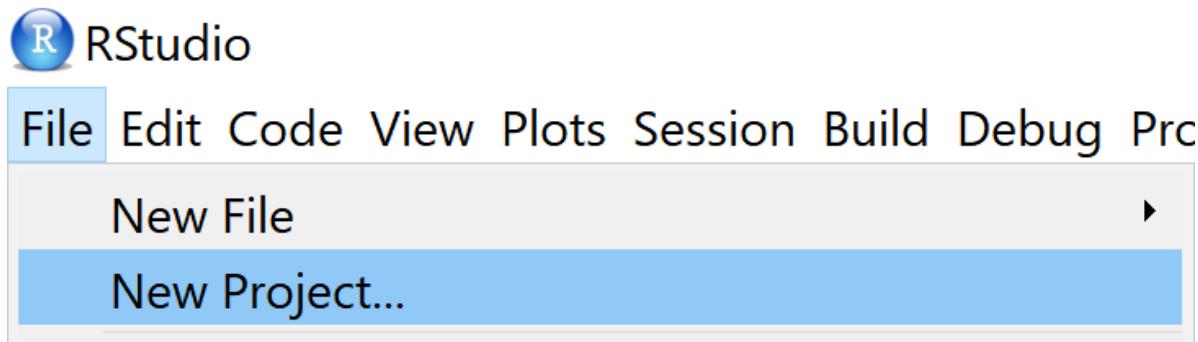


Click ‘New SSH key’

The screenshot shows the GitHub 'Personal settings' page. On the left, a sidebar lists 'Personal settings', 'Profile', 'Account', 'Emails', 'Notifications', 'Billing', and 'SSH and GPG keys'. The 'SSH and GPG keys' option is highlighted with an orange border. The main content area is titled 'SSH keys' and contains the message 'There are no SSH keys associated with your account.' with a link to 'generating SSH keys' and 'common SSH Problems'. A green button labeled 'New SSH key' is visible. Below it is a section titled 'GPG keys' with the message 'There are no GPG keys associated with your account.' and a link to 'generate a GPG key and add it to your account.', also featuring a green 'New GPG key' button.

Enter a name (e.g., RStudio) and paste the key in the correct window. Click ‘Add SSH Key’. This will allow you to send code from R Studio to your GitHub repositories without having to enter your GitHub login name and password every time. In other words, R Studio is now connected to your GitHub account and repository. You are now ready to create a **version controlled project** in R Studio.

**Restart RStudio.** In RStudio, go to File>New Project:



You get three choices. Choose the ‘Version Control’ option:

New Project

## Create Project



### New Directory

Start a project in a brand new working directory



### Existing Directory

Associate a project with an existing working directory



### Version Control

Checkout a project from a version control repository



Cancel

Choose the 'Git' option:

 Back

## Create Project from Version Control



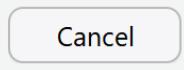
### Git

Clone a project from a Git repository 



### Subversion

Checkout a project from a Subversion repository 

 Cancel

We will be cloning the online GitHub repository we created. Cloning is a term used in Git that means creating a local copy of all files in your repository on your computer. You can copy-paste the URL from your GitHub repository (e.g., [https://github.com/Lakens/reproducibility\\_assignment](https://github.com/Lakens/reproducibility_assignment)). If you copy-paste this URL in the top field, it will automatically create a Project directory name that is similar to the name you gave your project on GitHub. You can select a folder on your computer by clicking the ‘Browse’ button to indicate where you want to save the local copy of your repository.

New Project

**Clone Git Repository**



Repository URL:

Project directory name:

Create project as subdirectory of:

Open in new session

Click ‘Create Project’. R will quickly download the files from your repository, and open the new project. You will see that the project creation was successful because the ‘Files’ tab in the RStudio interface shows we have downloaded some files from our GitHub repository (the README.md and LICENSE files). RStudio also created a .Rproj file and a .gitignore file. The **project file** is used to store information about the project, and that is required to use GitHub.

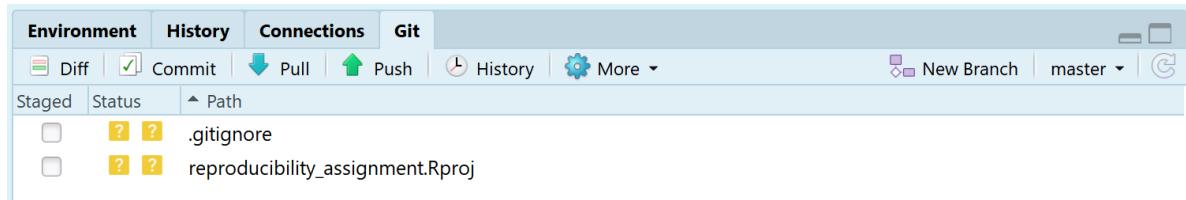
Files Plots Packages Help Viewer

New Folder Delete Rename More

C: > Users > Daniel > surfdrive > R > reproducibility\_assignment

	Name	Size	Modified
..			
<input type="checkbox"/>	.gitignore	44 B	Jun 17, 2018, 5:10 PM
<input type="checkbox"/>	LICENSE	1.1 KB	Jun 17, 2018, 5:10 PM
<input type="checkbox"/>	README.md	113 B	Jun 17, 2018, 5:10 PM
<input type="checkbox"/>	reproducibility_assignment.Rproj	218 B	Jun 17, 2018, 5:10 PM

We can also see this is a version control project in the top right of the interface, where there is now a ‘Git’ tab. If we click it, we see:



We see a range of buttons, such as the Diff, Commit, Pull, and Push buttons. These will be used to interact with GitHub. Many computer programmers interact with GitHub through the command line, such as:

```
$ git commit -m "This is a git commit message"
```

Learning to use git through the command line is not needed for most people who just want basic version control. Here, I will exclusively focus on version control and git through the menu options in RStudio. It is now time to create a file for which we want to control the versions we make of it.

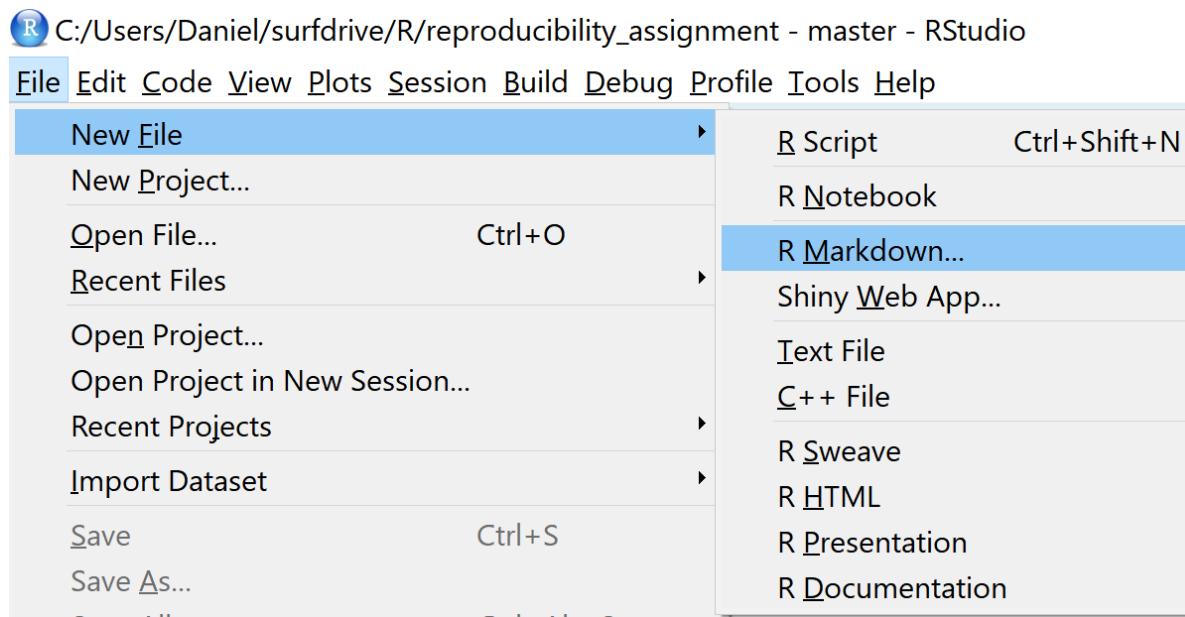
### 14.3 Step 3: Creating an R Markdown file

R Markdown files provide a way to save and execute code, while at the same time allowing you to create reports of your data analysis (and even full scientific articles that you can submit for publication!). A complete introduction to R Markdown is [available here](#). The main strength of R Markdown documents is that they allow you to create a fully reproducible document. This means that you do not just have some analysis code in an R script, but a manuscript that combines text and code and that you can **compile** to create a PDF or html version of the manuscript. HTML or PDF files have the advantage that people can read them with regular software. The R Markdown file contains code that performs the analyses **each time the document is compiled**. Instead of copy-pasting values from your analysis software into a word document, you combine code and text in the RMarkdown file to create a manuscript where every number or figure can be traced back to the exact code that generated it. That has the advantage that everyone can use your RMarkdown file and generate the same document (e.g., your manuscript) as you.

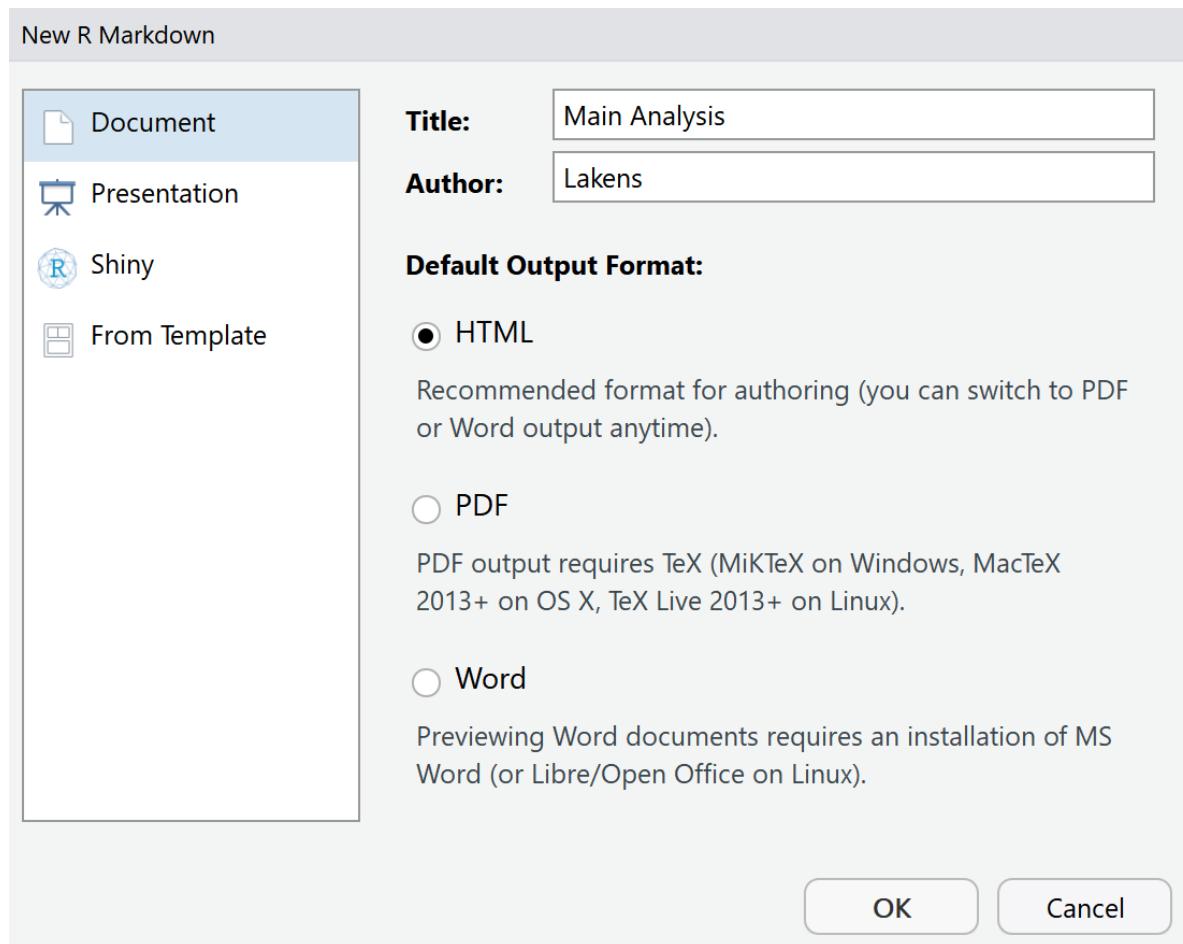
You can still make errors in the analysis if you use R Markdown files. The important difference is that you will be making programming errors that are stored in the R Markdown document. Compared to a typo when copying numbers from your analysis to a word document, errors in your analysis in your RMarkdown file will lead to the same document. Because the document is reproducible, all errors are reproducible as well. It is impossible to prevent all errors, but it

is possible to make them reproducible. This will make it easier to identify and correct errors. I understand you might worry about others seeing your errors if you allow them to see exactly what you have done. But **we all make mistakes**, and it is important for science to be able to identify and correct these mistakes. An important aspect of moving to a more reproducible workflow, and sharing all files underlying your manuscript publicly, is that we will have to learn to accept that **we all make errors**, and appreciate people who correct them (Bishop, 2018).

Let's start by creating a new R Markdown document in R Studio by clicking New File > R Markdown...



This gives you a new window where you can specify the title of your RMarkdown document and an author name. Enter the title ‘Main Analysis’, and feel free to change the Author subfield to anything you prefer. RMarkdown files can be compiled (also referred to as ‘knitted’) into an HTML file, a PDF document, or a word document. To generate PDF files you need to install MiKTeX which we won’t do for this example ([a good tutorial how to install MiKTeX is available here](#)). So leave the default output format to HTML and click OK.



Let's start by saving the new file: Click the save button, and save the file under the name 'main\_analysis.Rmd'. Because we are working in an R Studio project, the file will automatically be saved in the same folder as all other files in this project. If you look at the files tab in the bottom right pane, you will see the new file appear. Now let's take a look at the R Markdown file.

The R Markdown file by default includes several sections to get you started. First, there is a header section. In the header section, there is code that determines how the final document is rendered. This section is sensitive, in the sense that it needs to be programmed exactly right – including spaces and tabs – so it is not recommended to change it too much without looking up detailed documentation on how to change this section. If you want the technical details: An R Markdown file is fed to knitr software, which creates a normal markdown file, which then uses pandoc software to generate the specific document you requested. All of this happens automatically.

C:/Users/Daniel/surfdrive/R/reproducibility\_assignment - master - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

+ - Go to file/function Addins

```

1 --- } Header
2 title: "Main Analysis"
3 author: "Lakens"
4 date: "17 juni 2018"
5 output: html_document
6 ---
7
8 ``{r setup, include=FALSE} } Setup
9 knitr::opts_chunk$set(echo = TRUE)
10
11
12 ## R Markdown
13
14 This is an R Markdown document. Markdown is a simple formatting syntax for authoring
15 HTML, PDF, and MS Word documents. For more details on using R Markdown see
16 <http://rmarkdown.rstudio.com>.
17
18 ``{r cars} } R code
19 summary(cars)
20
21
22 ## Including Plots
23
24 You can also embed plots, for example:
25
26 ``{r pressure, echo=FALSE}
27 plot(pressure)
28
29
30 Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing
31 of the R code that generated the plot.

```

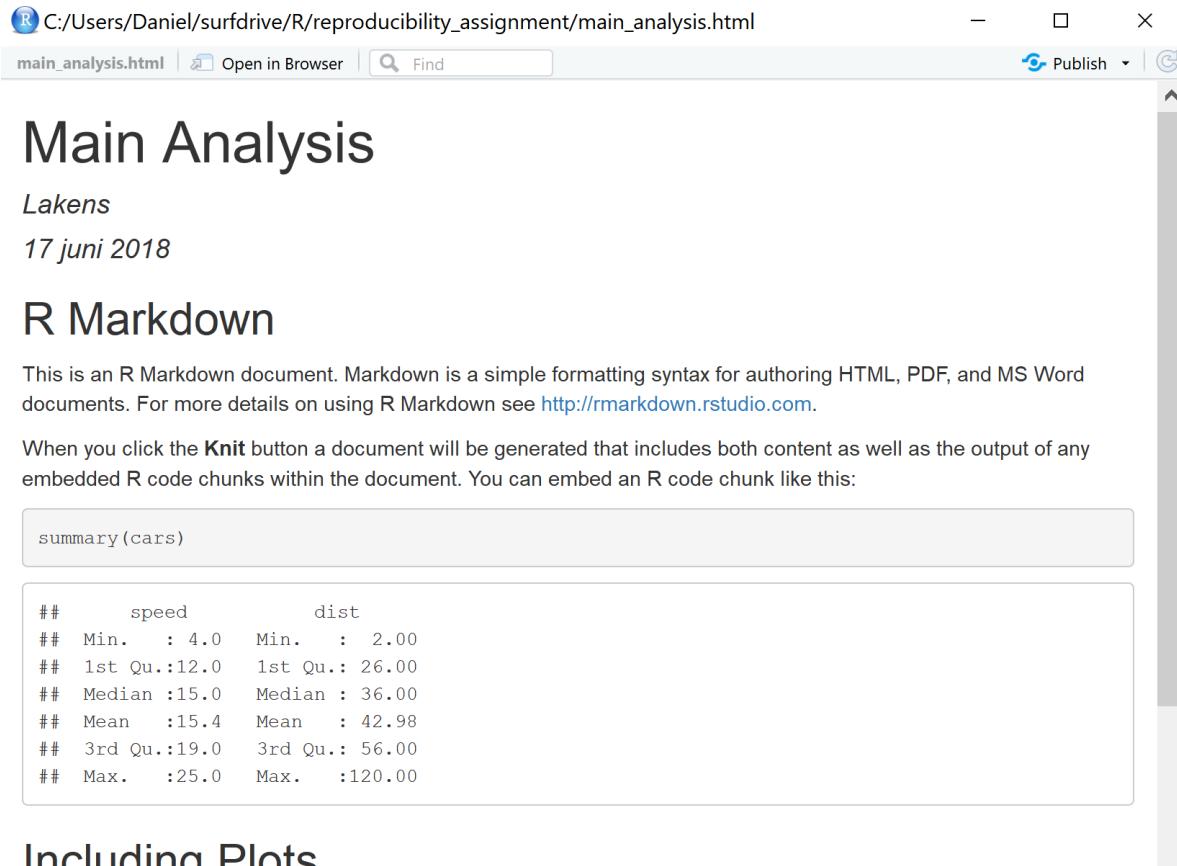
2:1 Main Analysis R Markdown

The header is followed by a **set-up** section where you can define general options for the entire R Markdown file. Then, we see the two main sections: **Markdown code**, which is a markup language in plain text formatting syntax that can be easily converted into HTML or other formats. Then, we see **R code** that is used to analyze data or create figures. To see the final result of this code, hit the



Knit button in the toolbar at the top of the pane.

Either a new window will appear that allows you to view the HTML file that was created, or your document will appear in the ‘viewer’ tab in RStudio. You see the formatted HTML document that combined both text and the output of R code.



The screenshot shows a browser window displaying an R Markdown document titled "main\_analysis.html". The title "Main Analysis" is at the top. Below it are author details: "Lakens" and "17 juni 2018". A section titled "R Markdown" follows. A note explains that R Markdown is a simple syntax for HTML, PDF, and MS Word documents, with a link to <http://rmarkdown.rstudio.com>. Another note says that clicking "Knit" generates a document including code output. A code chunk shows the summary of the "cars" dataset:

```
summary(cars)
```

```
##      speed         dist
## Min.   : 4.0   Min.   : 2.00
## 1st Qu.:12.0   1st Qu.: 26.00
## Median :15.0   Median : 36.00
## Mean    :15.4   Mean    : 42.98
## 3rd Qu.:19.0   3rd Qu.: 56.00
## Max.   :25.0   Max.   :120.00
```

## Including Plots

You can also embed plots, for example:



Close the window – we are now ready to analyze our data.

## 14.4 Step 4: Reproducible Data Analysis in R Studio

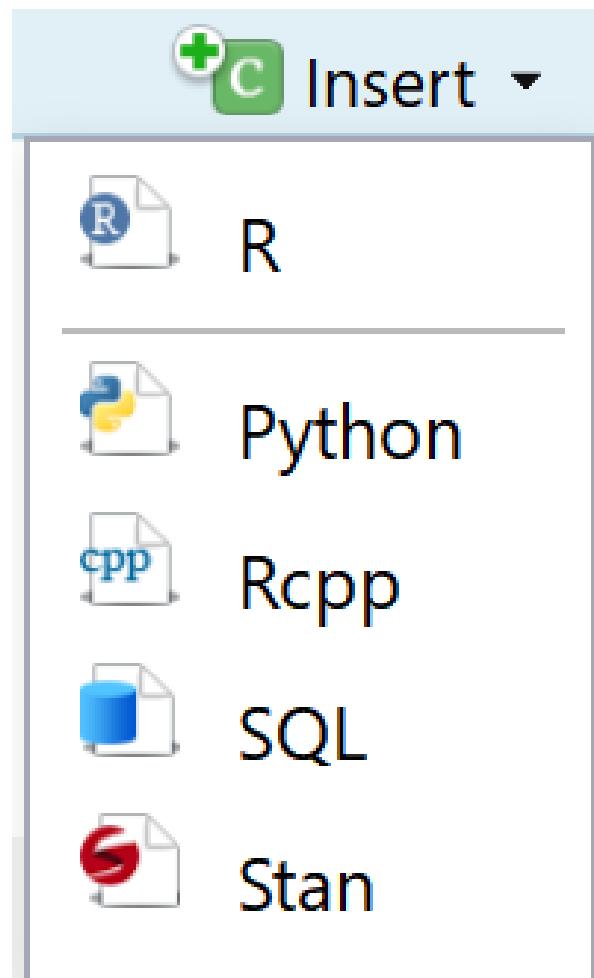
Delete all text from `## R Markdown` on down – only keep the header and set-up sections of the default document.

First, we need to analyze some data. We will download this data directly from an existing GitHub repository I created. Students in an introduction to psychology course performed a

simple Stroop experiment. During the Stroop experiment participants named the colors in a congruent trial (e.g., the word ‘red’ written in a red font) and incongruent trial (e.g., the word ‘red’ written in a green font). The time they took to name all words was recorded in seconds (e.g., 21.3 seconds) for both the congruent and incongruent trial. There are four columns in the dataset:

- Participant Number
- Response Time for Congruent Stimuli
- Response Time for Incongruent Stimuli
- Year of Data Collection

Click the button ‘+C Insert’ to insert code – a dropdown menu will be visible. Select R.



In the R Markdown file, you'll see a new section of R code that starts with three backticks followed by {r} and ends with three backticks. You can also just create these sections by manually typing in these two lines.

Copy-paste the code below – make sure to get all the text – and paste it between the start line and the end line of the R code chunk.

```
stroop_data <- read.table("https://raw.githubusercontent.com/Lakens/Stroop/master/stroop.txt"
sep = "\t", header = TRUE)

write.table(stroop_data, file = "stroop.csv", quote = F, row.names = F)
```

After copy-pasting the text, the code section should look like this:

```
```{r}
stroop_data <- read.table("https://raw.githubusercontent.com/Lakens/Stroop/master/stroop.txt",
sep = "\t", header = TRUE)

write.table(stroop_data, file = "stroop.csv", quote=F, row.names=F)
```
```

This code creates a data.frame called ‘stroop\_data’ that contains data, and then saves this data in a .csv file called ‘stroop.csv’. Click the Knit button to look at the document:



You should see something like:

# Main Analysis

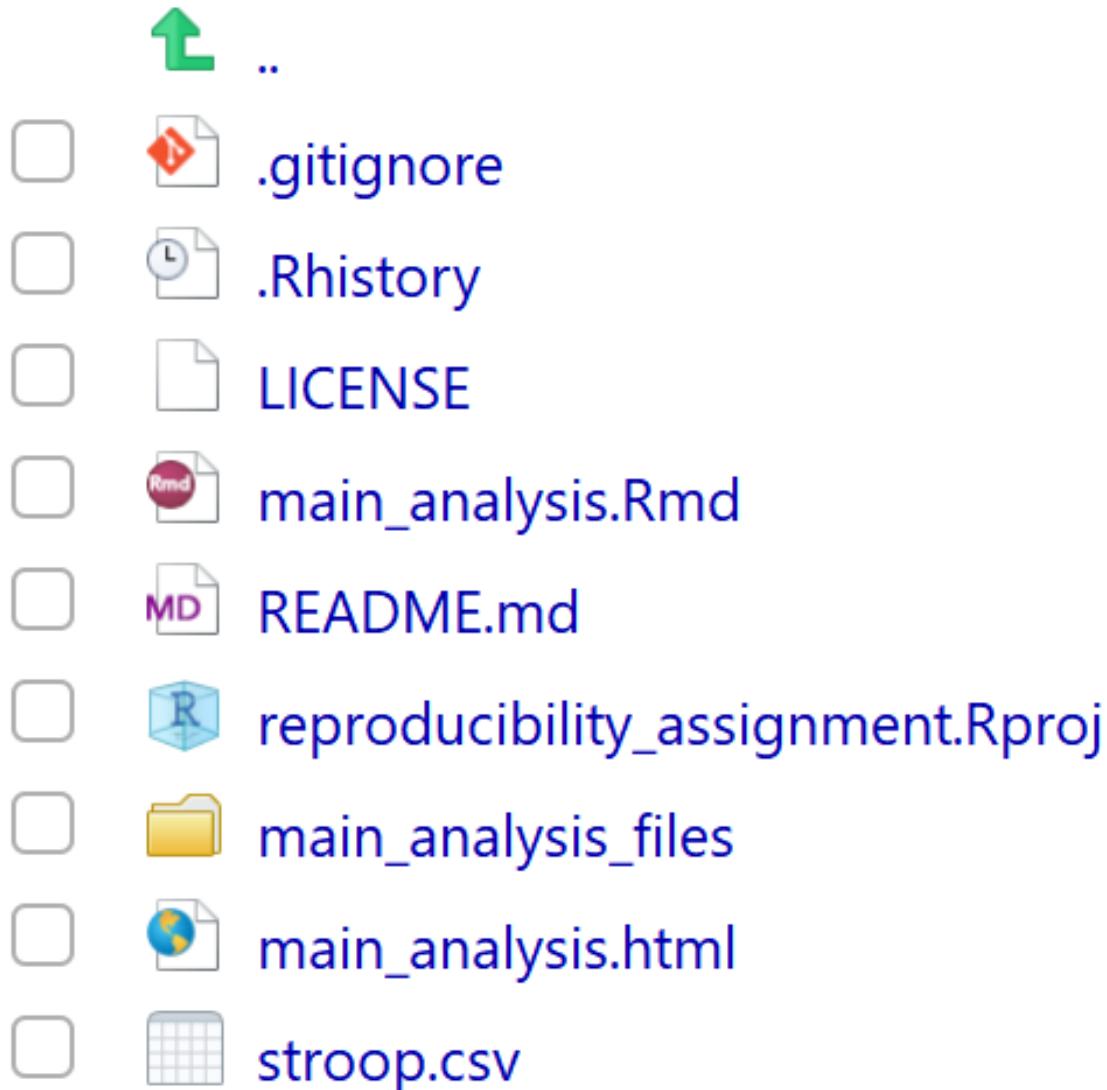
*Lakens*

*17 juni 2018*

```
stroop_data <- read.table("https://raw.githubusercontent.com/Lakens/Stroop/master/stroop.txt",
sep = "\t", header = TRUE)

write.table(stroop_data, file = "stroop.csv", quote=F, row.names=F)
```

This might not look very impressive – but the real action is in the file pane in the bottom right part of the screen. Close the window showing the HTML output and look at the file pane. You should now see several files:



One file is `stroop.csv` – this is our data file of the Stroop data that we downloaded from the internet and saved to our project folder, using R code.

There is really no need to keep downloading the file from the internet when we can also just load it from the local folder. So let's change the code. We won't completely delete this code –

we will just **comment it out** by placing a # in front of it. This way, we can still remember where we downloaded the file from, but we won't use the code.

Because it is always important to **provide comments in the code you write**, add this explanation above the line where we downloaded the code:

```
#run only once to download the data
```

Then, select the lines of code in the chunk, and press (on Windows) CTRL+SHIFT+C (or click 'Code' in the toolbar and then 'comment/uncomment lines'). This should add # in front of all lines, making it comments instead of code that is executed every time. You should end up with:

```
```{r}
# #run only once to download the data
# stroop_data <- read.table("https://raw.githubusercontent.com/Lakens/Stroop/master/stroop.txt", sep = "\t", header = TRUE)
#
# write.table(stroop_data, file = "stroop.csv", quote=F, row.names=F)
```

```

Now we need to add a line of code that we will run, and with which we will load the stroop.csv dataset from the local folder. Underneath the last commented out line of code, but within the R code block, add:

```
stroop_data <- read.csv("stroop.csv", sep = " ", header = TRUE)
```

```
```{r}
# #run only once to download the data
# stroop_data <- read.table("https://raw.githubusercontent.com/Lakens/Stroop/master/stroop.txt", sep = "\t", header = TRUE)
#
# write.table(stroop_data, file = "stroop.csv", quote=F, row.names=F)
stroop_data <- read.csv("stroop.csv", sep = " ", header = TRUE)
```

```

Click save, or press CTRL+S, to save the file. Knit the file. We see:

# Main Analysis

*Lakens*

17 juni 2018

```
# #run only once to download the data
# stroop_data <- read.table("https://raw.githubusercontent.com/Lakens/Stroop/master/stroop.txt",
#   sep = "\t", header = TRUE)
#
# write.table(stroop_data, file = "stroop.csv", quote=F, row.names=F)

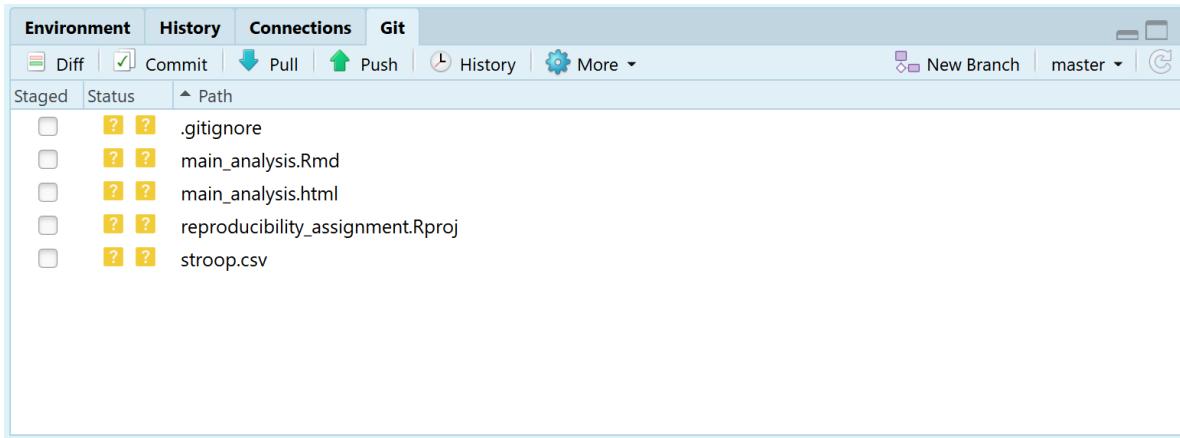
stroop_data <- read.csv("stroop.csv", sep = " ", header = TRUE)
```

Close the HTML file. We've done quite a lot of work. It would be a shame if this work was lost. So this seems to be the perfect time to save a version of our R Markdown file, not just locally, but also on GitHub.

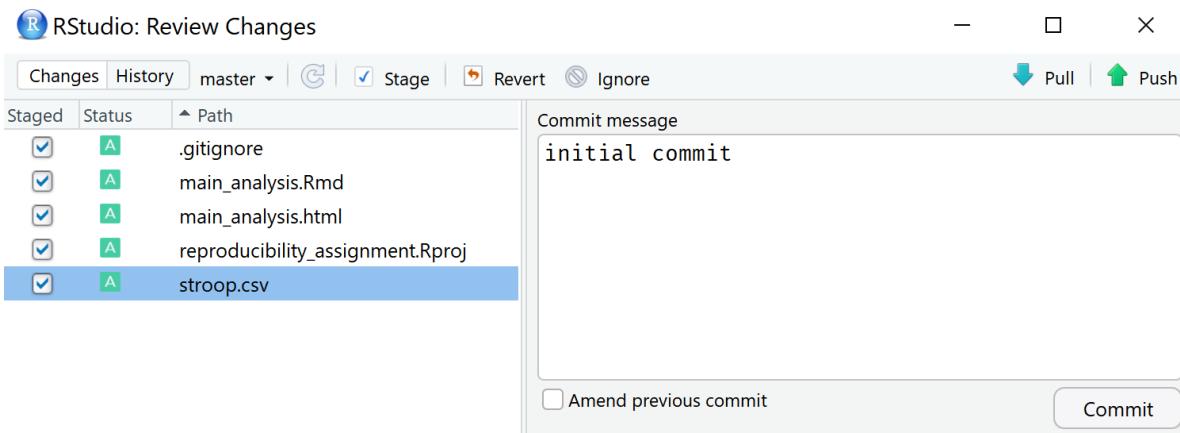
## 14.5 Step 5: Committing and Pushing to GitHub

It is time to store our changes in the cloud, on GitHub. This process takes two steps. First, we record the changes to the repository (aka the code and files we created), which is called ‘**commit**’. This does not require an internet connection, because we are just recording the changes locally. However, then we want to make sure these recorded changes are also stored on GitHub, which requires you to **push** the files to GitHub.

If we look at the Git tab in the top right pane in R Studio, we see the Commit button, the Push button, and we see a bunch of files. The status of these files is indicated by two question marks in yellow. These question marks indicate these files are not yet tracked by GitHub. Let's change this.



Click the commit button. A menu opens. You can choose to ‘stage’ the changes that have been made. Staging basically mean selecting which of the files you want to record, or **commit**. You can do this in several ways, such as double clicking each file, or selecting all files and clicking ‘Enter’. When staging all files, the yellow question marks change to a green ‘A’ symbol. Every commit should be accompanied by a **commit message** where you describe which changes you have made – after all, we are recording our changes. You can type in anything you want – it is common to write something like ‘initial commit’ the first time you commit changes. The menu should look like the screenshot below:



Now we are ready to **commit** these changes. Click the ‘Commit’ button. A new window opens that shows all changes that have been committed. We see that 5 files have changed. You can close this window and close the previous menu.

Git Commit

```
>>> C:/Program Files/Git/bin/git.exe commit -F C:/Users/Daniel/AppData/Local/Temp/Rtmp00000000/000000000000000000000000.Rmd
[master 4dbbace] initial commit
 5 files changed, 335 insertions(+)
 create mode 100644 .gitignore
 create mode 100644 main_analysis.Rmd
 create mode 100644 main_analysis.html
 create mode 100644 reproducibility_assignment.Rproj
 create mode 100644 stroop.csv
```

R Studio now reminds you that there is a difference between the local copy of your repository, and the remote version of the repository on GitHub. In the Git tab you see a reminder: “Your branch is ahead of ‘origin/master’ by 1 commit.”.



This means the files we have updated and recorded on our computer with a commit are not yet synchronized with the **remote repository** on GitHub. We can solve that by ‘pushing’ (aka synchronizing) the changes to the remote repository. Simply click the **push** button:



Another pop-up window appears:

Git Push

```
>>> C:/Program Files/Git/bin/git.exe push origin refs/heads/master
To https://github.com/Lakens/reproducibility_assignment
 9a2eaad..4dbbace  master -> master
```

Close

This window informs us there were no errors, and we successfully pushed the changes to the remote version of the repository. You can close this window.

You can check that you successfully pushed all files to GitHub by visiting the GitHub page for your repository in the browser. You should see something like:

The screenshot shows a GitHub repository page. At the top, it displays the repository name "Lakens / reproducibility\_assignment". To the right are buttons for "Unwatch" (with a count of 1), "Star" (0), "Fork" (0), and "Edit". Below the repository name are navigation links: "Code" (selected), "Issues 0", "Pull requests 0", "Projects 0", "Wiki", "Insights", and "Settings". A summary bar below these links shows "2 commits", "1 branch", "0 releases", "1 contributor", and "MIT". There is also a "New pull request" button. A "Branch: master" dropdown is visible. On the far right, there are buttons for "Create new file", "Upload files", "Find file", and "Clone or download" (which is highlighted in green). The main content area lists the repository's files: ".gitignore", "LICENSE", "README.md", "main\_analysis.Rmd", "main\_analysis.html", "reproducibility\_assignment.Rproj", and "stroop.csv". Each file entry includes its type (e.g., file icon), name, commit status (e.g., "initial commit"), and time of last commit (e.g., "6 minutes ago").

Congratulations on your first GitHub push! If you want to read a more extensive introduction to Git, see Vuorre & Curley (2018).

## 14.6 Step 6: Reproducible Data Analysis

So far, we have only read in data. The goal of an R Markdown file is to create a manuscript that contains a fully **reproducible data analysis**. In this chapter, I cannot teach you how to analyze data in R (but I can highly recommend learning it – there are plenty of excellent online resources). Instead of programming from scratch, visit [this raw text version of the R Markdown file](#) that will analyze the Stroop data. In the website, select all text (CTRL+A), copy it (CTRL+C). Then go to your main\_analysis.Rmd file in R Studio. Select all text (CTRL+A) and press delete. That's right – delete everything. You don't need to worry about losing anything – you have a **version controlled file** in your GitHub repository, which means you can always go back to a previous version! In the (now empty) main\_analysis.Rmd file, press CTRL+V and paste all text. The file should look like the first screenshot below.

This R Markdown file does a number of things, which we will explain in detail below. For example, it will automatically install libraries it needs, load the data, and create a report in HTML. You can press the Knit button, and the HTML document should load. You should see output as in the second screenshot below.

```

1 ---  

2 title: "Main Analysis"  

3 author: "Lakens"  

4 date: "20 juli 2018"  

5 output: html_document  

6 ---  

7  

8 ```{r global_options, echo=FALSE, warning=FALSE, message=FALSE,  

  include=FALSE}  

9 #This code will check if ggplot2 and reshape2 are installed.  

10 #If not, this code will install these packages.  

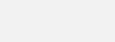
11 if(!require(ggplot2)){install.packages('ggplot2')}
```



```

12 library(ggplot2)  

13 if(!require(reshape2)){install.packages('reshape2')}
```



```

14 library(reshape2)  

15 knitr::opts_chunk$set(echo=FALSE, warning=FALSE, message=FALSE, include=TRUE)  

16 ...  

17  

18 # Introduction  

19  

20 Here, we analyze a simple dataset of a Stroop experiment. Students in an  

  introduction to psychology course completed an online Stroop task  

  (http://faculty.washington.edu/chudler/java/ready.html) and named the colors  

  in congruent trials (e.g., the word 'red' written in a red font) and in  

  incongruent trials (e.g., the word 'red' written in a green font). The time  

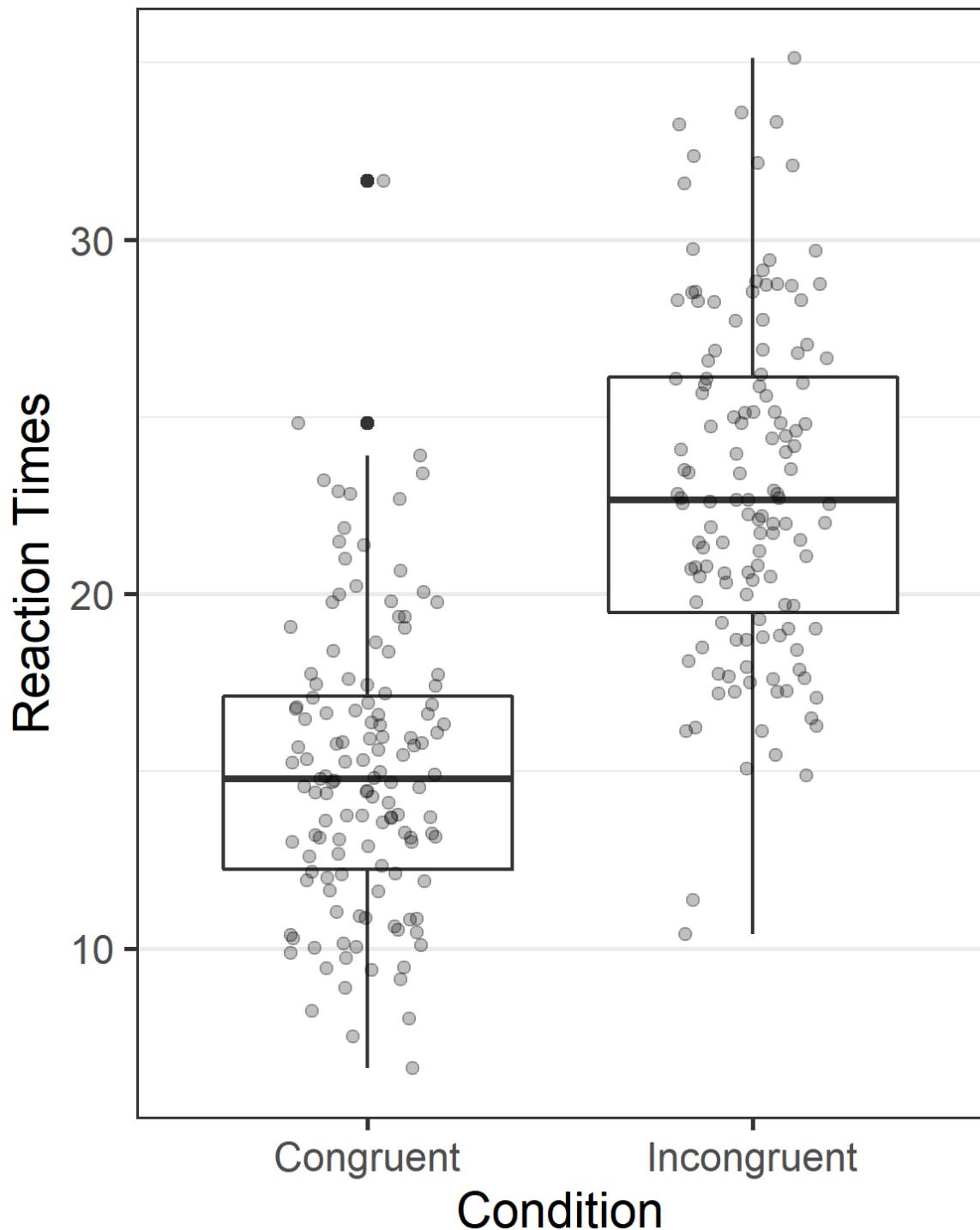
  they took to name all words was self-reported in seconds (e.g., 21.3 seconds)  

  for both the congruent and incongruent blocks. In this analysis, we are  

  interested in examining whether there is a Stroop effect.
```

## Results

The mean reaction time (in seconds) of participants in the Congruent condition ( $M = 15.1$ ,  $SD = 4.1$ ) was lower than the mean of participants in the Incongruent condition ( $M = 23$ ,  $SD = 4.78$ ,  $r = 0.37$ ). A dependent  $t$ -test indicated that based on our preregistered alpha level of 0.01 we could reject the null-hypothesis,  $t(130) = 18.04$ ,  $p < 0.001$ . As we can expect from the Stroop effect, the standardized effect size is very large, Hedges'  $g_{av} = 1.76$ . The congruency effect is very clear when we plot the data from the two groups.



It is important to note that none of the numbers that are in this text are static, or copy-pasted.

They are all calculated at the moment that the document is created, directly from the raw data. The same is true for the figures, which are created from the raw data the moment the manuscript is compiled. If you have access to the .Rmd (RMarkdown) file, you can perfectly **reproduce** the reported data analysis.

Since we have made substantial changes, this is the perfect moment to **commit** and **push** the changes to GitHub! Go to the Git tab in the top right pane. Click ‘Commit’. The window below will open. If the main\_analysis.Rmd file is selected, you will see red and green chunks of text. These tell you what was old (red) and what is new (green).

RStudio: Review Changes

Changes History master Stage Revert Ignore

Pull Push

Staged Status Path

M main\_analysis.Rmd

M main\_analysis.html

Commit message

Amend previous commit

Commit

Stage All Discard All

Stage chunk Discard chunk

```

@@ -1,22 +1,164 @@
1 1 ---
2 2 title: "Main Analysis"
3 3 author: "Lakens"
4 4 date: "17 juni 2018"
5 5 output: html_document
6 6 ---
7 7
8 8 ````{r setup, include=FALSE}
9 9 knitr::opts_chunk$set(echo = TRUE)
10 10 ````{r global_options, echo=FALSE, warning=FALSE, message=FALSE, include=FALSE}
11 11 knitr::opts_chunk$set(echo=FALSE, warning=FALSE, message=FALSE, include=TRUE)
12 12 #Introduction
13 13
14 14 Here, we analyze a simple dataset of a Stroop experiment. Students in an
   introduction to psychology course completed an online Stroop task
   (http://faculty.washington.edu/chudler/java/ready.html) and named the colors :
   congruent trials (e.g., the word 'red' written in a red font) and in
   incongruent trials (e.g., the word 'red' written in a green font). The time
   they took to name all words was self-reported in seconds (e.g., 21.3 seconds)
   for both the congruent and incongruent blocks. In this analysis, we are
   interested in examining whether there is a Stroop effect.
15
16 15 ````{r}
17 16 # #run only once to download the data
18 17 # stroop_data <-
19 18 read.table("https://raw.githubusercontent.com/Lakens/Stroop/master/stroop.txt"
20 20 sep = "\t", header = TRUE)

```

Select all files that have changed, and ‘stage’ them (for example by pressing enter). The checkboxes in front of the files, under the ‘Staged’ column, should be checked.

| Staged | Status | ▲ Path             |
|--------|--------|--------------------|
|        |        | main_analysis.Rmd  |
|        |        | main_analysis.html |

Type in a commit message, such as ‘update mean analysis’ in the ‘commit message’ field. Press the ‘Commit’ button. Close the window that pops up to inform you about the result of the commit. Then click ‘push’. Close the window that informs you about the push command, and close the commit window. You can always visit the GitHub repository online and look at the full history of your document to see all changes that have been made.

Let’s take a look at some sections of our new R Markdown document. First the header:

```
```{r global_options, echo=FALSE, warning=FALSE, message=FALSE,
include=FALSE}
knitr::opts_chunk$set(echo=FALSE, warning=FALSE, message=FALSE,
include=TRUE)
```
```

This sets general (global) options for the code chunks in the R Markdown file. The echo, warning, and message = FALSE hide the code chunks, warning messages, and other messages, where the ‘include=true’ will make all figures appear in the text. You can set some of these variables to TRUE, and hit Knit to see what they change. Sometimes you might want to share the HTML file with all code visible, for example when sharing with collaborators.

If you scroll down, you can see the introduction text, the code that generates the first figure, and the code that performs the analyses. These variables are used in the Results section. Let’s look at this section:

## #Results

```
The mean reaction time (in seconds) of participants in the Congruent condition (*M* = `r round(mean(stroop_data$Congruent), digits = 2)` , *SD* = `r round(sd(stroop_data$Congruent), digits = 2)` ) was lower than the mean of participants in the Incongruent condition (*M* = `r round(mean(stroop_data$Incongruent), digits = 2)` , *SD* = `r round(sd(stroop_data$Incongruent), digits = 2)` , *r* = `r round(cor(stroop_data$Congruent, stroop_data$Incongruent), digits = 2)` ). An independent *t*-test indicated we could reject the null-hypothesis, based on an alpha of 0.05, *t*(`r round(ttest_result$parameter, digits=2)` ) = `r round(ttest_result$statistic, digits=2)` , *p* `r ifelse(ttest_result$p.value > 0.001, " = ", " < ")` `r ifelse(ttest_result$p.value > 0.001, formatC(round(ttest_result$p.value, digits=3), digits=3, format="f"), "0.001")` . As we can expect from the Stroop effect, the standardized effect size is very large, Hedges' *g~av~* = `r round(d_unb, digits=2)` , 95% CI [`r round(ci_l_d_av, digits=2)` ;`r round(ci_u_d_av, digits=2)` ].
```

This section shows how you can **mix text and R code**. The start of this code is normal text. The `*M*` is still normal text (the `*` and `*` make sure the `M` is italicized, just as further down the `~av~` indicates these letters should be subscript), but then you see R code. In R Markdown you can embed R code within ‘`r`’. Any R code within the two backticks will be executed. In this case, the mean of the Congruent reaction times is calculated, and rounded to 2 digits. You can see this number in the text.

Learning to program takes time. You can see some things are quite tricky to program. For example, the code:

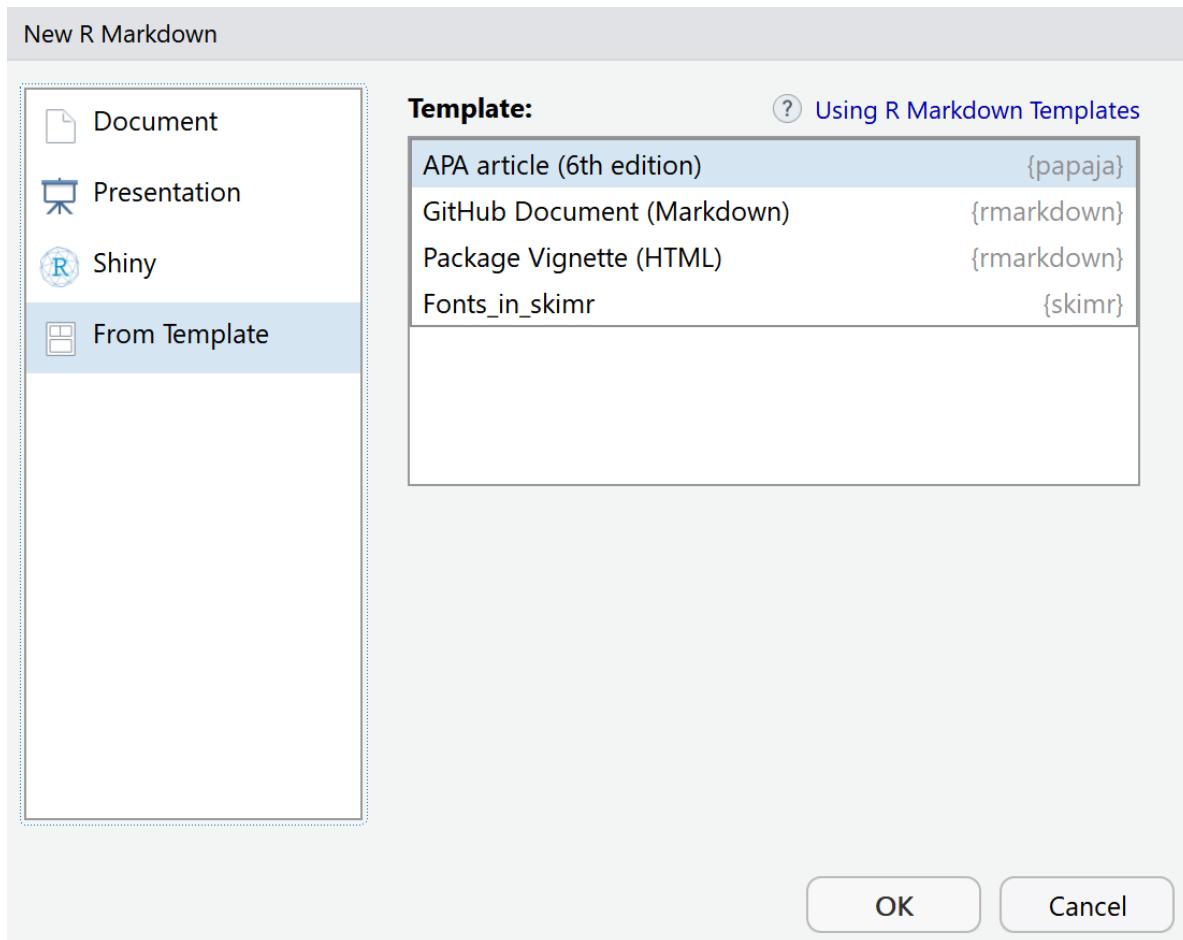
```
ifelse(ttest_result$p.value > 0.001, " = ", " < ")
ifelse(ttest_result$p.value > 0.001, formatC(round(ttest_result$p.value,
digits = 3), digits = 3, format = "f"), "0.001")
```

is a lot of code to make sure the exact  $p$ -value is reported, unless this  $p$ -value is smaller than 0.001, in which case ‘ $p < 0.001$ ’ is printed (the `papaja` package below makes reporting statistics a lot easier!). The first time you need to program something like this takes a lot of time – but remember you can re-use code in the future, and you can steal a lot of code from others! You can complete a full introduction to Markdown [here](#).

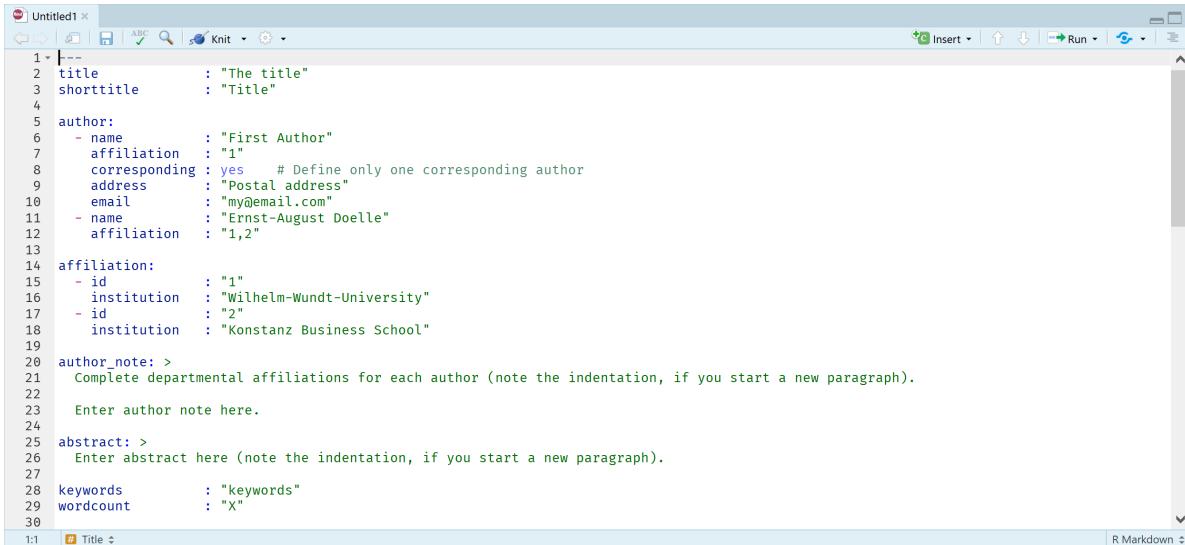
### 14.6.1 Extra: APA formatted manuscripts in `papaja`

If you want to write a reproducible manuscript in **APA style** (common in for example psychology) you might want to try out the R package `papaja` created by Frederik Aust. Install the

papaja package. **Restart RStudio.** Then, create a new R Markdown document, but instead of selecting the document option, select the ‘From Template’ option, and select the template APA article (6th edition) provided by the papaja package.



You will see a template with a lot of fields for you to fill in, such as the title, author names and affiliation, the author note, the abstract, etc. Papaja takes care that all this information ends up in a nice lay-out – exactly following the APA rules. This means that if you have installed MiKTeX (to be able to convert to PDF), you can knit the document to a pdf, and submit an APA formatted document that is completely reproducible. For a tutorial covering all options in papaja, including how to add citations: [https://crsh.github.io/papaja\\_man/index.html](https://crsh.github.io/papaja_man/index.html)



The screenshot shows the RStudio interface with a code editor window titled "Untitled1". The code is written in R Markdown, featuring syntax highlighting for R code and text. The code includes sections for title, author, affiliation, author note, abstract, keywords, and wordcount. It also contains comments and placeholder text for entering notes and abstracts.

```
1 -> ---  
2   title      : "The title"  
3   shorttitle : "Title"  
4  
5   author:  
6     - name    : "First Author"  
7     affiliation : "1"  
8     corresponding : yes # Define only one corresponding author  
9     address    : "Postal address"  
10    email      : "my@email.com"  
11    - name    : "Ernst-August Doele"  
12    affiliation : "1,2"  
13  
14   affiliation:  
15     - id      : "1"  
16     institution : "Wilhelm-Wundt-University"  
17     - id      : "2"  
18     institution : "Konstanz Business School"  
19  
20   author_note: >  
21     Complete departmental affiliations for each author (note the indentation, if you start a new paragraph).  
22  
23   Enter author note here.  
24  
25   abstract: >  
26     Enter abstract here (note the indentation, if you start a new paragraph).  
27  
28   keywords    : "Keywords"  
29   wordcount    : "X"  
30
```

## 14.7 Step 7: Organizing Your Data and Code

It is important to always **organize your data files and analysis files**. This helps others to quickly find the files you are looking for. In general, I recommend the **TIER protocol**: <https://www.projecttier.org/tier-protocol/>. If you share a datafile that is slightly larger than the one in this example, make sure you add a codebook so others understand what all variables mean. For a nice package that can help you to generate a machine readable codebook, see Arslan (2019).

When you are organizing your code, **take great care to make sure that any personally identifying information in your data is stored safely**. Open science is great, but you are responsible to share data responsibly. This means that you need to **ask participants permission to share their data** in the informed consent form (a [useful resource](#) is the Research Data Management Support page of the University of Utrecht). Whenever you collect personal data, make sure you [handle this data responsibly](#). Information specialists at your university library should be able to help.

## 14.8 Step 8: Archiving Your Data and Code

Although we have uploaded our data and code to GitHub, when you publish your article and want to **share your data and code**, it is important to remember that GitHub is not a data repository that guarantees long term data storage. GitHub is currently owned by Microsoft, and companies can choose to do with their free service whatever they want. This makes it less suitable to link to GitHub in scientific articles, because articles could be around for

decades from now. For scientific publications, you will want to link to a stable long-term data repository. For a **list of data repositories**, click [HERE](#). We will use the Open Science Framework (OSF) in this example as a stable data storage, because it is very easy to just integrate our GitHub repository within an OSF project.

Log in to the OSF at <https://osf.io/> (create an account if you haven't already done so). Click 'Create new project'. Give your project a name (for example 'Stroop Reproducible Analysis Assignment').

It is again important to add a **license** to your work, also on your OSF project. After having created a project, you can click 'Add a license':

License: Add a license

Choose a license:

Year:

Copyright Holders:

The MIT License (MIT)

Copyright (c) 2018 Daniel Lakens

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights

You can again choose an MIT license. You will have to fill in a year, and the copyright holders – in this case, that is you, so fill in your name (it will appear in the license text). Then click

Save

Although we could upload all our files to the OSF, we can also simply link our GitHub project to the OSF. In the menu bar of the OSF project, click on ‘Add-ons’. In the list, scroll to GitHub:

The screenshot shows a table titled 'Select Add-ons' with a search bar at the top. The left column is 'Categories' with options 'All', 'Citations', and 'Storage'. The right column lists external services with their icons and status: Dropbox (Enable), figshare (Enable), GitHub (Enable), GitLab (Enable), Google Drive (Enable), Mendeley (Enable), and OneDrive (Enable). A vertical scrollbar is visible on the right side of the table.

| Categories | Search...                           |
|------------|-------------------------------------|
| All        | Dropbox <a href="#">Enable</a>      |
| Citations  | figshare <a href="#">Enable</a>     |
| Storage    | GitHub <a href="#">Enable</a>       |
|            | GitLab <a href="#">Enable</a>       |
|            | Google Drive <a href="#">Enable</a> |
|            | Mendeley <a href="#">Enable</a>     |
|            | OneDrive <a href="#">Enable</a>     |

Follow the step-by-step guide to connect to GitHub provided by the OSF: <https://help.osf.io/hc/en-us/articles/360019929813-Connect-GitHub-to-a-Project>

Select your repository that contains the reproducibility assignment and click ‘Save’.

The screenshot shows the 'Configure Add-ons' page. It displays a GitHub connection status: 'GitHub authorized by Daniel Lakens' with a 'Disconnect Account' link. Below it, a 'Current Repo:' dropdown is set to 'Lakens/reproducibility\_assignment'. There are two green buttons: 'Save' and 'Create Repo'.

| Configure Add-ons                  |                                    |
|------------------------------------|------------------------------------|
| GitHub authorized by Daniel Lakens | <a href="#">Disconnect Account</a> |
| Current Repo:                      |                                    |
| Lakens/reproducibility_assignment  | <a href="#">Save</a>               |
|                                    | <a href="#">Create Repo</a>        |

Click the title of the OSF project page to go back to the main project page. You will now see in the ‘Files’ pane that the GitHub repository is linked:





OSFHOME ▾

Stroop Reproducible Analysis Assignme...

Files

Files

Click on a storage provider or drag and drop to upload

Name ^ v

- Stroop Reproducible Analysis Assignment
  - GitHub: Lakens/reproducibility\_assignment (master)
    - .gitignore
    - + Data
  - LICENSE
  - + Manuscript
    - README.md
    - reproducibility\_assignment.Rproj
  - OSF Storage

This is a good moment to click the ‘Make Public’ button in the top right of your project. After making the project public, people will be able to find it on the OSF. If you don’t want to make your project public just yet, but you do want to give others access to your files, you can create a ‘**View-only**’ link on the OSF. Go to ‘Contributors’ and click the +Add button next to View-only links. For a step-by-step guide, see [this tutorial](#).

## View-only Links + Add

Create a link to share this project so those who have the link can view—but not edit—the project.

You can use a view-only link to share access to your files only with reviewers. You can create an anonymized view-only link to hide your contributor names in the project - this is particularly useful in **blinded peer review**. Giving access to the files during peer review greatly helps reviewers – it will be a lot easier for them to answer any questions they might have about your materials, data, or code. Be aware it means that people you don’t know will have access to your files. So far, I don’t know of any negative experiences with this process, but it is important to be aware that others have access to your files before they are published.

The OSF page now just links to the files on the GitHub page. It does not independently store them. This means we do not yet have a **long term stable data storage solution**.

To create a snapshot of all files in the GitHub repository that will be stored for a long time, you have to create a **Registration** of your project. **We will not create a Registration of your project in this example.** Creating a registration starts several formal procedures: data in linked repositories (such as GitHub) are stored by the OSF, and the project appears in the list of registrations. You should only register when **you want to create a stable copy of your work**. Below you see an example of the files in an OSF project that has been registered. You see that the GitHub repository that was linked to the project has been turned into an Archive of GitHub – this creates a stable version of the project, as it was at the moment you registered.

| Name  | Modified |
|---|----------|
| Equivalence Testing for Psychological Research:...        |          |
| - OSF Storage   |          |
| + Archive of GitHub: Lakens-EquivalenceTe...              |          |
| - Equivalence Testing for Psychological Resea...          |          |
| - OSF Storage   |          |
| PREPRINT_Lakens_etal_EquivalenceTe... 2018-02-19 05:19 PM |          |
| + Equivalence Testing for Psychological Res...            |          |

A good moment to create a stable version of your project is when your manuscript is accepted for publication. You can create a Registration and use the Digital Object Identifier (DOI) to link to the code, data, and materials in the paper (you can add this link to the DOI to the manuscript as you check the proofs of your article before it is published). Note that it is recommended to link to the materials using the **DOI**. The DOI is a persistent link (meaning it will keep working) where a website address might change. A registration does not automatically get a DOI. After creating the Registration, you need to click the ‘Create DOI’ link to create a persistent object identifier.

## Contributors: Daniel Lakens

Date registered: 2016-11-10 04:52 PM

Date created: 2016-07-14 09:04 AM

[Create DOI](#)

Category:  Project

If you are ready to create a Registration, follow the instructions on the OSF: <https://help.osf.io/article/158-create-a-preregistration>. As an example of a Registration that was made to store all work related to one of my scientific publications, see <https://doi.org/10.17605/OSF.IO/9Z6WB> (this link is itself an example of how to link to an OSF project using a DOI).

#### **14.8.1 EXTRA: Sharing Reproducible Code on Code Ocean**

If you have used the workflow above to create a reproducible manuscript, you might want to make it easy for other people to explore your data and code. People can simply clone your GitHub repository – but this still requires them to install the software you used, and even if they have R installed, it requires them to install the packages you used to analyze your data. This can potentially lead to reproducibility problems. Packages in R update and change over time, and code that works on one machine might not run well on another computer. Furthermore, even when shared perfectly, downloading all files and getting the code up and running takes time. Several solutions exist, such as [Packrat](#), which is a dependency management system for R, or [Docker](#), which can create a container that works as a virtual machine that includes a computing environment including all the libraries, code and data that you need to reproduce an analysis (Wiebels & Moreau, 2021).

Here, I'll focus on a software solution that is designed to be easier to use than Docker, but provides many of the same benefits, called [Code Ocean](#). Code Ocean is a cloud-based computational reproducibility platform. You can create a computing capsule that runs online and contains all packages your code needs to run. Although Code Ocean does not (yet) guarantee long term storage of data and code, it is an interesting way to make your reproducible code available to fellow researchers, and makes it very easy for researchers (or reviewers) to make small changes to your code, and examine the results. Create a (free) account on CodeOcean. Go to the dashboard. Click the 'New Capsule' button and choose the option 'Import Git Repository.

+ New Capsule ▾



Create Blank Capsule



Import Git Repository

Enter the web address of your GitHub repository and click import.

X

## Import Git Repository

Create a new Code Ocean capsule from an existing git repository.

Enter the URL of any public Git repository:

`https://github.com/Lakens/reproducibility_assignment`

Cancel

Import

Click on ‘Environment’ in the left pane. In the middle pane, click the R icon to select the

programming language. At the time of writing this part of the chapter, the default language is 3.5.3, but you can click on ‘2 more versions’ and select the newest R version.

## Starter Environments

We've assembled some common languages and frameworks to get you up and running quickly. You can further customize these environments with multiple languages and additional packages in the next step.

 language: R

*By Language:*



### R (3.5.3)

Select

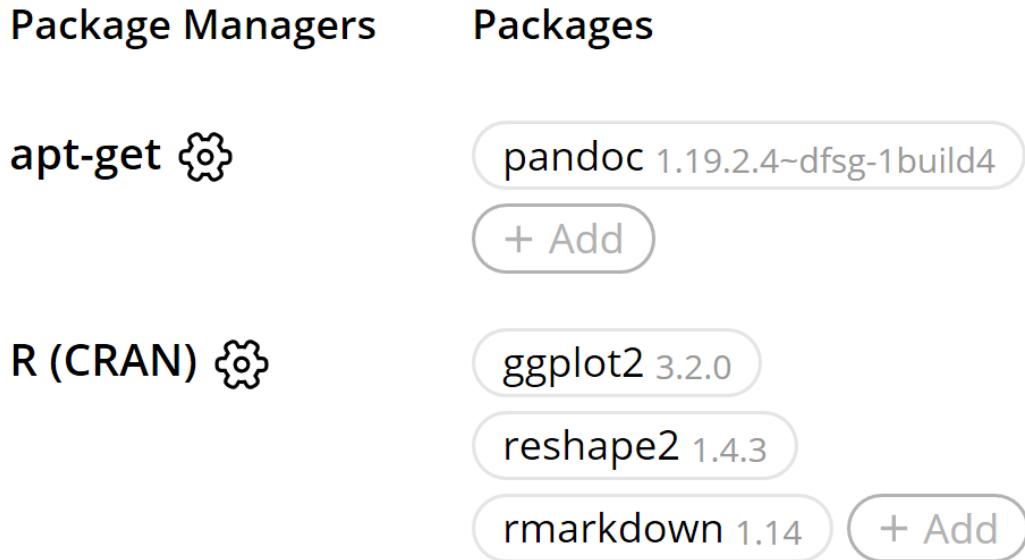
R is a language and environment for statistical computing and graphics

Ubuntu 18.04 R

[2 more versions >](#)

We need to set up the Code Ocean environment to contain the packages we need. Click +Add

behind apt-get, and type pandoc and click the return key two times (no need to specify a version of pandoc). Click +Add behind R (CRAN) and type in ggplot2, and click return twice (no need to select a specific version). Click +Add again, and type reshape2, click return twice. Click add once more, and type rmarkdown and click return twice.



In the left pane, drag and drop the main\_analysis.Rmd file into the ‘code’ folder, and the ‘stroop.csv’ file into the ‘data’ folder. Select the code folder, and click the + button in the bar on the top of the left pane to add a file. Name the file ‘run.sh’. This file will tell Code Ocean what to run. Select the file. The middle pane will be empty. Add the following code:

```
\#!/bin/bash  
Rscript -e "rmarkdown::render(input = 'main_analysis.Rmd', \  
output_dir = '../results', clean = TRUE)"
```

We need to make one final change. Select the main\_analysis.Rmd. Scroll down to line 28 where the data is read in. Change:

```
stroop_data \<- read.csv("stroop.csv", sep = " ", header = TRUE)  
to
```

```
stroop_data \<- read.csv("../data/stroop.csv", sep = " ", header = TRUE)
```

We are now all done! Click the ‘Reproducible Run’ button in the right pane. You will get output:

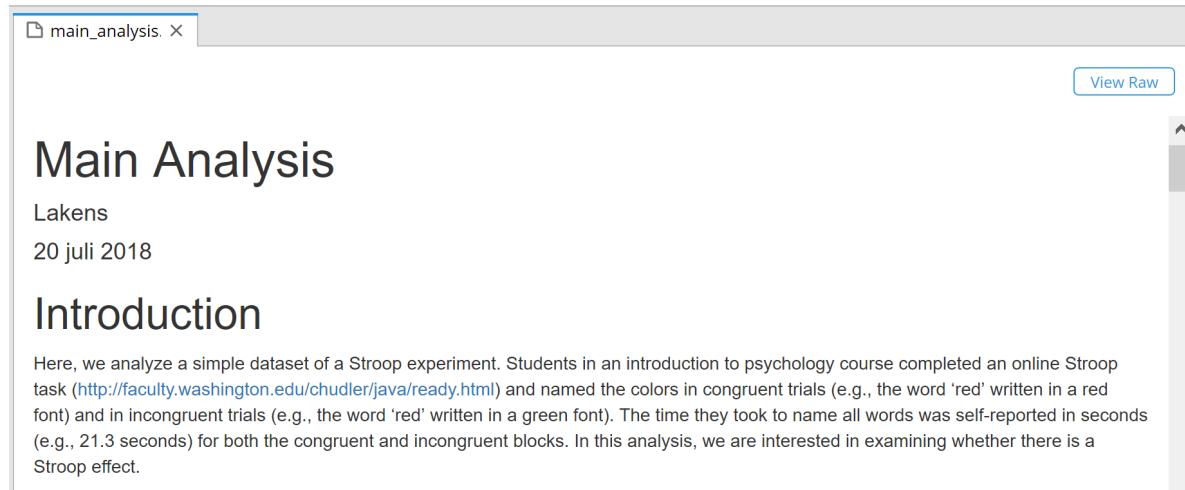
 Daniel Lakens ran  
Jul 24, 2019

⌚ 0:00:07 ✓

▼  Run 3995032

|  |           |
|--|-----------|
|  buildLog           | 856 B     |
|  main_analysis.html | 906.08 KB |
|  output             | 3.12 KB   |

Click on the ‘main\_analysis.html’. You will see the script has generated our reproducible results. Commit the changes. You could (in principle – for this example we won’t) make this completely reproducible container openly available and share it with your published paper.



main\_analysis.X

View Raw

## Main Analysis

Lakens

20 juli 2018

### Introduction

Here, we analyze a simple dataset of a Stroop experiment. Students in an introduction to psychology course completed an online Stroop task (<http://faculty.washington.edu/chudler/java/ready.html>) and named the colors in congruent trials (e.g., the word ‘red’ written in a red font) and in incongruent trials (e.g., the word ‘red’ written in a green font). The time they took to name all words was self-reported in seconds (e.g., 21.3 seconds) for both the congruent and incongruent blocks. In this analysis, we are interested in examining whether there is a Stroop effect.

There is now a completely reproducible analysis file online. Anyone cannot just reproduce your data analysis – they can go into the main\_analysis.Rmd file, and change anything they want in your code, and run it again. For example, let’s say you dislike the black straight line in the

first scatterplot, and you want it to be red. It is easy to change ‘black’ to ‘red’ in line 42, and re-run the analysis, and you will get a figure with a red line. Although that might in itself not be very exciting, the ability to easily re-analyze data might be useful in more realistic scenarios. For example, imagine you are reviewing a paper where the researchers do not plot the data. Without having to install any software, you can just type in `hist(stroop_data$Congruent)` after the data has been read in (e.g., on line 30), run the code again, and you will see a histogram for the reaction times in the Congruent condition. Give it a try.

## 14.9 Some points for improvement in computational reproducibility

We recently tried to computationally reproduce Registered Reports published in the psychological literature (Obels et al., 2020). We noticed some issues that, if solved, would easily improve the computational reproducibility of your work.

First, always add a codebook to data files. We already noted this above, and yes, it is a bit of work and not that fun to do, but it is very essential to include a codebook when you share data. Data is easier to understand and more reusable if variables and their values are clearly described. Researchers should ensure that the codebook and variable names are in the same language as the article.

Second, annotate code so it is clear what the code does. Well-annotated code makes clear what the analysis code does, in which order scripts should be run if there are multiple scripts (e.g., to pre-process the raw data, compute sum scores, analyze the results, and generate graphs), and which output each section of analysis code generates. Sometimes it might even be helpful to, from the final manuscript, copy-paste the sentences in the results section back into the code file, so it is very clear how sentences in the manuscript relate to the code file. It also helps to clearly structure code (e.g., using a README) so others know which output analysis code creates and in which order code should be run.

Third, check whether the code you shared still reproduces all analyses after revisions - researchers often make changes during the peer review process, but forget to update their analysis files.

Finally, remember that most code in R relies on specific libraries (also called packages). List all the packages that the code needs to run at the top of the script. Because packages update, it is necessary to report the version numbers of packages that were used (for example using `packrat`, or copying the output of the `sessionInfo()` function as a comment in the script). Remember that folder names and folder structures differ between computers, and therefore you should use relative locations (and not absolute paths like “`c:/user/myfolder/code`”). RStudio projects and the ‘here’ package provide an easy way to use relative paths. When multiple scripts are used in the analysis, include the order in which scripts should be performed on the data in a README file.

## 14.10 Conclusion

In this chapter, we used a number of platforms and software solutions, such as GitHub, the Open Science Framework, R Studio, R, and R Markdown. Following through the example here is not the same as being able to use these tools in your research. Learning to use these tools will take time. There will be many frustrations when the code or software doesn't work as you want, or when you have gotten your local and remote GitHub repositories so much out of sync you just need to delete everything on your local computer and re-download all files from GitHub (`git reset --hard [HEAD]` is your friend). There are a lot of resources available online to find answers, or to ask for help. In my experience, it is some work, but it is doable, even if you have very limited knowledge of programming. You can get a basic reproducible workflow up and running by simply using all the steps described here, and then learn new skills as you need them. Learning these skills is appreciated both within and outside of academia (which is useful for PhD students) and will quickly save you time (e.g., when recreating figures for a revision, or when analyzing very similar datasets in the future). A reproducible workflow also improves the quality of your scientific work, and makes it easier for other scientists to re-use your work in the future. For another open educational resource on making your scientific research accessible and reproducible, see [The Open Science Manual](#) by Claudio Zandonella Callegher and Davide Massidda.

# 15 Research Integrity

When doing research, it is important to be guided by responsible conduct of research, or more colloquially, **good research practices**. Good research practices are professional standards that have the goal to maximize the quality and reliability of research. On an abstract level beliefs about good research practices do not change substantially over time. But in practice the implementation good research practices changes as a function of social, political, and technological developments. For example, it is increasingly seen as a good research practice to share all data underlying the research you report. This was difficult before the internet, but has become much easier now free data repositories exist online. As a consequence, we increasingly see that research funders expect data collected with their grants to be open whenever possible.

A distinction is made between **research integrity** and **research ethics**. Research integrity is a set of principles based on professional standards. Research ethics is a set of moral principles, such as autonomy, beneficence, non-maleficence, and justice (Gillon, 1994). The principle of autonomy leads to research practices such as informed consent, the requirement to be truthful to participants, and confidentiality. From the principle of non-maleficence it follows that researchers should avoid research that harms participants, or research that is too burdensome (Varkey, 2021).

The professional standards in Codes of Conduct for Research Integrity vary slightly between documents (Komić et al., 2015). In this chapter, I will discuss both the [European Code of Conduct for Research Integrity](#), and the [Netherlands Code of Conduct for Research Integrity](#). Throughout this chapter, code of conduct for research integrity will be abbreviated to ‘code of conduct’. You might have to adhere to other codes of conduct, depending on where you work.

As the European code of conduct states, “A basic responsibility of the research community is to formulate the principles of research, to define the criteria for proper research behaviour, to maximise the quality and robustness of research, and to respond adequately to threats to, or violations of, research integrity.” Codes of conduct are always living documents, because what we believe is ‘proper research behavior’ changes over time. There are certain core principles you will see underlying all codes of conduct of research integrity, such as honesty, transparency, scrupulousness, accountability, reliability, respect, and independence. These underlying principles are translated into more specific behaviors that are considered proper behavior – both for researchers, as for research institutions. Have you ever read the code of conduct? If not, then your institution is already in violation of the code of conduct, as it is their responsibility

to “develop appropriate and adequate training in ethics and research integrity to ensure that all concerned are made aware of the relevant codes and regulations”.

The Dutch Code of Conduct for Research integrity states to “Conduct research that can be of scientific, scholarly and/or societal relevance”. Of course, you might perform a study for purely *educational* purposes, and the study you perform does not have to have any additional value (although it is always nice if it does). But researchers should prevent **research waste**, where they perform studies that have little to no value. Chalmers and Glasziou (2009) discuss four sources of research waste: Choosing the wrong questions for research, doing studies that are unnecessary, or poorly designed (which is why you need to evaluate the value of the information you will collect, as explained in the chapter on sample size justification), failure to report research promptly or at all (as explained in the chapter on **bias**), and biased or unusable reports of research (which can be prevented by reporting your study so it can be included in a future **meta-analysis**). The Dutch code of conduct also explicitly states researchers should “Make sure that your research design can answer the research question”. As you can see, many of the topics discussed in this textbook relate to preventing research waste, and are thereby related to research integrity.

Researchers should share their data when possible, which is also explained in the code of conduct: “As far as possible, make research findings and research data public subsequent to completion of the research. If this is not possible, establish valid reasons for their non-disclosure.” As discussed below, the General Data Protection Regulation (GDPR) requires European researchers to ask permission to share data that is collected. Old informed consent forms did not have such a question, and even often stated that data would be destroyed several years after data collection. This is a good example of updated professional standards, because nowadays, it is much more common to expect data to be available alongside the published article for perpetuity. You will therefore want to make sure you use updated consent forms that allow you to share the data you collect.

Younger researchers sometimes feel that their supervisors require them to act in ways that are not in line with the code of conduct. Some young researchers would not go along with such pressures, but others explicitly say they are willing to violate the code of conduct if this will get them closer to completing their PhD (van de Schoot et al., 2021). Others trust their supervisors to know what is the right thing to do, even though supervisors themselves might feel forced to act in ways that violate the code of conduct by *their* managers. Not surprisingly, pressuring people you have power over to violate the code of conduct is a violation of the code of conduct. For example, the Netherlands code of conduct states that “As a supervisor, principal investigator, research director or manager, refrain from any action which might encourage a researcher to disregard any of the standards in this chapter”.

Some researchers have noted how hypercompetition in science for research grants, as well as how researchers are individually rewarded for the number of published articles, can lead to unethical behavior (M. S. Anderson, Ronning, et al., 2007; Edwards & Roy, 2017). The Netherlands code of conduct stresses the importance of creating an open, safe, and inclusive research culture where researchers can discuss such pressures, as well as how to guarantee good

research practices are always followed. If you want to report and discuss suspected irregularities that you perceive as a violation of the code of conduct, universities typically have both internal and external confidential advisors that you can reach out to, and sometimes it is even possible to report these suspicions completely anonymously through services such as [SpeakUp](#). I would highly recommend, both for scientific integrity as well as for your own well-being, to discuss problematic behavior that you encounter with people you can trust, such as a confidential advisor.

We would not have problems with researchers violating the code of conduct if doing the right thing was always the easiest thing to do. Violating the code of conduct can come with immediate individual rewards, such as a higher probability of publishing a paper in a high impact journal, and it comes at long term collective costs for the reliability of scientific research, which can also impact the public's trust in science (Anvari & Lakens, 2018; Wingen et al., 2020). Social scientists might recognize this situation as a social dilemma, where what is best for the individual is not aligned with what is best for the collective. Changes in incentives structures can perhaps align individual and collective rewards. One way is to find and punish researchers who knowingly violate the code of conduct (for an example, see this story about [Brian Wansink](#)). New [bias detection tests](#) such as *p*-curve and *z*-curve analysis can also be used to identify researchers who have systematically used questionable research practices (discussed in the next section). In the end, even though it might sound idealistic, I believe all scientists should put science first. If you pursue a career in science at a public university you are paid by tax money to generate reliable knowledge. Nothing you do while pursuing additional goals, such as a successful career, should get in the way of the responsibility society has trusted you with, which is generating reliable and trustworthy knowledge.

## 15.1 Questionable Research Practices

Although in theory all researchers should follow the code of conduct for research integrity, many researchers do not. Researchers across scientific disciplines admit to certain practices that have been dubbed ‘questionable research practices’. This name is somewhat unfortunate, as most of these practices are not questionable at all, but directly violate the code of conduct. That they are nevertheless referred to as ‘questionable’ is mainly because many researchers were not aware of the problematic nature of these practices, and slowly needed to accept how problematic they always were.

Questionable research practices generally describe practices that violate the requirement from the code of conduct to “Make sure that the choice of research methods, data analysis, assessment of results and consideration of possible explanations is not determined by non-scientific or non-scholarly (e.g. commercial or political) interests, arguments or preferences.” In addition to commercial or political interests, many scientists have an interest in publishing scientific articles, as doing so is good for their career. Questionable research practices make it easier for researchers to publish their article, either because they increase the probability of being

able to report statistically significant results, or because they hide imperfections, which makes the results seem more convincing than they are. These practices come at the expense of the truth.

Researcher admit to engaging in questionable research practices, and depending on the community of researchers surveyed, several problematic practices are engaged in at least once by many scholars. Figure 15.1 summarizes the results from 14 different surveys (Agnoli et al., 2017; Bakker et al., 2021; Chin et al., 2021; Fiedler & Schwarz, 2016; Fraser et al., 2018; John et al., 2012; Latan et al., 2021; Makel et al., 2021; Moran et al., 2022; Motyl et al., 2017; Rabelo et al., 2020; Swift et al., 2022). However, coding of open ended questions suggest there is substantial measurement error when participants answer these items, so it is unclear whether the percentages in Figure 15.1 directly translate into the percentage of researchers actually engaging in questionable practices (Motyl et al., 2017).

Many researchers selectively publish only those results or analyses with significant results, despite the Dutch code of conduct stipulating that researchers should “Do justice to all research results obtained.” and the European code of conduct stating that “Authors and publishers consider negative results to be as valid as positive findings for publication and dissemination.” Registered Reports have been an important step in aligning research practices with the code of conduct when it comes to publishing null results.

Researchers also flexibly analyse their data by selectively reporting conditions, measures, covariates, and a host of other data analytic strategies that inflate the Type 1 error rate, and increase the probability of obtaining a statistically significant result. Preregistration has been an important step of increasing the transparency of data-driven choices in the analyses reported in scientific articles, and allows researchers to evaluate whether any deviations from the statistical analysis plan decrease the severity of the test, or increase it (Daniël Lakens, 2019). With increasing awareness of the problematic nature of these practices, hopefully we will see a strong decline in their occurrence, and researchers will learn correct approaches to maintain some flexibility in their analyses (for example by replacing optional stopping by [sequential analysis](#). Wigboldus & Dotsch (2016) make the important distinction between questionable research practices, and questionable reporting practices. Whenever in doubt, transparently reporting the decisions you made while analyzing data should give researchers all the information they need to evaluate the reported results.

```
Warning: Removed 1 row containing missing values or values outside the scale range
(`geom_bar()`).
```

## 15.2 Fabrication, Falsification, and Plagiarism

Beyond questionable research practices, fabricating data is making up results and recording them as if they were real, and falsification is manipulating/manipulating aspects of research,

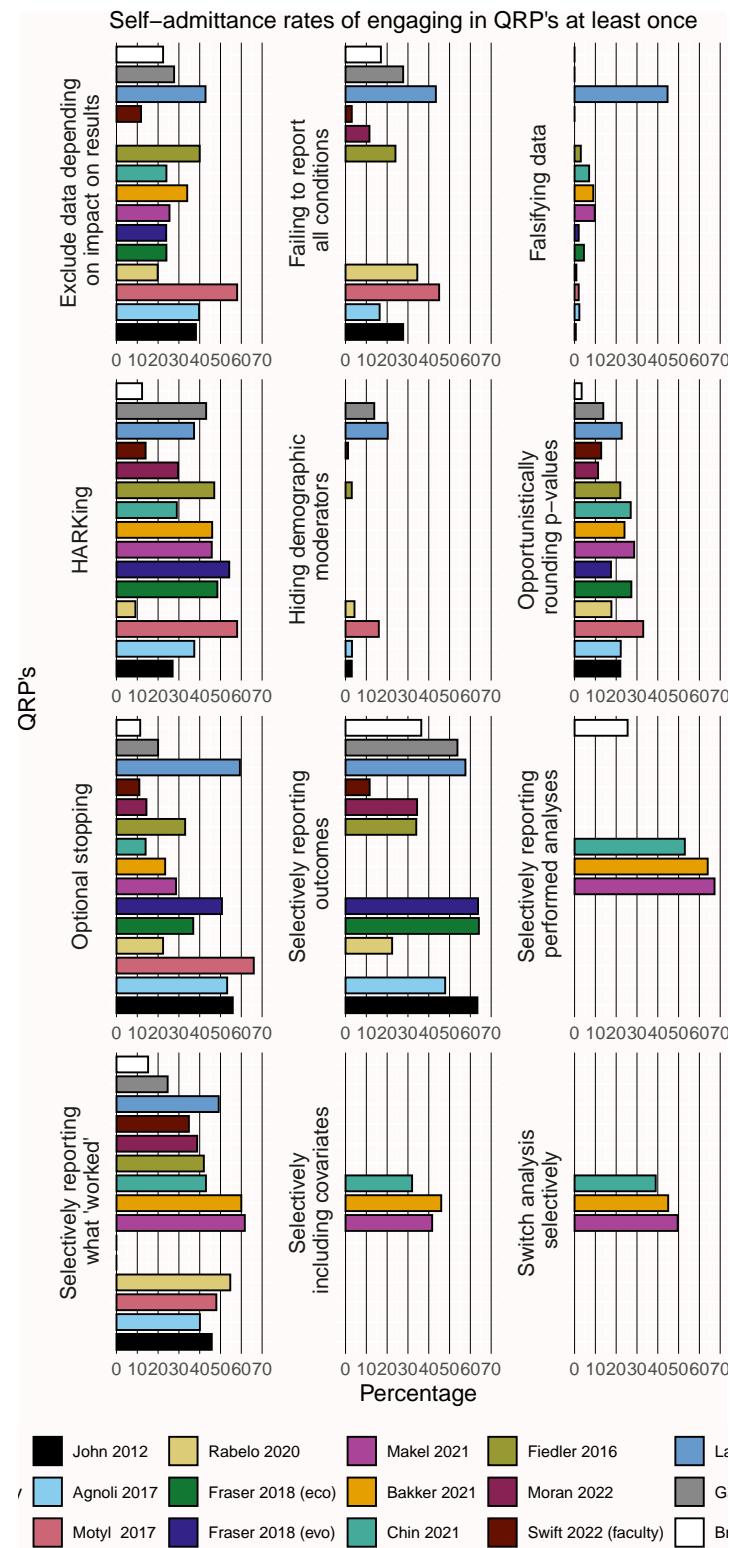


Figure 15.1: Self-admittance of engaging in a questionable research practices at least once from 14 surveys among a variety of samples of researchers.

including data, without any scientific justification. Data fabrication is a research practice that is outright dishonest. There have been a [substantial number of cases](#) where researchers have fabricated complete datasets in dozens of experiments. Some examples were already mentioned in the chapter on [bias detection](#)(#sec-bias). It can be difficult to prove fabrication, as researchers often keep bad records of data collection. For example, Susannah Cahalan makes a convincing case in her book '[The Great Pretender](#)' that the famous study by David Rosehan '[On being sane in insane places](#)' was largely made up. In the study, healthy confederates who pretended to hear voices were admitted as in-patients suffering from schizophrenia. Her detailed investigation raises severe doubts the study was performed as described (see also Scull (2023)).

One might hope falsification and fabrication is rare, but a recent large scale survey in The Netherlands yielded prevalence estimates around 4% (Gopalakrishna et al., 2022). Data fabrication can also occur on a smaller scale. Imagine collecting data for a study. As part of the study, it is your task to ask the age of participants and their gender, for the demographic statistics to be reported when describing the sample. After collecting all the data, you notice you have forgotten to collect the demographic data for two individuals. You might be tempted to, based on your memory, guess the demographic statistics of these two individuals, to not have to admit you have made a mistake during the data collection when you wrote up the demographic information. However, this would also constitute data fabrication. You should instead transparently mention a mistake was made. Mistakes happen, and it is important to create a culture where people can admit mistakes, so that we can learn from them and prevent them in the future (Bishop, 2018).

Note that it can be fine to **simulate** data to perform a power analysis – one should just not present such data as if it was collected from real participants. The Dutch code of conduct states: “Do not fabricate data or research results and do not report fabricated material as if it were fact. Do justice to all research results obtained. Do not remove or change results without explicit and proper justification. Do not add fabricated data during the data analysis.”

The European code of conduct defines plagiarisms as: using other people’s work and ideas without giving proper credit to the original source, thus violating the rights of the original author(s) to their intellectual outputs.” It is possible to re-use text, but the source should be cited, and quotation marks should be used to identify the text as a quote from another source. A special case of plagiarism is ‘self-plagiarism’ or text recycling where the same text by the same author is used in different articles. There is disagreement about how problematic this practice is (Bird & Sivilotti, 2008), which is to be expected, as there will always be some academics with a diverging opinion. In general, researchers are not supposed to re-use large portions of previous work and present it as new work just to increase their number of published articles. But many researchers believe it is perfectly fine to re-use descriptions from method sections if you need to communicate the same information in a new paper (Pemberton et al., 2019). The guidelines by the Committee on Publication Ethics (COPE) similarly [state](#):

The guidelines cover how to deal with text recycling both in a submitted manuscript

and a published article and include situations where text recycling may be acceptable as well as those where it is unlikely to be. For example, it may be entirely appropriate to have overlap in a methods section of a research article (referring to a previously used method) with citation of the original article. However, undisclosed overlap, or overlap in the results, discussion, or conclusions is unlikely to be acceptable.

Self-plagiarism is thus mainly seen as problematic when researchers use it to publish very similar content multiple times purely to make it look like they are more productive.

There are additional problematic research practices beyond fabrication, falsification, plagiarism, and QRP's. Gopalakrishna et al. (2022) also considered behavior such as insufficiently mentoring or supervising junior researchers, unfairly reviewing articles or grants, and inadequate note-taking of the research done as questionable practices.

### **15.3 Informed consent and data privacy**

When collecting data from participants outside of naturalistic observations in the public space, they should consent to participate in research. A consent form that participants read and sign before data collection is important both for research ethics, as for data privacy. The consent form explains the goal of the study, highlights that participation is voluntary and that participants can stop when they want, explains any risks and benefits (such as payment), informs them about data privacy issues, and details who participants can contact if there are any issues with the study.

Consent is also the legal basis for the use of personal data in the General Data Protection Regulation ([GDPR](#)). The consent form should identify the data controller and the contact details of the Data Protection Officer, and a description of the participants' rights (e.g., to withdraw the data up to a certain amount of time after the study), and information about where and how long data is stored and shared (Hallinan et al., 2023). According to the GDPR there are special categories of personal data that you can only collect with informed consent, such as racial or ethnic origin, political opinions, religious or philosophical beliefs, genetic or biometric data data, and questions about a person's sex life or sexual orientation. When collecting such data, it should be necessary for the research purpose. Data privacy officers at your university can assist you in this process.

Open data is important - but it is essential to maintain data privacy when sharing data in a public repository. This means carefully removing any personal identifiers (names, IP addresses, ID numbers of participants from data panels, etc) from the dataset before publicly sharing the data. If you use a version control system make sure that the identifying information is absent from the initial version of the data files you will share, as other users will not just have access to the latest version of the file, but also to the complete file history. For a good overview on the GDPR and research, see this information from [Groningen University](#).

## 15.4 Conflicts of Interest

A **conflict of interest** in research is any situation in which a researcher has interests in the outcome of the research that may lead to a personal advantage that can get in the way of generating true knowledge. A central feature of a conflict of interest is that there are two competing interests: one for doing good research, and one for failing to do good research. For example, a researcher might receive additional income as a consultant for a company, while working on a study that evaluates a product this company produces, such as a novel drug. If the study shows the drug does not have any benefits compared to existing drugs, a researcher might worry that honestly communicating this research finding will make the company decide to no longer hire their services as a consultant. Or a researcher might work for an advocacy organization and perform a study on the same topic that examines how many people are impacted by this topic, where high estimates might be in the interest of the advocacy organization. An argument can be made that scientists have a conflict of interest whenever they publish a scientific paper, as publishing is good for the career of a scientist, and studies are easier to publish when

Simply having a conflict of interest is not a violation of the code of conduct, as long as researchers are transparent about it. The European code of conduct states: “All authors disclose any conflicts of interest and financial or other types of support for the research or for the publication of its results.” Conflicts of interest can also emerge when you review the scientific work of peers (e.g., grant proposals, or scientific articles). Here, personal relationships can become a conflict of interest, either because you are very close friends with a researcher, or because you feel the other researcher is a rival or competitor. In those situations, you should again declare your conflict of interest, and an editor or grant review panel will typically try to find another reviewer.

## 15.5 Research ethics

Before you perform research most institutions require you to obtain permission from an **ethical review board** (ERB), or sometimes called an institutional review board (IRB). Specific types of research might be reviewed by specialized boards. For example, medical research is reviewed by a medical ethics review committee (METC), and animal research by an animal ethics committee. The goal of ethics review is to balance two goals: to protect subjects and to enable research that will benefit society (Whitney, 2016). The Declaration of Helsinki provides an important basis of the evaluation of research on human subjects. It highlights the right of individuals for self-determination and the right to make informed decisions whether they want to participate or stop participating in research.

The Declaration of Helsinki builds on the [Nuremberg Code](#), a set of ethical principles developed after the second world war in response to unethical research Nazi doctors performed on unconsenting prisoners in concentration camps (for an ethical discussion about whether this

unethical research should be used and cited, see Caplan (2021) and Moe (1984)). Another example of unethical experiments on human subjects is the [Tuskegee syphilis study](#) where 400 African American men with syphilis were included in a study to examine the effects of the disease when untreated. The men included in the study did not give consent to go untreated and did not receive their diagnosis. The study ended up continuing for 40 years. Although most studies that are performed at universities have a much lower risk of harm, it is still important to evaluate the possible harm to participants against the benefits for science. Researchers might show negative stimuli, or ask participants to remember events they experienced as negative, which can still be experienced as harmful. There might be equally effective alternatives that can be used when designing a study, that will still allow a researcher to answer their research question. In addition to preventing harm, researchers must inform participants about the study and ask for their consent to participate. The information in the informed consent should be truthful. If it is necessary to lie to participants in the informed consent about the study they will perform (for example, the participants believe they will interact with other participants, but these people are actually confederates and part of the study) this should be explained after data collection has been completed in a **debriefing**. Researchers should also maintain the **confidentiality** of participants. Take special care when collecting open questions when you plan to share the data in a public repository.

## 15.6 Test Yourself

**Q1:** Try to define ‘data fabrication’ in a single sentence. Start the sentence with ‘Data fabrication is any process through which’. Your definition should cover all forms of data fabrication that are dishonest, but it should not cover honest processes, such as *simulating* datasets.

**Q2:** Imagine you are analyzing your data, and one participant has entered an age of 117 in a text-entry question in an experiment they performed behind a computer. Although it is not impossible to have this age, it is perhaps more likely that the participant intended to enter the value 17. Should you change the value to 17? Now imagine you have measured the amount of time (in seconds) people browse a website using the system clock on your computer, which is extremely accurate, and time measurement is perfectly reliable. There is an experimental condition, and a control condition. There is no statistically significant difference between the two groups. However, if you change the data of one participant in the control condition from 117 seconds to 17 seconds, the difference between groups is statistically significant, and confirms the prediction you made when designing the study.

What is the difference between these two situations? Why is the second recoding of 117 to 7 a violation of the code of conduct for research integrity, according to the quote from the Netherlands Code of Conduct for Research Integrity three paragraphs above this question? If you write up the average age of participants after having changed the age of this one participant

from 117 to 17, what do you need to provide in addition to the statement ‘the mean age of participants was 20.4’ when this number is based on data you changed?

**Q3:** The practice of sometimes reporting results, but other times not reporting results is referred to as **selective reporting**. When it comes to selective reporting, it is again the intention of the researcher that matters. It might make sense to not report a study that was flawed (e.g., there was a programming mistake in the experiment, or all participants misunderstood the instructions and provided useless input). It might also make sense to not extensively report a study that was badly designed – for example, you thought a manipulation would have a specific effect, but the manipulation does not work as intended. However, even such data might be useful to others, and the knowledge that the manipulation you thought would have a specific effect has no effect might prevent others in the future of making the same mistake. It would at least sometimes be beneficial for science if such results were shared in some way. But, as we will see below, researchers also choose to selectively report studies based on whether the results were statistically significant or not.

A scientist performs several experiments, but only shares the results of those experiments that, after looking at the results, yield an outcome that supported their predictions. This scientist never shares the results of experiments that fail to support their predictions. How morally acceptable or unacceptable do you think the actions of this scientist are?

**Q4:** A scientist performs several experiments, but only shares the results of those experiments that, after looking at the results, are judged to have been well-designed. This scientist never shares the results of experiments that, after looking at the data, are judged to be badly designed. How morally acceptable or unacceptable do you think the actions of this scientist are?

**Q5:** A scientist performs one experiment in which several dependent variables are analyzed in multiple ways, but only shares the results of those analyses that, after looking at the results, yield an outcome that supported their predictions. This scientist never shares the results of analyses that fail to support their predictions. How morally acceptable or unacceptable do you think the actions of this scientist are?

Current practice is that researchers do selectively report studies. When Franco et al. (2014) examined what happened to 106 studies part of a large collaborative national representative survey, they found that if the results yielded non-significant effects, 31 studies were not written up, 7 were written up but not published yet, and 10 were published. When results showed strong (statistically significant) effects, only 4 had not been written up, 31 were written up but not yet published, and 56 were published. There is clear evidence researchers selectively report results that confirmed their hypotheses, as we discussed in the chapter on bias.

A recent study by Pickett and Roche (2017) examined the public perception of data fabrication, and selective reporting. Their results are summarized in the table below. As you can see, selective reporting is judged to be morally unacceptable by a large proportion of the public (71% believe it is morally unacceptable), and the majority of the public thinks there should be consequences when it is done (e.g., 73% believe such researchers should receive a funding

ban). How do these percentages in the study by Pickett and Roche reflect your own judgments about how morally acceptable or unacceptable selective reporting is?

**Table 1** Experimental findings: community members' evaluations of data falsification and fabrication ( $n = 415$ ) versus selective reporting ( $n = 406$ ) (Study 1)

| Variables                  | Falsification and fabrication<br>Support (%) | Selective reporting<br>Support (%) | $z$   | $p$   |
|----------------------------|--|------------------------------------|-------|-------|
| Morally unacceptable       | 96   | 71                                 | 9.59  | <.001 |
| Should be fired            | 96   | 63                                 | 11.75 | <.001 |
| Should receive funding ban | 93   | 73                                 | 7.75  | <.001 |
| Should be a crime          | 66   | 37                                 | 8.34  | <.001 |

Figure 15.2: Table from Pickett and Roche (2017) showing judgments of how moral selective reporting<sup>1</sup> and data fraud are in the eyes of members of the general public.

**Q6:** Assuming the results observed by Pickett and Roche, as well as the results by studies on questionable research practices summarized in Figure 15.1 are accurate and representative, there seems to be a large divide between current research practices, and what the general public think is morally acceptable. Do you think this divide is problematic? Do you think that if the general public was perfectly aware of current practices related to selective reporting, they would have a reason to evaluate the ways scientists work negatively, or do you think that with a good explanation of current practices, the general public would evaluate current practices positively?

**Q7:** Given that researchers admit to using questionable research practices, they must have some benefits. What are benefits of using questionable research practices?

**Q8:** What are downsides of using questionable research practices?

To improve research practices, we have seen many scientific fields move towards greater transparency. This includes sharing data and materials, clearer reporting of choices that were made during the data analysis, and pre-registering planned studies. It is almost impossible to prevent all fraud, but making research more transparent will make it easier to detect questionable research practices, such as selective reporting. At the same time, universities need to train people in research ethics, and make sure there is a climate where researchers (including you!) feel comfortable to do the right thing.

## 15.7 Grade Yourself

For this assignment, you will grade yourself. You will be able to check suggested answers below (which are an indication of what would be a good answer, although not exhaustive –

your answer might highlight correct important points not mentioned in the answers below). Read through the answers below and determine a grade for your own answers. Use a grading from 1 (very bad answer) to 10 (excellent answer). Be truthful and just.

**Answer Q1:** Data fabrication is any process through which data are generated that can pass for real data, but that are not based on real underlying observations that were actually made by a researcher. The data are nevertheless presented as if they are based real observations.

*Score yourself between 1 (no answer) to 10 (perfect answer) points. Your grade should be higher, the better you indicated fabricated data look similar to real observations, and that they are intentionally presented as if they are real.*

**Answer Q2:** The difference between the two cases is that in the second case, a researcher has the intention to generate an outcome that is in line with the outcome they want to observe. In terms of the quote by the Netherlands Code of Conduct of Research Integrity, what is missing is “*explicit and proper justification*”. What you need to provide if you report an average based on a 17 instead of a 117 is a footnote or statement indicating what you did (‘We changed one age value of 117 to 17’) and the justification for this (‘because we strongly suspected the value was a type on the participant was actually 17 years old’).

*Score yourself between 1 (no answer) to 10 (perfect answer) points. Your grade should be higher, the more aspects of the answer you provided (explaining the difference between the two cases based on the absence of a proper justification, specifying which aspect of the Netherlands Code of Conduct of Research Integrity is missing in the second case, and that you need to describe what you have changed, and the justification for changing it.*

**Q3, Q4, Q5, and Q6 are your personal opinion, and are not graded.**

**Answer Q7:** 1) because they are biased towards presenting support for their hypothesis to the world, 2) because they are much more strongly rewarded in their career for publishing results that ‘work’ than null results, and thus spend their time on the former, and 3) even if researchers would try to publish the results, journals are less likely to accept them for publication, 4) It is easier to publish a paper with a coherent story (only significant results). In general, we can expect the benefits of questionable research practices to be for individual scientists in the short run.

*Score yourself between 1 (no answer) to 10 (perfect answer) points. Your grade should be higher, the more of reasons you provided, including, but not limited to, the three above.*

**Answer Q8:** For an individual scientist, the risk is colleagues find out, and lose prestige (or in extreme cases, their job). Failures to replicate their work might also impact their prestige. For society, a downside is that scientific research is not as reliable as it should be. For science, a downside could be that the reputation of science, and the trust people place in science, is damaged. In general, we can expect the costs for questionable research practices are for society in the long run.

*Score yourself between 1 (no answer) to 10 (perfect answer) points. Your grade should be higher, the more of reasons you provided, including, but not limited to, the three above.*

If after all this work on research integrity you feel like you need something to cheer you up, [this video](#) might help.

# 16 Confirmation Bias and Organized Skepticism

I cannot give any scientist of any age better advice than this: The intensity of the conviction that a hypothesis is true has no bearing on whether it is true or not. The importance of the strength of our conviction is only to provide a proportionately strong incentive to find out if the hypothesis will stand up to critical evaluation.  
*Medawar, 1979, Advice to a Young Scientist*

Being a scientist is a rewarding but challenging career path. Doing science can lead to the intellectual satisfaction of making discoveries or increasing our understanding about important questions, the rewarding feeling of contributing to solutions to important problems society faces, interacting with stimulating colleagues, recognition from peers and the general public, as well as the possibility of a decent income if you become an internationally sought-after expert in your field. At the same time, it can be a difficult career that requires hard work, uncertainty about your future career, times where you have little success in advancing your knowledge, experiencing competitiveness or even animosity towards other scientists, and a feeling of pressure to achieve goals (National Academy of Sciences et al., 2009). Although science is a collective endeavor, scientists often has a strong personal commitment to their work. They are motivated to succeed, and disappointed if their work is not successful.

In his book “Modern science and the nature of life” William Beck (1957) writes:

Each successive step in the method of science calls for a greater emotional investment and adds to the difficulty of remaining objective. When the ego is involved, self-criticism may come hard (Who ever heard of two scientists battling to prove the other right?). One has always a vested interest in the successful outcome and, whether we enjoy admitting it or not, each of us feels the pressure to succeed, to blaze ‘new trails’ perhaps before we have mastered the old, to remain productive and therefore admired, to embark obsessively (as did Sigmund( upon a romantic crusade towards epic truth. It is apparent, therefore, how latent neurotic tendencies may impinge upon and distort the clean mandates of scientific method and may generate error, unrealistic values, anxiety, and – let’s face it, since science is done behind closed doors – dishonesty. Because scientists are human and science is not, as in all fields the thin thread of integrity is sometimes strained to break.

The recognition that science is a human activity has not gone unnoticed. In 1620 Francis Bacon wrote the book ‘Novum Organum’ (or ‘New Method’) which provided a first description of a modern scientific method, with a focus on empiricism and inductive reasoning. Bacon already realized more than 400 years ago that people are not passive observers, and provides a very early description of what we would now call *confirmation bias*:

The human understanding, when any proposition has been once laid down (either from general admission and belief, or from the pleasure it affords), forces everything else to add fresh support and confirmation; and although most cogent and abundant instances may exist to the contrary, yet either does not observe or despises them, or gets rid of and rejects them by some distinction, with violent and injurious prejudice, rather than sacrifice the authority of its first conclusions. It was well answered by him who was shown in a temple the votive tablets suspended by such as had escaped the peril of shipwreck, and was pressed as to whether he would then recognize the power of the gods, by an inquiry, But where are the portraits of those who have perished in spite of their vows? All superstition is much the same, whether it be that of astrology, dreams, omens, retributive judgment, or the like, in all of which the deluded believers observe events which are fulfilled, but neglect and pass over their failure, though it be much more common. But this evil insinuates itself still more craftily in philosophy and the sciences, in which a settled maxim vitiates and governs every other circumstance, though the latter be much more worthy of confidence. Besides, even in the absence of that eagerness and want of thought (which we have mentioned), it is the peculiar and perpetual error of the human understanding to be more moved and excited by affirmatives than negatives, whereas it ought duly and regularly to be impartial; nay, in establishing any true axiom the negative instance is the most powerful.

In a classic paper on confirmation bias, Nickerson (1998) defines confirmation bias as **the seeking or interpreting of evidence in ways that are partial to existing beliefs, expectations, or a hypothesis in hand**. The human factors that influence (or bias) scientific knowledge generation received relatively little attention from philosophers of science, even though it would be naïve to believe that scientists objectively pursue the truth. As the philosopher of science Chang (2022) writes:

There is a tendency in the philosophy of science to present the scientist as a ghostly being that just has degrees of belief in various descriptive statements, which are adjusted according to some rules of rational thinking (e.g., Bayes’s theorem) that remove any need for real judgement. Whatever does not fit into this bizarre and impoverished picture, we tend to denigrate as matters of ‘mere’ psychology or sociology.

The sociologist of science Robert Merton (1942) believed that “Four sets of institutional imperatives - universalism, communism, disinterestedness, organized scepticism - comprise the ethos of modern science.” *Universalism* means that “The acceptance or rejection of claims entering

the lists of science is not to depend on the personal or social attributes of their protagonist”. *Communism* means that “The substantive findings of science are a product of social collaboration and are assigned to the community”. Scientist do not own their theories – at best they receive recognition for developing their ideas. As Merton writes “Secrecy is the antithesis of this norm; full and open communication its enactment.” *Disinterestedness* occurs not on the individual level – a scientist can have a passions and motivations – but on the institutional level. The institution of science has disinterestedness as a norm, which means that claims should be truthful, and not spurious. According to Merton, scientists are subject to rigorous policing – scientists are accountable to their peers, who will check their work, and therefore only disinterestedness will lead to claims that survive scrutiny. And finally, *organized skepticism* means the “scrutiny of beliefs in terms of empirical and logical criteria”. Claims are only accepted after they have survived scrutiny by peers.

As with any norm, not all individuals subscribe to all norms, and more importantly, not everyone behaves in line with norms (at least not all the time). For example, a common norm is to be truthful when you talk to others. Yet, even if we subscribe to the norm to be truthful, we might not always tell the truth ourselves. And we might believe others lie even more often than we do. This is exactly the pattern that Anderson and colleagues (2007) found in a survey among US scientists (see Figure 16.1). Scientists subscribed to Mertonian norms (the maximum possible score is 12). They also admit not to always follow these norms in their own behavior, and they believe others follow these norms even less. The pattern is the opposite for counternorms (e,g, secrecy, self-interestedness, etc.).

When asked, scientists don’t see members of their own profession as being objective at all. In an interesting series of interviews with scientists involved in the Apollo moon landing, Mitroff (1974) concludes: “Every one of the scientists interviewed on the first round of interviews indicated that they thought the notion of the objective, emotionally disinterested scientist naïve”. His article is full of excellent quotes that illustrate this conclusion, such as:

Scientist B: The uninvolved, unemotional scientist is just as much a fiction as the mad scientist who will destroy the world for knowledge. Most of the scientists I know have theories and are looking for data to support them; they’re not sorting impersonally through the data looking for a theory to fit the data. You’ve got to make a clear distinction between not being objective and cheating. A good scientist will not be above changing his theory if he gets a preponderance of evidence that doesn’t support it, but basically he’s looking to defend it. Without [emotional] commitment one wouldn’t have the energy, the drive to press forward sometimes against extremely difficult odds. You don’t consciously falsify evidence in science but you put less priority on a piece of data that goes against you. No reputable scientist does this consciously but you do it subconsciously.

Scientist G: Every scientific idea needs a personal representative who will defend and nourish that idea so that it doesn’t suffer a premature death.

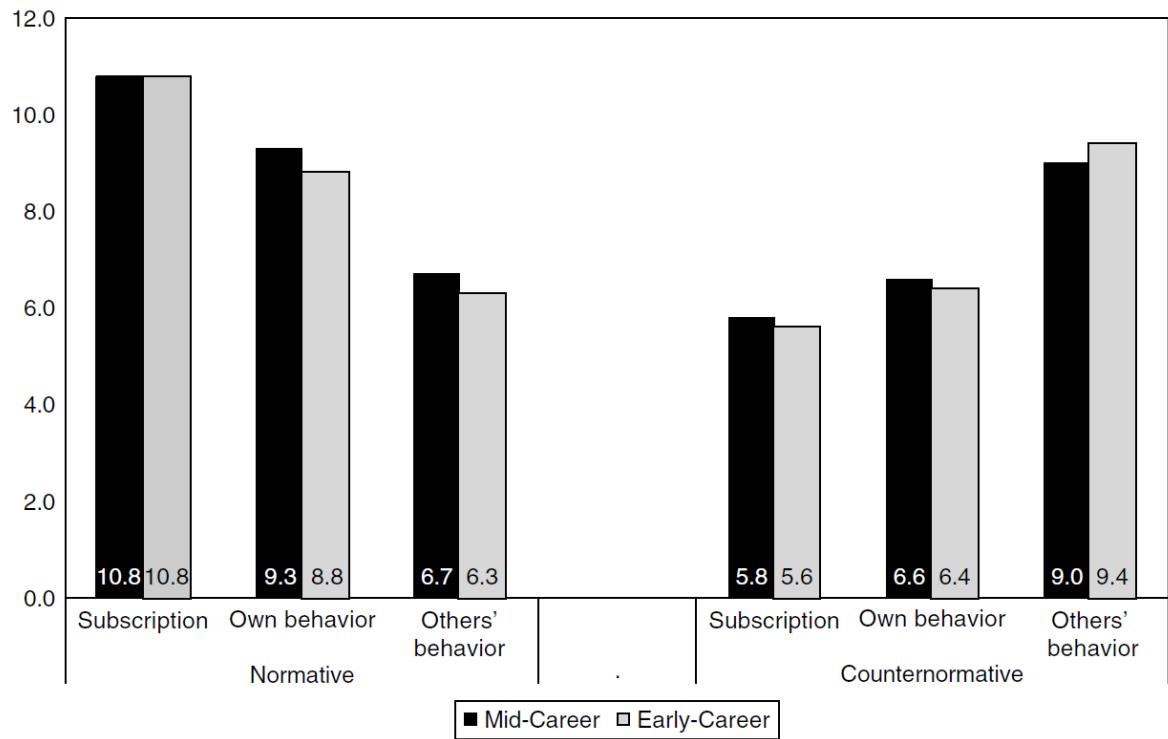


Figure 16.1: Means of Normative and Counternormative Subscription and Behavior from Anderson et al., (2007).

These interviews reveal that scientists believe a commitment to a specific idea or theory is a necessity if you want to motivate yourself to keep exploring an idea, even when the going gets tough, or to make sure an idea is not too easily dismissed. In other words, confirmation bias could even have a positive role to play.

Although there are now philosophers of science who recognize that science is a social process (Douglas, 2009; Longino, 1990), most researchers who study human factors in scientific research come from fields such as **psychology of science**, **sociology of science**, **science and technology studies**, or **meta-science**. Researchers in these fields try to describe ways in which researcher fall victim to confirmation bias, analyze the underlying mechanisms that cause confirmation, and propose interventions to reduce the effect of confirmation bias in science.

For example, Mahoney (1979) reviewed the literature related to the common textbook description of scientists as objective, rational, open-minded, intelligent, acting with integrity, and openly and cooperatively sharing knowledge, and concluded:

1. The scientist is not immune to perceptual biases and is frequently quite emotional in response to technical and epistemological matters.
2. It remains to be demonstrated that scientists are more logical than nonscientists in the conduct and interpretation of their work.
3. The scientist may sometimes be unreceptive to relevant data and – particularly in the case of theorists - prone to hasty speculation and dogmatic tenacity.
4. Although scientists generally have higher IQs than non-scientists, it remains to be demonstrated that small intellectual differences have a substantial impact on professional competence or contribution.
5. Reports of data fabrication and experimenter bias suggest that such phenomena are neither rare nor trivial.
6. Scientists tend to be secretive and suspicious until they have established a public priority claim to their work; disputes over personal credit and priority frequently result in bitter arguments.

So why does science still seem to work, given all these all too human limitations that scientists display? One way to look at science is as a method that groups of people use to make claims while implementing procedures that aim to reduce the role of confirmation bias. Although science encompasses much more than a set of rules to reduce confirmation bias, many practices, such as peer review, performing independent replication studies, and specifying the alpha level of a test before looking at the data, can only be understood from this perspective. Some scientists consider active attempts to resist confirmation bias an essential feature of good science. As Feynman (1974) writes: “The first principle is that you must not fool yourself - and you are the easiest person to fool.”

## 16.1 Confirmation bias in science

Wason (1960) created a simple task to examine how people test hypotheses. You first get a series of 3 numbers. Your task is to develop a hypothesis about the underlying rule that has generated these three numbers. You can then test the underlying rule by suggesting a new set of three numbers, and you will be told if the set of three numbers follows the rule you are supposed to discover, yes or no. Let's give it a try. I will give you the following 3 numbers: 2, 4, 8.

You can think of a rule that has generated this set of three numbers. To test your rule, you can provide a new set of three numbers. Take a moment to think of which 3 numbers you would want to suggest, and then you will hear if the numbers follow to the rule, or not. Let's say you have decided to suggest three numbers such as 3, 6, 12, or 5, 10, 20. These numbers are in line with a rule 'the first numbered is doubled, and then doubled again'. If you would have suggested three numbers like this, you would have heard they follow the rule you were supposed to discover. However, had you provided the three numbers 2, 3, 9, you would *also* have received the answer that this set of three numbers follows the underlying rule. The rule to be discovered was 'three numbers in increasing order of magnitude'.

If you are like most people who complete the Wason task, you tested a set of three numbers that would confirm the rule you had in mind. Having the rule confirmed tells you your rule might be correct, but that many other rules can also be correct. Testing a set of three numbers that you predict would not follow the rule, such as 1, 2, 3, and learning this set of three numbers actually follows to underlying rule, tells you with certainty that the rule you had in mind is incorrect. Confirming and falsifying predictions is both important, but people seem in general less inclined to try to prove themselves wrong. This knowledge about human psychology is useful to have, because we can use it to develop methods and procedures to counteract negative effects that arise from our inclination to want to confirm our hypotheses.

In his paper titled "Pathological Science" Langmuir (1989) discusses two examples of confirmation bias in physics. The first example is the Davis-Barnes effect, which described unexpected behavior of alpha particles interacting with electrons in a magnetic field, and the second example is N-rays, a hypothesized form of radiation inspired by the discovery of X-rays, described by French physicist Blondlot in 1903, and initially confirmed by others physicists. In both cases skepticism of the initial findings led other scientists to perform an on-site inspection of the experiment being performed, who concluded the results were due to observer error. As Langmuir writes: "These are cases where there is no dishonesty involved but where people are tricked into false results by a lack of understanding about what human beings can do to themselves in the way of being led astray by subjective effects, wishful thinking or threshold interactions."

There are also cases where dishonesty *is* involved. Sometimes scientists commit outright scientific fraud, and fabricate data, but it is not always clear where to draw the dividing line between intentional and unintentional bias. For example, in a famous case of the geneticist

Gregory Mendel who studied heredity in pea plants. Later re-analyses of his data by the statistician and geneticist Ronald Fisher revealed that his results are implausibly close to predicted outcomes (Ronald A. Fisher, 1936). Although there is agreement that the results are statistically implausible, it is difficult to pinpoint a cause. The statistical implausibility could be due to incorrectly reporting details of the experiment, classification errors, or even an assistant feeling some pressure to report results in line with expectations (Radick, 2022). One reason to embrace open science practices is so that the research community will benefit from greater transparency about what happened in situations where researchers raise doubts about the validity of results.

It is not just scientists who **fabricate data** – students do this as well. In an incredibly interesting paper documenting attempts to perform a replication study as a class assignment, Azrin and colleagues (1961) found that many of the students fabricated all or part of the data because following the experimental procedure was too difficult. In one class experiment, only a single student reported having trouble performing the experiment as it was supposed to be carried out. When students discussed the experiment later during the course, and the honest student admitted that they had tried to perform the experiment 6 times, but failed and gave up, 8 other students suddenly also admitted that they had problems following the experimental procedure, and had deviated substantially from the instructions. Even worse, in another class assignment replicating the same study, when one student asked “I’m having trouble with my experiment; can you tell me how you did yours?” 12 out of 19 students questioned this way readily admitted to fabricating data to this fellow student.

We can imagine many reasons why students would fabricate data, such as not wanting to admit they failed at following experimental instructions and feeling stupid, or simply fabricating data to not have to do any actual work. In a class I co-taught with a colleague many years ago students also fabricated data. We had asked them to collect data for a short survey from 10 friends of family members, just so that they would have real data to analyze during the course. At the time we did not realize the survey students created (also as part of the course) end up being much longer than a few minutes, nor did we realize that many students found it unpleasant to have to ask 10 people for a favor. None of the students told us they had difficulties following the instructions – instead many of them fabricated surveys until they could hand in 10 surveys. As teachers, we had obviously asked our students to complete an unreasonable task. But had a student honestly told us about the difficulty they experienced collecting the data, we would have adjusted the assignment (as we did the year after). The code of conduct for research integrity applies to staff and students. Whenever you feel pressure and are considering to violate the code of conduct (for example by fabricating data), don’t! Instead, bring the problem to the attention of a teacher, or a confidential advisor if you are more comfortable talking to someone else.

As discussed in the section on questionable research practices, sometimes researchers opportunistically use flexibility in their research methods to increase the probability of finding support for their hypotheses. It is often unclear to which extent researchers are aware of how problematic this behavior is, and therefore it is difficult to establish when this behavior is

simply dishonest, and when it is bias through a lack of understanding. These practices have been known for a long time. Kish (1959) already mentioned as one misuse of statistical tests: “First, there is”hunting with a shot-gun” for significant differences. [...] The keen-eyed researcher hunting through the results of one thousand random tosses of perfect coins would discover and display about fifty “significant” results (at the  $P = .05$  level). Perhaps the problem has become more acute now that high-speed computers allow hundreds of significance tests to be made.”

Barber (1976) reminds us that “Since experiments are designed and carried out by fallible individuals, they have as many pitfalls as other human endeavors” and provides an extensive overview of ways researchers might bias their conclusions. He lists many ways in which researchers can bias the results they observe, either as experimenter (e.g., treating people in the experimental condition slightly differently than people in the control condition) or as investigator (e.g., analyzing data in many different ways until a significant result has been observed). These concerns only received widespread attention in psychology at the start of the replication crisis, for example through the article ‘False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant’ (Simmons et al., 2011).

A final mechanism through which confirmation bias operates is known as **citation bias**, where authors selectively cite research that supports the claims they want to make, while ignoring evidence to the contrary. Citations are an important practice in science. They are used in the introduction of scientific articles to provide an overview of the knowledge that exists, and motivates the research that is performed. They also give credit to scientists who performed research that was useful, and the number of times articles are cited is often used as a metric to – in part – evaluate the excellence of research that scientists have done. It is therefore important that authors cite articles that deserve to be cited. Instead, authors often selectively cite the literature (Duyx et al., 2017). For example, statistically significant results are more likely to be cited than non-significant claims, which amplifies to already substantial effect of publication bias. Researchers might not cite criticism on their work, or on the measure they use, or the statistical analysis approach they use, to prevent peer reviewers from identifying possible weaknesses. Finally, scientists might have non-scientific reasons to cite some articles, and to not cite other articles. Researchers sometimes need to cite their own past work, but they are also biased towards citing themselves even when those citations are not the most relevant. Scientists might prefer to cite work by their friends or people from their in-group, and not cite work by scientists they feel a certain animosity towards, or members of an out-group.

The scientific literature on the Hawthorne effect provides one illustration of citation bias. In several studies in the Hawthorne Works electric plant, researchers tested if different lighting conditions would influence productivity. The Hawthorne effect is used to describe an overall increase in productivity when workers know they are being studied – irrespective of the experimental conditions. This interpretation has been widely *criticized*, but researchers predominantly cite positive interpretations, and ignore the criticism (Letrud & Hernes, 2019). Another famous example is a short letter in the New England Journal of Medicine stating that

addiction was rare in patients treated with narcotics, which was massively cited as support for the safety of opioids, which is believed to have contributed to the ongoing opioid crisis in Northern America (Leung et al., 2017). Citation bias can be prevented by always reading the literature you cite (a surprisingly trivial, but regrettably necessary, recommendation, increasingly important now that AI tools are being used that create fake references), and systematically searching the literature, instead of relying on the most cited papers on the first Google Scholar results page.

Citation bias can also be actively used to make a journal article more convincing to readers. Corneille et al. (2023) mention this trick, alongside a list of other techniques scientists can use to make their claims sounds more convincing than they are. It will often require quite some content expertise to recognize these tricks, such as citing work that is weak or known to be incorrect, making claims not backed up by evidence, making claims that generalize well beyond the evidence provided in the article, selectively quoting other work, or citing it out of context, or downplaying limitations of a study. As you develop yourself as a scientist, you will learn to identify these tricks, but they are often more difficult to identify for less experienced researchers. Researchers can be motivated to make their claims look more novel or convincing, because this can help them to get their work published in more prestigious scientific journals, which is good for their career.

## 16.2 Organized Skepticism

By now it is clear that there is a risk that scientists will let their biases influence the claims they make. Following Merton's (1942) notion of **organized skepticism** there are a number of practices in science that exist to, as far as possible, counteract biases by enabling claims to be subjected to critical scrutiny.

### 16.2.1 Error control

When William Gosset (also known as Student of Student's *t*-test) wrote an internal document for the Guinness brewery detailing the usefulness of using error probabilities to "enable us to form a judgment of-the number and nature of the fresh experiments necessary to establish or disprove various hypotheses which we are now entertaining." Already in this first paper on the topic Gosset recognizes that it is useful to specify the error rate that will be used to draw conclusions in some objective manner, as: "it is generally agreed that to leave the rejection of experiments entirely to the discretion-of the experimenter is dangerous, as he is likely to be biassed. Hence it has been proposed to adopt criterion depending on the probability of such a wide error occurring in the given number of observations." A similar point is made by the biostatistician Irwin Bross (1971): "He writes: "in conformity with the linguistic patterns in setting conventions it is natural to use a round number like 5%. Such round numbers serve to avoid any suggestion that the critical value has been gerrymandered or otherwise picked

to prove a point in a particular study.” In short, “If researchers are allowed to set the alpha level after looking at the data, there is a possibility that confirmation bias (or more intentional falsification-deflecting strategies)” (Uygun Tunç et al., 2023). The use of a fixed alpha level (in most fields of 5%) is therefore an example of organized skepticism. Claims need to pass a criterion that controls erroneous conclusions before taken seriously.

### 16.2.2 Preregistration

In the chapter on ‘Investigator Data Analysis Effects’ Barber (1976) presents a first example of a pitfall that concerns choosing the hypothesis after looking at the data: “A serious potential pitfall is present when investigators collect a large amount of data and have not pre-planned how they are to analyze the data. [...] The major problem here is that the investigator decides how the data are to be analyzed after he has “eyeballed” or studied the data.” A researcher can use a fixed alpha level before looking at the data, but this is not sufficient to control erroneous conclusions if they subsequently pick the test they want to perform after identifying interesting patterns in the data. The solution to this problem is the preregistration of an analysis plan. Once again, preregistration is a form of organized skepticism. Researchers are not simply trusted to report their planned analyses in an unbiased manner. Instead, they are asked to use a method of testing hypotheses where peers can scrutinize whether the analyses were indeed planned before the researchers had access to the data. It is perfectly possible that deviation from an analyses plan withstand scrutiny by peers (Daniël Lakens, 2019), but researchers should allow others to transparently evaluate if the tests were not chosen opportunistically. When researchers in a field are expected to preregister their research (such as in clinical trials) preregistration is an institutional implementation of organized skepticism. The topic of preregistration is discussed in more detail in the chapter on [preregistration and transparency](#).

### 16.2.3 Independent Replication Studies

After a study has been performed and a conclusion has been reached, a subsequent step where the claim is scrutinized is when other researchers try to independently replicate the finding. As Neher (1967) writes: “Individual researchers often fail to recognize crucial but subtle characteristics of their sample and of their study, mechanical errors of recording and calculation, and errors arising from the researchers’ own influence and biases.” Independent replication provides a method to explore the extent to which such characteristics caused an effect. The usefulness of independent replication in psychology was already pointed out by Mack (1951) and Lubin (1957). Independent replication is equally important in other fields, such as particle physics (Junk & Lyons, 2020).

If a finding can be independently replicated by other researchers, it becomes less likely that the original claim is impacted by subtle characteristics of the original study. It is also less likely that the original study suffered from more serious problems, such as fraud or inflated

Type 1 error rates due to flexibility in the data analysis. A successful independent replication can not completely take away such concerns. As Bakan (1967) warns: “If one investigator is interested in replicating the investigation of another investigator, he should carefully take into account the possibility of suggestion, or his willingness to accept the results of the earlier investigator (particularly if the first investigator has prestige for the second investigator). He should take careful cognizance of possible motivation for showing the earlier investigator to be in error, etc.” It is always possible that the researchers involved in the independent replication shared the same systematic biases, or simply happened to observe a Type 1 error as well, but with each successful independent replication such concerns become less likely. A non-successful independent replication is more difficult to interpret. The researcher performing the replication might have been motivated to botch the experiment because they wanted to find a non-significant result. There might have been actual differences between the study studies that need to be explored in subsequent studies. But failed independent replications raise questions about the generalizability of claims, and if multiple people fail in independently replicating a study, that is a cause for concern.

#### 16.2.4 Peer Review

The prototypical example of organized skepticism in science is the peer review process. As the philosopher of science Helen Longino (1990) writes: “I have argued both that criticism from alternative points of view is required for objectivity and that the subjection of hypotheses and evidential reasoning to critical scrutiny is what limits the intrusion of individual subjective preference into scientific knowledge. [...] Peer review is often pointed to as the standard avenue for such criticism”. Getting criticism is often emotionally distressing for people, and receiving negative peer reviews is not a fun experience. Perhaps surprisingly, we do not teach young people how to deal with criticism. When others list all the things they believe are wrong with the work young researchers have spent months or maybe even years of their lives on, we just expect them to learn how to deal with the accompanying emotions. It is logical to feel bad if you receive strong criticism - especially when you feel the criticism is not fair, or overly harsh. Over time, most - but not all - researchers learned to become detached from the evaluation of their work. Try not to take criticism personal. After all, it is part of organized skepticism. Try your best, and use valid criticism to improve your work where possible.

The **peer review** process works as follows. When a scientist has written a manuscript they submit it to the scientific journal of their choice. Journals have editors who process submitted manuscripts. An editor will first check if the manuscript seems like it would be of interest to their readership, and if it seems to be of sufficient quality. If so, the manuscript is sent out for peer review. Scientific peers with expertise on the topic discussed in the manuscript will be approached over email, and asked if they want to provide a review. Editors typically try to find at least two peer reviewers, but sometimes more. Peer reviewers get access to the manuscript, but they are typically not allowed to share it with others – in other words, in most cases the peer review process is confidential. Peer reviewers write their reviews for

free, as part of their job as a scientist, and they typically get a number of weeks to complete the review. The editor will then read the reviews, and decide if the manuscript is rejected (the editor declines to publish it), accepted (the manuscript is considered to be of sufficient quality to publish it), or if the manuscript needs to be revised (which means authors address criticism and suggestions by the peer reviewers, and resubmit the manuscript to the journal). Sometimes there will be multiple rounds of peer review before a manuscript is accepted.

Peer review is typically anonymous. The names of peer reviewers are not known to anyone except the editor. Researchers self-report that they would be less likely to review for a journal if their identity is made public, and anecdotally mention that signed reviews would make it more difficult to be honest about manuscripts they believe are poor quality (Mulligan et al., 2013). A more recent survey found that 50.8% of almost 3000 scientists believe that revealing the identity of reviewers would make peer review worse (Ross-Hellauer et al., 2017). Almost two-thirds of respondents believed reviewers would be less likely to deliver strong criticisms if their identity became known to the authors. The anonymity has positive, but also negative sides. As Longino (1996) writes “its confidentiality and privacy make it the vehicle for the entrenchment of established views.” Reviewers might try their best to keep certain findings, such as failures to replicate their work or claims that falsify predictions of theories they have proposed, out of the scientific literature. Scientists even have a running joke about ‘Reviewer 2’ - the reviewer who is always extremely critical about your manuscript, maybe even up to the point where the reviewer is rude and impolite, and will recommend that your manuscript should be rejected based on weak arguments. Note that there is no empirical support for the idea that reviewer 2 is actually more negative in general. Scientists share negative experiences with peer review in a Facebook group ‘Reviewer 2 Must be Stopped’.

Because the peer review process is central to whether the scientific manuscripts of scientists will be published, there is both a lot of criticism on peer review, concerns about the quality of peer review, attempts to fake peer review (e.g., an alliance between researchers who review their own papers, see C. Ferguson et al. (2014)), as well as experiments with improving peer review. Recent developments are open peer review, where the content of reviews is made available, and signed reviews, where authors are not anonymous but attach their names to the reviews they submit), among many other innovations in peer review. After high-quality peer review, a paper should be well-vetted (i.e., contain no mistakes or incorrect claims), but it is also possible the quality of the peer review is low, and a manuscript still contains mistakes or incorrect claims. Peer review is only as good as the peers. For example, when scientific peers review a manuscript, but none of the peers is well-trained in statistics, it is perfectly possible that a manuscript contains incorrect statistical inferences. Furthermore, with the increases in time-demands on academic staff, it might be increasingly difficult to find good reviewers who have the time to review a manuscript, or the reviewers might spend very little time carefully checking the manuscript. Furthermore, although it is slowly changing with the rise of open science (Vazire, 2017), peer reviewers often do not have access to the materials, data, and analysis scripts during peer review, and they have to trust those part of the process have been competently performed, which is not always the case. For these reasons, although peer review

plays an important role in science when it is done well, you can not trust that all peer reviewed manuscripts are free of mistakes or incorrect claims.

Peer review typically occurs when a study is submitted for publication, but it can also take place after a study is published. This is known as **post-publication peer review** and occurs, for example, on platforms such as [PubPeer](#). Scientists do not always appreciate additional scrutiny of their work, but post-publication peer review has often revealed flaws in published work that the original peer reviewers missed, and it therefore should be considered a valuable additional tool that facilitates organized skepticism.

### 16.2.5 Double-Checking Errors

As Friedlander (1964) writes: “Errors in research do occur. Their prevalence should be viewed with alarm rather than passive acceptance as an essential concomitant of humans conducting research.” Friedlander uses himself as an example of a researcher who made an error. He computed reliability scores in a factor analysis, and found these to be surprisingly and distressingly low. He repeated the calculation, now finding a higher reliability score. As Friedlander observed: “A combination of displeasure and”distrust” of these results, plus a high sense of commitment to a nearly completed study, prompted the writer to repeat the arithmetic process used in computing the reliability coefficients. Greater care was evident in the repeated calculations, for the writer was rewarded with reliability coefficients all above .70! An additional repeat left these coefficients undamaged. Had the researcher not been displeased and surprised with the low reliability coefficients, it is doubtful that he would have repeated his calculations; a Type II error would have been committed.” Rosenthal (1966) provides an overview of several studies where researchers made recording errors when writing down responses by participants that were in the direction of their hypotheses. In short, errors happen, and they are more likely to happen in ways that support the researchers’ hypothesis.

We all make errors, and we might not check errors if we observe results in the predicted direction. One way to prevent biased double-checking is to double-check all analyses we perform. For example, Wichters (2011) writes: “my close colleagues and I have implemented a ‘co-pilot’ model for our statistical analyses, in which we share data between us for double-checking and preventing embarrassing errors.” Strand (2023) similarly writes: “If we start with the assumption that mistakes will happen even when people are trying to avoid them, we must come up with methods of checking our work to find those mistakes”. She explains how building the habit to double-check work within a research collaboration will function as one layer of protection in Reason’s ‘Swiss cheese’ model of accident causation. Implementing checks in all projects also reduces the idea that work is checked due to a lack of trust, as it simply becomes part of how a group operates. Errors can also be prevented by implementing other tools, such as computationally reproducible manuscripts that prevent copy-paste errors (Rouder et al., 2019; Strand, 2023).

### **16.2.6 The Devil's Advocate**

The Devil's advocate is a person who takes on the role of the skeptic and argues against the accepted or desired position, regardless of whether they believe in their arguments or not. The practice originates in the Catholic church where it was used while deciding to declare a person a Saint, where the *advocatus diaboli* argued against the canonization of a candidate, and opposed 'God's advocate' (*advocatus Dei*). The idea behind creating an explicit role for a Devil's Advocate that is assigned to one person in a group is that people in general do not like to give criticism because they fear interpersonal backlash. This is, as we saw above, also the reason that peer review is typically anonymous. When groups make decisions, no one is anonymous. By assigning a specific individual to the role of a Devil's Advocate, there is at least one person who will actively raise criticism, while they are shielded from any negative interpersonal consequences because it is their assigned duty to raise these criticisms. Additional benefits are that Devil's advocates promote a diversity of viewpoints, and counter the pressure to conform.

Of course, a Devil's Advocate needs to be listened to, and their role should not be merely ceremonial (an accusation Christopher Hitchens made when he was interviewed by the Vatican as a Devil's Advocate during the decision about the beatification of Mother Theresa). It should not be possible to state that your decision procedure used a Devil's Advocate, which was subsequently ignored, to pretend you prevented bias in decision making. Transparency about which criticism was raised, and how it was addressed, can help. Another issue is that a Devil's Advocate needs to have sufficient knowledge about good counter-arguments to be successful. Research shows that an authentic minority dissent (i.e., including some individuals who actually hold different views than the majority) might lead to higher quality decisions than a Devil's Advocate (Nemeth et al., 2001).

### **16.2.7 Adversarial Collaborations**

One way to guarantee that there is sufficient expertise among individuals arguing different sides of a debate is to organize a collaboration between disagreeing scientists (Rosenthal, 1966). Rosenthal writes: "For the resolution of theoretical and empirical issues important enough to engage the interest of two or more competent and disagreeing scientists, it seems worthwhile to coordinate their efforts more efficiently. At the design stage the opponents might profitably collaborate in the production of a research plan which by agreement would provide a resolution of the difference of opinion. At the stage of data collection, too, the opponents may collaborate either in person or by means of assistants provided by both scientists." If the two parties in such a collaboration each have their own preferred outcome, such as opposing theoretical predictions, research projects where both sides of a debate work together to resolve disagreements empirically are called *adversarial collaborations* (Mellers et al., 2001). An excellent example of a large international **adversarial collaboration** to design and conduct an

experiment that best tested and clarified disagreements among experts in the field about the facial feedback hypothesis was conducted by Coles et al. (2022).

For an adversarial collaboration to be successful researchers need to be able to design an experiment that will be able to differentiate between theories, following the principles of *strong inference* (Platt, 1964). This may not always be possible. Furthermore, there is often a lot of auxiliary assumptions that will need to be tested before any critical test of different theories can be performed. Finally, researchers involved in such a project might try to resist the ability of the study to falsify their theory, for example by remaining strategically ambiguous about which results would be, and which results would not be, predicted by their theory (Frankenhuis et al., 2022). Despite these difficulties, adversarial collaborations hold great promise to resolve longstanding debates in the field where relatively little progress is made.

Beyond new empirical studies, it can also be beneficial to write collaborative review papers with a larger team of researchers with different viewpoints. The journal “Psychological Science in the Public Interest” has such collaborative review articles as its main aim since the year 2000 (Ceci & Bjork, 2000). A carefully selected “blue-ribbon” team (i.e., a team consisting of exceptional researchers in the area) representing a range of viewpoints is instructed to provide a fair and balanced state of the art review on a specific topic. Such reviews can still be adversarial in nature (Crusius et al., 2020).

### 16.2.8 Red Team Science

All else equal, scientists should trust studies and theories that have been more critically evaluated. The more that a scientific product has been exposed to processes designed to detect flaws, the more that researchers can trust the product (Mayo, 1996). Yet, there are barriers to adopting critical approaches in science. Researchers are susceptible to biases, such as confirmation bias, or they may gain a competitive advantage for jobs, funding, and promotions by sacrificing rigor in order to produce larger quantities of research. And even if researchers are transparent enough to allow others to critically examine their materials, code, and ideas, there is little incentive for others—including peer reviewers—to do so. We can only trust findings in a field if there are self-correcting mechanisms that guarantee critical appraisal that will identify and correct erroneous conclusions (Vazire & Holcombe, 2022).

Finding ways to prove ourselves wrong is a scientific ideal, but it is rarely scientific practice. Openness to critique is nowhere near as widespread as researchers like to think. Scientists rarely implement procedures to receive and incorporate push back. Most formal mechanisms are tied to peer-review, which typically happens after the research is completed and the manuscript written up, but it is likely more beneficial to receive peer feedback before the data is collected (Daniël Lakens, 2023).

In science, “red teams” can be used in the form of a group of diverse scientific critics who criticize a research project from all angles and even act to counteract the biases of the original authors, in order to improve the final product. Red teams are used in the software industry

to identify security flaws before they can be discovered and exploited by malefactors (Zenko, 2015). Similarly, teams of scientists should engage with red teams at each phase of a research project and incorporate their criticism (Daniël Lakens, 2020). The logic is similar to the Registered Report publication system — in which protocols are reviewed before the results are known — except that criticism is not organized by journals or editors, but within a larger collaboration. Ideally, there is a larger amount of speedier communication between researchers and their red team than peer review allows, resulting in higher quality preprints and submissions for publication. Red Team members can be chosen because each member has an important expertise – e.g., a content expert, a statistical expert, a measurement expert, etc.) representing a much greater diversity and expertise that can typically be accomplished in peer review. Red teams are especially useful for highly sensitive or expensive research projects. They have not been used a lot in science, but some first steps are being taken to explore their usefulness.

### 16.2.9 Blinding

Knowledge that is not available to researchers can also not bias them. For example, some journals allow authors to submit an anonymized manuscript, without author names or any other hints of the identity of the authors, to prevent this knowledge from influencing the evaluation of reviewers about the manuscript.

**Double-blind studies**, where neither the participant nor the experimenter knows whether participants are in the experimental or control conditions, have the goal to prevent participant effects and experimenter effects (Rosenthal, 1966).

To prevent researchers from being biased during the analyses of their data they can rely on methods of **blind analysis**, where the data file they analyze no longer has any identifying information about which observations belong to which condition (MacCoun & Perlmutter, 2015). A colleague uninvolved in the data analysis will create an adjusted data according to one of several possible blinding strategies. The researchers will perform the analyses, and when all analyses are performed, there is an ‘unblinding party’ where the data is unblinded, and the researchers learn whether their predictions are supported on the unblinded data, or not.

### 16.2.10 Separating Theorists from Experimentalists

Another way to reduce experimenter bias is to introduce a task division in science between those individuals who develop the hypotheses, and those who test them. Such a distinction between theorists and experimentalists is common in many fields, and it could be fruitful if some tensions exist between both sides. As Moscovici (1972) writes: “Experiment and theory do not stand in a transparent relation to one another; it is the role of the theory to make experimentation unnecessary, and the role of experimentation to render the theory impossible”.

Rosenthal (1966) discusses the possibility of a professional experimenter whose only job it is to collect high quality data for other researchers who have developed the hypothesis to test: “The emotional investment of the professional experimenter would be in collecting the most accurate data possible. That is the performance dimension on which his rewards would be based. His emotional investment would not be in obtaining data in support of his hypothesis. Hypotheses would remain the business of the principal investigator and not of the data collector. There might, in general, be less incentive to obtain biased data by the professional experimenter than by the scientist-experimenter or the graduate student-experimenter.”

This distinction is commonly present in other fields, such as in experimental physics. As Junk and Lyons (2020) note that there is specialization in experimental particle physics between theorists and experimentalists. One benefit is that models are fully defined by theorists before they are tested. “The second benefit is that experimentalists almost never test theories that they themselves invented, helping to reduce possible effects of confirmation bias.” In psychology, a separation between experimentalists and theorists does not exist, but a similar divide between those who collect the data and those who interpret it theoretically can be achieved by letting other researchers write the discussion section of papers (Schoenegger & Pils, 2023): “Outsourcing the discussion section to papers not written by the authors of the original papers plausibly reduces personal biases across the board”.

### **16.2.11 Method of multiple working hypotheses**

In many scientific fields there is currently no tradition of specialization, and individual scientists do all tests involved in the research process – theorizing, experimental design, measurement development, data collection, data analysis, and reporting scientific results. In 1980 T. C. Chamberlin already observed how scientists tend to develop a preference for certain theories or explanations:

The moment one has offered an original explanation for a phenomenon which seems satisfactory, that moment affection for his intellectual child springs into existence; and as the explanation grows into a definite theory, his parental affections cluster about his intellectual offspring, and it grows more and more dear to him, so that, while he holds it seemingly tentative, it is still lovingly tentative, and not impartially tentative. So soon as this parental affection takes possession of the mind, there is a rapid passage to the adoption of the theory. There is an unconscious selection and magnifying of the phenomena that fall into harmony with the theory and support it, and an unconscious neglect of those that fail of coincidence. The mind lingers with pleasure upon the facts that fall happily into the embrace of the theory, and feels a natural coldness toward those that seem refractory. Instinctively there is a special searching-out of phenomena that support it, for the mind is led by its desires. There springs up, also, an unconscious pressing of the theory to make it fit the facts, and a pressing of the facts to make them fit the theory.

To prevent such affective processing from biasing knowledge generation Chamberlin proposes the **method of multiple working hypotheses**: Instead of entertaining and testing a single hypothesis, a scientist actively develops a large number of working hypotheses (Chamberlin, 1890). The idea is that none of these hypotheses has any preferential status, and a scientist can more objectively examine which is best corroborated by the data. Chamberlain writes: “The effort is to bring up into view every rational explanation of new phenomena, and to develop every tenable hypothesis respecting their cause and history. The investigator thus becomes the parent of a family of hypotheses: and, by his parental relation to all, he is forbidden to fasten his affections unduly upon any one.” If it is not possible to separate the theorists and the experimentalists, at least a single scientist can try to mentally embrace multitudes of theoretical ideas at the same time.

Platt (1964) was inspired by Chamberlin when developing his ideas on **strong inference**: “It seems to me that Chamberlin has hit on the explanation - and the cure - for many of our problems in the sciences. The conflict and exclusion of alternatives that is necessary to sharp inductive inference has been all too often a conflict between men, each with his single Ruling Theory. But whenever each man begins to have multiple working hypotheses, it becomes purely a conflict between ideas. It becomes much easier then for each of us to aim every day at conclusive disproofs - at strong inference – without either reluctance or combativeness. In fact, when there are multiple hypotheses which are not anyone’s “personal property” and when there are crucial experiments to test them, the daily life in the laboratory takes on an interest and excitement it never had, and the students can hardly wait to get to work to see how the detective story will come out.” Of course this approach requires that researchers become experts in each theoretical model, and have the skill and expertise required to test all different hypotheses.

## 16.3 Conclusion

As Reif (1961) observed: “The work situation of the scientist is not just a quiet haven for scholarly activity, ideally suited to those of introverted temperament. The pure scientist, like the businessman or lawyer, works in a social setting, and like them, he is subject to appreciable social and competitive pressures.”

It has been widely recognized that science is a human endeavor. Scientists have motivations and desires that might bias the claims they make. At the same time, these motivations and desires might make individuals stick to a hypothesis long enough to make a new discovery, where most other researchers would have already given up on the idea. There are certain practices in science on an institutional level and an individual level that can be used to prevent motivations and desires from leading us astray. These human factors are part of science, and we need to design science in such a way that we achieve efficient and reliable knowledge generation. It's important to be aware of the role confirmation bias plays in science, and how you can use some

of the practices described in this chapter to prevent confirmation bias from fooling yourself. It is worth keeping in mind the warning by Johann Wolfgang von Goethe from 1792:

Thus we can never be too careful in our efforts to avoid drawing hasty conclusions from experiments or using them directly as proof to bear out some theory. For here at this pass, this transition from empirical evidence to judgment, cognition to application, all the inner enemies of man lie in wait: imagination, which sweeps him away on its wings before he knows his feet have left the ground; impatience; haste; self-satisfaction; rigidity; formalistic thought; prejudice; ease; frivolity; fickleness—this whole throng and its retinue. Here they lie in ambush and surprise not only the active observer but also the contemplative one who appears safe from all passion.

To explore the topics in this chapter further, you can listen to the HPS podcast episode on [Collective Objectivity](#) with Fiona Fidler, or the Nullius in Verba podcast episode on [Confirmation Bias](#) and [Skepticism](#). You can also read the book [Nobody's Fool](#) by Daniel Simons and Christopher Chabris, and watch [this video](#) on confirmation bias.

## 17 Replication Studies

In 2015 a team of 270 authors published the results of a research project where they replicated 100 studies (Open Science Collaboration, 2015). The original studies had all been published in three psychology journals in the year 2008. The authors of the replication project selected the last study of papers that could feasibly be replicated, performed a study with high power to detect the observed effect size, and attempted to design the best possible replication study. They stayed close to the original study where possible, but deviated where this was deemed necessary. For the original studies published in 2008 97 of the 100 studies were interpreted as significant. Given an estimated 92% power for the effect sizes observed in the original studies  $97 \times 0.92 = 89$  of the replication studies could be expected to observe a significant effect, if the effects in the original studies were at least as large as reported. Yet, only 35 out of the 97 original studies that were significant replicated, for a replication rate of 36%. This result was a surprise for most researchers, and led to the realization that it is much more difficult to replicate findings than one might intuitively think. This result solidified the idea of a *replication crisis*, a sudden loss of confidence in the reliability of published results, which led to confusion and uncertainty about how scientists worked. Since 2015 the field of **metascience** has emerged to use empirical methods to study science itself and identify some of the causes of low replicability rates, and develop possible solutions to increase it.

At the same time, the authors of the replication project acknowledged that a single replication of 100 studies is just the starting point of trying to understand why it was so difficult to replicate findings in the literature. They wrote in their conclusion:

After this intensive effort to reproduce a sample of published psychological findings, how many of the effects have we established are true? Zero. And how many of the effects have we established are false? Zero. Is this a limitation of the project design? No. It is the reality of doing science, even if it is not appreciated in daily practice. Humans desire certainty, and science infrequently provides it. As much as we might wish it to be otherwise, a single study almost never provides definitive resolution for or against an effect and its explanation. The original studies examined here offered tentative evidence; the replications we conducted offered additional, confirmatory evidence. In some cases, the replications increase confidence in the reliability of the original results; in other cases, the replications suggest that more investigation is needed to establish the validity of the original findings. Scientific progress is a cumulative process of uncertainty reduction that can only succeed if science itself remains the greatest skeptic of its explanatory claims.

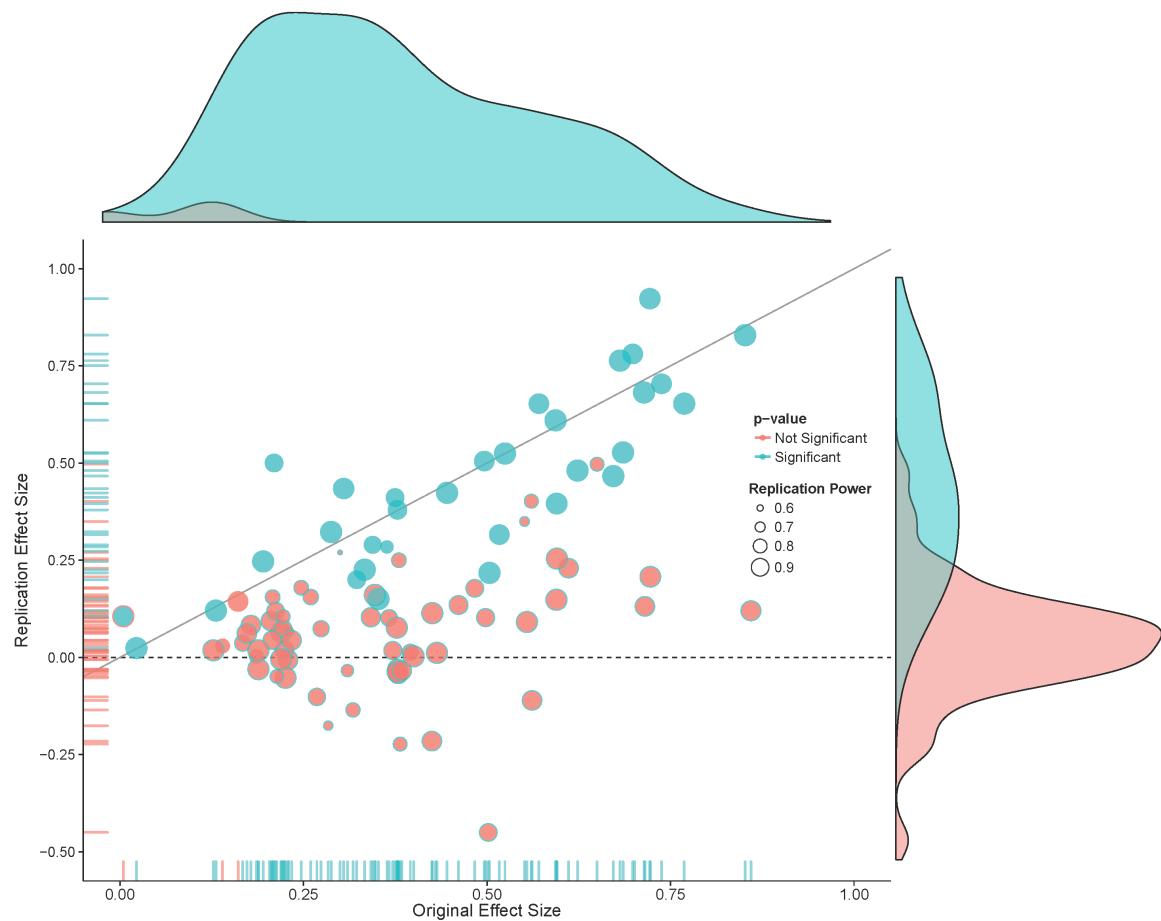


Figure 17.1: Results of the Reproducibility Project:Psychology, with replications in green, and non-replications (based on  $p > .05$ ) in red.

A replication study is an experiment where the methods and procedures in a previous study are repeated by collecting new data. Typically, the term **direct replication study** is used when the methods and measures are as similar to the earlier study as possible. In a **conceptual replication study** a researcher intentionally introduces differences with the original study with the aim to test the generalization of the effect, either because they aim to systematically explore the impact of this change, or because they are not able to use the same methods and procedures. It is important to distinguish replication, where new data is collected, from **reproducibility**, where the same data is used to reproduce the reported results. In reproducibility checks the goal is to examine the presence of errors in the analysis files. Confusingly, the large-scale collaborative research project where 100 studies in psychology were replicated, and which is often considered an important contributor to the **replication crisis**, was called the Reproducibility Project: Psychology (Open Science Collaboration, 2015). It should have been called the Replication Project: Psychology. Our apologies.

The goal of direct replication studies is first of all to identify Type 1 or Type 2 errors in the literature. In any original study with an alpha of 5% and a statistical power of 80% there is a probability of making an erroneous claim. Direct replication studies (especially those with low error rates) have the goal to identify these errors in the scientific literature, which is an important part of having a reliable knowledge base. It is especially important given that the scientific literature is biased, which increases the probability that a published claim is a Type 1 error. Schmidt (2009) also notes that claims can be erroneous because they were based on fraudulent data. Although direct replication studies can not identify fraud, they can point out erroneous claims due to fraud.

The second important goal of direct replications is to identify factors in the study that were deemed irrelevant, but were crucial in generating the observed data (Tunç & Tunç, 2023). For example, researchers might discover that an original and replication study yielded different results because the experimenter in one study treated the participants in a much more friendly manner than the experimenter in the other study. Such details are typically not reported in the method section, as they are deemed irrelevant, but direct replication studies might identify such factors. This highlights how replication studies are never identical in the social sciences. They are designed to be similar in all ways that a researcher thinks will matter. But not all factors deemed irrelevant will be irrelevant (and vice versa).

The goal of conceptual replication studies is to examine the generalizability of effects (Sidman, 1960). Researchers intentionally vary factors in the experiment to examine if this leads to variability in the results. Sometimes this variability is theoretically predicted, and sometimes researchers simply want to see what will happen. I would define a direct replication as a study where a researcher has the goal to not introduce variability in the effect size compared to the original study, while in a conceptual replication variability is introduced intentionally with the goal to test the generalizability of the effect. This distinction means that it is possible that what one researcher aims to be a direct replication is seen by a peer as a conceptual replication. For example, if a researcher sees no reason why an effect tested in Germany would not be identical in The Netherlands, they will consider it a direct replication. A peer

might believe there are theoretically relevant differences between the two countries that make this a conceptual replication, as they would interpret the study as intentionally introducing variability by not keeping an important factor constant. The only way to clarify which factors are deemed theoretically relevant is to write a **constraints on generalizability** statement in the discussion where researchers specify which contexts they theoretically expect the effect to replicate in, and where variability in the results would not be considered problematic for their original claim (Simons et al., 2017).

It should be clear that the goals of replication studies are rather modest. At the same time, they are essential for a well-functioning empirical science. They provide a tool to identify false positive or false negative results, and can reveal variability across contexts that might falsify theoretical predictions, or lead to the generation of new theories. Being able to systematically replicate and extend a basic effect is one of the most important ways in which scientists develop and test theories. Although there are other approaches to generating reliable knowledge, such as **triangulation** where the same theory is tested in different but complementary ways, in practice the vast majority of scientifically established claims are based on replicable effects.

Not all researchers agree that their science has inter-subjectively repeatable observations. In what is called the ‘crisis in social psychology’ Gergen (1973) argued social psychology was not a cumulative science:

It is the purpose of this paper to argue that social psychology is primarily an historical inquiry. Unlike the natural sciences, it deals with facts that are largely nonrepeatable and which fluctuate markedly over time. Principles of human interaction cannot readily be developed over time because the facts on which they are based do not generally remain stable. Knowledge cannot accumulate in the usual scientific sense because such knowledge does not generally transcend its historical boundaries.

The belief that basic claims in psychology are not repeatable events lead to **social constructivism**. This approach did not become particularly popular, but it is useful to know of its existence. It is a fact that human behavior can change over time. It is also true that many psychological mechanisms have enabled accurate predictions for more than a century, and this is unlikely to change. Still, some researchers might believe they are studying unrepeatable events, and if so, they can state why they believe this to be the case. These researchers give up the aim to build theories upon which generalizable predictions can be made, and they will have to make a different argument for the value of their research. Luckily, one does not need to be a social constructivist to acknowledge that the world changes. We should not expect all direct replications to yield the same result. For example, in the classic ‘foot-in-the-door’ effect study, Freedman and Fraser (1966) first called residents in a local community over the phone to ask them to answer some questions (the small request). If participants agreed, they were asked a larger request, which consisted of “five or six men from our staff coming into your home some morning for about 2 hours to enumerate and classify all the household products that you have. They will have to have full freedom in your house to go through the cupboards and

storage places.” The idea that anyone would nowadays agree to such a request when called by a stranger over the telephone seems highly improbable. Repeating this procedure will not lead to more than 50% of respondents agreeing to this request. But the theories we build should be able to account for why some findings no longer replicate by specifying necessary conditions for the effect to be observed. This is the role of theory, and if researchers have good theories, they should be able to continue to make predictions, even if some aspects of the world change. As De Groot (1969, p. 89) writes: “If one knows something to be true, he is in a position to predict; where prediction is impossible there is no knowledge”.

## 17.1 Why replication studies are important

Over the last half century, researchers have repeatedly observed that replication studies were rarely performed or published. In an editorial in the Journal of Personality and Social Psychology, Greenwald (Greenwald, 1976) writes: “There may be a crisis in personality and social psychology, associated with the difficulty often experienced by researchers in attempting to replicate published work. A precise statement of the magnitude of this problem cannot be made, since most failures to replicate do not receive public report”. A similar concern about the replicability of findings is expressed by Epstein (Epstein, 1980, p. 790): “Not only are experimental findings often difficult to replicate when there are the slightest alterations in conditions, but even attempts at exact replication frequently fail.” Neher (1967, p. 262) concludes: “The general adoption of independent replication as a requirement for acceptance of findings in the behavioral sciences will require the efforts of investigators, readers, and publishing editors alike. It seems clear that such a policy is both long overdue and crucial to the development of a sound body of knowledge concerning human behavior.” Lubin (Lubin, 1957) suggests that, where relevant, manuscripts that demonstrate the replicability of findings should receive a higher publication priority. N. C. Smith (1970, p. 974) notes how replication studies are neglected: “The review of the literature on replication and cross-validation research has revealed that psychologists in both research”disciplines” have tended to ignore replication research. Thus, one cannot help but wonder what the impact might be if every investigator repeated the study which he believed to be his most significant contribution to the field.” One problem in the past was the difficulty of describing the methods and analyses in sufficient detail to allow others to repeat the study as closely as possible (Mack, 1951). For example, Pereboom (1971, p. 442) writes: “Related to the above is the common difficulty of communicating all important details of a psychological experiment to one’s audience. [...] Investigators attempting to replicate the work of others are painfully aware of these informational gaps.” Open science practices, such as sharing **computationally reproducible** code and materials, are an important way to solve this problem, and a lot of progress has been made over the last decade to address these problems.

Many researchers have suggested that performing replication studies should be common practice. Lykken (1968, p. 159) writes: “Ideally, all experiments would be replicated before publication but this goal is impractical”. Loevinger (1968, p. 455) makes a similar point:

“Most studies should be replicated prior to publication. This recommendation is particularly pertinent in cases where the results are in the predicted direction, but not significant, or barely so, or only by one-tailed tests”. Samelson (1980, p. 623) notes in the specific context of Watson’s ‘Little Albert’ study: “Beyond this apparent failure of internal criticism of the data is another one that is even less debatable: the clear neglect of a cardinal rule of scientific method, that is, replication”.

The idea that replication is a ‘cardinal rule’ or ‘cornerstone’ of the scientific method follows directly from a methodological falsificationist philosophy of science. Popper (2002) discusses how we increase our confidence in theories that make predictions that withstand attempts to falsify the theory. To be able to falsify predictions, predictions need to rule out certain observable data patterns. For example, if our theory predicts that people will be faster at naming the colour of words when their meaning matches the colour (e.g., “blue” written in blue instead of “blue” written in red), the observation that people are not faster (or even slower) would falsify our prediction. A problem is that given variability in observed data any possible data pattern will occur, exactly as often as dictated by chance. This means that in the long run, just based on chance, a study will show people are *slower* at naming the colour of words when their meaning matches the colour. This fluke will be observed by chance, even though our theory is correct. Popper realized this was a problem for his account of falsification, because it means that “probability statements will not be falsifiable”. After all, if all possible data patterns have a non-zero probability, even if they are extremely rare, they are not logically ruled out. This would make falsification impossible if we demanded that science works according to perfectly formal logical rules.

The solution is to admit that science does not work following perfectly formal rules. And yet, as Popper acknowledges, science still works. Therefore, instead of abandoning the idea of falsification, Popper proposes a more pragmatic approach to falsification. He writes: “It is fairly clear that this ‘practical falsification’ can be obtained only through a methodological decision to regard highly improbable events as ruled out — as prohibited.” The logical follow-up question is then “Where are we to draw the line? Where does this ‘high improbability’ begin?” Popper argues that even if any low probability event *can* occur, *they are not reproducible at will*. Any single study can reveal any possible effect, but a prediction should be considered falsified if we fail to see “the predictable and reproducible occurrence of systematic deviations”. This is why replication is considered a ‘cardinal rule’ in methodological falsificationism: When observations are probabilistic, only the replicable occurrence of low probability events can be taken as the falsification of a prediction. A single  $p < 0.05$  is not considered sufficient; only if close replication studies repeatedly observe a low probability event does Popper allow us to ‘practically falsify’ probabilistic predictions.

## 17.2 Direct versus conceptual replications

As Schmidt (2009) writes “There is no such thing as an exact replication.” However, it is possible to 1) repeat an experiment where a researcher stays as closely as possible to the original study, 2) repeat an experiment where there it is likely there is some variation in factors that are deemed irrelevant, and 3) knowingly vary aspects of a study design. Popper agrees: “We can never repeat an experiment precisely — all we can do is to keep certain conditions constant, within certain limits.” One of the first extensive treatments of replication comes from Sidman (1960). He distinguishes direct replications from **systematic replications** (which I refer to here as conceptual replications). Sidman writes:

Where direct replication helps to establish generality of a phenomenon among the members of a species, systematic replication can accomplish this and, at the same time, extend its generality over a wide range of different situations.

If the same result is observed when systematically varying auxiliary assumptions we build confidence in the general nature of the finding, and therefore, in the finding itself. The more robust an effect is to factors that are deemed irrelevant, the less likely it is that the effect is caused by a confound introduced by one of these factors. If a prediction is confirmed across time, locations, in different samples of participants, by different experimenters, and with different measures of the same variable, then the likelihood of confounds underlying all these interrelated effects decreases. If conceptual replications are successful, they yield more information than a direct replication, because a conceptual replication generalizes the finding beyond the original context. However, this benefit is only present if the conceptual replication is successful. Sidman warns:

But this procedure is a gamble. If systematic replication fails, the original experiment will still have to be redone, else there is no way of determining whether the failure to replicate stemmed from the introduction of new variables in the second experiment, or whether the control of relevant factors was inadequate in the first one.

Sometimes there is no need to choose between a direct replication and a conceptual replication, as both can be performed. Researchers can perform **replication and extension studies** where an original study is replicated, but additional conditions are added that test a novel hypotheses. Sidman (1960) refers to this as the **baseline technique** where an original effect is always part of the experimental design, and variations are tested against the baseline effect. Replication and extension studies are one of the best ways to build cumulative knowledge and develop strong scientific theories (Bonett, 2012).

It is often difficult to reach agreement on whether a replication study is a direct replication or a conceptual replication, because researchers can disagree about whether changes are theoretically relevant or not. Some authors have chosen to resolve this difficulty by defining a replication study as “a study for which any outcome would be considered diagnostic evidence

about a claim from prior research” (Nosek & Errington, 2020). However, this definition is too broad to be useful in practice. It has limited use when theoretical predictions are vague (which regrettably holds for most research lines in psychology), and it does not sufficiently acknowledge that it is important to specify falsifiable auxiliary hypotheses. The specification of falsifiable auxiliary hypothesis in replication studies lies at the core of the **Systematic Replications Framework** (Tunç & Tunç, 2023). The idea is that researchers should specify the auxiliary hypotheses that are assumed to be relevant. In principle the number of auxiliary hypotheses is infinite, but as Uygun-Tunç and Tunç (2023) write: “it is usually assumed that the exact color of the lab walls, the elevation of the lab above the sea level, the exact design of the chairs used by the subjects, the humidity of the room that the study takes place or many other minute details do not significantly influence the study outcomes.” These factors are relegated to the **ceteris paribus clause**, meaning that any differences on these factors are not considered relevant, and for all purposes studies can be treated as ‘all equal’ - even if the color of the walls differs between an original and a replication study. No replication study is exactly the same as the original study (Schmidt, 2009), and some differences in auxiliary hypotheses are meaningfully different. The challenge is to identify which auxiliary hypotheses explain failures to replicate an original study.

A direct replication study that yields no statistically significant effect can have three interpretations (Schmidt, 2009; Tunç & Tunç, 2023). First, it is possible that the replication study yielded a Type 2 error. Second, it is possible that the original study was a Type 1 error. Third, some of the auxiliary assumptions that have been relegated to the ceteris paribus clause actually matter more than researchers might have thought. To resolve disagreements between the results of original and replication studies researchers should perform a set of studies that systematically varies those auxiliary hypotheses that are most crucial for a theoretical viewpoint. Resolving inconsistencies in science is an effortful process that can be facilitated by engaging in an **adversarial collaboration**, where two teams join forces to resolve inconsistencies (Mellers et al., 2001). Opposing teams can point out the most crucial auxiliary hypotheses to test, and severely test different theories.

### 17.3 Analyzing Replication Studies.

There are multiple ways to statistically analyze a replication study (S. F. Anderson & Maxwell, 2016). The most straightforward statistical approach is also the least common: Testing whether the effect sizes in the original and replication studies are statistically different from each other. Two effect sizes from an independent *t*-test can be tested against each other using:

$$Z_{Diff} = \frac{\delta_1 - \delta_2}{\sqrt{V_{\delta_1} + V_{\delta_2}}}$$

where the difference between the two Cohen's  $d$  effect sizes is divided by the standard error of the difference score, based on the square root of the combined variances of the effect sizes (Borenstein, 2009, formula 19.6 and 19.7). This formula provides a hint why researchers rarely test for differences between effect sizes. The standard error of the difference score is based on the variance in the original and replication study. If the original study has a small sample size, the variance will be large, and a test of the difference between effect sizes can have very low power.

Let's take a look at an original non-preregistered study which observed an effect size of  $d = 0.6$  in an independent  $t$ -test with 30 participants in each group. A preregistered replication study is performed which observed an effect size of 0 with 150 participants in each group. Testing the two effect sizes against each other can be achieved in three ways that are in principle identical (although the exact implementation, such as computing Cohen's  $d$  or Hedges'  $g$ , might yield slightly different  $p$ -values). The first is to compute the corresponding  $p$ -value for the  $Z$ -test in the formula above. The other two ways are implemented in the `metafor` package and consist of a heterogeneity analysis and a moderator analysis. These two approaches are mathematically identical. In the heterogeneity analysis we test if the variability in effect sizes (even though there are only two) is greater than expected based only on random variation.

```

d1 <- escalc(n1i = 30,
              n2i = 30,
              di = 0.6,
              measure = "SMD")
d2 <- escalc(n1i = 100,
              n2i = 100,
              di = 0.0,
              measure = "SMD")
metadata <- data.frame(yi = c(d1$yi, d2$yi),
                        vi = c(d1$vi, d2$vi),
                        study = c("original", "replication"))

# Test based on heterogeneity analysis
res_h <- rma(yi,
               vi,
               data = metadata,
               method = "FE")
res_h

# The moderator test would be: rma(yi, vi, mods = ~study, method = "FE", data = metadata, dig

```

Fixed-Effects Model ( $k = 2$ )

```
I^2 (total heterogeneity / total variability): 74.45%
H^2 (total variability / sampling variability): 3.91
```

```
Test for Heterogeneity:
Q(df = 1) = 3.9146, p-val = 0.0479
```

Model Results:

| estimate | se     | zval   | pval   | ci.lb   | ci.ub  |
|----------|--------|--------|--------|---------|--------|
| 0.1322   | 0.1246 | 1.0607 | 0.2888 | -0.1121 | 0.3765 |

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The result yields  $p$ -value of 0.048, which is statistically significant, and therefore allows us to statistically conclude that the two effect sizes differ from each other. We can also perform this test as a moderator analysis.

```
res_mod <- rma(yi,
                 vi,
                 mods = ~study,
                 method = "FE",
                 data = metadata,
                 digits = 3)
```

This result yields the same  $p$ -value of 0.048. This same test was recently repackaged by Spence and Stanley (2024) as a prediction interval, but this approach is just a test of the difference between effect sizes.

It is worth pointing out that in the example above the difference between the effect sizes was large, the replication study had a much larger sample size than the original study, but the differences was only just statistically significant. As we see in Figure 17.2 the original study had large uncertainty, which as explained above is part of the variance of the estimate of the difference between effect sizes. If the replication study had yielded an effect size that was even slightly in the positive direction (which should happen 50% of the time if the true effect size is 0) the difference would no longer have been statistically significant. In general, power is low when original studies have small sample sizes. Despite this limitation, directly testing the difference between effect sizes is the most intuitive and coherent approach to claiming that a study failed to replicate.

It is also possible to test the difference between correlations, where the formula is:

$$Z = \frac{\ln(1 + r) - \ln(1 - r)}{2}$$

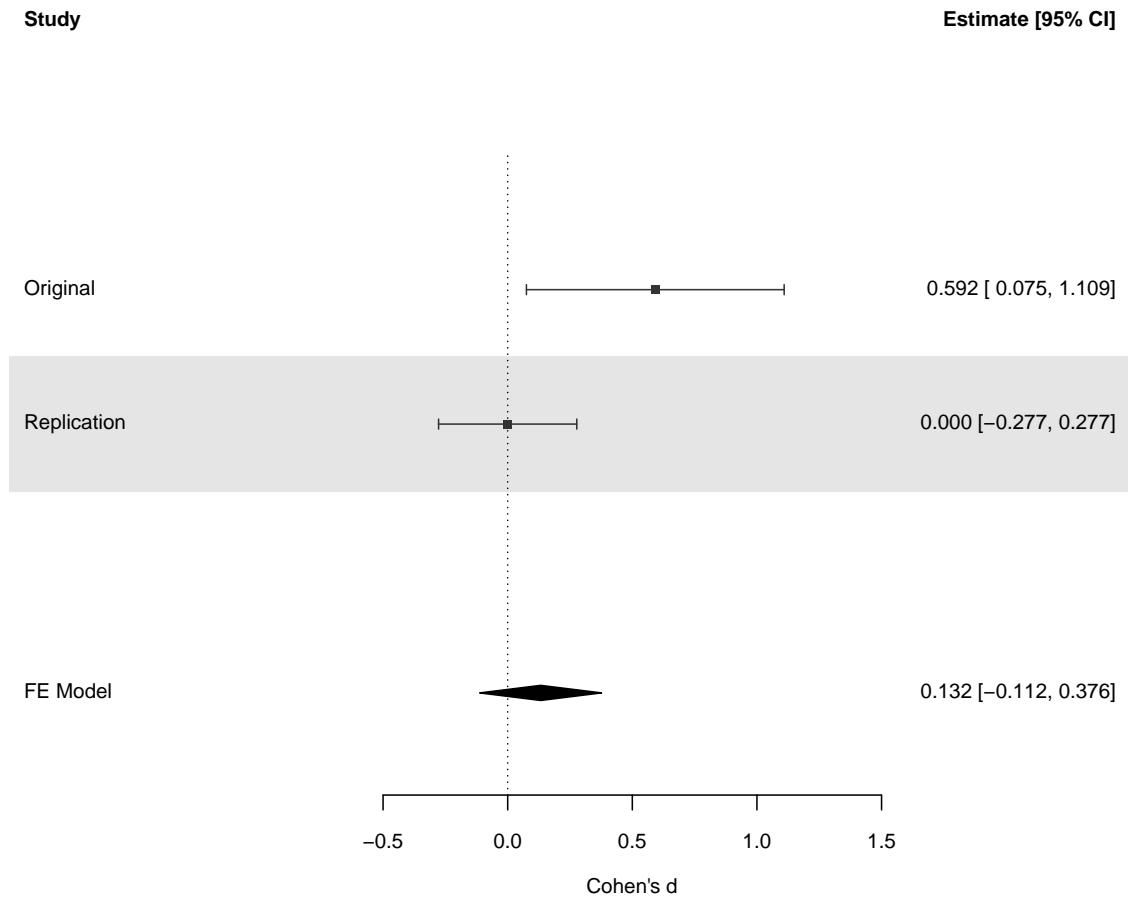


Figure 17.2: Forest plot for an original study ( $N = 60$ ,  $d = 0.6$ ) and a replication study ( $N = 200$ ,  $d = 0$ ).

This test can for example be performed in the cocor package in R:

```
library(cocor)
cocor.indep.groups(n1 = 30, r1.jk = .4, n2 = 200, r2.hm = .01)
```

Results of a comparison of two correlations based on independent groups

```
Comparison between r1.jk = 0.4 and r2.hm = 0.01
Difference: r1.jk - r2.hm = 0.39
Group sizes: n1 = 30, n2 = 200
Null hypothesis: r1.jk is equal to r2.hm
Alternative hypothesis: r1.jk is not equal to r2.hm (two-sided)
Alpha: 0.05

fisher1925: Fisher's z (1925)
z = 2.0157, p-value = 0.0438
Null hypothesis rejected

zou2007: Zou's (2007) confidence interval
95% confidence interval for r1.jk - r2.hm: 0.0102 0.6888
Null hypothesis rejected (Interval does not include 0)
```

We see that the difference between an original study with 30 participants that observed an effect of  $r = 0.4$  and a replication study that observed an effect of  $r = 0.01$  is just statistically significant. We can use the pwrss package (M. Bulus & Polat, 2023) to examine the sample size we would need to achieve sufficient power for such difference between correlations with 90% power.

```
library(pwrss)
pwrss.z.2corrs(r1 = 0.4, r2 = 0.01,
                power = .90, alpha = 0.05,
                alternative = "not equal")
```

```
Difference between Two Correlations
(Independent Samples z Test)
H0: r1 = r2
HA: r1 != r2
-----
Statistical power = 0.9
n1 = 126
```

```

n2 = 126
-----
Alternative = "not equal"
Non-centrality parameter = 3.242
Type I error rate = 0.05
Type II error rate = 0.1

```

The same calculation can be performed in G\*Power.

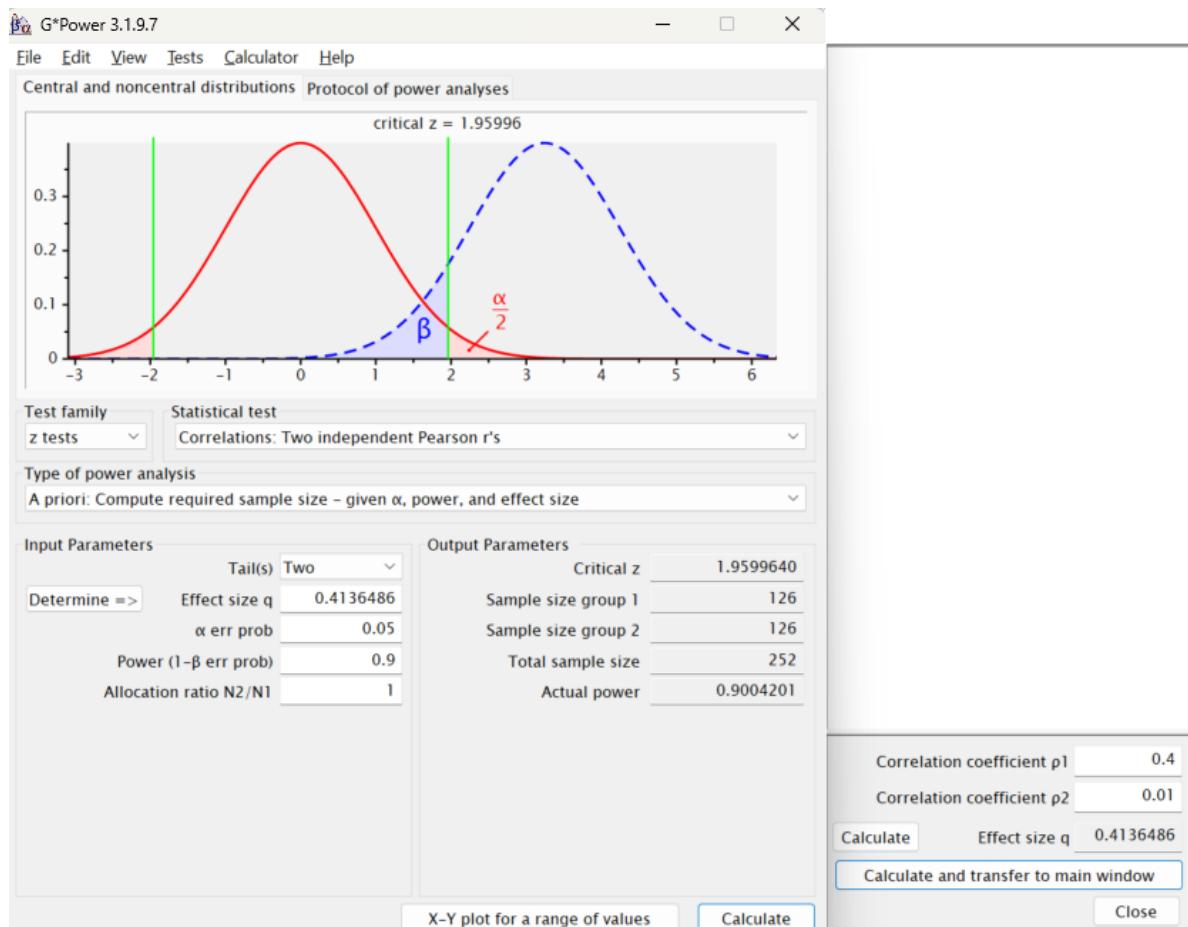


Figure 17.3: Power analysis for the difference between two independent correlations in G\*Power.

Although a statistical difference between effect sizes is one coherent approach to decide whether a study has been replicated, researchers are sometimes interested in a different question: Was there a significant result in the replication study? In this approach to analyzing replication studies there is no direct comparison with the effect observed in the original study. The

question is therefore not so much ‘is the original effect replicated?’ but ‘if we repeat the original study is a statistically significant effect observed?’. In other words, we are not testing whether an effect has been replicated, but whether a predicted effect has been observed. Another way of saying this is that we are not asking whether the observed effect replicated, but whether the original ordinal claim of the presence of a non-zero effect is replicated. In the example in Figure 17.2 the replication study has an effect size of 0, so there is no statistically significant effect, and the original effect did not replicate (in the sense that repeated the procedure did not yield a significant result).

Let’s take a step back, and consider which statistical test best reflects the question ‘did this study replicate’. On the one hand it seems reasonable to consider an effect ‘not replicated’ if there is a statistically significant difference between the effect sizes. However, this can mean that a non-significant effect in the replication studies leads to a ‘replication’ just because the effect size estimate is not statistically smaller than the original effect size. On the other hand, it seems reasonable to consider an effect replicated if it is statistically significant in the replication study. These two approaches can lead to a conflict, however. Some statistically significant effects are statistically smaller than the original study. Should these be considered ‘replicated’ or not? We might want to combine both statistical tests, and consider a study a replication if the effect is both statistically different from 0 (i.e.,  $p < .05$  in a traditional significance test), and the difference in effect sizes is not statistically different from 0 (i.e.,  $p > .05$  for a test of heterogeneity in a meta-analysis of both effect sizes). We can perform both tests, and only consider a finding replicated if both conditions are met. Logically, a finding should then be considered as a non-replication if the opposite is true (i.e.,  $p > .05$  for significance test of the replication study, and  $p < .05$  for the test of heterogeneity).

However, due to the low power of the test for a difference between the effect sizes in the original and replication study, it is not easy to meet the bar of an informative result. In practice, many replication studies that do not yield a statistically significant result will also not show a statistically significant difference as part of the test of heterogeneity. In a preprint (that we never bothered to resubmit after we received positive reviews) Richard Morey and I (Richard D. Morey & Lakens, 2016) summarized this problem in our title “Why most of psychology is statistically unfalsifiable”. If we want to have an informative test for replication following the criteria outlined above, most tests will be uninformative, and lead to inconclusive results.

This problem becomes even bigger if we admit that we often consider some effects too small to matter. S. F. Anderson & Maxwell (2016) mention an additional approach to analyzing replication studies where we do not just test the difference in effect sizes between an original and replication study, but combine this with an equivalence test to examine if the difference - if any - is too small to matter. With very large sample sizes tiny differences between studies can become statistically significant, which would result in the conclusion that a study did not replicate, even if the difference is considered too small to matter.

Anderson and Maxwell also point out the importance of incorporating a [smallest effect size of interest](#) when interpreting whether a replication study yields a null effect. For example,

McCarthy et al. (2018) replicated a hostility priming study originally reported to have an effect size of  $d = 3.01$ . In a multi-lab replication they observed an effect very close to zero:  $d = 0.06$ , 95% CI = [0.01, 0.12]. This effect was statistically different from the original, but statistically significant. McCarthy and colleagues argue this effect, even if significant, is too small to matter, and conclude “Our results suggest that the procedures we used in this replication study are unlikely to produce an assimilative priming effect that researchers could practically and routinely detect. Indeed, to detect priming effects as small as the 0.08-scale-unit difference we observed (which works out to approximately  $d = 0.06$ , 95% CI = [0.01, 0.12]), a study would need 4,362 participants in each priming condition to have 80% power with an alpha set to .05.”

Should we consider the replication study with  $d = 0.06$  a successful replication, just because the effect is significant in a very large sample? As explained in the chapter on equivalence testing, effects very close to zero are often considered equivalent to the absence of an effect because the effect is practically insignificant, theoretically smaller than expected, or impossible to reliably investigate given the resources that researchers have available. As McCarthy et al. (2018) argue, the very small effect they observed is no longer practically feasible to study, and psychologists would therefore consider such effects equivalent to a null result. It is recommended to specify a smallest effect of interest in replication studies, where possible together with the original author (see the example by Richard D. Morey et al. (2021) below) and evaluate replication studies with an equivalence test against the smallest effect size of interest.

As noted above, although ideally we directly test the difference between effect sizes, the uncertainty in the original effect size estimate when sample sizes are small can make lead to a test with very low power. As we can never increase the sample size of original studies, and we also can not do science if we can not conclude claims in the literature can not be replicated, we will need to be pragmatic and develop alternative statistical procedures to claim an effect can not be replicated. As the large uncertainty in the original effect sizes are the main cause of the problem, a logical solution is to move to statistical approaches where this uncertainty is not part of the test. An intuitive approach would be to test if the effect size in the replication study is statistically smaller than the effect in the original study. In the example above we would test whether the 95% confidence interval around the effect size in the replication study excludes the effect in the original study (i.e., 0.6, or 0.592 when converted to Hedges'  $g$ ). In essence this is a two-sided test against the original effect size. This is the equivalent of changing a independent  $t$ -test where two groups are compared to a one-sample  $t$ -test where the 95% confidence interval in one group is compared against the observed mean in the other group, ignoring uncertainty in this other group. We do not accept this approach when we compare two means, and we should not accept this approach when we compare two standardized mean differences from an original and replication study. The problem is that this test will too easily reject the null-hypothesis that the two effects are equal, because it ignores the variability in one of the two effect sizes.

An alternative approach is to test against a more conservative effect size, and consider an

original finding not replicated if this conservative effect size can be rejected. In practice researchers are primarily interested in the question whether the effect size in the replication study is statistically smaller than the effect size in the original study. This question is answered by an *inferiority test*, which is statistically significant if the 90% confidence interval around the effect size in the replication study does not contain the conservative effect size estimate. One implementation of an inferiority test is the *small telescopes* approach (Simonsohn, 2015). In the small telescopes approach a test is performed against the effect the original study had 33% power to detect. In the example in Figure 17.2 the original study had 33% power to detect and effect of  $d = 0.4$ .

```
pwr::pwr.t.test(
  n = 30,
  sig.level = 0.05,
  power = 0.33,
  type = "two.sample",
  alternative = "two.sided"
)
```

Two-sample t test power calculation

```
n = 30
d = 0.3988825
sig.level = 0.05
power = 0.33
alternative = two.sided
```

NOTE: n is number in \*each\* group

The 95% confidence interval of the replication study shows effects larger than 0.277 can be rejected, so a 90% confidence interval would be able to reject even smaller effect sizes. Therefore, the small telescopes approach would allow researchers to conclude that the original effect could not be replicated. The small telescopes is popular especially when replicating small original studies, as the smaller the sample size, the larger the effect size the study had 33% power to detect, and this the easier it is to reject in a replication study. This is the opposite of what happens if we want to test the two effect sizes against each other, where smaller original studies reduce the power of the test.

A fair point of criticism of the small telescopes approach is that the convention to test against an effect the original study had 33% power to detect is completely arbitrary. Ideally researchers would specify the *smallest effect size of interest* and perform an inferiority test against the smallest effect that would actually matter. Specifying a smallest effect size of interest is

difficult, however. In some cases researchers involved in the original study might specify the smallest effect they care about. For example, in a multi-lab replication study of the action-sentence compatibility effect the researchers who published the original study stated that they considered an effect of 50 milliseconds or less theoretically negligible (Richard D. Morey et al., 2021). The large sample size allowed the authors to conclusively reject the presence of effects as large or larger than were considered theoretically negligible. Another approach is to set the smallest effect size of interest based on theoretical predictions. Alternatively, researchers might set the smallest effect size of interest to be larger than the **crud factor**, which is the effect size in a literature due to theoretically uninteresting systematic noise (Orben & Lakens, 2020). For example, C. J. Ferguson & Heene (2021) suggests a smallest effect size of interest of  $r = 0.1$  (or  $d = 0.2$ ) because such effect sizes can be observed for nonsensical variables, and are likely purely driven by (methodological) confounds.

Another approach that has been proposed is to combine the effect size of the original study and the replication study in a meta-analysis, and test whether the meta-analytic effect is statistically different from zero. This approach is interesting if there is no bias, but when there is bias the weakness of this approach outweighs its usefulness. As publication bias and selection bias **inflated effect sizes** the meta-analytic effect size will also be inflated. This in turn inflates the probability that the meta-analytic effect size is statistically significant, when there is no effect.

## 17.4 Replication studies or lower alpha levels?

Statistically minded researchers sometimes remark that there is no difference in the Type 1 error probability when a single study with a lower alpha level is used to test a hypothesis (say  $0.05 \times 0.05 = 0.0025$ ) compared to when the same hypothesis is tested in two studies, each at an alpha level of 0.05. This is correct, but in practice it is not possible to replace the function of replication studies to decrease the Type 1 error probability with directly lowering alpha levels in single studies. Lowering the alpha level to a desired Type 1 error probability would require that scientists 1) can perfectly predict how important a claim will be in the future, and 2) are able to reach consensus on the Type 1 error rate they collectively find acceptable for each claim. Neither is true in practice. First, it is possible that a claim becomes increasingly important in a research area, for example because a large number of follow-up studies cite the study and assume the claim is true. This increase in importance might convince the entire scientific community that it is worthwhile if the Type 1 error probability is reduced by performing a replication study, as the importance of the original claim has made a Type 1 error more costly (Isager et al., 2023). For claims no one builds on, no one will care about reducing the Type 1 error probability. How important a claim will be for follow-up research can not be predicted in advance.

The second reason to reduce the probability of Type 1 errors through replications is that there are individual differences between researchers in when they believe a finding is satisfactorily

demonstrated. As Popper writes: “Every test of a theory, whether resulting in its corroboration or falsification, must stop at some basic statement or other which we *decide to accept*.” “This procedure has no natural end. Thus if the test is to lead us anywhere, nothing remains but to stop at some point or other and say that we are satisfied, for the time being.” Different researchers will have different thresholds for how satisfied they are with the error probability associated with a claim. Some researchers are happy to accept a claim when the Type 1 error probability is 10%, while more skeptical researchers would like to see it reduced to 0.1% before building on a finding. Using a lower alpha level for a single study does not give scientists the flexibility to lower Type 1 error probabilities as the need arises, while performing replication studies does. It is a good idea to think carefully about the desired Type 1 error rate for studies, and if Type 1 errors are costly, researchers can decide to use a lower alpha level than the default level of 0.05 (Maier & Lakens, 2022). But replication studies will remain necessary in practice to further reduce the probability of a Type 1 error as claims increase in importance, or for researchers who are more skeptical about a claim.

There is another reason why it is more beneficial to use independent replication studies to reduce the error rate, compared to performing studies with a lower alpha level. If others are able to independently replicate the same effect, it becomes less likely that the original finding was due to systematic error. Systematic errors do not average out to zero in the long run (as random errors do). There can be many sources of systematic error in a study. One source is the measures that are used. For example, if a researcher uses a weighing scale that is limited to a maximum weight of 150 kilo, they might fail to identify an increase in weight, while different researchers who use a weighting scale with a higher maximum weight will identify the difference. An experimenter might be another source of systematic error. An observed effect might not be due to a manipulation, but due to the way the experimenter treats participants in different conditions. Other experimenters who repeat the manipulation, but do not show the same experimenter bias, would observe different results.

When a researcher repeats their own experiment this is referred to as a **self-replication**, while in an **independent replication** other researchers repeat the experiment. As explained above, self-replication can reduce the Type 1 error rate of a claim, while independent replication can in addition reduce the probability of a claim being caused by a systematic error. Both self-replication and independent replication are useful in science. For example, research collaborations such as the Large Hadron Collider at CERN prefer to not only replicate studies with the same detector, but also replicate studies across different detectors. The Large Hadron Collider has four detectors (ATLAS, CMS, ALICE, and LHCb). Experiments can be self-replicated in the same detector by collecting more data (referred to by physicists as a ‘replica’), but they can also be replicated in a different detector. As Junk and Lyons (2020) note, in self-replications: “The statistical variations are expected to be different in the replica and the original, but the sources of systematic errors are expected to be unchanged.” One way to examine systematic errors is to perform the same experiment in different detectors. The detectors at CERN are not exact copies of each other, and by performing studies in two different detectors, the research collaboration increases its confidence in the reliability of the

conclusions if a replication study yields the same observations, despite minor differences in the experimental set-up.

The value of an independent replication is not only the reduction in the probability of a Type 1 error, but at the same time the reduction in concerns about systematic error influencing the conclusion (Neher, 1967). As already highlighted by Mack (1951): “Indeed, the introduction of different techniques gives the replication the additional advantage of serving as a check on the validity of the original research”. Similarly, Lubin (1957) notes: “Our confidence in a study will be a positive monotonic function of the extent to which replication designs are used which vary these supposedly irrelevant factors”. To conclude, although both self-replications (a replication by the same researchers) and independent replication (a replication by different researchers) reduce the probability of a false positive claim, independent replications have the added benefit of testing the generalizability of the findings across factors that are deemed irrelevant to observe the effect.

From a purely statistical level is it interesting to ask whether it is less costly in terms of the required sample size to perform a single study at an alpha level of 0.0025, or two studies at an alpha level of 0.05. If the sample size required to achieve a desired power is much lower for a single test at an alpha of 0.0025 this would be a reason to perform larger single studies to achieve a low Type 1 error rate instead of performing two smaller studies at an alpha of 0.05. For an independent *t*-test the more efficient approach depends on the power of the test and whether the test is two-sided or one-sided. In Figure 17.4 a ratio of 1 means the sample size required to reach 70% to 95% power for an alpha level of 0.0025 is identical to two single studies with an alpha level of 0.05, and the two approaches would be equally efficient for a two-sided independent *t*-test. We see the ratio is below 1 for studies with high power. For example, the total sample size required to detect an effect of  $d = 0.5$  with an alpha level of 0.0025 and with 80% power is 244. With an alpha of 0.05 the total required sample size is 128 in each study, which makes the ratio  $244/(128+128) = 0.95$ , and the number of observations saved is  $(2*128)-244 = 12$ .

However, there are strong arguments in favor of [directional tests](#) in general, and especially for replication studies. When we examine the same scenario for a one-sided test the ratios change, and one larger study at 0.0025 is slightly more efficient for smaller effect sizes when the desired statistical power is 90%, but for 80% it is more efficient to perform two tests at 0.05. For example, the total sample size required to detect an effect of  $d = 0.5$  with an alpha level of 0.0025 and with 80% power is 218. With an alpha of 0.05 the total required sample size is 102, which makes the ratio  $218/(102+102) = 1.07$ , and the number of observations saved (now when performing two studies) is  $(2*102)-218 = 14$ .

The statistical power in these calculations is identical for both scenarios, but one needs to reflect on how the two studies at an alpha of 0.05 will be analyzed. If the two studies are direct replications, a fixed effect meta-analysis of both effect sizes has the same power as a single study with twice the sample size, and therefore a meta-analysis at an alpha level of 0.0025 would be identical to a single study with an alpha of 0.0025. Altogether, the choice of whether to perform two studies at an alpha of 0.05 or a single study at 0.0025 depends

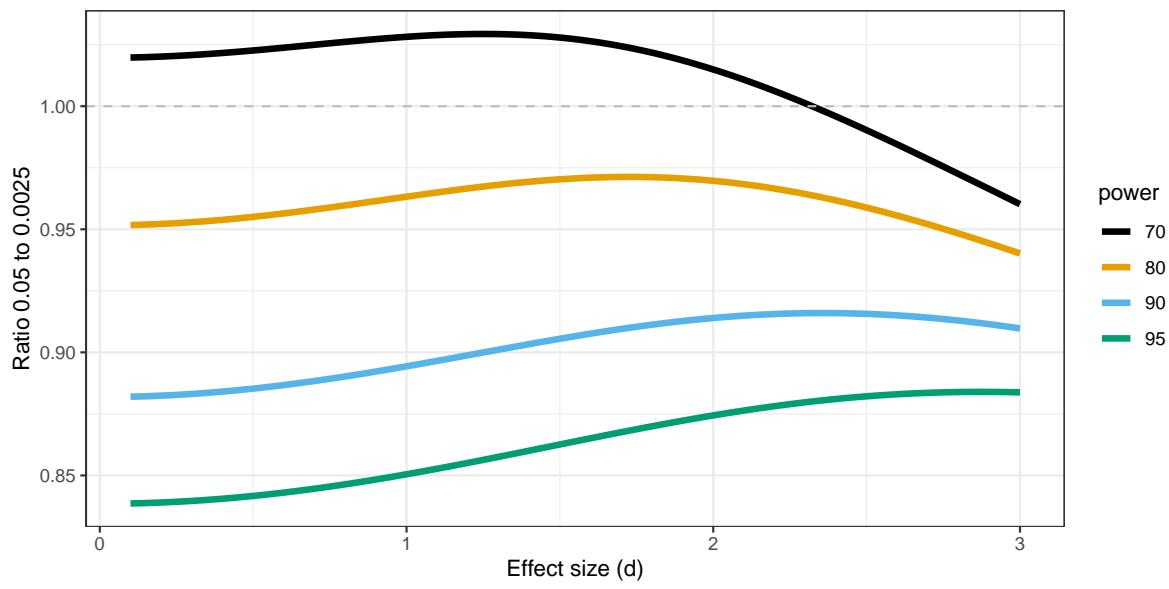


Figure 17.4: Ratio of the sample size required to run one study at an alpha of 0.0025 or two studies at an alpha of 0.05 for a two-sided independent  $t$ -test.

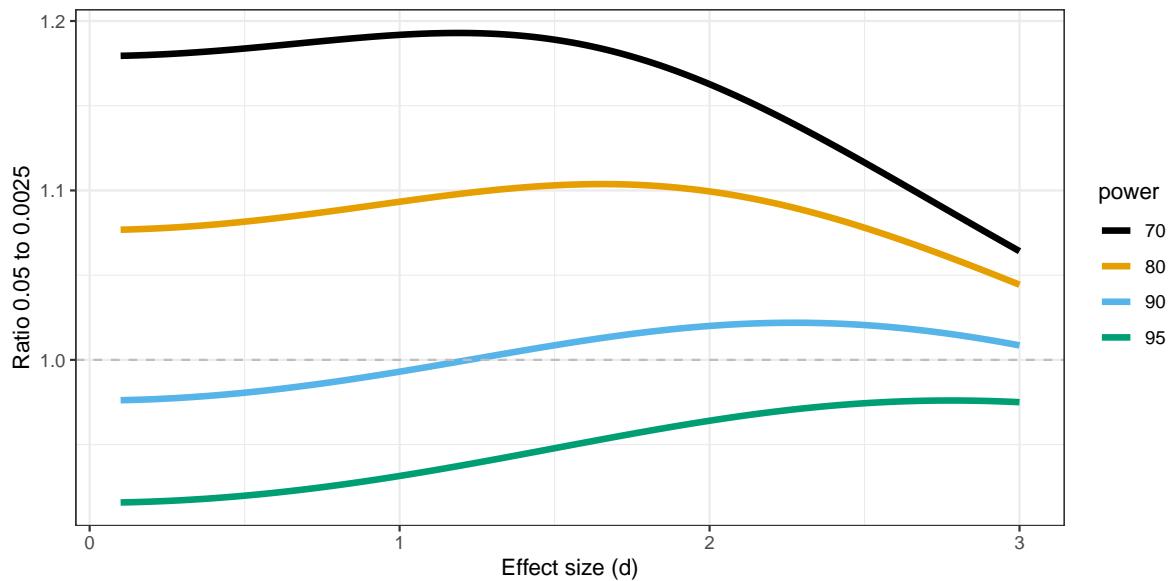


Figure 17.5: Ratio of the sample size required to run one study at an alpha of 0.0025 or two studies at an alpha of 0.05 for a one-sided independent  $t$ -test.

on the directionality of the test and the desired power. There is no clear-cut advantage of a single study at an alpha level of 0.0025, and given the non-statistical benefits of independent replication studies, and the fact that replication studies should arguably always be directional tests, a case can be made to prefer the two study approach. This does require that scientists collaborate, and that all studies are shared regardless of the outcome (for example as Registered Reports).

## 17.5 When replication studies yield conflicting results

As mentioned above, there are three possible reasons for a non-replication: The replication has yielded a Type 1 error, the original study was a Type 1 error, or there is a difference between the two studies. The probability of a Type 2 error can be reduced by performing an a-priori power analysis, or even better, by performing a power analysis for an equivalence test against a smallest effect size of interest. Even then, there is always a probability that the result in a replication study is a false negative.

Some researchers strongly believe failures to replicate published findings can be explained by the presence of hitherto unidentified, or ‘hidden’, moderators (Stroebe & Strack, 2014). There has been at least one example of researchers who were able to provide modest support for the idea that a previous failure to replicate a finding was due to how personally relevant a message in the study was (Luttrell et al., 2017). It is difficult to reliably identify moderator variables that explain failures to replicate published findings, but easy to raise them as an explanation when replication studies do not observe the same effect as the original study. Especially in the social sciences some of these potential moderators are practically impossible to test, such as the fact that society has changed over time. This is an age-old problem, already identified by Galileo in [The Assayer](#), one of the first books on the scientific method. In this book, Galileo discusses the claim that Babylonians cooked eggs by whirling them in a sling, which turned out to be impossible to replicate, and writes:

‘If we do not achieve an effect which others formerly achieved, it must be that we lack something in our operation which was the cause of this effect succeeding, and if we lack one thing only, then this alone can be the true cause. Now we do not lack eggs, or slings, or sturdy fellows to whirl them, and still they do not cook, but rather cool down faster if hot. And since we lack nothing except being Babylonians, then being Babylonian is the cause of the egg hardening.’

At the same time, some failures to replicate *are* due to a difference in auxiliary hypotheses. In the most interesting study examining whether failures to replicate are due to differences in auxiliary assumptions, Ebersole and colleagues (2020) performed additional replications of 10 studies that were replicated in the Reproducibility Project: Psychology (RP:P, Open Science Collaboration (2015)). When the replication studies were designed the authors of original studies were approached for feedback on the design of the replication study. For each of these

10 studies the original authors had raised concerns, but these were not incorporated in the replication study. For example, authors raised the concern that the replication study included participants who had taken prior psychology or economics courses, or who had participated in prior psychology studies. The authors predicted that the effects should be larger in ‘naive’ participants. Other authors pointed out the possibility that the stimuli were not sufficiently pilot tested, the fact that data was collected in a different country, differences in the materials or stimuli, or they pointed out differences in screen resolution of the computer set-up. All these concerns involve predictions about the effects of auxiliary hypotheses, and a team of 172 researchers collaborated with original authors to examine these auxiliary hypotheses in Many Labs 5 (Ebersole et al., 2020). What makes this project especially interesting is that large replication studies were performed both of the RP:P version of the study, as of the revised protocol that addressed the concerns raised by the researchers.

The results of this project provide a nice illustration of how difficult it is to predict whether findings will replicate, until you try to replicate them (Miller, 2009). Two of the studies that did not replicate in the RP:P also did not replicate when a larger new sample was collected, but did replicate with the revised protocol. However, these replication effect sizes were arguably trivially small (Albarracin et al., Study 5 and Shnabel & Nadler in Figure 17.6). A third study (Van Dijk et al) showed a similar pattern but only just failed to show a significant effect in the revised protocol. A fourth study (Albarracin et al., Study 7) was just significant in the original study, and the RP:P study found a very similar effect size which was just non-significant. A meta-analysis pooling these two studies would have yielded a significant meta-analytic effect. And yet, surprisingly, both the replication of the RP:P and the revised protocol based on feedback by the original authors yielded clear null results. A fifth study (Crosby et al) found the same pattern in the original and RP:P study as the second study. Neither the much larger RP:P replication, nor the replication based on the revised protocol yielded a significant result. And yet, the pattern of effect sizes is extremely similar in all four studies, and a meta-analysis across all studies reveals a small but statistically significant effect. In total six of the studies can clearly be regarded as a non-replication where the original authors’ concerns did not matter, as only the original study showed a significant effect, and none of the replication studies yielded a significant result.

Note that we can not conclude that the concerns the authors raised in the other four studies mattered. Despite the large sample sizes, only one statistical difference between the RP:P protocol and the revised protocol was observed (Payne et al.), and here the changes suggested by the authors led to an effect sizes were even *further* away from the effect size in the original study, if we test the effect sizes directly against each other in a test for heterogeneity:

```
# Payne, Burkley, & Stokes (2008)
# Note that the CI in the figure is wide because there is considerable variability across the

r1 <- escalc(ni = 545,
              ri = 0.05,
```

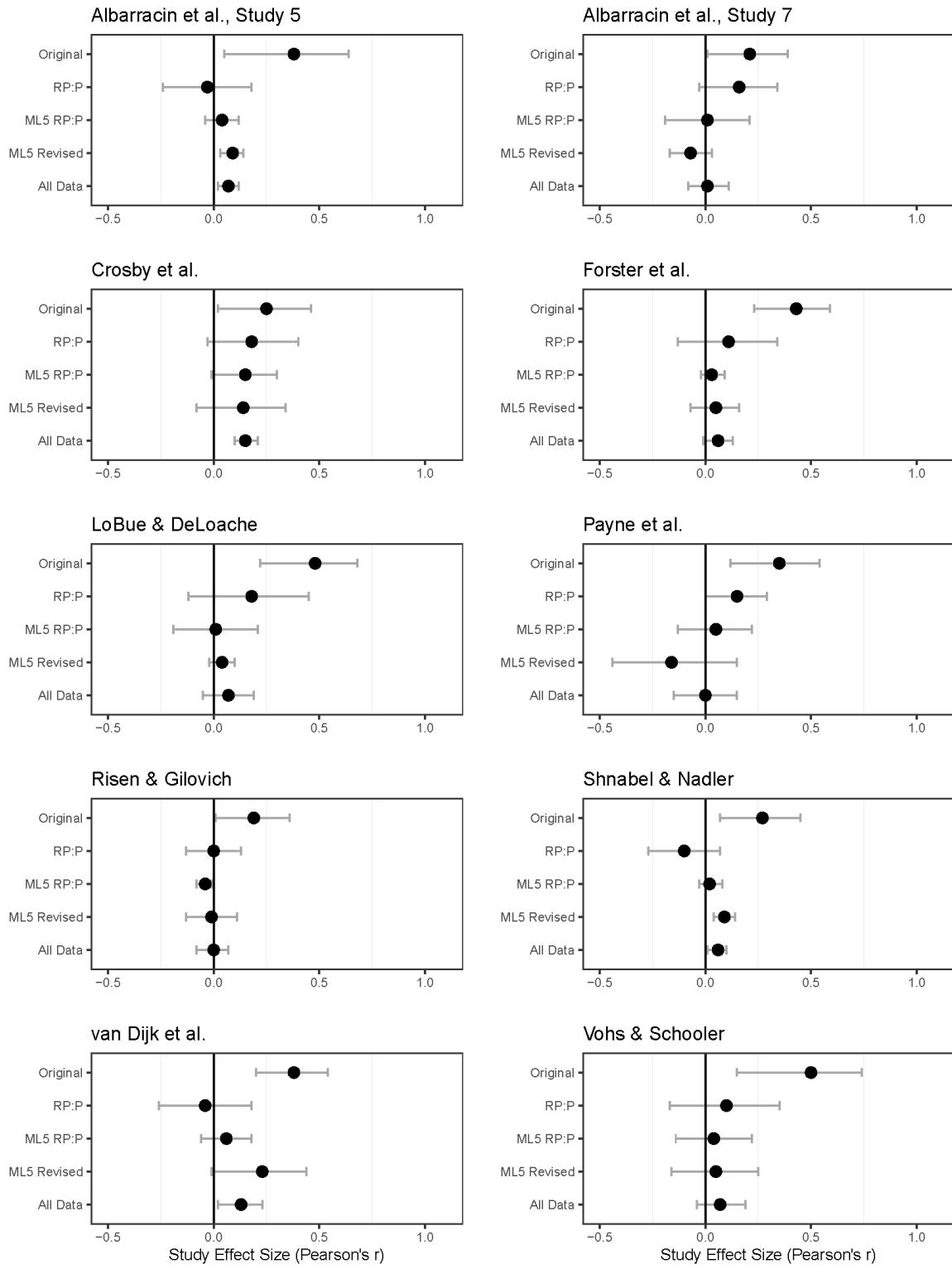


Figure 17.6: Forest plot for the original study, RP:P replication study, the larger replication of the RP:P replication study, and the revised protocol based on author feedback, and a meta-analysis of all data, for all 10 studies in Many Labs 5  
535

```

        measure = "ZCOR")
r2 <- escalc(ni = 558,
              ri = -0.16,
              measure = "ZCOR")
metadata <- data.frame(yi = c(r1$yi, r2$yi),
                        vi = c(r1$vi, r2$vi),
                        study = c("original", "replication"))

# Test based on heterogeneity analysis
res_h <- rma(yi,
               vi,
               data = metadata,
               method = "FE")
res_h
```

Fixed-Effects Model (k = 2)

$I^2$  (total heterogeneity / total variability): 91.84%  
 $H^2$  (total variability / sampling variability): 12.26

Test for Heterogeneity:  
 $Q(df = 1) = 12.2578$ , p-val = 0.0005

Model Results:

| estimate | se     | zval    | pval   | ci.lb   | ci.ub  |   |
|----------|--------|---------|--------|---------|--------|---|
| -0.0569  | 0.0302 | -1.8854 | 0.0594 | -0.1161 | 0.0023 | . |

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The other difference between the RP:P protocol and the revised protocol were not statistically significant, but in one other study the effect size was even further away from the original effect size after the changes to the protocol, and in the other two studies the effect size was more similar as to the original study. Altogether, it seems authors who raise concerns about replication studies do not have a high success rate in predicting which auxiliary hypotheses influence the observed effect size. This is a very important conclusion of Many Labs 5.

## 17.6 Why are replication studies so rare?

“It is difficult to deny that there is more thrill, and usually more glory, involved in blazing a new trail than in checking the pioneer’s work” (Mack, 1951). Throughout history, researchers have pointed out that the reward structures value novel research over replication studies (Fishman & Neigher, 1982; Koole & Lakens, 2012). Performing replication studies is a social dilemma: It is good for everyone if scientists perform replication studies, but it is better for an individual scientist to perform a novel study than a replication study. Exact numbers of how many replication studies are performed are difficult to get, as there is no complete database that keeps track of all replication studies (but see [Curate Science](#), [Replication WIKI](#) or the [Replication Database](#)).

Although the reward structures have remained the same, there are some positive developments. At the start of the replication crisis failed replications of Bem’s pre-cognition study were desk-rejected by the editor of JPSP, Eliot Smith, who stated “This journal does not publish replication studies, whether successful or unsuccessful” and “We don’t want to be the Journal of Bem Replication” (Aldhous, 2011). This led to public outcry, and numerous journals have started to explicitly state they accept replication studies. An increasing number of studies are accepting Registered Report publications, which can also be replication studies, and Peer Community Inn: Registered Reports initiative is publishing replication studies. The APA has created specific [reporting guidelines for replication studies](#). Some science funders have developed [grants for replication research](#). At the same time, replication studies are still rewarded less than novel work, which means researchers who want to build a career are still pushed towards novel research instead of replication studies. So despite positive developments, in many disciplines there is still some way to go before replication studies become a normal aspect of scientific research.

# References

- Abelson, P. (2003). The Value of Life and Health for Public Policy. *Economic Record*, 79, S2–S13. <https://doi.org/10.1111/1475-4932.00087>
- Aberson, C. L. (2019). *Applied Power Analysis for the Behavioral Sciences* (2nd ed.). Routledge.
- Aert, R. C. M. van, & Assen, M. A. L. M. van. (2018). *Correcting for Publication Bias in a Meta-Analysis with the P-uniform\* Method*. <https://doi.org/10.31222/osf.io/zqjir9>
- Agnoli, F., Wicherts, J. M., Veldkamp, C. L. S., Albiero, P., & Cubelli, R. (2017). Questionable research practices among italian research psychologists. *PLOS ONE*, 12(3), e0172792. <https://doi.org/10.1371/journal.pone.0172792>
- Akker, O. van den, Bakker, M., Assen, M. A. L. M. van, Pennington, C. R., Verweij, L., Elsherif, M., Claesen, A., Gaillard, S. D. M., Yeung, S. K., Frankenberger, J.-L., Krautter, K., Cockcroft, J. P., Kreuer, K. S., Evans, T. R., Heppel, F., Schoch, S. F., Korbacher, M., Yamada, Y., Albayrak-Aydemir, N., ... Wicherts, J. (2023, May 10). *The effectiveness of preregistration in psychology: Assessing preregistration strictness and preregistration-study consistency*. <https://doi.org/10.31222/osf.io/h8xjw>
- Albers, C. J., Kiers, H. A. L., & Ravenzwaaij, D. van. (2018). Credible Confidence: A Pragmatic View on the Frequentist vs Bayesian Debate. *Collabra: Psychology*, 4(1, 1), 31. <https://doi.org/10.1525/collabra.149>
- Albers, C. J., & Lakens, D. (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of Experimental Social Psychology*, 74, 187–195. <https://doi.org/10.1016/j.jesp.2017.09.004>
- Aldhous, P. (2011). *Journal rejects studies contradicting precognition*. New Scientist. <https://www.newscientist.com/article/dn20447-journal-rejects-studies-contradicting-precognition/>
- Aldrich, J. (1997). R.A. Fisher and the making of maximum likelihood 1912-1922. *Statistical Science*, 12(3), 162–176. <https://doi.org/10.1214/ss/1030037906>
- Allison, D. B., Allison, R. L., Faith, M. S., Paultre, F., & Pi-Sunyer, F. X. (1997). Power and money: Designing statistically powerful studies while minimizing financial costs. *Psychological Methods*, 2(1), 20–33. <https://doi.org/10.1037/1082-989X.2.1.20>
- Altman, D. G., & Bland, J. M. (1995). Statistics notes: Absence of evidence is not evidence of absence. *BMJ*, 311(7003), 485. <https://doi.org/10.1136/bmj.311.7003.485>
- Altoè, G., Bertoldo, G., Zandonella Callegher, C., Toffalini, E., Calcagnì, A., Finos, L., & Pastore, M. (2020). Enhancing Statistical Inference in Psychological Research via Prospective and Retrospective Design Analysis. *Frontiers in Psychology*, 10. <https://www.frontiersin.org/articles/10.3389/fpsyg.2019.02893>

- Anderson, M. S., Martinson, B. C., & De Vries, R. (2007). Normative dissonance in science: Results from a national survey of US scientists. *Journal of Empirical Research on Human Research Ethics*, 2(4), 3–14. <http://jre.sagepub.com/content/2/4/3.short>
- Anderson, M. S., Ronning, E. A., De Vries, R., & Martinson, B. C. (2007). The perverse effects of competition on scientists' work and relationships. *Science and Engineering Ethics*, 13(4), 437–461. <http://link.springer.com/article/10.1007/s11948-007-9042-5>
- Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychological Science*, 28(11), 1547–1562. <https://doi.org/10.1177/0956797617723724>
- Anderson, S. F., & Maxwell, S. E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods*, 21(1), 1–12. <https://doi.org/10.1037/met0000051>
- Anvari, F., Kievit, R., Lakens, D., Pennington, C. R., Przybylski, A. K., Tiokhin, L., Wiernik, B. M., & Orben, A. (2021). Not all effects are indispensable: Psychological science requires verifiable lines of reasoning for whether an effect matters. *Perspectives on Psychological Science*. <https://doi.org/10.31234/osf.io/g3vtr>
- Anvari, F., & Lakens, D. (2018). The replicability crisis and public trust in psychological science. *Comprehensive Results in Social Psychology*, 3(3), 266–286. <https://doi.org/10.1080/23743603.2019.1684822>
- Anvari, F., & Lakens, D. (2021). Using anchor-based methods to determine the smallest effect size of interest. *Journal of Experimental Social Psychology*, 96, 104159. <https://doi.org/10.1016/j.jesp.2021.104159>
- Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, 73(1), 3. <https://doi.org/10.1037/amp0000191>
- Armitage, P., McPherson, C. K., & Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society: Series A (General)*, 132(2), 235–244.
- Arslan, R. C. (2019). How to Automatically Document Data With the codebook Package to Facilitate Data Reuse. *Advances in Methods and Practices in Psychological Science*, 2515245919838783. <https://doi.org/10.1177/2515245919838783>
- Azrin, N. H., Holz, W., Ulrich, R., & Goldiamond, I. (1961). The control of the content of conversation through reinforcement. *Journal of the Experimental Analysis of Behavior*, 4, 25–30. <https://doi.org/10.1901/jeab.1961.4-25>
- Babbage, C. (1830). *Reflections on the Decline of Science in England: And on Some of Its Causes*. B. Fellowes. <http://archive.org/details/reflectionsonde00mollgoog>
- Bacchetti, P. (2010). Current sample size conventions: Flaws, harms, and alternatives. *BMC Medicine*, 8(1), 17. <https://doi.org/10.1186/1741-7015-8-17>
- Baguley, T. (2004). Understanding statistical power in the context of applied research. *Applied Ergonomics*, 35(2), 73–80. <https://doi.org/10.1016/j.apergo.2004.01.002>
- Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British*

- Journal of Psychology*, 100(3), 603–617. <https://doi.org/10.1348/000712608X377117>
- Baguley, T. (2012). *Serious stats: A guide to advanced statistics for the behavioral sciences*. Palgrave Macmillan.
- Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, 66(6), 423–437. <https://doi.org/10.1037/h0020412>
- Bakan, D. (1967). *On method: Toward a reconstruction of psychological investigation*. San Francisco, Jossey-Bass. [http://archive.org/details/onmethodtowardre0000baka\\_y0b7](http://archive.org/details/onmethodtowardre0000baka_y0b7)
- Bakker, B. N., Kokil, J., Dörr, T., Fasching, N., & Lelkes, Y. (2021). Questionable and Open Research Practices: Attitudes and Perceptions among Quantitative Communication Researchers. *Journal of Communication*, 71(5), 715–738. <https://doi.org/10.1093/joc/jqab031>
- Ball, K., Berch, D. B., Helmers, K. F., Jobe, J. B., Leveck, M. D., Marsiske, M., Morris, J. N., Rebok, G. W., Smith, D. M., & Tennstedt, S. L. (2002). Effects of cognitive training interventions with older adults: A randomized controlled trial. *Jama*, 288(18), 2271–2281.
- Barber, T. X. (1976). *Pitfalls in Human Research: Ten Pivotal Points*. Pergamon Press. <https://books.google.com?id=UBN9AAAAMAAJ>
- Bartoš, F., & Schimmack, U. (2020). *Z-Curve.2.0: Estimating Replication Rates and Discovery Rates*. <https://doi.org/10.31234/osf.io/urgtn>
- Bauer, P., & Kieser, M. (1996). A unifying approach for confidence intervals and testing of equivalence and difference. *Biometrika*, 83(4), 934–937. <http://biomet.oxfordjournals.org/content/83/4/934.short>
- Bausell, R. B., & Li, Y.-F. (2002). *Power Analysis for Experimental Research: A Practical Guide for the Biological, Medical and Social Sciences* (1st edition). Cambridge University Press.
- Beck, W. S. (1957). *Modern Science and the nature of life* (First Edition). Harcourt, Brace.
- Becker, B. J. (2005). Failsafe N or File-Drawer Number. In *Publication Bias in Meta-Analysis* (pp. 111–125). John Wiley & Sons, Ltd. <https://doi.org/10.1002/0470870168.ch7>
- Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100(3), 407–425. <https://doi.org/10.1037/a0021524>
- Bem, D. J., Utts, J., & Johnson, W. O. (2011). Must psychologists change the way they analyze their data? *Journal of Personality and Social Psychology*, 101(4), 716–719. <https://doi.org/10.1037/a0024777>
- Bender, R., & Lange, S. (2001). Adjusting for multiple testing—when and how? *Journal of Clinical Epidemiology*, 54(4), 343–349. <http://www.sciencedirect.com/science/article/pii/S0895435600003140>
- Benjamini, Y. (2016). It's Not the p-values' Fault. *The American Statistician: Supplemental Material to the ASA Statement on P-Values and Statistical Significance*, 70, 1–2. [http://amstat.tandfonline.com/doi/suppl/10.1080/00031305.2016.1154108/suppl\\_file/utas\\_a\\_1154108\\_sm5354.pdf](http://amstat.tandfonline.com/doi/suppl/10.1080/00031305.2016.1154108/suppl_file/utas_a_1154108_sm5354.pdf)
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289–300. <http://www.jstor.org/stable/2346101>

- Ben-Shachar, M. S., Lüdecke, D., & Makowski, D. (2020). Effectsize: Estimation of Effect Size Indices and Standardized Parameters. *Journal of Open Source Software*, 5(56), 2815. <https://doi.org/10.21105/joss.02815>
- Berger, J. O., & Bayarri, M. J. (2004). The Interplay of Bayesian and Frequentist Analysis. *Statistical Science*, 19(1), 58–80. <https://doi.org/10.1214/088342304000000116>
- Berkeley, G. (1735). *A defence of free-thinking in mathematics, in answer to a pamphlet of Philalethes Cantabrigiensis entitled Geometry No Friend to Infidelity. Also an appendix concerning mr. Walton's Vindication of the principles of fluxions against the objections contained in The analyst. By the author of The minute philosopher* (Vol. 3). <https://www.maths.tcd.ie/pub/HistMath/People/Berkeley/Defence/Defence.html>
- Bird, S. B., & Sivilotti, M. L. A. (2008). Self-plagiarism, recycling fraud, and the intent to mislead. *Journal of Medical Toxicology*, 4(2), 69–70. <https://doi.org/10.1007/BF03160957>
- Bishop, D. V. M. (2018). Fallibility in Science: Responding to Errors in the Work of Oneself and Others. *Advances in Methods and Practices in Psychological Science*, 2515245918776632. <https://doi.org/10.1177/2515245918776632>
- Bland, M. (2015). *An introduction to medical statistics* (Fourth edition). Oxford University Press.
- Bonett, D. G. (2012). Replication-Extension Studies. *Current Directions in Psychological Science*, 21(6), 409–412. <https://doi.org/10.1177/0963721412459512>
- Borenstein, M. (Ed.). (2009). *Introduction to meta-analysis*. John Wiley & Sons.
- Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *The Journal of Applied Psychology*, 100(2), 431–449. <https://doi.org/10.1037/a0038047>
- Bozarth, J. D., & Roberts, R. R. (1972). Signifying significant significance. *American Psychologist*, 27(8), 774.
- Bretz, F., Hothorn, T., & Westfall, P. H. (2011). *Multiple comparisons using R*. CRC Press.
- Bross, I. D. (1971). Critical levels, statistical language and scientific inference. In *Foundations of statistical inference* (pp. 500–513). Holt, Rinehart and Winston.
- Brown, G. W. (1983). Errors, Types I and II. *American Journal of Diseases of Children*, 137(6), 586–591. <https://doi.org/10.1001/archpedi.1983.02140320062014>
- Brown, N. J. L., & Heathers, J. A. J. (2017). The GRIM Test: A Simple Technique Detects Numerous Anomalies in the Reporting of Results in Psychology. *Social Psychological and Personality Science*, 8(4), 363–369. <https://doi.org/10.1177/1948550616673876>
- Brunner, J., & Schimmack, U. (2020). Estimating Population Mean Power Under Conditions of Heterogeneity and Selection for Significance. *Meta-Psychology*, 4. <https://doi.org/10.15626/MP.2018.874>
- Bryan, C. J., Tipton, E., & Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behaviour*, 1–10. <https://doi.org/10.1038/s41562-021-01143-3>
- Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*, 2(1), 16. <https://doi.org/10.5334/joc.72>
- Brysbaert, M., & Stevens, M. (2018). Power Analysis and Effect Size in Mixed Effects Models:

- A Tutorial. *Journal of Cognition*, 1(1). <https://doi.org/10.5334/joc.10>
- Buchanan, E. M., Scofield, J., & Valentine, K. D. (2017). *MOTE: Effect Size and Confidence Interval Calculator*. (Version 0.0.0.9100.).
- Bulus, Metin, & Dong, N. (2021). Bound Constrained Optimization of Sample Sizes Subject to Monetary Restrictions in Planning Multilevel Randomized Trials and Regression Discontinuity Studies. *The Journal of Experimental Education*, 89(2), 379–401. <https://doi.org/10.1080/00220973.2019.1636197>
- Bulus, M., & Polat, C. (2023). *pwrss R paketi ile istatistiksel güç analizi [Statistical power analysis with pwrss R package]*. <https://osf.io/ua5fc>
- Burriss, R. P., Troscianko, J., Lovell, P. G., Fulford, A. J. C., Stevens, M., Quigley, R., Payne, J., Saxton, T. K., & Rowland, H. M. (2015). Changes in women's facial skin color over the ovulatory cycle are not detectable by the human visual system. *PLOS ONE*, 10(7), e0130093. <https://doi.org/10.1371/journal.pone.0130093>
- Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5), 365–376. <https://doi.org/10.1038/nrn3475>
- Button, K. S., Kounali, D., Thomas, L., Wiles, N. J., Peters, T. J., Welton, N. J., Ades, A. E., & Lewis, G. (2015). Minimal clinically important difference on the Beck Depression Inventory - II according to the patient's perspective. *Psychological Medicine*, 45(15), 3269–3279. <https://doi.org/10.1017/S0033291715001270>
- Caplan, A. L. (2021). How Should We Regard Information Gathered in Nazi Experiments? *AMA Journal of Ethics*, 23(1), 55–58. <https://doi.org/10.1001/amajethics.2021.55>
- Carter, E. C., & McCullough, M. E. (2014). Publication bias and the limited strength model of self-control: Has the evidence for ego depletion been overestimated? *Frontiers in Psychology*, 5. <https://doi.org/10.3389/fpsyg.2014.00823>
- Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for Bias in Psychology: A Comparison of Meta-Analytic Methods. *Advances in Methods and Practices in Psychological Science*, 2(2), 115–144. <https://doi.org/10.1177/2515245919847196>
- Cascio, W. F., & Zedeck, S. (1983). Open a New Window in Rational Research Planning: Adjust Alpha to Maximize Statistical Power. *Personnel Psychology*, 36(3), 517–526. <https://doi.org/10.1111/j.1744-6570.1983.tb02233.x>
- Ceci, S. J., & Bjork, R. A. (2000). Psychological Science in the Public Interest: The Case for Juried Analyses. *Psychological Science in the Public Interest*, 11(3), 177–178. <https://doi.org/10.1111/1467-9280.00237>
- Cevolani, G., Crupi, V., & Festa, R. (2011). Verisimilitude and belief change for conjunctive theories. *Erkenntnis*, 75(2), 183. <http://link.springer.com/article/10.1007/s10670-011-9290-2>
- Chalmers, I., & Glasziou, P. (2009). Avoidable waste in the production and reporting of research evidence. *The Lancet*, 374(9683), 86–89.
- Chamberlin, T. C. (1890). The Method of Multiple Working Hypotheses. *Science*, ns-15(366), 92–96. <https://doi.org/10.1126/science.ns-15.366.92>
- Chambers, C. D., & Tzavella, L. (2022). The past, present and future of Registered Reports.

- Nature Human Behaviour*, 6(1, 1), 29–42. <https://doi.org/10.1038/s41562-021-01193-7>
- Chang, H. (2022). *Realism for Realistic People: A New Pragmatist Philosophy of Science*. Cambridge University Press. <https://doi.org/10.1017/9781108635738>
- Chang, M. (2016). *Adaptive Design Theory and Implementation Using SAS and R* (2nd edition). Chapman and Hall/CRC.
- Chatziathanasiou, K. (2022). *Beware the Lure of Narratives: “Hungry Judges” Should not Motivate the Use of “Artificial Intelligence” in Law* (SSRN Scholarly Paper ID 4011603). Social Science Research Network. <https://doi.org/10.2139/ssrn.4011603>
- Chin, J. M., Pickett, J. T., Vazire, S., & Holcombe, A. O. (2021). Questionable Research Practices and Open Science in Quantitative Criminology. *Journal of Quantitative Criminology*. <https://doi.org/10.1007/s10940-021-09525-6>
- Cho, H.-C., & Abe, S. (2013). Is two-tailed testing for directional research hypotheses tests legitimate? *Journal of Business Research*, 66(9), 1261–1266. <https://doi.org/10.1016/j.jbusres.2012.02.023>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). L. Erlbaum Associates.
- Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, 45(12), 1304–1312. <https://doi.org/10.1037/0003-066X.45.12.1304>
- Cohen, J. (1994). The earth is round ( $p < .05$ ). *American Psychologist*, 49(12), 997–1003. <https://doi.org/10.1037/0003-066X.49.12.997>
- Coles, N. A., March, D. S., Marmolejo-Ramos, F., Larsen, J. T., Arinze, N. C., Ndukaiche, I. L. G., Willis, M. L., Foroni, F., Reggev, N., Mokady, A., Forscher, P. S., Hunter, J. F., Kaminski, G., Yüvrük, E., Kapucu, A., Nagy, T., Hajdu, N., Tejada, J., Freitag, R. M. K., ... Liuzza, M. T. (2022). A multi-lab test of the facial feedback hypothesis by the Many Smiles Collaboration. *Nature Human Behaviour*, 6(12, 12), 1731–1742. <https://doi.org/10.1038/s41562-022-01458-9>
- Colling, L. J., Szűcs, D., De Marco, D., Cipora, K., Ulrich, R., Nuerk, H.-C., Soltanlou, M., Bryce, D., Chen, S.-C., Schroeder, P. A., Henare, D. T., Chrystall, C. K., Corballis, P. M., Ansari, D., Goffin, C., Sokolowski, H. M., Hancock, P. J. B., Millen, A. E., Langton, S. R. H., ... McShane, B. B. (2020). Registered Replication Report on Fischer, Castel, Dodd, and Pratt (2003). *Advances in Methods and Practices in Psychological Science*, 3(2), 143–162. <https://doi.org/10.1177/2515245920903079>
- Colquhoun, D. (2019). The False Positive Risk: A Proposal Concerning What to Do About p-Values. *The American Statistician*, 73, 192–201. <https://doi.org/10.1080/00031305.2018.1529622>
- Cook, J., Hislop, J., Adewuyi, T., Harrild, K., Altman, D., Ramsay, C., Fraser, C., Buckley, B., Fayers, P., Harvey, I., Briggs, A., Norrie, J., Fergusson, D., Ford, I., & Vale, L. (2014). Assessing methods to specify the target difference for a randomised controlled trial: DELTA (Difference ELicitation in TriAls) review. *Health Technology Assessment*, 18(28). <https://doi.org/10.3310/hta18280>
- Cook, T. D. (2002). P-Value Adjustment in Sequential Clinical Trials. *Biometrics*, 58(4), 1005–1011.
- Cooper, H. (2020). *Reporting quantitative research in psychology: How to meet APA Style*

- Journal Article Reporting Standards (2nd ed.).* American Psychological Association. <https://doi.org/10.1037/0000178-000>
- Cooper, H. M., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). Russell Sage Foundation.
- Copay, A. G., Subach, B. R., Glassman, S. D., Polly, D. W., & Schuler, T. C. (2007). Understanding the minimum clinically important difference: A review of concepts and methods. *The Spine Journal*, 7(5), 541–546. <https://doi.org/10.1016/j.spinee.2007.01.008>
- Corneille, O., Havemann, J., Henderson, E. L., IJzerman, H., Hussey, I., Orban de Xivry, J.-J., Jussim, L., Holmes, N. P., Pilacinski, A., Beffara, B., Carroll, H., Outa, N. O., Lush, P., & Lotter, L. D. (2023). Beware “persuasive communication devices” when writing and reading scientific articles. *eLife*, 12, e88654. <https://doi.org/10.7554/eLife.88654>
- Correll, J., Mellinger, C., McClelland, G. H., & Judd, C. M. (2020). Avoid Cohen’s “Small,” “Medium,” and “Large” for Power Analysis. *Trends in Cognitive Sciences*, 24(3), 200–207. <https://doi.org/10.1016/j.tics.2019.12.009>
- Cousineau, D., & Chiasson, F. (2019). *Superb: Computes standard error and confidence interval of means under various designs and sampling schemes* [Manual].
- Cowles, M., & Davis, C. (1982). On the origins of the .05 level of statistical significance. *American Psychologist*, 37(5), 553. <http://psycnet.apa.org/journals/amp/37/5/553/>
- Cox, D. R. (1958). Some Problems Connected with Statistical Inference. *Annals of Mathematical Statistics*, 29(2), 357–372. <https://doi.org/10.1214/aoms/117706618>
- Cribbie, R. A., Gruman, J. A., & Arpin-Cribbie, C. A. (2004). Recommendations for applying tests of equivalence. *Journal of Clinical Psychology*, 60(1), 1–10.
- Crusius, J., Gonzalez, M. F., Lange, J., & Cohen-Charash, Y. (2020). Envy: An Adversarial Review and Comparison of Two Competing Views. *Emotion Review*, 12(1), 3–21. <https://doi.org/10.1177/1754073919873131>
- Crüwell, S., Apthorp, D., Baker, B. J., Colling, L., Elson, M., Geiger, S. J., Lobentanzer, S., Monéger, J., Patterson, A., Schwarzkopf, D. S., Zaneva, M., & Brown, N. J. L. (2023). What’s in a Badge? A Computational Reproducibility Investigation of the Open Data Badge Policy in One Issue of Psychological Science. *Psychological Science*, 09567976221140828. <https://doi.org/10.1177/09567976221140828>
- Cumming, G. (2008). Replication and *p* Intervals: *p* Values Predict the Future Only Vaguely, but Confidence Intervals Do Much Better. *Perspectives on Psychological Science*, 3(4), 286–300. <https://doi.org/10.1111/j.1745-6924.2008.00079.x>
- Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge. [https://books.google.nl/books?hl=nl&lr=&id=1W6laNc7Xt8C&oi=fnd&pg=PR1&dq=cumming+new+statistics&ots=PuKTSNb51T&sig=U\\_k1y1YyREjLxP-FMfw4ood9OI](https://books.google.nl/books?hl=nl&lr=&id=1W6laNc7Xt8C&oi=fnd&pg=PR1&dq=cumming+new+statistics&ots=PuKTSNb51T&sig=U_k1y1YyREjLxP-FMfw4ood9OI)
- Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science*, 25(1), 7–29. <https://doi.org/10.1177/0956797613504966>
- Cumming, G., & Calin-Jageman, R. (2016). *Introduction to the New Statistics: Estimation, Open Science, and Beyond*. Routledge. <https://books.google.com?id=KR8xDQAAQBAJ>
- Cumming, G., & Maillardet, R. (2006). Confidence intervals and replication: Where will the next mean fall? *Psychological Methods*, 11(3), 217–227. <https://doi.org/10.1037/1082->

989X.11.3.217

- Danziger, S., Levav, J., & Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, 108(17), 6889–6892. <https://doi.org/10.1073/PNAS.1018033108>
- de Groot, A. D. (1969). *Methodology* (Vol. 6). Mouton & Co.
- de Heide, R., & Grünwald, P. D. (2017). *Why optional stopping is a problem for Bayesians*. <http://arxiv.org/abs/1708.08278>
- DeBruine, L. M., & Barr, D. J. (2021). Understanding Mixed-Effects Models Through Data Simulation. *Advances in Methods and Practices in Psychological Science*, 4(1), 2515245920965119. <https://doi.org/10.1177/2515245920965119>
- Delacre, M., Lakens, D., Ley, C., Liu, L., & Leys, C. (2021). *Why Hedges' g\*s based on the non-pooled standard deviation should be reported with Welch's t-test*. <https://doi.org/10.31234/osf.io/tu6mp>
- Delacre, M., Lakens, D., & Leys, C. (2017). Why Psychologists Should by Default Use Welch's *t*-test Instead of Student's *t*-test. *International Review of Social Psychology*, 30(1). <https://doi.org/10.5334/irsp.82>
- Detsky, A. S. (1990). Using cost-effectiveness analysis to improve the efficiency of allocating funds to clinical trials. *Statistics in Medicine*, 9(1-2), 173–184. <https://doi.org/10.1002/sim.4780090124>
- Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. Palgrave Macmillan.
- Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, 5. <https://doi.org/10.3389/fpsyg.2014.00781>
- Dmitrienko, A., & D'Agostino Sr, R. (2013). Traditional multiplicity adjustment methods in clinical trials. *Statistics in Medicine*, 32(29), 5172–5218. <https://doi.org/10.1002/sim.5990>
- Dodge, H. F., & Romig, H. G. (1929). A Method of Sampling Inspection. *Bell System Technical Journal*, 8(4), 613–631. <https://doi.org/10.1002/j.1538-7305.1929.tb01240.x>
- Dongen, N. N. N. van, Doorn, J. B. van, Gronau, Q. F., Ravenzwaaij, D. van, Hoekstra, R., Haucke, M. N., Lakens, D., Hennig, C., Morey, R. D., Homer, S., Gelman, A., Sprenger, J., & Wagenmakers, E.-J. (2019). Multiple Perspectives on Inference for Two Simple Statistical Scenarios. *The American Statistician*, 73, 328–339. <https://doi.org/10.1080/00031305.2019.1565553>
- Douglas, H. E. (2009). *Science, policy, and the value-free ideal*. University of Pittsburgh Press.
- Dubin, R. (1969). *Theory building*. Free Press. <http://catalog.hathitrust.org/api/volumes/oclc/160506.html>
- Duhem, P. (1954). *The aim and structure of physical theory*. Princeton University Press.
- Dupont, W. D. (1983). Sequential stopping rules and sequentially adjusted P values: Does one require the other? *Controlled Clinical Trials*, 4(1), 3–10. [https://doi.org/10.1016/S0197-2456\(83\)80003-8](https://doi.org/10.1016/S0197-2456(83)80003-8)
- Duyx, B., Urlings, M. J. E., Swaen, G. M. H., Bouter, L. M., & Zeegers, M. P. (2017). Scientific citations favor positive results: A systematic review and meta-analysis. *Journal of Clinical Epidemiology*, 88, 92–101. <https://doi.org/10.1016/j.jclinepi.2017.06.002>
- Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks,

- J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B. V., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J. A., Conway, J. G., ... Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, 67, 68–82. <https://doi.org/10.1016/j.jesp.2015.10.012>
- Ebersole, C. R., Mathur, M. B., Baranski, E., Bart-Plange, D.-J., Buttrick, N. R., Chartier, C. R., Corker, K. S., Corley, M., Hartshorne, J. K., IJzerman, H., Lazarević, L. B., Rabagliati, H., Ropovik, I., Aczel, B., Aeschbach, L. F., Andrighetto, L., Arnal, J. D., Arrow, H., Babincak, P., ... Nosek, B. A. (2020). Many Labs 5: Testing Pre-Data-Collection Peer Review as an Intervention to Increase Replicability. *Advances in Methods and Practices in Psychological Science*, 3(3), 309–331. <https://doi.org/10.1177/2515245920958687>
- Eckermann, S., Karnon, J., & Willan, A. R. (2010). The Value of Value of Information. *PharmacoEconomics*, 28(9), 699–709. <https://doi.org/10.2165/11537370-00000000-00000>
- Edwards, M. A., & Roy, S. (2017). Academic Research in the 21st Century: Maintaining Scientific Integrity in a Climate of Perverse Incentives and Hypercompetition. *Environmental Engineering Science*, 34(1), 51–61. <https://doi.org/10.1089/ees.2016.0223>
- Elson, M., Mohseni, M. R., Breuer, J., Scharkow, M., & Quandt, T. (2014). Press CRTT to measure aggressive behavior: The unstandardized use of the competitive reaction time task in aggression research. *Psychological Assessment*, 26(2), 419–432. <https://doi.org/10.1037/a0035569>
- Epstein, S. (1980). The stability of behavior: II. Implications for psychological research. *American Psychologist*, 35(9), 790–806. <https://doi.org/10.1037/0003-066X.35.9.790>
- Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, 28(1), 1–11. <https://doi.org/10.3758/BF03203630>
- Eysenck, H. J. (1978). An exercise in mega-silliness. *American Psychologist*, 33(5), 517–517. <https://doi.org/10.1037/0003-066X.33.5.517.a>
- Fanelli, D. (2010). “Positive” Results Increase Down the Hierarchy of the Sciences. *PLoS ONE*, 5(4). <https://doi.org/10.1371/journal.pone.0010068>
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). GPower 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175–191. <https://doi.org/10.3758/BF03193146>
- Ferguson, C. J. (2014). Comment: Why meta-analyses rarely resolve ideological debates. *Emotion Review*, 6(3), 251–252. <http://journals.sagepub.com/doi/abs/10.1177/1754073914523046>
- Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories publication bias and psychological science’s aversion to the null. *Perspectives on Psychological Science*, 7(6), 555–561. <http://pps.sagepub.com/content/7/6/555.short>
- Ferguson, C. J., & Heene, M. (2021). Providing a lower-bound estimate for psychology’s “crud factor”: The case of aggression. *Professional Psychology: Research and Practice*, 52(6), 620–626. <https://doi.org/http://dx.doi.org/10.1037/pro0000386>
- Ferguson, C., Marcus, A., & Oransky, I. (2014). Publishing: The peer-review scam. *Nature*, 515(7528, 7528), 480–482. <https://doi.org/10.1038/515480a>

- Ferron, J., & Onghena, P. (1996). The Power of Randomization Tests for Single-Case Phase Designs. *The Journal of Experimental Education*, 64(3), 231–239. <https://doi.org/10.1080/00220973.1996.9943805>
- Feyerabend, P. (1993). *Against method* (3rd ed). Verso.
- Feynman, R. P. (1974). Cargo cult science. *Engineering and Science*, 37(7), 10–13.
- Fiedler, K. (2004). Tools, toys, truisms, and theories: Some thoughts on the creative cycle of theory formation. *Personality and Social Psychology Review*, 8(2), 123–131. [https://doi.org/10.1207/s15327957pspr0802\\_5](https://doi.org/10.1207/s15327957pspr0802_5)
- Fiedler, K., & Schwarz, N. (2016). Questionable Research Practices Revisited. *Social Psychological and Personality Science*, 7(1), 45–52. <https://doi.org/10.1177/1948550615612150>
- Field, S. A., Tyre, A. J., Jonzén, N., Rhodes, J. R., & Possingham, H. P. (2004). Minimizing the cost of environmental management decisions by optimizing statistical thresholds. *Ecology Letters*, 7(8), 669–675. <https://doi.org/10.1111/j.1461-0248.2004.00625.x>
- Fisher, Ronald Aylmer. (1935). *The design of experiments*. Oliver And Boyd; Edinburgh; London.
- Fisher, Ronald A. (1936). Has Mendel's work been rediscovered? *Annals of Science*, 1(2), 115–137.
- Fisher, Ronald A. (1956). *Statistical methods and scientific inference*: Vol. viii. Hafner Publishing Co.
- Fishman, D. B., & Neigher, W. D. (1982). American psychology in the eighties: Who will buy? *American Psychologist*, 37(5), 533–546. <https://doi.org/10.1037/0003-066X.37.5.533>
- Fraley, R. C., & Vazire, S. (2014). The N-Pact Factor: Evaluating the Quality of Empirical Journals with Respect to Sample Size and Statistical Power. *PLOS ONE*, 9(10), e109019. <https://doi.org/10.1371/journal.pone.0109019>
- Francis, G. (2014). The frequency of excess success for articles in Psychological Science. *Psychonomic Bulletin & Review*, 21(5), 1180–1187. <https://doi.org/10.3758/s13423-014-0601-x>
- Francis, G. (2016). Equivalent statistics and data interpretation. *Behavior Research Methods*, 1–15. <https://doi.org/10.3758/s13428-016-0812-3>
- Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, 345(6203), 1502–1505. <https://doi.org/10.1126/SCIENCE.1255484>
- Frankenhuis, W. E., Panchanathan, K., & Smaldino, P. E. (2022). Strategic ambiguity in the social sciences. *Social Psychological Bulletin*. <https://www.psycharchives.org/en/item/e5bb9192-80a4-4ae4-9cda-5d144008196e>
- Fraser, H., Parker, T., Nakagawa, S., Barnett, A., & Fidler, F. (2018). Questionable research practices in ecology and evolution. *PLOS ONE*, 13(7), e0200303. <https://doi.org/10.1371/journal.pone.0200303>
- Freedman, J. L., & Fraser, S. C. (1966). Compliance without pressure: The foot-in-the-door technique. *Journal of Personality and Social Psychology*, 4(2), 195–202. <https://doi.org/10.1037/h0023552>
- Freiman, J. A., Chalmers, T. C., Smith, H., & Kuebler, R. R. (1978). The importance of beta, the type II error and sample size in the design and interpretation of the randomized control

- trial. Survey of 71 "negative" trials. *The New England Journal of Medicine*, 299(13), 690–694. <https://doi.org/10.1056/NEJM197809282991304>
- Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*, 1(4), 379–390. <https://doi.org/10.1037/1082-989X.1.4.379>
- Fricker, R. D., Burke, K., Han, X., & Woodall, W. H. (2019). Assessing the Statistical Analyses Used in Basic and Applied Social Psychology After Their p-Value Ban. *The American Statistician*, 73, 374–384. <https://doi.org/10.1080/00031305.2018.1537892>
- Fried, B. J., Boers, M., & Baker, P. R. (1993). A method for achieving consensus on rheumatoid arthritis outcome measures: The OMERACT conference process. *The Journal of Rheumatology*, 20(3), 548–551.
- Friede, T., & Kieser, M. (2006). Sample size recalculation in internal pilot study designs: A review. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 48(4), 537–555. <https://doi.org/10.1002/bimj.200510238>
- Friedlander, F. (1964). Type I and Type II Bias. *American Psychologist*, 19(3), 198–199. <https://doi.org/10.1037/h0038977>
- Fugard, A. J. B., & Potts, H. W. W. (2015). Supporting thinking on sample sizes for thematic analyses: A quantitative tool. *International Journal of Social Research Methodology*, 18(6), 669–684. <https://doi.org/10.1080/13645579.2015.1005453>
- Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, 2(2), 156–168. <https://doi.org/10.1177/2515245919847202>
- Gannon, M. A., de Bragança Pereira, C. A., & Polpo, A. (2019). Blending Bayesian and Classical Tools to Define Optimal Sample-Size-Dependent Significance Levels. *The American Statistician*, 73, 213–222. <https://doi.org/10.1080/00031305.2018.1518268>
- Gelman, A., & Carlin, J. (2014). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science*, 9(6), 641–651. <https://pdfs.semanticscholar.org/8658/058a37d6e2d0d58b4da7efbc5a653544ee64.pdf>
- Gergen, K. J. (1973). Social psychology as history. *Journal of Personality and Social Psychology*, 26, 309–320. <https://doi.org/10.1037/h0034436>
- Gerring, J. (2012). Mere Description. *British Journal of Political Science*, 42(4), 721–746. <https://doi.org/10.1017/S0007123412000130>
- Gillon, R. (1994). Medical ethics: Four principles plus attention to scope. *BMJ*, 309(6948), 184. <https://doi.org/10.1136/bmj.309.6948.184>
- Glöckner, A. (2016). The irrational hungry judge effect revisited: Simulations reveal that the magnitude of the effect is overestimated. *Judgment and Decision Making*, 11(6), 601–610.
- Glover, S., & Dixon, P. (2004). Likelihood ratios: A simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin & Review*, 11(5), 791–806.
- Goldacre, B., DeVito, N. J., Heneghan, C., Irving, F., Bacon, S., Fleminger, J., & Curtis, H. (2018). Compliance with requirement to report results on the EU Clinical Trials Register: Cohort study and web resource. *BMJ*, 362, k3218. <https://doi.org/10.1136/bmj.k3218>
- Good, I. J. (1992). The Bayes/Non-Bayes compromise: A brief review. *Journal of the American Statistical Association*, 87(419), 597–606. <https://doi.org/10.2307/2290192>
- Goodyear-Smith, F. A., van Driel, M. L., Arroll, B., & Del Mar, C. (2012). Analysis of

- decisions made in meta-analyses of depression screening and the risk of confirmation bias: A case study. *BMC Medical Research Methodology*, 12, 76. <https://doi.org/10.1186/1471-2288-12-76>
- Gopalakrishna, G., Riet, G. ter, Vink, G., Stoop, I., Wicherts, J. M., & Bouter, L. M. (2022). Prevalence of questionable research practices, research misconduct and their potential explanatory factors: A survey among academic researchers in The Netherlands. *PLOS ONE*, 17(2), e0263023. <https://doi.org/10.1371/journal.pone.0263023>
- Gosset, W. S. (1904). *The Application of the "Law of Error" to the Work of the Brewery* (1 vol 8; pp. 3–16). Arthur Guinness & Son, Ltd.
- Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, 7(4), 493–498. <https://doi.org/10.1111/2041-210X.12504>
- Green, S. B. (1991). How Many Subjects Does It Take To Do A Regression Analysis. *Multivariate Behavioral Research*, 26(3), 499–510. [https://doi.org/10.1207/s15327906mbr2603\\_7](https://doi.org/10.1207/s15327906mbr2603_7)
- Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, 82(1), 1–20. <http://psycnet.apa.org/journals/bul/82/1/1>
- Greenwald, A. G. (Ed.). (1976). An editorial. *Journal of Personality and Social Psychology*, 33(1), 1–7. <https://doi.org/10.1037/h0078635>
- Grünwald, P., de Heide, R., & Koolen, W. (2019). *Safe Testing*. <http://arxiv.org/abs/1906.07801>
- Gupta, S. K. (2011). Intention-to-treat concept: A review. *Perspectives in Clinical Research*, 2(3), 109–112. <https://doi.org/10.4103/2229-3485.83221>
- Hacking, I. (1965). *Logic of Statistical Inference*. Cambridge University Press.
- Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Angono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., Bruyneel, S., Calvillo, D. P., Campbell, W. K., Cannon, P. R., Carlucci, M., Carruth, N. P., Cheung, T., Crowell, A., De Ridder, D. T. D., Dewitte, S., ... Zwienenberg, M. (2016). A Multilab Preregistered Replication of the Ego-Depletion Effect. *Perspectives on Psychological Science*, 11(4), 546–573. <https://doi.org/10.1177/1745691616652873>
- Hallahan, M., & Rosenthal, R. (1996). Statistical power: Concepts, procedures, and applications. *Behaviour Research and Therapy*, 34(5), 489–499. [https://doi.org/10.1016/0005-7967\(95\)00082-8](https://doi.org/10.1016/0005-7967(95)00082-8)
- Hallinan, D., Boehm, F., Külpmann, A., & Elson, M. (2023). Information Provision for Informed Consent Procedures in Psychological Research Under the General Data Protection Regulation: A Practical Guide. *Advances in Methods and Practices in Psychological Science*, 6(1), 25152459231151944. <https://doi.org/10.1177/25152459231151944>
- Halpern, J., Brown Jr, B. W., & Hornberger, J. (2001). The sample size for a clinical trial: A Bayesian decision theoretic approach. *Statistics in Medicine*, 20(6), 841–858. <https://doi.org/10.1002/sim.703>
- Halpern, S. D., Karlawish, J. H., & Berlin, J. A. (2002). The continuing unethical conduct of underpowered clinical trials. *Jama*, 288(3), 358–362. <https://doi.org/doi:10.1001/jama.288.3.358>
- Hand, D. J. (1994). Deconstructing Statistical Questions. *Journal of the Royal Statistical So-*

- ciety. *Series A (Statistics in Society)*, 157(3), 317–356. <https://doi.org/10.2307/2983526>
- Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., Mohr, A. H., Clayton, E., Yoon, E. J., Tessler, M. H., Lenne, R. L., Altman, S., Long, B., & Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal *Cognition*. *Open Science*, 5(8), 180448. <https://doi.org/10.1098/rsos.180448>
- Harms, C., & Lakens, D. (2018). Making 'null effects' informative: Statistical techniques and inferential frameworks. *Journal of Clinical and Translational Research*, 3, 382–393. <https://doi.org/10.18053/jctres.03.2017S2.007>
- Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, D. D. (2021). *Doing Meta-Analysis with R: A Hands-On Guide*. Chapman and Hall/CRC. <https://doi.org/10.1201/9781003107347>
- Hauck, D. W. W., & Anderson, S. (1984). A new statistical procedure for testing equivalence in two-group comparative bioavailability trials. *Journal of Pharmacokinetics and Biopharmaceutics*, 12(1), 83–91. <https://doi.org/10.1007/BF01063612>
- Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods*, 6(3), 203–217. <https://doi.org/10.1037/1082-989X.6.3.203>
- Hempel, C. G. (1966). *Philosophy of natural science* (Nachdr.). Prentice-Hall.
- Hilgard, J. (2021). Maximal positive controls: A method for estimating the largest plausible effect size. *Journal of Experimental Social Psychology*, 93. <https://doi.org/10.1016/j.jesp.2020.104082>
- Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical Benchmarks for Interpreting Effect Sizes in Research. *Child Development Perspectives*, 2(3), 172–177. <https://doi.org/10.1111/j.1750-8606.2008.00061.x>
- Hodges, J. L., & Lehmann, E. L. (1954). Testing the Approximate Validity of Statistical Hypotheses. *Journal of the Royal Statistical Society. Series B (Methodological)*, 16(2), 261–268. <https://doi.org/10.1111/j.2517-6161.1954.tb00169.x>
- Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician*, 55(1), 19–24. <https://doi.org/10.1198/000313001300339897>
- Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I<sup>2</sup> index? *Psychological Methods*, 11(2), 193. <http://psycnet.apa.org/journals/met/11/2/193/>
- Hung, H. M. J., O'Neill, R. T., Bauer, P., & Kohne, K. (1997). The Behavior of the P-Value When the Alternative Hypothesis is True. *Biometrics*, 53(1), 11–22. <https://doi.org/10.2307/2533093>
- Hunt, K. (1975). Do we really need more replications? *Psychological Reports*, 36(2), 587–593.
- Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., & Williams, C. C. (2008). Gender Similarities Characterize Math Performance. *Science*, 321(5888), 494–495. <https://doi.org/10.1126/science.1160364>
- Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, 2(8), e124. <https://doi.org/10.1371/journal.pmed.0020124>
- Ioannidis, J. P. A., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials*, 4(3), 245–253. <https://doi.org/10.1177/1740774507079441>

- Isager, P. M., van Aert, R. C. M., Bahník, Š., Brandt, M. J., DeSoto, K. A., Giner-Sorolla, R., Krueger, J. I., Perugini, M., Ropovik, I., van 't Veer, A. E., Vranka, M., & Lakens, D. (2023). Deciding what to replicate: A decision model for replication study selection under resource and knowledge constraints. *Psychological Methods*, 28(2), 438–451. <https://doi.org/10.1037/met0000438>
- Iyengar, S., & Greenhouse, J. B. (1988). Selection Models and the File Drawer Problem. *Statistical Science*, 3(1), 109–117. <http://www.jstor.org/stable/2245925>
- Jaeschke, R., Singer, J., & Guyatt, G. H. (1989). Measurement of health status: Ascertaining the minimal clinically important difference. *Controlled Clinical Trials*, 10(4), 407–415. [https://doi.org/10.1016/0197-2456\(89\)90005-6](https://doi.org/10.1016/0197-2456(89)90005-6)
- Jeffreys, H. (1939). *Theory of probability* (1st ed). Oxford University Press.
- Jennison, C., & Turnbull, B. W. (2000). *Group sequential methods with applications to clinical trials*. Chapman & Hall/CRC.
- Johansson, T. (2011). Hail the impossible: P-values, evidence, and likelihood. *Scandinavian Journal of Psychology*, 52(2), 113–125. <https://doi.org/10.1111/j.1467-9450.2010.00852.x>
- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23(5), 524–532. <https://doi.org/10.1177/0956797611430953>
- Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences*, 110(48), 19313–19317. <https://doi.org/10.1073/pnas.1313476110>
- Jones, L. V. (1952). Test of hypotheses: One-sided vs. Two-sided alternatives. *Psychological Bulletin*, 49(1), 43–46. <https://doi.org/10.1037/h0056832>
- Jostmann, N. B., Lakens, D., & Schubert, T. W. (2009). Weight as an Embodiment of Importance. *Psychological Science*, 20(9), 1169–1174. <https://doi.org/10.1111/j.1467-9280.2009.02426.x>
- Jostmann, N. B., Lakens, D., & Schubert, T. W. (2016). A short history of the weight-importance effect and a recommendation for pre-testing: Commentary on Ebersole et al. (2016). *Journal of Experimental Social Psychology*, 67, 93–94. <https://doi.org/10.1016/j.jesp.2015.12.001>
- Julious, S. A. (2004). Sample sizes for clinical trials with normal data. *Statistics in Medicine*, 23(12), 1921–1986. <https://doi.org/10.1002/sim.1783>
- Junk, T., & Lyons, L. (2020). Reproducibility and Replication of Experimental Particle Physics Results. *Harvard Data Science Review*, 2(4). <https://doi.org/10.1162/99608f92.250f995b>
- Kaiser, H. F. (1960). Directional statistical decisions. *Psychological Review*, 67(3), 160–167. <https://doi.org/10.1037/h0047595>
- Kaplan, R. M., & Irvin, V. L. (2015). Likelihood of Null Effects of Large NHLBI Clinical Trials Has Increased over Time. *PLOS ONE*, 10(8), e0132382. <https://doi.org/10.1371/journal.pone.0132382>
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430), 773–795. <https://doi.org/10.1080/01621459.1995.10476572>
- Keefe, R. S. E., Kraemer, H. C., Epstein, R. S., Frank, E., Haynes, G., Laughren, T. P., McNulty, J., Reed, S. D., Sanchez, J., & Leon, A. C. (2013). Defining a Clinically Mean-

- ingful Effect for the Design and Interpretation of Randomized Controlled Trials. *Innovations in Clinical Neuroscience*, 10, 4S–19S. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3719483/>
- Kelley, K. (2007). Confidence Intervals for Standardized Effect Sizes: Theory, Application, and Implementation. *Journal of Statistical Software*, 20(8). <https://doi.org/10.18637/JSS.V020.I08>
- Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, 17(2), 137–152. <https://doi.org/10.1037/a0028086>
- Kelley, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods*, 11(4), 363–385. <https://doi.org/10.1037>
- Kelter, R. (2021). Analysis of type I and II error rates of Bayesian and frequentist parametric and nonparametric two-sample hypothesis tests under preliminary assessment of normality. *Computational Statistics*, 36(2), 1263–1288. <https://doi.org/10.1007/s00180-020-01034-7>
- Kenett, R. S., Shmueli, G., & Kenett, R. (2016). *Information Quality: The Potential of Data and Analytics to Generate Knowledge* (1st edition). Wiley.
- Kennedy-Shaffer, L. (2019). Before  $p < 0.05$  to Beyond  $p < 0.05$ : Using History to Contextualize p-Values and Significance Testing. *The American Statistician*, 73, 82–90. <https://doi.org/10.1080/00031305.2018.1537891>
- Kenny, D. A., & Judd, C. M. (2019). The unappreciated heterogeneity of effect sizes: Implications for power, precision, planning of research, and replication. *Psychological Methods*, 24(5), 578–589. <https://doi.org/10.1037/met0000209>
- Keppel, G. (1991). *Design and analysis: A researcher's handbook*, 3rd ed (pp. xiii, 594). Prentice-Hall, Inc.
- Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, 2(3), 196–217. [https://doi.org/10.1207/s15327957pspr0203\\_4](https://doi.org/10.1207/s15327957pspr0203_4)
- King, M. T. (2011). A point of minimal important difference (MID): A critique of terminology and methods. *Expert Review of Pharmacoeconomics & Outcomes Research*, 11(2), 171–184. <https://doi.org/10.1586/erp.11.9>
- Kish, L. (1959). Some Statistical Problems in Research Design. *American Sociological Review*, 24(3), 328–338. <https://doi.org/10.2307/2089381>
- Kish, L. (1965). *Survey Sampling*. Wiley.
- Komić, D., Marušić, S. L., & Marušić, A. (2015). Research Integrity and Research Ethics in Professional Codes of Ethics: Survey of Terminology Used by Professional Organizations across Research Disciplines. *PLOS ONE*, 10(7), e0133662. <https://doi.org/10.1371/journal.pone.0133662>
- Koole, S. L., & Lakens, D. (2012). Rewarding replications A sure and simple way to improve psychological science. *Perspectives on Psychological Science*, 7(6), 608–614. <https://doi.org/10.1177/1745691612462586>
- Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, 49(4), 241–253. <https://doi.org/10.3102/0013189X20912798>
- Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, 6(3), 299–312.

- Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, 142(2), 573–603. <https://doi.org/10.1037/a0029146>
- Kruschke, J. K. (2014). *Doing Bayesian Data Analysis, Second Edition: A Tutorial with R, JAGS, and Stan* (2 edition). Academic Press.
- Kruschke, J. K. (2018). Rejecting or Accepting Parameter Values in Bayesian Estimation. *Advances in Methods and Practices in Psychological Science*, 1(2), 270–280. <https://doi.org/10.1177/2515245918771304>
- Kruschke, J. K., & Liddell, T. M. (2017). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*. <https://doi.org/10.3758/s13423-016-1221-4>
- Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press.
- Kuipers, T. A. F. (2016). Models, postulates, and generalized nomic truth approximation. *Synthese*, 193(10), 3057–3077. <https://doi.org/10.1007/s11229-015-0916-9>
- Lakatos, I. (1978). *The methodology of scientific research programmes: Volume 1: Philosophical papers*. Cambridge University Press.
- Lakens, Daniël. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4. <https://doi.org/10.3389/fpsyg.2013.00863>
- Lakens, Daniël. (2014). Performing high-powered studies efficiently with sequential analyses: Sequential analyses. *European Journal of Social Psychology*, 44(7), 701–710. <https://doi.org/10.1002/ejsp.2023>
- Lakens, Daniël. (2017). Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science*, 8(4), 355–362. <https://doi.org/10.1177/1948550617697177>
- Lakens, Daniël. (2019). The value of preregistration for psychological science: A conceptual analysis. *Japanese Psychological Review*, 62(3), 221–230. [https://doi.org/10.24602/sjpr.62.3\\_221](https://doi.org/10.24602/sjpr.62.3_221)
- Lakens, Daniël. (2020). Pandemic researchers — recruit your own best critics. *Nature*, 581(7807, 7807), 121–121. <https://doi.org/10.1038/d41586-020-01392-8>
- Lakens, Daniël. (2021). The practical alternative to the p value is the correctly used p value. *Perspectives on Psychological Science*, 16(3), 639–648. <https://doi.org/10.1177/1745691620958012>
- Lakens, Daniël. (2022a). Sample Size Justification. *Collabra: Psychology*. <https://doi.org/10.31234/osf.io/9d3yf>
- Lakens, Daniël. (2022b). Why P values are not measures of evidence. *Trends in Ecology & Evolution*, 37(4), 289–290. <https://doi.org/10.1016/j.tree.2021.12.006>
- Lakens, Daniël. (2023). Is my study useless? Why researchers need methodological review boards. *Nature*, 613(7942, 7942), 9–9. <https://doi.org/10.1038/d41586-022-04504-8>
- Lakens, Daniel. (2023, December 18). *When and How to Deviate from a Preregistration*. <https://doi.org/10.31234/osf.io/ha29k>
- Lakens, Daniël, Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., Buchanan, E. M., Caldwell, A. R., Calster, B., Carlsson, R., Chen, S.-C., Chung, B., Colling, L. J., Collins, G. S.,

- Crook, Z., ... Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour*, 2, 168–171. <https://doi.org/10.1038/s41562-018-0311-x>
- Lakens, Daniël, & Caldwell, A. R. (2021). Simulation-Based Power Analysis for Factorial Analysis of Variance Designs. *Advances in Methods and Practices in Psychological Science*, 4(1). <https://doi.org/10.1177/2515245920951503>
- Lakens, Daniël, & DeBruine, L. (2020). *Improving Transparency, Falsifiability, and Rigour by Making Hypothesis Tests Machine Readable*. <https://doi.org/10.31234/osf.io/5xcd4>
- Lakens, Daniël, & Etz, A. J. (2017). Too True to be Bad: When Sets of Studies With Significant and Nonsignificant Findings Are Probably True. *Social Psychological and Personality Science*, 8(8), 875–881. <https://doi.org/10.1177/1948550617693058>
- Lakens, Daniël, Hilgard, J., & Staaks, J. (2016). On the reproducibility of meta-analyses: Six practical recommendations. *BMC Psychology*, 4, 24. <https://doi.org/10.1186/s40359-016-0126-3>
- Lakens, Daniël, McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2020). Improving Inferences About Null Effects With Bayes Factors and Equivalence Tests. *The Journals of Gerontology: Series B*, 75(1), 45–57. <https://doi.org/10.1093/geronb/gby065>
- Lakens, Daniël, Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2), 259–269. <https://doi.org/10.1177/2515245918770963>
- Lan, K. K. G., & DeMets, D. L. (1983). Discrete Sequential Boundaries for Clinical Trials. *Biometrika*, 70(3), 659. <https://doi.org/10.2307/2336502>
- Langmuir, I., & Hall, R. N. (1989). Pathological Science. *Physics Today*, 42(10), 36–48. <https://doi.org/10.1063/1.881205>
- Latan, H., Chiappetta Jabbour, C. J., Lopes de Sousa Jabbour, A. B., & Ali, M. (2021). Crossing the Red Line? Empirical Evidence and Useful Recommendations on Questionable Research Practices among Business Scholars. *Journal of Business Ethics*, 1–21. <https://doi.org/10.1007/s10551-021-04961-7>
- Laudan, L. (1981). *Science and Hypothesis*. Springer Netherlands. <https://doi.org/10.1007/978-94-015-7288-0>
- Laudan, L. (1986). *Science and Values: The Aims of Science and Their Role in Scientific Debate*.
- Lawrence, J. M., Meyerowitz-Katz, G., Heathers, J. A. J., Brown, N. J. L., & Sheldrick, K. A. (2021). The lesson of ivermectin: Meta-analyses based on summary data alone are inherently unreliable. *Nature Medicine*, 27(11, 11), 1853–1854. <https://doi.org/10.1038/s41591-021-01535-y>
- Leamer, E. E. (1978). *Specification Searches: Ad Hoc Inference with Nonexperimental Data* (1 edition). Wiley.
- Lehmann, E. L., & Romano, J. P. (2005). *Testing statistical hypotheses* (3rd ed). Springer.
- Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician*, 55(3), 187–193. <https://doi.org/10.1198/000313001317098149>
- Lenth, R. V. (2007). Post hoc power: Tables and commentary. *Iowa City: Department of Statistics and Actuarial Science, University of Iowa*. <https://pdfs.semanticscholar.org/fbf8/cab4b59e54c6a3ed39ba3656f35ef86c5ee3.pdf>

- Leon, A. C., Davis, L. L., & Kraemer, H. C. (2011). The Role and Interpretation of Pilot Studies in Clinical Research. *Journal of Psychiatric Research*, 45(5), 626–629. <https://doi.org/10.1016/j.jpsychires.2010.10.008>
- Letrud, K., & Hernes, S. (2019). Affirmative citation bias in scientific myth debunking: A three-in-one case study. *PLOS ONE*, 14(9), e0222213. <https://doi.org/10.1371/journal.pone.0222213>
- Leung, P. T. M., Macdonald, E. M., Stanbrook, M. B., Dhalla, I. A., & Juurlink, D. N. (2017). A 1980 Letter on the Risk of Opioid Addiction. *New England Journal of Medicine*, 376(22), 2194–2195. <https://doi.org/10.1056/NEJMcl700150>
- Levine, T. R., Weber, R., Park, H. S., & Hullett, C. R. (2008). A communication researchers' guide to null hypothesis significance testing and alternatives. *Human Communication Research*, 34(2), 188–209. <http://onlinelibrary.wiley.com/doi/10.1111/j.1468-2958.2008.00318.x/full>
- Leys, C., Delacre, M., Mora, Y. L., Lakens, D., & Ley, C. (2019). How to Classify, Detect, and Manage Univariate and Multivariate Outliers, With Emphasis on Pre-Registration. *International Review of Social Psychology*, 32(1), 5. <https://doi.org/10.5334/irsp.289>
- Linden, A. H., & Hönekopp, J. (2021). Heterogeneity of Research Results: A New Perspective From Which to Assess and Promote Progress in Psychological Science. *Perspectives on Psychological Science*, 16(2), 358–376. <https://doi.org/10.1177/1745691620964193>
- Lindley, D. V. (1957). A statistical paradox. *Biometrika*, 44(1/2), 187–192.
- Lindsay, D. S. (2015). Replication in Psychological Science. *Psychological Science*, 26(12), 1827–1832. <https://doi.org/10.1177/0956797615616374>
- Loevinger, J. (1968). The "information explosion.". *American Psychologist*, 23(6), 455–455. <https://doi.org/10.1037/h0020800>
- Longino, H. E. (1990). *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press. <https://books.google.com?id=S8fIbD19BisC>
- Louis, T. A., & Zeger, S. L. (2009). Effective communication of standard errors and confidence intervals. *Biostatistics*, 10(1), 1–2. <https://doi.org/10.1093/biostatistics/kxn014>
- Lovakov, A., & Agadullina, E. R. (2021). Empirically derived guidelines for effect size interpretation in social psychology. *European Journal of Social Psychology*, 51(3), 485–504. <https://doi.org/10.1002/ejsp.2752>
- Lubin, A. (1957). Replicability as a publication criterion. *American Psychologist*, 12, 519–520. <https://doi.org/10.1037/h0039746>
- Luttrell, A., Petty, R. E., & Xu, M. (2017). Replicating and fixing failed replications: The case of need for cognition and argument quality. *Journal of Experimental Social Psychology*, 69, 178–183. <https://doi.org/10.1016/j.jesp.2016.09.006>
- Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, 70, 151–159. <https://doi.org/10.1037/h0026141>
- Lyons, I. M., Nuerk, H.-C., & Ansari, D. (2015). Rethinking the implications of numerical ratio effects for understanding the development of representational precision and numerical processing across formats. *Journal of Experimental Psychology: General*, 144(5), 1021–1035. <https://doi.org/10.1037/xge0000094>
- MacCoun, R., & Perlmutter, S. (2015). Blind analysis: Hide results to seek the truth. *Nature*,

- 526(7572, 7572), 187–189. <https://doi.org/10.1038/526187a>
- Mack, R. W. (1951). The Need for Replication Research in Sociology. *American Sociological Review*, 16(1), 93–94. <https://doi.org/10.2307/2087978>
- Mahoney, M. J. (1979). Psychology of the scientist: An evaluative review. *Social Studies of Science*, 9(3), 349–375. <https://doi.org/10.1177/030631277900900304>
- Maier, M., & Lakens, D. (2022). Justify your alpha: A primer on two practical approaches. *Advances in Methods and Practices in Psychological Science*. <https://doi.org/10.31234/osf.io/ts4r6>
- Makel, M. C., Hodges, J., Cook, B. G., & Plucker, J. A. (2021). Both Questionable and Open Research Practices Are Prevalent in Education Research. *Educational Researcher*, 50(8), 493–504. <https://doi.org/10.3102/0013189X211001356>
- Marshall, B., Cardon, P., Poddar, A., & Fontenot, R. (2013). Does Sample Size Matter in Qualitative Research?: A Review of Qualitative Interviews in Research. *Journal of Computer Information Systems*, 54(1), 11–22. <https://doi.org/10.1080/08874417.2013.11645667>
- Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed). Lawrence Erlbaum Associates.
- Maxwell, S. E., Delaney, H. D., & Kelley, K. (2017). *Designing Experiments and Analyzing Data: A Model Comparison Perspective, Third Edition* (3 edition). Routledge.
- Maxwell, S. E., & Kelley, K. (2011). Ethics and sample size planning. In *Handbook of ethics in quantitative methodology* (pp. 179–204). Routledge.
- Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample Size Planning for Statistical Power and Accuracy in Parameter Estimation. *Annual Review of Psychology*, 59(1), 537–563. <https://doi.org/10.1146/annurev.psych.59.103006.093735>
- Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. University of Chicago Press.
- Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars*. Cambridge University Press.
- Mayo, D. G., & Spanos, A. (2011). Error statistics. *Philosophy of Statistics*, 7, 152–198. <http://books.google.com/books?hl=en&lr=&id=mPG5RupkTX0C&oi=fnd&pg=PA153&dq=%22%EF%AC%81t.+This+is+understandable+for+two+main+reasons:+%EF%AC%81rst,+it+is+not+the+job%22+%22What+is+the+role+of+probability+in+making+reliable%22+%22serve+both+to+direct+and+justify+the+use+of+formal+statistical+tools.+Next%22+&ots=w0BijrIuke&sig=hNF3O36y8VTq9WNt8k5vp9kCIM0>
- Mazzolari, R., Porcelli, S., Bishop, D. J., & Lakens, D. (2022). Myths and methodologies: The use of equivalence and non-inferiority tests for interventional studies in exercise physiology and sport science. *Experimental Physiology*, 107(3), 201–212. <https://doi.org/10.1113/EP090171>
- McCarthy, R. J., Skowronski, J. J., Verschuere, B., Meijer, E. H., Jim, A., Hoogesteyn, K., Orthey, R., Acar, O. A., Aczel, B., Bakos, B. E., Barbosa, F., Baskin, E., Bègue, L., Ben-Shakhar, G., Birt, A. R., Blatz, L., Charman, S. D., Claesen, A., Clay, S. L., ... Yıldız, E. (2018). Registered Replication Report on Srull and Wyer (1979). *Advances in Methods and Practices in Psychological Science*, 1(3), 321–336. <https://doi.org/10.1177>

2515245918777487

- McElreath, R. (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan* (Vol. 122). CRC Press. <http://books.google.com/books?hl=en&lr=&id=mDo0CwAAQBAJ&oi=fnd&pg=PP1&dq=%22Maximum+entropy%22+%2213.+Adventures+in%22+%22sure+that+planks+won%22%80%99t+slip+at+busy%22+%22Ordered+categorical%22+%2214.+Missing+Data+and+Other%22+%22Other+count%22+%22Continuous+categories+and+the+Gaussian%22+&ots=fD6h3iyM2m&sig=pEXBn5Ex-ps9Fhw-b6juTVKIngY>
- McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: The case of r and d. *Psychological Methods*, 11(4), 386–401. <https://doi.org/10.1037/1082-989X.11.4.386>
- McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, 111(2), 361–365. <https://doi.org/10.1037/0033-2909.111.2.361>
- McGuire, W. J. (2004). A Perspectivist Approach to Theory Construction. *Personality and Social Psychology Review*, 8(2), 173–182. [https://doi.org/10.1207/s15327957pspr0802\\_11](https://doi.org/10.1207/s15327957pspr0802_11)
- McIntosh, R. D., & Rittmo, J. Ö. (2021). Power calculations in single-case neuropsychology: A practical primer. *Cortex*, 135, 146–158. <https://doi.org/10.1016/j.cortex.2020.11.005>
- Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, 103–115. <http://www.jstor.org/stable/186099>
- Meehl, P. E. (1978). Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology. *Journal of Consulting and Clinical Psychology*, 46(4), 806–834. <https://doi.org/10.1037/0022-006X.46.4.806>
- Meehl, P. E. (1990a). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, 1(2), 108–141. [https://doi.org/10.1207/s15327965pli0102\\_1](https://doi.org/10.1207/s15327965pli0102_1)
- Meehl, P. E. (1990b). Why Summaries of Research on Psychological Theories are Often Uninterpretable: *Psychological Reports*, 66(1), 195–244. <https://doi.org/10.2466/pr0.1990.66.1.195>
- Meehl, P. E. (2004). Cliometric metatheory III: Peircean consensus, verisimilitude and asymptotic method. *The British Journal for the Philosophy of Science*, 55(4), 615–643.
- Melara, R. D., & Algom, D. (2003). Driven by information: A tectonic theory of Stroop effects. *Psychological Review*, 110(3), 422–471. <https://doi.org/10.1037/0033-295X.110.3.422>
- Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science*, 12(4), 269–275. <https://doi.org/10.1111/1467-9280.00350>
- Merton, R. K. (1942). A Note on Science and Democracy. *Journal of Legal and Political Sociology*, 1, 115–126. <https://heinonline.org/HOL/Page?handle=hein.journals/jolegpo1&id=115&div=&collection=>
- Meyners, M. (2012). Equivalence tests – A review. *Food Quality and Preference*, 26(2), 231–245. <https://doi.org/10.1016/j.foodqual.2012.05.003>
- Meyvis, T., & Van Osselaer, S. M. J. (2018). Increasing the Power of Your Study by Increasing the Effect Size. *Journal of Consumer Research*, 44(5), 1157–1173. <https://doi.org/10.1093/jcr/ucx110>
- Millar, R. B. (2011). *Maximum likelihood estimation and inference: With examples in R, SAS,*

- and ADMB.* Wiley.
- Miller, J. (2009). What is the probability of replicating a statistically significant effect? *Psychonomic Bulletin & Review*, 16(4), 617–640. <https://doi.org/10.3758/PBR.16.4.617>
- Miller, J., & Ulrich, R. (2019). The quest for an optimal alpha. *PLOS ONE*, 14(1), e0208631. <https://doi.org/10.1371/journal.pone.0208631>
- Mitroff, I. I. (1974). Norms and Counter-Norms in a Select Group of the Apollo Moon Scientists: A Case Study of the Ambivalence of Scientists. *American Sociological Review*, 39(4), 579–595. <https://doi.org/10.2307/2094423>
- Moe, K. (1984). Should the Nazi Research Data Be Cited? *The Hastings Center Report*, 14(6), 5–7. <https://doi.org/10.2307/3561733>
- Moran, C., Link to external site, this link will open in a new window, Richard, A., Link to external site, this link will open in a new window, Wilson, K., Twomey, R., Link to external site, this link will open in a new window, Coriou, A., & Link to external site, this link will open in a new window. (2022). I know it's bad, but I have been pressured into it: Questionable research practices among psychology students in Canada. *Canadian Psychology/Psychologie Canadienne*. <https://doi.org/10.1037/cap0000326>
- Morey, Richard D. (2020, June 12). *Power and precision* [Blog]. <https://medium.com/@richarddmorey/power-and-precision-47f644ddea5e>
- Morey, Richard D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, 23(1), 103–123. <http://link.springer.com/article/10.3758/s13423-015-0947-8>
- Morey, Richard D., Kaschak, M. P., Díez-Álamo, A. M., Glenberg, A. M., Zwaan, R. A., Lakens, D., Ibáñez, A., García, A., Gianelli, C., Jones, J. L., Madden, J., Alifano, F., Bergen, B., Bloxsom, N. G., Bub, D. N., Cai, Z. G., Chartier, C. R., Chatterjee, A., Conwell, E., ... Ziv-Crispel, N. (2021). A pre-registered, multi-lab non-replication of the action-sentence compatibility effect (ACE). *Psychonomic Bulletin & Review*. <https://doi.org/10.3758/s13423-021-01927-8>
- Morey, Richard D., & Lakens, D. (2016). *Why most of psychology is statistically unfalsifiable*. [https://raw.githubusercontent.com/richarddmorey/psychology\\_resolution/master/paper/response.pdf](https://raw.githubusercontent.com/richarddmorey/psychology_resolution/master/paper/response.pdf)
- Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11), 2074–2102. <https://doi.org/10.1002/sim.8086>
- Morse, J. M. (1995). The Significance of Saturation. *Qualitative Health Research*, 5(2), 147–149. <https://doi.org/10.1177/104973239500500201>
- Moscovici, S. (1972). Society and theory in social psychology. In *Context of social psychology* (pp. 17–81).
- Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., Grahe, J. E., McCarthy, R. J., Musser, E. D., & Antfolk, J. (2018). The Psychological Science Accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science*, 1(4), 501–515. <https://doi.org/10.1177/2515245918797607>
- Motyl, M., Demos, A. P., Carsel, T. S., Hanson, B. E., Melton, Z. J., Mueller, A. B., Prims,

- J. P., Sun, J., Washburn, A. N., Wong, K. M., Yantis, C., & Skitka, L. J. (2017). The state of social and personality science: Rotten to the core, not so bad, getting better, or getting worse? *Journal of Personality and Social Psychology*, 113, 34–58. <https://doi.org/10.1037/pspa0000084>
- Mrozek, J. R., & Taylor, L. O. (2002). What determines the value of life? A meta-analysis. *Journal of Policy Analysis and Management*, 21(2), 253–270. <https://doi.org/10.1002/pam.10026>
- Mudge, J. F., Baker, L. F., Edge, C. B., & Houlahan, J. E. (2012). Setting an Optimal That Minimizes Errors in Null Hypothesis Significance Tests. *PLOS ONE*, 7(2), e32734. <https://doi.org/10.1371/journal.pone.0032734>
- Mullan, F., & Jacoby, I. (1985). The town meeting for technology: The maturation of consensus conferences. *JAMA*, 254(8), 1068–1072. <https://doi.org/10.1001/jama.1985.03360080080035>
- Mulligan, A., Hall, L., & Raphael, E. (2013). Peer review in a changing world: An international study measuring the attitudes of researchers. *Journal of the American Society for Information Science and Technology*, 64(1), 132–161. <https://doi.org/10.1002/asi.22798>
- Murphy, K. R., & Myors, B. (1999). Testing the hypothesis that treatments have negligible effects: Minimum-effect tests in the general linear model. *Journal of Applied Psychology*, 84(2), 234–248. <https://doi.org/10.1037/0021-9010.84.2.234>
- Murphy, K. R., Myors, B., & Wolach, A. H. (2014). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (Fourth edition). Routledge, Taylor & Francis Group.
- National Academy of Sciences, National Academy of Engineering, & Institute of Medicine. (2009). *On being a scientist: A guide to responsible conduct in research: Third edition*. The National Academies Press. <https://doi.org/10.17226/12192>
- Neher, A. (1967). Probability Pyramiding, Research Error and the Need for Independent Replication. *The Psychological Record*, 17(2), 257–262. <https://doi.org/10.1007/BF03393713>
- Nemeth, C., Brown, K., & Rogers, J. (2001). Devil's advocate versus authentic dissent: Stimulating quantity and quality. *European Journal of Social Psychology*, 31(6), 707–720. <https://doi.org/10.1002/ejsp.58>
- Neyman, J. (1957). "Inductive Behavior" as a Basic Concept of Philosophy of Science. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, 25(1/3), 7. <https://doi.org/10.2307/1401671>
- Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 231(694-706), 289–337. <https://doi.org/10.1098/rsta.1933.0009>
- Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2), 175–220.
- Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, 5(2), 241–301. <https://doi.org/10.1037//1082-989X.5.2.241>

- Niiniluoto, I. (1998). Verisimilitude: The Third Period. *The British Journal for the Philosophy of Science*, 49, 1–29. <https://academic.oup.com/bjps/article-abstract/49/1/1/1448878/Verisimilitude-The-Third-Period>
- Niiniluoto, I. (1999). *Critical Scientific Realism*. Oxford University Press. [https://books.google.com?id=Ng\\_p\\_3XCHxAC](https://books.google.com?id=Ng_p_3XCHxAC)
- Norman, G. R., Sloan, J. A., & Wyrwich, K. W. (2004). The truly remarkable universality of half a standard deviation: Confirmation through another look. *Expert Review of Pharmacoeconomics & Outcomes Research*, 4(5), 581–585.
- Nosek, B. A., & Errington, T. M. (2020). What is replication? *PLOS Biology*, 18(3), e3000691. <https://doi.org/10.1371/journal.pbio.3000691>
- Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, 45(3), 137–141. <https://doi.org/10.1027/1864-9335/a000192>
- Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2015). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*. <https://doi.org/10.3758/s13428-015-0664-2>
- Nuijten, M. B., & Wicherts, J. (2023, January 31). *The effectiveness of implementing statcheck in the peer review process to avoid statistical reporting errors*. <https://doi.org/10.31234/osf.io/bxau9>
- Nunnally, J. (1960). The place of statistics in psychology. *Educational and Psychological Measurement*, 20(4), 641–650. <https://doi.org/10.1177/001316446002000401>
- O'Donnell, M., Nelson, L. D., Ackermann, E., Aczel, B., Akhtar, A., Aldrovandi, S., Alshaif, N., Andringa, R., Aveyard, M., Babincak, P., Balatekin, N., Baldwin, S. A., Banik, G., Baskin, E., Bell, R., Białobrzeska, O., Birt, A. R., Boot, W. R., Braithwaite, S. R., ... Zrubka, M. (2018). Registered Replication Report: Dijksterhuis and van Knippenberg (1998). *Perspectives on Psychological Science*, 13(2), 268–294. <https://doi.org/10.1177/1745691618755704>
- Obels, P., Lakens, D., Coles, N. A., Gottfried, J., & Green, S. A. (2020). Analysis of Open Data and Computational Reproducibility in Registered Reports in Psychology. *Advances in Methods and Practices in Psychological Science*, 3(2), 229–237. <https://doi.org/10.1177/2515245920918872>
- Oddie, G. (2013). The content, consequence and likeness approaches to verisimilitude: Compatibility, trivialization, and underdetermination. *Synthese*, 190(9), 1647–1687. <https://doi.org/10.1007/s11229-011-9930-8>
- Okada, K. (2013). Is Omega Squared Less Biased? A Comparison of Three Major Effect Size Indices in One-Way Anova. *Behaviormetrika*, 40(2), 129–147. <https://doi.org/10.2333/bhmk.40.129>
- Olejnik, S., & Algina, J. (2003). Generalized Eta and Omega Squared Statistics: Measures of Effect Size for Some Common Research Designs. *Psychological Methods*, 8(4), 434–447. <https://doi.org/10.1037/1082-989X.8.4.434>
- Olsson-Collentine, A., Wicherts, J. M., & van Assen, M. A. L. M. (2020). Heterogeneity in direct replications in psychology and its association with effect size. *Psychological Bulletin*, 146(10), 922–940. <https://doi.org/10.1037/bul0000294>

- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716–aac4716. <https://doi.org/10.1126/science.aac4716>
- Orben, A., & Lakens, D. (2020). Crud (Re)Defined. *Advances in Methods and Practices in Psychological Science*, 3(2), 238–247. <https://doi.org/10.1177/2515245920917961>
- Parker, R. A., & Berman, N. G. (2003). Sample Size. *The American Statistician*, 57(3), 166–170. <https://doi.org/10.1198/0003130031919>
- Parkhurst, D. F. (2001). Statistical significance tests: Equivalence and reverse tests should reduce misinterpretation. *Bioscience*, 51(12), 1051–1057. [https://doi.org/10.1641/0006-3568\(2001\)051%5B1051:SSTEAR%5D2.0.CO;2](https://doi.org/10.1641/0006-3568(2001)051%5B1051:SSTEAR%5D2.0.CO;2)
- Parsons, S., Kruijt, A.-W., & Fox, E. (2019). Psychological Science Needs a Standard Practice of Reporting the Reliability of Cognitive-Behavioral Measurements. *Advances in Methods and Practices in Psychological Science*, 2(4), 378–395. <https://doi.org/10.1177/2515245919879695>
- Pawitan, Y. (2001). *In all likelihood: Statistical modelling and inference using likelihood*. Clarendon Press ; Oxford University Press.
- Pemberton, M., Hall, S., Moskovitz, C., & Anson, C. M. (2019). Text recycling: Views of North American journal editors from an interview-based study. *Learned Publishing*, 32(4), 355–366. <https://doi.org/10.1002/leap.1259>
- Pereboom, A. C. (1971). Some Fundamental Problems in Experimental Psychology: An Overview. *Psychological Reports*, 28(2). <https://doi.org/10.2466/pr0.1971.28.2.439>
- Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *Bmj*, 316(7139), 1236–1238. <http://www.bmjjournals.org/content/316/7139/1236.short>
- Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science*, 9(3), 319–332. <https://doi.org/10.1177/1745691614528519>
- Perugini, M., Gallucci, M., & Costantini, G. (2018). A Practical Primer To Power Analysis for Simple Experimental Designs. *International Review of Social Psychology*, 31(1), 20. <https://doi.org/10.5334/irsp.181>
- Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2007). Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. *Statistics in Medicine*, 26(25), 4544–4562. <https://doi.org/10.1002/sim.2889>
- Phillips, B. M., Hunt, J. W., Anderson, B. S., Puckett, H. M., Fairey, R., Wilson, C. J., & Tjeerdema, R. (2001). Statistical significance of sediment toxicity test results: Threshold values derived by the detectable significance approach. *Environmental Toxicology and Chemistry*, 20(2), 371–373. <https://doi.org/10.1002/etc.5620200218>
- Pickett, J. T., & Roche, S. P. (2017). Questionable, Objectionable or Criminal? Public Opinion on Data Fraud and Selective Reporting in Science. *Science and Engineering Ethics*, 1–21. <https://doi.org/10.1007/s11948-017-9886-2>
- Platt, J. R. (1964). Strong Inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science*, 146(3642), 347–353. <https://doi.org/10.1126/science.146.3642.347>
- Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2), 191–199. <https://doi.org/10.1093/biomet/64.2.191>

- Polanin, J. R., Hennessy, E. A., & Tsuji, S. (2020). Transparency and Reproducibility of Meta-Analyses in Psychology: A Meta-Review. *Perspectives on Psychological Science*, 15(4), 1026–1041. <https://doi.org/10.1177/1745691620906416>
- Popper, K. R. (2002). *The logic of scientific discovery*. Routledge.
- Primbs, M., Pennington, C. R., Lakens, D., Silan, M. A., Lieck, D. S. N., Forscher, P., Buchanan, E. M., & Westwood, S. J. (2022). Are Small Effects the Indispensable Foundation for a Cumulative Psychological Science? A Reply to Götz et al. (2022). *Perspectives on Psychological Science*. <https://doi.org/10.31234/osf.io/6s8bj>
- Proschan, M. A. (2005). Two-Stage Sample Size Re-Estimation Based on a Nuisance Parameter: A Review. *Journal of Biopharmaceutical Statistics*, 15(4), 559–574. <https://doi.org/10.1081/BIP-200062852>
- Proschan, M. A., Lan, K. K. G., & Wittes, J. T. (2006). *Statistical monitoring of clinical trials: A unified approach*. Springer.
- Psillos, S. (1999). *Scientific realism: How science tracks truth*. Routledge.
- Quertemont, E. (2011). How to Statistically Show the Absence of an Effect. *Psychologica Belgica*, 51(2), 109–127. <https://doi.org/10.5334/pb-51-2-109>
- Rabelo, A. L. A., Farias, J. E. M., Sarmet, M. M., Joaquim, T. C. R., Hoersting, R. C., Victorino, L., Modesto, J. G. N., & Pilati, R. (2020). Questionable research practices among Brazilian psychological researchers: Results from a replication study and an international comparison. *International Journal of Psychology*, 55(4), 674–683. <https://doi.org/10.1002/ijop.12632>
- Radick, G. (2022). Mendel the fraud? A social history of truth in genetics. *Studies in History and Philosophy of Science*, 93, 39–46. <https://doi.org/10.1016/j.shpsa.2021.12.012>
- Reif, F. (1961). The Competitive World of the Pure Scientist. *Science*, 134(3494), 1957–1962. <https://doi.org/10.1126/science.134.3494.1957>
- Rice, W. R., & Gaines, S. D. (1994). 'Heads I win, tails you lose': Testing directional alternative hypotheses in ecological and evolutionary research. *Trends in Ecology & Evolution*, 9(6), 235–237. [https://doi.org/10.1016/0169-5347\(94\)90258-5](https://doi.org/10.1016/0169-5347(94)90258-5)
- Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One Hundred Years of Social Psychology Quantitatively Described. *Review of General Psychology*, 7(4), 331–363. <https://doi.org/10.1037/1089-2680.7.4.331>
- Richardson, J. T. E. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review*, 6(2), 135–147. <https://doi.org/10.1016/j.edurev.2010.12.001>
- Rijnsoever, F. J. van. (2017). (I Can't Get No) Saturation: A simulation and guidelines for sample sizes in qualitative research. *PLOS ONE*, 12(7), e0181689. <https://doi.org/10.1371/journal.pone.0181689>
- Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, 113(3), 553–565. <https://doi.org/http://dx.doi.org/10.1037/0033-2909.113.3.553>
- Rogers, S. (1992–1993). How a publicity blitz created the myth of subliminal advertising. *Public Relations Quarterly*, 37(4), 12. <https://www.proquest.com/docview/222493951/abstract/99E0E14044C24A96PQ/1>

- Ropovik, I., Adamkovic, M., & Greger, D. (2021). Neglect of publication bias compromises meta-analyses of educational research. *PLOS ONE*, 16(6), e0252415. <https://doi.org/10.1371/journal.pone.0252415>
- Rosenthal, R. (1966). *Experimenter effects in behavioral research*. Appleton-Century-Crofts.
- Rosnow, R. L., & Rosenthal, R. (2009). Effect Sizes: Why, When, and How to Use Them. *Zeitschrift Für Psychologie / Journal of Psychology*, 217(1), 6–14. <https://doi.org/10.1027/0044-3409.217.1.6>
- Ross-Hellauer, T., Deppe, A., & Schmidt, B. (2017). Survey on open peer review: Attitudes and experience amongst editors, authors and reviewers. *PLOS ONE*, 12(12), e0189311. <https://doi.org/10.1371/journal.pone.0189311>
- Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, 21(2), 301–308. <http://link.springer.com/article/10.3758/s13423-014-0595-4>
- Rouder, J. N., Haaf, J. M., & Snyder, H. K. (2019). Minimizing Mistakes in Psychological Science. *Advances in Methods and Practices in Psychological Science*, 2(1), 3–11. <https://doi.org/10.1177/2515245918801915>
- Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16(2), 225–237. <https://doi.org/10.3758/PBR.16.2.225>
- Royall, R. (1997). *Statistical Evidence: A Likelihood Paradigm*. Chapman and Hall/CRC.
- Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, 57(5), 416–428. <https://doi.org/10.1037/h0042040>
- Rücker, G., Schwarzer, G., Carpenter, J. R., & Schumacher, M. (2008). Undue reliance on I(2) in assessing heterogeneity may mislead. *BMC Medical Research Methodology*, 8, 79. <https://doi.org/10.1186/1471-2288-8-79>
- Samelson, F. (1980). J B Watson's Little Albert, Cyril Burt's twins, and the need for a critical science. *American Psychologist*, 35(7), 619–625. <https://doi.org/10.1037/0003-066X.35.7.619>
- Sarafoglou, A., Kovacs, M., Bakos, B., Wagenmakers, E.-J., & Aczel, B. (2022). A survey on how preregistration affects the research workflow: Better science but more work. *Royal Society Open Science*, 9(7), 211997. <https://doi.org/10.1098/rsos.211997>
- Scheel, A. M., Schijen, M. R. M. J., & Lakens, D. (2021). An Excess of Positive Results: Comparing the Standard Psychology Literature With Registered Reports. *Advances in Methods and Practices in Psychological Science*, 4(2), 25152459211007467. <https://doi.org/10.1177/25152459211007467>
- Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2021). Why Hypothesis Testers Should Spend Less Time Testing Hypotheses. *Perspectives on Psychological Science*, 16(4), 744–755. <https://doi.org/10.1177/1745691620966795>
- Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, 17(4), 551–566. <https://doi.org/10.1037/a0029487>
- Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, 13(2), 90–100. <https://doi.org/10.1037/a0015108>
- Schnuerch, M., & Erdfelder, E. (2020). Controlling decision errors with minimal costs: The

- sequential probability ratio t test. *Psychological Methods*, 25(2), 206–226. <https://doi.org/10.1037/met0000234>
- Schoemann, A. M., Boulton, A. J., & Short, S. D. (2017). Determining Power and Sample Size for Simple and Complex Mediation Models. *Social Psychological and Personality Science*, 8(4), 379–386. <https://doi.org/10.1177/1948550617715068>
- Schoenegger, P., & Pils, R. (2023). Social sciences in crisis: On the proposed elimination of the discussion section. *Synthese*, 202(2), 54. <https://doi.org/10.1007/s11229-023-04267-3>
- Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, 22(2), 322–339. <https://doi.org/10.1037/MET0000061>
- Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6), 657–680. <http://link.springer.com/article/10.1007/BF01068419>
- Schulz, K. F., & Grimes, D. A. (2005). Sample size calculations in randomised trials: Mandatory and mystical. *The Lancet*, 365(9467), 1348–1353. [https://doi.org/10.1016/S0140-6736\(05\)61034-3](https://doi.org/10.1016/S0140-6736(05)61034-3)
- Schumi, J., & Wittes, J. T. (2011). Through the looking glass: Understanding non-inferiority. *Trials*, 12(1), 106. <https://doi.org/10.1186/1745-6215-12-106>
- Schweder, T., & Hjort, N. L. (2016). *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*. Cambridge University Press. <https://doi.org/10.1017/CBO9781139046671>
- Scull, A. (2023). Rosenhan revisited: Successful scientific fraud. *History of Psychiatry*, 0957154X221150878. <https://doi.org/10.1177/0957154X221150878>
- Seaman, M. A., & Serlin, R. C. (1998). Equivalence confidence intervals for two-group comparisons of means. *Psychological Methods*, 3(4), 403–411. <https://doi.org/http://dx.doi.org.dianus.libr.tue.nl/10.1037/1082-989X.3.4.403>
- Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, 105(2), 309–316. <https://doi.org/10.1037/0033-295X.105.2.309>
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.
- Shafer, G. (1976). *A mathematical theory of evidence*. Princeton University Press.
- Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3), 289–310.
- Sidman, M. (1960). *Tactics of Scientific Research: Evaluating Experimental Data in Psychology* (New edition). Cambridge Center for Behavioral.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, 22(11), 1359–1366. <https://doi.org/10.1177/0956797611417632>
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2013, January 17–19). *Life after P-Hacking*. Meeting of the Society for Personality and Social Psychology, New Orleans, LA. [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2205186](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2205186)

- Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on Generality (COG): A Proposed Addition to All Empirical Papers. *Perspectives on Psychological Science*, 12(6), 1123–1128. <https://doi.org/10.1177/1745691617708630>
- Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, 26(5), 559–569. <https://doi.org/10.1177/0956797614567341>
- Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, 143(2), 534. <http://psycnet.apa.org/journals/xge/143/2/534/>
- Smart, R. G. (1964). The importance of negative results in psychological research. *Canadian Psychologist / Psychologie Canadienne*, 5a(4), 225–232. <https://doi.org/10.1037/h0083036>
- Smith, N. C. (1970). Replication studies: A neglected aspect of psychological research. *American Psychologist*, 25(10), 970–975. <https://doi.org/10.1037/h0029774>
- Smithson, M. (2003). *Confidence intervals*. Sage Publications.
- Sotola, L. K. (2022). Garbage In, Garbage Out? Evaluating the Evidentiary Value of Published Meta-analyses Using Z-Curve Analysis. *Collabra: Psychology*, 8(1), 32571. <https://doi.org/10.1525/collabra.32571>
- Spanos, A. (1999). *Probability theory and statistical inference: Econometric modeling with observational data*. Cambridge University Press.
- Spanos, A. (2013). Who should be afraid of the Jeffreys-Lindley paradox? *Philosophy of Science*, 80(1), 73–93. <https://doi.org/10.1086/668875>
- Spellman, B. A. (2015). A Short (Personal) Future History of Revolution 2.0. *Perspectives on Psychological Science*, 10(6), 886–899. <https://doi.org/10.1177/1745691615609918>
- Spence, J. R., & Stanley, D. J. (2024). Tempered Expectations: A Tutorial for Calculating and Interpreting Prediction Intervals in the Context of Replications. *Advances in Methods and Practices in Psychological Science*, 7(1), 25152459231217932. <https://doi.org/10.1177/25152459231217932>
- Spiegelhalter, D. (2019). *The Art of Statistics: How to Learn from Data* (Illustrated edition). Basic Books.
- Spiegelhalter, D. J., Freedman, L. S., & Blackburn, P. R. (1986). Monitoring clinical trials: Conditional or predictive power? *Controlled Clinical Trials*, 7(1), 8–17. [https://doi.org/10.1016/0197-2456\(86\)90003-6](https://doi.org/10.1016/0197-2456(86)90003-6)
- Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, 5(1), 60–78. <https://doi.org/10.1002/jrsm.1095>
- Stanley, T. D., Doucouliagos, H., & Ioannidis, J. P. A. (2017). Finding the power to reduce publication bias: Finding the power to reduce publication bias. *Statistics in Medicine*. <https://doi.org/10.1002/sim.7228>
- Steiger, J. H. (2004). Beyond the F Test: Effect Size Confidence Intervals and Tests of Close Fit in the Analysis of Variance and Contrast Analysis. *Psychological Methods*, 9(2), 164–182. <https://doi.org/10.1037/1082-989X.9.2.164>
- Sterling, T. D. (1959). Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance—Or Vice Versa. *Journal of the American Statistical Association*, 54(285), 30–34. <https://doi.org/10.2307/2282137>

- Stewart, L. A., & Tierney, J. F. (2002). To IPD or not to IPD?: Advantages and Disadvantages of Systematic Reviews Using Individual Patient Data. *Evaluation & the Health Professions*, 25(1), 76–97. <https://doi.org/10.1177/0163278702025001006>
- Stodden, V., Seiler, J., & Ma, Z. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences*, 115(11), 2584–2589. <https://doi.org/10.1073/pnas.1708290115>
- Strand, J. F. (2023). Error tight: Exercises for lab groups to prevent research mistakes. *Psychological Methods*, No Pagination Specified–No Pagination Specified. <https://doi.org/10.1037/met0000547>
- Stroebe, W., & Strack, F. (2014). The Alleged Crisis and the Illusion of Exact Replication. *Perspectives on Psychological Science*, 9(1), 59–71. <https://doi.org/10.1177/1745691613514450>
- Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18(6), 643–662.
- Swift, J. K., Link to external site, this link will open in a new window, Christopherson, C. D., Link to external site, this link will open in a new window, Bird, M. O., Link to external site, this link will open in a new window, Zöld, A., Link to external site, this link will open in a new window, Goode, J., & Link to external site, this link will open in a new window. (2022). Questionable research practices among faculty and students in APA-accredited clinical and counseling psychology doctoral programs. *Training and Education in Professional Psychology*, 16(3), 299–305. <https://doi.org/10.1037/tep0000322>
- Taper, M. L., & Lele, S. R. (2011). Philosophy of Statistics. In P. S. Bandyopadhyay & M. R. Forster (Eds.), *Evidence, evidence functions, and error probabilities* (pp. 513–531). Elsevier, USA.
- Taylor, D. J., & Muller, K. E. (1996). Bias in linear model power and sample size calculation due to estimating noncentrality. *Communications in Statistics-Theory and Methods*, 25(7), 1595–1610. <https://doi.org/10.1080/03610929608831787>
- Teare, M. D., Dimairo, M., Shephard, N., Hayman, A., Whitehead, A., & Walters, S. J. (2014). Sample size requirements to estimate key design parameters from external pilot randomised controlled trials: A simulation study. *Trials*, 15(1), 264. <https://doi.org/10.1186/1745-6215-15-264>
- Tendeiro, J. N., & Kiers, H. A. L. (2019). A review of issues about null hypothesis Bayesian testing. *Psychological Methods*. <https://doi.org/10.1037/met0000221>
- Tendeiro, J. N., Kiers, H. A. L., Hoekstra, R., Wong, T. K., & Morey, R. D. (2024). Diagnosing the Misuse of the Bayes Factor in Applied Research. *Advances in Methods and Practices in Psychological Science*, 7(1), 25152459231213371. <https://doi.org/10.1177/25152459231213371>
- ter Schure, J., & Grünwald, P. D. (2019). *Accumulation Bias in Meta-Analysis: The Need to Consider Time in Error Control*. <http://arxiv.org/abs/1905.13494>
- Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine*, 22(13), 2113–2126. <https://doi.org/10.1002/sim.1461>
- Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect

- sizes. *Psychology in the Schools*, 44(5), 423–432. <https://doi.org/10.1002/pits.20234>
- Tunç, D. U., & Tunç, M. N. (2023). A Falsificationist Treatment of Auxiliary Hypotheses in Social and Behavioral Sciences: Systematic Replications Framework. *Meta-Psychology*, 7. <https://doi.org/10.15626/MP.2021.2756>
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(4), 327–352. <https://doi.org/10.1037/0033-295X.84.4.327>
- Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, 76(2), 105–110. <https://doi.org/10.1037/h0031322>
- Ulrich, R., & Miller, J. (2018). Some properties of p-curves, with an application to gradual publication bias. *Psychological Methods*, 23(3), 546–560. <https://doi.org/10.1037/met0000125>
- Uygun Tunç, D., & Tunç, M. N. (2022). A Falsificationist Treatment of Auxiliary Hypotheses in Social and Behavioral Sciences: Systematic Replications Framework. *Meta-Psychology*. <https://doi.org/10.31234/osf.io/pdm7y>
- Uygun Tunç, D., Tunç, M. N., & Lakens, D. (2023). The epistemic and pragmatic function of dichotomous claims based on statistical hypothesis tests. *Theory & Psychology*, 33(3), 403–423. <https://doi.org/10.1177/09593543231160112>
- Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How Many Studies Do You Need?: A Primer on Statistical Power for Meta-Analysis. *Journal of Educational and Behavioral Statistics*, 35(2), 215–247. <https://doi.org/10.3102/1076998609346961>
- van de Schoot, R., Winter, S. D., Griffioen, E., Grimmelikhuijsen, S., Arts, I., Veen, D., Grandfield, E. M., & Tummers, L. G. (2021). The Use of Questionable Research Practices to Survive in Academia Examined With Expert Elicitation, Prior-Data Conflicts, Bayes Factors for Replication Effects, and the Bayes Truth Serum. *Frontiers in Psychology*, 12. <https://www.frontiersin.org/article/10.3389/fpsyg.2021.621547>
- van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods*, 22(2), 217–239. <https://doi.org/10.1037/met0000100>
- Van Fraassen, B. C. (1980). *The scientific image*. Clarendon Press ; Oxford University Press.
- van 't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology*, 67, 2–12. <https://doi.org/10.1016/j.jesp.2016.03.004>
- Varkey, B. (2021). Principles of Clinical Ethics and Their Application to Practice. *Medical Principles and Practice: International Journal of the Kuwait University, Health Science Centre*, 30(1), 17–28. <https://doi.org/10.1159/000509119>
- Vazire, S. (2017). Quality Uncertainty Erodes Trust in Science. *Collabra: Psychology*, 3(1), 1. <https://doi.org/10.1525/collabra.74>
- Vazire, S., & Holcombe, A. O. (2022). Where Are the Self-Correcting Mechanisms in Science? *Review of General Psychology*, 26(2), 212–223. <https://doi.org/10.1177/10892680211033912>
- Verschueren, B., Meijer, E. H., Jim, A., Hoogesteyn, K., Orthey, R., McCarthy, R. J., Skowronski, J. J., Acar, O. A., Aczel, B., Bakos, B. E., Barbosa, F., Baskin, E., Bègue, L., Ben-Shakhar, G., Birt, A. R., Blatz, L., Charman, S. D., Claesen, A., Clay, S. L., ... Yıldız, E. (2018). Registered Replication Report on Mazar, Amir, and Ariely (2008). *Advances in*

- Methods and Practices in Psychological Science*, 1(3), 299–317. <https://doi.org/10.1177/2515245918781032>
- Viamonte, S. M., Ball, K. K., & Kilgore, M. (2006). A Cost-Benefit Analysis of Risk-Reduction Strategies Targeted at Older Drivers. *Traffic Injury Prevention*, 7(4), 352–359. <https://doi.org/10.1080/15389580600791362>
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *J Stat Softw*, 36(3), 1–48. <http://dx.doi.org/10.18637/jss.v036.i03>
- Vohs, K. D., Schmeichel, B. J., Lohmann, S., Gronau, Q. F., Finley, A. J., Ainsworth, S. E., Alquist, J. L., Baker, M. D., Brizi, A., Bunyi, A., Butschek, G. J., Campbell, C., Capaldi, J., Cau, C., Chambers, H., Chatzisarantis, N. L. D., Christensen, W. J., Clay, S. L., Curtis, J., ... Albarracín, D. (2021). A Multisite Preregistered Paradigmatic Test of the Ego-Depletion Effect. *Psychological Science*, 32(10), 1566–1581. <https://doi.org/10.1177/0956797621989733>
- Vosgerau, J., Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2019). 99% impossible: A valid, or falsifiable, internal meta-analysis. *Journal of Experimental Psychology. General*, 148(9), 1628–1639. <https://doi.org/10.1037/xge0000663>
- Vuorre, M., & Curley, J. P. (2018). Curating Research Assets: A Tutorial on the Git Version Control System. *Advances in Methods and Practices in Psychological Science*, 1(2), 219–236. <https://doi.org/10.1177/2515245918754826>
- Wacholder, S., Chanock, S., Garcia-Closas, M., El ghormli, L., & Rothman, N. (2004). Assessing the Probability That a Positive Report is False: An Approach for Molecular Epidemiology Studies. *JNCI Journal of the National Cancer Institute*, 96(6), 434–442. <https://doi.org/10.1093/jnci/djh075>
- Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, 14(5), 779–804. <https://doi.org/10.3758/BF03194105>
- Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Albohn, D. N., Allard, E. S., Benning, S. D., Blouin-Hudon, E.-M., Bulnes, L. C., Caldwell, T. L., Calin-Jageman, R. J., Capaldi, C. A., Carfagno, N. S., Chasten, K. T., Cleeremans, A., Connell, L., DeCicco, J. M., ... Zwaan, R. A. (2016). Registered Replication Report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, 11(6), 917–928. <https://doi.org/10.1177/1745691616674458>
- Wagenmakers, E.-J., Wetzel, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, 100(3), 426–432. <https://doi.org/10.1037/a0022790>
- Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2), 117–186. <https://www.jstor.org/stable/2240273>
- Waldron, S., & Allen, C. (2022). Not all pre-registrations are equal. *Neuropsychopharmacology*, 47(13, 13), 2181–2183. <https://doi.org/10.1038/s41386-022-01418-x>
- Wang, B., Zhou, Z., Wang, H., Tu, X. M., & Feng, C. (2019). The p-value and model specification in statistics. *General Psychiatry*, 32(3), e100081. <https://doi.org/10.1136/gpsych-2019-100081>
- Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quar-*

- terly Journal of Experimental Psychology, 12(3), 129–140. <https://doi.org/10.1080/17470216008416717>
- Wassmer, G., & Brannath, W. (2016). *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*. Springer International Publishing. <https://doi.org/10.1007/978-3-319-32562-0>
- Weinshall-Margel, K., & Shapard, J. (2011). Overlooked factors in the analysis of parole decisions. *Proceedings of the National Academy of Sciences*, 108(42), E833–E833. <https://doi.org/10.1073/pnas.1110910108>
- Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority* (2nd ed). CRC Press.
- Westberg, M. (1985). Combining Independent Statistical Tests. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 34(3), 287–296. <https://doi.org/10.2307/2987655>
- Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, 143(5), 2020–2045. <https://doi.org/10.1037/xge0000014>
- Westlake, W. J. (1972). Use of Confidence Intervals in Analysis of Comparative Bioavailability Trials. *Journal of Pharmaceutical Sciences*, 61(8), 1340–1341. <https://doi.org/10.1002/JPS.2600610845>
- Whitney, S. N. (2016). *Balanced Ethics Review*. Springer International Publishing. <https://doi.org/10.1007/978-3-319-20705-6>
- Wicherts, J. M. (2011). Psychology must learn a lesson from fraud case. *Nature*, 480(7375, 7375), 7–7. <https://doi.org/10.1038/480007a>
- Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., Aert, V., M, R. C., Assen, V., & M, M. A. L. (2016). Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. *Frontiers in Psychology*, 7. <https://doi.org/10.3389/fpsyg.2016.01832>
- Wiebels, K., & Moreau, D. (2021). Leveraging Containers for Reproducible Psychological Research. *Advances in Methods and Practices in Psychological Science*, 4(2), 25152459211017853. <https://doi.org/10.1177/25152459211017853>
- Wigboldus, D. H. J., & Dotsch, R. (2016). Encourage Playing with Data and Discourage Questionable Reporting Practices. *Psychometrika*, 81(1), 27–32. <https://doi.org/10.1007/s11336-015-9445-1>
- Williams, R. H., Zimmerman, D. W., & Zumbo, B. D. (1995). Impact of Measurement Error on Statistical Power: Review of an Old Paradox. *The Journal of Experimental Education*, 63(4), 363–370. <https://doi.org/10.1080/00220973.1995.9943470>
- Wilson, E. C. F. (2015). A Practical Guide to Value of Information Analysis. *PharmacoEconomics*, 33(2), 105–121. <https://doi.org/10.1007/s40273-014-0219-x>
- Wilson VanVoorhis, C. R., & Morgan, B. L. (2007). Understanding power and rules of thumb for determining sample sizes. *Tutorials in Quantitative Methods for Psychology*, 3(2), 43–50. <https://doi.org/10.20982/tqmp.03.2.p043>
- Winer, B. J. (1962). *Statistical principles in experimental design*. New York : McGraw-Hill. <https://trove.nla.gov.au/version/39914160>
- Wingen, T., Berkessel, J. B., & Englich, B. (2020). No Replication, No Trust? How Low

- Replicability Influences Trust in Psychology. *Social Psychological and Personality Science*, 11(4), 454–463. <https://doi.org/10.1177/1948550619877412>
- Wiseman, R., Watt, C., & Kornbrot, D. (2019). Registered reports: An early example and analysis. *PeerJ*, 7, e6232. <https://doi.org/10.7717/peerj.6232>
- Wittes, J., & Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine*, 9(1-2), 65–72. <https://doi.org/10.1002/sim.4780090113>
- Wong, T. K., Kiers, H., & Tendeiro, J. (2022). On the Potential Mismatch Between the Function of the Bayes Factor and Researchers' Expectations. *Collabra: Psychology*, 8(1), 36357. <https://doi.org/10.1525/collabra.36357>
- Wynants, L., Calster, B. V., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., Bonten, M. M. J., Dahly, D. L., Damen, J. A., Debray, T. P. A., Jong, V. M. T. de, Vos, M. D., Dhiman, P., Haller, M. C., Harhay, M. O., Henckaerts, L., Heus, P., Kammer, M., Kreuzberger, N., ... Smeden, M. van. (2020). Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. *BMJ*, 369, m1328. <https://doi.org/10.1136/bmj.m1328>
- Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science*, 12(6), 1100–1122. <https://doi.org/10.1177/1745691617693393>
- Yuan, K.-H., & Maxwell, S. (2005). On the Post Hoc Power in Testing Mean Differences. *Journal of Educational and Behavioral Statistics*, 30(2), 141–167. <https://doi.org/10.3102/10769986030002141>
- Zabell, S. L. (1992). R. A. Fisher and Fiducial Argument. *Statistical Science*, 7(3), 369–387. <https://doi.org/10.1214/ss/1177011233>
- Zenko, M. (2015). *Red Team: How to Succeed By Thinking Like the Enemy* (1st edition). Basic Books.
- Zumbo, B. D., & Hubley, A. M. (1998). A note on misconceptions concerning prospective and retrospective power. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(2), 385–388. <https://doi.org/10.1111/1467-9884.00139>

# Change Log

The current version of this textbook is 1.5.1

This version has been compiled on June 28, 2024.

This version was generated from Git commit #c852c7b4. All version controlled changes can be found on [GitHub](#).

This page documents the changes to the textbook that were more substantial than fixing a typo.

## **17.6.0.0.1 \* Updates**

*June 20, 2024:*

Added CH17

*March 22, 2024:*

Updated CI figure in CH7, incorporated feedback LeFoll, updated references

*February 21, 2024:*

Added section on subgroup analyses, updates section on heterogeneity in CH 11

*February 15, 2024:*

Expanded section on “Why Effect Sizes Selected for Significance are Inflated” in CH 6

*January 10, 2024:*

Added the section ‘Deviating from a Preregistration’ in CH 13

*October 15, 2023:*

Incorporated extensive edits by Nick Brown in CH 4-6.

*September 6, 2023:*

Incorporated extensive edits by Nick Brown in CH 1-3.

*August 27, 2023:*

Add CH 16 on confirmation bias and organized skepticism. Add Bakan 1967 quote to CH 13.

*August 12, 2023:*

Added section on why standardized effect sizes hinder the interpretation of effect sizes in CH 6. Added Spanos 1999 to CH 1. Split up the correct interpretation of  $p$  values for significant and non-significant results CH 1. Added new Statcheck study CH 12. Added Platt quote CH 5.

*July 21, 2023:*

Added “Why Effect Sizes Selected for Significance are Inflated” section to CH 6, moved main part of “The Minimal Statistically Detectable Effect” from CH 8 to CH 6, replaced Greek characters by latex, added sentence bias is expected for papers that depend on main hypothesis test in CH 12.

*July 13, 2023:*

Updated Open Questions in CH 1, 2, 3, 4, 6, 7, 8 and 9. Added a figure illustrating how confidence intervals become more narrow as N increases in CH 7.

*July 7, 2023:*

Added this change log page.

*June 12, 2023:*

Added an updated figure from Carter & McCullough, 2014, in the chapter in bias detection, now generated from the raw data.

*May 5, 2023:*

Added the option to download a PDF and epub version of the book.

*March 19, 2023:*

Updated CH 5 with new sections on falsification, severity, and risky predictions, and a new final section on verisimilitude.

*March 3, 2023:*

Updated book to Quarto. Added webexercises to all chapters.

*February 27, 2023:*

Added a section “Dealing with Inconsistencies in Science” to CH 5.

*October 4, 2022:*

Added CH 15 on research integrity.