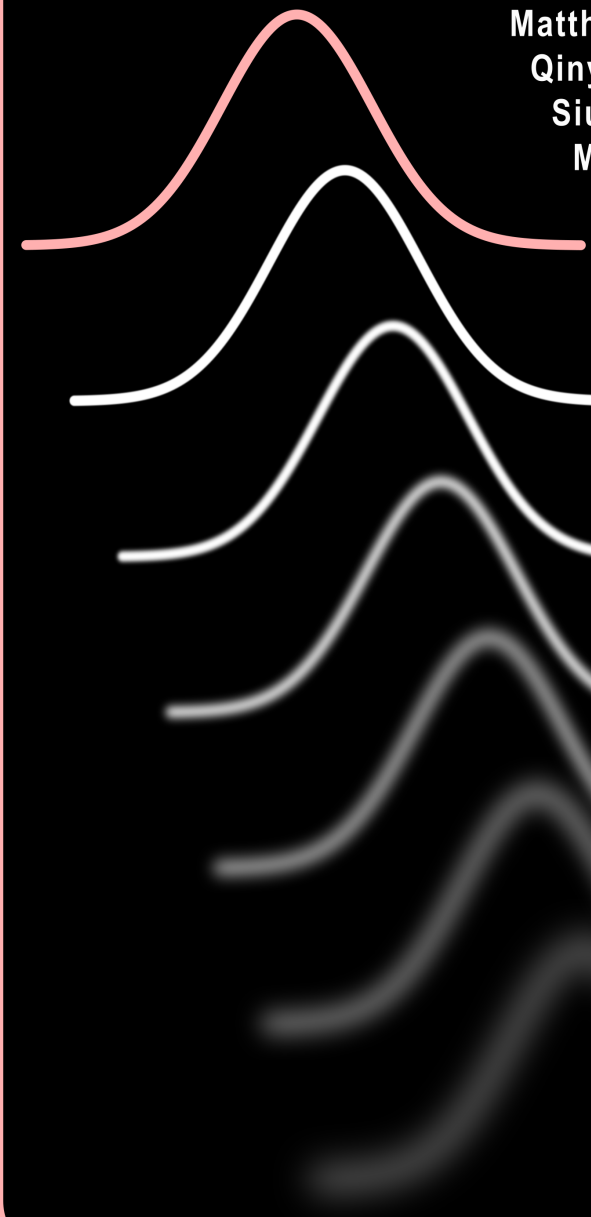
Guide to Effect Sizes and Confidence Intervals

Matthew B. Jané
Qinyu Xiao
Siu Kit Yeung
Mattan S. Ben-Shachar
Aaron R. Caldwell
Denis Cousineau
Daniel J. Dunleavy
Mahmoud Elsherif
Blair T. Johnson
David Moreau
Paul Riesthuis
Lukas Röseler
James Steele
Felipe F. Vieira
Mircea Zloteanu
Gilad Feldman



Guide to Effect Sizes and Confidence Intervals

Matthew B. Jané Qinyu Xiao Siu Kit Yeung
Flavio Azevedo Mattan S. Ben-Shachar Aaron R. Caldwell
Denis Cousineau Daniel J. Dunleavy Mahmoud Elsherif
Tylor J. Harlow Blair T. Johnson David Moreau
Paul Riesthuis Lukas Röseler James Steele
Felipe F. Vieira Mircea Zloteanu Gilad Feldman

2024

Table of contents

Welcome

This effect sizes and confidence intervals collaborative guide aims to provide academics, students and researchers with hands-on, step-by-step instructions for calculating effect sizes and confidence intervals for common statistical tests used in the behavioral, cognitive and social sciences, particularly when original data are not available and when reported information is incomplete. It also introduces general background information on effect sizes and confidence intervals, as well as useful R packages for their calculation. Many of the methods and procedures described in this Guide are based on R or R-based Shiny Apps developed by the science community. We were motivated to focus on R as we aim to maximize the reproducibility of our research outcomes and encourage the most reproducible study planning and data analysis workflow, though we also document other methods whenever possible for the reference of our readers. We regularly update this open educational resource, as packages are updated frequently and new packages are developed from time to time in this rapidly changing Open Scholarship era.

Introduction

Effect sizes and confidence intervals are critical metrics for interpreting results and quantifying the magnitude of findings in scientific research. However, calculating these values can be challenging, particularly when original data are unavailable or results are incompletely reported in prior publications. To address this need, our collaborative guide provides hands-on instructions for calculating effect sizes and confidence intervals for common statistical tests in the behavioral, cognitive, and social sciences. Our guide includes background information on these concepts as well as recommendations for useful R packages that can automate many of these computations. R is emphasized due to its capabilities for reproducible analyses; however, we also cover alternative methods for those without expertise in R. This guide is intended to be an evolving open educational resource, updated as new methods and packages become available in this fast-changing era of open scholarship. By compiling these applied instructions, our goal is to enable students and researchers to easily obtain these metrics, facilitating robust and transparent quantification of results, as well as cumulative scientific progress.

Guidelines for contribution

All are encouraged to contribute to this Guide. Please note that this Guide is in continuous development such that it will remain a work in progress for an indefinite period of time. This is intended because we hope the Guide to always reflect the state of the art on the topics of effect sizes and confidence intervals. To contribute, there are now two options:

1. You can suggest edits and make comments in the following google doc: mgto.org/effectsizeguide.
2. You can suggest edits directly in the online book using Hypothes.is. To do this you will need to create a free account on hypothes.is (hypothes.is/signup; this will take about a minute). Then when you navigate to the online book, you can open the panel on the top right of the screen. There you can suggest edits and create comments with code and latex!

Notes

- Please use the headings and style as set forth in this document. You can use keyboard shortcuts such as Ctrl + Alt + 1/2/3. The normal text is in Times New Roman font, font size 11. The codes are formatted using the Code Blocks add-on of Google Docs, github theme, font size 8.
- Use the Suggesting mode rather than the Editing mode. Suggesting is now the default mode for this document. Therefore, please do not hesitate to correct mistakes or modify the contents directly.
- Add a comment to the document if you find anything missing or improper, or if you feel that things are better organized in a different way. We appreciate your suggestions. If you have any questions, please also add a comment. We will reply and seek to clarify in the document body.
- Please make proper citations (in APA 7th format) and provide relevant links when you refer to any source that is not your own.

Credit and authorship

If you believe you have made sufficient contribution that qualifies you as an author, and you would like to be listed as an author of this Guide, please do not hesitate and list your name and contact information below. The administrators (M. B. J., Q. X., S. K. Y., and G. F.) of this Guide will verify your contribution and add you to the author list. We welcome comments from any person, regardless of whether they want to be an author. You are also welcome to request content to be added to this Guide (please see the Things to add to the guide section in the end).

The authorship order is such that M. B. J. and Q. X. will be the first two authors, S. K. Y. will be second author, and G. F. will be the last and the corresponding author. All other contributors will be listed alphabetically in the middle and are all considered joint third authors. Contributors are by default given investigation, writing - original draft, and writing - review & editing CRediT authorship roles. It is possible to take on more roles if contributors prefer. Any change in this authorship order rule will have to be approved by all who are already listed as an author.

Cite this guide

Cite this guide with the following citation:

APA:

Jané, M.B., Xiao, Q., Yeung, S., *Azevedo, F., *Ben-Shachar, M.S., *Caldwell, A.R., *Cousineau, D., *Dunleavy, D.J., *Elsherif, M., *Harlow T.J., *Johnson, B., *Moreau, D., *Riesthuis, P., *Röseler, L., *Steele, J., *Vieira, F.F., *Zloteanu, M., & ^Feldman, G. (2024). Guide to Effect Sizes and Confidence Intervals. http://dx.doi.org/10.17605/OSF.IO/D8C4G

BibTeX:

```
@misc{jané2024,
  title={Guide to Effect Sizes and Confidence Intervals},
  url={osf.io/d8c4g},
  DOI={10.17605/OSF.IO/D8C4G},
  publisher={OSF},
  author={Jané, Matthew B and Xiao, Qinyu and Yeung, Siu Kit and Azevedo, Flavio and Ben-S
  year={2024}
}
```

1 Defining Effect Sizes

Effect sizes quantify the magnitude of effects (i.e., strength of a relationship, size of a difference), which are the outcomes of empirical research. Many types effect sizes also provide information on the direction of the association (e.g., a positive or negative correlation).

Effect sizes are by no means a new concept. However, reporting them remained largely optional for many years, and only until recently does it become a community standard: scientists now see reporting effect sizes (in addition to the traditional statistical significance) as a must and journals also start to require such reporting. Notably, in 2001 and 2010, The Publication Manual of the American Psychological Association 5th and 6th editions emphasized that it is “almost always necessary” (Divine et al. 2018) to report effect sizes (APA 2010, 34; see Fritz, Morris, and Richler 2012, which provides a comprehensive summary on history and importance of effect size reporting).

Effects sizes can be grouped in broad categories as (1) raw effect sizes, and (2) standardized effect sizes. The raw effect sizes are a summary of the results that are expressed in the same units as the raw data. For example, when kilograms are measured, a raw effect size reports a measure in kilograms. Consider the effect of a diet on a treatment group; a control group receives no diet. The change in weight can be expressed as the mean difference between the groups. This measure is also in kg and so is a raw effect size. Standardized effect sizes expressed on standardized units such as standard deviations, percentages, or ratios. Standardized effect sizes tend to be more comparable across studies that use different measures or unit scales.

2 Benchmarks

2.1 Introduction to benchmarks

What makes an effect size "large" or "small" is completely dependent on the context of the study in question. However, it can be useful to have some loose criterion in order to guide researchers in effectively communicating effect size estimates. Jacob Cohen (1988), the pioneer of estimation statistics, suggested many conventional benchmarks (i.e., how we refer to an effect size other than using a number) that we currently use. However, Cohen (1988) noted that labels such as "small", "medium", and "large" are relative, and in referring to the size of an effect, the discipline, the context of research, as well as the research method and goals, should take precedence over benchmarks any time it's possible. There are general differences in effect sizes across different disciplines, and within each discipline, effect sizes differ depending on study designs and research methods (Schäfer and Schwarz 2019) and goals; as Glass, McGaw, and Smith (1981) explains:

> Depending on what benefits can be achieved at what cost, an effect size of 2.0 might be "poor" and one of .1 might be "good."

Therefore, it is crucial to recognize that benchmarks are only general guidelines, and importantly, out of context. They also tend to attract controversy (Glass, McGaw, and Smith 1981; Kelley and Preacher 2012; Harrell 2020). Note that field-specific empirical benchmarks have been suggested by researchers. For social psychology, these alternative benchmarks obtained through meta-analyzing the literature (for example, this and this; see this Twitter/X thread for a summary) are typically smaller than what Cohen put forward. Although such field-specific effect size distributions can provide an overview of the observed effect sizes, it does not provide a good interpretation of the magnitude of the effect (see Panzarella, Beribisky, and Cribbie 2021). To examine the magnitude of the effect, the specific context of the study at hand needs to be taken into account (pp. 532-535, Cohen 1988). Please refer to the table below:

Effect Size	Reference	Small	Medium	Large
Mean Differences				

Effect Size	Reference	Small	Medium	Large
Cohen's d or Hedges' g	Cohen (1988)[1]	0.20	0.50	0.80
		0.18	0.37	0.60
	Lovakov and Agadullina (2021)[2]	0.15	0.36	0.65
Correlational				
Correlation Coefficient (r)	Cohen (1988)	.10	.30	.50
	Richard, Bond Jr., and Stokes-Zoota (2003)[3][4]	.10	.20	.30
	Lovakov and Agadullina (2021)	.12	.24	.41
	Paterson et al. (2016)	.12	.20	.31
	Bosco et al. (2015)	.09	.18	.26
Cohen's f^2		.02	.25	.40
eta-squared (η^2)	Cohen (1988)	.01	.06	.14
Cohen's f	Cohen (1988)	.10	.25	.40
Categorical				
Cohen's w	Cohen (1988)	0.10	0.30	0.50
Phi	Cohen (1988)	.10	.30	.50

[1] Sawilowsky (2009) expanded Cohen's benchmarks to include very small effects (d = 0.01), very large effects (d = 1.20), and huge effects (d = 2.0). It has to be noted that very large and huge effects are very rare in experimental social psychology.

[2] According to this recent meta-analysis on the effect sizes in social psychology studies, "It is recommended that correlation coefficients of .1, .25, and .40 and Hedges' g (or Cohen's d) of 0.15, 0.40, and 0.70 should be interpreted as small, medium, and large effects for studies in social psychology.

[3] Note, for paired samples, this does not refer to the probability of an increase/decrease in paired samples but rather the probability of a randomly sampled value of X. This is also referred to as the "relative" effect in the literature. Therefore, the results will differ from the concordance probability provided below.

[4] These benchmarks are also recommended by Gignac and Szodorai (2016). Funder and Ozer (2019) expanded them to also include very small effects (r = .05) and very large effects (r = .40 or greater). According to them, […] an effect-size r of .05 indicates an effect that is very small for the explanation of single events but potentially consequential in the not-very-long run, an effect-size r of .10 indicates an effect that is still small at the level of single events but potentially more ultimately consequential, an effect-size r of .20 indicates a medium effect that is of some explanatory and practical use even in the short run and therefore even more important, and an effect-size r of .30 indicates a large effect that is potentially powerful in both the short and the long run. A very large effect size (r = .40 or greater) in the context of psychological research is likely to be a gross overestimate that will rarely be found in a large sample or in a replication." But see here for controversies with this paper.

Effect Size	Reference	Small	Medium	Large
Cramer's V		[5]		
Cohen's h	Cohen (1988)	0.2	0.5	0.8

It should be noted that small/medium/large effects do not necessarily mean that they have small/medium/large practical implications (for details see, Coe 2012; Pogrow 2019). These benchmarks are more relevant for guiding our expectations. Whether they have practical importance depends on contexts. To assess practical importance, it will always be desirable for standardized effect sizes to be translated to increase/decrease in raw units (or any meaningful units) or a Binomial Effect Size Display (roughly, differences in proportions such as success rate before and after intervention). The reporting of unstandardardized effect sizes is not only beneficial for interpretation but they are also more robust and more easy to compute (Baguley 2009). Additionally, a useful tool to examine, for example, the magnitude of a Cohen's d is by examining U3, percentage overlap, probability of superiority, and numbers needed to treat (For nice visualizations see https://rpsychologist.com/cohend/, Magnusson 2023).

To further assess the practical importance of observed effect sizes, it is necessary to establish the smallest effect size of interest for each specific field (SESOI, Lakens, Scheel, and Isager 2018). Cohen's benchmarks, field-specific benchmarks, or published findings are not preferred to establish the SESOI because they do not convey information about the practical relevance/magnitude of an effect size (Panzarella, Beribisky, and Cribbie 2021). Recent developments in various areas of research in psychology have been taken to establish the SESOI through anchor-based methods (Anvari and Lakens 2021), consensus-methods (Riesthuis et al. 2022), and cost-benefit analyses (see Otgaar et al. 2022, 2023). These approaches are frequently implemented successfully in medical research (e.g., HEIJDE et al. 2001) and recommendations are to, ideally, implement the various methods simultaneously to obtain a precise estimate of the smallest effect size of interest (termed minimally clinically important difference in the medical literature, Bonini et al. 2020). Interestingly, the minimally clinically important difference (MCID, smallest effect which patients perceive as beneficial [or harmful], McGlothlin and Lewis 2014) is sometimes even deemed as a low bar and other measures are encouraged such as patient acceptable symptomatic state (PASS, level of symptoms a patients allows while still accept their symptom state, this can be used to examine whether a certain treatment leads to a state that patients consider acceptable, Daste et al. 2022), substantial clinical benefit (SCB, effect that leads patient to self-report significant improvements, Wellington et al. 2023), and maximal outcome improvement (MOI, similar to MCID, PASS, and SCB, except that the scores are normalized by the maximal improvement possible for each patient, Beck et al. 2020; Rossi, Brand, and Lubowitz 2023).

[5] The benchmarks for Cramer's V are dependent on the size of the contingency table on which the effect is calculated. According to Cohen, use benchmarks for phi coefficient divided by the square root of the smaller dimension minus 1. For example, a medium effect for a Cramer's V from a 4 by 3 table would be .3 / sqrt(3 - 1) = .21.

Please also note that only zero means no effect. An effect of the size .01 is an effect, but a very small (Sawilowsky 2009), and likely unimportant one. It makes sense to say that "we failed to find evidence for rejecting the null hypothesis," or "we found evidence for only a small/little/weak-to-no effect" or "we did not find a meaningful effect". **It does not make sense to say, "we found no effect."** Purely by the random nature of our universe, it is hard to imagine that we can obtain a sharp zero-effect result. This is also related to the crud factor, which refers to the idea that "everything correlates with everything else" (Orben and Lakens 2020, 1; Meehl 1984), but the practical implication of very weak/small correlations between some variables may be limited, and whether the effect is reliably detected depends on statistical power.

2.2 Why Contextualizing Effect Sizes Matters

Interpreting effect sizes is not straightforward. Many researchers default to benchmarks–labeling an effect size as "small," "medium," or "large" based on cut-offs (e.g. Cohen's $d \approx 0.2, 0.5, 0.8$) or corresponding correlation values ($r \approx 0.1, 0.3, 0.5$).

While Cohen's effect size conventions are widely taught and convenient, they can be misleading if applied uncritically. Even Cohen himself cautioned that these cut-offs were arbitrary and intended as a last resort in the absence of domain-specific guidance (Cohen, 1988; Lakens, 2013). What qualifies as a "small" effect in one discipline might be average or even large in another. For instance, although r=0.30r = 0.30r=0.30 is classified as a medium correlation by Cohen's rule of thumb, empirical surveys in applied psychology suggest that typical effects in the field often fall between r=0.20r = 0.20r=0.20 and r=0.30r = 0.30r=0.30, making such values relatively large in context (Richard, Bond, & Stokes-Zoota, 2003; Funder & Ozer, 2019).

Although it can be a useful rule of thumb, solely relying on fixed benchmarks across disciplines risks misrepresenting findings. Many published effects that meet Cohen's "medium" threshold would be considered large when compared to empirical distributions within their sub-fields. Conversely, effects below Cohen's d 0.20 (or 0.10 r) can still be meaningful, particularly when aggregated over time or applied broadly (as later sections will illustrate). Ultimately, an effect size's significance depends on its research context, including measurement scales, base rates, and what is typical for the domain (Hill et al. 2008).

Example 2.1 (Elections)**.** Why a "Small" Effect Can Be a Big Deal To grasp why context is crucial, consider a voter turnout scenario. Suppose a new get-out-the-vote message increases voter turnout by only 2 percentage points among those who receive it. In raw terms, this is a small effect – most people's behavior didn't change. Yet in a national election, a 2% swing can decide the winner. A marginal change that seems trivial in percentage terms can have enormous real-world consequences when an outcome (like an election) is on a knife's edge. Similarly, a medical example: a daily aspirin regimen might only reduce the absolute risk of heart attack by a fraction of a percent for an individual (a tiny effect size, say $r \approx 0.03$). But

if millions of people take aspirin, thousands of heart attacks could be prevented. What looks "small" by statistical convention can be life-saving at scale. These examples illustrate that magnitude labels (small/medium/large) are not value judgments – a "small" effect can matter a great deal if the context amplifies its impact (large population, repeated over time, high stakes decision), whereas a "large" effect in a trivial context might not matter at all. Throughout this chapter, we will see many such cases where small effects accumulate into big outcomes and why researchers must interpret effect sizes with context in mind.

2.2.1 Lessons from Empirical Benchmarks

Hill, Bloom, Black, and Lipsey (2008) argued that these generic classifications lack empirical grounding and advocate for context-driven benchmarks. Their work exemplifies the need for compelling alternatives: evaluating effect sizes relative to normative expectations, policy-relevant gaps, and observed results from similar interventions.

2.2.1.1 Normative Expectations for Growth

One approach to contextualizing effect sizes is by comparing them to expected growth in the absence of an intervention. Using nationally normed standardized test data, the authors examined average student learning gains across K–12 education. The results highlight how expected academic progress varies significantly by grade level:

- Grade 1–2: Average annual gain of $d = 0.97$ (reading) and $d = 1.03$ (math).
- Grade 5–6: Gains decline to $d = 0.32$ (reading) and $d = 0.41$ (math).
- Grade 11–12: Minimal expected gains of $d = 0.06$ (reading) and $d = 0.01$ (math).

This empirical benchmark suggests that an intervention producing may be negligible in early grades but relatively substantial in high school. Contextualizing effect sizes within natural developmental trajectories is crucial for assessing their substantive significance.

2.2.1.2 *Policy-Relevant Performance Gaps*

Another way to interpret effect sizes is by comparing them to existing disparities in educational outcomes. Using National Assessment of Educational Progress (NAEP) data, the authors quantified achievement gaps by race, socioeconomic status, and gender:

- Black-White gap in reading: $d = -0.83$ (Grade 4), $d = -0.67$ (Grade 12).

- SES gap (free vs. reduced-price lunch eligibility) in math: $d = -0.85$ (Grade 4).
- Gender gap in reading: $d = -0.18$ (Grade 4), $d = -0.44$ (Grade 12).

If an educational intervention produces an effect of $d = 0.10$, it must be interpreted relative to these gaps. A $d = 0.10$ effect would have little impact in closing the Black-White achievement gap but could significantly influence gender-based disparities in literacy.

2.2.1.3 *Observed Effect Sizes from Similar Interventions*

Finally, the authors suggest using historical effect sizes from randomized controlled trials (RCTs) and meta-analyses as a benchmark for assessing new interventions. Their synthesis of 61 RCTs found that:

- Elementary school interventions typically yield $d = 0.33$.
- Middle school interventions average $d = 0.51$.
- High school interventions produce smaller effects, averaging $d = 0.27$.

A separate meta-analysis of 76 meta-analyses found that across all grades, effect sizes cluster between $d = 0.20$ and $d = 0.30$. This suggests that any new intervention achieving an effect size within this range is performing on par with prior research, while larger effects may indicate an exceptionally successful intervention.

2.2.2 Implications for Effect Size Interpretation

Hill et al. (2008)'s framework underscores the importance of contextualizing effect sizes within empirical reality rather than relying on arbitrary thresholds. Their approach offers three key takeaways:

1. ***Compare intervention effects to natural growth rates***—an impact in early grades may be trivial, while the same effect in high school is more meaningful.
2. ***Evaluate effect sizes against real-world disparities***—a policy-relevant benchmark (e.g., racial achievement gaps) provides a clearer sense of whether an intervention is impactful.
3. ***Situate new findings within prior research***—meta-analytic evidence helps gauge whether an observed effect is typical or remarkable within a given field.

By embedding effect sizes in empirical benchmarks, researchers can move beyond rigid classifications and provide a more nuanced, context-sensitive interpretation of intervention impacts.

3 Reporting Effect Sizes

When reporting effect sizes, it is important to provide sufficient detail and context to ensure transparency, convey directionality, and indicate precision. Transparency involves clearly documenting procedures and data so that others can reproduce your effect size calculations. Next, for directional effects like Cohen's d, make sure to define the direction of comparison and align it with your hypothesis. Finally, indicate the precision of the estimate, typically by reporting confidence intervals. Narrower confidence intervals reflect more precision, while wider intervals reflect greater uncertainty (Winter, 2019). Factors like sample size, variability, and study design influence precision. Reporting effect sizes thoughtfully with transparency, directionality, and precision, enables readers to accurately interpret the meaningfulness and implications of your results. In the following sections, we provide recommendations to optimize reporting on each of these factors.

> **i** Not all CIs are created equal.
>
> Confidence Intervals only indicate parameter precision under specific assumptions. Some have even titled this issue as the precision fallacy (Morey et al. 2016). For the same data, CIs can be computed in various ways resulting in wildly different intervals (see the submarine example in Morey et al. 2016). Such CIs are computed by inverting hypothesis tests (using the p-value obtained from a model); see this discussion by Gelman (2011). Under this approach, the CI reflects the data and model (+assumptions), not just the parameter estimate. If one is using an improper model, the associated CI will be misleading and its width will not reflect precision or uncertainty. The solution is to compute CIs based on the data at hand, such as constructing parametric (if the distribution is known) or non-parametric (empirical distribution) bootstrapped CIs, or understand that your CIs are conditional on the model you used. That said, for CIs computed for effect sizes like Cohen's d, which assume a Gaussian distribution, the precision fallacy should not be a problem and can be used to infer precision (see this forum discussion).

3.1 Transparency

When reporting effect sizes and their calculations, you should prioritize transparency and reproducibility. No matter what tool you used to calculate your effect size (R is the most recommended tool here), you must make sure that others can easily follow your procedures and

obtain the same results. This means that if you use online calculators (which is discouraged) or standalone programs (JAMOVI is most recommended; you can also use JASP, which however does not allow access to syntax at this moment), you should include screenshots that capture the input and output, with clear explanations. If you use R, Python or other programming languages, you should copy-and-paste your codes into your supplementary document (or submit your scripts to open online repositories), ideally with annotations and comments explaining the codes. inputs and outputs.

3.2 Directionality

Some effect sizes are directional (e.g., Cohen's d, Pearson correlations r), which means that they can be positive or negative. Their signs carry important information, and therefore cannot be omitted. When you report these effect sizes, make it clear what is compared to what (i.e., the direction of comparison). Better still, make sure your comparison is inline with the theory. For instance, a theory predicts that your group X should score higher on an item than your Group Y,[1] you should hypothesize accordingly that Group X will have a higher mean than Group Y on the item, and subtract mean(Y) from mean(X) (rather than the other way around) to obtain the mean difference. You should then expect your t statistic to be positive, and your d value as well. In other words, avoid reporting anything like t = -5.14, d = 0.36, where the signs of the statistics do not match.

3.3 Precision

Effect sizes may be very precisely estimated from the available data, the used methodology, and how the population was sampled. It might also be estimated with little confidence on the resulting number. This may be the case for example when the sample is very small, when the population displays a lot of variability, when a between-group design is used instead of a paired-sample design, and finally, when clustered sampling is used instead of randomized sampling. Precision can be estimated using various tools, but probably the most commonly used one is the Confidence intervals. This interval has a confidence level, frequently 95%.

[1] Of course, if a theory/effect predicts Group X has a higher mean than Group Y, then it also predicts the reverse, i.e., Group Y has a lower mean than Group X. But theories/effects are commonly articulated in a certain way. It is more common that we say, for example, people prefer the status quo rather than that people do not prefer the non-status quo, when we refer to the status quo bias. Consider another "theory": teenagers get taller when they get older. It just does not make sense to say the same thing reversely, i.e., teenagers get shorter when they get younger, because people cannot get younger, at least in the 2020s.

4 Interpreting Confidence Intervals

What is the correct interpretation of a confidence interval? Imagine you conducted a study where you compared two groups. You obtained a Cohen's d = 0.3, 95% CI [0.2, 0.4]. How do you interpret this confidence interval?

Confidence intervals are yielded by a certain procedure, such that when the procedure is repeatedly applied to a series of hypothetical datasets drawn from the studied population/populations, it yields intervals that contain the true parameter value (in our example, it means the true difference between the two groups) in 95% of the cases. For the effect estimate and confidence intervals to be valid, the data and test must meet the assumptions of the estimating procedure.

In colloquial terms, if we conduct this research over and over (repeating the same sampling procedure, administering the same experimental manipulation, conducting the same statistical analysis, etc.), because of sampling variability (our samples are slightly different at each time), we will get different Cohen's d values. For each of these d values, we calculate a 95% interval. Then, among all these many intervals, we expect that 95% of them will contain the true d, which we never know exactly.

There is also a common criticism levied against the confidence interval interpretation: "There is a 95% probability that the true parameter exists within the 95% confidence interval". However this criticism is unwarranted in the specific case of a single observed confidence interval, that is, as long as there is a single realized confidence interval sampled from the population, this interpretation is fine (Vos and Holbert 2022). It is important to note however, this interpretation is incorrect when there are multiple realized confidence intervals randomly sampled from the same population. The criticized interpretation also tends to be more practical than the interpretation using repeated sampling, the following example described by Vos and Holbert (2022) illustrates this,

> The distinction between these interpretations can be understood with the simple example of the probability of rolling a '6' with a fair die. The probability is 1/6 because if you roll the die repeatedly the proportion of times that the face with '6' comes up will be come very close to 1/6. Or, the probability is 1/6 because it is equivalent to a random selection from an urn where exactly one of 6 balls is labelled with '6'. The distinction in this simple example is less useful since repeatedly rolling a die is less problematic than repeatedly conducting the same randomized trial.

For further reading on confidence interpretations, see Hoekstra et al. (2014) and Morey et al. (2016).

5 Reporting Confidence Intervals

Confidence intervals must be calculated and reported for every effect size that you obtained and mentioned in your manuscript. If you are doing a replication and your target article/study did not report CIs for its effect sizes, you should calculate CIs and report them.

Normally, we calculate 95% confidence intervals (i.e., 95% of such intervals are expected to contain the true parameter value if we conduct an infinite number of identical studies).

> **Alpha level**
>
> The confidence interval depends on the alpha level, that is, the proportion of CIs upon repeated sampling that will not contain the true parameter. If the true effect is zero (or null), the the alpha level represents the false positive rate (i.e., the rate of observing a significant effect when there is none). The 95% CI is based on an alpha level of .05, however researchers can choose any value (between 0 and 1), as long as it is properly justified (Lakens 2022).

Nonetheless, for some effect sizes (e.g., eta-squared, partial eta-squared, R-squared), we calculate 90% confidence intervals. This is because η^2 is squared and always positive, and F-tests are one-sided. Reporting 95% CI for eta squared may result in situations in which the CI includes zero but the p-value falls below .05, whereas reporting 90% CI prevents such a problem. For further information regarding this issue, read Daniel Lakens blog on confidence intervals and Steiger (2004).

Confidence intervals should be reported immediately after an effect size, e.g., Cohen's d = 0.40, 95% CI [0.20, 0.60]. After the first time reporting them in a manuscript, every subsequent CI can be simply denoted by brackets without the "95% CI" preceding it.

Unless you are measuring something that is meaningful in real life (e.g., income, years of experience, amount that a person is willing to donate), please make sure that the CI you calculated is a CI of the effect size, not of other statistics, such as the test statistics or mean difference in raw units.

If you see that the effect size estimate is not included within your CI, you likely have an issue, check carefully. For means and for difference in means, the estimate should be precisely the midpoint of your CI; for other statistics (e.g., correlation, proportion, frequency, standard deviation), one arm might be longer than the other so the estimate may not be the midpoint.

For further reading related to the calculation and reporting of effect sizes and confidence intervals, see Steiger (2004) and Lakens (2014).

6 Using R

6.1 Why Use R?

We strongly recommend using open-source software such as R or Python for computing effect sizes and confidence intervals. In this guide, we focus on R, which has several advantages:

- **Reproducibility**: R syntax can be shared to allow others to reproduce your analyses. This promotes transparency and reliability in research.
- **Flexibility**: CRAN repositories contain thousands of user-contributed packages for specialized statistical techniques. This allows calculating a diverse range of effect size and CI metrics.
- **Free and open source**: R is free to download and use. The open source nature means community-driven innovation and packages.
- **Visualizations**: R makes it easy to create publication-quality graphics to visualize your results.
- **Scripting**: Automating analyses through R scripts improves efficiency and consistency.
- **Range of packages**: Packages like `effectsize`, `MBESS`, `metafor`, and more contain a variety of effect size and CI functions.

Many (if not all) of these advantages are shared with Python and a number of other programming languages. While online calculators or GUI software can also allow calculating confidence intervals and effect sizes, open-source software such as R provide transparency, reproducibility, and access to a vast array of techniques. In the case of R, the learning curve is well worth it for doing robust, state-of-the-art effect size and confidence interval estimation.

6.2 Useful R Packages

The following R packages are handy for effect size and CI calculations, conversions among different effect sizes, and conversion of test statistics to effect sizes. If you use one of the packages below, please make sure you cite them to give the authors their due credit! To obtain citations for packages, you can use the `citation()` function and input the name of the package as a string.

- `MOTE` (Buchanan et al. 2019): This is a highly recommended package for calculating effect sizes, which is capable of handling a wide variety of effect sizes in the difference family (the d family) and variance-overlap family (r, eta, omega, epsilon). The functions also provide non-central confidence intervals for each effect size and output in APA style in LaTeX. `MOTE` has an online shiny application (doomlab.shinyapps.io/mote/). The CRAN project can be found here: cran.r-project.org/package=MOTE.
- `effectsize` (Ben-Shachar, Lüdecke, and Makowski 2020): This package is particularly useful in data analysis. A major advantage of this package is that it takes in many different model objects and directly outputs effect sizes and CIs. It also implements conversions between a wide array of indices and features functions to perform automated effect size interpretations based on existing benchmark thresholds. The CRAN project can be found here: cran.r-project.org/package=effectsize.
- `MBESS` (Kelley 2022): One of the most comprehensive and useful packages for effect size and confidence interval calculations. It provides functions that can calculate ESs and CIs from test statistics and the p-value. The CRAN project can be found here: cran.r-project.org/package=MBESS.
- `metafor` (Viechtbauer 2010): Probably the most comprehensive meta-analysis package currently available. Includes the function, `escalc()`, that calculates various types of effect sizes from test-statistics, summary statistics, and more. The CRAN project can be found here: cran.r-project.org/package=metafor.
- `psych` (William Revelle 2023): One of the most comprehensive and general packages for common statistical procedures in psychology research. It also includes some effect size and CI calculation functions (e.g., `cohen.d()`). The CRAN project can be found here: cran.r-project.org/package=psych.
- `esc` (Lüdecke 2019): This package can help convert among different effect sizes (pp. 4-12 in the reference manual). It's also helpful when only incomplete information (e.g., only descriptives, or only p-values) have been provided in the paper, and we want to calculate effect sizes from them. Another package that provides similar conversion functions is the `compute.es` package. The CRAN project can be found here: cran.r-project.org/package=esc.
- `psychmeta` (Dahlke and Wiernik 2019): This package is mainly used for psychometric meta-analyses. It has a function for converting different effect sizes/test statistics (convert_es, p. 38 in the reference manual), including r, d, t-statistic (and its p-value), F (and its p-value in two-group one-way ANOVA), chi-squared (one degree of freedom), etc., to r, d and the common language effect sizes (CLES, A, AUC). The CRAN project can be found here cran.r-project.org/package=psychmeta.
- `effsize` (Torchiano 2020): This is a relatively lightweight package that handles d, g, Cliff delta, and Vargha-Delaney A). The CRAN project can be found here: cran.r-project.org/package=effsize.

- `MAd` (W. T. Hoyt 2014): This package is a collection of functions for conducting a meta-analysis with mean differences data. It also provides conversion functions. The CRAN project can be found here: cran.r-project.org/package=MAd.
- `TOSTER` (Läkens 2017; Caldwell 2022): This package is designed for equivalence testing. It contains many functions to test for differences in effect sizes along with other useful functions for effect size comparisons. The CRAN project can be found here: cran.r-project.org/package=TOSTER.
- `DeclareDesign` (Blair et al. 2019): This simulation framework can be used to assess whether procedures for calculating confidence intervals are valid and can be used for arbitrary designs. The `diagnose_design()` function calculates coverage for designs with estimation strategies that produce confidence intervals. The CRAN project can be found here: cran.r-project.org/package=DeclareDesign.
- `statpsych` (Bonett 2024): This R package has many functions for computing confidence intervals, power analyses, and simulations for all different kinds of effect sizes. cran.r-project.org/web/packages/statpsych/statpsych.pdf

Part I

Standardized Effect Sizes

7 Mean Differences

T-tests are the most commonly used statistical tests for examining differences between group means or examining a group mean against a constant. Calculating effect sizes for t-tests are fairly straightforward. Nonetheless, there are cases where statistical information for the calculation of effect sizes are missing (which happens quite often in older articles), and therefore we document methods that make use of partial information (e.g., only the mean and standard deviation, or only the t-statistic and degrees of freedom) for the calculation. There are multiple types of effect sizes used to calculate standardized mean differences (d), yet researchers very often do not identify which type of d value they are reporting (see Lakens 2013). Here we document the equations and code necessary for calculating each type of d value compiled across multiple sources (Becker 1988; Cohen 1988; Lakens 2013; Caldwell 2022; Glass, McGaw, and Smith 1981). A d value calculated from a sample will also contain sampling error, therefore we will also include equations to calculate the standard error. The standard error allows us to then calculate the confidence interval. For each variant of d in the sections below, the 95% confidence interval is calculated in the following way, that is,

$$CI_d = d \pm 1.96 \times SE \tag{7.1}$$

Lastly, we will supply example R code so you can apply to your own data.

Here is a table for every effect size discussed in this chapter:

Type	Description	Section
Single Group Design		Section 7.2
d_s - Single Group	Standardized mean difference for comparing a single group to some constant	Section 7.2
Two Independent Groups Design		Section 7.3

Type	Description	Section
d_p - Pooled Standard Deviation	Uses the average within-group standard deviation to standardize the mean difference. Can be calculated directly from a independent sample t-test. Assumes homogeneity of variance between groups.	Section 7.3.1
d_Δ - Control Group Standard Deviation	Uses the standard deviation of the control group to standardize the mean difference (often referred to as Glass's Delta). Does not assume homogeneity of variance between treatment/intervention and control group.	Section 7.3.2
Repeated Measures (Paired Groups) Design		Section 7.4
d_z - Difference score standard deviation	Uses the standard deviation of difference scores (also known as change scores) to standardize the within person mean difference (i.e., pre/post change).	Section 7.4.1
d_{rm} - Repeated measures	Uses the within-person standard deviation that utilizes a correction to d_z to reduce the impact of the pre/post correlation on the effect size. Assumes homogeneity of variance between conditions.	Section 7.4.2
d_{av} - Average variance	Uses the pooled variance between conditions (pre/post test). Does not use the correlation between conditions. Assumes homogeneity of variance between conditions.	Section 7.4.3
d_b - Becker's *d*	Uses the pre-test standard deviation to standardize the pre/post mean difference. Does not assume homogeneity of variance between pre-test and post-test.	Section 7.4.4
Pre-Post-Control Design		Section 7.5

Type	Description	Section
d_{PPC1} - Separate pre-test standard deviations	Defined as the difference between the Becker's *d* between the treatment and control group. Particularly, standardizing the mean pre/post change by the pre-test of the respective group.	Section 7.5.1
d_{PPC2} - Pooled pre-test standard deviation	Standardizes the difference in mean changes between treatment and control group. Assumes homogeneity of variance between the pre-test of the control and treatment condition.	Section 7.5.2
d_{PPC3} - Pooled pre-test and post-test standard deviation	Pools the standard deviation between pre-test and post-test in treatment and control condition. Assumes homogeneity of variance between pre/post-test scores *and* treatment and control conditions. Confidence intervals are not easy to compute.	Section 7.5.3
Mean Ratios		Section 9.2.8
$lnRR_{\text{ind}}$ - Response ratio between independent groups	The ratio between the means between two groups. Does not use the standard deviation in the effect size formula.	Section 7.7.1
$lnRR_{\text{dep}}$ - Response ratio between dependent groups	The ratio between the means between conditions (i.e., repeated measures). Does not use the standard deviation in the effect size formula.	Section 7.7.2

7.1 Reporting a t-test with effect size and CI

Whatever effect size you choose to report, you can report it alongside the t-test statistics (i.e., t-value and the p value). For example,

> The treatment group had a significantly higher mean than the control group (*t* = 2.76, *p* = .009, *n* = 35, *d* = 0.47, 95% CI [0.11, 0.81]).

7.2 Single Group Designs

For a single-group design, we want to compare the mean of that group to some constant, C (i.e., a target value). The standardized mean difference for a single group can be calculated by (equation 2.3.3, Cohen 1988),

$$d_s = \frac{M - C}{S}, \tag{7.2}$$

where the standardizer (S) is the sample standard deviation. The interpretation of d_s is therefore how many standard deviations is the mean away from the target value, C. A positive d_s value would indicate that the mean is larger than the target value C, whereas a negative d_s value would denote the mean is less than the C. The corresponding standard error for d_s can then be calculated with (see documentation for Caldwell 2022),

$$SE_{d_s} = \sqrt{\frac{1}{n} + \frac{d_s^2}{2n}}, \tag{7.3}$$

where n denotes the sample size. In R, we can use the `d.single.t()` function from the `MOTE` package (Buchanan et al. 2019) to calculate the single group standardized mean difference.

```
# Install packages if not already installed:
# install.packages('MOTE')
# Cohen's d for one group

# For example:
# Sample Mean = 30.4, SD = 22.53, N = 96
# Target Value, C = 15

library(MOTE)

stats <- d.single.t(
  m = 30.4,
  u = 15,
  sd = 22.53,
  n = 96
)

SE <- sqrt(1/stats$n + stats$d^2/(2*stats$n))

# print just the d value and confidence intervals
```

```r
data.frame(d = round(stats$d,3),
           SE = round(SE,3),
           ci.lb = round(stats$dlow,3),
           ci.ub = round(stats$dhigh,3))
```

```
      d    SE ci.lb ci.ub
1 0.684 0.113  0.46 0.904
```

As you can see, the output shows that the effect size is d_s = 0.68, 95% CI [0.46, 0.90].

7.3 Two Independent Groups Design

7.3.1 Standardize by Pooled Standard Deviation (d_p)

For a design that consists of two independent groups (we can denote these as group A and group B), the standardized mean difference can be calculated by (equation 5.1, Glass, McGaw, and Smith 1981),

$$d_p = \frac{M_A - M_B}{S_p}, \tag{7.4}$$

where S_p is the pooled standard deviation defined as (pp. 108, Glass, McGaw, and Smith 1981),

$$S_p = \sqrt{\frac{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}{n_A + n_B - 2}}. \tag{7.5}$$

Using the pooled standard deviation as the standardizer characterizes the classic formulation of Cohen's d. This formulation requires the assumption that the variances (likewise the standard deviations) are equal between groups in the population. If this assumption is not met then it is recommended to use some of the other d value formulations later in this section. We can interpret the d_p value as the number of standard deviations the mean of group A is away from the mean of group B. A positive d_p value would indicate that the mean of group A is larger than the mean of group B and vice versa for a negative d_p value.

Cohen's d_p is related to the t-statistic from an independent samples t-test. In fact, we can calculate the d_p value from the t-statistic with the following formula (equation 5.3, Glass, McGaw, and Smith 1981):

$$d_p = t\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}. \tag{7.6}$$

The corresponding standard error of d_p is,

$$SE_{d_p} = \sqrt{\frac{n_A + n_B}{n_A n_B} + \frac{d_p^2}{2(n_A + n_B)}}. \tag{7.7}$$

In R, we can use the `escalc()` function from the `metafor` package to calculate the two group standardized mean difference (using the `measure = "SMD"` argument).

```
# use metafor package
library(metafor)

## Standardized mean difference for two independent groups

# Given means and SDs
# For example:
# Group A Mean = 30.4, SD = 22.53, N = 96
# Group B Mean = 21.4, SD = 19.59, N = 96

stats <- escalc(
  measure = "SMD",
  m1i = 30.4,
  m2i = 21.4,
  sd1i = 22.53,
  sd2i = 19.59,
  n1i = 96,
  n2i = 96,
  var.names = c("d", "variance") # add informative labels
)

# print output
summary(stats)
```

```
       d variance    sei     zi   pval  ci.lb  ci.ub
1 0.4246   0.0213 0.1460 2.9093 0.0036 0.1386 0.7107
```

```r
# Given t-test statistics
# For example:
# t = 2.954, nA = 96, nB = 96

stats <- escalc(
  measure = "SMD",
  ti = 2.954,
  n1i = 96,
  n2i = 96,
  var.names = c("d", "variance") # add informative labels
)

# print output
summary(stats)
```

```
      d variance    sei     zi   pval  ci.lb  ci.ub
1 0.4247   0.0213 0.1460 2.9097 0.0036 0.1386 0.7108
```

The output for both examples show that the effect size is d_p = 0.42, 95% CI [0.14, 0.71].

7.3.2 Standardize by Control Group Standard Deviation (d_Δ)

When two groups differ substantially in their standard deviations, we can instead standardize by one of the two available group's standard deviation, typically this is the control or reference group. In our scenario let's suppose that group B is the control/reference group and therefore we can use the standard deviation of group B (S_B) as the standardizer, such that,

$$d_\Delta = \frac{M_A - M_B}{S_B}. \tag{7.8}$$

This formulation is commonly referred to as Glass' Δ (Glass 1981). The standard error for d_Δ can be defined as,

$$SE_{d_\Delta} = \sqrt{\frac{n_A + n_B}{n_A n_B} + \frac{d_\Delta^2}{n_B + 1}} \tag{7.9}$$

Standardizing by the control group standard deviation rather than pooling (as we did in the previous section with d_p) results in less degrees of freedom ($df = n_C - 1$) and therefore a

larger standard error. In R, we can use the `escalc()` function from the `metafor` package to calculate d_{Δ} (using the `measure = "SMD1"` argument). Since we have already loaded in the `metafor` package, we do not need to do it again.

```
# Glass' delta (standardize by control group)
# given difference score means and SDs

# For example:
# group A Mean = 30.4, SD = 22.53, N = 96
# group B Mean = 21.4, SD = 19.59, N = 96

stats <- escalc(
  measure = "SMD1",
  m1i = 30.4,
  m2i = 21.4,
  sd2i = 19.59, # Note: use sd2i for whichever group needs to be standardized
  n1i = 96,
  n2i = 96,
  var.names = c("d", "variance") # add informative labels
)

# print the SDM value and confidence intervals
summary(stats)
```

```
      d variance    sei     zi   pval  ci.lb  ci.ub
1 0.4558   0.0219 0.1480 3.0788 0.0021 0.1656 0.7459
```

7.4 Repeated Measures Designs

In a repeated-measures design, the same subjects (or items, etc.) are measured on two or more separate occasions, or in multiple conditions within a single session, and we want to know the mean difference between those occasions or conditions (Baayen, Davidson, and Bates 2008; Barr et al. 2013). An example of this would be in a pre/post comparison where subjects are tested before and after undergoing some treatment (see Figure 7.1 for a visualization). A standardized mean difference in a repeated-measures design can take on a few different forms that we define below.

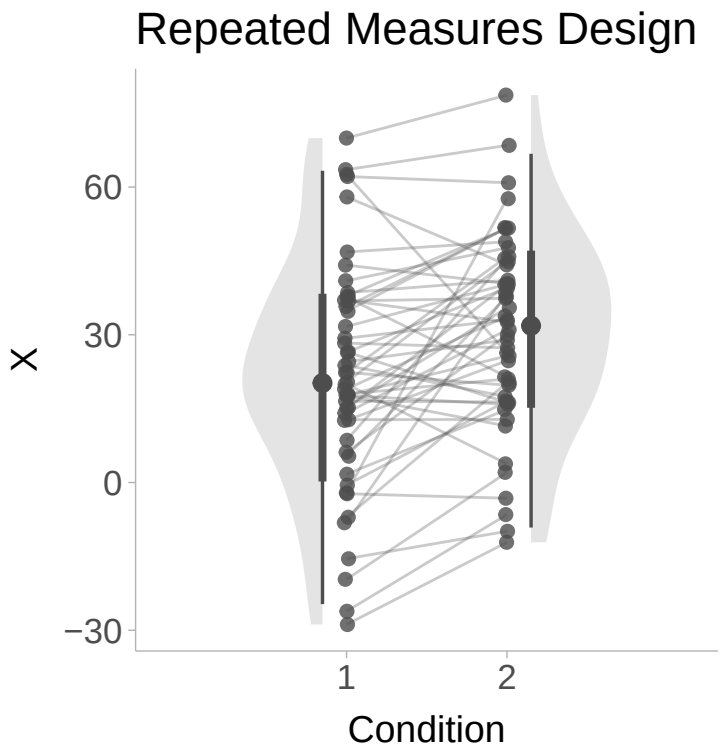


Figure 7.1: Figure displaying simulated data of a repeated measures design, the x-axis shows the condition (e.g., pre-test and post-test) and y-axis is the scores. Lines indicate within person pre/post change.

7.4.1 Difference Score d (d_z)

Instead of comparing the means of two sets of scores, a within subject design allows us to subtract the scores obtained in condition 1 from the scores in condition 2. The means and standard deviations of difference scores ($X_{\text{diff}} = X_2 - X_1$) can be treated similarly to that of a single group design (if the target value was zero, i.e., $C = 0$) such that (equation 2.3.5, Cohen 1988),

$$d_z = \frac{M_{\text{diff}}}{S_{\text{diff}}} \tag{7.10}$$

A positive d_z value would indicate that the scores in condition 2 are, on average, larger than scores than condition 1 and vice versa for a negative d_z value. A convenient aspect of d_z is that it has a straight-forward relationship with the paired t-statistic, $d_z = \frac{t}{\sqrt{n}}$. This makes it very useful for power analyses. If the standard deviation of difference scores are not accessible, then it can be calculated using the standard deviation of condition 1 (S_1), the standard deviation of condition 2 (S_2), and the correlation between conditions (r) (equation 2.3.6, Cohen 1988):

$$S_{\text{diff}} = \sqrt{S_1^2 + S_2^2 - 2rS_1S_2} \tag{7.11}$$

It is important to note that when the correlation between groups is large, then the d_z value will also be larger, whereas a small correlation will return a smaller d_z value. The standard error of d_z can be calculated similarly to the single group design such that,

$$SE_{d_z} = \sqrt{\frac{1}{n} + \frac{d_z^2}{2n}} \tag{7.12}$$

In R, we can use the `escalc()` function from the `metafor` package to calculate d_z (using the `measure = "SMCC"` argument).

```
# Cohen's dz for difference scores


# From paired t-test
# For example:
# t = 10.70, N = 96
stats <- escalc(
  measure = "SMCC",
  ti = 10.70,
```

```
  ni = 96,
  var.names = c("d", "variance") # add informative labels
)

# print output
summary(stats)
```

```
      d variance    sei     zi   pval  ci.lb  ci.ub
1 1.0834   0.0165 0.1286 8.4267 <.0001 0.8314 1.3354
```

```
# given difference score means and SDs
# For example:
# Difference Score Mean = 21.4, SD = 19.59, N = 96
stats <- escalc(
  measure = "SMCC",
  m1i = 21.4,
  m2i = 0, # per documentation, this value should be set to zero
  sd1i = 19.59,
  sd2i = 0, # per documentation, this value should be set to zero
  ri = 0, # per documentation, this value should be set to zero
  ni = 96,
  var.names = c("d", "variance") # add informative labels
)

# print output
summary(stats)
```

```
      d variance    sei     zi   pval  ci.lb  ci.ub
1 1.0837   0.0165 0.1286 8.4283 <.0001 0.8317 1.3358
```

The output shows that the effect size is d_z = 1.08, 95% CI [0.83, 1.34].

7.4.2 Repeated Measures d (d_{rm})

For a within-group design, we want to compare the means of scores obtained from condition 1 and condition 2. The repeated measures standardized mean difference between the two conditions can be calculated by (equation 9, Lakens 2013),

$$d_{rm} = \frac{M_2 - M_1}{S_w}. \tag{7.13}$$

The standardizer here is the within-subject standard deviation, S_w. The within-subject standard deviation can be defined as,

$$S_w = \sqrt{\frac{S_1^2 + S_2^2 - 2rS_1S_2}{2(1-r)}}. \tag{7.14}$$

We can also express S_w in terms of the standard deviation of difference scores (S_{diff}),

$$S_w = \frac{S_{\text{diff}}}{\sqrt{2(1-r)}}. \tag{7.15}$$

Furthermore, we can even express d_{rm} in terms of the difference score standardized mean difference (d_z),

$$d_{rm} = d_z \times \sqrt{2(1-r)}. \tag{7.16}$$

Ultimately, d_{rm} is more appropriate as an effect size estimate for use in meta-analysis whereas d_z is more appropriate for power analysis (Lakens 2013). The standard error for d_{rm} can be computed as,

$$SE_{d_{rm}} = \sqrt{\left(\frac{1}{n} + \frac{d_{rm}^2}{2n}\right) \times 2(1-r)} \tag{7.17}$$

In R, we can use the `d.ind.t.rm` function from the `MOTE` package to calculate the repeated measures standardized mean difference (d_{rm}).

```
# Cohen's d for repeated measures
# given means and SDs and correlation
library(MOTE)

# For example:
# Condition 1 Mean = 30.4, SD = 22.53, N = 96
# Condition 2 Mean = 21.4, SD = 19.59, N = 96
# correlation between conditions: r = .40
```

```r
stats <- d.dep.t.rm(
  m1 = 30.4,
  m2 = 21.4,
  sd1 = 22.53,
  sd2 = 19.59,
  r = .40,
  n = 96,
  a = 0.05
)

SE = sqrt( (1/stats$n + stats$d^2/(2*stats$n)) * 2*(1-stats$r))

# print just the d value and confidence intervals
data.frame(d = round(stats$d,3),
           SE = round(SE,3),
           ci.lb = round(stats$dlow,3),
           ci.ub = round(stats$dhigh,3))
```

```
      d    SE ci.lb ci.ub
1 0.425 0.117 0.215 0.633
```

The output shows that the effect size is d_{rm} = 0.43, 95% CI [0.22, 0.63].

7.4.3 Average Variance d (d_{av})

The problem with d_z and d_{rm}, is that they require the correlation between conditions. In practice, correlations between conditions are frequently not reported. An alternative estimator of d in repeated measures design is to simply use the classic variation of Cohen's d_p (i.e., pooled standard deviation). In a repeated measures design, the sample size does not change between conditions, therefore weighting the variance of condition 1 and condition 2 by their respective degrees of freedom is an unnecessary step. Instead, we can standardize by the square root of the average the variances of condition 1 and 2 (see equation 5, Algina and Keselman 2003):

$$d_{av} = \frac{M_2 - M_1}{\sqrt{\frac{S_1^2 + S_2^2}{2}}} \tag{7.18}$$

This formulation is convenient especially when the correlation between conditions is not present, however without the correlation it fails to take into account the consistency of change

between conditions. However, the consistency of scores is taken into account in the standard error of the d_{av} which can be expressed as (equation 9, Algina and Keselman 2003),

$$SE_{d_{av}} = \sqrt{\frac{2(S_1^2 + S_2^2 - 2rS_1S_2)}{n(S_1^2 + S^2)}} \tag{7.19}$$

As we might expect, the higher the correlation (the more consistent the change in scores between conditions) the smaller the standard error. In R, we can use the `escalc()` function from the `metafor` package to calculate the average variance standardized mean difference (d_{av}; using the `measure = "SMCRP"` argument).

```
# Cohen's d for repeated measures (average variance)
# given means and SDs

# For example:
# Condition 1 Mean = 30.4, SD = 22.53, N = 96
# Condition 2 Mean = 21.4, SD = 19.59, N = 96
# Correlation between conditions: r = .50

stats <- escalc(
  measure = "SMCRP",
  m1i = 30.4,
  m2i = 21.4,
  sd1i = 22.53,
  sd2i = 19.59,
  ri = .50,
  ni = 96,
  var.names = c("d", "variance") # add informative labels
)

# print just the d value and confidence intervals
summary(stats)
```

```
      d variance    sei     zi   pval  ci.lb  ci.ub
1 0.4242   0.0110 0.1049 4.0442 <.0001 0.2186 0.6298
```

The output shows that the effect size is d_{av} = 0.42, 95% CI [0.22, 0.63].

7.4.4 Becker's d (d_b)

An even simpler variant of repeated measures d value comes from Becker (1988). Becker's d standardizes simply by the pre-test standard deviation (we will denote the pre-test with condition 1) when the comparison is a pre/post design,

$$d_b = \frac{M_2 - M_1}{S_1}. \tag{7.20}$$

A convenient aspect of Becker's d is in the use of the raw score standard deviation (S_1) as the standardizer. This allows us to interpret d_b in units of standard deviations of pre-test scores, whereas for d_z and d_{rm} the interpretation of the effect size units are less clear.

We can also obtain the standard error with (equation 13, Becker 1988),

$$SE_{d_b} = \sqrt{\frac{2(1-r)}{n} + \frac{d_b^2}{2n}} \tag{7.21}$$

Notice that even though the formula for calculating d_b did not include the correlation coefficient, the standard error does. Using the `escalc()` function, we can calculate Becker's formulation of standardized mean difference (using the `measure = "SMCR"` argument).

```
# Cohen's d for repeated measures standardized with pre-test SD (becker's d)
# given means, the pre-test SDs, and the correlation

# For example:
# Pre-test Mean = 21.4, SD = 19.59, N = 96
# Post-test Mean = 30.4, N = 96
# Correlation between conditions: r = .40

# NOTE: MEANS FLIPPED SO THAT M2 - M1
# (by default escalc does M1 - M2)
stats <- escalc(
  measure = "SMCR",
  m1i = 30.4, # post-test mean
  m2i = 21.4, # pre-test mean
  sd1i = 22.53, # pre-test SD
  ri = .50,
  ni = 96,
  var.names = c("d", "variance") # add informative labels
)
```

```
# print just the d value and confidence intervals
summary(stats)
```

```
      d variance    sei     zi   pval  ci.lb  ci.ub
1 0.3963   0.0112 0.1060 3.7389 0.0002 0.1886 0.6040
```

The output shows that the effect size is d_b = 0.40, 95% CI [0.19, 0.60].

7.4.5 Comparing Repeated Measures d values

Figure 7.2 shows repeated measures designs with a high ($r = .95$) and low ($r = .05$) correlations between conditions. Let us fix the standard deviations and means for both conditions and only vary the correlation. Now we can compare the repeated measures estimators based on these two conditions shown in Figure 7.2:

- High correlation:
  - $d_z = 1.24$
  - $d_{rm} = 0.39$
  - $d_{av} = 0.43$
  - $d_b = 0.40$
- Low correlation:
  - $d_z = 0.31$
  - $d_{rm} = 0.43$
  - $d_{av} = 0.43$
  - $d_b = 0.40$

We notice that the correlation greatly influences d_z more than any other estimator. The d_{rm} value has very little change, whereas d_{av} and d_b do not take into account the correlation at all.

7.5 Pretest-Posttest-Control Group Designs

In many areas of research both between and within group factors are incorporated. For example, in research involving the examination of the effects of an intervention often a sample is randomised into two seperate groups (intervention and control) and then they are measured on the outcome of interest both before (pre-test) and after (post-test) the intervention/control

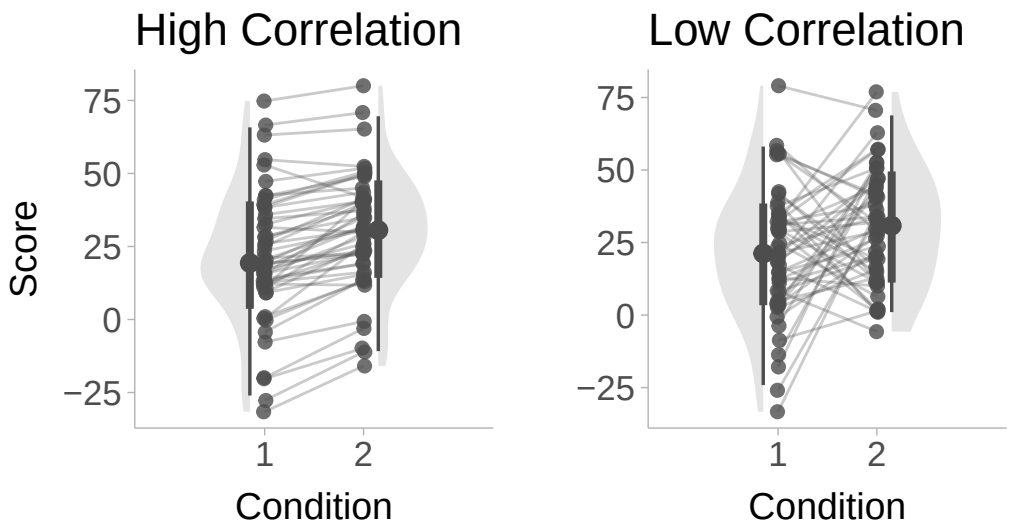


Figure 7.2: Figure displaying simulated data of a repeated measures design, the x-axis shows the condition (e.g., pre-test and post-test) and y-axis is the scores. Left panel shows a high pre/post correlation (r = .95) and right panel shows a low correlation condition (r = .05). Lines indicate within person pre/post change.

period. In these types of 2x2 (group x time) study designs it is usually the difference between the standardized mean change for the intervention/treatment (T) and control (C) groups that is of interest. For a visualization of a pretest-posttest-control group design see Figure 7.3.

Morris (2008) details three effect sizes for this pretest-posttest-control (PPC).

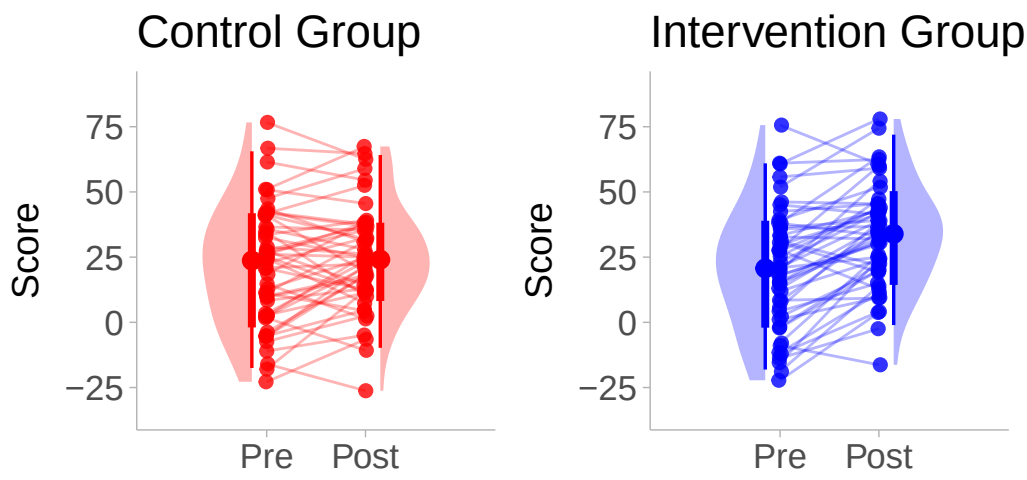


Figure 7.3: Illustration of a pre-post control design. Left panel shows the pre-post difference in the control group and right panel shows the pre-post difference in the intervention/treatment group. Lines indicate within person pre/post change.

7.5.1 PPC1 - separate pre-test standard deviations

The separate pre-test (i.e., baseline) standard deviations are used to standardize the pre/post mean difference in the intervention group and the control group respectively (see equation 4, Morris 2008),

$$d_T = \frac{M_{T,\text{post}} - M_{T,\text{pre}}}{S_{T,\text{pre}}} \tag{7.22}$$

$$d_C = \frac{M_{C,\text{post}} - M_{C,\text{pre}}}{S_{C,\text{pre}}} \tag{7.23}$$

Note that these effect sizes are identical to the Becker's d formulation of the SMD (see Section 7.4.4). Therefore the pretest-posttest-control group effect size is simply the difference between the intervention and control pre/post SMD (equation 15, Becker 1988),

$$d_{PPC1} = d_T - d_C \tag{7.24}$$

The asymptotic standard error of d_{PPC1} was first derived by Becker (1988) and can be expressed as the square root of the sum of the sampling variances (equation 16, Becker 1988),

$$SE_{d_{PPC1}} = \sqrt{\left[\frac{2(1-r_T)}{n_T} + \frac{d_T}{2n_T}\right] + \left[\frac{2(1-r_C)}{n_C} + \frac{d_C}{2n_C}\right]}. \tag{7.25}$$

Note that this is an approximate formula for the standard error, for an exact solution see Morris (2000). We can calculate d_{PPC1} and it's confidence intervals using the `metafor` package:

```
# Example:

# Control Group (N = 90)
## Pre-test Mean = 20, SD = 6
## Post-test Mean = 25, SD = 7
## Pre/post correlation = .50

# Intervention Group (N = 90)
## Pre-test Mean = 20, SD = 5
## Post-test Mean = 27, SD = 8
## Pre/post correlation = .50


# calculate the observed standardized mean difference
# treatment group effect
dT <- escalc(measure = "SMCR",
             m1i = 27,
             m2i = 20,
             sd1i = 8,
             sd2i = 5,
             ri = .50,
             ni = 90)
```

```r
# control group effect
dC <- escalc(measure = "SMCR",
             m1i = 25,
             m2i = 20,
             sd1i = 7,
             sd2i = 6,
             ri = .50,
             ni = 90)

# calculate d and SE
dPPC1 <- dT$yi - dC$yi
SE <- sqrt(dT$vi + dC$vi)


# print the d value and confidence intervals
data.frame(d = round(dPPC1,3),
           SE = round(SE,3),
           ci.lb = round(dPPC1 - 1.96*SE,3),
           ci.ub = round(dPPC1 + 1.96*SE,3))
```

```
     d    SE  ci.lb ci.ub
1 0.159 0.171 -0.176 0.494
```

The output shows a pre-post intervention effect of d_{PPC1} = 0.16 95% CI [-0.18, 0.49].

7.5.2 PPC2 - pooled pre-test standard deviations

The pooled pre-test (i.e., baseline) standard deviations can be used to standardized the difference in pre/post change between intervention and control groups such that (equation 8, Morris 2008),

$$d_{PPC2} = \frac{(M_{T,\text{post}} - M_{T,\text{pre}}) - (M_{C,\text{post}} - M_{C,\text{pre}})}{S_{p,\text{pre}}} \tag{7.26}$$

where

$$S_{p,\text{pre}} = \sqrt{\frac{(n_T - 1)S^2_{T,\text{pre}} + (n_C - 1)S^2_{C,\text{pre}}}{n_T + n_C - 2}}. \tag{7.27}$$

The distribution of d_{PPC2} was described by Morris (2008) and can be expressed as (adapted from equation 24, Morris 2008),

$$SE_{d_{PPC2}} = \sqrt{2\left(1-\hat{\rho}\right)\left(\frac{n_T+n_C}{n_T n_C}\right)\left(\frac{n_T+n_C-2}{n_T+n_C-4}\right)\left[1+\frac{d_{PPC2}^2}{2\left(1-\hat{\rho}\right)\left(\frac{n_T+n_C}{n_T n_C}\right)}\right]-\frac{d_{PPC2}^2}{CF^2}}. \tag{7.28}$$

Note the original equation shown in the paper by Morris (2008) uses the population pre/post correlation ρ, however in the equation above we replace ρ with the sample size weighted average of the Pearson correlation in the treatment and control group (i.e., $\hat{\rho} = \frac{n_T r_T + n_C r_C}{n_T + n_C}$). Also, CF is the correction factor that can be found in the following section on small sample bias.

We can use base R to obtain d_{PPC2} and confidence intervals:

```
# Example:

# Control Group (N = 90)
## Pre-test Mean = 20, SD = 6
## Post-test Mean = 25, SD = 7
## Pre/post correlation = .50
M_Cpre <- 20
M_Cpost <- 25
SD_Cpre <- 6
SD_Cpost <- 7
rC <- .50
nC <- 90

# Intervention Group (N = 90)
## Pre-test Mean = 20, SD = 5
## Post-test Mean = 27, SD = 8
## Pre/post correlation = .50
M_Tpre <- 20
M_Tpost <- 27
SD_Tpre <- 5
SD_Tpost <- 8
rT <- .50
nT <- 90

# calculate the observed standardized mean difference
dPPC2 <- ((M_Tpost- M_Tpre) - (M_Cpost - M_Cpre)) / sqrt( ( (nT - 1)*(SD_Tpre^2) + (nC - 1
```

```r
# calculate the standard error
rho <- (nT*rT+nC*rC)/(nT + nC)
CF <- gamma((nT+nC-2)/2) / ( sqrt((nT+nC-2)/2) * gamma(((nT+nC-2)-1)/2) )
SE <-  sqrt(2*(1-rho) * (nT+nC)/(nT*nC) * (nT+nC-2)/(nT+nC-4) * (1 + (dPPC2^2 / (2*(1 - rh

# print the d value and confidence intervals
data.frame(d = round(dPPC2,3),
           SE = round(SE,3),
           ci.lb = round(dPPC2 - 1.96*SE,3),
           ci.ub = round(dPPC2 + 1.96*SE,3))
```

```
     d    SE ci.lb ci.ub
1 0.362 0.151 0.066 0.658
```

The output shows a pre-post intervention effect of d_{PPC2} = 0.36 95% CI [0.07, 0.66].

7.5.3 PPC3 - pooled pre- and post-test

The two previous effect sizes (PPC1 and PPC2) only use the pretest standard deviation and ignore the post-test standard deviation. However, if we are happy to assume that pretest and posttest variances are homogeneous[1] the pooled pre-test and post-test standard deviations can be used to standardize the difference in pre/post change between intervention and control groups, such that (equation 8, Morris 2008),

$$d_{PPC3} = \frac{(M_{T,\text{post}} - M_{T,\text{pre}}) - (M_{C,\text{post}} - M_{C,\text{pre}})}{S_{p,\text{pre-post}}}, \tag{7.29}$$

where,

$$S_{p,\text{pre-post}} = \sqrt{\frac{(n_T - 1)\left(S^2_{T,\text{pre}} + S^2_{T,\text{post}}\right) + (n_C - 1)\left(S^2_{C,\text{pre}} + S^2_{C,\text{post}}\right)}{2(n_T + n_C - 2)}}. \tag{7.30}$$

The standard error for d_{PPC3} is currently unknown. An option to estimate this standard error is to use a non-parametric or parametric bootstrap by repeatedly sampling the raw data,

[1] Note, this may not be the case especially where there is a mean-variance relationship and one (usually the intervention) group has a higher posttest mean score.

or if the raw data is not available resample simulated data. We can do this in base R by simulating pre/post data using the `mvrnorm()` function from the `MASS` package (Venables and Ripley 2002):

```
# Install the package below if not done so already
# install.packages(MASS)

# Example:

# Control Group (N = 90)
## Pre-test Mean = 20, SD = 6
## Post-test Mean = 25, SD = 7
## Pre/post correlation = .50
M_Cpre <- 20
M_Cpost <- 25
SD_Cpre <- 6
SD_Cpost <- 7
rC <- .50
nC <- 90

# Intervention Group (N = 90)
## Pre-test Mean = 20, SD = 5
## Post-test Mean = 27, SD = 8
## Pre/post correlation = .50
M_Tpre <- 20
M_Tpost <- 27
SD_Tpre <- 5
SD_Tpost <- 8
rT <- .50
nT <- 90

# simulate data
set.seed(1) # set seed for reproducibility
boot_dPPC3 <- c()
for(i in 1:1000){
  # simulate control group pre-post data
  data_C <- MASS::mvrnorm(n = nC,
                          # input observed means
                          mu = c(M_Cpre,M_Cpost),
                          # input observed covariance matrix
                          Sigma = data.frame(pre = c(SD_Cpre^2, rC*SD_Cpre*SD_Cpost),
                                             post = c(rC*SD_Cpre*SD_Cpost,SD_Cpost^2)))
```

```
  # simulate intervention group pre-post data
  data_T <- MASS::mvrnorm(n = nT,
                          # input observed means
                          mu = c(M_Tpre,M_Tpost),
                          # input observed covariance matrix
                          Sigma = data.frame(pre = c(SD_Tpre^2, rT*SD_Tpre*SD_Tpost),
                                             post = c(rT*SD_Tpre*SD_Tpost,SD_Tpost^2)))

  # calculate the mean difference in pre/post change (the numerator)
  MeanDiff <- (mean(data_T[,2]) - mean(data_T[,1])) - (mean(data_C[,2]) - mean(data_C[,1]))

  # calculate the pooled pre-post standard deviation (the denominator)
  S_Pprepost <-  sqrt( ( (nT - 1)*(sd(data_T[,1])^2+sd(data_T[,2])^2) + (nC - 1)*(sd(data_

  # calculate the standardized mean difference for each bootstrap iteration
  boot_dPPC3[i] <- MeanDiff / S_Pprepost
}

# calculate bootstrapped standard error
SE <- sd(boot_dPPC3)

# calculate the observed standardized mean difference
dPPC3 <- ((M_Tpost- M_Tpre) - (M_Cpost - M_Cpre)) / sqrt( ( (nT - 1)*(SD_Tpre^2+SD_Tpost^2

#print the d value and confidence intervals
data.frame(d = round(dPPC3,3),
           SE = round(SE,3),
           ci.lb = round(dPPC3 - 1.96*SE,3),
           ci.ub = round(dPPC3 + 1.96*SE,3))
```

```
      d    SE ci.lb ci.ub
1 0.303 0.153 0.003 0.604
```

The output shows a pre-post intervention effect of d_{PPC3} = 0.30 95% CI [0.003, 0.60].

7.6 Small Sample Bias in d values

All the estimators of d listed above are biased estimates of the population d value, specifically they all over-estimate the population value in small sample sizes. To adjust for this bias, we can apply a correction factor based on the degrees of freedom. The degrees of freedom will

largely depend on the estimator used. The degrees of freedom for each estimator is listed below:

- Single Group design (d_s): $df = n - 1$
- Between Groups - Pooled Standard Deviation (d_p): $df = n_A + n_B - 2$
- Between Groups - Control Group Standard Deviation (d_Δ): $df = n_B - 1$
- Repeated Measures - all types (d_z, d_{rm}, d_{av}, d_b): $df = n - 1$
- Pretest-Posttest-Control Separate Standard Deviation (d_{PPC1}): $df_C = n_C - 1,\ df_T = n_T - 1$
- Pretest-Posttest-Control Pooled Pretest Standard Deviation (d_{PPC2}): $df = n_T + n_C - 2$
- Pretest-Posttest-Control Pooled Pretest and Posttest Standard Deviation (d_{PPC3}): $df = 2(n_T + n_C - 2)$

With the appropriate degrees of freedom, we can use the following correction factor, CF, to obtain an unbiased estimate of the population standardized mean difference:

$$CF = \frac{\Gamma\left(\frac{df}{2}\right)}{\Gamma\left(\frac{df-1}{2}\right)\sqrt{\frac{df}{2}}} \tag{7.31}$$

Where $\Gamma(\cdot)$ is the gamma function. An approximation of this complex formula given by Hedges (1981) can be written as $CF \approx 1 - \frac{3}{4 \cdot df - 1}$. In R, this can be calculated using,

```
# Example (independent groups d_p):
# Group 1 sample size = 20
# Group 2 sample size = 18
n1 <- 20
n2 <- 18

# calculate degrees of freedom
df <- n1 + n2 - 2

# calculate correction factor
CF <- gamma(df/2) / ( sqrt(df/2) * gamma((df-1)/2) )

# print
CF
```

```
[1] 0.9789964
```

This correction factor can then be applied to any of the standardized mean difference variants mentioned above,

$$d^* = d \times CF \tag{7.32}$$

The corrected d value, d^*, is commonly referred to as Hedges' g or just g. To avoid notation confusion we will just add an asterisk to d to denote the correction. Note that in the case of d_{PPC1}, we must apply CF to both d_C and d_T such that, $d^*_{PPC1} = d_T \times CF_T - d_C \times CF_C$. We also need to correct the standard error for d^* using the same correction factor,

$$SE_{d^*} = SE_d \times CF \tag{7.33}$$

These standard errors can then be used to calculate the confidence interval of the corrected d value,

$$CI_{d*} = d^* \pm 1.96 \times SE_{d^*} \tag{7.34}$$

It is very important to note that the `escalc()` function automatically applies the small sample correction by default, therefore any code that utilizes `escalc()` *do not* also apply the correction factor.

```
# Example:
# Cohen's d = .50, SE = .10

d = .50
SE = .10

# correct d value and CIs small sample bias
d_corrected <- d * CF
SE_corrected <- SE * CF
ci.lb_corrected <- d_corrected - 1.96*SE_corrected
ci.ub_corrected <- d_corrected + 1.96*SE_corrected

# print just the d value and confidence intervals
data.frame(d = round(d_corrected,3),
           SE = SE_corrected,
           ci.lb = round(ci.lb_corrected,3),
           ci.ub = round(ci.ub_corrected,3))
```

```
      d         SE ci.lb ci.ub
1 0.489 0.09789964 0.298 0.681
```

The output shows that the corrected effect size is d^* = 0.50, 95% CI [0.30, 0.68].

7.7 Ratios of Means

Another common approach, particularly within the fields of ecology and evolution, is to take the natural logarithm of the ratio between two means; the so-called Response Ratio (LRR). This is sometimes more favorable as, due to its construction using the standard deviation in some form as a denominator, the various versions of standardized mean differences are impacted by the estimate of this parameter for which studies are often less powered compared to mean magnitudes (Yang et al. 2022). For the LRR however the standard deviation only impacts its variance estimation and not the point estimate. A limitation of the LRR however is that it is limited to data that are observed on a ratio scale (i.e., have an absolute zero and instances of it are related ordinally and additively meaning both means will be positive).

Although strictly speaking the LRR is not a difference in means in an additive sense as the above standardized mean difference effect sizes are, it can in one sense be considered to reflect the difference in means on the multiplicative scale. In fact, after calculation it is often transformed to reflect the percentage difference or change between means: $100 \times \exp(LRR) - 1$. However, this can introduce transformation induced bias because a non-linear transformation of a mean value is not generally equal to the mean of the transformed value. In the context of meta-analysis, when combining LRR estimates across studies a correction factor can be applied: $100 \times \exp(LRR + 0.5S^2_{\text{total}}) - 1$, where S^2_{total} is the variance of all LRR values.

Similarly to the various standardized mean differences, there are varied calculations for the LRR dependent upon the study design being used (see Senior, Viechtbauer, and Nakagawa 2020).

7.7.1 Response Ratio for Independent Groups (LRR_{ind})

When calculating the response ratio for two independent groups (group A and B). The LRR can be calculated as follows,

$$LRR_{\text{ind}} = \ln\left(\frac{M_A}{M_B}\right) + CF \tag{7.35}$$

Where M_A and M_B are the means for group A and B, respectively. CF is the small sample correction factor calculated as,

$$CF = \frac{S_A^2}{2n_A M_A^2} - \frac{S_B^2}{2n_B M_B^2}. \tag{7.36}$$

Where n_A and n_B are the sample sizes. The standard error can be calculated as,

$$SE_{LRR_{\text{ind}}} = \sqrt{\frac{S_A^2}{n_A M_A^2} + \frac{S_B^2}{n_B M_B^2} + \frac{S_A^4}{2n_A^2 M_A^4} + \frac{S_B^4}{2n_B^2 M_B^4}} \tag{7.37}$$

Using R we can easily calculate this effect size using the `escalc()` function in the `metafor` package (Viechtbauer 2010):

```
# LRR for two independent groups
# given means and SDs

# For example:
# Group A Mean = 30.4, Standard deviation = 22.53, Sample size = 96
# Group B Mean = 21.4, Standard deviation = 19.59, Sample size = 96

# calculate lnRRind and standard error
LRRind <- escalc(measure = "ROM",
                 m1i = 30.4,
                 m2i = 21.4,
                 sd1i = 22.53,
                 sd2i = 19.59,
                 n1i = 96,
                 n2i = 96)

summary(LRRind)
```

```
     yi     vi    sei     zi   pval  ci.lb  ci.ub
1 0.3511 0.0145 0.1202 2.9203 0.0035 0.1154 0.5867
```

The example shows a response ratio of LRR_{ind} = 0.35 95% CI [0.12, 0.59].

7.7.2 Response Ratio for Dependent Groups (LRR_{dep})

When we have dependent samples (e.g., a pre/post comparison), the LRR can be calculated as follows,

$$LRR_{\text{dep}} = \ln\left(\frac{M_2}{M_1}\right) + CF \tag{7.38}$$

Where M_1 and M_2 are the means for conditions 1 and 2, respectively. Where CF is the small sample correct factor calculated as,

$$CF = \frac{S_2^2}{2nM_2^2} - \frac{S_1^2}{2nM_1^2} \tag{7.39}$$

The standard error can then be calculated as,

$$SE_{LRR_{\text{dep}}} = \sqrt{\frac{S_1^2}{nM_1^2} + \frac{S_2^2}{nM_2^2} + \frac{S_1^4}{2n^2M_1^4} + \frac{S_2^4}{2n^2M_2^4} + \frac{2rS_1S_2}{nM_1M_2} + \frac{r^2S_1^2S_2^2(M_1^4 + M_2^4)}{2n^2M_1^4M_2^4}} \tag{7.40}$$

Using R we can easily calculate this effect size using the `escalc()` function from the `metafor` package as follows:

```
# LRR for two dependent groups
# given means and SDs


# For example:
# Mean 1 = 30.4, Standard deviation 1 = 22.53
# Mean 2 = 21.4, Standard deviation 2 = 19.59
# Sample size = 96
# Correlation = 0.4

# calculate lnRR and standard error
LRRdep <- escalc(measure = "ROMC",
                 m1i = 30.4,
                 m2i = 21.4,
                 sd1i = 22.53,
                 sd2i = 19.59,
                 ni = 96,
```

```
                  ri = .40)

summary(LRRdep)
```

```
      yi     vi    sei     zi   pval  ci.lb  ci.ub
1 0.3511 0.0088 0.0938 3.7429 0.0002 0.1672 0.5349
```

The example shows a response ratio of LRR_{dep} = 0.35 95% CI [0.17, 0.53].

8 Correlation between Two Continuous Variables

To quantify the relationship between two continuous variables, the most common method is to use a Pearson correlation coefficient (denoted with the letter r). The pearson correlation takes the covariance between a continuous independent (X) and dependent (Y) variable and standardizes it by the standard deviations of X and Y,

$$r = \frac{\text{Cov}(X, Y)}{S_X S_Y}.$$

We can visualize what a correlation between two variables looks like with scatter plots. Figure 8.1 shows scatter plots with differing levels of correlation.

The standard error of the Pearson correlation coefficient is,

$$SE_r = \sqrt{\frac{\left(1 - r^2\right)^2}{n - 1}}$$

Unlike Cohen's d and other effect size measures, The correlation coefficient is bounded by -1 and positive 1, with positive 1 being a perfectly positive correlation, -1 being a perfectly negative correlation, and zero indicating no correlation between the two variables. The bounding has the consequence of making the confidence interval asymmetric around r (e.g., if the correlation is positive, the lower bound is farther away from r than the upper bound is). It is important to note that with a correlation of zero, the confidence interval is symmetric and approximately normal. To obtain the confidence intervals of r, we first need to apply a Fisher's Z transformation. A Fisher's Z transformation is a hyperbolic arctangent transformation of a Pearson correlation coefficient and can be computed as,

$$Z_r = \text{arctanh}(r)$$

The Fisher Z transformation ensures Z_r has a symmetric and approximately normal sampling distribution. This then allows us to calculate the confidence interval from the standard error of Z_r ($SE_{Z_r} = \frac{1}{\sqrt{n-3}}$),

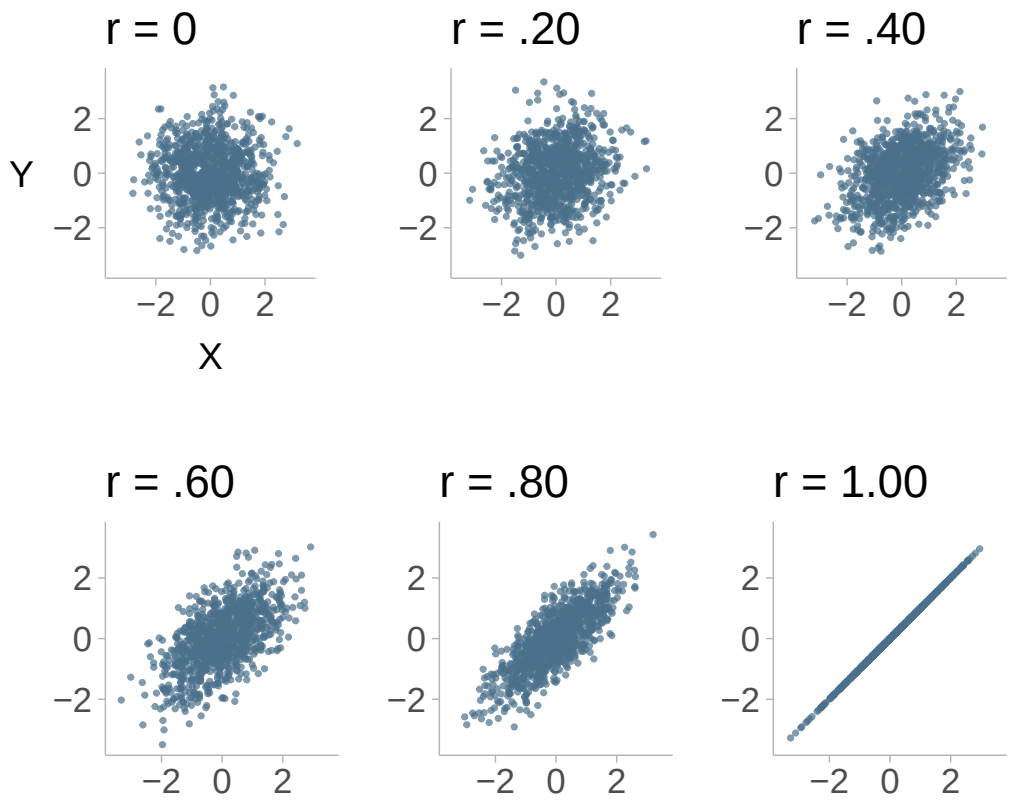


Figure 8.1: Simulated data from a bivariate normal distribution displaying 6 different correlations, r = 0, .20, .40, .60, .80, and 1.00.

$$CI_r = Z_r \pm 1.96 \times SE_{Z_r}$$

We can also back-transform the confidence interval of the Fisher's Z_r value to obtain the confidence of the Pearson correlation,

$$CI_r = \tanh(CI_{Z_r})$$

In R, the full process of obtaining the correlation and confidence intervals can be done quite easily with base R using the `cor.test()` function. Let's use the `palmerpenguins` data set (Horst, Hill, and Gorman 2020) to look at the correlation between flipper length and body mass.

```
library(palmerpenguins)

# compute pearson correlation
cor.test(x = penguins$flipper_length_mm,
         y = penguins$body_mass_g,
         use = "pairwise.complete.obs", # use rows with non-missing data
         method = "pearson",
         conf.level = .95)
```

```
    Pearson's product-moment correlation

data:  penguins$flipper_length_mm and penguins$body_mass_g
t = 32.722, df = 340, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.843041 0.894599
sample estimates:
      cor
0.8712018
```

The output shows that the correlation and its confidence intervals are r = 0.87, 95% CI [0.84, 0.89]. Note that you can also use the `cor()` function to simply get the correlation without all the extra information. We can also use the `escalc()` function in R to compute the correlation and CIs if all we have is the correlation and sample size (i.e., no raw data).

```
library(metafor)
```

```r
# get pearson correlation and sample size
r <- .871
n <- 341

# calculate z-transformed correlation
stats <- escalc(measure = "ZCOR",
                ri = r,
                ni = n)

# get pearson correlation and CI
summary(stats, transf = tanh)
```

```
     yi  ci.lb  ci.ub
1 0.8710 0.8428 0.8945
```

8.1 Rank-Order (Spearman) Correlation

If we want to get the monotonic relationship between two variables rather than the strictly linear relationship, we can look at the correlation between ranks of X and Y. A Spearman rank-order correlation is actually just the Pearson correlation on ranks such that,

$$r_s = \frac{\text{Cov}(\text{rank}(X), \text{rank}(Y))}{S_{\text{rank}(X)} S_{\text{rank}(Y)}}$$

Viewing Figure 8.2 we can see that non-linear monotonic relationships are captured well with Spearman's correlation, whereas Pearson's correlation is only describing the linear relationship between the two.

We can use base R to calculate Spearman correlations. Let's use the `palmerpenguins` data set (Horst, Hill, and Gorman 2020) to look at the correlation between flipper length and body mass.

```r
library(palmerpenguins)

# compute spearman correlation
cor(x = penguins$flipper_length_mm,
    y = penguins$body_mass_g,
    use = "pairwise.complete.obs", # use rows with non-missing data
    method = "spearman")
```

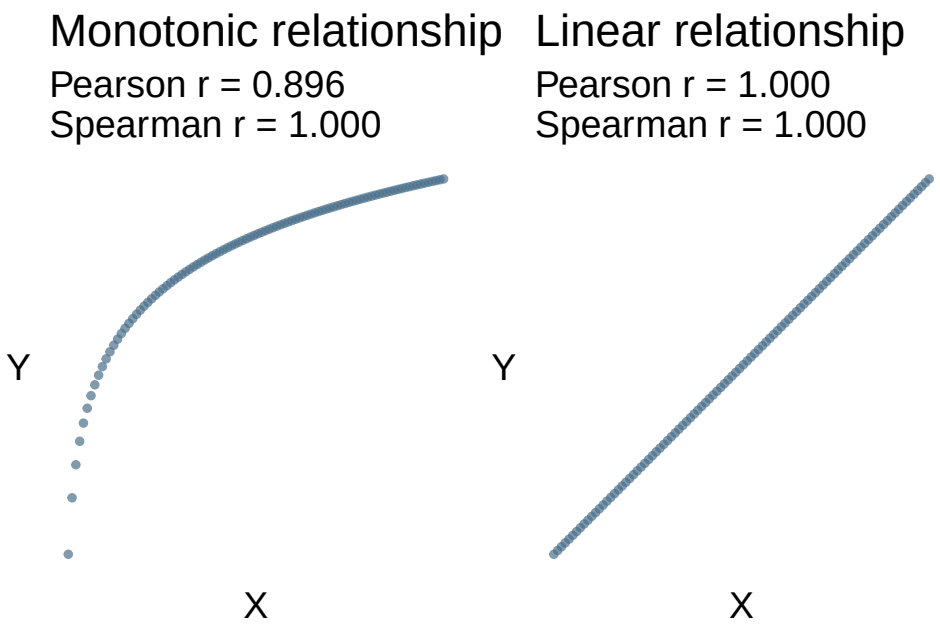


Figure 8.2: Scatter plots for two directly related variables. Left panel shows a monotonic relationship between X and Y. Right panel shows a linear relationship between X and Y

```
[1] 0.8399741
```

8.2 Concordance-Discordance (Kendall) Correlation

Similar to Spearman's rank-order correlation, Kendall's correlation (Kendall 1945) is a rank-based correlation that measures the ordinal association between variables. The equation for Kendall's correlation is expressed as,

$$r_\tau = \frac{2}{n(n-1)} \sum_{i<j} \text{sign}(x_i - x_j) \cdot \text{sign}(y_i - y_j)$$

Where $i = 1...n$ and $j = 1...n$ denote the observations and $\text{sign}(\cdot)$ returns the sign of whatever is in the parentheses such that a positive value returns 1, a negative value returns –1 and zero returns 0. Similar to Pearson's and Spearman's correlations we can compute Kendall's correlation in base R. We will use the `palmerpenguins` data set again.

```
library(palmerpenguins)

# compute kendall correlation
cor(x = penguins$flipper_length_mm,
    y = penguins$body_mass_g,
    use = "pairwise.complete.obs", # use rows with non-missing data
    method = "kendall")
```

```
[1] 0.6604675
```

9 Effect Sizes for Categorical Variables

For relationships between categorical variables, there are many variations of effect sizes that one can use. Commonly used effect size measures for statistical procedures on categorical data include: phi coefficient (ϕ), Cramer's V, Cohen's h, Cohen's ω, odds ratio (OR), risk difference (RD), and relative risk (RR).

Here is a table for every effect size discussed in this chapter:

Type	Description	Section
ϕ - phi coefficient	Pearson correlation between two binary variables (i.e., 2x2 contingency tables).	Section 9.2.1
V - Cramer's *V*	Measures the association between categorical variables. Similar to a ϕ coefficient, but meant for contingency tables larger than 2x2.	Section 9.2.2
h - Cohen's *h*	Pearson correlation between two binary variables. Difficult to interpret.	Section 9.2.3
w - Cohen's *w*	Association between two categorical variables and it is computed identically to the ϕ coefficient. If computed on a 2x2 contingency table, it will have an identical value to ϕ.	Section 9.2.4
- Ben-Shachar's Fei	A correction to Cohen's w for one dimensional count tables.	Section 9.2.5
OR - Odds Ratio	Ratio of odds of an event occurring between treatment and control groups	Section 9.2.6
RD - Risk Difference	Difference between proportions in treatment and control groups.	Section 9.2.7
RR - Relative Risk	Ratio of proportions in the treatment and control groups.	Section 9.2.8

9.1 One Sample Proportion Test

A one sample proportion test is used when we want to assess the difference between an observed proportion p and some fixed proportion of interest p_0. In this case like this we can obtain the test statistic with the following formula:

$$z = \frac{p - p_0}{\sqrt{\frac{p(1-p)}{n}}}, \tag{9.1}$$

where n is the sample size. Note that this is only valid if the proportion of interest is chance ($p_0 = .50$) because the sampling distribution with a proportion of .50 is normal. However if the proportion of interest is not .50, then we should instead compute Cohen's h (see Section 9.2.3 for details), which transforms the scale so that the distributions are normal regardless of the proportion. The test-statistic with Cohen's h,

$$z = h\sqrt{n} \tag{9.2}$$

Let's try testing the proportion against chance ($p_0 = .50$) in R. We can then calculate the *p*-value in base R by using the `pnorm()` function:

```
# Example:
p <- .7 # observed proportion
p0 <- .5 # proportion of interest
n <- 50 # sample size

# test statistic
z <- (p-p0) / sqrt(p*(1-p)/n)

# two tailed test p-value
pval <- 2*(1-pnorm(z))

# output results
data.frame(z,pval)
```

```
        z        pval
1 3.086067 0.002028231
```

Results show a significant difference from chance with z = 3.09 and *p*-value = .002.

9.2 Effect Sizes

9.2.1 Phi Coefficient (ϕ)

Phi coefficient (ϕ) is a measure of association between two binary variables (therefore, it ONLY applies to 2 by 2 contingency tables, i.e., each variable has only two levels). It is a special case of the Pearson correlation coefficient and an r for two binary variables is equal to phi. Note that unlike r that ranges from -1 to 1, ϕ ranges from 0 to 1. Also, the sign of r indicates the direction of association, whereas to get the direction of an association given a 2 by 2 contingency table, we need to look at the table itself; ϕ only provides a measure of strength. The 2 by 2 contingency table is illustrated by Table 9.2.

Table 9.2: Contingency table between two binary variables

	$X = 0$	$X = 1$
$Y = 0$	n_{00}	n_{10}
$Y = 1$	n_{01}	n_{11}

The sample sizes within each cell provide us with the necessary information to estimate the relationship between the two variables. A large phi coefficient would be expected to have relatively large sample sizes in the diagonal cells (n_{00} and n_{11}) and relatively low sample sizes in the off-diagonal cells (n_{01} and n_{10}). To calculate phi, it can be calculated from the cells of the contingency table directly (adapted from equation 1, Guilford 1965),

$$\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{(n_{00} + n_{01})(n_{10} + n_{11})(n_{00} + n_{10})(n_{01} + n_{11})}} \tag{9.3}$$

or more conveniently, from the χ^2-statistic (equation 7.2.5, Cohen 1988),

$$\phi = \sqrt{\frac{\chi^2}{n}} \tag{9.4}$$

Where n is the total sample size (i.e., the sum of all the cells). Using the `effectsize` package in R, we can calculate the the phi coefficient using the `phi` function directly from the contingency table:

```
library(effectsize)

# Example contingency table:
#  40  17
```

```r
#  11  45

# contingency table
contingency_table <- matrix(c(40, 11,
                              17, 45),ncol = 2)

# calculate phi
phi(contingency_table, alternative = "two.sided")
```

```
Phi (adj.) |       95% CI
-------------------------
0.50       | [0.31, 0.69]
```

In our example we obtained a phi coefficient of ϕ = .50 95% CI [0.31, 0.69]. Note that `phi()` applies a small sample correction factor by default, this is preferable, but if you want to remove it you can add the following argument: `adjust=FALSE`.

9.2.2 Cramer's V

Cramer's V, sometimes also referred to as Cramer's phi (ϕ), is a generalized effect size measure of the association between two nominal variables. It applies to contingency tables of any size (2×2, 3×3, 3×4, 5×3, etc.). Cramer's V on a 2×2 contingency table is equivalent to the phi coefficient. For an illustration of a higher order contingency table, Table 9.3 represents a 3×4 contingency table of two variables.

Table 9.3: Contingency table between two categorical variables

	$X = 0$	$X = 1$	$X = 2$	$X = 3$
$Y = 0$	n_{00}	n_{10}	n_{21}	n_{31}
$Y = 1$	n_{01}	n_{11}	n_{21}	n_{31}
$Y = 2$	n_{02}	n_{12}	n_{22}	n_{32}

The value of Cramer's V ranges from 0 to 1 and can interpreted in a similar way to the phi coefficient. Again we can use the χ^2 statistic to compute the value, however, since there can be more than 2 levels to each variable, we also need to take into account the number of levels, k, of the variable with the least number of levels (e.g., a 3×4 contingency table, k would be equal to 3). Cramer's V is defined as (equation 7.2.6, Cohen 1988),

$$V = \sqrt{\frac{\chi^2}{n(k-1)}} \tag{9.5}$$

Note that Cramer's V reduces to the phi coefficient for 2×2 contingency tables. The standard error of a Cramer's V is similar to that of a Pearson correlation and a ϕ coefficient.

$$SE_V = \sqrt{\frac{(1-V^2)^2}{n-1}} \tag{9.6}$$

Where n is the total sample size (i.e., the sum of all cells). Like the pearson correlation, we can not calculate the confidence interval directly from the standard error, instead, we must convert V to a Fisher's Z statistic, $Z_V = \text{arctanh}(V)$. We can then calculate the 95% confidence interval for V by back-transforming the confidence interval for Z_V:

$$SE_{Z_V} = \frac{1}{\sqrt{n-3}} \tag{9.7}$$

$$CI_V = \tanh(Z_V \pm 1.96 \times SE_{Z_V}) \tag{9.8}$$

Using the `cramers_v` function in the `effectsize` package (Ben-Shachar, Lüdecke, and Makowski 2020), we can calculate Cramer's V and it's 95% confidence interval using the Fisher's Z method described above. For the example, we can example data from a 3 $\times$ 3 contingency table.

```
# Example contingency table:
#  40  14  12
#  11  27   9
#   5  10  34

library(effectsize)

# contingency table
contingency_table <- matrix(c(40, 11,  5,
                              14, 27, 10,
                              12,  9, 34),ncol = 3)

# calculate cramer's v
cramers_v(contingency_table,
          alternative = "two.sided")
```

```
Cramer's V (adj.) |       95% CI
--------------------------------
0.43              | [0.30, 0.53]
```

In our example we obtained a Cramer's V of V = .43 95% CI [.30, .53]. Note that `cramers_v()` applies a small sample correction factor by default, this is preferable, but if you want to remove it you can add the following argument: `adjust=FALSE`.

9.2.3 Cohen's h

Cohen's h is a measure of distance between two proportions or probabilities. It is sometimes also referred to as the "difference between arcsines". For a given proportion p_1, its arcsine transformation is given by (equation 6.2.1, Cohen 1988):

$$\psi_1 = 2 \cdot \text{arcsin}(\sqrt{p_1}). \tag{9.9}$$

Cohen's h is the difference between the arcsine transformations of two proportions (equation 6.2.2, Cohen 1988):

$$h = \psi_1 - \psi_2 \tag{9.10}$$

Cohen's h is commonly used for the power analysis of proportion tests. In fact, it is the required effect size measure in the program *G Power* (Faul et al. 2009). We can calculate the standard error of Cohen's h,

$$SE_h = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{9.11}$$

Since the sampling distribution of h is symmetric, we can calculate the confidence intervals from the standard error,

$$CI_h = h \pm 1.96 \times SE_h \tag{9.12}$$

To calculate Cohen's h, we can use the `cohens_h()` function in the `effectsize` package in R.

```
# Example proportions: p1 = .45, p2 = .30

# contingency table
```

```r
contingency_table <- matrix(c(40, 11,
                              14, 27),ncol = 2)

# calculate cohen's h
cohens_h(contingency_table)
```

```
Cohen's h |       95% CI
------------------------
0.93      | [0.52, 1.34]
```

From the example, the R code outputted a Cohen's h value of h = .93 95% CI [0.52, 1.34].

9.2.4 Cohen's w

Cohen's w is a measure of association analogous to the phi coefficient, but on tables that are larger than 2x2. Although Cohen's w is useful for power analyses, it is not so useful as a stand-alone effect size. As Cohen (1988) states (pp. 221):

> As a measure of association, [Cohen's w] lacks familiarity and convenience

Cohen's w has the exact same formula as the phi coefficient with the only difference being that the χ^2 statistic comes from a contingency table of any size (equation 7.2.5, Cohen 1988),

$$w = \sqrt{\frac{\chi^2}{n}} \tag{9.13}$$

And can also be calculated directly from Cramer's V (equation 7.2.7, Cohen 1988),

$$w = V \times \sqrt{k - 1} \tag{9.14}$$

Where k is the number of categories in the variable with the least number of categories. We can use the `cohens_w()` function in the `effectsize` package (Ben-Shachar, Lüdecke, and Makowski 2020).

```r
# Example contingency table
# 40 14 15
# 11 27 10
#  5 22 30

# create contingency table
```

```r
contingency_table <- matrix(c(40, 11, 15,
                              14, 27, 10,
                               5, 22, 30),ncol = 3)

# calculate cohen's w
cohens_w(contingency_table,
         alternative = "two.sided")
```

```
Cohen's w |       95% CI
------------------------
0.53      | [0.36, 0.66]
```

```r
## Calculate cohen's w from cramer's v
# get cramer's v
v <- cramers_v(contingency_table,
               alternative = "two.sided",
               adjust = FALSE)$Cramers_v

# number of categories in variable with least categories
k <- min(dim(contingency_table))

# calculate cohen's w from cramer's v
(w <- v * sqrt(k-1))
```

```
[1] 0.5277187
```

From the example code, the `cohens_w` function returned Cohen's w value of w = .45 95% CI [0.24, 0.65]. We were also able to recover Cohen's w from Cramer's V.

9.2.5 Ben-Shachar's Fei ( )

Ben-Shachar et al. (2023) introduced a new effect size for one-dimensional tables of counts/proportions that they label with the Hebrew letter, .פ Ben-Shachar's פ is a correction to Cohen's w that adjusts for the expected value and consequently bounds the value between 0 and 1. The equation for פ is defined as,

$$\text{פ} = \sqrt{\frac{\chi^2}{n\left(\frac{1}{\min(P_E)} - 1\right)}} \tag{9.15}$$

Where min(P_E) is the smallest expected probability. The formula for Ben-Schachar's ɒ can be also be expressed in terms of Cohen's ω,

$$\text{ɒ} = \frac{\omega}{\sqrt{\left(\frac{1}{\max(P_E)} - 1\right)}} \tag{9.16}$$

In R, we can calculate Ben-Shachar's ɒ using the `fei()` function in the `effectsize` package (Ben-Shachar, Lüdecke, and Makowski 2020).

```
# Example:
# Observed counts: 20, 50, 100 (observed proportions: .12, .29, .59)
# Expected proportions: .5, .2, .3

# observed counts
observed_counts <- c(20,50,100)

# expected probabilities
expected_probabilities <- c(.5,.2,.3)

# calculate fei
fei(observed_counts,
    p = expected_probabilities,
    alternative = "two.sided")
```

```
Fei  |       95% CI
------------------
0.39 | [0.31, 0.47]

- Adjusted for uniform expected probabilities.
```

From the example code, the `fei` function returned Ben-Shachar's ɒ value of .39 95% CI [0.31, 0.47].

9.2.6 Odds Ratio (OR)

Odds ratio measures the effect size between two binary variables. It is commonly used in medical and behavioral intervention research, and notably, in meta-analysis.

Let's imagine a study conducted to investigate the association between smoking and the development of major depressive disorder (MDD). The study includes a sample of 251 individuals,

categorizing them into two groups: 125 smokers and 126 non-smokers. The researchers are interested in understanding the odds of having major depressive disorder (MDD) among smokers compared to non-smokers. Say we find that 25 smokers were diagnosed with MDD while 100 were not, but in the non-smoker group, 12 individuals were diagnosed with MDD while 120 were not. The odds ratio would then be:

$$OR = \frac{25/100}{12/120} = \frac{.25}{.10} = 2.50 \tag{9.17}$$

In general, we can can compute the odds-ratio from a contingency table between binary variables X (i.e., the treatment) and Y (i.e., the outcome; see Table 9.4).

Table 9.4: Contingency table between two binary variables

	$X = T$	$X = C$
$Y = 0$	n_{T0}	n_{C0}
$Y = 1$	n_{T1}	n_{C1}

Ultimately, we want to compare the outcome between the treatment group (T) and the control group (C). Therefore we can compute the odds ratio as,

$$OR = \frac{n_{T1}/n_{T0}}{n_{C1}/n_{C0}} \tag{9.18}$$

The standard distribution of the odds-ratio is asymmetric. To calculate confidence intervals, we can first convert the odds ratio to a log odds ratio ($LOR = \log(OR)$). Then we can calculate the standard error of the log odds ratio,

$$SE_{LOR} = \sqrt{\frac{1}{n_{T0}} + \frac{1}{n_{T1}} + \frac{1}{n_{C0}} + \frac{1}{n_{C1}}} \tag{9.19}$$

With the standard error of the log odds ratio we can then calculate the confidence interval of the odds ratio by back-transforming using the exponential function,

$$CI_{OR} = \exp(LOR \pm 1.96 \times SE_{LOR}) \tag{9.20}$$

In R, we can use the `metafor` package (Viechtbauer 2010) to calculate the odds ratio and it's confidence interval:

```r
# load in metafor package
library(metafor)
```

```
Loading required package: Matrix
```

```
Loading required package: metadat
```

```
Loading required package: numDeriv
```

```
Loading the 'metafor' package (version 4.8-0). For an
introduction to the package please type: help(metafor)
```

```r
# Example:
# Treatment Group: 10 diseased, 43 healthy
# Control Group:  24 diseased, 41 healthy

# compute log odds ratio
LOR <- escalc(measure = "OR",
              ai = 10,
              bi = 43,
              ci = 24,
              di = 41)

# get estimate and CI of odds ratio
summary(LOR, transf = exp)
```

```
     yi  ci.lb  ci.ub
1 0.3973 0.1693 0.9321
```

The code output for this example shows an odds ratio of OR = 0.40 95% CI [0.17, 0.93].

9.2.7 Risk Difference (RD)

Risk difference can be used to interpret the difference between two proportions. We can use the contingency table from Table 9.4, and calculate a risk difference between the treatment group and the control group. First, calculate the proportion of cases where the outcome is $Y = 1$ *within* the control group and the treatment group:

$$p_C = \frac{n_{C1}}{n_{C0} + n_{C1}} = \frac{n_{C1}}{n_C} \tag{9.21}$$

$$p_T = \frac{n_{T1}}{n_{T0} + n_{T1}} = \frac{n_{T1}}{n_T} \tag{9.22}$$

Then using these proportions we can calculate the risk difference (RD),

$$RD = p_T - p_C. \tag{9.23}$$

The corresponding standard error is,

$$SE_{RD} = \sqrt{\frac{p_T(1-p_T)}{n_T} + \frac{p_C(1-p_C)}{n_C}} \tag{9.24}$$

Where n_T and n_C are the total sample sizes *within* the treatment and control group, respectively. Using the standard error we can compute the 95% confidence intervals,

$$CI_{RD} = RD \pm 1.96 \times SE_{RD} \tag{9.25}$$

The risk difference formula is fairly simple, so we can compute it using base R.

```
# Example:
# Treatment Group: 10 diseased, 43 healthy, 53 total
# Control Group:  24 diseased, 41 healthy, 65 total

# calculate risk difference
stats <- escalc(measure = "RD",
                ai = 10, n1i = 53,
                ci = 24, n2i = 65,
                var.names=c("RD", "variance"))

# display results
```

```
summary(stats)
```

```
       RD variance    sei      zi   pval   ci.lb   ci.ub
1 -0.1806   0.0065 0.0804 -2.2444 0.0248 -0.3382 -0.0229
```

The risk difference in this example is RD = .15 95% CI [-.06, -.36].

9.2.8 Relative Risk (RR)

The relative risk, often referred to as the "risk ratio," calculates the ratio between the proportion of cases in the treatment group and the proportion of cases in the control group. It provides a straightforward interpretation: "individuals receiving the treatment have a RR times higher odds of experiencing the outcome compared to controls." To calculate relative risk, first we need to calculate the proportion of outcome cases in the treatment and control group,

$$p_C = \frac{n_{C1}}{n_{C0} + n_{C1}} = \frac{n_{C1}}{n_C},$$
$$p_T = \frac{n_{T1}}{n_{T0} + n_{T1}} = \frac{n_{T1}}{n_T}.$$

Then we can calculate the relative risk,

$$RR = \frac{p_T}{p_C}. \tag{9.26}$$

To get the confidence intervals. We will need to first get the standard error of the log relative risk $LRR = \log(RR)$.The corresponding standard error of the log relative risk can be computed as,

$$SE_{LRR} = \sqrt{\frac{1}{n_{C1}} + \frac{1}{n_{T1}} - \frac{1}{n_T} - \frac{1}{n_C}} \tag{9.27}$$

The confidence intervals can be computed from the standard error,

$$CI_{RR} = \exp(LRR \pm 1.96 \times SE_{LRR}) \tag{9.28}$$

To compute relative risk, we can simply use the equations above in base R.

```r
# Example:
# Treatment Group: 10 diseased, 43 healthy, 53 total
# Control Group:  24 diseased, 41 healthy, 65 total

# calculate log relative risk
stats <- escalc(measure = "RR",
                ai = 10, n1i = 53,
                ci = 24, n2i = 65,
                var.names=c("RR", "variance"))

# display results
summary(stats, transf = exp)
```

```
     RR  ci.lb  ci.ub
1 0.5110 0.2688 0.9714
```

The example shows a relative risk of $RR = 0.51 95% CI [0.32, 0.70].

10 Effect Sizes for ANOVAs

10.1 ANOVAs

For ANOVAs/F-tests, you will always need to report two kinds of effects: the omnibus effect of the factor(s) and the effect of planned contrasts or post hoc comparisons.

For instance, imagine that you are comparing three groups/conditions with a one-way ANOVA. The ANOVA will first return an F-statistic, the degrees of freedom, and the associated p-value. Here, you need to calculate the size of this omnibus factor effect as either eta-squared, partial eta-squared, or generalized eta-squared.

Suppose the omnibus effect is significant. You now know that there is at least one group that differs from the others. However, you may want to know which group(s) differ from the others, and how much they differ. Therefore, you conduct post hoc comparisons on these groups. Post hoc comparisons can be conducted pairwise between each group (or level of a factor), which allows you to obtain a t-statistic and p-value for each comparison. Consequently, you can calculate and report a standardized mean difference for each comparison of interest.

Imagine that you are conducting a two-by-two factorial ANOVA for a treatment outcome following two drug interventions (drug A and B).

Table 10.1: Experimental conditions

	Drug A	No Drug A
Drug B	A and B	B
No Drug B	A	none

You can obtain the omnibus effects that encapsulates the main effects and interaction effects for these drug interventions on the outcome. Note that lower-order effects are not directly interpretable given higher-order (interaction) effects. Depending on your research question, you can compare the outcome between any two of the four conditions in Table 10.1, reporting the standardized mean difference. It's these pairwise standardized mean differences that allow us to interpret lower order effects.

10.2 ANOVA tables

An ANOVA table generally consists of the grouping factors (+ residuals), the sum of squares, the degrees of freedom, the mean square, the F-statistic, and the p-value. Using base R, we can construct an ANOVA table using the `aov()` function to generate the ANOVA model and then using `summary.aov()` to extract the table. For an example case, we will use the `palmerpenguins` data set package and we will investigate the differences in the body mass (the outcome) of three penguin species (the predictor/grouping variable):

```
library(palmerpenguins)

# construct anova model
# formula structure: outcome ~ grouping variable
ANOVA_mdl <- aov(body_mass_g ~ species,
                 data = penguins) # dataset

# extract ANOVA table
ANOVA_table <- summary.aov(ANOVA_mdl)
ANOVA_table
```

```
             Df    Sum Sq  Mean Sq F value Pr(>F)
species       2 146864214 73432107   343.6 <2e-16 ***
Residuals   339  72443483   213698
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
2 observations deleted due to missingness
```

By default, `summary.aov()` does not report the η^2 value (i.e., the variance explained by species), however we will discuss this more in Section 10.7.1. The results show that the mean body mass between the three penguin species (Adelie, Gentoo, Chinstrap) differ significantly from one another (see Figure 10.1).

10.3 One-way between-subjects ANOVA

One-way between-subject ANOVA is an extension of independent-samples t-tests. The null hypothesis is that all k means of k independent groups are identical, whereas the alternative hypothesis is that there are at least one mean that differs from the other groups. The assumptions include: (1) independence of observations, (2) normality of residuals, and (3) equality (or homogeneity) of variances (homoscedasticity).[1]

[1] There are variants of ANOVAs that can have each of these assumptions violated.

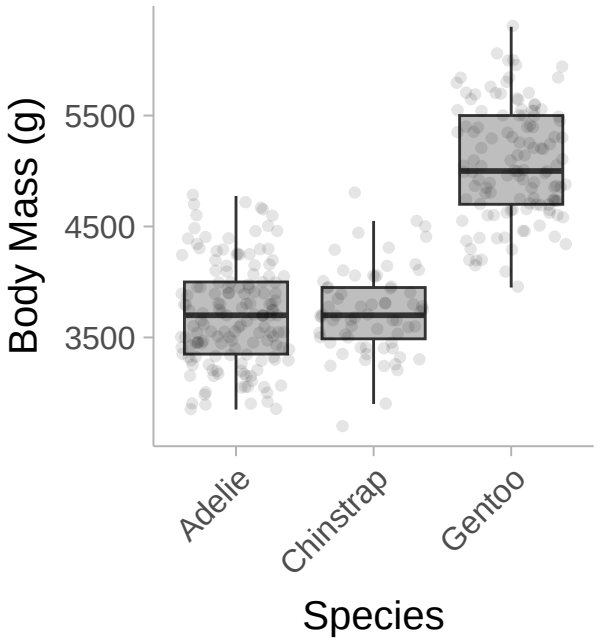


Figure 10.1: Box plots showing the distribution of body mass by species.

Note, sometimes you may encounter a between-subject one-way ANOVA which compares only two conditions, particularly when the paper is old. This is essentially a t-test, and the F-statistic is just t^2. It is preferable to report Cohen's d as a effect size in two group comparisons as opposed to η^2 and, fortunately, we can convert η^2 to a Cohen's d. It is recommended that subsequent analyses (e.g., power analysis) of two group comparisons are based on Cohen's d.

10.3.1 Determining degrees of freedom for one-way ANOVAs

Please refer to the following table to determine the degrees of freedom for ANOVA effects, if they are not reported or if you suspect that they have been misreported.

Degrees of freedom	
Between subjects ANOVA	
Effect	$k-1$
Error	$n-k$
Total	$n-1$

10.3.2 Calculating eta-squared from F-statistic and degrees of freedom

Using the formula below, we can calculate η^2 of an ANOVA model using the F-statistic and the degrees of freedom,

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

Where df_{error} and df_{effect} denotes the degrees of freedom for the residual error term and the main effect, respectively. In R, we can use the `F_to_eta2()` function from the `effectsize` package (Ben-Shachar, Lüdecke, and Makowski 2020):

```
library(effectsize)

n = 154 # number of subjects
k = 3 # number of groups
f = 84.3 # F-statistic

# get degrees of freedom
df_effect = k - 1
df_error = n - k

# calculate eta-squared
F_to_eta2(f = f,
          df = df_effect,
          df_error = df_error,
          alternative = 'two.sided') # obtain two sided CIs
```

```
Eta2 (partial) |       95% CI
------------------------------
0.53           | [0.42, 0.61]
```

The results of this example show an η^2 value of .53 95% CI [.42, .61]. This suggests that the grouping variable accounts for 53% of the total variation in the outcome.

10.3.3 Calculating eta-squared from an ANOVA table

Let's use the table from the ANOVA model in Section 10.2 that pertains to the effect of penguin species on body mass:

Table 10.3: One-way ANOVA table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
species	2	146864214	73432107.1	343.6263	0
Residuals	339	72443483	213697.6	NA	NA

From this table we can use the sum of squares from the grouping variable (species) and the total sum of squares ($SS_{\text{total}} = SS_{\text{effect}} + SS_{\text{error}}$) to calculate the η^2 value using the following equation:

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

In R, we can use the `eta.full.SS()` or `eta.F()` function in the `MOTE` package (Buchanan et al. 2019) to obtain η^2 from an ANOVA table.

```
library(MOTE)

# calculate eta-squared with eta.full.SS
eta <- eta.full.SS(dfm = 2,  # effect degrees of freedom
                   dfe = 339, # error degrees of freedom
                   ssm = 146864214, # sum of squares for the effect
                   sst = 146864214 + 72443483, # total sum of squares
                   Fvalue = 343.6263,
                   a = .05)

# display results
data.frame(eta_squared = round(eta$eta,3),
           ci.lb = round(eta$etalow,3),
           ci.ub = round(eta$etahigh,3))
```

```
  eta_squared ci.lb ci.ub
1        0.67 0.606 0.722
```

```
# alternative approach with eta.F
eta <- eta.F(dfm = 2,  # effect degrees of freedom
             dfe = 339, # error degrees of freedom
             Fvalue = 343.6263,
             a = .05)
```

```r
  # display results
  data.frame(eta_squared = round(eta$eta,3),
             ci.lb = round(eta$etalow,3),
             ci.ub = round(eta$etahigh,3))
```

```
  eta_squared ci.lb ci.ub
1        0.67 0.606 0.722
```

The example code outputs η^2 = .67 [.61, .72]. This suggests that species accounts for 67% of the total variation in body mass between penguins.

10.3.4 Calculating Cohen's d for post-hoc comparisons

In an omnibus ANOVA, the p-value is telling us whether the means from all groups come from the same population mean, however this does not inform us about *which* group means differ or by how much. Using the same example as before, let's say we want to answer a specific question such as: what is the difference in body mass between Adelie penguins and Gentoo penguins? To answer this question, we can calculate the raw mean difference between the two groups. In R, we can do that with the following code:

```r
  library(tidyverse)

  # get means for each species
  stats <- penguins %>%
    summarize(mean = mean(body_mass_g, na.rm = TRUE),
              .by = species)

  # display means
  stats
```

```
# A tibble: 3 x 2
  species    mean
  <fct>     <dbl>
1 Adelie    3701.
2 Gentoo    5076.
3 Chinstrap 3733.
```

```r
  # get raw mean difference
  stats$mean[2] - stats$mean[1]
```

```
[1] 1375.354
```

Based on the mean difference, Gentoo penguins are on average 1375 grams heavier than Adelie penguins in total body mass. We can also calculate a standardized mean difference using the `escalc()` function in the `metafor` package (Viechtbauer 2010).

```
library(metafor)

# Means, SDs, and sample sizes for each species
stats <- penguins %>%
  summarize(mean = mean(body_mass_g, na.rm = TRUE),
            sd = sd(body_mass_g, na.rm = TRUE),
            n = n(),
            .by = species)

# show table of stats
stats
```

```
# A tibble: 3 x 4
  species    mean    sd     n
  <fct>     <dbl> <dbl> <int>
1 Adelie    3701.  459.   152
2 Gentoo    5076.  504.   124
3 Chinstrap 3733.  384.    68
```

```
# calculate standardized mean difference
effect_size <- escalc(measure = 'SMD',
                      m1i = stats$mean[2],
                      m2i = stats$mean[1],
                      sd1i = stats$sd[2],
                      sd2i = stats$sd[1],
                      n1i = stats$n[2],
                      n2i = stats$n[1],
                      var.names = c("d","variance"))

# display results
summary(effect_size)
```

```
      d variance    sei      zi   pval  ci.lb  ci.ub
1 2.8602   0.0295 0.1716 16.6629 <.0001 2.5237 3.1966
```

The standardized mean difference between Adelie and Gentoo penguins is d = 2.86 95% CI [2.52, 3.19], demonstrating that Gentoo penguins have body mass 2.86 standard deviations larger than Adelie penguins.

We can also quantify contrasts from summary statistics reported from the ANOVA table and the group means (labeled group A and B). We can calculate the standardized mean difference using the means from both groups and the mean squared error (MSE; i.e., the mean square of the residuals) the following equation:

$$d = \frac{M_A - M_B}{\sqrt{MSE}}$$

This method gives a standardized mean difference equivalent to the Cohen's d with the pooled standard deviation in the denominator (see chapter on mean differences). Therefore if we obtain the mean squared errors from Section 10.3.3 and we obtain the means (means: Gentoo = 5076, Adelie = 3701), we can calculate the standardized mean difference as: $\frac{5076-3701}{\sqrt{213697.6}} = \frac{1375}{462.27} = 2.974$. The discrepency between the standardized mean difference provided by the `escalc()` function is due to the fact that the function automatically applies a small sample correction factor (i.e., Hedges' g) thus reducing the overall effect (see the small sample bias section in the mean differences chapter).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
species	2	146864214	73432107.1	343.6263	0
Residuals	339	72443483	213697.6	NA	NA

Beware the assumptions.

Note that this method is ONLY valid when you are willing to assume equal variances among groups (homoscedasticity), and when you conduct a Fisher's one-way ANOVA (rather than Welch's). This method is also impractical if you are calculating from reported statistics, and MSE is not reported (which is typically the case).
If you are unwilling to assume homogeneity of variances, you should know it also that it also makes little sense to conduct a Fisher's ANOVA in such situations. You may want to switch to Welch's ANOVA, which does not assume homoscedasticity. For comparisons between groups with unequal variances you may want to use an alternative standardized effect size measure, such as Glass' delta.

10.4 One-way repeated measures ANOVA

One-way repeated measures ANOVA (rmANOVA) is an extension of paired-samples t-tests, with the difference being it can be used in two or more groups.

10.4.1 Determining degrees of freedom for one-way rmANOVA

Please refer to the following table to determine the degrees of freedom for repeated measure ANOVA effects.

Degrees of freedom	
Within-subject ANOVA (repeated measures)	
Effect	$k - 1$
Error-between	$(n - 1) \times (k - 1)$
Error-within	$(n - 1) \cdot (k - 1)$
Total (within)	$n \cdot (k - 1)$

10.4.2 Eta-squared from rmANOVA statistics

Commonly, we use eta-squared (η^2) or partial eta-squared (η_p^2) as the effect size measure for one-way rmANOVAs, for which these two are in fact equal. Let's construct an rmANOVA model using example data from the `datarium` package (Kassambara 2019). The `selfesteem` data set simply shows self-esteem scores over three repeated measurements within the same subjects (see Figure 10.2).

Let's get the ANOVA results in R,

```
# load in data
data("selfesteem", package = "datarium")

# reformat data
selfesteem <- selfesteem %>%
  pivot_longer(cols = c("t1","t2","t3"),
               names_to = "time",
               values_to = "self_esteem") %>%
  rename(subject = id)

# display data
head(selfesteem)
```

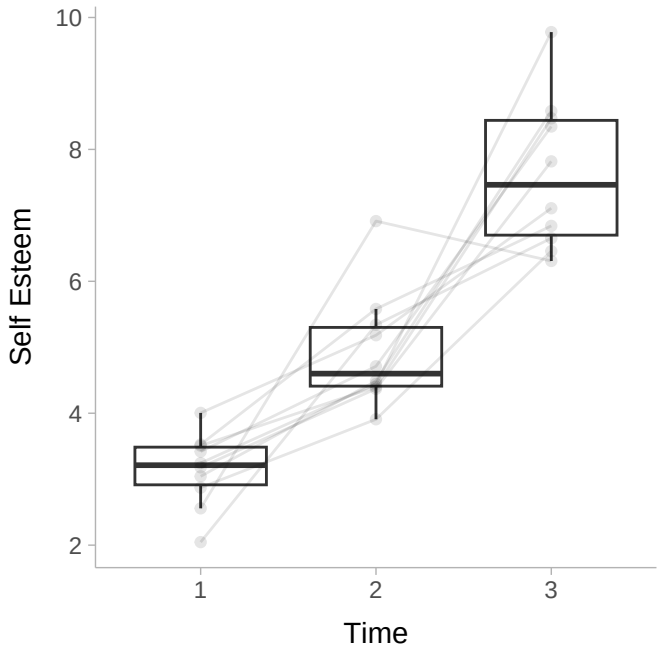


Figure 10.2: Box plot of self esteem at each time point. Lines connect subjects across time points.

```
# A tibble: 6 x 3
  subject time  self_esteem
    <int> <chr>       <dbl>
1       1 t1           4.01
2       1 t2           5.18
3       1 t3           7.11
4       2 t1           2.56
5       2 t2           6.91
6       2 t3           6.31
```

```
# fit ANOVA
rmANOVA_mdl = aov(formula = self_esteem ~ time + Error(subject),
                  data = selfesteem)

# display ANOVA table
summary(rmANOVA_mdl)
```

```
Error: subject
```

```
          Df  Sum Sq Mean Sq F value Pr(>F)
Residuals  1 0.07667 0.07667

Error: Within
          Df Sum Sq Mean Sq F value   Pr(>F)
time       2 102.46   51.23   63.07 1.06e-10 ***
Residuals 26  21.12    0.81
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There are two tables displayed here, the table on top displays the between subject effects and the table below shows the within subject effects. The equations and functions to calculate η^2 mentioned in the one-way between-subjects ANOVAs section also apply here:

$$\eta^2 = \frac{df_{\text{effect}} \times F}{df_{\text{effect}} \times F + df_{\text{error-within}}},$$

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

Note that here SS_{total} does not include $SS_{\text{error-between}}$ because we are not interested in the between-subject variance since we are conducting a rmANOVA. In other words, between-subjects variance can be large or small, but we do not care about it when we examine whether there is an effect within subjects. Therefore the total sum of squares can be defined as

$$SS_{\text{total}} = SS_{\text{effect}} + SS_{\text{error-within}}.$$

As a result, we can calculate η^2 from the rmANOVA table as,

$$\eta^2 = \frac{102.46}{21.12 + 102.46} = .83$$

We can plug the rmANOVA model into the `eta_squared()` function from the `effectsize` package in R (Ben-Shachar, Lüdecke, and Makowski 2020) to calculate η^2.

```r
library(effectsize)

# calculate eta-squared
eta_squared(rmANOVA_mdl,
            alternative = "two.sided")
```

```
# Effect Size for ANOVA (Type I)

Group  | Parameter | Eta2 (partial) |       95% CI
-------------------------------------------------
Within |      time |           0.83 | [0.69, 0.89]
```

As expected, we find the same point-estimate from our hand calculation. To calculate η^2 from the F-statistic and degrees of freedom we can use the `MOTE` package (Buchanan et al. 2019) as we did in Section 10.3.3:

```
library(MOTE)

# calculate eta squared
eta <- eta.full.SS(dfm = 2,  # effect degrees of freedom
                   dfe = 26, # error degrees of freedom
                   ssm = 102.46, # sum of squares for the effect
                   sst = 102.46 + 21.12, # total sum of squares
                   Fvalue = 63.07,
                   a = .05)

# display results
data.frame(eta_squared = round(eta$eta,3),
           ci.lb = round(eta$etalow,3),
           ci.ub = round(eta$etahigh,3))
```

```
  eta_squared ci.lb ci.ub
1       0.829 0.644  0.91
```

The results show η^2 = .83 95% CI [.64, .91]. Note there is a slight discrepancy in the calculation of confidence intervals between the `MOTE` and `effectsize` package.

10.5 Two-way between-subjects ANOVA

Two-way between-subjects ANOVA is used when there are two predictor grouping variables in the model. Note again that "between-subjects" means that each group contains different subjects.

10.5.1 Determining degrees of freedom

Please refer to the following table to determine the degrees of freedom for two-way ANOVA effects (Morse 2018). Note that k_1 is the number of groups in the first variable, and k_2 is the number of groups in the second variable.

Degrees of freedom	
Within subjects ANOVA	
Main Effect (of one variable)	$k_1 - 1$ or $k_2 - 1$
Interaction Effect	$(k_1 - 1) \times (k_2 - 1)$
Error	$n - k_1 \cdot k_2$
Total	$n - 1$

10.5.2 Eta-squared from two-way ANOVA statistics

For two-way ANOVAs we can obtain the partial eta-squared (η_p^2) for each predictor in the model. Let's construct our ANOVA model using data from the `palmerpenguins` dataset (Horst, Hill, and Gorman 2020). In this example we want to see how the species and the sex of the penguin explains variance in body mass (see Figure 10.3).

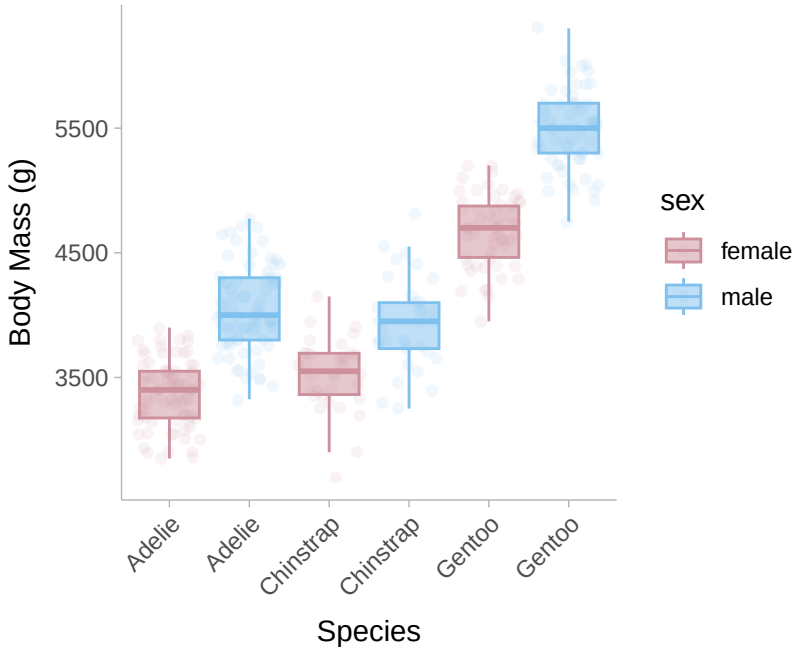


Figure 10.3: Box plots of body mass for each penguin species and sex.

Let's load in the data set and fit the two-way ANOVA model and display the table.

```
library(palmerpenguins)

ANOVA2_mdl <- aov(body_mass_g ~ species + sex + species:sex,
                  data = penguins)

summary(ANOVA2_mdl)
```

```
             Df    Sum Sq  Mean Sq F value   Pr(>F)
species       2 145190219 72595110 758.358  < 2e-16 ***
sex           1  37090262 37090262 387.460  < 2e-16 ***
species:sex   2   1676557   838278   8.757 0.000197 ***
Residuals   327  31302628    95727
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
11 observations deleted due to missingness
```

The results show that species, sex, and the interaction between the two account for substantial variance in body mass. We can obtain the contributions of species, sex, and their interaction by computing the partial eta-squared value (η_p^2) for each. To do this, we can use similar formulas to η^2 from the one-way ANOVA. The difference between the formulas for η_p^2 and η^2 is that η_p^2 does not use the total sum of squares in the denominator, instead it uses the residual sum of squares (SS_{error}) and the sum of squares from the effect of interest (SS_{effect}). For example,

$$\text{For species:} \quad \eta_p^2 = \frac{SS_{\text{effect}}}{SS_{\text{effect}} + SS_{\text{error}}} = \frac{145190219}{145190219 + 31302628} = .82$$

$$\text{For sex:} \quad \eta_p^2 = \frac{SS_{\text{effect}}}{SS_{\text{effect}} + SS_{\text{error}}} = \frac{37090262}{37090262 + 31302628} = .54$$

$$\text{For sex} \times \text{species:} \quad \eta_p^2 = \frac{SS_{\text{effect}}}{SS_{\text{effect}} + SS_{\text{error}}} = \frac{1676557}{1676557 + 31302628} = .05$$

We can also easily do this in R using the `eta_squared` function in the `effectsize` package (Ben-Shachar, Lüdecke, and Makowski 2020) and setting the argument `partial = TRUE`.

```
# calculate partial eta-squared
eta_squared(ANOVA2_mdl,
            partial = TRUE,
```

```
              alternative = "two.sided")
```

```
# Effect Size for ANOVA (Type I)

Parameter   | Eta2 (partial) |       95% CI
-----------------------------------------
species     |           0.82 | [0.79, 0.85]
sex         |           0.54 | [0.48, 0.60]
species:sex |           0.05 | [0.01, 0.10]
```

The results show the η_p^2 value and 95% CIs for each term in the ANOVA model.

10.6 Two-way repeated measures ANOVA

A two-way repeated measures ANOVA (rmANOVA) would indicate that subjects are exposed to each condition along two variables.

10.6.1 Determing degrees of freedom

Please refer to the following table to determine the degrees of freedom for two-way rmANOVA effects (Morse 2018). Note that k_1 is the number of groups in the first variable, and k_2 is the number of groups in the second variable.

Degrees of freedom	
Between subjects ANOVA	
Main Effect (of one variable)	$k_1 - 1$ or $k_2 - 1$
Interaction Effect	$(k_1 - 1) \times (k_2 - 1)$
Error-between	$(k_1 \cdot k_2) - 1$
Error-within	$(n - 1) \times (k_1 \cdot k_2 - 1)$
Total	$n - 1$

10.6.2 Eta-squared from Two-way rmANOVA

For a two-way repeated measures ANOVA, we can use the `weightloss` data set from the `datarius` package (Kassambara 2019). This data set contains a diet condition and a control condition that tracked subjects across time (3 time points) for each of condition (see Figure 10.4).

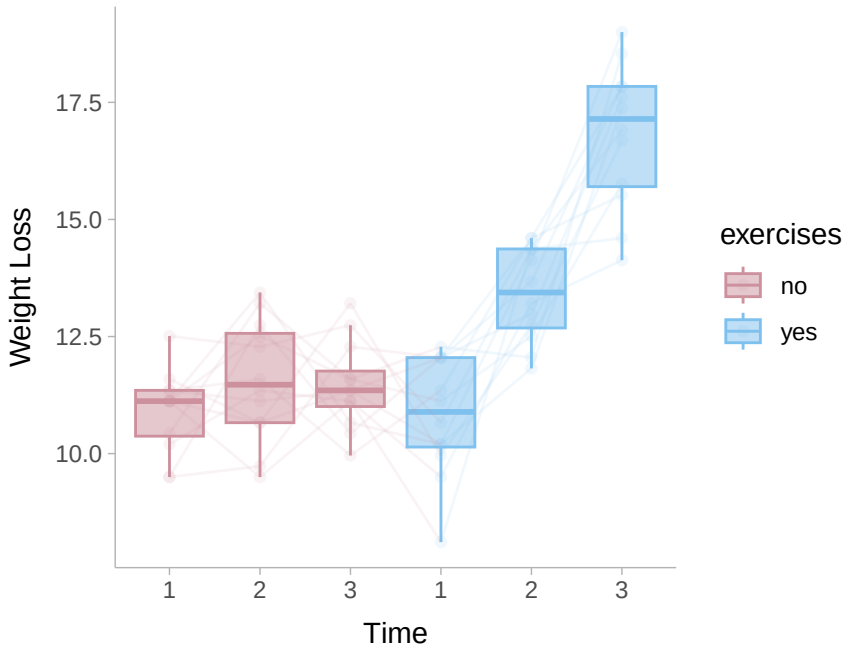


Figure 10.4: Box plots showing the weight loss across time points and conditions. Lines connect subjects across conditions.

Let's conduct a two-way rmANOVA in R.

```
### load in data
data("weightloss", package = "datarium")

# reformat data
weightloss <- weightloss  %>%
  pivot_longer(cols = c("t1","t2","t3"),
               names_to = "time",
               values_to = "weight_loss") %>%
  rename(subject = id) %>%
  filter(diet=='no')

# fit ANOVA model
rmANOVA2_mdl = aov(formula = weight_loss ~ time + exercises + time:exercises + Error(subje
                   data = weightloss)

# display ANOVA table
summary(rmANOVA2_mdl)
```

```
Error: subject
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals 11  20.64   1.877

Error: Within
               Df Sum Sq Mean Sq F value   Pr(>F)
time            2 129.26   64.63   50.57 3.45e-13 ***
exercises       1 101.03  101.03   79.05 3.16e-12 ***
time:exercises  2  92.55   46.28   36.21 9.26e-11 ***
Residuals      55  70.29    1.28
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the table and graph above, we can see that there is substantial within-person change in weight loss under the exercise condition and no discernible increase in weight loss without exercising. This suggests that there is a substantial interaction effect. Like we did in the between-subjects two-way ANOVA, we can calculate the partial eta squared values from the ANOVA table

$$\text{For time:} \quad \eta_p^2 = \frac{SS_{\text{effect}}}{SS_{\text{effect}} + SS_{\text{error-within}}} = \frac{129.26}{129.26 + 70.29} = .65$$

$$\text{For exercise:} \quad \eta_p^2 = \frac{SS_{\text{effect}}}{SS_{\text{effect}} + SS_{\text{error-within}}} = \frac{101.03}{101.03 + 70.29} = .59$$

$$\text{For sex} \times \text{species:} \quad \eta_p^2 = \frac{SS_{\text{effect}}}{SS_{\text{effect}} + SS_{\text{error-within}}} = \frac{92.55}{92.55 + 70.29} = .57$$

Remember for the partial eta-squared, the denominator is not the total sum of squares rather it is the effect sum of squares and the error. In the repeated measures ANOVA, the error should only be for the within subject error because the variance between subjects is not something we are interested in. We can also calculate this in R using the `eta_squared()` function again.

```r
# calculate partial eta-squared
eta_squared(rmANOVA2_mdl,
            partial = TRUE,
            alternative = "two.sided")
```

```
# Effect Size for ANOVA (Type I)

Group  |      Parameter | Eta2 (partial) |       95% CI
```

```
-----------------------------------------------------
Within |           time |           0.65 | [0.49, 0.75]
Within |      exercises |           0.59 | [0.42, 0.70]
Within | time:exercises |           0.57 | [0.39, 0.69]
```

The results show the η_p^2 value and 95% CIs for each term in the rmANOVA model.

10.7 Effect Sizes for ANOVAs

ANOVA (Analysis of Variance) is a statistical method used to compare means across multiple groups or conditions. It is mostly used when the outcome variable is continuous and the predictor variables are categorical. Commonly used effect size measures for ANOVAs / F-tests include: eta-squared (η^2), partial eta-squared (η_p^2), generalized eta-squared (η_G^2), omega-squared (ω^2), partial omega-squared (ω), generalized omega-squared (ω_G^2), Cohen's f.

Type	Description	Section
η^2 - eta-squared	Measures the variance explained of the whole ANOVA model.	Section 10.7.1
η_p^2 - Partial eta-squared	Measures the variance explained by a specific factor in the model.	Section 10.7.2
η_G^2 - Generalized eta-squared	Similar to η^2, but uses the sum of squares of all non-manipulated variables in the calculation. This allows meta-analysts to compare η_G across different designs.	Section 10.7.3
$\omega^2, \omega_p^2, \omega_G^2$ - Omega squared corrections	Corrections to bias observed in η^2 measures. Can be interpreted in the same way as η^2.	Section 10.7.4
f - Cohen's *f*	This effect size can be interpreted as the average Cohen's d between each group.	Section 10.7.5

10.7.1 Eta-Squared (η^2)

Eta-squared is the ratio between the between-group variance and the total variance. It describes the proportion of the total variability in the data that are accounted for by a factor. Therefore, it is a measure of *variance explained*. To calculate eta-squared (η^2) we need to first calculate the total sum of squares (SS_{total}) and the effect sum of squares (SS_{effect}),

$$SS_{\text{total}} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{10.1}$$

Where $\bar{y}$ is the grand mean (i.e., the mean of all data points collapsed across groups) and y_i is the outcome value for observation i. To calculate the sum of squares of the effect, we can use predicted outcome values ($\widehat{y}_i$) rather than the raw outcome values (y_i). In the case of categorical predictors, $\widehat{y}_i$ is equal to the mean of the outcome *within* that individual's respective group. Therefore the sum of squares of the effect can be calculated using the following formula:

$$SS_{\text{effect}} = \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2. \tag{10.2}$$

Now we can calculate the η^2 value,

$$\eta^2 = \frac{SS_{\text{effect}}}{SS_{\text{total}}} \tag{10.3}$$

The standard error of η^2 can be approximated from Olkin and Finn (1995):

$$SE_{\eta^2} = \sqrt{\frac{4\eta^2\left(1-\eta^2\right)^2\left(n+k-1\right)^2}{\left(n^2-1\right)\left(3+n\right)}} \tag{10.4}$$

The sampling distribution for η^2 is asymmetric as all the values are bounded in the range, 0 to 1. The confidence interval surrounding η^2 will likewise be asymmetric so instead of calculating the confidence interval from the standard error, we can instead use a non-central F-distribution using the degrees of freedom between groups (e.g., for three groups: $df_b = k - 1 = 3 - 1 = 2$) and the degrees of freedom within groups (e.g., for 100 subjects and three groups: $df_b = n - k = 100 - 3 = 97$) to obtain the confidence intervals (this is done automatically in the `effectsize` package). Another option is to use bootstrapping procedure (i.e., resampling the observed data points to construct a sampling distribution around η^2, see Kirby and Gerlanc 2013) and then take the .025 and .975 quantiles of that distribution.

In R, we can calculate η^2 from a one-way ANOVA using the `palmerpenguins` data set. The `aov` function in base R allows the analyst to model an ANOVA with categorical predictors on the right side (species) of the `~` and the outcome on the left side (body mass of penguin). We can then use the `eta_squared` function in the `effectsize` package to calculate the point estimate and confidence intervals.

```r
# Example:
# group: species
# outcome: body mass

# One-Way ANOVA
mdl1 <- aov(data = penguins,
            body_mass_g ~ species)

# calculate eta-squared
eta_squared(mdl1,
            partial = FALSE,
            alternative = "two.sided")
```

```
# Effect Size for ANOVA (Type I)

Parameter | Eta2 |       95% CI
------------------------------
species   | 0.67 | [0.62, 0.71]
```

The species of the penguin explains the majority of the variation in body mass showing an eta-squared value of η^2 = .67 [.62, .71]. Let us now do the same thing with a two-way ANOVA, using both `species` and `sex` as our categorical predictors.

```r
# Example:
# group: species and sex
# outcome: body mass

# Two-Way ANOVA
mdl2 <- aov(data = penguins,
            body_mass_g ~ species + sex)

# calculate eta-squared
eta_squared(mdl2,
            partial = FALSE,
            alternative = "two.sided")
```

```
# Effect Size for ANOVA (Type I)

Parameter | Eta2 |       95% CI
------------------------------
species   | 0.67 | [0.62, 0.72]
```

```
sex       | 0.17 | [0.10, 0.24]
```

Notice that the η^2 does not change for species since the sum of squares is divided by the total sum of squares rather than the residual sum of squares (see partial eta squared). The example shows an eta-squared value for species of η^2 = .67 95% CI [.62, .72] and for sex η^2 = .17 95% CI [.10, .24].

10.7.2 Partial Eta-Squared (η_p^2)

Partial eta-squared is the most commonly reported effect size measure for F-tests. It describes the proportion of variability associated with an effect when the variability associated with all other effects identified in the analysis has been removed from consideration (hence, it is "partial"). If you have access to an ANOVA table, the partial eta-squared for an effect is calculated as:

$$\eta_p^2 = \frac{SS_{\text{effect}}}{SS_{\text{effect}} + SS_{\text{error}}} \tag{10.5}$$

There are two things to take note of here:

1. In a one-way ANOVA (one categorical predictor), partial eta-squared and eta-squared are equivalent since $SS_{\text{total}} = SS_{\text{effect}} + SS_{\text{error}}$
2. If there are multiple predictors, the denominator will only include the sum of squares of the effect of interest (+ residuals) rather than the effect of all predictors (which is the case for the non-partial eta squared).

In R, let us compare the partial eta-squared values for a one-way ANOVA and a two-way ANOVA using the `eta_squared` function in the `effectsize` package.

```
# Example:
# group: species
# outcome: body mass


# one-way ANOVA
mdl1 <- aov(data = penguins,
            body_mass_g ~ species)

# calculate eta-square
eta_squared(mdl1,
             partial = TRUE,
```

```
              alternative = "two.sided")
```

```
For one-way between subjects designs, partial eta squared is equivalent
  to eta squared. Returning eta squared.
```

```
# Effect Size for ANOVA

Parameter | Eta2 |       95% CI
--------------------------------
species   | 0.67 | [0.62, 0.71]
```

The species of the penguin explains the majority of the variation in body mass showing a partial eta-squared value of η^2 = η_p^2 = .67 95% CI [.62, .71]. Let us now do the same thing with a two-way ANOVA, using both `species` and `sex` as our categorical predictors.

```
# Example:
# group: species and sex
# outcome: body mass

# two-way ANOVA
mdl2 <- aov(data = penguins,
            body_mass_g ~ species + sex)

# calculate eta-squared
eta_squared(mdl2,
             partial = TRUE,
             alternative = "two.sided")
```

```
# Effect Size for ANOVA (Type I)

Parameter | Eta2 (partial) |       95% CI
----------------------------------------
species   |           0.81 | [0.78, 0.84]
sex       |           0.53 | [0.46, 0.59]
```

Once we run a two-way ANOVA, the partial eta-squared value for species begins to differ. The example shows a partial eta-squared value for species of η_p^2 = .81 95% CI [.78, .84] and for sex η^2 = .53 95% CI [.46, .59].

10.7.3 Generalized Eta-Squared (η^2_G)

Generalized eta-squared was devised to allow effect size comparisons across studies with different designs, which eta-squared and partial eta-squared cannot help with (refer to for details). If you can (either you are confident that you calculated it right, or the statistical software that you use just happens to return this measure), report generalized eta-squared in addition to eta-squared or partial eta-squared. The biggest advantage of generalized eta-squared is that it facilitates meta-analysis, which is important for the accumulation of knowledge. To calculate generalized eta-squared, the denominator should be the sums of squares of all the non-manipulated variables (i.e., variance of purely individual differences in the outcome rather than individual differences in treatment effects). Note the formula will depend on the design of the study. In R, the `eta_squared` function in the `effectsize` package supports the calculation of generalized eta-squared by using the `generalized=TRUE` argument.

10.7.4 Omega squared corrections (ω^2, ω^2_p)

Similar to Hedges' correction for small sample bias in standardized mean differences, η^2 is also biased. We can apply a correction to η^2 and obtain a relatively unbiased estimate of the population proportion of variance explained by the predictor. To calculate ω, we need to calculate the within group mean squared errors:

$$MSE_{\text{within}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Where the predicted values of the outcome, $\hat{y}_i$, are the mean value for the individual's respective group. The formula for ω^2 is,

$$\omega^2 = \frac{SS_{\text{effect}} - (k-1) \times MSE_{\text{within}}}{SS_{\text{total}} + MSE_{\text{within}}} \tag{10.6}$$

Where k is the number of groups in the predictor (effect) variable. For partial omega-squared values, we need the mean squared error of the effect and the residuals which can easily be calculated from their sum of squares:

$$MSE_{\text{effect}} = \frac{SS_{\text{effect}}}{n} \tag{10.7}$$

$$MSE_{\text{error}} = \frac{SS_{\text{error}}}{n} \tag{10.8}$$

Then to calculate the partial omega squared we can use the following formula:

$$\omega_p^2 = \frac{(k-1)(MSE_{\text{effect}} - MSE_{\text{error}})}{(k-1) \times MSE_{\text{effect}} + (n-k-1) \times MSE_{\text{error}}} \tag{10.9}$$

In R, we can use the `omega_squared` function in the `effectsize` package to calculate both ω^2 and ω_p^2. For the first example we will use a one-way ANOVA.

```
# Example:
# group: species
# outcome: body mass

# One-Way ANOVA
mdl1 <- aov(data = penguins,
            body_mass_g ~ species)

# omega-squared
omega_squared(mdl1,
              partial = FALSE,
              alternative = "two.sided")
```

```
# Effect Size for ANOVA (Type I)

Parameter | Omega2 |       95% CI
---------------------------------
species   |   0.67 | [0.61, 0.71]
```

```
# partial omega-squared
omega_squared(mdl1,
              partial = TRUE,
              alternative = "two.sided")
```

```
For one-way between subjects designs, partial omega squared is
  equivalent to omega squared. Returning omega squared.

# Effect Size for ANOVA

Parameter | Omega2 |       95% CI
---------------------------------
species   |   0.67 | [0.61, 0.71]
```

The species of the penguin explains the majority of the variation in body mass, showing an omega-squared value of ω^2 = .67 95% CI [.61, .71]. Note that the partial and non-partial omega squared values do not show a difference as expected in a one-way ANOVA. Let us now do the same thing with a two-way ANOVA, using both `species` and `sex` as our categorical predictors.

```
# Example:
# group: species and sex
# outcome: body mass

# Two-Way ANOVA
mdl2 <- aov(data = penguins,
            body_mass_g ~ species + sex)

# omega-squared
omega_squared(mdl2,
              partial = FALSE,
              alternative = "two.sided")
```

```
# Effect Size for ANOVA (Type I)

Parameter | Omega2 |       95% CI
---------------------------------
species   |   0.67 | [0.62, 0.72]
sex       |   0.17 | [0.10, 0.24]
```

```
# partial omega-squared
omega_squared(mdl2,
              partial = TRUE,
              alternative = "two.sided")
```

```
# Effect Size for ANOVA (Type I)

Parameter | Omega2 (partial) |       95% CI
-----------------------------------------
species   |             0.81 | [0.78, 0.84]
sex       |             0.53 | [0.46, 0.58]
```

Once we run a two-way ANOVA, the partial eta-squared value for species diverge. The example shows a partial eta-squared value for species of ω_p^2 = .81 95% CI[.78, .84] and for sex ω^2 = .53 95% CI [.46, .58].

10.7.5 Cohen's f

Cohen's f is defined as the ratio of the between-group standard deviation to the within-group standard deviation (note that ANOVA assumes equal variances among groups). Cohen's f is the effect size measure asked for by G*Power for power analysis for F-tests. This can be calculated easily from the eta-squared value,

$$f = \sqrt{\frac{\eta^2}{1-\eta^2}} \tag{10.10}$$

or by the ω^2 value,

$$f = \sqrt{\frac{\omega^2}{1-\omega^2}} \tag{10.11}$$

Cohen's f can be interpreted as "the average Cohen's d (i.e., standardized mean difference) between groups". Note that there is no directionality to this effect size (f is always greater than zero), therefore two studies showing the same f with the same groups, can have very different patterns of group mean differences. Also note that Cohen's f is often reported as f^2. The confidence intervals for Cohen's f can be computed from the upper bounds and lower bounds of the confidence intervals from eta-square or omega-square using the formulas to calculate f (e.g., for the upper bound $f_{UB} = \sqrt{\frac{\eta^2_{UB}}{1-\eta^2_{UB}}}$).

In R, we can use the `cohens_f()` function in the `effectsize` package to calculate Cohen's f. We can also get f from eta^2 value using the `eta2_to_f()` function. We will again use example data from the `palmerpenguins` package.

```
# Example:
# group: species
# outcome: body mass

# fit ANOVA model
mdl <- aov(data = penguins,
           body_mass_g ~ species)

# calculate Cohen's f from anova table
cohens_f(mdl,alternative = "two.sided")
```

```
For one-way between subjects designs, partial eta squared is equivalent
  to eta squared. Returning eta squared.
```

```
# Effect Size for ANOVA

Parameter | Cohen's f |       95% CI
-------------------------------------
species   |      1.42 | [1.27, 1.57]
```

```
  # code for calculating f from eta2
  # get eta squared values
  eta_values <- eta_squared(mdl,alternative = "two.sided")
```

```
For one-way between subjects designs, partial eta squared is equivalent
  to eta squared. Returning eta squared.
```

```
  # calculate f from eta
  data.frame(f = eta2_to_f(eta_values$Eta2),
             ci.lb = eta2_to_f(eta_values$CI_low),
             ci.ub = eta2_to_f(eta_values$CI_high))
```

```
         f    ci.lb    ci.ub
1 1.423831 1.271129 1.573348
```

In the example above, the difference in body mass between the three penguin species was very large showing a Cohen's f of 1.42 95% CI [1.27, 1.57].

10.8 Reporting ANOVA results

For ANOVAs/F-tests, you will always need to report two kinds of effects: the omnibus effect of the factor(s) and the effect of planned contrasts or post hoc comparisons.

For instance, imagine that you are comparing three groups/conditions with a one-way ANOVA. The ANOVA will first return an F-statistic, the degrees of freedom, and the associated p-value. Here, you need to calculate the size of this omnibus factor effect in eta-squared, partial eta-squared, or generalized eta-squared. Suppose the omnibus effect is significant. You now know that there is at least one group that differs from the others. You want to know which group(s) differ from the others, and how much they differ. Therefore, you conduct post hoc comparisons on these groups. Because post hoc comparisons compare each group with the others in pairs, you will get a *t*-statistic and p-value for each comparison. For this, you need to calculate and report the standardized mean difference (d).

Imagine instead that you have two independent variables or factors, and you conduct a two-by-two factorial ANOVA. The first thing to do then is look at the interaction. If the interaction is significant, you again report the associated omnibus effect size measures, and proceed to analyze the simple effects. Depending on your research question, you compare the levels of one independent variable on each level of the other independent variable. You will report d for these simple effects. If the interaction is not significant, you look at the main effects and report the associated omnibus effect. You then proceed to analyze the main effect by comparing the levels of one independent variable while collapsing/aggregating the levels of the other independent variable. You will report d for these pairwise comparisons.

Note that lower-order effects are not directly interpretable if you have higher order effects. If you have an interaction in a two-way ANOVA, you cannot interpret the main effects directly. If you have a three-way interaction in a three-way ANOVA, you cannot interpret the main effects or the two-way interactions directly, regardless of whether they are significant or not.

In R, we can use the `summary` function to display the anova table. We can also append the table to include, for example, partial omega squared values and their respective confidence intervals.

```
# ANOVA mdl
mdl <- aov(data = penguins,
           body_mass_g ~ species + sex)

# calculate partial omega-squared values
omega_values <- omega_squared(mdl, alternative = "two.sided")

# create table of partial omega-squared values
omega_table <- data.frame(omega_sq = round(c(omega_values$Omega2_partial,NA),3),
                      ci.lb = round(c(omega_values$CI_low,NA),3),
                      ci.ub = round(c(omega_values$CI_high,NA),3))

# append omega values to summary of anova table
cbind(summary(mdl)[[1]],
      omega_table)
```

```
             Df    Sum Sq    Mean Sq  F value        Pr(>F) omega_sq ci.lb
species       2 145190219 72595109.6 724.2080 3.079053e-121    0.813 0.781
sex           1  37090262 37090261.8 370.0121  8.729411e-56    0.526 0.457
Residuals   329  32979185   100240.7       NA            NA       NA    NA
            ci.ub
species     0.838
sex         0.585
Residuals      NA
```

11 Differences in Variability

Occasionally researchers would like to compare the variations between two conditions or groups rather than the mean. Two commonly used effect sizes are the natural logarithm of the variability ratio (LVR) and the coefficient of variance ratio ($LCVR$). The latter of these can be useful when there may be a mean-variance relationship present (i.e., variances tend to increase with mean values). An LVR or $LCVR$ of zero therefore would indicate no difference in variation between the two groups, an LVR or $LCVR$ of >0 would indicate larger variance in group A, and an LVR or $LCVR$ of <0 would indicate larger variance in group B (reference group). There are both independent and dependent versions of these effect sizes (see Senior, Viechtbauer, and Nakagawa 2020). To obtain confidence intervals of the LVR or $LCVR$, we multiply the standard error of $LVR/LCVR$ by 1.96 similarly to other effect size statistics,

$$CI_{LVR} = LVR \pm 1.96 \cdot SE_{LVR} \tag{11.1}$$

$$CI_{LCVR} = LCVR \pm 1.96 \cdot SE_{LCVR} \tag{11.2}$$

Here is a table for every effect size discussed in this chapter:

Type	Description	Section
Variability Ratios (VR)		Section 11.1
LVR_{ind} - Natural Logarithm of variability ratio for independent groups	Used to compare the standard deviations (i.e., the variability) between two groups.	Section 11.1.1
LVR_{dep} - Natural Logarithm of variability ratio for dependent groups	Used to compare the standard deviations (i.e., the variability) between paired groups (i.e., repeated measures designs).	Section 11.1.2
Coefficient of Variation Ratios (CVR)		Section 11.2

Type	Description	Section
$LCVR_{\text{ind}}$ - Natural Logarithm of coefficient variation ratio for independent groups	Used to compare the variation between two groups. More useful than a variability ratio (LVR_{ind}) when there is a relationship between the mean and variance.	Section 11.2.1
$LCVR_{\text{dep}}$ - Natural Logarithm of coefficient variation ratio for dependent groups	Used to compare the variation between paired groups (i.e., repeated measures). More useful than a variability ratio (LVR_{dep}) when there is a relationship between the mean and variance.	Section 11.2.1

11.1 Variability Ratios

11.1.1 Natural Logarithm of Variability Ratio for Independent Groups (LVR_{ind})

The variability ratio for independent groups (denoted as group A and group B) can be calculated by taking the natural logarithm of the *standard deviation* within one group divided by the standard deviation in another group,

$$LVR_{\text{ind}} = \ln\left(\frac{S_A}{S_B}\right) + CF \tag{11.3}$$

Where CF is a small sample correction factor calculated as,

$$CF = \frac{1}{2(n_A - 1)} - \frac{1}{2(n_B - 1)} \tag{11.4}$$

A LVR of zero therefore would indicate no difference in variation between the two groups, a LVR of >0 would indicate larger variance in group A, and LVR of <0 would indicate larger variance in group B. The standard error of the LVR can be calculated as,

$$SE_{LVR_{\text{ind}}} = \sqrt{\frac{n_A}{2(n_A - 1)^2} + \frac{n_B}{2(n_B - 1)^2}} \tag{11.5}$$

In R, we can use the `escalc()` function from the `metafor` package (Viechtbauer 2010) as follows:

```r
library(metafor)

# Example:
# Group A: standard deviation = 4.5, sample size = 50
# Group B: standard deviation = 3.5, sample size = 50

# calculate the variability ratio
LVRind <- escalc(
    measure = "VR",
    sd1i = 4.5,
    sd2i = 3.5,
    n1i = 50,
    n2i = 50,
    var.names = c("LVRind","variance")
  )

# display results
summary(LVRind)
```

```
  LVRind variance    sei     zi   pval   ci.lb  ci.ub
1 0.2513   0.0204 0.1429 1.7592 0.0785 -0.0287 0.5313
```

From the example, we obtain a natural log variability ratio of LVR_{ind} = 0.25 95% CI [-0.03, 0.53].

11.1.2 Natural Logarithm of Variability Ratio for Dependent Groups (LVR_{dep})

The variability ratio for dependent groups (denoted as groups 1 and 2; e.g., pre-post comparisons) can similarly be calculated by taking the natural logarithm of the *standard deviation* within one group divided by the *standard deviation* in another group,

$$LVR_{\text{dep}} = \ln\left(\frac{S_2}{S_1}\right) \tag{11.6}$$

Note, the correction factor is irrelevant due to the fact that the conditions will have the same sample size ($n = n_1 = n_2$). The standard error for which can be calculated as,

$$SE_{LVR_{\text{dep}}} = \sqrt{\frac{n}{n-1} - \frac{r^2}{n-1} + \frac{r^4\left(S_A^8 + S_B^8\right)}{2(n-1)^2 S_A^4 + S_B^4}}. \tag{11.7}$$

Where r is the correlation between the two conditions. In R, we can use the `escalc()` function from the `metafor` package as follows:

```
# Example:
# Condition 1: standard deviation = 4.5
# Condition 2: standard deviation = 3.5
# Sample size = 50
# Correlation = 0.4

# calculate variability ratio
LVRdep <- escalc(
  measure = "VRC",
  sd1i = 4.5,
  sd2i = 3.5,
  ni = 50,
  ri = .40,
  var.names = c("LVRdep","variance")
)

summary(LVRdep)
```

```
  LVRdep variance    sei     zi   pval   ci.lb  ci.ub
1 0.2513   0.0171 0.1309 1.9194 0.0549 -0.0053 0.5079
```

The output shows a LVR_{dep} of 0.25 95% CI [-0.01, 0.51].

11.2 Coefficient of Variation Ratios

11.2.1 Natural Logarithm of Coefficient of Variation Ratio for independent groups (LCVR_ind)

The coefficient of variation ratio for independent groups can be calculated by taking the natural logarithm of the coefficient of variation within one group divided by the coefficient of variation in another group,

$$LCVR_{\text{ind}} = \ln\left(\frac{CV_A}{CV_B}\right) + CF \tag{11.8}$$

Where $CV_A = S_A/M_A$, $CV_B = S_B/M_B$, and M indicates the mean of the respective group. The correction factor, CF, is a small sample size bias correction factor that combines that from the LRR (presented earlier) and the LVR calculated as,

$$CF = \frac{1}{2(n_A - 1)} - \frac{1}{2(n_B - 1)} + \frac{S_A^2}{2(n_A M_A^2)} + \frac{S_B^2}{2(n_B M_B^2)} \tag{11.9}$$

In R, we can use the `escalc()` function from the `metafor` package as follows:

```
# Example:
# Group A: mean = 22.4, standard deviation = 4.5, sample size = 50
# Group B: mean = 20.1, standard deviation = 3.5, sample size = 50

# calculate variability ratio
LCVRind <- escalc(
  measure = "CVR",
  m1i = 22.4,
  m2i = 20.1,
  sd1i = 4.5,
  sd2i = 3.5,
  n1i = 50,
  n2i = 50,
  var.names= c("LCVRind", "variance")
)


summary(LCVRind)
```

```
  LCVRind variance    sei     zi   pval   ci.lb  ci.ub
1  0.1430   0.0218 0.1477 0.9679 0.3331 -0.1466 0.4325
```

The output shows a $LCVR_{\text{ind}}$ of 0.14 95% CI [-0.15, 0.43].

11.2.2 Natural Logarithm of Coefficient of Variation Ratio for independent groups ($LCVR_{\text{dep}}$)

The coefficient of variation ratio for dependent groups can be similarly calculated by taking the natural logarithm of the coefficient of variation within one group divided by the coefficient of variation in another group,

$$LCVR_{\text{dep}} = \ln\left(\frac{CV_2}{CV_1}\right) + CF \tag{11.10}$$

Where $CV_1 = S_1/M_1, CV_2 = S_2/M_2$ and the correction factor is calculated as,

$$CF = \frac{S_2^2}{2nM_2^2} - \frac{S_1^2}{2nM_1^2} \tag{11.11}$$

The standard error of the $LCVR_{\text{dep}}$ can be calculated as,

$$SE_{LCVR_{\text{dep}}} = \sqrt{\frac{S_1^2}{nM_1^2} + \frac{S_2^2}{nM_2^2} + \frac{S_1^4}{2n^2M_1^4} + \frac{S_2^4}{2n^2M_2^4} + \frac{2rS_1S_2}{nM_1M_2} + \frac{r^2S_1^2S_2^2(M_1^4 + M_2^4)}{2n^2M_1^4M_2^4}} \tag{11.12}$$

In R, we can simply use the `metafor` packages `escalc()` function as follows:

```
# Example:
# Group 1: standard deviation = 4.5
# Group 2: standard deviation = 3.5
# Sample size = 50
# Correlation = 0.4

# calculate coefficient of variability ratio
LCVRdep <- escalc(
  measure = "CVRC",
  m1i = 22.4,
  m2i = 20.1,
  sd1i = 4.5,
  sd2i = 3.5,
  ni = 50,
  ri = .40,
  var.names = c("LCVRdep", "variance")
)
```

```r
# display results
summary(LCVRdep)
```

```
  LCVRdep variance    sei     zi   pval   ci.lb  ci.ub
1  0.1430   0.0180 0.1342 1.0658 0.2865 -0.1200 0.4059
```

The output shows a $LCVR_{\text{dep}}$ of 0.14 95% CI [-0.12, 0.41].

12 Non-Parametric Tests

Sometimes the assumptions of parametric models (e.g., normality of model residuals) are suspect. This is often the case in psychology when using ordinal scales. In these cases a "non-parametric" approach may be helpful. A statistical test being non-parametric means that the parameters (i.e., mean and variance for "normal" Gaussian model) are not estimated; despite popular belief the data themselves are *never* non-parametric. Additionally, these tests are *not* tests of the median (Divine et al. 2018). Rather one can consider them as rank based or proportional odds tests. If the scores you are analyzing are not metric (i.e., ordinal) due to the use of a Likert-Scale, and you still use parametric tests such as t-tests, you run the risk of a high false-positive probability (e.g., Liddell and Kruschke 2018). Note that in German, scale anchors have been developed that are very similar to Likert scale, but can be interpreted as metric (e.g., Rohrmann 2007).

We will briefly discuss here two groups of tests that can be applied to independent and paired samples. Then we present three effect sizes that can accompany these tests as well as their calculations and examples in R.

Here is a table for every effect size discussed in this chapter:

Type	Description	Section
Rank-Biserial Correlation		Section 12.3.1
r_{rb} (dependent groups) - Rank-biserial correlation on dependent groups	A measure of dominance between dependent groups (i.e., repeated measure designs).	Section 12.3.1.1
r_{rb} (independent groups) - Rank-biserial Correlation on independent groups	A measure of dominance between two independent groups.	Section 12.3.1.2
Concordance Probability		Section 12.3.2

Type	Description	Section
p_c - Concordance probability	A simple transformation of the rank-biserial correlation and it represents the probability of superiority in one group relative to the other group. This section shows R code for both independent and dependent samples.	Section 12.3.2
Wilcoxon-Mann-Whitney Odds		Section 12.3.3
O_{WMW} - Wilcoxon-Mann-Whitney Odds	Also known as the Generalized Odds Ratio, it transforms the concordance probability to an Odds Ratio. This section shows R code for both independent and dependent samples.	Section 12.3.3

12.1 Wilcoxon-Mann-Whitney tests

A non-parametric alternative to the t-test is the Wilcoxon-Mann-Whitney (WMW) group of tests. When comparing two independent samples this is called a Wilcoxon rank-sum test, but sometimes referred to as a Mann-Whitney U Test. When using it on paired samples, or one sample, it is a signed rank test. These are generally referred to as tests of "symmetry" (Divine et al. 2018), which can be performed in base R.

```
library(tidyverse)

# load data
data(sleep)

# structure data as two data columns
dat <- sleep %>%
  pivot_wider(names_from = group,
              values_from = extra,
              id_cols = ID) %>%
  rename(g1 = `1`, g2 = `2`) %>%
  as.data.frame()

# view data
head(dat)
```

```
  ID   g1   g2
1  1  0.7  1.9
2  2 -1.6  0.8
3  3 -0.2  1.1
4  4 -1.2  0.1
5  5 -0.1 -0.1
6  6  3.4  4.4
```

```r
# wilcoxon signed-rank test
wilcox.test(x = dat$g1,
            y = dat$g2,
            paired = TRUE)
```

```
    Wilcoxon signed rank test with continuity correction

data:  dat$g1 and dat$g2
V = 0, p-value = 0.009091
alternative hypothesis: true location shift is not equal to 0
```

```r
# wilcoxon rank-sum test
wilcox.test(x = dat$g1,
            y = dat$g2,
            paired = FALSE)
```

```
    Wilcoxon rank sum test with continuity correction

data:  dat$g1 and dat$g2
W = 25.5, p-value = 0.06933
alternative hypothesis: true location shift is not equal to 0
```

The p-value is lower in the signed-rank test because the correlation between conditions is accounted for whereas the rank-sum test assumes the two groups are independent samples.

12.2 Brunner-Munzel Tests

Brunner-Munzel's tests can be used instead of the WMW tests. The primary reason is the interpretation of the test (Munzel and Brunner 2002; Brunner and Munzel 2000; Neubert and

Brunner 2007). Recently, Karch (2021) argued that the Mann-Whitney test is not a decent test of equality of medians, distributions or stochastic equality. The Brunner-Munzel test, on the other hand, provides a sensible approach to test for stochastic equality.

The Brunner-Munzel tests measure a rank based "relative effect" or "stochastic superiority probability". The test statistic ($\hat{p}$) is essentially the probability of a value in one condition being greater than other while splitting the ties[1]. However, Brunner-Munzel tests can not be applied to the single group or one-sample designs.

$$\hat{p} = P(X < Y) + \frac{1}{2} \cdot P(X = Y) \tag{12.1}$$

These tests are relatively new so there are very few packages that offer Brunner-Munzel tests. Moreover, Karch (2021) argues that the stochastic superiority effect size ($\hat{p}$) offers a nuanced way to interpret group differences by visualizing observations as competitors in a contest. Propounded by scholars like Cliff (1993) and Divine et al. (2018), it views each observation from one group in a duel with every observation from another. If an observation from the first group surpasses its counterpart, it "wins," and the group garners a point; tied observations yield half a point to each group. Other interpretations, like transforming p to the Wilcoxon-Mann-Whitney (WMW) odds or Cliff's δ offer deeper insights. There are implementations of the Brunner-Munzel test in a few packages in R (i.e. `lawstat`, `rankFD`, and `brunnermunzel`). Karch (2021) recommends the `brunnermunzel.permutation.test()` function from the `brunnermunzel` package (Ara 2022). The `TOSTER` R package can also provide coverage (Läkens 2017; Caldwell 2022).

```
library(TOSTER)

# Paired samples
data(sleep)

# When sample sizes are small
# a permutation version should be used.
# When this is done a seed should be set.
set.seed(1)

# calculate brunner-munzel test
brunner_munzel(extra ~ group,
               data = sleep,
```

[1] Note, for paired samples, this does not refer to the probability of an increase/decrease in paired sample but rather the probability that a randomly sampled value of X will be greater/less than Y. This is also referred to as the "relative" effect in the literature. Therefore, the results will differ from the concordance probability. The test-statistic can be computed as follows for conditions X and Y,

```
                paired = TRUE,
                perm = TRUE)
```

```
    Paired Brunner-Munzel permutation test

data:  extra by group
t-observed = -3.7266, df = 9, p-value = 0.003906
alternative hypothesis: true relative effect is not equal to 0.5
95 percent confidence interval:
 0.1233862 0.3866138
sample estimates:
p(X>Y) + .5*P(X=Y)
             0.255
```

```
  # two sample test
  brunner_munzel(extra ~ group,
                data = sleep,
                paired = FALSE,
                perm = TRUE)
```

```
    two-sample Brunner-Munzel permutation test

data:  extra by group
t-observed = -2.1447, df = 16.898, p-value = 0.0506
alternative hypothesis: true relative effect is not equal to 0.5
95 percent confidence interval:
 0.00412703 0.50031995
sample estimates:
p(X>Y) + .5*P(X=Y)
             0.255
```

12.3 Rank-Based Effect Sizes

Since the mean and standard deviation are not estimated for a WMW or Brunner-Munzel test, it would be inappropriate to present a standardized mean difference (e.g., Cohen's d) to accompany these tests. Instead, a rank based effect size (i.e., based on the ranks of the observed values) can be reported to accompany the non-parametric statistical tests.

12.3.1 Rank-Biserial Correlation

The rank-biserial correlation (r_{rb}) is considered a measure of dominance. The correlation represents the difference between the proportion of favorable and unfavorable pairs or signed ranks. Larger values indicate that more of X is larger than more of Y, with a value of (−1) indicates that all observations in the second, Y, group are larger than the first, X, group, and a value of (+1) indicates that all observations in the first group are larger than the second.

12.3.1.1 Dependent Groups

To get the rank-biserial correlation between dependent groups/conditions we can use the following procedure:

1. Calculate difference scores between conditions for each person i:

$$D_i = X_i - Y_i$$

2. Calculate the positive and negative rank sums:

$$\text{When } D_i > 0, \quad R_{\oplus} = \sum_{i=1} -1 \cdot \text{sign}(D_i) \cdot \text{rank}(|D_i|)$$

$$\text{When } D_i < 0, \quad R_{\ominus} = \sum_{i=1} -1 \cdot \text{sign}(D_i) \cdot \text{rank}(|D_i|)$$

3. We can set a constant, H, to be -1 when the positive rank sum is greater than or equal to the negative rank sum ($R_{\oplus} \geq R_{\ominus}$).

$$H = \begin{cases} -1 & R_{\oplus} \geq R_{\ominus} \\ 1 & R_{\oplus} < R_{\ominus} \end{cases}$$

4. Calculate rank-biserial correlation:

$$r_{rb} = 4H \times \left| \frac{\text{min}(R_{\oplus}, R_{\ominus}) - .5 \times (R_{\oplus} + R_{\ominus})}{n(n+1)} \right| \tag{12.2}$$

5. The confidence intervals can then be calculated by Z-transforming the correlation,

$$Z_{rb} = \text{arctanh}(r_{rb}).$$

6. Calculate the standard error of the Z-transformed correlation,

$$SE_{Z_{rb}} = \frac{SE_{r_{rb}}}{1 - r_{rb}^2}. \tag{12.3}$$

7. Then the confidence interval can be calculated and then back-transformed,

$$CI_{r_{rb}} = \tanh(Z_{rb} \pm 1.96 \cdot SE_{Z_{rb}}). \tag{12.4}$$

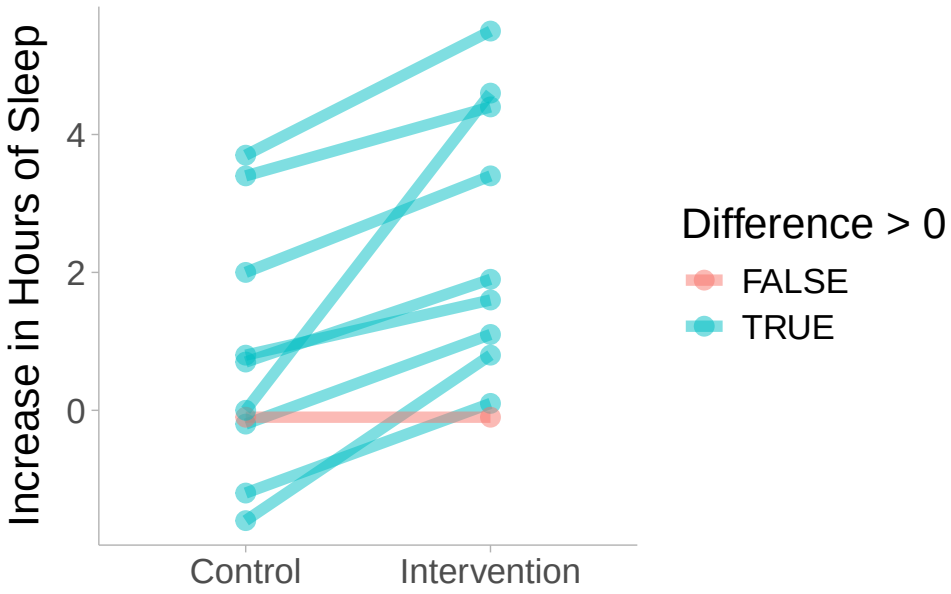
In R, we can use the `ses_calc()` function in `TOSTER` package (Läkens 2017). For the following example, we will calculate the rank-biserial correlation in the sleep dataset:

```
# Dependent groups
# When this is done a seed should be set.
set.seed(1)

# calculate rank-biserial correlation
ses_calc(extra ~ group,
         data = sleep,
         paired = TRUE)
```

```
                         estimate lower.ci  upper.ci conf.level
Rank-Biserial Correlation 0.9818182 0.928369 0.9954785       0.95
```

The example shows a rank-biserial correlation is r_{rb} = .982 [.938, .995]. This suggests that nearly every individual in the sample showed an increase in condition 2 relative to condition 1. As you can see from the figure below, only one individual showed a decline (individual shown in red).



12.3.1.2 Dependent Groups

To get the rank-biserial correlation between independent groups we can use the following procedure:

1. Calculate the ranks for each observation across all observations in groups 1 and 2

$$R = \text{rank}(X_1, X_2)$$

2. Rank ordered observations R are split into R_1 and R_2 which are the ranked observations associated with groups 1 and 2, respectively. Then calculate the rank sums from each group

$$U_1 = \left(\sum_{i=1}^{n_1} R_{1i}\right) - n_1 \cdot \frac{n_1 + 1}{2}$$

$$U_2 = \left(\sum_{i=1}^{n_2} R_{2i}\right) - n_2 \cdot \frac{n_2 + 1}{2}$$

3. Calculate the rank-biserial correlation

$$r_{rb} = \frac{U_1}{n_1 n_2} - \frac{U_2}{n_1 n_2} \tag{12.5}$$

4. The confidence intervals can then be calculated by transforming the estimate.

$$Z_{rb} = \text{arctanh}(r_{rb})$$

5. Calculate the standard error of the Z-transformed correlation

$$SE_{Z_{rb}} = \frac{SE_{r_{rb}}}{1 - r_{rb}^2} \tag{12.6}$$

6. Then the confidence interval can be calculated and then back-transformed.

$$CI_{r_{rb}} = \tanh(Z_{rb} \pm 1.96 \cdot SE_{Z_{rb}}) \tag{12.7}$$

In R, we can use `ses_calc` in the `TOSTER` package can be utilized to calculate r_{rb}.

```
library(likert)

# Two Sample
# install the janitor package for data cleaning
# clean and import data from likert
data(mass, package = "likert")
df_mass = mass %>%
  as.data.frame() %>%
  janitor::clean_names()

# function needs input as a numeric
# ordered factors can be converted to ranks
# Again, the warning can be ignored
set.seed(24111)
ses_calc(
  rank(math_relates_to_my_life) ~ gender,
  data = df_mass,
  paired = FALSE
)
```

```
                          estimate   lower.ci   upper.ci conf.level
Rank-Biserial Correlation -0.452381 -0.7831567 0.07794462       0.95
```

The example shows a rank-biserial correlation is r_{rb} = -.45 95% CI [-.78, .08].

12.3.2 Concordance Probability

In the two sample case, concordance probability is the probability that a randomly chosen subject from one group has a response that is larger than that of a randomly chosen subject from the other group. In the two sample case, this is roughly equivalent to the statistic of the Brunner-Munzel test. In the paired sample case, it is the probability that a randomly chosen difference score (D) will have a positive (+) sign plus 0.5 times the probability of a tie (no/zero difference). The concordance probability can go by many names such as the c-index, the non-parametric probability of superiority, or the non-parametric common language effect size (CLES).

The calculation of concordance can be derived from the rank-biserial correlation,

$$p_c = \frac{r_{rb} + 1}{2}. \tag{12.8}$$

In R, we can use the `ses_calc()` function again along with the `sleep` data set. For repeated measures experiments, the concordance probability in dependent groups can be calculated utilizing the `paired=TRUE` argument in the `ses_calc()` function:

```
# calculate concordance probability
ses_calc(extra ~ group,
         data = sleep,
         paired = TRUE,
         ses = "c")
```

```
            estimate  lower.ci  upper.ci conf.level
Concordance 0.9909091 0.9641845 0.9977392       0.95
```

The results show a very high concordance probability of .991 95% CI [.996, .998]. For two independent groups, the concordance probability can be calculated similarly without specifying the `paired` argument:

```
# calculate concordance probability
ses_calc(rank(math_relates_to_my_life) ~ gender,
         data = df_mass,
         ses = "c")
```

```
             estimate  lower.ci  upper.ci conf.level
Concordance 0.2738095 0.1084217 0.5389723       0.95
```

The results show a concordance probability of .27 95% CI [.11, .54].

12.3.3 Wilcoxon-Mann-Whitney Odds

The Wilcoxon-Mann-Whitney odds (O'Brien and Castelloe 2006), also known as the "Generalized Odds Ratio" (Agresti 1980), essentially transforms the concordance probability into an odds ratio. The log odds can be converted from the concordance by taking the logit of the concordance,

$$\log(O_{WMW}) = \text{logit}(p_c) \tag{12.9}$$

The exponential value of the log-odds will provide the odds on a more interpretable scale,

$$O_{WMW} = \exp\left[\text{logit}(p_c)\right] \tag{12.10}$$

In R, we can calculate O_{WMW} by using the `ses_calc()` function from the `TOSTER` package:

```
# calculate WMW odds
ses_calc(extra ~ group,
         data = sleep,
         paired = TRUE,
         ses = "odds")
```

```
         estimate lower.ci upper.ci conf.level
WMW Odds      109 26.92087 441.3305       0.95
```

The results shows an extremely large O_{WMW} of 109 95% CI [26.9, 441.3]. We can also calculate O_{WMW} in independent groups using the same function:

```
# calculate WMW odds
ses_calc(rank(math_relates_to_my_life) ~ gender,
         data = df_mass,
         ses = "odds")
```

```
          estimate  lower.ci upper.ci conf.level
WMW Odds 0.3770492 0.1216064 1.169067       0.95
```

The results shows an extremely large O_{WMW} of .377 95% CI [0.12, 1.17].

13 Regression

Regression is a method of predicting an outcome variable from one or more predictor variables.

13.1 Regression Overview

In a simple linear regression there is only one predictor (x) and one outcome (y) in the regression model,

$$y = b_0 + b_1 x + e \tag{13.1}$$

where b_0 is the intercept coefficient, b_1 is the slope coefficient, and e is the residual error term that has a variance of Var$(e) = \sigma^2$. We can fit this model on the `palmerpenguins` data set (Horst, Hill, and Gorman 2020) using flipper length as our predictor variable and bill length as our outcome variable (see Figure 13.1).

For a simple linear regression we can obtain an unstandardized regression coefficient by finding the optimal value of b_0 and b_1 that minimizes the variance in e, namely, σ^2 (i.e., this finds the best fit line that maximally reduces error). In a multiple regression we can model y as a function of multiple predictor variables such that,

$$y = b_0 + b_1 x_1 + b_2 x_2 + ... + e \tag{13.2}$$

Where the coefficients are all estimated jointly to minimize the error variance. The line produced by the regression equation is our predicted values of y, however it can also be interpreted as the mean of y given some value of x (i.e., the conditional mean). In a regression equation we can construct more complex models that include non-linear terms such as interactions or polynomials (or any sort of function of x). For example, we can create a model where we include a main effect, x_1 and a quadratic polynomial term x_1^2,

$$y = b_0 + b_1 x + b_2 x^2 + e \tag{13.3}$$

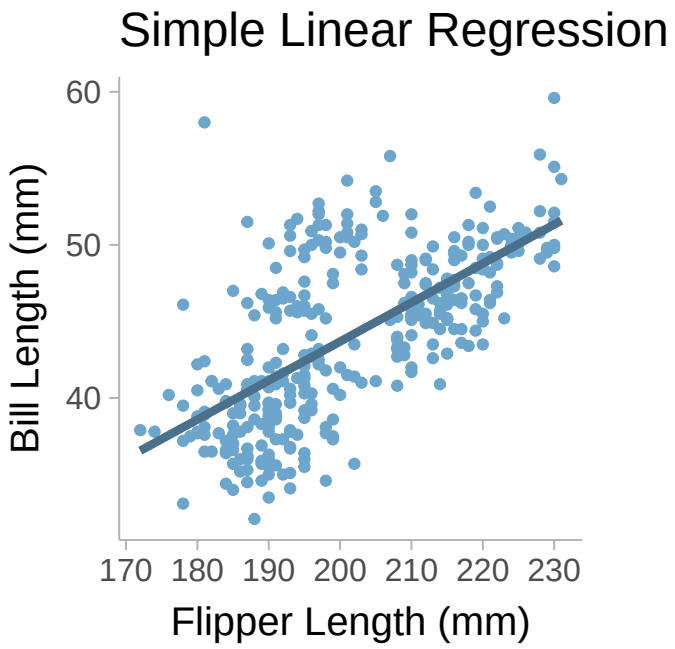


Figure 13.1: Linear regression for the palmerpenguins data set. Plot shows a positive association beteen flipper length and bill length.

When we fit this new model to the `palmerpenguins` data set and we can see that the model adds slight curvature to the regression line which is what we would expect second degree (quadratic) polynomial (see Figure 13.2).

13.2 Effect Sizes for a Linear Regression

If we want to calculate the variance explained in the outcome by all the predictor variables, we can compute an R^2 value. The R^2 value can be interpreted one of two ways:

1. The variance in y explained by the predictor variables.
2. The square of the correlation between predicted y values and observed (actual) y values (the square root of R^2 will give us the correlation between predicted and observed y values).

We can construct a linear regression model quite easily in base R using the `lm()` function. We will continue to use the `palmerpenguins` data set for the example. In this example we will predict bill length with two predictor variables: flipper length and bill depth.

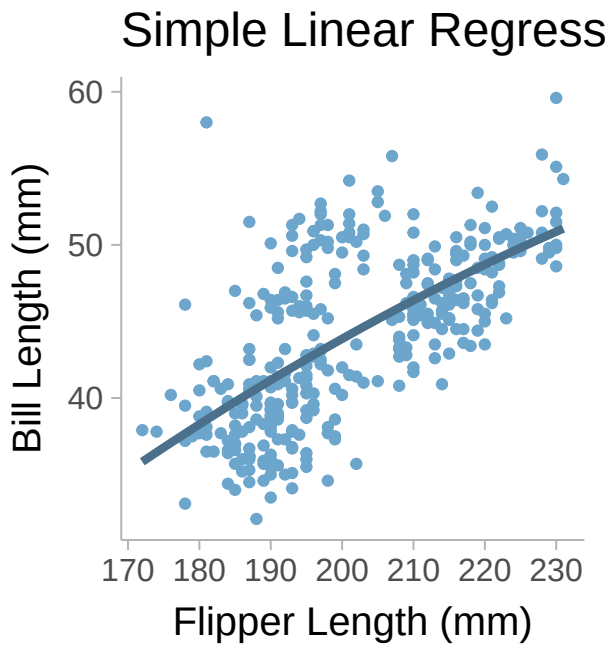


Figure 13.2: Polynomial (quadratic) regression for the palmerpenguins data set.

```
library(palmerpenguins)

# construct multiple regression model
mdl <- lm(bill_length_mm ~ flipper_length_mm + bill_depth_mm,
          data = penguins)

# output regression summary
summary(mdl)
```

```
Call:
lm(formula = bill_length_mm ~ flipper_length_mm + bill_depth_mm,
    data = penguins)

Residuals:
     Min       1Q   Median       3Q      Max
-10.8831  -2.7734  -0.3268   2.3128  19.7630

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)       -28.14701    5.51435  -5.104 5.54e-07 ***
flipper_length_mm   0.30569    0.01902  16.073  < 2e-16 ***
bill_depth_mm       0.62103    0.13543   4.586 6.38e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.009 on 339 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.4638,    Adjusted R-squared:  0.4607
F-statistic: 146.6 on 2 and 339 DF,  p-value: < 2.2e-16
```

We will notice that the linear regression summary returns two R^2 values. The first one is the traditional R^2 and the other is the adjusted R^2. The adjusted R^2_{adj} applies a correction factor since R^2 is biased when there are multiple predictor variables and/or a small sample size. If we want to know the contribution for each term in the regression model, we can also use semi-partial sr^2 values that are similar to partial eta-squared in the ANOVA section of this book. In R, we can calculate sr^2 with the `r2_semipartial()` function in the `effectsize` package (Ben-Shachar, Lüdecke, and Makowski 2020):

```
library(effectsize)

# get semipartial sr^2 values for each predictor
r2_semipartial(mdl,alternative = "two.sided")
```

```
Term              |  sr2 |       95% CI
----------------------------------------
flipper_length_mm | 0.41 | [0.33, 0.49]
bill_depth_mm     | 0.03 | [0.01, 0.06]
```

A standardized effect size for each term could also be calculated from standardizing the regression coefficients. Standardized regression coefficients are calculated by re-scaling the predictor and outcome variables to be z-scores (i.e., setting the mean and variance to be zero and one, respectively).

```
# construct multiple regression model with scaled variables
stand_mdl <- lm(scale(bill_length_mm) ~ scale(flipper_length_mm) + scale(bill_depth_mm),
                data = penguins)

# display summary of regression model
summary(stand_mdl)
```

```
Call:
lm(formula = scale(bill_length_mm) ~ scale(flipper_length_mm) +
    scale(bill_depth_mm), data = penguins)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9934 -0.5080 -0.0599  0.4236  3.6199

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              -2.328e-15  3.971e-02   0.000        1
scale(flipper_length_mm)  7.873e-01  4.899e-02  16.073  < 2e-16 ***
scale(bill_depth_mm)      2.246e-01  4.899e-02   4.586 6.38e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7344 on 339 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.4638,    Adjusted R-squared:  0.4607
F-statistic: 146.6 on 2 and 339 DF,  p-value: < 2.2e-16
```

Alternatively, we can use the `standardise` function in the `effectsize` package:

```
# standardize coefficients
standardise(mdl)
```

```
Call:
lm(formula = bill_length_mm ~ flipper_length_mm + bill_depth_mm,
    data = data_std)

Coefficients:
      (Intercept)  flipper_length_mm      bill_depth_mm
        4.335e-16          7.873e-01          2.246e-01
```

13.3 Pearson correlation vs regression coefficients in simple linear regressions

A slope coefficient in a simple linear regression model can be defined as the covariance between predictor x and outcome y divided by the variance in x,

$$b_1 = \frac{\text{Cov}(x, y)}{S_x^2}$$

Where S_x is the standard deviation of x (the square of the standard deviation is the variance). A Pearson correlation is defined as,

$$r = \frac{\text{Cov}(x, y)}{S_x S_y}$$

We can see that these formulas are quite similar, in fact we can express r as a function of b_1 such that,

$$r = b_1 \frac{S_x}{S_y} \tag{13.4}$$

Which means that if $S_x = S_y$ then $r = b_1$. Furthermore, if the regression coefficient is standardized this would make the outcome and predictor variable to both have a variance of 1, thus making $S_x = S_y = 1$. Therefore a standardized regression coefficient in a simple linear regression with one predictor is equal to a Pearson correlation.

13.4 Multi-Level Regression models

We can allow the regression coefficients such as the intercept and slope to vary with respect to some grouping variable. For example, let's say we think that the intercept will vary between the different species of penguins when we look at the relationship between bill length and flipper length. Using the `lme4` package (Bates et al. 2015), we can construct a model that allows the intercept coefficient to vary between species.

```
library(palmerpenguins)
library(lme4)

# construct multi-level regression model
```

```r
# the 1 in (1 | species) denotes the intercept
ml_mdl <- lmer(bill_length_mm ~ flipper_length_mm + (1 | species),
               data = penguins)

# display summary of model
summary(ml_mdl)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: bill_length_mm ~ flipper_length_mm + (1 | species)
   Data: penguins

REML criterion at convergence: 1640.6

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.5568 -0.6666  0.0109  0.7020  4.7678

Random effects:
 Groups   Name        Variance Std.Dev.
 species  (Intercept) 20.06    4.479
 Residual              6.74    2.596
Number of obs: 342, groups:  species, 3

Fixed effects:
                  Estimate Std. Error t value
(Intercept)        1.81165    4.97514   0.364
flipper_length_mm  0.21507    0.02113  10.177

Correlation of Fixed Effects:
            (Intr)
flppr_lngt_ -0.854
```

Note in the table that we have random effects and fixed effects. The random effects shows the grouping (categorical) variable that the parameter is allowed to vary on and then it shows the parameter that is varying, which in our case is the intercept coefficient. It also includes the variance of the intercept, which is the extent to which the intercept varies between species. For the fixed effect terms, we see the intercept displayed as well as the slope, this shows the **mean** of the intercept across species and, since the slope is equal across species, the slope is just a single (fixed) value. Notice that in Figure 13.3, the slopes are fixed and equal between each species and only the intercepts (i.e., the vertical height of each line) differs.

Extending the same model, we can also allow the slopes to vary between species.

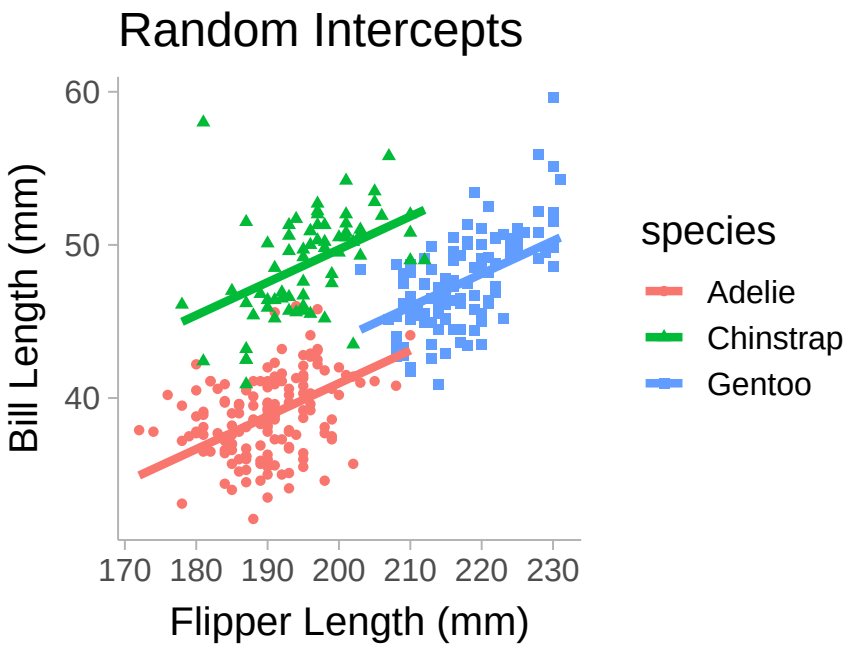


Figure 13.3: Multi-level regression model allowing each species to have their own intercept.

```
library(palmerpenguins)
library(lme4)

# construct multi-level model with random slope and intercept
ml_mdl <- lmer(bill_length_mm ~ flipper_length_mm + (1 + flipper_length_mm | species),
               data = penguins)

# display regression results
summary(ml_mdl)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: bill_length_mm ~ flipper_length_mm + (1 + flipper_length_mm |
    species)
   Data: penguins

REML criterion at convergence: 1638.2

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.6326 -0.6657  0.0083  0.6843  4.9531
```

```
Random effects:
 Groups   Name              Variance  Std.Dev. Corr
 species  (Intercept)       3.0062118 1.73384
          flipper_length_mm 0.0007402 0.02721  -0.61
 Residual                   6.6886861 2.58625
Number of obs: 342, groups:  species, 3

Fixed effects:
                  Estimate Std. Error t value
(Intercept)        1.56035    4.32870   0.360
flipper_length_mm  0.21609    0.02623   8.237

Correlation of Fixed Effects:
            (Intr)
flppr_lngt_ -0.863
optimizer (nloptwrap) convergence code: 0 (OK)
unable to evaluate scaled gradient
Model failed to converge: degenerate  Hessian with 1 negative eigenvalues
```

Varying the slope will include `flipper_length_mm` in the random effects terms. Also note that the summary returns the correlation between random effect terms, which may be useful to know if there is a strong relationship between the intercept and slope across species. Figure 13.4 shows the model allowing slope and intercept to vary between species.

13.4.1 Marginal and Conditional R^2

For multi-level models, we can compute a conditional R^2 and a marginal R^2. These can be interpreted as,

- **Marginal** R^2: the variance explained solely by the fixed effects,
- **Conditional** R^2: the variance explained in the whole model, including both the fixed effects and random effects terms.

In R, we can use the `MuMIn` package (Bartoń 2023) to compute both the marginal and conditional R^2:

```r
library(MuMIn)

# calculate conditional and marginal R2
r.squaredGLMM(ml_mdl)
```

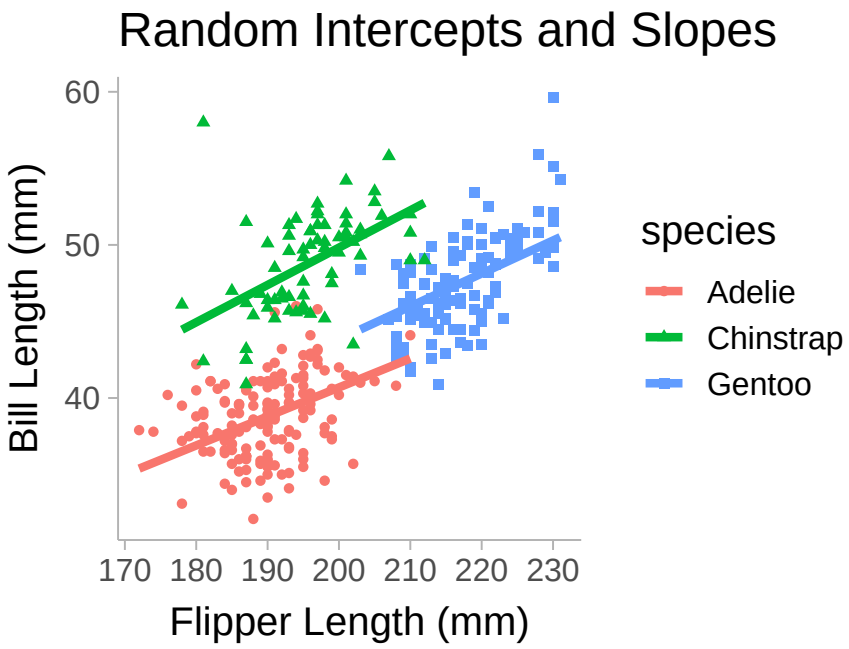


Figure 13.4: Multi-level regression model allowing each species to have their own intercept and slope.

```
           R2m       R2c
[1,] 0.2470201 0.8210591
```

As we can see the marginal R^2 is .25, whereas the conditional R^2 is .82.

14 Artifacts and Bias in Effect Sizes

14.1 Resources

Effect size estimates such as correlation coefficients and Cohen's d values can be severely biased due to various statistical artifacts such as measurement error and selection effects (e.g., range restriction). Methods have been developed to correct for the bias in effect sizes and thus these corrections are called "artifact corrections". Artifact correction formulas can be complex and therefore readers are referred to other resources listed below:

- Jané (2023) : An open-access textbook that contains equations and R code for various types of artifact corrections. In beta version.
- Hunter and Schmidt (1990) : Classic textbook on the topic of artifact corrections. Hunter and Schmidt pioneered the methodology for artifact correction style meta-analyses.
- Wiernik and Dahlke (2020) : A paper that serves as a condensed version of Hunter and Schmidt's book. It contains most of the equations necessary to correct effect sizes.
- Dahlke and Wiernik (2019) : An R package called `psychmeta` that allows meta-analysts to conduct artifact correction meta-analyses. Contains all the functions one would need to correct effect sizes for artifacts in R.

14.2 Correcting for Measurement Error

If we have reliability estimates of the variables of interest, we can correct a Pearson correlation or a standardized mean difference (Cohen's d) for measurement error. Non-differential measurement error attenuates Pearson correlations and Cohen's d, therefore we can apply correction factors to adjust for this bias. For a pearson correlation, we can use the correction for attenuation first developed by Spearman (1904),

$$r_c = \frac{r_{\text{obs}}}{\sqrt{r_{xx'} r_{yy'}}} \tag{14.1}$$

where r_c is the corrected correlation, r_{obs} is the observed correlation, $r_{xx'}$ is the reliability of x, and $r_{yy'}$ is the reliability of y. Reliability coefficients can be estimated a number of different

ways, however the two of the most common estimators are Cronbach's Alpha and test-retest reliability. Alpha measures the internal consistency of a set of sub-component measurements (e.g., question item responses on a questionnaire) while test-retest reliability measures the stability over time.

A Cohen's d can be corrected similarly to a correlation coefficient, however since d reflects the difference in a continuous variable (y) between two groups, we just need to correct for reliability in the continuous variable,

$$d_c = \frac{d_{\text{obs}}}{\sqrt{r_{yy'}}}.$$

However in the case of a Cohen's d, it is important that $r_{yy'}$ is the pooled within-group reliability (calculate pooled reliability the same way you calculate the pooled standard deviation for denominator of Cohen's d). If all you have is the total sample reliability (more commonly reported) you can follow this three step process (Wiernik and Dahlke 2020),

1. Convert the d value to a point-biserial correlation (see section on conversions)
2. Correct the point-biserial correlation using Equation 14.1 (setting $r_{xx'} = 1$)
3. Convert it back to a Cohen's d

Note that confidence intervals for r_c and d_c must also be corrected such that,

$$CI_{r_c} = \left[\frac{r_{\text{lower-bound}}}{\sqrt{r_{xx'} r_{yy'}}}, \frac{r_{\text{upper-bound}}}{\sqrt{r_{xx'} r_{yy'}}} \right]$$

and

$$CI_{d_c} = \left[\frac{d_{\text{lower-bound}}}{\sqrt{r_{yy'}}}, \frac{d_{\text{upper-bound}}}{\sqrt{r_{yy'}}} \right].$$

14.3 Correcting for Range Restriction

Range restriction corrections can be quite complex depending on the underlying selection process. The process for correcting Pearson correlations and Cohen's d for range restriction is laid out in table 3 of Wiernik and Dahlke (2020).

Part II

Converting Between Effect Sizes

15 Converting to Cohen's d

15.1 From Independent Samples t-statistic

To calculate a between subject standardized mean difference (d_p, i.e., pooled standard deviation standardizer), we can use the sample size in each group (n_A and n_B) as well as the t-statistic from an independent sample t-test and plug it into the following formula:

$$d_p = t\sqrt{\frac{1}{n_A} + \frac{1}{n_B}} \tag{15.1}$$

Note that this only works for Student's t-tests and not Welch's t-test (unless the sample sizes are equal). Using the `escalc()` function in the `metafor` package we can convert t to d_p or the associated *p*-value from the t-test to d_p.

```
library(metafor)

# Example:
# independent samples t-statistic = 3.25
# nA = 50, nB = 40

# calculate dp from t
stats <- escalc(measure = "SMD",
                ti = 3.25,
                n1i = 50,
                n2i = 40,
                var.names = c("dp","variance"))

# display results
summary(stats)
```

```
     dp variance    sei     zi   pval  ci.lb  ci.ub
1 0.6835   0.0476 0.2182 3.1331 0.0017 0.2559 1.1111
```

```r
# Example:
# p-value = .0016
# nA = 50, nB = 40

# calculate dp from p-value
stats <- escalc(measure = "SMD",
                pi = .0016,
                n1i = 50,
                n2i = 40,
                var.names = c("dp","variance"))

# display results
summary(stats)
```

```
     dp variance    sei     zi   pval  ci.lb  ci.ub
1 0.6850   0.0476 0.2182 3.1395 0.0017 0.2574 1.1127
```

15.2 From Paired Sample t-statistic

To calculate a within-subject standardized mean difference (d_z, i.e., difference score standardizer), we can use the sample size (n) as well as the t-statistic from a paired sample t-test and plug it into the following formula:

$$d_z = \frac{t}{\sqrt{n}} \tag{15.2}$$

Note that if we want to get the repeated measures d_{rm} value we can use $d_{rm} = d_z \times \sqrt{2(1-r)}$, but this will require the correlation between conditions. Using the `escalc()` function in the `metafor` package we can convert t to d_z.

```r
# Example:
# paired t-statistic = 3.25
# n = 50

# calculate dz from t-statistic
stats <- escalc(measure = "SMCC",
                ti = 3.25,
                ni = 50,
```

```
                  var.names = c("dz","variance"))

# display results
summary(stats)
```

```
     dz variance    sei     zi   pval  ci.lb  ci.ub
1 0.4525   0.0220 0.1485 3.0477 0.0023 0.1615 0.7436
```

```
# Example:
# p-value = 3.25
# n = 50

# calculate dz from p-value
stats <- escalc(measure = "SMCC",
                pi = .0021,
                ni = 50,
                var.names = c("dz","variance"))

# display results
summary(stats)
```

```
     dz variance    sei     zi   pval  ci.lb  ci.ub
1 0.4523   0.0220 0.1485 3.0462 0.0023 0.1613 0.7433
```

15.3 From Pearson Correlation

If a Pearson correlation is calculated between a continuous score and a dichotomous score (grouping variable with groups A and B), this is considered a point-biserial correlation. The point-biserial correlation can be converted into a d_p value using the following formula:

$$d_p = \frac{r}{\sqrt{1-r^2}}\sqrt{\frac{n_A+n_B-2}{n_A}+\frac{n_A+n_B-2}{n_B}}. \tag{15.3}$$

If the group sample sizes are equal ($n_A = n_B$), then the equation simplifies to be,

$$d_p = \frac{r\sqrt{4}}{\sqrt{1-r^2}} \tag{15.4}$$

Using the `escalc()` function in the `metafor` package we can convert r to d_p.

```
# Example:
# r = .40
# nA = 50, nB = 40

# calculate dp
stats <- escalc(measure = "SMD",
                ri = .4,
                n1i = 50,
                n2i = 40,
                var.names = c("dp","variance"))

# display results
summary(stats)
```

```
     dp variance    sei     zi   pval  ci.lb  ci.ub
1 0.8611   0.0491 0.2216 3.8852 0.0001 0.4267 1.2955
```

15.4 From Odds-Ratio

An odds-ratio from a contingency table can also be converted to a d_p. Note that this formula is an approximation:

$$d_p = \frac{\log(OR)\sqrt{3}}{\pi} \tag{15.5}$$

This method is the called the logit method as it uses the variance of the logistic distribution, however, it may actually be preferable to use the Cox logit method which is simply,

$$d_p = \frac{\log(OR)}{1.65} \tag{15.6}$$

Using the `oddsratio_to_d` function in the `effectsize` package we can convert OR to d_p with the logit method. we will use base R to calculate d_p with the Cox logit method.

```
library(effectsize)

# Example:
# OR = 1.46
OR <- 1.46

# calculate dp using logit method
oddsratio_to_d(OR = OR)
```

```
[1] 0.2086429
```

```
# calculate dp using cox logit method
OR / 1.65
```

```
[1] 0.8848485
```

16 Converting to Pearson Correlation

16.1 From t-statistic

From a t statistic calculated from a correlational test or an independent samples t-test, we can calculate the correlation coefficient using,

$$r = \frac{t}{\sqrt{t^2 + n - 2}}. \tag{16.1}$$

Where n is the sample size. Using the `escalc()` function in the `metafor` package (Viechtbauer 2010) we can convert t to r and we can convert the p-value from the test to r.

```
library(metafor)

# Example:
# t = 4.14, n = 50

# calculate correlation
# note: measure = "ZCOR" will give z-transformed correlations
stats <- escalc(measure = "COR",
                ti = 4.14,
                ni = 50,
                var.names = c("r","variance"))

# display results
summary(stats)
```

```
      r variance    sei     zi   pval  ci.lb  ci.ub
1 0.5130   0.0111 0.1053 4.8728 <.0001 0.3066 0.7193
```

```
# Example:
# p = .00014, n = 50
```

```
# using the p-value from the test
stats <- escalc(measure = "COR",
                pi = .00014,
                ni = 50)

# display results
summary(stats)
```

```
     yi     vi    sei     zi   pval  ci.lb  ci.ub
1 0.5129 0.0111 0.1053 4.8714 <.0001 0.3065 0.7192
```

16.2 From Cohen's d

From a between groups (i.e., groups A and B) Cohen's d value (d_p), we can calculate the correlation coefficient with the following formula:

$$r = \frac{d_p}{\sqrt{d_p^2 + \frac{n_A+n_B-2}{n_A} + \frac{n_A+n_B-2}{n_B}}} \tag{16.2}$$

Using the `d_to_r` function in the `effectsize` package we can convert d_p to r.

```
library(effectsize)

# Example:
# d = 0.60, nA = 50, nB = 70

# calculate and display correlation
d_to_r(d = 0.60,
       n1 = 50,
       n2 = 70)
```

```
[1] 0.2858532
```

16.3 From Odds-Ratio

The correlation coefficient from an odds ratio can be calculated with the following formula:

$$r = \frac{\log(OR) \times \sqrt{3}}{\pi \sqrt{\frac{3\log(OR)^2}{\pi^2} + \frac{n_1+n_2-2}{n_1} + \frac{n_1+n_2-2}{n_2}}} \tag{16.3}$$

Using the `oddsratio_to_r()` function in the `effectsize` package we can convert OR to r.

```
library(effectsize)

# Example:
# OR = 2.21, n1 = 50, n2 = 70

# calculate and display correlation
oddsratio_to_r(OR = 2.21,
               n1 = 50,
               n2 = 70)
```

```
[1] 0.2124017
```

17 Converting to Odds Ratio

17.1 From Cohen's d

We can calculate an odds-ratio from a between groups cohen's d (d_p):

$$OR = \exp\left(\frac{d_p \pi}{\sqrt{3}}\right) \tag{17.1}$$

Where $\exp(\cdot)$ is an exponential transformation (this inverses the logarithm). Using the `d_to_oddsratio()` function in the `effectsize` package we can convert d to OR.

```
library(effectsize)

# Example:
# d = 0.60, nA = 50, nB = 70

# calculate and display odds ratio
d_to_oddsratio(d = 0.60,
               n1 = 50,
               n2 = 70)
```

```
[1] 2.969162
```

17.2 From a Correlation

We can calculate an odds ratio from a point-biserial correlation using the following formula:

$$OR = \exp\left(\frac{r\pi\sqrt{\frac{n_A+n_B-2}{n_A} + \frac{n_A+n_B-2}{n_B}}}{\sqrt{3(1-r^2)}}\right) \tag{17.2}$$

When sample sizes are equal, this equation can be simplified to be approximately,

$$OR = \exp\left(\frac{r\pi\sqrt{4}}{\sqrt{3(1-r^2)}}\right) \tag{17.3}$$

Using the `r_to_oddsratio()` function in the `effectsize` package we can convert r to OR.

```
library(effectsize)

# Example:
# r = .50, n1 = 50, n2 = 70

# calculate and display odds ratio
r_to_oddsratio(r = .50,
               n1 = 50,
               n2 = 70)
```

```
[1] 8.120527
```

18 Interpreting Effect Sizes in Social Sciences

19 Conclusion

19.1 Limitations and Future Directions

While this guide covers a wide range of effect size and confidence interval methods, there are some limitations to note. First, our instructions focus specifically on applications in behavioral, cognitive, and social science research. The techniques may need to be adapted for other scientific domains. Second, we only cover free and open source options, so proprietary software packages are not discussed. Finally, as new methods and R packages arise, the guide will need to be continually updated, perhaps in a similar manner as Parsons et al. (2022) Open Scholarship terms after publication.

In the future, we aim to expand the guide by collaborating with experts in other fields to include discipline-specific recommendations. We also plan to incorporate new R packages and techniques as they emerge. Readers are encouraged to consult the cited packages' documentation and peer-reviewed sources to further explore limitations and assumptions of the covered techniques.

19.2 Conclusion

Robust quantification of study results is a central pillar of open and reproducible science. With this collaborative collection of applied instructions, our guide aims to make calculating effect sizes and confidence intervals more accessible. We hope these resources empower both young researchers and experienced scholars across a variety of disciplines to incorporate these crucial statistical practices into their workflows. In our view, more widespread and thoughtful adoption of these methods will greatly strengthen the collective rigor, transparency, and impact of scientific research.

References

Agresti, Alan. 1980. “Generalized Odds Ratios for Ordinal Data.” *Biometrics*, 59–67.

Algina, James, and H. J. Keselman. 2003. “Approximate Confidence Intervals for Effect Sizes.” *Educational and Psychological Measurement* 63 (4): 537–53. https://doi.org/10.1177/0013164403256358.

Anvari, Farid, and Daniël Lakens. 2021. “Using Anchor-Based Methods to Determine the Smallest Effect Size of Interest.” *Journal of Experimental Social Psychology* 96: 104159.

APA. 2010. *Publication Manual of the American Psychological Association*. American Psychological Association. https://thuvienso.hoasen.edu.vn/handle/123456789/8327.

Ara, Toshiaki. 2022. *Brunnermunzel: (Permuted) Brunner-Munzel Test*. https://CRAN.R-project.org/package=brunnermunzel.

Baayen, R Harald, Douglas J Davidson, and Douglas M Bates. 2008. “Mixed-Effects Modeling with Crossed Random Effects for Subjects and Items.” *Journal of Memory and Language* 59 (4): 390–412.

Baguley, Thom. 2009. “Standardized or Simple Effect Size: What Should Be Reported?” *British Journal of Psychology* 100 (3): 603–17.

Barr, Dale J, Roger Levy, Christoph Scheepers, and Harry J Tily. 2013. “Random Effects Structure for Confirmatory Hypothesis Testing: Keep It Maximal.” *Journal of Memory and Language* 68 (3): 255–78.

Bartoń, Kamil. 2023. *MuMIn: Multi-Model Inference*. https://CRAN.R-project.org/package=MuMIn.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. “Fitting Linear Mixed-Effects Models Using lme4.” *Journal of Statistical Software* 67 (1): 1–48. https://doi.org/10.18637/jss.v067.i01.

Beck, Edward C., Anirudh K. Gowd, Joseph N. Liu, Brian R. Waterman, Kristen F. Nicholson, Brian Forsythe, Adam B. Yanke, Brian J. Cole, and Nikhil N. Verma. 2020. “How Is Maximum Outcome Improvement Defined in Patients Undergoing Shoulder Arthroscopy for Rotator Cuff Repair? A 1-Year Follow-up Study.” *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 36 (7): 1805–10. https://doi.org/10.1016/j.arthro.2020.02.047.

Becker, Betsy J. 1988. “Synthesizing Standardized Mean-Change Measures - UConn Library.” *British Journal of Mathematical and Statistical Psychology* 41 (2): 257278. https://doi.org/https://doi.org/10.1111/j.2044-8317.1988.tb00901.x.

Ben-Shachar, Mattan S., Daniel Lüdecke, and Dominique Makowski. 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.

Ben-Shachar, Mattan S., Indrajeet Patil, Rémi Thériault, Brenton M. Wiernik, and Daniel Lüdecke. 2023. "Phi, Fei, Fo, Fum: Effect Sizes for Categorical Data That Use the Chi-Squared Statistic." *Mathematics* 11 (9): 1982. https://doi.org/10.3390/math11091982.

Blair, Graeme, Jasper Cooper, Alexander Coppock, and Macartan Humphreys. 2019. "Declaring and Diagnosing Research Designs." *American Political Science Review* 113: 838–59. https://declaredesign.org/paper.pdf.

Bonett, Douglas G. 2024. *Statpsych: Statistical Methods for Psychologists*. https://CRAN.R-project.org/package=statpsych.

Bonini, Matteo, Marcello Di Paolo, Diego Bagnasco, Ilaria Baiardini, Fulvio Braido, Marco Caminati, Elisiana Carpagnano, et al. 2020. "Minimal Clinically Important Difference for Asthma Endpoints: An Expert Consensus Report." *European Respiratory Review* 29 (156).

Bosco, Frank A., Herman Aguinis, Kulraj Singh, James G. Field, and Charles A. Pierce. 2015. "Correlational Effect Size Benchmarks." *Journal of Applied Psychology* 100 (2): 431–49. https://doi.org/10.1037/a0038047.

Brunner, Edgar, and Ullrich Munzel. 2000. "The Nonparametric Behrens-Fisher Problem: Asymptotic Theory and a Small-Sample Approximation." *Biometrical Journal* 42 (1): 17–25. https://doi.org/10.1002/(SICI)1521-4036(200001)42:1%3C17::AID-BIMJ17%3E3.0.CO;2-U.

Buchanan, Erin M., Amber Gillenwaters, John E. Scofield, and K. D. Valentine. 2019. *MOTE: Measure of the Effect: Package to Assist in Effect Size Calculations and Their Confidence Intervals*. http://github.com/doomlab/MOTE.

Caldwell, Aaron R. 2022. "Exploring Equivalence Testing with the Updated TOSTER r Package." *PsyArXiv*. https://doi.org/10.31234/osf.io/ty8de.

Cliff, Norman. 1993. "Dominance Statistics: Ordinal Analyses to Answer Ordinal Questions." *Psychological Bulletin* 114 (3): 494.

Coe, R. 2012. "It's the Effect Size, Stupid What Effect Size Is and Why It Is Important." In. https://www.semanticscholar.org/paper/It%27s-the-Effect-Size%2C-Stupid-What-effect-size-is-it-Coe/c5ac87df5d6e0e6b6de2f745284835c2a368b0f7.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Academic Press.

Dahlke, Jeffrey A., and Brenton M. Wiernik. 2019. "psychmeta: An r Package for Psychometric Meta-Analysis." *Applied Psychological Measurement* 43 (5): 415–16. https://doi.org/10.1177/0146621618795933.

Daste, Camille, Hendy Abdoul, Frantz Foissac, Marie-Martine Lefèvre-Colau, Serge Poiraudeau, François Rannou, and Christelle Nguyen. 2022. "Patient Acceptable Symptom State for Patient-Reported Outcomes in People with Non-Specific Chronic Low Back Pain." *Annals of Physical and Rehabilitation Medicine* 65 (1): 101451. https://doi.org/10.1016/j.rehab.2020.10.005.

Divine, George W, H James Norton, Anna E Barón, and Elizabeth Juarez-Colunga. 2018. "The Wilcoxon–Mann–Whitney Procedure Fails as a Test of Medians." *The American Statistician* 72 (3): 278–86.

Faul, Franz, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. "Statistical Power

Analyses Using G*Power 3.1: Tests for Correlation and Regression Analyses." *Behavior Research Methods* 41 (4): 1149–60. https://doi.org/10.3758/BRM.41.4.1149.

Fritz, Catherine O., Peter E. Morris, and Jennifer J. Richler. 2012. "Effect Size Estimates: Current Use, Calculations, and Interpretation." *Journal of Experimental Psychology: General* 141 (1): 2–18. https://doi.org/10.1037/a0024338.

Funder, David C., and Daniel J. Ozer. 2019. "Evaluating Effect Size in Psychological Research: Sense and Nonsense." *Advances in Methods and Practices in Psychological Science* 2 (2): 156–68. https://doi.org/10.1177/2515245919847202.

Gelman, Andrew. 2011. "Why It Doesn't Make Sense in General to Form Confidence Intervals by Inverting Hypothesis Tests | Statistical Modeling, Causal Inference, and Social Science." https://statmodeling.stat.columbia.edu/2011/08/25/why_it_doesnt_m/.

Gignac, Gilles E., and Eva T. Szodorai. 2016. "Effect Size Guidelines for Individual Differences Researchers." *Personality and Individual Differences* 102 (November): 74–78. https://doi.org/10.1016/j.paid.2016.06.069.

Glass, Gene V. 1981. "Meta-Analysis in Social Research." *(No Title)*. https://cir.nii.ac.jp/crid/1130000795088566912.

Glass, Gene V., Barry McGaw, and Mary L. Smith. 1981. "Meta-Analysis in Social Research." *(No Title)*. https://cir.nii.ac.jp/crid/1130000795088566912.

Guilford, J. P. 1965. "The Minimal Phi Coefficient and the Maximal Phi." *Educational and Psychological Measurement* 25 (1): 3–8. https://doi.org/10.1177/001316446502500101.

Harrell, Frank. 2020. "Author Checklist - Data Analysis." https://discourse.datamethods.org/t/author-checklist/3407.

Hedges, Larry V. 1981. "Distribution Theory for Glass's Estimator of Effect Size and Related Estimators." *Journal of Educational Statistics* 6 (2): 107–28. https://doi.org/10.3102/10769986006002107.

HEIJDE, DÉSIRÉE van der, MARISSA Lassere, JOHN Edmonds, JOHN Kirwan, VIBEKE Strand, and Maarten Boers. 2001. "Minimal Clinically Important Difference in Plain Films in RA: Group Discussions, Conclusions, and Recommendations. OMERACT Imaging Task Force." *The Journal of Rheumatology* 28 (4): 914–17.

Hill, Carolyn J, Howard S Bloom, Alison Rebeck Black, and Mark W Lipsey. 2008. "Empirical Benchmarks for Interpreting Effect Sizes in Research." *Child Development Perspectives* 2 (3): 172–77.

Hoekstra, Rink, Richard D. Morey, Jeffrey N. Rouder, and Eric-Jan Wagenmakers. 2014. "Robust Misinterpretation of Confidence Intervals." *Psychonomic Bulletin & Review* 21 (5): 1157–64. https://doi.org/10.3758/s13423-013-0572-3.

Horst, Allison Marie, Alison Presmanes Hill, and Kristen B Gorman. 2020. *Palmerpenguins: Palmer Archipelago (Antarctica) Penguin Data*. https://doi.org/10.5281/zenodo.3960218.

Hunter, John E., and Frank L. Schmidt. 1990. *Methods of meta-analysis: correcting error and bias in research findings*. Newbury Park: Sage Publications.

Jané, Matthew B. 2023. *Artifact Corrections for Effect Sizes: Implementation in r and Application to Meta-Analysis*. (n.p.). https://matthewbjane.quarto.pub/artifact-corrections-for-effect-sizes/.

Karch, Julian D. 2021. "Psychologists Should Use Brunner-Munzel's Instead of Mann-

Whitney's u Test as the Default Nonparametric Procedure." *Advances in Methods and Practices in Psychological Science* 4 (2): 2515245921999602.
Kassambara, Alboukadel. 2019. *Datarium: Data Bank for Statistical Analysis and Visualization*. https://CRAN.R-project.org/package=datarium.
Kelley, Ken. 2022. *MBESS: The MBESS r Package*. https://CRAN.R-project.org/package=MBESS.
Kelley, Ken, and Kristopher J. Preacher. 2012. "On Effect Size." *Psychological Methods* 17 (2): 137–52. https://doi.org/10.1037/a0028086.
Kendall, Maurice G. 1945. "The Treatment of Ties in Ranking Problems." *Biometrika* 33 (3): 239–51.
Kirby, Kris N, and Daniel Gerlanc. 2013. "BootES: An r Package for Bootstrap Confidence Intervals on Effect Sizes." *Behavior Research Methods* 45: 905–27.
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://www.frontiersin.org/articles/10.3389/fpsyg.2013.00863.
———. 2014. "The 20." http://daniellakens.blogspot.com/2014/06/calculating-confidence-intervals-for.html.
———. 2022. "Sample Size Justification." *Collabra: Psychology* 8 (1): 33267. https://doi.org/10.1525/collabra.33267.
Lakens, Daniël, Anne M Scheel, and Peder M Isager. 2018. "Equivalence Testing for Psychological Research: A Tutorial." *Advances in Methods and Practices in Psychological Science* 1 (2): 259–69.
Läkens, Daniel. 2017. "Equivalence Tests: A Practical Primer for t-Tests, Correlations, and Meta-Analyses." *Social Psychological and Personality Science* 1: 1–8. https://doi.org/10.1177/1948550617697177.
Liddell, Torrin M., and John K. Kruschke. 2018. "Analyzing Ordinal Data with Metric Models: What Could Possibly Go Wrong?" *Journal of Experimental Social Psychology* 79 (November): 328–48. https://doi.org/10.1016/j.jesp.2018.08.009.
Lovakov, Andrey, and Elena R. Agadullina. 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.
Lüdecke, Daniel. 2019. *Esc: Effect Size Computation for Meta Analysis (Version 0.5.1)*. https://doi.org/10.5281/zenodo.1249218.
Magnusson, Kristoffer. 2023. "A Causal Inference Perspective on Therapist Effects."
McGlothlin, Anna E., and Roger J. Lewis. 2014. "Minimal Clinically Important Difference: Defining What Really Matters to Patients." *JAMA* 312 (13): 1342–43. https://doi.org/10.1001/jama.2014.13128.
Meehl, Paul E. 1984. "Radical Behaviorism and Mental Events: Four Methodological Queries." *Behavioral and Brain Sciences* 7 (4): 563–64. https://doi.org/10.1017/S0140525X00027308.
Morey, Richard D., Rink Hoekstra, Jeffrey N. Rouder, Michael D. Lee, and Eric-Jan Wagenmakers. 2016. "The Fallacy of Placing Confidence in Confidence Intervals." *Psychonomic Bulletin & Review* 23 (1): 103–23. https://doi.org/10.3758/s13423-015-0947-8.

Morris, Scott B. 2000. “Distribution of the Standardized Mean Change Effect Size for Meta-Analysis on Repeated Measures.” *British Journal of Mathematical and Statistical Psychology* 53 (1): 17–29.

———. 2008. “Estimating Effect Sizes From Pretest-Posttest-Control Group Designs.” *Organizational Research Methods* 11 (2): 364–86. https://doi.org/10.1177/1094428106291059.

Morse, David. 2018. “How to Calculate Degrees of Freedom When Using Two Way ANOVA with Unequal Sample Size?”

Munzel, Ullrich, and Edgar Brunner. 2002. “An Exact Paired Rank Test.” *Biometrical Journal* 44 (5): 584–93. https://doi.org/10.1002/1521-4036(200207)44:5%3C584::AID-BIMJ584%3E3.0.CO;2-9.

Neubert, Karin, and Edgar Brunner. 2007. “A Studentized Permutation Test for the Non-Parametric Behrens–fisher Problem.” *Computational Statistics & Data Analysis* 51 (10): 5192–5204. https://doi.org/10.1016/j.csda.2006.05.024.

O’Brien, Ralph G, and John Castelloe. 2006. “Exploiting the Link Between the Wilcoxon-Mann-Whitney Test and a Simple Odds Statistic.” In *Proceedings of the Thirty-First Annual SAS Users Group International Conference*, 209–31. Citeseer.

Olkin, Ingram, and Jeremy D. Finn. 1995. “Correlations Redux.” *Psychological Bulletin* 118 (1): 155–64. https://doi.org/10.1037/0033-2909.118.1.155.

Orben, Amy, and Daniël Lakens. 2020. “Crud (Re)Defined.” *Advances in Methods and Practices in Psychological Science* 3 (2): 238–47. https://doi.org/10.1177/2515245920917961.

Otgaar, Henry, Paul Riesthuis, Tess Neal, Jason Chin, Irena Boskovic, and Eric Rassin. 2023. “If Generalization Is the Grail, Practical Relevance Is the Nirvana: Considerations from the Contribution of Psychological Science of Memory to Law.” *Henry Otgaar, Paul Riesthuis, Tess MS Neal, Jason M. Chin, Irena Boskovic & Eric Rassin,“If Generalization Is the Grail, Practical Relevance Is the Nirvana: Considerations from the Contribution of Psychological Science of Memory to Law”(accepted 2023) Journal of Applied Research in Memory and Co*.

Otgaar, Henry, Paul Riesthuis, Johannes G Ramaekers, Maryanne Garry, and Lilian Kloft. 2022. “The Importance of the Smallest Effect Size of Interest in Expert Witness Testimony on Alcohol and Memory.” *Frontiers in Psychology* 13: 980533.

Panzarella, Emily, Nataly Beribisky, and Robert A Cribbie. 2021. “Denouncing the Use of Field-Specific Effect Size Distributions to Inform Magnitude.” *PeerJ* 9: e11383.

Paterson, Ted A., P. D. Harms, Piers Steel, and Marcus Credé. 2016. “An Assessment of the Magnitude of Effect Sizes: Evidence From 30 Years of Meta-Analysis in Management.” *Journal of Leadership & Organizational Studies* 23 (1): 66–81. https://doi.org/10.1177/1548051815614321.

Pogrow, Stanley. 2019. “How Effect Size (Practical Significance) Misleads Clinical Practice: The Case for Switching to Practical Benefit to Assess Applied Research Findings.” *The American Statistician* 73 (sup1): 223–34. https://doi.org/10.1080/00031305.2018.1549101.

Richard, F. D., Charles F. Bond Jr., and Juli J. Stokes-Zoota. 2003. “One Hundred Years of Social Psychology Quantitatively Described.” *Review of General Psychology* 7 (4): 331–

63. https://doi.org/10.1037/1089-2680.7.4.331.
Riesthuis, Paul, Ivan Mangiulli, Nick Broers, and Henry Otgaar. 2022. “Expert Opinions on the Smallest Effect Size of Interest in False Memory Research.” *Applied Cognitive Psychology* 36 (1): 203–15.
Rohrmann, Bernd. 2007. “Verbal Qualifiers for Rating Scales: Sociolinguistic Considerations and Psychometric Data.” *Project Report, University of Melbourne/Australia*.
Rossi, Michael J, Jefferson C Brand, and James H Lubowitz. 2023. “Minimally Clinically Important Difference (MCID) Is a Low Bar.” *Arthroscopy: The Journal of Arthroscopic & Related Surgery*. Elsevier.
Sawilowsky, Shlomo. 2009. “New Effect Size Rules of Thumb.” *Journal of Modern Applied Statistical Methods* 8 (2). https://doi.org/10.22237/jmasm/1257035100.
Schäfer, Thomas, and Marcus A. Schwarz. 2019. “The Meaningfulness of Effect Sizes in Psychological Research: Differences Between Sub-Disciplines and the Impact of Potential Biases.” *Frontiers in Psychology* 10. https://www.frontiersin.org/articles/10.3389/fpsyg.2019.00813.
Senior, Alistair M., Wolfgang Viechtbauer, and Shinichi Nakagawa. 2020. “Revisiting and Expanding the Meta-Analysis of Variation: The Log Coefficient of Variation Ratio.” *Research Synthesis Methods* 11 (4): 553–67. https://doi.org/10.1002/jrsm.1423.
Spearman, C. 1904. “The Proof and Measurement of Association Between Two Things.” *International Journal of Epidemiology* 39 (5): 1137–50. https://doi.org/10.1093/ije/dyq191.
Steiger, James 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): 164–82. https://doi.org/10.1037/1082-989X.9.2.164.
Torchiano, Marco. 2020. *Effsize: Efficient Effect Size Computation*. https://doi.org/10.5281/zenodo.1480624.
Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with s*. Fourth. New York: Springer. https://www.stats.ox.ac.uk/pub/MASS4/.
Viechtbauer, Wolfgang. 2010. “Conducting Meta-Analyses in R with the metafor Package.” *Journal of Statistical Software* 36 (3): 1–48. https://doi.org/10.18637/jss.v036.i03.
Vos, Paul, and Don Holbert. 2022. “Frequentist Statistical Inference Without Repeated Sampling.” *Synthese* 200 (2): 89. https://doi.org/10.1007/s11229-022-03560-x.
W. T. Hoyt, A. C. Del Re &. 2014. *MAd: Meta-Analysis with Mean Differences*. *R Package*. https://CRAN.R-project.org/package=MAd.
Wellington, Ian J., Annabelle P. Davey, Mark P. Cote, Benjamin C. Hawthorne, Caitlin G. Dorsey, Patrick M. Garvin, James C. Messina, Cory R. Hewitt, and Augustus D. Mazzocca. 2023. “Substantial Clinical Benefit Values Demonstrate a High Degree of Variability When Stratified by Time and Geographic Region.” *JSES International* 7 (1): 153–57. https://doi.org/10.1016/j.jseint.2022.10.003.
Wiernik, Brenton M., and Jeffrey A. Dahlke. 2020. “Obtaining Unbiased Results in Meta-Analysis: The Importance of Correcting for Statistical Artifacts.” *Advances in Methods and Practices in Psychological Science* 3 (1): 94–123. https://doi.org/10.1177/2515245919885611.
William Revelle. 2023. *Psych: Procedures for Psychological, Psychometric, and Person-*

ality Research. Evanston, Illinois: Northwestern University. https://CRAN.R-project.org/package=psych.

Yang, Yefeng, Helmut Hillebrand, Malgorzata Lagisz, Ian Cleasby, and Shinichi Nakagawa. 2022. "Low Statistical Power and Overestimated Anthropogenic Impacts, Exacerbated by Publication Bias, Dominate Field Studies in Global Change Biology." *Global Change Biology* 28 (3): 969–89. https://doi.org/10.1111/gcb.15972.