Guide to Effect Sizes and Confidence Intervals

Matthew B. Jané
Qinyu Xiao
Siu Kit Yeung
Daniel J. Dunleavy
Lukas Röseler
Mahmoud Elsherif
Denis Cousineau
Aaron R. Caldwell
Blair T. Johnson
David Moreau
Felipe F. Vieira
Mircea Zloteanu
Gilad Feldman

Guide to Effect Sizes and Confidence Intervals

Matthew B. Jané   Qinyu Xiao   Siu Kit Yeung
Daniel J. Dunleavy   Lukas Röseler   Mahmoud Elsherif
Denis Cousineau   Aaron R. Caldwell   Blair T. Johnson
David Moreau   Felipe F. Vieira   Mircea Zloteanu
Gilad Feldman

2023

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, use the following google doc: .

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.

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 our empirical research. 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 a standardized scale which has no longer any unit but which have a universal interpretation. A z score is an example of a standardized measure. This document is concerned exclusively on standardized effect sizes.

2 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 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 thread for a summary) are typically smaller than what Cohen put forward. Please refer to the table below:

Effect Size	Reference	Small	Medium	Large
Mean Differences				
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				

[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.

Effect Size	Reference	Small	Medium	Large
Correlation Coefficient (r)	Cohen (1988)	.10	.30	.50
	Richard, Bond Jr., and Stokes-Zoota (2003)[34]	.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 q				
Cohen's f	Cohen (1988)	.10	.25	.40
Categorical				
Cohen's omega	Cohen (1988)	0.10	0.30	0.50
Phi	Cohen (1988)	.10	.30	.50
Cramer's V		[5]		
Odds ratio				
Relative risk				
Risk difference				
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

[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.

[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.

rate before and after intervention).

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.

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 effect sizes 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.

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). 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 one of the following:

1. Your effect size estimate does not fall in your confidence interval: you certainly have an issue. This can occur for example, if the width of the CI can be smaller than the tolerance of the optimizer, resulting in CIs of width 0.
2. One of your CI bound is "infinite"
3. Your effect size estimate is not included within your CI (for comparison between two groups): You 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 calculaton 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 does some conversion. 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.

7 Artifacts and Bias in Effect Sizes

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.
- 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 for conducting artifact correction meta-analyses. Contains all the functions one would need to correct effect sizes for artifacts in R.

Part I

Standardized Effect Sizes

8 Standardized Effect Sizes for 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 is fairly straightforward. Nonetheless, there are cases where crucial figures for the calculation are missing (which happens quite often in older articles), and therefore we document methods that make use of partial information (e.g., only the M and the SD, or only the t-statistic and df) for the calculation. There are multiple types of effect sizes used to calculate standardized mean differences (i.e., Cohen's 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 show the equations to calculate the standard error. The standard allows us to then calculate the confidence interval. For each formulation in the sections below, the confidence interval will be able to be calculated in the same way, that is,

$$CI_d = d \pm 1.96 \times SE$$

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

8.1 Single Group Designs

For a single group design, we have one group and 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,

$$d_s = \frac{M - C}{S_1}$$

A positive d_s value would indicate that the mean of group 1 is larger than the target value, C. This formulation assumes that the sample is drawn from a normal distribution. The standardizer (i.e., the denominator) is the sample standard deviation. The corresponding standard error for d_s is,

$$SE_{d_s} = \sqrt{\frac{1}{n} + \frac{d_1^2}{2n}}.$$

In R, we can use the `d.single.t` function from the `MOTE` package 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
)

# print just the d value and confidence intervals
data.frame(d = apa(stats$d),
           dlow = apa(stats$dlow),
           dhigh = apa(stats$dhigh))
```

```
      d  dlow dhigh
1 0.684 0.460 0.904
```

As you can see, the output shows that the effect size is d_s = 0.68, 95% CI [0.46, 0.90]. Note the `apa` function in `MOTE` takes a value and returns an APA formatted effect size value (i.e., leading zero and three decimal places).

8.2 Two Groups Design

8.2.1 Standardize by Pooled Standard Deviation (d_p)

For a two group design (i.e., between-groups design), we want to compare the means of two groups (group 1 and group 2). The standardized mean difference between two groups can be calculated by,

$$d_p = \frac{M_1 - M_2}{S_p}.$$

A positive d_p value would indicate that the mean of group 1 is larger than the mean of group 2. Dividing the mean difference by the pooled standard deviation, S_p, is the classic formulation of Cohen's d. The pooled standard deviation, S_p, can be calculated as the square root of the average variance (weighted by the degrees of freedom, $df = n - 1$) of group 1 and group 2:

$$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

Note that the term *variance* refers to the square of the standard deviation (S^2). Cohen's d_p has 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:

$$d = t\sqrt{\frac{2(n_1 + n_2)}{n_1 n_2 (n_1 + n_2 - 2)}}.$$

The corresponding standard error of d_p is,

$$SE_{d_p} = \sqrt{\frac{n_1 + n_2}{n_1 n_2} + \frac{d_p^2}{2(n_1 + n_2)}}.$$

In R, we can use the `d.ind.t` function from the `MOTE` package to calculate the two group standardized mean difference. Since we have already loaded in the `MOTE` package, we do not need to again.

```
# Cohen's d for two independent groups
# given means and SDs

# For example:
# Group 1 Mean = 30.4, SD = 22.53, N = 96
# Group 2 Mean = 21.4, SD = 19.59, N = 96

stats <- d.ind.t(
  m1 = 30.4,
  m2 = 21.4,
  sd1 = 22.53,
  sd2 = 19.59,
  n1 = 96,
  n2 = 96,
  a = 0.05
)

# print just the d value and confidence intervals
data.frame(d = apa(stats$d),
           dlow = apa(stats$dlow),
           dhigh = apa(stats$dhigh))
```

```
      d  dlow dhigh
1 0.426 0.140 0.712
```

The output shows that the effect size is d_p = 0.43, 95% CI [0.14, 0.71].

8.2.2 Standardize by Control Group Standard Deviation (d_Δ)

When two groups differ substantially in their standard deviations, we can instead standardize by the control group standard deviation (S_C), such that,

$$d_\Delta = \frac{M_T - M_C}{S_C}.$$

Where the subscripts, T and C, denotes the treatment group and control group, respectively. This formulation is commonly referred to as Glass' Δ (Glass 1981). The standard error for d_Δ can be defined as,

$$SE_{d_\Delta} = \sqrt{\frac{n_T + n_C}{n_T n_C} + \frac{d_\Delta^2}{n_C + 1}}$$

Notice that when we only standardize by the standard deviation of the control group (rather than pooling), we he will have less degrees of freedom ($df = n_C - 1$) and therefore more sampling error than we do when we divide by the pooled standard deviation ($df = n_T + n_C - 2$).In R, we can use the `delta.ind.t.diff` function from the `MOTE` package to calculate d_Δ.

```
# Cohen's dz for difference scores
# given difference score means and SDs

# For example:
# Control group Mean = 30.4, SD = 22.53, N = 96
# Treatment group Mean = 21.4, SD = 19.59, N = 96
# correlation between conditions: r = .40

stats <- delta.ind.t(
  m1 = 30.4,
  m2 = 21.4,
  sd1 = 22.53,
  sd2 = 19.59,
  n1 = 96,
  n2 = 96,
  a = 0.05
)

# print just the d value and confidence intervals
data.frame(d = apa(stats$d),
           dlow = apa(stats$dlow),
           dhigh = apa(stats$dhigh))
```

```
      d  dlow dhigh
1 0.399 0.140 0.712
```

8.3 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 8.1 for a visualization). A standardized mean difference in a repeated-measures design can take on a few different forms that we define below.

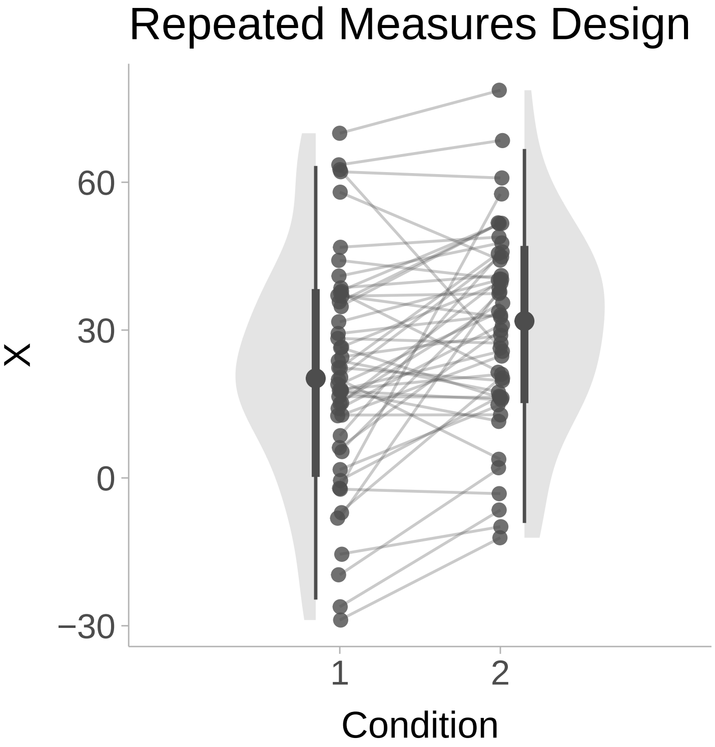

Figure 8.1: Figure displaying simulated data of a repeated measures design, the x-axis shows the condition (e.g., pre-test and post-test) and y-axis is the scores. Lines connect the change within subject from one condition to the next.

8.3.1 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. These difference scores ($X_{\text{diff}} = X_2 - X_1$) can be used similarly to the single group design (if the target value was zero, i.e., $C = 0$) such that,

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

Where the difference between this formulation and the single group design is the nature of the scores (difference scores rather than raw scores). The convenient thing about d_z is that it has a straight-forward relationship with the 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):

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

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,

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

In R, we can use the `d.ind.t.diff` function from the `MOTE` package to calculate d_z.

```r
# Cohen's dz for difference scores
# given difference score means and SDs

# For example:
# Difference Score Mean = 21.4, SD = 19.59, N = 96

library(MOTE)

stats <- d.dep.t.diff(
  m = 21.4,
  sd = 19.59,
  n = 96,
  a = 0.05
)

# print just the d value and confidence intervals
data.frame(d = apa(stats$d),
           dlow = apa(stats$dlow),
           dhigh = apa(stats$dhigh))
```

```
      d  dlow dhigh
1 1.092 0.837 1.344
```

The output shows that the effect size is d_z = 1.09, 95% CI [0.84, 1.34].

8.3.2 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,

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

A positive d_{rm} value would indicate that the mean of condition 2 is larger than the mean of condition 1. The standardizer here is the within-subject standard deviation, S_w. The within-subject standard deviation can be defined as,

$$S_{\text{diff}} = \sqrt{\frac{S_1^2 + S_2^2 - 2rS_1S_2}{2(1-r)}}.$$

We can also express S_w in terms of S_{diff},

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

Furthermore, we can even express d_{rm} in terms of d_z,

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

Ultimately the 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,

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

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}).

```r
# Cohen's d for repeated measures
# given means and SDs and correlation

# 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

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

# print just the d value and confidence intervals
data.frame(d = apa(stats$d),
           dlow = apa(stats$dlow),
           dhigh = apa(stats$dhigh))
```

```
      d  dlow dhigh
1 0.425 0.215 0.633
```

The output shows that the effect size is d_{rm} = 0.42, 95% CI [0.21, 0.63].

8.3.3 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 Cohen's d in repeated measures design is to simply use the classic variation of cohen's d (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 (i.e., $df = n - 1$) is an unnecessary step. Instead, we can standardize by the square root of the average the variances of condition 1 and 2:

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

This formulation is convenient especially when the correlation is not present, however without the correlation it fails to take into account the consistency of change between conditions. The standard error of the d_{av} can be expressed as,

$$SE_{d_{av}} = \sqrt{\frac{2}{n} + \frac{d_{av}^2}{4n}}$$

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 (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

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

# print just the d value and confidence intervals
data.frame(d = apa(stats$d),
           dlow = apa(stats$dlow),
           dhigh = apa(stats$dhigh))
```

```
      d  dlow dhigh
1 0.427 0.217 0.635
```

The output shows that the effect size is d_{av} = 0.43, 95% CI [0.22, 0.64].

8.3.4 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 when the comparison is a pre/post design,

$$d_b = \frac{M_{\text{post}} - M_{\text{pre}}}{S_{\text{pre}}}.$$

The convenient interpretation of “change in baseline standard deviations” can be quite useful. We can also obtain the standard error with,

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

Notice that even though the formula for calculating d_b did not include the correlation coefficient, the standard error does.

In base R, we can calculate Becker’s formulation of standardized mean difference using the equations above.

```r
# Install the package below if not done so already
# install.packages(escalc)
# Cohen's d for repeated measures (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

Mpre <- 21.4
Mpost <- 30.4
Spre <- 19.59
r <- .40
n <- 96
a <- 0.05

d <- (Mpost - Mpre) / Spre

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

stats <- data.frame(d = d,
                    dlow = d - 1.96*SE,
                    dhigh = d + 1.96*SE)

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

```
data.frame(d = apa(d),
           dlow = apa(d - 1.96*SE),
           dhigh = apa(d + 1.96*SE))
```

```
      d  dlow dhigh
1 0.459 0.231 0.688
```

The output shows that the effect size is d_{rm} = 0.46, 95% CI [0.23, 0.69].

8.3.5 Comparing Repeated Measures d values

Figure 8.2 shows repeated measures designs with a high ($r =$ .95) and low ($r =$ .05) correlation between conditions. Let us fix the standard deviations and means for both conditions (i.e., high and low correlation) and only vary the correlation. Now we can compare the repeated measures estimators based on these two conditions shown in Figure 8.2:

- 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.

8.4 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:

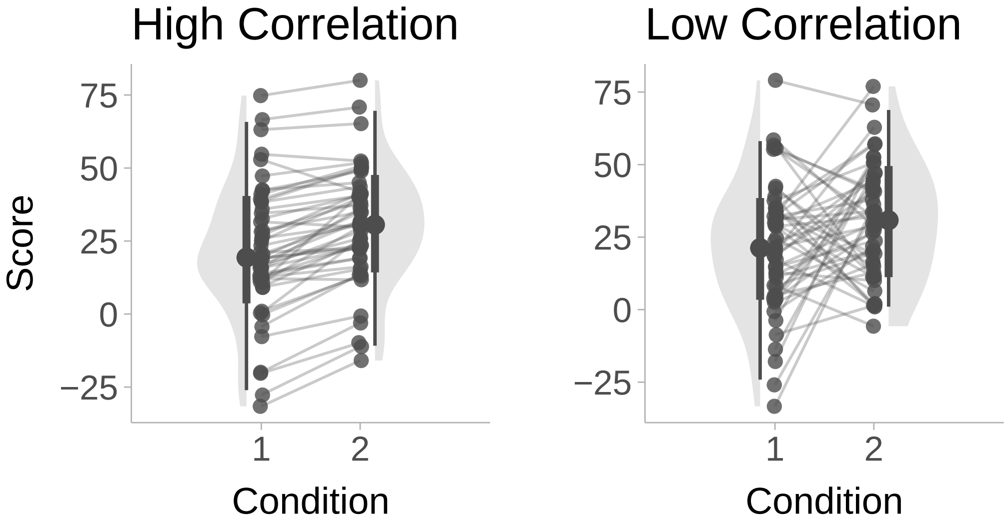

Figure 8.2: Figure displaying simulated data of a repeated measures design, the x-axis shows the condition (e.g., pre-test and post-test) and y-axis is the scores. Lines connect the change within subject from one condition to the next.

- Single Group design (d_s): $df = n - 1$
- Between Groups - Pooled Standard Deviation (d_p): $df = n_1 + n_2 - 2$
- Between Groups - Control Group Standard Deviation (d_Δ): $df = n_C - 1$
- Repeated Measures - all types (d_z, d_{rm}, d_{av}, d_b): $df = n - 1$

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:

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

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,

```r
# Example:
# Group 1 sample size = 20
# Group 2 sample size = 18

n1 <- 20
n2 <- 18

df <- n1 + n2 - 2

CF <- gamma(df/2) / ( sqrt(df/2) * gamma((df-1)/2) )

CF
```

```
[1] 0.9789964
```

This correction factor can then be applied to any of the estimators mentioned above,

$$d^* = d \times CF$$

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. We also need to correct the standard error for d^*

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

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

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

```
# 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
dlow_corrected <- d_corrected - 1.96*SE_corrected
dhigh_corrected <- d_corrected + 1.96*SE_corrected

# print just the d value and confidence intervals
data.frame(d = apa(d),
           dlow = apa(dlow_corrected),
           dhigh = apa(dhigh_corrected))
```

```
      d  dlow dhigh
1 0.500 0.298 0.681
```

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

9 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 9.1 shows scatter plots with differing levels of correlation.

The standard error of the Pearson correlation coefficient is,

$$SE_r = \frac{1 - r^2}{\sqrt{n-1}}$$

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. Instead, 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 ($SE_{Z_r} = \frac{1}{\sqrt{n-3}}$). We can also back-transform the confidence into a Pearson correlation scale,

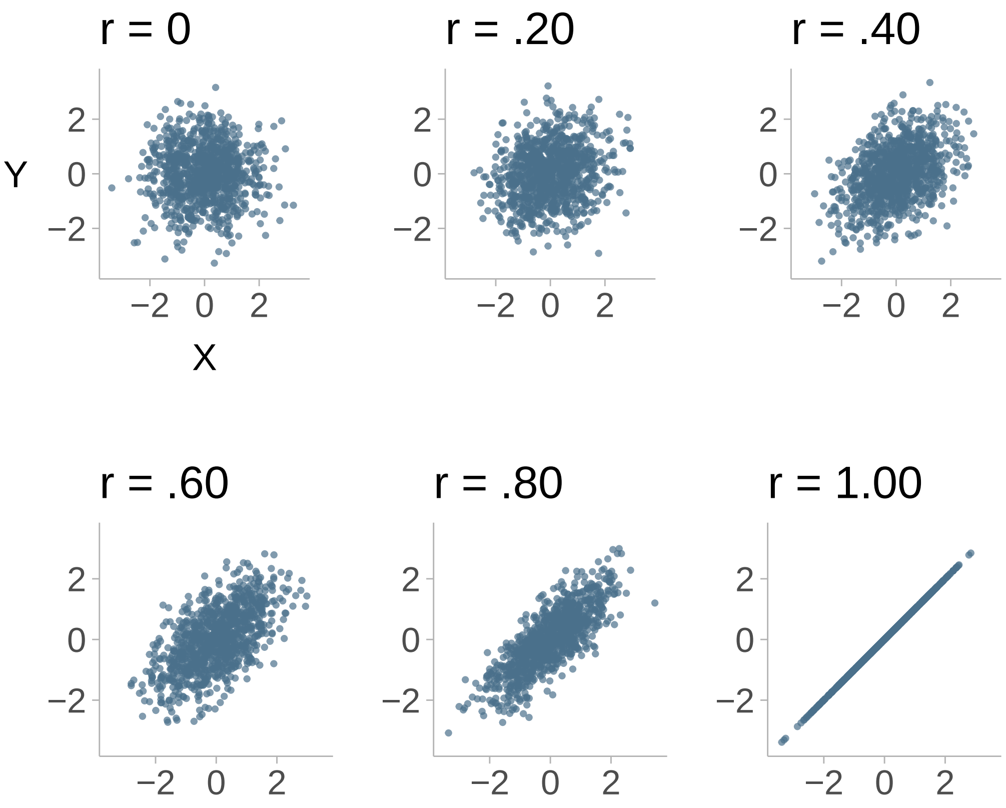

Figure 9.1: Simulated data from a bivariate normal distribution displaying 6 different correlations, r = 0, .20, .40, .60, .80, and 1.00.

$$CI_r = \text{tanh}(Z_r \pm 1.96 \times SE_{Z_r})$$

We can then back-transform the upper bound and lower bound into the upper and lower bound of r by taking the hyperbolic tangent (the inverse of the arctangent).

In R, the full process of obtaining confidence intervals can be done quite easily. Note if you have raw data for X and Y, then you can compute the correlation with base R, `cor(X,Y)`.

```r
# example: r = .50, n = 50
r <- .50
n <- 50

# compute Zr
Zr <- atanh(r)

# calculate standard error of Zr
SE_Zr <- 1/sqrt(n-3)

# compute confidence interval of Zr
Zlow <- Zr - 1.96 * SE_Zr
Zhigh <- Zr + 1.96 * SE_Zr

# backtransform CI of Z to CI of Pearson correlation
rlow <- tanh(Zlow)
rhigh <- tanh(Zhigh)

# print pearson correlation and confidence intervals
data.frame(r = MOTE::apa(r),
           rlow = MOTE::apa(rlow),
           rhigh = MOTE::apa(rhigh))
```

```
      r  rlow rhigh
1 0.500 0.257 0.683
```

The output shows that the correlation and its confidence intervals are r = 0.50, 95% CI [0.26, 0.68].

10 Effect Sizes for Categorical Variables

For dichotomous relationships that involve proportions, 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).

10.1 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, phi 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; phi only provides a measure of strength. The 2 by 2 contingency table is illustrated by Table 10.1.

Table 10.1: 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,

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

or more conveniently, from the χ^2-statistic,

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

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

```
# Example contingency table:
#  40  17
#  11  45

library(effectsize)

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

phi_coefficient <- phi(contingency_table, alternative = "two.sided")

phi_coefficient
```

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

In our example we obtained a phi coefficient of ϕ = .50 [0.31, 0.69].

10.2 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 10.2 represents a 3×4 contingency table of two variables.

Table 10.2: Contingency table between two categorical variables

	$X = 0$	$X = 1$	$X = 2$	$X = 3$
$Y = 0$	n_{00}	n_{10}	n_{21}	n_{31}

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

Similarly to the phi coefficient, the value of Cramer's V ranges from 0 to 1 and can interpreted in a similar way to a 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),

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

The standard error of a Cramer's V is similar to that of a Pearson correlation and a ϕ coefficient.

$$SE_V = \sqrt{\frac{1 - V^2}{n - 2}}$$

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:

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

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

Using the `ufs` package (Peters and Gruijters 2023), 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$×$3 contingency table.

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

```r
library(ufs)
```

```

Attaching package: 'ufs'

The following object is masked from 'package:effectsize':

    arr
```

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


V <- cramersV(contingency_table)
CI <- confIntV(contingency_table)

# print pearson correlation and confidence intervals
data.frame(V = MOTE::apa(V$output$cramersV),
           Vlow = MOTE::apa(CI$output$confIntV.fisher[1]),
           Vhigh = MOTE::apa(CI$output$confIntV.fisher[2]))
```

```
      V  Vlow Vhigh
1 0.442 0.309 0.558
```

In our example we obtained a Cramer's V of V = .44 [.31, .56].

10.3 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, its arcsine transformation is given by:

$$\psi = 2 \cdot \text{arcsin}(\sqrt{p}).$$

Cohen's h is the difference between the arcsine transformations of two proportions:

$$h = \psi_1 - \psi_2$$

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

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

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

$$CI_h = h \pm 1.96 \times SE_h$$

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

```r
# install package if not done so already
# install.packages('effectsize')
# Example proportions: p1 = .45, p2 = .30

library(effectsize)

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

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 [0.52, 1.34].

10.4 Cohen's omega (ω)

Although Cohen's ω 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 ω] lacks familiarity and convenience

Cohen's ω 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,

$$\omega = \sqrt{\frac{\chi^2}{n}}$$

And can also be calculated directly from Cramer's V,

$$\omega = V \times \sqrt{k-1}$$

Where k is the number of categories in the variable with the least number of categories.

```r
# Example contingency table


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

cohens_w(contingency_table,
         alternative = "two.sided")
```

```
Cohen's w |       95% CI
------------------------
0.45      | [0.24, 0.65]
```

From the example code, the `cohens_w` function returned Cohen's ω value of ω = .45 [0.24, 0.65].

10.5 Ben-Shachar's Fei ( )

Ben-Shachar et al. (2023) introduced a new effect size for contingency tables that they label with the Hebrew letter, .פ Ben-Shachar's פ is a correction to Cohen's ω that adjusts for the expected value and consequently bounds the value between 0 and 1 (Cohen's ω is only bounded between 0 and 1 when marginal distributions are uniform).

Cohen's ω 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,

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

Where $\text{min}(P_E)$ is the smallest expected probability in the contingency table. An expected probability for a given cell is the proportion of the total sample that would be expected to exist in that cell if the two variables were independent. The formula for Ben-Schacher's פ can be also be expressed in terms of Cohen's ω,

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

In R, we can calculate Ben-Shacher's פ using the `effectsize` package.

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

observed_counts <- c(20,50,100)
expected_probabilities <- c(.5,.2,.3)


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 `cohens_w` function returned Ben-Shacher's פ value of .39 [0.31, 0.47].

10.6 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:

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

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 10.3).

Table 10.3: 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 ($X = T$) and the control group ($X = C$). Therefore we can compute the odds ratio as,

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

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,

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

With the standard error of the log odds ratio we can then calculate the log odds ratio,

$$LOR_{UP} = LOR - 1.96 \times SE_{LOR}$$

Then the final step we can convert these back to OR confidence intervals by taking the exponential transformation of the upper and lower bounds: $OR_{UP} = \text{exp}(LOR_{UP})$ and $OR_{UP} = \text{exp}(LOR_{UP})$.

In R, we can use the `effectsize` package to calculate the odds ratio and it's confidence interval:

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


contingency_table <- matrix(c(10, 24,
                              43, 41),ncol = 2)

oddsratio(contingency_table,
          alternative = "two.sided")
```

```
Odds ratio |        95% CI
-------------------------
0.40       | [0.17, 0.93]
```

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

10.7 Risk Difference (RD)

Risk difference can be used to interpret the difference between two proportions. If we use the contingency table from Table 10.3, and calculate a risk difference between the treatment group and the control group. We can first calculate the proportion of cases where the outcome is $Y = 1$ *within* the control group and the treatment group:

$$p_C = \frac{n_{C1}}{n_{C0} + n_{C1}}$$

$$p_T = \frac{n_{T1}}{n_{T0} + n_{T1}}$$

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

$$RD = p_T - p_C.$$

The corresponding standard error is,

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

Where n_C and n_T are the total sample sizes *within* the control and treatment group, respectively. The standard error can then be used to compute the 95% confidence intervals,

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

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

```
# Example:
# Treatment group: proportion of cases = .5, sample size = 40
# Control group: proportion of cases = .3, sample size = 45

pT <- .50
pC <- .30
nT <- 40
nC <- 45

RD <- pT - pC

SE <- sqrt( pC*(1-pC)/nC + pT*(1-pT)/nT )

# compute 95% CIs
RDlow <- RD - 1.96*SE
RDhigh <- RD + 1.96*SE

data.frame(
```

```r
    RD = MOTE::apa(RD),
    RDlow = MOTE::apa(RDlow),
    RDhigh = MOTE::apa(RDhigh)
    )
```

```
     RD  RDlow RDhigh
1 0.200 -0.005  0.405
```

10.8 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 riskm, first we need to calculate the proportion of outcome cases in the treatment and control group

$$p_C = \frac{n_{C1}}{n_{C0} + n_{C1}}$$

$$p_T = \frac{n_{T1}}{n_{T0} + n_{T1}}$$

Then we can calculate the relative risk,

$$RR = \frac{p_T}{p_C}$$

The corresponding standard error can be computed as,

$$SE_{RR} = \sqrt{\frac{p_T}{n_T} + \frac{p_C}{n_C}}$$

The confidence intervals can be computed from the standard error,

$$CI_{RR} = RR \pm 1.96 \times SE_{RR}$$

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

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

pT <- 10/(43+10)
pC <- 24/(41+24)
nT <- 53
nC <- 65

RR <- pT / pC

SE <- sqrt(pT/nT + pC/nC)

RRlow <- RR - 1.96*SE
RRhigh <- RR + 1.96*SE

# print pearson correlation and confidence intervals
data.frame(RR = MOTE::apa(RR),
           RRlow = MOTE::apa(RRlow),
           RRhigh = MOTE::apa(RRhigh))
```

```
     RR RRlow RRhigh
1 0.511 0.323  0.699
```

11 Effect Sizes for ANOVAs

ANOVA (Analysis of Variance) is a statistical method used to compare means across multiple groups. 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.

11.1 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 particular 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}),

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

Where $\bar{y}$ is the grand mean (i.e., the mean of all data points collapsed across groups). To calculate the sum of squares of the effect, we can take the predicted y values ($\hat{y}_i$). In the case of categorical predictors, $\hat{y}_i$ 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:

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

Now we can calculate the eta-squared value,

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

The standard error of eta-square can be approximated from Olkin and Finn (1995):

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

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. 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. The R code below will compute the proper confidence interval.

Where n is the total sample size and k is the number of predictors. In R, we can calculate η^2 from a one-way ANOVA using the penguin data set from the `palmerpenguins` data package. 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

library(palmerpenguins)
library(effectsize)

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

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)

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 [.62, .72] and for sex η^2 = .17 [.10, .24].

11.2 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:

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

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 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.

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


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

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 [.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)

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 eta-squared value for species begins to differ. The example shows a partial eta-squared value for species of η_p^2 = .81 [.78, .84] and for sex η^2 = .53 [.46, .59].

11.3 Generalized Eta-Squared (η_G^2)

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.

11.4 Omega squared corrections (ω^2, ω_p^2)

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:

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

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

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

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

$$MS_{\text{effect}} = \frac{SS_{\text{effect}}}{n}$$

$$MS_{\text{error}} = \frac{SS_{\text{error}}}{n}$$

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

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

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

library(palmerpenguins)
```

```
# 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 [.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 eta-squared value for species diverge. The example shows a partial eta-squared value for species of ω_p^2 = .81 [.78, .84] and for sex ω^2 = .53 [.46, .58].

11.4.1 Cohen's f

Cohen's f is defined as the ratio of the standard deviations of the group means and the common standard deviation within each of the groups (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,

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

or by the ω^2 value,

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

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. Note that Cohen's f is also 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 $f_{UP} = \sqrt{\frac{\eta^2_{UP}}{1-\eta^2_{UP}}}$).

In R, we can use the `cohens_f` function in the `effectsize` package to calculate Cohen's f. We will again use example data from the `palmerpenguins` package.

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

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

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]
```

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 [1.27, 1.57].

11.5 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 Cohen's d or Hedges' g.

Imagine 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 IV on each level of the other IV. You will report d or g 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 IV while collapsing/aggregating the levels of the other IV. You will report d or g for these pairwise comparisons.

Note that lower-order effects are not directly interpretable if higher-order effects are significant. If you have a significant interaction in a two-way ANOVA, you cannot interpret the main effects directly. If you have a significant 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

```r
# 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 = MOTE::apa(c(omega_values$Omega2_partial,NA)),
                     omega_low = MOTE::apa(c(omega_values$CI_low,NA)),
                     omega_high = MOTE::apa(c(omega_values$CI_high,NA)))

# 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 omega_low
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
            omega_high
species          0.838
sex              0.585
Residuals           NA
```

12 Differences in Variance

Occasionally researchers would like to compare the variance between two conditions or groups rather than the mean.

12.1 Variance Ratio (VR)

The variance ratio can be calculated by taking the variance within one group and dividing it by the variance in another group,

$$VR = \frac{S_1^2}{S_2^2}$$

A VR of 1 would indicate no difference between the variances, a VR of >1 would indicate larger variance in group 1, and VR of <1 would indicate larger variance in group 2. To obtain confidence intervals of a variance ratio, we need to first take the log-transform of the variance ratio and calculate the standard error of the log-variance ratio (LVR),

$$LVR = \text{log}(VR)$$

$$SE_{LVR} = \sqrt{\frac{1}{2(n_1 - 1)} + \frac{1}{2(n_2 - 1)}}$$

The 95% confidence intervals for the variance ratio can then be calculated by back-transforming the CI of the log-variance ratio,

$$CI_{VR} = \text{exp}(LVR \pm 1.96 \times SE_{LVR})$$

In R, we can simply use base R to calculate the VR and it's confidence intervals

```
# Example:
# Group 1: standard deviation = 4.5, sample size = 50
# Group 2: standard deviation = 3.5, sample size = 50
```

```
SD1 <- 4.5
SD2 <- 3.5
n1 <- n2 <- 50

VR <- SD1^2 / SD2^2

# calculate confidence interval
LVR <- log(VR)
SE_LVR <- sqrt(1/(2*(n1-1)) + 1/(2*(n2-1)))

VRlow <- exp(LVR - SE_LVR)
VRhigh <- exp(LVR + SE_LVR)

# print the VR value and confidence intervals
data.frame(VR = MOTE::apa(VR),
           VRlow = MOTE::apa(VRlow),
           VRhigh = MOTE::apa(VRhigh))
```

```
     VR VRlow VRhigh
1 1.653 1.433  1.907
```

From the example, we obtain a variance ratio of VR = 1.65 [1.43, 1.91].

13 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 than 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)).

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., Liddel & 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, 1978).

We will briefly discuss here two groups of tests that can be applied to the independent and paired samples then present 3 effect sizes that can accompany these tests as well as their calculations and examples in R.

13.1 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).

```
# Paired samples ----

data(sleep)

# wilcoxon test
wilcox.test(extra ~ group,
```

```r
            data = sleep,
            paired = TRUE)
```

```

    Wilcoxon signed rank test with continuity correction

data:  extra by group
V = 0, p-value = 0.009091
alternative hypothesis: true location shift is not equal to 0
```

```r
# Two Sample ------
# data import 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
wilcox.test(rank(math_relates_to_my_life) ~ gender,
            data = df_mass)
```

```

    Wilcoxon rank sum test with continuity correction

data:  rank(math_relates_to_my_life) by gender
W = 23, p-value = 0.1104
alternative hypothesis: true location shift is not equal to 0
```

13.2 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.

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

These tests are relatively new so there are very few packages offer Brunner-Munzel. 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. This concept can be further elucidated through a bubble plot, where placement above, below, or on the diagonal indicates the dominance of one group’s observation over the other. 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. The `TOSTER` R package can also provide coverage (Läkens 2017; Caldwell 2022).

```r
# Install package for data cleaning
# install.packages('janitor')
library(janitor)
```

```
Attaching package: 'janitor'


The following objects are masked from 'package:stats':

    chisq.test, fisher.test
```

```r
# Paired samples
library(TOSTER)
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.

```r
# When sample sizes are small
# a permutation version should be used.
# When this is done a seed should be set.
set.seed(2124)
brunner_munzel(extra ~ group,
               data = sleep,
               paired = TRUE,
               perm = TRUE)
```

```

    Paired Brunner-Munzel permutation test

data:  extra by group
t = -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
```

```r
# Two Sample
# data import from likert
data(mass, package = "likert")
df_mass = mass |>
  as.data.frame() |>
  clean_names()

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

```
    two-sample Brunner-Munzel permutation test

data:  rank(math_relates_to_my_life) by gender
t = -2.1665, df = 17.953, p-value = 0.0642
alternative hypothesis: true relative effect is not equal to 0.5
95 percent confidence interval:
 0.04761905 0.54961243
sample estimates:
p(X<Y) + .5*P(X=Y)
         0.2738095
```

13.3 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.

13.3.1 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.

Paired Samples Calculation

1. Calculate difference scores between pairs:

$$D = X_2 - X_1$$

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 rank positive rank sum is greater than or equal to the negative rank sum ($R_{\oplus} \geq R_{\ominus}$) or we can set H to 1 when the rank positive rank sum is less than the negative rank sum ($R_{\oplus} < R_{\ominus}$).

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

4. Calculate rank-biserial correlation:

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

5. For paired samples, or one sample, the standard error is calculated as the following:

$$SE_{r_{rb}} = \sqrt{\frac{(2 \cdot nd^3 + 3 \cdot nd^2 + nd)/6}{(nd^2 + nd)/2}}$$

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

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

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

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

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

$$CI_{rb} = \text{tanh}(Z_{rb} \pm 1.96 \cdot SE_{Z_{rb}})$$

Two Sample Calculation

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

$$R = \text{rank}(X)$$

2. 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 rank biserial correlation

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

4. For independent samples, the standard error is calculated as the following:

$$SE_{rb} = \sqrt{\frac{n_1 + n_2 + 1}{3n_1 n_2}}$$

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

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

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

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

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

$$CI_{rb} = \text{tanh}(Z_{rb} \pm 1.96 \cdot SE_{Z_{rb}})$$

Calculation in R

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

```
# Paired samples

data(sleep)
library(TOSTER)

# When sample sizes are small
# a permutation version should be used.
# When this is done a seed should be set.
set.seed(2124)
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
```

```
# Two Sample
# data import from likert
data(mass, package = "likert")
df_mass = mass |>
  as.data.frame() |>
  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
```

13.3.2 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. It is also referred to as the c-index, the non-parametric probability of superiority, or the non-parametric common language effect size (CLES).

Calculation

The calculation of concordance can be derived from the rank-biserial correlation. The concordance probability (p_c) can be converted from the correlation.

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

Calculation in R

```
# Paired samples

data(sleep)

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
```

```
# Two Sample
# data import from likert
data(mass, package = "likert")
df_mass = mass |>
  as.data.frame() |>
  janitor::clean_names()
```

```
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
```

13.3.3 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.

Calculation

The odds can be converted from the concordance by taking the logit of the concordance. This will provide the log odds. The exponential value of the log-odds will provide the odds on a more interpretable scale.

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

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

Calculation in R

```
# Paired samples ----

data(sleep)

TOSTER::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
```

```r
# Two Sample ------
# data import from likert
data(mass, package = "likert")
df_mass = mass |>
  as.data.frame() |>
  janitor::clean_names()

TOSTER::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
```

Part II

Converting Between Effect Sizes

14 Converting to Cohen's d

14.1 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_1 and n_2) as well as the t-statistic from an independent sample t-test and plug it into the following formula:

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

Using the `t_to_d` function in the `effectsize` package we can convert t to d_p.

```r
# Example:
# unpaired t-statistic = 3.25
# n1 = 50, n2 = 40

library(effectsize)

t <- 3.25
n1 <- 50
n2 <- 40

t_to_d(t, df_error = n1+n2-2, paired = FALSE)
```

```
d    |       95% CI
-------------------
0.69 | [0.26, 1.12]
```

14.2 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 in each group (n_1 and n_2) as well as the t-statistic from an paired sample t-test and plug it into the following formula:

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

Using the `t_to_d` function in the `effectsize` package we can convert t to d_z.

```r
# Example:
# paired t-statistic = 3.25
# n = 50

t <- 3.25
n <- 50

t_to_d(t, df_error = n-1, paired = TRUE)
```

```
d    |       95% CI
-------------------
0.46 | [0.17, 0.76]
```

14.3 From Pearson Correlation

If a Pearson correlation is calculated between a continuous score and a dichotomous score, this is considered a point-biserial correlation. The point-biserial correlation can be converted into a d_p value using the following formula:

$$d_p = \frac{r}{\sqrt{1-r^2}}\sqrt{\frac{n_1+n_2-2}{n_1}+\frac{n_1+n_2-2}{n_2}}$$

Or if sample sizes within each group are unknown (or equal), the equatio simplifies to be approximately,

$$d_p \approx \frac{r\sqrt{4}}{\sqrt{1-r^2}}$$

Using the `r_to_d` function in the `effectsize` package we can convert r to d_p.

```
# Example:
# r = 3.25
# n1 = 50, n2 = 40

r <- .50
n1 <- 50
n2 <- 40

r_to_d(r = r, n1 = n1, n2 = n2)
```

```
[1] 1.148913
```

14.4 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:

$$d_p = \frac{\log(OR)\sqrt{3}}{\pi}$$

Using the `oddsratio_to_d` function in the `effectsize` package we can convert OR to d_p.

```
# Example:
# OR = 1.62

OR <- 1.46

oddsratio_to_d(OR = OR)
```

```
[1] 0.2086429
```

15 Converting to Pearson Correlation

15.1 From t-statistic

From a t statistic calculated from a correlational test, we can calculate the correlation coefficient using the following formula:

$$r = \sqrt{\frac{t^2}{t^2 + n - 2}}$$

Using the `t_to_r` function in the `effectsize` package we can convert t to r.

```
# Example:
# t = 4.14, n = 50

library(effectsize)

t <- 4.14
n <- 50

t_to_r(t = t, df = n-2)
```

```
r    |       95% CI
-------------------
0.51 | [0.28, 0.67]
```

15.2 From Cohen's d

From a between groups Cohen's d value (d_p), we can calculate the correlation coefficient from the following formula:

$$r = \frac{d_p}{\sqrt{d_p^2 + \frac{n_1+n_2-2}{n_1} + \frac{n_1+n_2-2}{n_2}}}$$

Using the `d_to_r` function in the `effectsize` package we can convert d_p to r.

```r
# Example:
# d = 0.60, n1 = 50, n2 = 70

d <- 0.60
n1 <- 50
n2 <- 70

d_to_r(d = d, n1 = n1, n2 = n2)
```

```
[1] 0.2858532
```

15.3 From Odds-Ratio

The correlation coefficient from an odds ratio can be calculated with the following formula:

$$r = \frac{\log(OR) \times \sqrt{3}}{\pi\sqrt{\frac{3\log(OR)^2}{\pi^2} + \frac{n_1+n_2-2}{n_1} + \frac{n_1+n_2-2}{n_2}}}$$

Using the `oddsratio_to_r` function in the `effectsize` package we can convert OR to r.

```r
# Example:
# OR = 2.21, n1 = 50, n2 = 70

OR <- 2.21
n1 <- 50
n2 <- 70

oddsratio_to_r(OR=OR, n1 = n1, n2 = n2)
```

```
[1] 0.2124017
```

16 Converting to Odds Ratio

16.1 From Cohen's d

We can calculate an odds-ratio from a between groups cohen's d (d_p):

$$OR = \exp\left(\frac{d_p \pi}{\sqrt{3}}\right)$$

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.

```r
# Example:
# d = 0.60, n1 = 50, n2 = 70

library(effectsize)

d <- 0.60
n1 <- 50
n2 <- 70

d_to_oddsratio(d = d, n1 = n1, n2 = n2)
```

```
[1] 2.969162
```

16.2 From a Pearson Correlation

We can calculate an odds ratio from a Pearson correlation using the following formula:

$$OR = \exp\left(\frac{r\pi\sqrt{\frac{n_1+n_2-2}{n_1} + \frac{n_1+n_2-2}{n_2}}}{\sqrt{3(1-r^2)}}\right)$$

When sample sizes are equal, this equation can be simplified to be approximately,

$$OR = \exp\left(\frac{r\pi\sqrt{4}}{\sqrt{3(1-r^2)}}\right)$$

Using the `r_to_oddsratio` function in the `effectsize` package we can convert d to OR.

```r
# Example:
# r = .50, n1 = 50, n2 = 70

r <- .40
n1 <- 50
n2 <- 70

r_to_oddsratio(r = r, n1 = n1, n2 = n2)
```

```
[1] 4.870584
```

Part III

Conclusion

17 Conclusion

17.1 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.

17.2 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.

APA. 2010. *Publication Manual of the American Psychological Association*. American Psychological Association. https://thuvienso.hoasen.edu.vn/handle/123456789/8327.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

Peters, Gjalt-Jorn Ygram, and Stefan Gruijters. 2023. *Ufs: A Collection of Utilities*. https://ufs.opens.science.

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.

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.

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.

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.

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 Personality Research*. Evanston, Illinois: Northwestern University. https://CRAN.R-project.org/package=psych.