Principles and Implementation of Effect Size Confidence Intervals

Authors: Wang Jun, Song Qiongya, Xu Yuepei, Binbin Jia, Hu Chuanpeng, Hu Chuanpeng

Date: 2019-04-15T00:00:00+00:00

Abstract

In the context of the reproducibility crisis in psychology, reporting effect sizes and their confidence intervals is gradually becoming a new standard required by the mainstream psychology community; however, researchers may lack adequate understanding of effect size confidence intervals. To enhance researchers' understanding and application of effect size confidence intervals, this article introduces the fundamental principles underlying the confidence intervals for the most commonly used effect size metrics in psychological research—Cohen' s d and ^{2}—namely, when the alternative hypothesis (H1) is true, the noncentrality parameters of the corresponding noncentral distributions must be estimated through iterative estimation to construct confidence intervals for Cohen' s d and ^{2}. Specifically, Cohen' s d corresponds to the noncentral t-distribution, whereas ^{2} corresponds to the noncentral F-distribution. Using existing computer programs, confidence intervals for Cohen' s d and ^{2} can be calculated; for example, with R and JASP, as demonstrated separately in this article. Reporting effect size confidence intervals not only facilitates better statistical inference for researchers but also promotes knowledge accumulation across the scientific community; therefore, the methods introduced in this article are of significant importance to researchers.

Full Text

Calculating Confidence Intervals of Cohen' s d and Eta-squared: A Practical Primer

WANG Jun[1], SONG Qiongya[1], XU Yuepei[2], JIA Binbin[3], HU Chuan-Peng ,

[1]Department of Psychology, Sun Yat-Sen University, Guangzhou, 510006, China
[2]College of Education, Shanghai Normal University, Shanghai, 200234, China

[3]Shanghai University of Sport, Shanghai, 200438, China
Neuroimaging Center, Focus Program Translational Neuroscience (FTN), Johannes Gutenberg University Medical Centre Mainz, 55131 Mainz, Germany
Deutsches Resilienz Zentrum (DRZ), University Medical Centre of the Johannes Gutenberg University, 55131 Mainz, Germany

Abstract

Amid psychology' s replication crisis, reporting effect sizes (ES) and their confidence intervals (CIs) is becoming the new standard required by mainstream psychological journals. However, researchers may lack adequate understanding of effect size confidence intervals. To enhance comprehension and application of these methods, this paper introduces the fundamental principles underlying confidence intervals for the most commonly used effect size metrics in psychological research—Cohen' s d and ^{2}. Both require iterative estimation of non-centrality parameters from non-central distributions when the alternative hypothesis (H) is true. Specifically, Cohen' s d corresponds to the non-central t-distribution, while ^{2} corresponds to the non-central F-distribution. We demonstrate how existing software programs, such as R and JASP, can compute confidence intervals for Cohen' s d and ^{2}. Reporting effect size confidence intervals not only facilitates better statistical inference but also promotes cumulative knowledge building across the scientific community, making the methods introduced here highly significant for researchers.

Keywords: effect size; confidence interval; Cohen' s d; Eta squared; R

1. Introduction

Statistical inference is essential for researchers to derive logical conclusions from data and test research hypotheses. Null hypothesis significance testing (NHST) represents the most widely used statistical inference method in psychological research (Cumming et al., 2007). However, its reliance on $p < 0.05$ as the criterion for statistical significance has indirectly contributed to excessively high false-positive rates in psychology, and p-values are heavily influenced by sampling characteristics, making them unsuitable for comparing results across replication studies or different experiments (胡传鹏, 王非, 过继成思, 宋梦迪, 隋洁, 彭凯平, 2016). In recent years, widespread concern about research reproducibility in psychology has renewed scholarly attention to the limitations of NHST (Kline, 2004; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). To address these shortcomings, new statistical approaches have been gradually introduced into psychological research, including estimation-based statistics (Cumming, 2012, 2014), Bayesian factors (胡传鹏, 孔祥祯, Wagenmakers, Ly, 彭凯平, 2018; Wagenmakers et al., 2018), and likelihood methods (Etz, 2018). Among these, estimation-based statistics have been recommended by researchers both domestically and internationally due to their intuitive interpretability and ability to compensate for NHST' s deficiencies. The emphasis on effect sizes and their con-

fidence intervals (CIs) is gradually becoming a mandatory reporting standard in major international and domestic psychology journals (APA Publications Communications Board Working Group on Journal Article Reporting Standards, 2008; Appelbaum, Cooper, Kline, Mayo-Wilson, Nezu, & Rao, 2018; Cumming, 2014).

Nevertheless, compared to NHST, which has dominated psychology for decades, the use of effect sizes and confidence intervals remains limited, with very few studies reporting effect size confidence intervals (Fritz, Morris, & Richler, 2012). Although Chinese researchers have introduced the concept of effect size extensively (胡竹菁, 2010; 卢谢峰, 唐源鸿, 曾凡梅, 2011; 郑昊敏, 温忠麟, 吴艳, 2011), they rarely discuss confidence intervals for effect sizes.

Notably, both psychology students and professionals still harbor misconceptions about confidence intervals (胡传鹏等, 2016; Hoekstra, Morey, Rouder, & Wagenmakers, 2014). For instance, 胡传鹏等人 (2016) surveyed Chinese researchers' understanding of CIs by presenting a hypothetical study with a 95% confidence interval of [0.1, 0.4] for an effect. Respondents evaluated six statements: (A) The probability that the true mean is greater than 0 is at least 95%; (B) The probability that the true mean equals 0 is less than 5%; (C) The null hypothesis that the true mean equals 0 is likely incorrect; (D) There is a 95% probability that the true mean lies between 0.1 and 0.4; (E) We are 95% confident that the true mean lies between 0.1 and 0.4; and (F) If we repeated the experiment, the true mean would fall between 0.1 and 0.4 in 95% of cases. All six statements represent common misinterpretations of confidence intervals (Hoekstra et al., 2014), yet most respondents judged at least some of them to be correct (see Figure 1 [Figure 1: see original paper], data from Lyu, Peng, & Hu, 2018). The correct interpretation is that if we repeatedly conducted the experiment and calculated confidence intervals each time, approximately 95% of these intervals would contain the true mean. Therefore, the observed interval [0.1, 0.4] is just one of many possible intervals, and whether it includes the true value remains unknown (Cumming, 2014).

To deepen researchers' understanding of effect sizes and their confidence intervals and to facilitate accurate calculation and reporting, this paper first introduces the principles of effect size confidence intervals and their advantages. We then use two common effect sizes—Cohen' s d and Eta squared (²)—as examples to explain the principles underlying their confidence intervals and demonstrate implementation in open-source software such as R and JASP. It is important to note that the effect sizes discussed here are not limited to standardized metrics like Cohen' s d. According to Cumming (2014), an effect size refers to the magnitude of any effect of interest to researchers, which can be either standardized or unstandardized with original units. Moreover, standardized effect sizes are not inherently superior to unstandardized ones; researchers should select metrics that appropriately reflect the information in their data, as unstandardized effect sizes are sometimes more interpretable.

2. Advantages of Reporting Effect Sizes and Their Confidence Intervals

Compared to p-values from NHST, reporting effect sizes and their confidence intervals provides more detailed and multifaceted information about results. Specifically, this approach offers several key advantages.

First, it enables comparison of error magnitudes across different experiments. Suppose a researcher conducts three experiments with effect sizes and confidence intervals as shown in Figure 2 [Figure 2: see original paper]. Using traditional NHST, one would conclude that Experiments 1 and 3 show significant results ($p < 0.05$), with both group means differing significantly from zero, while Experiment 2 yields $p > 0.05$, indicating no significant difference from zero. In this scenario, the conclusions drawn from Experiments 1 and 3 would appear nearly identical. However, NHST cannot answer crucial questions: How large are the differences between group means? What is the sampling error? Which experiment provides the most reliable evidence for the hypothesis?

In conventional reporting practices, researchers often supplement NHST with unstandardized point estimates (e.g., means) and standard errors to address these limitations. Reporting effect sizes (in this case, mean differences) and their confidence intervals achieves the same goal. Figure 2 reveals that although both Experiments 1 and 3 are statistically significant, Experiment 1 shows a smaller effect size with less variability, whereas Experiment 3 demonstrates a larger effect size with greater variability. By analyzing effect sizes and confidence intervals, researchers can draw more nuanced conclusions about Experiments 1 and 3.

Second, effect sizes and confidence intervals help researchers reach correct conclusions. When evaluating results based solely on effect sizes and confidence intervals, most researchers can draw logically sound conclusions when comparing different studies. However, the proportion of correct conclusions decreases when relying only on NHST and effect sizes (Coulson, Healey, Fidler, & Cumming, 2010; Lyu et al., 2018). Unlike NHST' s dichotomous thinking, reporting effect sizes and confidence intervals encourages an "estimation" and "quantitative" orientation (Cumming & Fidler, 2009). This mindset predisposes researchers to ask more quantitative questions. Using Figure 2 again as an example, although Experiment 2' s result is not statistically significant, its effect size and confidence interval show the same trend as Experiments 1 and 3. This observation prompts deeper reflection: Could excessive "noise" in Experiment 2' s data have caused the non-significant result?

Third, this approach reveals richer information about studies. In Figure 2, Experiment 1' s effect size is actually quite small, suggesting minimal practical difference between the two groups. However, perhaps due to small sampling error and large sample size, Experiment 1' s confidence interval is very narrow, allowing researchers to conclude significant differences with high confidence. This exemplifies the disconnect between statistical significance and practical signifi-

cance. Conversely, for Experiment 2, although its confidence interval includes zero, its point estimate for the effect size is the highest among the three experiments, indicating that excessive "noise" in the data led to large variability and a wide confidence interval. Experiment 3' s results are more ideal, with both effect size and confidence interval at reasonable levels.

Finally, because effect sizes are not sample-dependent (卢谢峰等, 2011), they are more suitable for cross-experiment syntheses and meta-analyses than sample-dependent p-values. From a frequentist perspective, any individual study can be viewed as an independent sample providing one estimate of population parameters, making single studies potentially limited in scope. However, by accumulating data across multiple studies, researchers can conduct meta-analyses to estimate population parameters more precisely. Meta-analysis not only increases sample size and statistical power but also narrows confidence intervals, yielding more accurate estimates of population effect sizes (Cumming, 2012). Compared to p-values, effect sizes and confidence intervals facilitate meta-analytic statistics, and the process of quantitatively reporting them inherently embodies meta-analytic thinking.

These advantages have led to widespread recommendations for reporting effect sizes and confidence intervals. The American Psychological Association (APA) Publication Manual (6th edition) recommends reporting effect sizes and their confidence intervals, and the 2018 Journal Article Reporting Standards introduced in *American Psychologist* also endorse this practice (Appelbaum et al., 2018).

In summary, although reporting effect sizes and confidence intervals has gained broad support in current research, the actual application of effect size confidence intervals remains limited (Fritz et al., 2012). A primary reason may be that researchers know little about effect size confidence intervals and lack appropriate tools for implementation (for example, SPSS, a commonly used statistical software in psychology, does not output confidence intervals for common effect size metrics). To address this issue, we next use Cohen' s d and Eta squared (^{2}) as examples to explain the principles and calculation formulas for their confidence intervals and demonstrate how to compute them using open-source software.

3. Standardized Mean Difference (Cohen' s d)

Cohen originally defined d using the population standard deviation as the standardization unit. However, since population standard deviations are typically unknown in practice, the more common approach uses the sample standard deviation (hereafter described using sample standard deviation s). Cohen' s d represents the ratio of the difference between the sample mean and the null hypothesis (H) mean to the standard deviation:

$$d = \frac{M - \mu_0}{s} \tag{3.1}$$

where s denotes the sample standard deviation and μ_0 represents the reference value against which d is measured. Cohen' s d can thus be understood as how many standard deviations s the sample mean M differs from the reference value μ_0. Depending on the research purpose, various formulas exist for calculating Cohen' s d; for details, see Cumming (2014), Hedges (1981), and Lakens (2013).

3.1 Principles of Cohen' s d Confidence Intervals

Understanding Cohen' s d confidence intervals requires first comprehending the distribution of t-values under two scenarios: when the null hypothesis (H) is true (i.e., no effect) and when the alternative hypothesis (H) is true. Suppose we randomly draw countless samples of size N from a normal distribution (μ_0, σ). For any given sample, we can calculate its mean M and standard deviation s. To test whether this sample belongs to the standard normal population, we can conduct a one-sample t-test based on the null hypothesis $H_0 : \mu = \mu_0$, calculating the t-value using:

$$t = \frac{M - \mu_0}{s/\sqrt{N}} \tag{3.2}$$

When the null hypothesis is true, if we repeatedly draw samples of size N and conduct t-tests, these t-values will form a t-distribution with $df = N-1$ degrees of freedom. This t-distribution is centered at 0 and symmetric. In this case, we can also view the t-statistic as the distance between M and μ_0 measured in units of $s/\sqrt{N}$ (standard error). For each sample, we can use the t-distribution table to calculate p-values and conduct hypothesis tests.

However, if the null hypothesis (H) is false, then the alternative hypothesis (H) is true, meaning $\mu \neq \mu_0$. In this situation, we are actually sampling from a population with mean μ_1, so the sample means M calculated from countless draws will be closer to μ_1 than to μ_0. If we still use the above formula for the t-test, the t-values calculated from repeated sampling will no longer follow a t-distribution centered at 0 with two-sided symmetry, but rather a skewed non-central t-distribution whose center is not at zero. For such a non-central t-distribution, the parameters include not only degrees of freedom (df) but also a non-centrality parameter Δ (read: delta). Δ can be viewed as the distance between μ_1 and μ_0 measured in standard error units. Under otherwise identical conditions, larger Δ values indicate that the center of this non-central t-distribution deviates further from 0 (as shown in Figure 3 [Figure 3: see original paper], where ncp denotes the value of Δ in R software).

Combining equations (3.1) and (3.2), we obtain:

$$d = \frac{t}{\sqrt{N}} \tag{3.3}$$

Equation (3.1) shows that d represents the distance between M and μ_0 measured in units of s (standard deviation), while equation (3.2) shows that t represents the distance between M and μ_0 measured in units of $s/\sqrt{N}$ (standard error). Equation (3.3) demonstrates that Cohen' s d has a one-to-one correspondence with the t-value. Therefore, the sampling distribution of Cohen' s d is also a non-central t-distribution, which must be used when calculating confidence intervals for Cohen' s d.

Since the t-value follows a non-central t-distribution when the alternative hypothesis (H) is true, d also follows a non-central t-distribution in this case. This means that d' s confidence interval is asymmetric, with unequal distances from the center to the upper and lower bounds. Consequently, we need to use iterative approximations to construct d' s confidence interval. We can explain this in detail using the following figure.

Suppose we have a population effect of Cohen' s d = 1.21 and need to construct its 95% confidence interval (as shown in Figure 4 [Figure 4: see original paper]). This means that if we construct such intervals countless times, approximately 95% of them will contain the true value of 1.21. When centered at the lower bound d_L, the sampling distribution of d rejects d_L in favor of the true value with a probability of 2.5% (dark gray area). Similarly, when centered at the upper bound d_U, the sampling distribution rejects d_U in favor of the true value with the same 2.5% probability (light gray area). This means that the sum of probabilities of containing the true value for distributions centered at the upper and lower bounds equals exactly 5%. Moving either bound toward the center increases the probability of containing the true value. Likewise, to estimate a 99% confidence interval, the upper and lower bounds would be further from the center compared to the 95% interval, with the sum of probabilities for distributions centered at these bounds equal to 1% (0.005 in each tail).

Exploratory Software for Confidence Intervals (ESCI), developed by Geoff Cumming, consists of a series of Excel files that can perform complex statistical calculations using familiar Microsoft Excel software, including calculating Cohen' s d and its confidence intervals (Cumming, 2001). Using ESCI provides more intuitive understanding of the relationship between interval bounds and d values. In ESCI, moving the distribution centered at the lower bound d_L leftward decreases d_L and reduces the region to the right of the true value, meaning the p-value corresponding to the true value decreases and the probability of rejecting d_L in favor of the true value becomes smaller. Conversely, moving the distribution centered at d_L rightward increases d_L and expands the region to the right of the true value, increasing the probability of rejecting d_L in favor of the true value. To obtain an accurate 95% confidence interval, we need to shift the distribution centered at d_L such that the region to its right exceeding the true value is 0.025, while simultaneously shifting the distribution centered at d_U such that the region to its left exceeding the true value is also 0.025. The resulting d_L and d_U constitute the lower and upper bounds of our confidence interval.

Since both curves are non-central t-distributions, we can adjust them by changing d values to slide them left or right. This continuous adjustment to achieve the desired interval is called iterative estimation. In essence, while keeping degrees of freedom constant, we substitute different non-centrality parameters (denoted as Δ or sometimes in some studies) for calculations and make further adjustments. When calculating confidence intervals, we continuously adjust Δ to shift the non-central t-distribution until the critical values on the curve fall exactly within the two-tailed range of 0.025 and 0.975, yielding the confidence interval for Cohen' s d. So how do we determine the non-centrality parameters for distributions centered at the upper and lower bounds of the confidence interval?

For single-sample studies, the non-centrality parameter Δ is calculated as:

$$\Delta = \frac{\mu_1 - \mu_0}{\sigma/\sqrt{N}}$$

Combining this with equation (3.1), we obtain:

$$\Delta = d \times \sqrt{N} \tag{3.4}$$

ESCI uses equation (3.4) to convert between Cohen' s d and the non-centrality parameter Δ, which can then be used to calculate the non-central t-distribution. Therefore, we can derive the confidence interval for Cohen' s d as:

$$d_L = \frac{\Delta_L}{\sqrt{N}}, \quad d_U = \frac{\Delta_U}{\sqrt{N}}$$

Similarly, for two-sample studies, the non-centrality parameter Δ is calculated as:

$$\Delta = \frac{\mu_1 - \mu_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

We can then calculate the effect size and confidence interval for two-sample studies:

$$d = \frac{M_1 - M_2}{s_p} \tag{3.5}$$

where s_p is the pooled standard deviation:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{3.6}$$

The relationship between d and the non-centrality parameter becomes:

$$\Delta = d \times \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \tag{3.7}$$

Thus, the confidence interval for Cohen' s d in two-sample studies is:

$$d_L = \Delta_L \times \sqrt{\frac{n_1 + n_2}{n_1 n_2}}, \quad d_U = \Delta_U \times \sqrt{\frac{n_1 + n_2}{n_1 n_2}} \tag{3.8}$$

For more details on the principles of Cohen' s d confidence intervals, refer to Cumming (2012), Chapter 11.

3.2. Examples and Software Implementation

In research practice, researchers do not need to perform iterative estimation manually. Currently, numerous mature packages in R (R Core Team, 2018) can calculate confidence intervals for Cohen' s d. JASP, a user-friendly software developed based on R for traditional statistical and Bayesian analyses (Wagenmakers et al., 2015; 胡传鹏等, 2018), can also compute Cohen' s d confidence intervals. (For SPSS plugins to calculate Cohen' s d confidence intervals, see: http://dl.dropbox.com/u/1857674/CIstuff/CI.html; for ESCI developed in Microsoft Excel to calculate Cohen' s d confidence intervals, see: https://thenewstatistics.com/itns/esci.)

We will demonstrate using JASP' s example dataset "Kitchen Rolls" (available at: https://osf.io/q9387/). Topolinski and Sparenberg (2012) found that the direction of rotating a paper roll could change individuals' scores on personality measures of openness. Wagenmakers et al. (2015) conducted a replication study, and the data used here come from that replication. This example dataset includes openness scores on a personality scale for two groups of participants, where one group rotated a kitchen roll clockwise while completing the questionnaire and the other group rotated it counterclockwise. In the data analysis, the average NEO PI-R score serves as the dependent variable, participant group (clockwise vs. counterclockwise) as the independent variable, and an independent samples t-test is used for data analysis.

3.2.1. Calculating Cohen' s d Confidence Intervals Using JASP After opening the sample data in JASP, select T-Tests → Independent Samples T-Test to access the interface. Import the required variables into the corresponding boxes (similar to SPSS), and select the desired statistical options in the lower interface. Under Additional Statistics, you can check the Effect Size and Confidence interval options. The results calculated according to formulas (3.5)-(3.8) provide the effect size Cohen' s d and its confidence interval.

Figure 5 [Figure 5: see original paper] shows the JASP independent samples t-test interface (left) and results (right). The results indicate that the dependent

variable meets assumptions of normality and homogeneity of variance, so the Student t-test is selected for analysis. The results show no significant difference in average NEO PI-R scores between the two groups ($t(100) = 0.754$, $p = 0.453$), with Cohen' s d = 0.149, 95% CIs [-0.240, 0.538].

3.2.2. Calculating Cohen' s d Confidence Intervals Using R Multiple packages in R can perform independent samples t-tests, such as car and MBESS. If we use the t.test function from the car package, we find no significant difference in average NEO PI-R scores between the two groups, $t(100) = 0.754$, $p = 0.453$ (of course, t-values and p-values can also be obtained using JASP or SPSS). After obtaining the t-value, we can calculate Cohen' s d confidence interval using the following commands:

```
library("MBESS")  # Load MBESS package
# Define relevant parameters and calculate 95% CI for Cohen's d
MBESS::ci.smd(ncp = 0.75361, n.1 = 48, n.2 = 54, conf.level = 0.95)
```

Here, ncp (non-centrality parameter) is the t-value, and n.1 and n.2 represent the sample sizes of the two groups. MBESS uses formulas (3.5)-(3.8) to obtain the results.

3.3. Results Reporting and Interpretation

As demonstrated above, using two different software packages to estimate the difference in personality scale scores between participants who rotated the roll clockwise versus counterclockwise yielded identical 95% confidence intervals. Both outputs indicate no significant difference in average NEO PI-R scores between groups, with identical estimates for the effect size and its 95% confidence interval—effect size d = 0.149, 95% CI [-0.240, 0.538]. Based on these results, we can conclude: The current data fail to reject the null hypothesis, meaning we cannot infer that clockwise versus counterclockwise rotation significantly affects NEO PI-R scores. (Note that both $p > 0.05$ and the fact that Cohen' s d' s confidence interval includes zero do not support the conclusion that the null hypothesis is true—that is, we cannot use p-values to support a conclusion of no difference between groups, because p-value calculations are predicated on the assumption that the null hypothesis is true. To provide evidence for the null hypothesis, other statistical approaches are needed.)

4. Effect Sizes and Confidence Intervals in ANOVA

Another common effect size metric in psychological research is Eta-squared (^{2}) in analysis of variance (ANOVA) (Fritz et al., 2012), first proposed by Pearson (1905). It can be understood as the proportion of total variance explained by one or more factors (including interactions) (Cohen & Cohen, 2010). The formula for ^{2} is:

$$\eta^2 = \frac{SS_{between}}{SS_{total}} \tag{4.1}$$

It is crucial to note that the effect size metric ^{2} (partial eta-squared) output by SPSS is widely used in psychology but has a different meaning from ^{2} and can cause confusion. Research indicates that many researchers easily confuse ^{2} and ^{2}, which can have serious consequences. For example, in meta-analysis, incorrectly using ^{2} instead of ^{2} can introduce substantial bias (Levine & Hullett, 2002). Additionally, misusing ^{2} and ^{2} is detrimental to theory construction (Pierce, Block, & Aguinis, 2004). Therefore, when reporting ^{2}, researchers must clearly specify which metric is being reported. (When uncertainty exists about whether ^{2} or ^{2} is reported in a paper, one can sum the effect sizes for each factor; if the result equals 1, it is ^{2}; if greater than 1, it is ^{2}.) Furthermore, with small sample sizes (when the ratio of independent variables to sample size is less than 1:10), ^{2} becomes a more recommended effect size metric (卢谢峰等, 2011). Of course, ^{2} is another similar effect size statistic; see Maxwell & Delaney (2004) for details. Below, we focus on explaining the calculation of ^{2} confidence intervals based on formula 4.1.

4.1. Principles of ^{2} Confidence Interval Calculation

Understanding ^{2} confidence intervals also requires comprehension of non-central distributions related to relevant parameters. Here, constructing ^{2} confidence intervals involves the distribution of F-values in ANOVA and another effect size metric, Cohen' s f. Using the simplest one-way between-subjects ANOVA design as an example, total variance can be decomposed into between-group variance and within-group variance:

$$SS_{total} = \sum_{j=1}^{k}\sum_{i=1}^{n}(X_{ij} - \bar{X})^2$$

where X represents observed values, j denotes group levels (with k groups total), and n represents the number of subjects per group (with n subjects in each group). The F-value is calculated as:

$$F = \frac{MS_{between}}{MS_{error}} = \frac{SS_{between}/df_1}{SS_{error}/df_2}$$

where $df_1 = k - 1$ and $df_2 = nk - df_1 - 1$. The effect size for between-group treatment is:

$$\eta^2 = \frac{SS_{between}}{SS_{between} + SS_{error}} \tag{4.2}$$

Another effect size metric for between-group effects, Cohen's f, can be calculated as:

$$f = \sqrt{\frac{\eta^2}{1-\eta^2}} \tag{4.3}$$

The F-distribution has a very close relationship with the ^{2} distribution. By definition, the ^{2} distribution is the distribution of the sum of squares of samples independently drawn from a standard normal distribution. That is, assuming we have p random variables $\{X_i\}$ drawn from a standard normal distribution (N(0,1)), then:

$$\chi^2 = \sum_{i=1}^{p} X_i^2$$

This follows a central ^{2} distribution with k-1 degrees of freedom (note that "central" here does not mean the distribution is centrally symmetric, but rather that it is the distribution of squared sums drawn from a centrally symmetric distribution).

Referring back to the F-value formula in ANOVA, if we divide both numerator and denominator by $\sigma^2_{between}$ (variance due to treatment) and σ^2_{error} (variance due to error) respectively (under the null hypothesis in ANOVA, we assume treatment variance equals error variance, i.e., $\sigma^2_{between} = \sigma^2_{error}$, so they cancel out), the numerator and denominator of the F-value (F(df , df), hereafter abbreviated as F) each correspond to a ^{2} distribution:

$$F(df_1, df_2) = \frac{(SS_{between}/df_1)/\sigma^2_{between}}{(SS_{error}/df_2)/\sigma^2_{error}}$$

In ANOVA, the null hypothesis states that between-group means are equal and experimental errors follow a normal distribution N(0, σ^2_{error}). In this case, both the numerator and denominator correspond to central ^{2} distributions, and the resulting F-distribution is also central.

When the null hypothesis is false, between-group means are not equal, the numerator corresponds to a non-central ^{2} distribution, while the denominator (experimental error) still corresponds to a central ^{2} distribution. The F-distribution consequently becomes non-central, denoted as F(df , df ,). In fact, the central distribution is a special case of the non-central distribution. The non-centrality parameter ncp determines the specific shape of the distribution, as illustrated by the central F(2, 52, ncp = 0) distribution (black) and the non-central F(2, 52, ncp = 1) distribution (red) shown in Figure 6 [Figure 6: see original paper].

Calculating effect sizes presupposes that H is false (between-group means are not equal), and the corresponding F-distribution is non-central. If we calculate ² confidence intervals based on the non-central F-distribution, we encounter the same problem as with Cohen' s d: the non-centrality parameters differ for the F-distributions at the upper and lower bounds of the confidence interval. Therefore, estimating ² confidence intervals also requires using the inversion confidence interval principle (Steiger & Fouladi, 1996).

We proceed through three stages: statistical test → non-centrality parameter → effect size statistic. First, we need to establish the relationship between the statistical test value (F-value in ANOVA), the non-centrality parameter, and the effect size ². From formula 4.3, we derive $SS_{between} = \eta^2 \times (SS_{between} + SS_{error})$, which leads to:

$$F = \frac{\eta^2/(1-\eta^2) \times df_2}{df_1}$$

When the null hypothesis is false, the estimated non-centrality parameter of the F(df , df) distribution (non-centrality parameters may be denoted by different symbols, commonly or) is calculated as (Smithson, 2001):

$$\delta = \frac{F \times df_1}{df_2} \tag{4.5}$$

Combining formula (4.5), we obtain the non-centrality parameter estimate:

$$\delta = \frac{\eta^2}{1-\eta^2} \times df_2 \tag{4.6}$$

Rearranging gives:

$$\eta^2 = \frac{\delta}{\delta + df_2} \tag{4.7}$$

Thus, we have established the relationship between ² and the F-value and its non-centrality parameter. Next, we can use the confidence interval inversion principle to calculate ²' s confidence interval. Given a sample F(5,194), we need to construct a 100(1-)% (=0.05) two-sided confidence interval (as shown in Figure 7 [Figure 7: see original paper]).

The lower bound corresponds to the /2 point on the right side of F(5,194), and the upper bound corresponds to the /2 point on the left side of F(5,194). After obtaining the non-centrality parameters corresponding to the upper and lower bounds, we can convert them to ²' s confidence interval using:

$$\eta_L^2 = \frac{\delta_L}{\delta_L + df_2} \tag{4.9}$$

$$\eta_U^2 = \frac{\delta_U}{\delta_U + df_2} \tag{4.10}$$

This completes the estimation of ²' s confidence interval.

It is worth noting that for ANOVA effect size confidence intervals, reporting 90% confidence intervals is typically sufficient. The reason is that while mean differences can be positive or negative, ² and R² are squared values and therefore only positive. When calculating 95% confidence intervals, the result may include 0 while the p-value is less than .05, creating a contradiction between the confidence interval and p-value (see Karl Wuensch' s explanation: http://core.ecu.edu/psyc/wuenschk/spss/spss-programs.htm). Moreover, Steiger (2004) notes that 95% and 90% confidence intervals for mean comparisons have equivalent statistical power, and since ² cannot be less than 0, the lower bound of a confidence interval that is not significantly different from 0 (typically not containing 0) should start at 0 (Steiger, 2004).

4.2. Implementing ² and Its Confidence Interval in R

Again, we will use sample data provided by JASP to demonstrate how to calculate ²'s 90% CI in R. The datasets Tooth Growth and Bugs are used to illustrate implementation for between-subjects and within-subjects designs respectively (for SPSS implementation, see: http://core.ecu.edu/psyc/wuenschk/spss/spss-programs.htm).

4.2.1. Between-Subjects Design ² and CI Implementation in R The Tooth Growth data come from a two-way completely randomized design where 60 guinea pigs were randomly assigned to 6 treatment conditions to study how different types of supplements (vitamin C, VC, and orange juice, OJ) at different ascorbic acid doses (0.5mg, 1mg, and 2mg) affect tooth growth in guinea pigs, with tooth length as the dependent variable.

First, obtain the necessary statistics for CI calculation using statistical software. You can use R' s built-in aov function or other packages with statistical capabilities (such as ez, car, etc.). Note that different packages or functions in R use different types of sums of squares: aov defaults to Type I SS, ezANOVA defaults to Type II SS (though you can adjust the type in R; see https://cran.r-project.org/web/packages/ez/ez.pdf), while SPSS defaults to Type III SS (adjustable in SPSS model options). When sample sizes are equal across groups, different SS types yield similar results, but with unbalanced data, careful consideration of SS type is necessary as they produce different statistical results; interested readers may refer to Langsrud (2003). A more convenient approach is to use JASP directly for statistical analysis to obtain the necessary statistics.

For this data, we obtain F(2,54) = 92. Then, in R, load the MBESS package and input the relevant statistics to calculate the CI:

```
library("MBESS")  # Load MBESS package
ci.pvaf(F.value=92, df.1=2, df.2=54, N=60, conf.level=.90)  # Input F-value and df to calcul
```

4.2.2. Within-Subjects Design ^{2} and CI Implementation in R The Bugs data come from a two-way mixed design examining hostility indices toward different types of bug pictures (not scary/not disgusting, not scary/disgusting, scary/not disgusting, and scary/disgusting) across genders (male, female), using a 10-point rating scale indicating desire to kill or drive away the bugs (Ryan, Wilde, & Crist, 2013). Using JASP, we obtain F(2.64, 224.48) (note that for within-subjects designs violating sphericity assumptions, corrected degrees of freedom are used). Then in R, use the following commands:

```
# Load MBESS package
library("MBESS")
# Input F-value and degrees of freedom
Lims <- conf.limits.ncf(F.value=20.14, conf.level=0.90, df.1=2.64, df.2=224.48)
# Calculate lower limit of 90% CI
Lower.lim <- Lims$Lower.Limit / (Lims$Lower.Limit + df.1 + df.2 + 1)
# Calculate upper limit of 90% CI
Upper.lim <- Lims$Upper.Limit / (Lims$Upper.Limit + df.1 + df.2 + 1)
```

4.3. Results Reporting and Interpretation

Interpretation of ^{2} and its confidence interval primarily follows ^{2}' s definition as the proportion of total variance explained by experimental effects. Therefore, ^{2}' s magnitude indicates the effectiveness of manipulating independent variables in a specific experimental study. Larger ^{2} values indicate stronger relationships between relevant variables, though whether this relationship is correlational or causal depends primarily on the experimental design (e.g., quasi-experimental vs. experimental designs). However, since ^{2} confidence intervals cannot be less than 0, their interpretation differs from Cohen' s d confidence intervals—we cannot use inclusion of 0 as a basis for rejecting or accepting the null hypothesis. Moreover, as ANOVA represents a special case under the general linear model, it is often only the first step in examining variable relationships. Therefore, we typically use ^{2} and its confidence interval as an index of experimental manipulation effectiveness, with specific between-group comparisons (e.g., post-hoc tests following significant main effects, simple effects analysis following significant interactions) being the focus of researcher attention, where Cohen' s d from t-tests can again serve as an effect size metric for evaluating the reliability of between-group differences.

5. Summary

In recent years, psychology's replication crisis has profoundly impacted the field, and changes in statistical reporting standards constitute a crucial component of evolving journal article reporting standards (刘宇等, 2018; Appelbaum et al., 2018; Levitt, Bamberg, Creswell, Frost, Josselson, & Suárez-Orozco, 2018). As two of the most commonly used effect size metrics in estimation-based statistics, Cohen' s d and ^{2} hold significant importance for researchers (Fritz et al., 2012). This paper explains the principles underlying confidence intervals for these two effect sizes and uses real data to demonstrate their implementation in R and JASP (all demonstration data and code available at: https://osf.io/4ameb/), which may prove helpful to researchers.

Although this paper does not address another common effect size metric—the confidence interval for correlation coefficients—their calculation and implementation are relatively mature in both JASP and R, and readers can consult relevant resources. For more on confidence interval principles, see Smithson (2003).

It is important to recognize that every statistical method has its advantages and disadvantages (Rouder, Morey, Verhagen, Province, & Wagenmakers, 2016). For psychological science, no single statistical method can resolve the replication crisis (胡传鹏等, 2016; 刘佳, 霍涌泉, 陈文博, 解诗薇, 王静, 2018). For researchers and the field as a whole, the most critical task is to fully understand the assumptions and limitations of each statistical method; otherwise, false positives cannot be truly avoided. The content introduced in this paper may help researchers meet new reporting standards and provide richer information in their results.

References

Note: The references section was already in English in the original document and has been preserved exactly as provided.

Note: Figure translations are in progress. See original paper for figures.