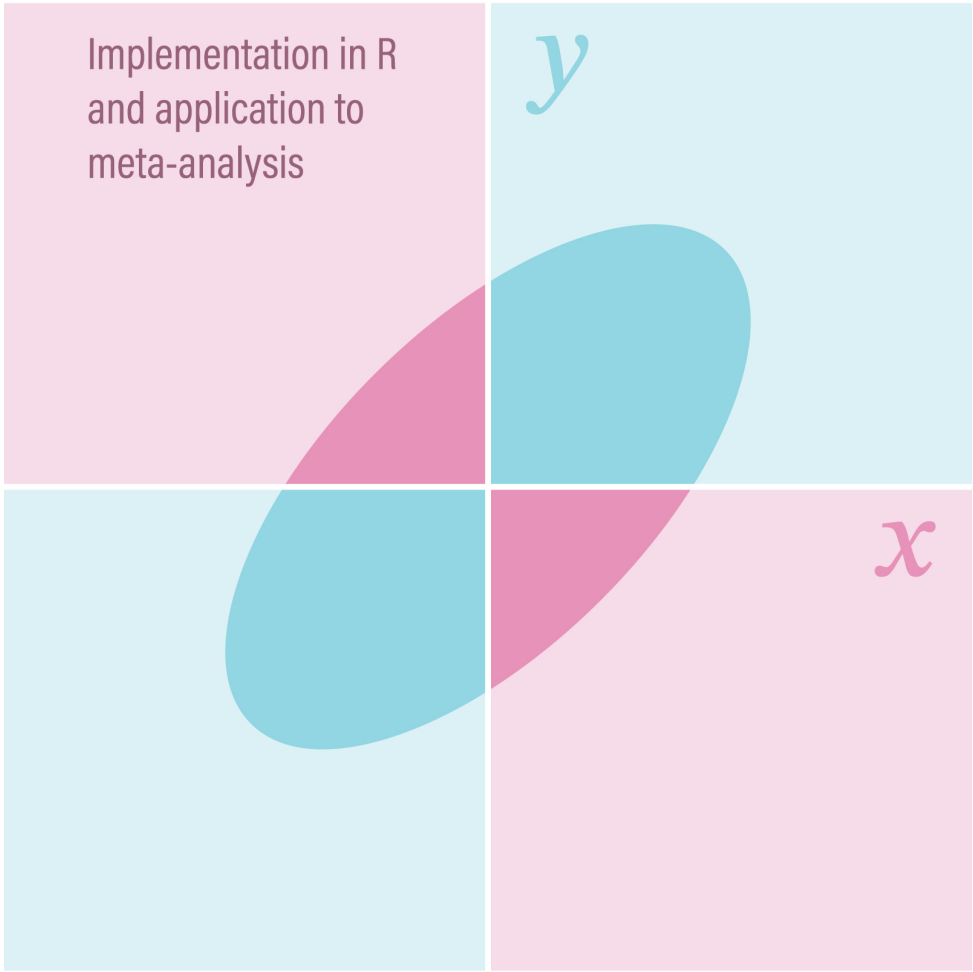
ARTIFACT CORRECTIONS

for EFFECT SIZES

Implementation in R
and application to
meta-analysis



Matthew B.
Jané

Artifact Corrections for Effect Sizes

Seeing Reality for What it is

Matthew B. Jané

2023-10-08

Table of contents

1 Greetings

Welcome to the living open source textbook ***Artifact Corrections for Effect Sizes***. This textbook covers all the essential equations and code needed to correct for biases in our effect size estimates. It will also hopefully provide readers with a deeper understanding, appreciation, and intuition for these seemingly complex formulas. It also covers how to apply these corrections to meta-analysis.

What are Statistical Artifacts?

In this book statistical artifacts will be defined broadly as **any source of contamination that induces bias in research findings**. Artifacts are present in virtually every research study, so it is crucial that we address them.

Open and Living Textbook

A living textbook is one that constantly updates with new features and is open to changes from others. This book will contain modern methods and cutting-edge techniques for artifact corrections, so in order to keep this book up-to-date it needs to grow as the research grows. New features, such as interactive figures will be added soon

It is important that this book is both open-source and open-access. All the figures, code, and documents are available in a github repository. The current maintainer of the book is Matthew B. Jané. This work is under a CC-BY license, therefore if you use any part of this work in your own work, it is important that you acknowledge it and cite it as follows:

1.0.0.0.1 * APA

Jané, M. 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/

1.0.0.0.2 * BibTeX

```
@book{MatthewBJane2023,
  title     = "Artifact Corrections for Effect Sizes: Implementation in R and Application
  author    = "Jané, Matthew B.",
```

```
  year      = 2023,
  publisher = "(n.p.)",
  url       = {https://matthewbjane.quarto.pub/artifact-corrections-for-effect-sizes/}
}
```

Contributions

Please feel free to contribute to this textbook, if your contribution makes it to the published version of this book, your name will be included in the contributor list below with a description of your work.

Name	Affiliation	Role
Velu Immonen, BS	Solent University	Designed cover and twitter preview.

2 Dedication

In Loving Memory of Haley Jané

My companion, whose love and presence have filled my life with joy and comfort.

3 Effect Sizes and Notation

3.1 What are Effect Sizes?

Effect sizes are statistics that measure the magnitude of a relationship between variables. It's important to remember that effect sizes are a valuable tool, enabling researchers to extract meaningful insights from data, rather than being the ultimate objective themselves. Effect sizes aide in researcher's ability to draw meaningful inferences from data and therefore it is crucial that they are accurate. Correlation coefficients and standardized mean differences are two of the most common effect sizes and so they will be the primary focus of this book.

3.2 Defining the Target Effect Size

It is important to clearly define the quantity that we aim to estimate (i.e., the target effect size). A well-defined target effect size guides our statistical analyses and allows us to link our empirical observations to a theoretically meaningful quantity (Lundberg, Johnson, and Stewart 2021). A target effect size is composed of two major components: the populations of interest and the variables of interest. If the methodology within a study produces effect size estimates that do not accurately reflect the variables and population of interest, then the findings can be severely biased relative to the target quantity. We can illustrate this with an example:

For example, let's say we want to know the correlation between student motivation and stress among high-school students. Suppose we investigate this correlation by administering a survey to a sample of senior students at a private high school. The survey consists of a two questions asking student's to rate their level of motivation and stress on a scale of 1-10. Now we can compare how the target population differs from what we obtained from our study and what potential sources of contamination may exist when estimating the effect size (see Table 3.1)

Table 3.1: Comparison of the target population and variables and what the example study obtained.

	Target	Study	Potential sources of contamination
Population	High-School Students	Seniors High-School students	• **Range Restriction.** Senior student's at private universities may have a more narrow range of characteristics such as stress and motivation than other classes (e.g., freshmen) and schools (e.g., public)
Variables	Motivation and Stress	Self-report 1-10 scale with questions referring to motivation and stress	• **Coarseness.** Motivation and Stress is likely naturally continuous variable rather than a discrete variable with 10 levels. • **Unreliability.** The scale may have various interpretations between people that does not reflect true individual differences in motivation.

Knowing how the study effect size differs from our intended target can guide us on which corrections we want to apply in order to achieve an effect size estimate that more accurately reflects our target effect size.

3.3 Effect Sizes

Let's start by differentiating between a population effect size and a sample estimate. The population effect size, represented by the Greek letter θ, characterizes the entire population, which encompasses all possible observations of interest. In contrast, a sample effect size (we will denote this as h) is an estimate of this population value, estimated by a random sample of observations.

The population effect size, θ, is a fixed value that does not change from sample to sample. Therefore if we were to take a random sample, i, from the population and calculate an estimate, h_i, we will find that h_i does not exactly reflect the population value. We can model h_i with the following equation (equation 1, Borenstein et al. 2010),

$$h_i = \theta + \varepsilon_i, \tag{3.1}$$

where ε_i represents sampling error. Sampling error occurs because a random sample of observations will not be identical to the entire population. Therefore this sampling error will cause the effect size estimate, h_i, to vary from sample to sample (i.e., $h_1 \neq h_2 \neq h_3$...). We can visualize this variability in Figure 3.1.

If h_i is an unbiased estimator of θ, then the average of h_i across infinite repeated samples from the population (i.e., the expectation, denoted with $\mathbb{E}_i[\cdot]$), would equal θ. For this to be true, we have to assume that the expectation of ε_i is equal to zero, such that,

$$\mathbb{E}_i[h_i] = \mathbb{E}_i[\theta + \varepsilon_i] \tag{3.2}$$
$$= \mathbb{E}_i[\theta] + \mathbb{E}_i[\varepsilon_i] \tag{3.3}$$
$$= \theta + 0 \tag{3.4}$$
$$= \theta. \tag{3.5}$$

Because θ is fixed and has no variance across samples, the variance in the effect size estimates across repeated samples is entirely attributable to the variance in sampling error such that,

$$\text{var}(h_i) = \text{var}(\theta) + \text{var}(\varepsilon_i) \tag{3.6}$$
$$= 0 + \text{var}(\varepsilon_i) \tag{3.7}$$
$$= \text{var}(\varepsilon_i). \tag{3.8}$$

Since we don't have access to all possible samples, the variance of h_i (or ε_i) must be estimated. Estimates of sampling error variance are usually a function of the sample size (i.e., the number of observations in a given sample; will be denoted by n). The square root of the estimated sampling error variance estimate can be defined as the standard error (Borenstein et al. 2010, equation 10),

$$se(h_i) = \sqrt{\widehat{\text{var}}(h_i)} = \sqrt{\widehat{\text{var}}(\varepsilon_i)},$$

where the hat indicates a sample estimate $\widehat{(\cdot)}$.

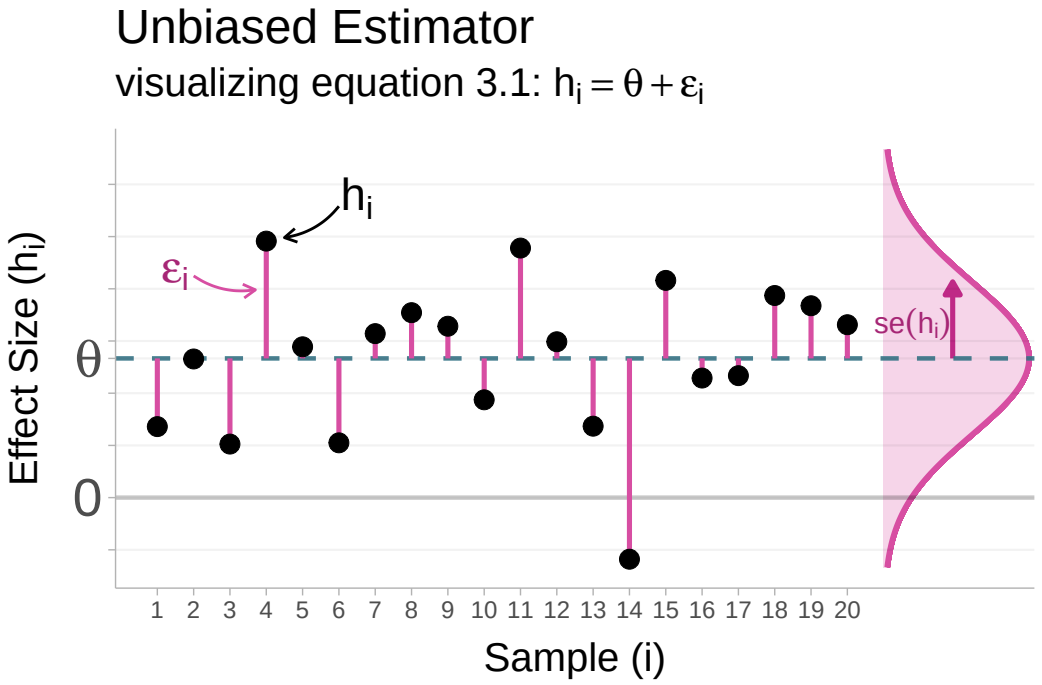


Figure 3.1: Illustration of the equation described in Equation 3.1. This figure shows the distribution of h_i. The dashed blue line denote the target population effect size, θ. The red lines denote random sampling errors, ε_i which represent the difference between the estimates h_i and the population value, $\alpha\theta$.

3.4 Effect Sizes and Study Artifacts

In practice, effect sized estimated within a study are often biased relative to the target population effect size. Specifically, the expectation of the effect size estimate over replications is not equal to the target population effect size. These sources of bias is what we will refer to as study artifacts. The target effect size is the quantity that we want to estimate which will depend on the research goal. We will continue with the same notation as the previous section, therefore θ will represent the target population effect size. We can define bias as the difference between the target population effect size, θ, and the expectation of the effect size estimate contaminated by artifacts, h_i, such that,

$$\text{bias} = \mathbb{E}_i[h_i] - \theta. \tag{3.9}$$

To account for this bias in contaminated effect size estimates, we can append Equation 3.1 to now include an artifact attenuation/inflation factor we can denote as α (J. E. Hunter and Schmidt 2015,

adapted from the third equation of p. 134 and the second equation of p. 139):

$$h_i = \alpha\theta + \varepsilon_i. \tag{3.10}$$

Therefore, assuming the expectation of the sampling errors (ε_i) is zero, we can compute the expectation of a contaminated effect size estimate as:

$$\begin{aligned}
\mathbb{E}[h_i] &= \mathbb{E}[\alpha\theta + \varepsilon_i] && (3.11)\\
&= \mathbb{E}[\alpha\theta] + \mathbb{E}[\varepsilon_i] && (3.12)\\
&= \alpha\theta + 0 && (3.13)\\
&= \alpha\theta. && (3.14)\\
& && (3.15)
\end{aligned}$$

If $\alpha = 1$ then that would indicate that h_i is an unbiased estimate of θ (see Figure 3.1), if $\alpha > 1$ that would indicate that h_i is on average *inflated* relative to θ (see Figure 3.2) and if $\alpha < 1$ then h_i is *attenuated* relative to θ (see Figure 3.3). If we divide both sides of Equation 3.10 by α such that,

$$\frac{h_i}{\alpha} = \theta + \frac{\varepsilon_i}{\alpha},$$

we see that the expectation of h_i/α is equal to θ,

$$\begin{aligned}
\mathbb{E}\left[\frac{h_i}{\alpha}\right] &= \mathbb{E}\left[\theta + \frac{\varepsilon_i}{\alpha}\right] && (3.16)\\
&= \mathbb{E}\left[\theta\right] + \mathbb{E}\left[\frac{\varepsilon_i}{\alpha}\right] && (3.17)\\
&= \theta + \frac{\mathbb{E}\left[\varepsilon_i\right]}{\alpha} && (3.18)\\
&= \theta + \frac{0}{\alpha} && (3.19)\\
&= \theta. && (3.20)\\
& && (3.21)
\end{aligned}$$

In principle, if we knew the value of α then we would be able to obtain an unbiased estimate of the target effect size θ. In practice, α is usually not known and must be estimated, we will denote an estimate of α with the English letter a. Obtaining α or a is non-trivial and is the primary objective of this book. The procedure depends on the type of effect size, the research goal, the type of artifact, as well as other

considerations. It is important to note that a will itself contain sampling error (sampling errors for a will be denoted by ξ_i), and therefore the estimate of a_i for a given sample/study i can be defined as,

$$a_i = \alpha + \xi_i.$$

> Cases where α is known
>
> In the chapters on Small Samples, Artificial Dichotomization, and Scale Coarseness we will see that, given some statistical assumptions, α can be a known (fixed) quantity.

Now we can correct the contaminated effect size estimate with the artifact attenuation/inflation factor, we can define a corrected effect size for a given study i as,

$$h_{c_i} = \frac{h_i}{\alpha} \approx \frac{h_i}{a_i}.$$

We can also see that the artifact factor can be obtained by taking the ratio of the corrected effect size to the uncorrected effect size:

$$\alpha = \frac{h_i}{h_{c_i}} \quad \text{or} \quad a_i = \frac{h_i}{h_{c_i}}. \tag{3.22}$$

A corrected correlation can be modeled as function of the target population value and ε_{c_i} (J. E. Hunter and Schmidt 2015, adapted from equation seen on p. 134),

$$h_{c_i} = \theta + \varepsilon_{c_i}.$$

The sampling variance of h_{c_i} can be quite complicated to calculate and it will depend on how a_i is obtained, each artifact correction chapter will cover how to estimate the sampling variance and standard error for the corrected effect size. If we know the precise value of α or if a_i is fixed and does not vary, the sampling variance would be a simple transform of the sampling variance of the contaminated effect size h_i such that (J. E. Hunter and Schmidt 2015, equation 3.27),

$$\text{var}\left(h_{c_i}\right) = \frac{\text{var}\left(h_i\right)}{\alpha^2} = \frac{\text{var}\left(h_i\right)}{\left(\frac{h_i}{h_{c_i}}\right)^2}.$$

The standard error can be expressed as the square root of the estimated sampling variance,

$$se(h_{c_i}) = \sqrt{\widehat{\text{var}}(h_{c_i})}$$

Standard errors as defined here are sample estimates. The expectation of the standard error is equal to the square root of the true sampling variance such that, $\mathbb{E}\left[se(h_{c_i})\right] = \sqrt{\text{var}(h_{c_i})}$.

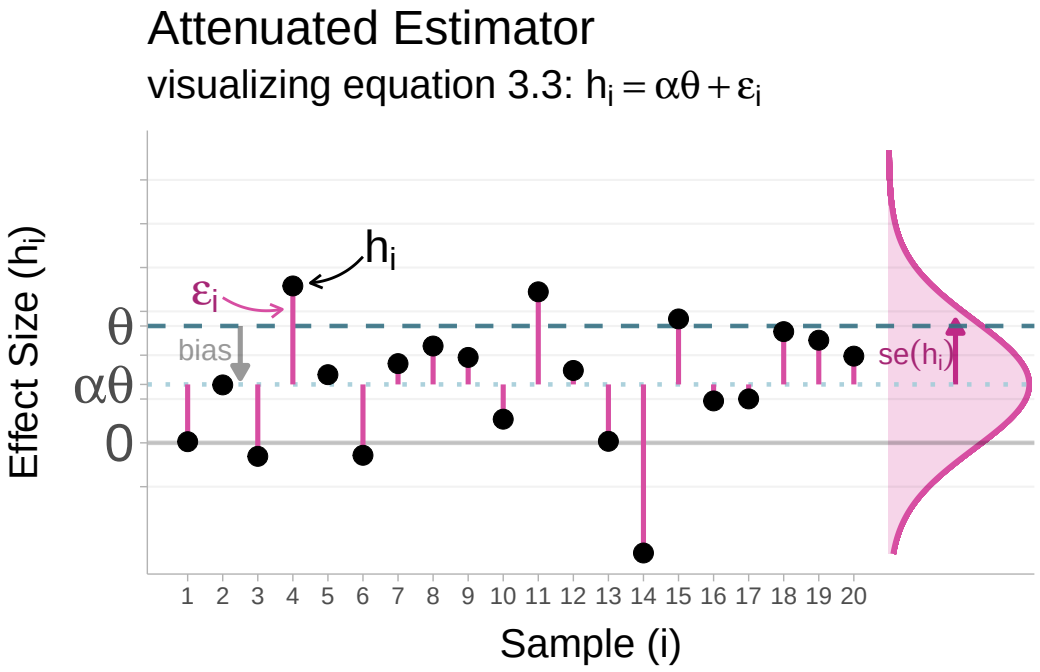


Figure 3.2: Illustration of the equation described in Equation 3.10. This figure shows the distribution of h_i when $\alpha < 1$ (i.e., attenuation). The dotted line denotes the contaminated population effect size, $\alpha\theta$, whereas the dashed blue line denotes the target population effect size, θ. The red lines denote random sampling errors, ε which represent the difference between the contaminated estimates h_i and the contaminated population value, $\alpha\theta$. The grey arrow indicates direction and magnitude of the bias in the effect size estimator (defined by Equation 5.6).

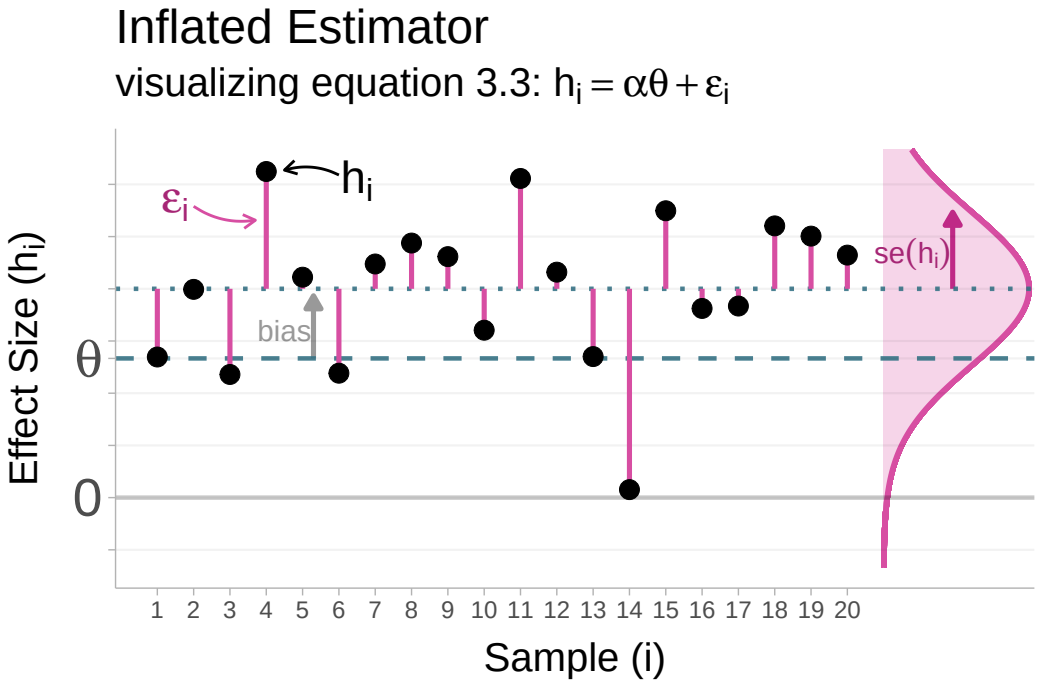


Figure 3.3: Illustration of the equation described in Equation 3.10. This figure shows the distribution of h_i when $\alpha > 1$ (i.e., inflation). The dotted line denotes the contaminated population effect size, $\alpha\theta$, whereas the dashed blue line denotes the target population effect size, θ. The red lines denote random sampling errors, ε which represent the difference between the contaminated estimates h_i and the contaminated population value, $\alpha\theta$. The grey arrow indicates direction and magnitude of the bias in the effect size estimator (defined by Equation 5.6).

3.5 Correlations

A correlation describes the relationship between two continuous variables. The population correlation (ρ) between two variables, X and Y, can be defined as the covariance (σ_{XY}) divided by standard deviations of X (σ_X) and Y (σ_Y) (Cooper, Hedges, and Valentine 2009, equation 11.21),

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

A correlation can range anywhere between -1 to 1. To visualize a positive and negative correlation see Figure 3.4.

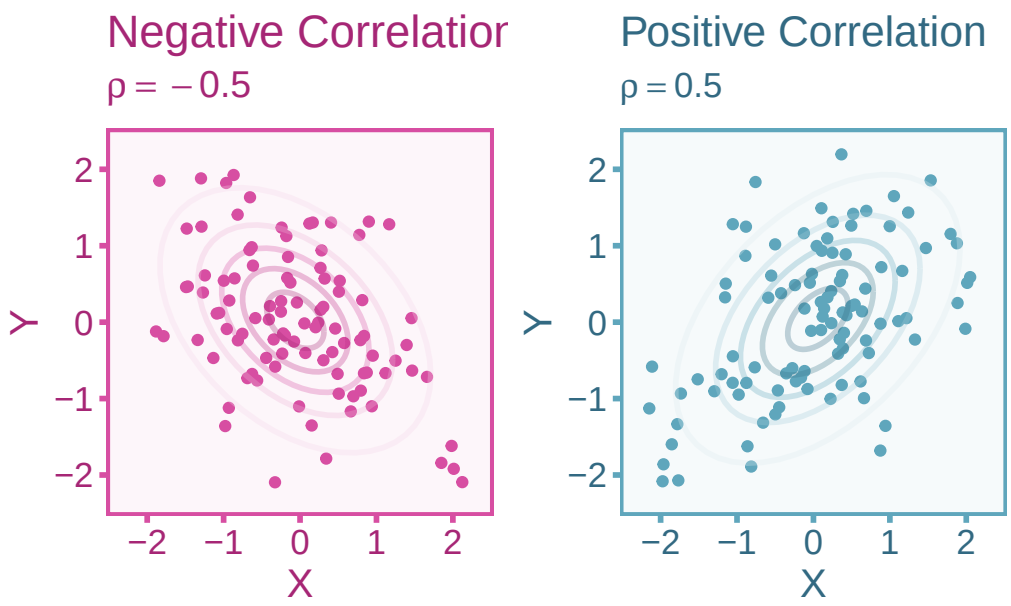


Figure 3.4: Diagram illustrating a correlation between two variables (X and Y). The left panel shows a negative correlation and the right panel shows a positive correlation. The ellipses show the contour of a bivariate normal distribution which the data points are sampled from.

A sample estimate of the population correlation can be denoted as r and consists of sample estimates of the covariance and the standard deviations:

$$r = \frac{S_{XY}}{S_X S_Y}. \tag{3.23}$$

This is referred to as the Pearson correlation coefficient. The Pearson correlation coefficient was first introduced by Auguste Bravais (1844), then later developed further by Karl Pearson, lending itself to the name. The sample estimates of S_{XY}, S_X, and S_Y can be computed with the following formulations:

$$S_{XY} = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})(Y_j - \overline{Y}) \tag{3.24}$$

$$S_X = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})^2} \tag{3.25}$$

$$S_Y = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \overline{Y})^2}. \tag{3.26}$$

Where n is the number of observations (i.e., the sample size) and $\overline{X}$ and $\overline{Y}$ denote the sample mean of X and Y, respectively. The Pearson correlation estimator (r) can then be calculated with the formula (Pearson 1895),

$$r = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_{j=1}^{n}(X_i - \overline{X})(Y_j - \overline{Y})}{\sqrt{\sum_{j=1}^{n}(X_j - \overline{X})^2}\sqrt{\sum_{j=1}^{n}(Y_j - \overline{Y})^2}}.$$

Note that the $\frac{1}{n-1}$ term from each of the formulas for S_{XY}, S_X, S_Y, cancels out.

Similar to Equation 3.1, we can model the relationship between a estimate correlation r_i for sample i and the population correlation (J. E. Hunter and Schmidt 2015,equation 2.1),

$$r_i = \rho + \varepsilon_i. \tag{3.27}$$

Where ε_i is the sampling error term. In the absence of artifacts and assuming finite covariance and variance, the Pearson correlation, r, is an asymptotically unbiased (i.e., as n approaches infinity) estimator of the population correlation, ρ (see the *Assumptions* box in this section). The sampling variance for a Pearson correlation is defined as (Pearson and Filon 1898, 174),

$$\text{var}(r_i) = \frac{\left(1-\rho^2\right)^2}{n}. \tag{3.28}$$

The sampling variance for a correlation coefficient assumes bivariate normality. In a study contaminated by artifacts, the contaminated correlation, r_i, is biased relative to the target population correlation, ρ. Bias can be accounted for by incorporating an artifact inflation/attenuation factor into Equation 3.27,

$$r_i = \alpha\rho + \varepsilon_i. \tag{3.29}$$

Note that the population value (or the expected value) of r_i is now equal to $\alpha\theta$, therefore the sampling variance for the contaminated correlation would be (Cooper, Hedges, and Valentine 2009, combination of equation 17.5 and equation 17.3),

$$\text{var}(r_i) = \frac{\left(1 - (\alpha\rho)^2\right)^2}{n}.$$

Assumptions

A Pearson correlation is an asymptotically unbiased estimator assuming that the variances and covariance are finite (Hyndman 2010). It is often thought that a Pearson correlation strictly assumes bivariate normality, however this is not true. Bivariate normality ensures that the Pearson correlation provides a complete description of the association between two variables. This was described succinctly in a Stack Exchange answer by Rob Hyndman (2010):

> Pearson's correlation is a measure of the linear relationship between two continuous random variables. It does not assume normality although it does assume finite variances and finite covariance. When the variables are bivariate normal, Pearson's correlation provides a complete description of the association.

Although a Pearson correlation is asymptotically unbiased regardless of the distribution, the assumption of bivariate normality *is* required for sampling variance. Specifically, Equation 3.28 is only true if the X and Y are bivariate normal. When distributions are not at-least approximately normal, bootstrapping is recommended to obtain variance estimates (see Bishara and Hittner 2017).

Since the population value is not known in practice, we can estimate the sampling variance by replacing $\alpha\rho$ with the contaminated correlation r_i and divide by the degrees of freedom $(n-1)$ rather (Cooper, Hedges, and Valentine 2009, equation 17.5),

$$\widehat{\text{var}}(r_i) = \frac{\left(1 - r_i^2\right)^2}{n-1}. \tag{3.30}$$

The standard error can be computed as the square root of the estimated sampling variance, $se(r_i) = \sqrt{\widehat{\text{var}}(r_i)}$.

We can obtain a corrected value of a correlation coefficient by dividing by the artifact attenuation/inflation factor:

$$r_{c_i} = \frac{r_i}{\alpha} \approx \frac{r_i}{a_i}. \tag{3.31}$$

As mentioned earlier, if the artifact factor is fixed and known then we can estimate the sampling variance of r_{c_i} by adjusting the estimated variance of r_i (J. E. Hunter and Schmidt 2015, adapted from equation 3.20),

$$\widehat{\text{var}}(r_{c_i}) = \frac{\widehat{\text{var}}(r_i)}{\alpha^2}.$$

Thus the standard error is $se(r_i) = \sqrt{\widehat{\text{var}}(r_{c_i})}$.

Applied Example in R

Let's load in the `iris` data set that contains various physical measurements of three species of plants. We can also subset the data set to only look at the Setosa species:

```
# load in data
data("iris")

# subset rows to only include setosa species
df <- subset(iris, df$species == 'setosa')

# view first 6 plants
head(iris)
```

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

Let's say we want to know the correlation between petal length and sepal length. We can use the `cor()` function in base R to obtain the Pearson correlation coefficient:

```
r <- cor(iris$Petal.Length,iris$Sepal.Length)
r
```

```
[1] 0.8717538
```

The result shows a strong positive correlation between petal and sepal length ($r = .87$). To compute the sampling variance and the standard error, we can use the `var_error_r()` function in the `psychmeta` package (Dahlke and Wiernik 2019)

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

# compute sampling variance
var_r <- var_error_r(r, n = nrow(iris), correct_bias = FALSE)

# compute standard error from sampling variance
se_r <- sqrt(var_r)

# print results
data.frame(r,var_r,se_r)
```

```
         r        var_r       se_r
1 0.8717538 0.0003867233 0.01966528
```

3.6 Standardized Mean Differences

Standardized mean differences are used to quantify the difference between the average value between two groups. In the population, the standardized mean difference (δ), the difference between the mean of group A (μ_A) and group B (μ_B) is standardized by the population within group standard deviation (σ) (Cooper, Hedges, and Valentine 2009, equation 11.15),

$$\delta = \frac{\mu_A - \mu_B}{\sigma}.$$

Dividing by the population within-group standard deviation assumes that the standard deviations within both groups are fixed and equal (i.e., $\sigma = \sigma_A = \sigma_B$).

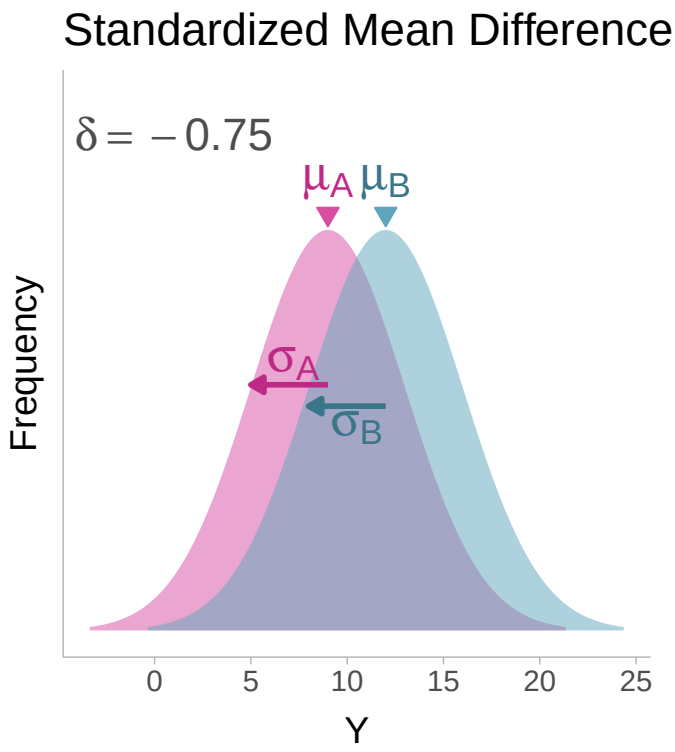


Figure 3.5: Diagram illustrating a standardized mean difference in the population between two normal distributions. The mean and standard deviation of group A is $\mu_A = 9$ and $\sigma_A = 4$, respectively. Whereas mean and standard deviation of group B is $\mu_A = 12$ and $\sigma_A = 4$, respectively. Therefore the standardized mean difference is $\delta = (9-12)/4 = -.75$. Note that $\sigma_A = \sigma_B$.

Cohen (1988) developed a sample estimator of δ that is commonly referred to as Cohen's d, we will use the term standardized mean difference or d value instead. Within a sample, we can estimate δ with the sample estimator (d) (Cooper, Hedges, and Valentine 2009, equation 11.96),

$$d = \frac{\overline{Y}_A - \overline{Y}_B}{S_p}. \tag{3.32}$$

Where S_p is the pooled sample standard deviation and the bars indicate the sample mean. The pooled standard deviation computes a weighted average (weighted by the within-group degrees of freedom, e.g., $n_A - 1$) of the within-group sample variance and then takes the square root (equation 12.12, Cooper, Hedges, and Valentine 2009),

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

Where n_A and n_B represents the sample size within group A and group B, respectively. Similar to a Pearson correlation, the sample standardized mean difference (d value) is an asymptotically unbiased estimator (i.e., unbiased as the sample size approaches infinity; see the *Assumptions* box in this section). In the absence of artifacts, we can model a d value from a sample, i (J. E. Hunter and Schmidt 2015, first equation on p. 292),

$$d_i = \delta + \varepsilon_i. \tag{3.33}$$

The variance in d_i is (Hedges and Olkin 1985, equation 8),

$$\text{var}(d_i) = \frac{n}{n_A n_B} + \frac{\delta}{2n}, \tag{3.34}$$

where n is the total sample size ($n = n_A + n_B$).

Assumptions

Cohen's (1988) estimator of the standardized mean difference (Equation 4.1) is an asymptotically unbiased estimator (as n approaches infinity; see *Small Sample* chapter for bias in finite samples) of the population value as long as two assumptions are met: 1) within-group population variances are finite and 2) the within-group population variances are fixed and equal between both groups. It is often thought that the within-group distributions must be strictly normal, however this is only a requirement for the sampling variance. Specifically, Equation 3.34 is only true if the distribution within each group is normal. When distributions are not at-least approximately normal, bootstrapping is recommended to obtain variance estimates.
Interpretation of a standardized mean difference may be affected by non-normal distributions, because overlap statistics as well as other common language interpretations are often derived from the assumption that both distributions are normally distributed.

In a study contaminated by artifacts, we can append Equation 3.33 to include an artifact inflation/attenuation factor (α). Thus we can define a contaminated d value as,

$$d_i = \alpha\delta + \varepsilon_i. \tag{3.35}$$

The population value of the *contaminated* standardized mean difference is $\alpha\delta$, where the *target* population value is δ. To estimate the sampling error variance of the contaminated d value, we can append Equation 3.34 to use the population value of a contaminated correlation, $\alpha\delta$,

$$\text{var}(d_i) = \frac{n}{n_A n_B} + \frac{\alpha\delta}{2n}.$$

Since the contaminated population standardized mean difference, $\alpha\delta$, is unavailable, we can replace $\alpha\delta$ with it's estimate, d_i (equation 7.23, J. E. Hunter and Schmidt 2015),

$$\widehat{\text{var}}(d_i) = \left(\frac{n-1}{n-3}\right)\left(\frac{n}{n_A n_B} + \frac{d_i}{2n}\right). \tag{3.36}$$

Note that the multiplicative term, $(n-1)/(n-3)$, should be included because using the sample d value instead of the population value in Equation 3.36 produces biased variance estimates in small sample sizes (Goulet-Pelletier and Cousineau 2018). The standard error of the contaminated d value can thus be defined as, $se(d_i) = \widehat{\text{var}}(d_i)$. The artifact attenuation/inflation factor, α (or the estimate, a_i) can then be used to correct the d value for bias induced by artifact contamination,

$$d_{c_i} = \frac{d_i}{\alpha} \approx \frac{d_i}{a_i}.$$

Occasionally, corrections can not be applied to the contaminated d value directly. In such cases, we may have to correct d by first converting to a *point-biserial* correlation (i.e., a Pearson correlation between a dichotomous group variable and continuous variable), correcting the correlation, and then converting back to a (corrected) d value (J. E. Hunter and Schmidt 2015). The three-step procudure can be applied as follows:

1. Convert d to r using the proportion of subjects in group A or group B ($p = n_A/n$ or $p = n_B/n$) (Wiernik and Dahlke 2020, equation 9),

$$r_i = \frac{d_i}{\sqrt{\frac{1}{p_i(1-p_i)} + d_i}}.$$

2. Correct the point-biserial correlation,

$$r_{c_i} = \frac{r_i}{\alpha} \approx \frac{r_i}{a_i}.$$

3. Convert r_c back to a d value. If the proportions of group membership are contaminated by artifacts, then we need the true group proportions in the target population (p_i^*) to convert back to d. If the the true group proportions are unavailable, then we can use the observed proportions as estimates (Wiernik and Dahlke 2020, equation 10)

$$d_{c_i} = \frac{r_{c_i}}{\sqrt{p_i^*(1-p_i^*)(1-r_{c_i}^2)}} \approx \frac{r_{c_i}}{\sqrt{p_i(1-p_i)(1-r_{c_i}^2)}}.$$

The sampling variance of the corrected standardized mean difference can be quite complicated when there is sampling error in the artifact factor. However if a_i is fixed, we can use the same three-step procedure to adjust the sampling variance (Wiernik and Dahlke 2020, table 3):

1. Convert $\widehat{\text{var}}(d_i)$ to $\widehat{\text{var}}(r_i)$,

$$\widehat{\text{var}}(r_i) = \frac{\widehat{\text{var}}(d_i)}{\left(1 + d_i^2 p_i \left[1 - p_i\right]\right)^2 \left(d_i^2 + \frac{1}{p_i(1-p_i)}\right)}.$$

2. Estimate the sampling variance of the corrected point-biserial correlation,

$$\widehat{\text{var}}(r_{c_i}) = \frac{\widehat{\text{var}}(r_i)}{\alpha^2} = \frac{\widehat{\text{var}}(r_i)}{\left(\frac{r_i}{r_{c_i}}\right)^2}.$$

3. Then convert the sampling variance of r_c back to the sampling variance of a d value,

$$\widehat{\text{var}}(d_{c_i}) = \frac{\widehat{\text{var}}(r_{c_i})}{p_i^* \left(1 - p_i^*\right) \left(1 - r_{c_i}^2\right)^3}.$$

Alternatively, we can squeeze the three-step procedure into a single equation (Wiernik and Dahlke 2020, equation 28),

$$\widehat{\text{var}}(d_{c_i}) = \frac{\widehat{\text{var}}(d_i) \left(\frac{r_{c_i}}{r_i}\right)^2}{\left(1 + d_i^2 p_i \left[1 - p_i\right]\right)^2 \left(d_i^2 + \frac{1}{p_i(1-p_i)}\right) p_i^* \left(1 - p_i^*\right) \left(1 - r_{c_i}^2\right)^3}$$

Applied Example in R

Let's load in a dataset for a plant growth experiment:

```
# load in data
data("PlantGrowth")

# view first 6 plants
head(PlantGrowth)
```

```
  weight group
1   4.17  ctrl
2   5.58  ctrl
3   5.18  ctrl
4   6.11  ctrl
5   4.50  ctrl
6   4.61  ctrl
```

We see that the data set contains two variables, weight of the plant and the experimental group. There are three two experimental groups present: a control group and two treatment groups. If we want to obtain the standardized mean difference between the each treatment group and the control group, we can use the `cohen.d` function `psych` package (William Revelle 2023):

```
library(psych)

# estimate standardized mean difference for first treatment group
d1 <- cohen.d(weight ~ group,
              data = subset(PlantGrowth, group == 'ctrl' | group == 'trt1'))
d1$cohen.d
```

```
          lower     effect     upper
weight -1.44938 -0.5615655 0.3411078
```

```
# estimate standardized mean difference for first treatment group
d2 <- cohen.d(weight ~ group,
              data = subset(PlantGrowth, group == 'ctrl' | group == 'trt2'))
d2$cohen.d
```

```
            lower    effect    upper
weight 0.05850126 1.005987 1.929048
```

We see that the first treatment showed a negative effect on plant growth ($d = -0.56$) and the second treatment shows a positive effect ($d = 1.01$). For our purposes we want the estimate and the standard error of the estimate, therefore we can extract that information:

```
data.frame(d = d1$cohen.d[,'effect'],
           se.d = d1$se)
```

```
           d      se.d
1 -0.5615655 0.8952442
```

```
  data.frame(d = d2$cohen.d[,'effect'],
             se.d = d2$se)
```

```
         d      se.d
1 1.005987 0.9352736
```

Part I

Artifact Corrections

4 Small Samples

4.1 Introduction

The purpose of sample statistics is to draw meaningful inferences about the population. However, effect size estimators such as Pearson's correlation coefficient and Cohen's d are biased in small sample sizes. This small sample bias is an artifact and can be adjusted with the appropriate correction factor.

4.2 When Correcting alongside other Artifacts

The small sample bias should always be corrected for prior to applying any other artifact correction. It is independent of all other artifact corrections and therefore the corrected effect sizes in this section can be treated as the uncorrected effect sizes in other sections.

4.3 Correcting Standardized Mean Differences for Small Sample Bias

4.3.1 Defining the Target Quantity

Our quantity of interest is the population standardized mean difference, δ, between groups A and B. We can model the relationship between the population standardized mean difference and the estimate (d),

$$d = a\delta + e.$$

Where a is an attenuation/inflation factor and e is our sampling error term. Ultimately, we can obtain an unbiased estimate of the population standardized mean difference by correcting the sample standardized mean difference as follows,

$$d_c = \frac{d}{\hat{a}}.$$

4.3.2 Artifact Correction for *d*

As the sample size approaches infinity, Cohen's estimator of the standardized mean difference is unbiased (Hedges 1981; Cohen 2013). However, in small sample sizes Cohen's estimator is inflated, that is, on average, it overestimates the population standardized mean difference. To see why this is the case, we can first define the population standardized mean difference between group A and group B such that,

$$\delta = \frac{\mu_A - \mu_B}{\sigma}.$$

Where $\bar{y}_A$ and $\bar{y}_B$ are the observed arithmetic means of group A and group B, respectively. A sample estimate of the standardized mean difference is,

$$d = \frac{\overline{Y}_A - \overline{Y}_B}{S_p} \tag{4.1}$$

Where S_p is the pooled standard deviation (i.e., weighted average within-group standard deviation). The estimator, d, is an asymptotically unbiased estimate of δ. We can denote this asymptotic relationship as,

$$\mathbb{E}_i[d_i] \overset{n}{\asymp} \delta.$$

However, d is a biased estimator of δ when the sample size is finite. Particularly, the smaller the sample size, the larger the bias. We can see that in Figure 4.1, d tends to over-estimate δ. Therefore, we can apply an artifact inflation factor, a, to capture this over-estimation,

$$\mathbb{E}_i[d_i] = a\delta.$$

The reason for this bias in d values is two-fold:

- Standard deviations tend to be attenuated in small sample sizes. This is due to the fact that although variance (squared standard deviation) is an unbiased estimator of the population variance, the square root of the variance (i.e., the standard deviation, S_p) is a biased estimate of the population standard deviation ($\mathbb{E}_i[\sqrt{S_{p_i}}] \neq \sqrt{\sigma}$, Holtzman 1950).
- A ratio is biased in small sample sizes (Kempen and Vliet 2000), therefore the ratio between the mean difference and the standard deviation (see Equation 4.1) will likewise be biased.

To obtain an unbiased estimate of the population standardized mean difference, we need to first estimate the artifact inflation factor, a. In this case, the artifact inflation factor has been mathematically derived previously by Hedges (1989). For other types of artifacts, a is unknown in practice and must be estimated, however, for small sample bias the exact value of a is known. The precise value of a is a function of sample size (equation 6e, Hedges 1989),

$$a = \frac{\Gamma\left(\frac{n-3}{2}\right)\sqrt{\frac{n-2}{2}}}{\Gamma\left(\frac{n-2}{2}\right)}.$$

Where $\Gamma(\cdot)$ denotes the gamma function. The gamma function is factorial function generalized to non-integers (note that a factorial function on integers would look something like: $3! = 3 \cdot 2 \cdot 1$, Taboga 2021). There is also an approximation of a that is more computationally trivial (re-arrangement of the first formula on pp. 114, Hedges 1989):

$$a \approx \frac{4n-9}{4n-12}$$

However, with the advent of computers, this approximation formula is unnecessary. We can see in Figure 4.2 that there is notable bias when sample size is below 20. Furthermore, the bias is most pronounced when the sample d value is larger (there is no bias at $d = 0$).

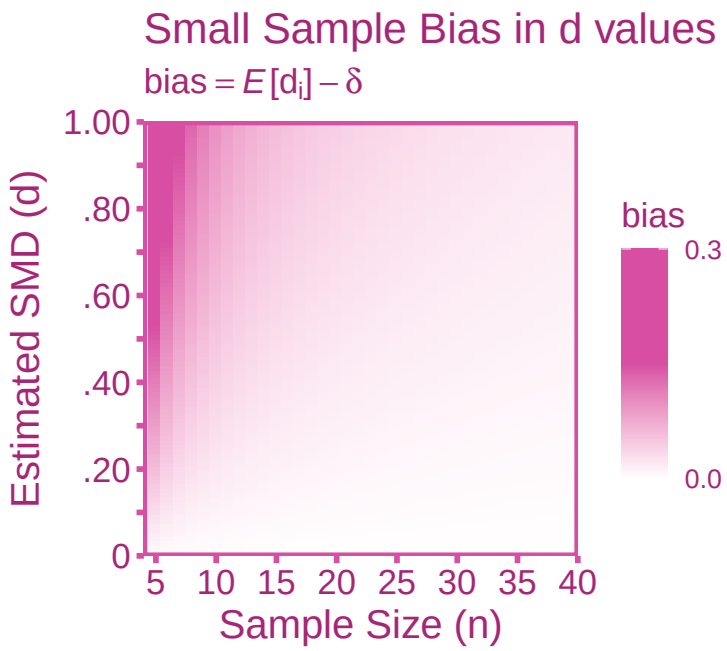


Figure 4.1: Plot showing the bias in the standardized mean difference computed in small samples. The X-axis is the sample size (n, the vertical bars are indicative of each integer). The Y-axis is the the estimated standardized mean difference (d). The dark pink coloring indicates more bias.

Using a, we can correct the d value such that,

$$d_c = \frac{d}{a} = \frac{d}{\left[\frac{\Gamma(\frac{n-3}{2})\sqrt{\frac{n-2}{2}}}{\Gamma(\frac{n-2}{2})}\right]}. \tag{4.2}$$

To obtain the standard error of d_c we can apply the same correction as above to the standard error of d (se),

$$se(d_c) = \frac{se(d)}{a} = \frac{se(d)}{\left[\frac{\Gamma(\frac{n-3}{2})\sqrt{\frac{n-2}{2}}}{\Gamma(\frac{n-2}{2})}\right]}. \tag{4.3}$$

Applied Example in R

Let's say we want to compute the mean difference of test scores between two classes. Class 1 has 11 students and class 2 has 10 students. We then obtain a standardized mean difference between the two classes of $d = .43$. In R, we can use the `correct_d_bias()` function to correct the point estimate (using Equation 4.2) and then `var_error_g` function to correct the error variance and thus the standard error (using Equation 4.3). Both of these functions are found in the `psychmeta` package (Dahlke and Wiernik 2019).

```
library(psychmeta)

d <- .43
n1 <- 11
n2 <- 10

# correct
dc <- correct_d_bias(d = d,
                     n = n1+n2)

var_corrected <- var_error_g(g = dc,
                             n1 = n1,
                             n2 = n2)

# print results
cbind(dc=dc, se = sqrt(var_corrected))
```

```
           dc       se
[1,] 0.4144264 0.424262
```

The output shows a corrected standardized mean difference of $d_c = 0.41$ ($se(d_c) = 0.42$)

4.4 Correcting for Small Sample Bias in Correlations

4.4.1 Defining the Estimand

Our quantity of interest is the population correlation, ρ. We can model the relationship between the population correlation and our sample estimate (r) with,

$$r = a\rho + e$$

Where a is our small sample biasing factor and e is our sampling error term. Ultimately, we can obtain an unbiased estimate of the population correlation by correcting the observed correlation as follows,

$$r_c = \frac{r}{a}$$

4.4.2 Artifact Correction for *r*

Let's first define the correlation in the population as the covariance between X and Y (σ_{XY}) standardized by the product of the standard deviation of X (σ_X) and Y (σ_Y):

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

The sample estimate can be defined as (S denoting the sample estimates of σ),

$$r = \frac{S_{XY}}{S_X S_Y}$$

Asymptotically, the expectation of a sample correlation is equal to the population correlation,

$$\mathbb{E}_i[r_i] \overset{n}{\asymp} \rho.$$

However, similar to standardized mean differences, correlations are also biased in small sample sizes (Olkin and Pratt 1958). Unlike standardized mean differences, Pearson correlations are *under*-estimated, rather than over-estimated. Therefore, an artifact *attenuation* factor, a, can account for this bias,

$$\mathbb{E}_i[r_i] = a\rho.$$

Because the attenuation factor, a, has been mathematically derived directly Olkin and Pratt (equation 2.3, 1958), there is no need to obtain a noisy estimate of a. Therefore a can be defined exactly as,

$$a = \frac{1}{F\left(\frac{1}{2}, \frac{1}{2}; \frac{n-1}{2}; 1 - r^2\right)} \tag{4.4}$$

Where $F(\cdot)$ is the hypergeometric function (for background see the Wikipedia page:). The hypergeometric function is a complicated and iterative function which which can be defined in terms of $\Gamma(\cdot)$ functions (plugging in values into equation 2.2, Olkin and Pratt 1958)

$$F\left(\frac{1}{2},\frac{1}{2};\frac{n-1}{2};1-r^2\right)=\sum_{z=0}^{\infty}\frac{\Gamma\left(\frac{1}{2}+z\right)\Gamma\left(\frac{n-1}{2}\right)\left(1-r^2\right)^z}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-1}{2}+z\right)z!}$$

Although this formula is complex, it can be easily done in R so do not worry! There is also an approximation of a that is much simpler than Equation 4.4,

$$a\approx\frac{2(n-3)}{2n-r^2-5}$$

Then we can correct the point-estimate the sampling variance for small sample bias. I will emphasize that approximations are not necessary if a computer is available. We can see in Figure 4.2 that there is notable bias when sample size is below 15. Furthermore, the bias is most pronounced when the sample correlation around .60 (there is no bias at $r=0$ and $r=\pm1$).

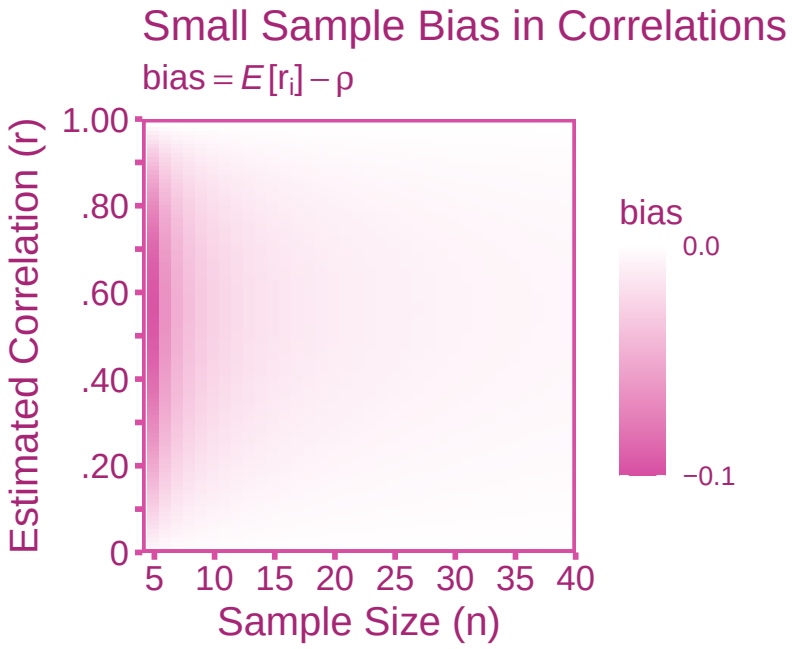


Figure 4.2: Plot showing the bias in the correlations computed in small samples. The X-axis is the sample size (n, the vertical bars are indicative of each integer). The Y-axis is the the estimated Pearson correlation (r). The dark pink coloring indicates more bias.

To correct for small sample bias, we can divide the sample correlation r by the attenuation factor a,

$$r_c=\frac{r}{a}=\frac{r}{\left[\frac{1}{F\left(\frac{1}{2},\frac{1}{2};\frac{n-1}{2};1-r^2\right)}\right]} \tag{4.5}$$

Where the standard error of r_c can be obtained by correcting the observed standard error (se),

$$se(r_c) = \frac{se(r)}{a} = \frac{se(r)}{\left[\frac{1}{F(\frac{1}{2},\frac{1}{2};\frac{n-1}{2};1-r^2)}\right]} \tag{4.6}$$

Applied Example in R

Let's say we compute a correlation of $r = .43$ between test scores and self-reported motivation in a small sample of students. In R, we can use the `escalc` function in the `metafor` package (Viechtbauer 2010). Setting `measure = "UCOR"` will apply the small sample correction for the point estimate (Equation 4.5) and the error variance (i.e., square of the standard error, Equation 4.6). Note in order to run this function you must have the `gsl` package already installed.

```
# install.packages('gsl')
library(metafor)

r <- .43
n <- 11

# correct the correlation
escalc(measure = 'UCOR',
       ri = r,
       ni = n,
       var.names = c("rc", "se.2"),
       digits = 3)
```

```
    rc  se.2
1 0.454 0.063
```

The output shows a corrected correlation between test scores and motivation of $r_c = .454$ ($se(r)^2 = .063$).

5 Unreliability

5.1 Introduction

In general terms, measurement is the process of quantifying an attribute or characteristic of something. In scientific measurement, the measurand is the quantity or the attribute we intend to measure. In the psychological sciences, measurands usually take the form of constructs such as intelligence or anxiety. Often the goal of measurement is to produce quantities (i.e., scores) that accurately reflect the measurand. However, quantities that do not reflect a *real* attribute can still have useful predictive value (e.g., socio-economic status). It is important to note that measures are not all created equal, some perform better than others. Ideally, measures should produce scores that are consistent and repeatable, this is referred to as the *reliability* of a measure. A high quality measure should produce highly reliable scores. This section will review what reliability is in theory, how to estimate reliability, and how to correct effect sizes for measurement error.

5.2 Reliability in True Score Theory

True score theory (or classical test theory) is a mathematical formalization of observed scores obtained from a measurement procedure. Observed scores, X_m, is defined as a score obtained upon measurement m. The true score model assumes that each individual, has a true score, T, that stays constant over repeated measurements. Variation in observed scores over repeated measurements is due to measurement-specific error, E_m,

$$X_m = T + E_m.$$

Here, measurements are *strictly parallel*. Strictly parallel measurements have the following four properties (p. 69, Haertel 2006):

1. Measurements have identical specifications. That is, each measurement uses the same measurement procedure.
2. The distribution of observed scores for each measurement are identical: $f(X_1) = f(X_2) = ...$
3. Any set of two measurements are assumed to covary the same as any other set of two measurements: $\sigma_{X_1X_2} = \sigma_{X_2X_3} = \sigma_{X_1X_3} = ...$
4. Each measurement equally covaries with any other variable: $\sigma_{X_1Y} = \sigma_{X_2Y} = ...$

True scores can be defined as the expected value (i.e., the mean) of observed scores over repeated measurements such that, $\mathbb{E}_m[X_m] = T$. Given this assumption, it follows that the average of the resultant errors is zero across repeated measurements, $\mathbb{E}_m[E_m] = 0$. It also follows that the covariance between observed scores from measurement to measurement must only be attributable to the variation in true scores ($\sigma_{XX'} = \sigma_T^2$) and therefore true scores and errors are independent ($\sigma_{ET} = 0$). The independence between true scores and errors provide convenient parsing of the variance in observed scores,

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2. \tag{5.1}$$

In practice, the goal is to obtain observed scores that closely resemble true scores, therefore it is important to minimize measurement error variance (σ_E^2). If $\sigma_E^2 = 0$, then the scores can be said to have perfect reliability, that is, observed scores do not vary upon repeated measurements and are thus identical to true scores. In practice, this is virtually never the case. In true score theory, *reliability* can be defined as the correlation between observed scores in parallel measurements,

$$\rho_{XX'} = \frac{\sigma_{XX'}}{\sigma_X \sigma_{X'}} = \frac{\sigma_T^2}{\sigma_X^2}.$$

The reliability can also be expressed by the square of the correlation between observed scores and true scores. To understand why this is the case, note that the covariance between parallel forms of a measure is equivalent to the covariance between observed scores and true scores, $\sigma_{XT} = \sigma_{(T+E)T} = \sigma_T^2 + \sigma_{TE} = \sigma_T^2 = \sigma_{XX'}$ (Haertel 2006),

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\left(\sigma_T^2\right)^2}{\sigma_X^2 \sigma_T^2} = \frac{\sigma_{XT}^2}{\sigma_X^2 \sigma_T^2} = \rho_{XT}^2. \tag{5.2}$$

To visualize how reliability relates true scores and observed scores see the structural diagram Figure 5.1.

It is important to emphasize that true scores are expected values over repeated observations and they do not necessarily correspond to an actual, tangible attribute of interest (Borsboom and Mellenbergh 2002). As a result, every measurement has a true score, regardless of whether it gauges a concrete attribute or not. For example, if we construct a test by summing the responses to the items: "how many languages can you confidently hold a conversation in?" and "Estimate the number of photos you've taken in the last year across all devices". Even in such a nonsensical measure, the test's composite score retains a true score, but this true score does not mirror a tangible reality.

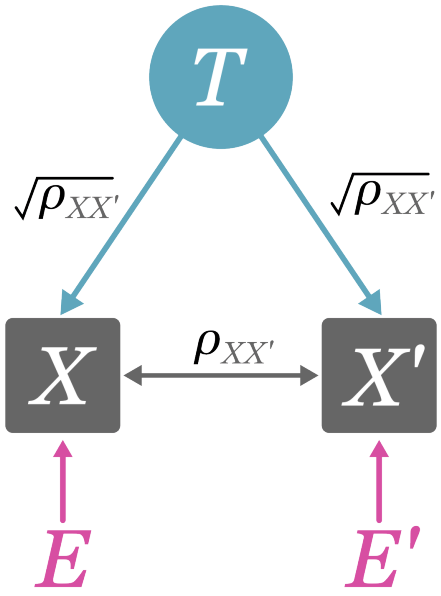


Figure 5.1: Structural diagram illustrating the relationship between true scores, observed scores, and error scores. The pink circle labeled t indicates the true scores, the blue squares labeled with X and X' represent observed scores on parallel measurements, and the red E denotes error. Correlations between T, X, and X' are in terms of reliability ($\rho_{XX'}$). Note that $\rho_{XX'} = \rho^2_{XT}$.

5.3 Reliability vs Validity

Reliability and validity are distinct properties in measurement. Validity pertains to whether the scores produced by a measure reflects the quantities it is intended to measure (p. 14 Kelley 1927). According to Borsboom, Mellenbergh, and Van Heerden (2004), a measure is *valid* if both of the following statements are true:

1. The attribute exists.
2. Variations in the attribute causally produce variations in the outcomes of the measurement procedure.

Borsboom's formulation of validity is simpler and more practical than other formulations such as Cronbach and Meehl's (1955) nomological network approach to validity. It is important to note that even if an attribute does not exist (statement 1), scores may still provide predictive utility. For example, socio-economic status (SES) is a formative quantity that is constructed from a composite of education, income, occupation status, etc. Although SES is not causal to these indicators, SES can still be used as a predictor of important life outcomes.

5.4 Estimating Reliability

In practice, reliability must be estimated through indirect methods, this is due to the fact that true scores and errors are unknown. There are many reliability estimators that can be used, however we will go over a selection of internal consistency estimators as well as test-retest stability estimators.

5.4.1 Internal Consistency Estimators

Taking multiple measurements and then averaging tends to provide a more stable estimate of true values. For instance, let's consider the case of Francis Galton (1907), who conducted a study involving 787 individuals estimating the weight of an ox. On average, each person's estimate deviated by approximately 37 pounds from the actual weight of the ox, which was recorded as 1198 pounds. However, when all the guesses were averaged together, the combined estimate was 1207 pounds, just a 9 pound difference from the true value. This principle can be extended to broader applications, such as measuring psychological constructs. If we were to assess someone's level of extraversion using ratings from their mother, father, friend, and sibling, the average of their combined assessments would yield a more reliable score compared to relying solely on a single evaluator. So to create a more stable composite score (X), we can take the score from κ items (x) and sum them such that,

$$X = x_1 + x_2 + ... + x_\kappa.$$

The most commonly reported reliability estimator in the psychological sciences is coefficient alpha, also referred to as Cronbach's alpha. Coefficient alpha, along with other internal consistency estimators, serves the purpose of assessing the reliability of composite scores comprising multiple item scores. Coefficient alpha was first derived by Lee Cronbach (see equation 13, Cronbach 1951) only requires three parameters to calculate, the number of measurements (κ), the sample variances of each item ($S^2_{x_m}$), and the variance of the composite score (S^2_X). Coefficient alpha will estimate the reliability of the composite observed score ($r_{XX'}$),

$$_\alpha r_{XX'} = \frac{\kappa}{\kappa - 1}\left(1 - \frac{\sum_{m=1}^{\kappa} S^2_{x_m}}{S^2_X}\right). \tag{5.3}$$

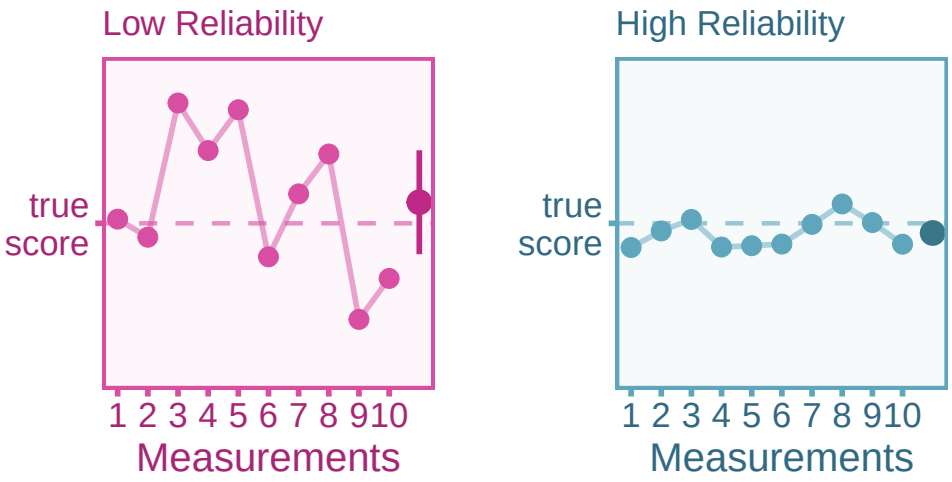


Figure 5.2: Figures showing the observed scores upon 10 repeated measurements and the composite observed score for a single person (the true score is denoted with the dashed line). The left panel shows 10 observed scores with a lot of variation (i.e., low reliability). The composite score (dark red dot with error bars), shows wide error bars illustrating the low precision of the observed score score. The right panel also shows 10 observed scores with little variation (i.e., high reliability). The composite score (dark blue dot with error bars), shows narrow error bars illustrating the high precision of the observed score.

With tighter assumptions (see Haertel 2006), the formula for coefficient alpha can be simplified to just two parameters: the number of measurements (κ) and the average correlation between measured scores ($\bar{r}_{x_i x_j}$, where $i \neq j$). This formula is known as Spearman-Brown's prophecy (see equation III of Charles Spearman 1910; or the last equation on page 299 of Brown 1910),

$$_{\text{sb}}r_{XX'} = \frac{\kappa \bar{r}_{x_i x_j}}{1 + (\kappa - 1)\bar{r}_{x_i x_j}} \tag{5.4}$$

This can be simplified further if we have two observed item scores. This formulation is a variation of split-half reliability:

$$_{\text{sh}}r_{XX'} = \frac{2r_{x_1x_2}}{1 + r_{x_1x_2}} \tag{5.5}$$

All of these reliability estimators measure internal consistency, therefore they do not account for error outside of the measurement-specific error. There are other sources of error that internal consistency reliability estimates do not account for, such as transient error or rater-specific error.

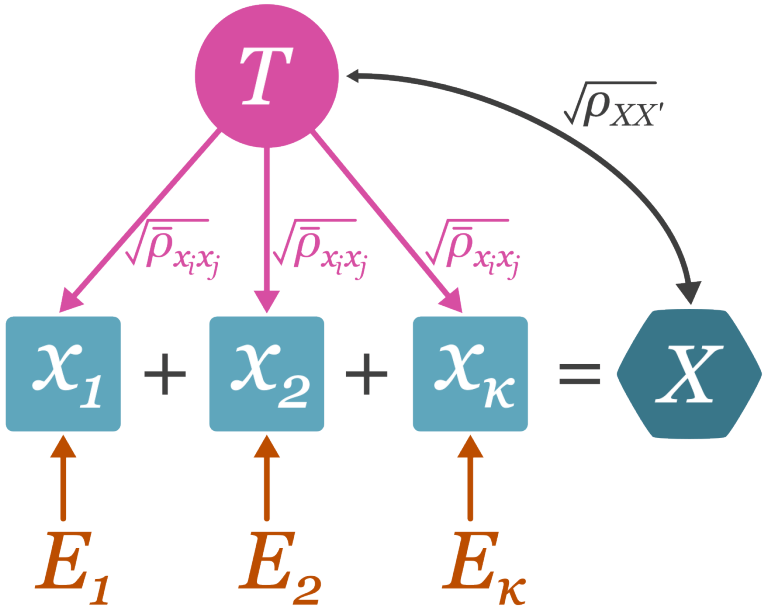


Figure 5.3: Structural model illustrating internal consistency. The pink circle labeled T indicates the true scores, the blue squares, $x_{1\ldots\kappa}$, represent the observed sub-scores across multiple measurements, and the red e denotes error. The dark blue hexagon, X, indicates a composite score as a sum of the observed scores ($\varkappa_{1\ldots\kappa}$). Note that $\rho_{XX'} = \rho_{XT}$.

Applied Example

Using the `anxiety` data set from the `irr` package (Gamer et al. 2019), we can estimate the reliability of three raters evaluating the anxiety in a sample of 20 individuals. Let us first load in the data:

```r
library(irr)

# load in data
data("anxiety", package = "irr")

# print data
head(anxiety, 4)
```

```
  rater1 rater2 rater3
1      3      3      2
2      3      6      1
3      3      4      4
4      4      6      4
```

We can also use the `alpha()` function from the `psych` package (William Revelle 2023) to estimate coefficient alpha too. The output of `alpha` provides a lot of useful information, and it reports two types of alphas: standardized and raw. The standardized alpha is based on the correlations between items rather than the covariance between items, which is useful if items are on completely different scales (standardized alpha equivalent to the Spearman-Brown prophecy formula found in Equation 5.4). Also you will notice that the standardized alpha and). If the ratings are on the same scale then the raw_alpha is preferred.

```r
library(psych)

reliability <- alpha(anxiety)

rXX <- reliability$total$raw_alpha

rXX
```

```
[1] 0.4525862
```

The output shows a very low reliability between raters ($r_{XX'} = .45$).

5.4.2 Test-Retest Stability Estimator

Transient errors represent fluctuations in observed scores over time. These fluctuations, even if they are systematic (e.g., fatigue over the course of a single day), add extraneous within-person variance that can mask individual differences. Considering transient fluctuations as error depends on the research goal, so it is important for researchers to take care in considering which variance components should be

considered error in their study (see Section 5.4.3). To estimate test-retest reliability, we can compute the correlation between the measurement at time 1 (X_{t_1}) and the second measurement at time 2 (X_{t_2}),

$$_{\text{tr}}r_{XX'} = \text{corr}(X_{t_1} X_{t_2}).$$

Note that calculating the pearson correlation coefficient between time-points ignores systematic changes (e.g., practice effects). We can visualize test-retest reliability in Figure 5.4 where the top panels show the correlations between time points and the bottom panels show the within-person change between time-points for scores with high and low reliability.

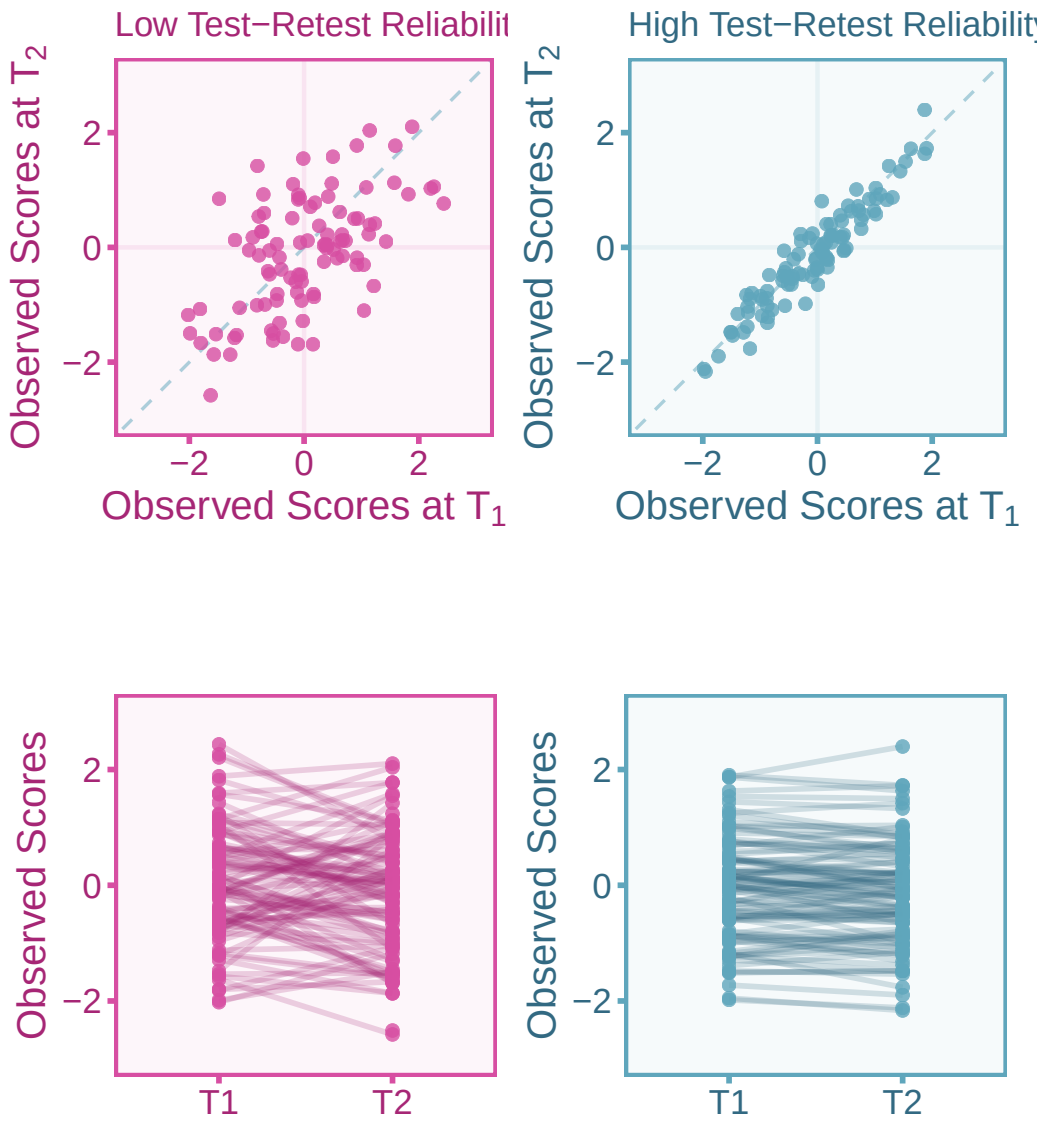


Figure 5.4: Illustrating test-retest reliability. Top-left and top-right panels show the correlation between observed scores at both time-points for a measure that has low and high reliability, respectively. Bottom-left and bottom-right panels show the within-person change from time-point 1 to time-point 2 for scores with low and high reliability, respectively.

5.4.3 Sources of Measurement Error

Measurement error can itself be broken down into multiple sources of error. Depending on the study, different sources of error may be more relevant than others. Different reliability estimators will account for different sources of error therefore it is important for researchers to choose the right reliability estimator for their study. A description of four of the most common sources of error is adapted from table 1 of Wiernik and Dahlke (2020):

1. Random Response Error: Genuine randomness in responses. Examples include: motor errors and variation in response time.
2. Time/Environment-Specific (Transient) Error: Fluctuations in scores as a result of the specific time or environment of the measurement. For instance, if researchers administered an ability test to a sample of undergraduate students throughout the course of a day, the student's who complete the test at the end of the day will likely perform worse than participant's who completed due to fatigue rather than ability. Errors due to illness, mood, hunger, environmental distractors, etc. all fall under the umbrella of transient errors.
3. Instrument-Specific Error: Error due to the specific content or make-up of the measurement instrument. For example, a psychological scale using Likert items may show participant's idiosyncratic interpretations of questions and response options rather than their standing on the latent construct.
4. Rater/Observer-Specific Error: Errors induced by idiosyncratic biases of individual raters and rater by ratee interactions (e.g., Teacher A gives higher grades to students who stay after class).

Different estimators of reliability account for different sources of measurement error therefore depending on the research design, it is important to carefully choose which reliability is most relevant for your use case. Note that even if two estimators account for the same types of measurement error, they likely hold different assumptions that may be violated in a given research context.

Table 5.1: List of reliability coefficients and the sources of error they account for. The sources of error are denoted by the columns labeled 1-4, corresponding to each of the four sources of error.

Estimator	Description	1	2	3	4
Coefficient Alpha	Internal consistency coefficient for composite measures.	X		X	

Estimator	Description	1	2	3	4
Coefficient Omega	Internal consistency coefficient for composite measures with specified factor structure.	X		X	
Split-Half	Internal consistency coefficient for measurements that are split into two halves.	X		X	
Kuder-Richardson 20	Internal consistency when observed scores are binary (special case of coefficient alpha).	X		X	
Item Response Theory Reliability	Reliability coefficient derived from item response theory (as opposed to classical test theory)	X		X	
Inter-Rater/Inter-Observer Reliability	Consistency in scoring between raters/observers.	X			X
Test-Retest	Stability coefficient for repeated measurements across time	X	X		

Estimator	Description	1	2	3	4
Delayed Alpha	Average of all possible split-half reliabilities	X	X	X	
G-Coefficient	Reliability coefficient derived from generalizability theory (G-theory). Can incorporate any source of error if enough data is present.	X	X	X	X

5.5 Correction for Bias in Correlations

5.5.1 Defining the Target Quantity

Continuing with our emphasis on clearly defining our quantity of interest prior to applying any corrections, let us define it here. Our target is the population correlation between true scores of our independent and dependent variables. We can define the observed scores of the independent and dependent variables X and Y as,

$$X = T + E_X$$

$$Y = U + E_Y.$$

Where T and U are the true scores for the independent and dependent variables, respectively. The population true score correlation can thus be be denoted by, ρ_{TU}, and can be defined as the standardized covariance,

$$\rho_{TU} = \frac{\sigma_{TU}}{\sigma_T \sigma_U}.$$

In a given study, we will only have access to the observed scores of the independent and dependent variables, therefore the study correlation is r_{XY}. The relationship between the observed correlation and the true population correlation can be defined as,

$$r_{XY} = a\rho_{TU} + \varepsilon.$$

Where a is the artifact attenuation factor (we will see that measurement error attenuates rather than inflates the correlation). An unbiased estimate of the true score population correlation (ρ_{TU}) can then be calculated by dividing the observed score correlation by an estimate of the artifact attenuation factor,

$$r_{TU} = \frac{r_{XY}}{\hat{a}}.$$

The measurement model can be visualized in Figure 5.5.

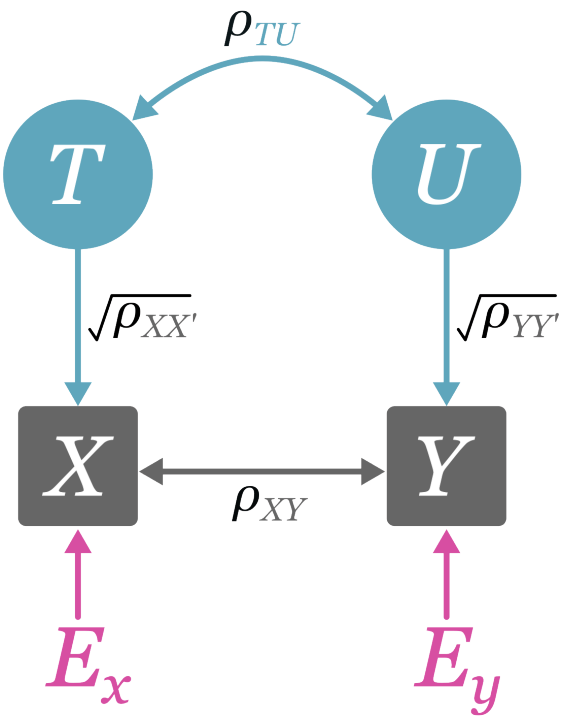


Figure 5.5: This figure shows the relationship between the true scores, observed scores, and error scores. The true score correlation is denoted by the curved arrow connecting the circles indicating true score variables, T and U.

5.5.2 Artifactual Correction for Unreliability

Measurement error induces systematic bias in effect size estimates such as correlation coefficients C. Spearman (1904). In the population, let us assume there is some factor a that accounts for the systematic bias in observed score correlations (ρ_{XY}) relative to true score correlations (ρ_{TU}), such that

$$\rho_{XY} = a\rho_{TU}.$$

Since the correlation is defined as the covariance standardized by the standard deviations, the population correlation between true scores, T and U, is defined as,

$$\rho_{TU} = \frac{\sigma_{TU}}{\sigma_T \sigma_U}.$$

Likewise the correlation between the observed scores, X and Y, would be the observed covariance divided by the observed standard deviations,

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

However, if we assume that there is no covariance between errors in X and Y ($\sigma_{E_X E_Y} = 0$), then the covariance between observed scores is only attributable to the covariance between true scores, therefore $\sigma_{XY} = \sigma_{TU}$. This means that the observed score correlation can be expressed as,

$$\rho_{XY} = \frac{\sigma_{TU}}{\sigma_X \sigma_Y}. \tag{5.6}$$

Now the only difference between the observed score correlation and the true score correlation is the standard deviations in the denominator. In the presence of measurement error, the observed score standard deviations (σ_X and σ_Y) will be larger than the true score standard deviations (σ_T and σ_U). Using the definition of reliability, we can show how the observed variance is inflated compared to the true variance as a function of reliability. Since the reliability is defined as the ratio of true variance to total observed variance (see Equation 5.2), we can see how reliability inflates the observed variance,

$$\begin{aligned}
\sigma_X^2 &= \sigma_T^2 \left(\frac{\sigma_X^2}{\sigma_T^2} \right) \\
&= \sigma_T^2 \left(\frac{1}{\rho_{XX'}} \right) \\
&= \frac{\sigma_T^2}{\rho_{XX'}}.
\end{aligned}$$

Therefore the observed standard deviation is,

$$\sigma_X = \frac{\sigma_T}{\sqrt{\rho_{XX'}}}. \tag{5.7}$$

Since the reliability, and its square root, will be less than 1, then the observed score variance will be larger than the If we use the definition of an observed score correlation (Equation 5.6), then we can replace σ_X and σ_Y with $\frac{\sigma_T}{\sqrt{\rho_{XX'}}}$ and $\frac{\sigma_U}{\sqrt{\rho_{YY'}}}$, respectively. Now we can see how the observed score correlation differs from the true score correlation:

$$\begin{aligned}
\rho_{XY} &= \frac{\sigma_{TU}}{\left[\frac{\sigma_T}{\sqrt{\rho_{XX'}}}\right]\left[\frac{\sigma_U}{\sqrt{\rho_{YY'}}}\right]} \\
&= \frac{\sigma_{TU}}{\sigma_T\sigma_U}\sqrt{\rho_{XX'}}\sqrt{\rho_{YY'}} \\
&= \rho_{TU}\sqrt{\rho_{XX'}}\sqrt{\rho_{YY'}}
\end{aligned}$$

This attenuation formula was first derived by Charles Spearman (1904). See Figure 5.6 for a visualization of the attenuation. Note that this formulation requires that there is no correlation between E_X and E_Y ($\rho_{E_X E_Y} = 0$).

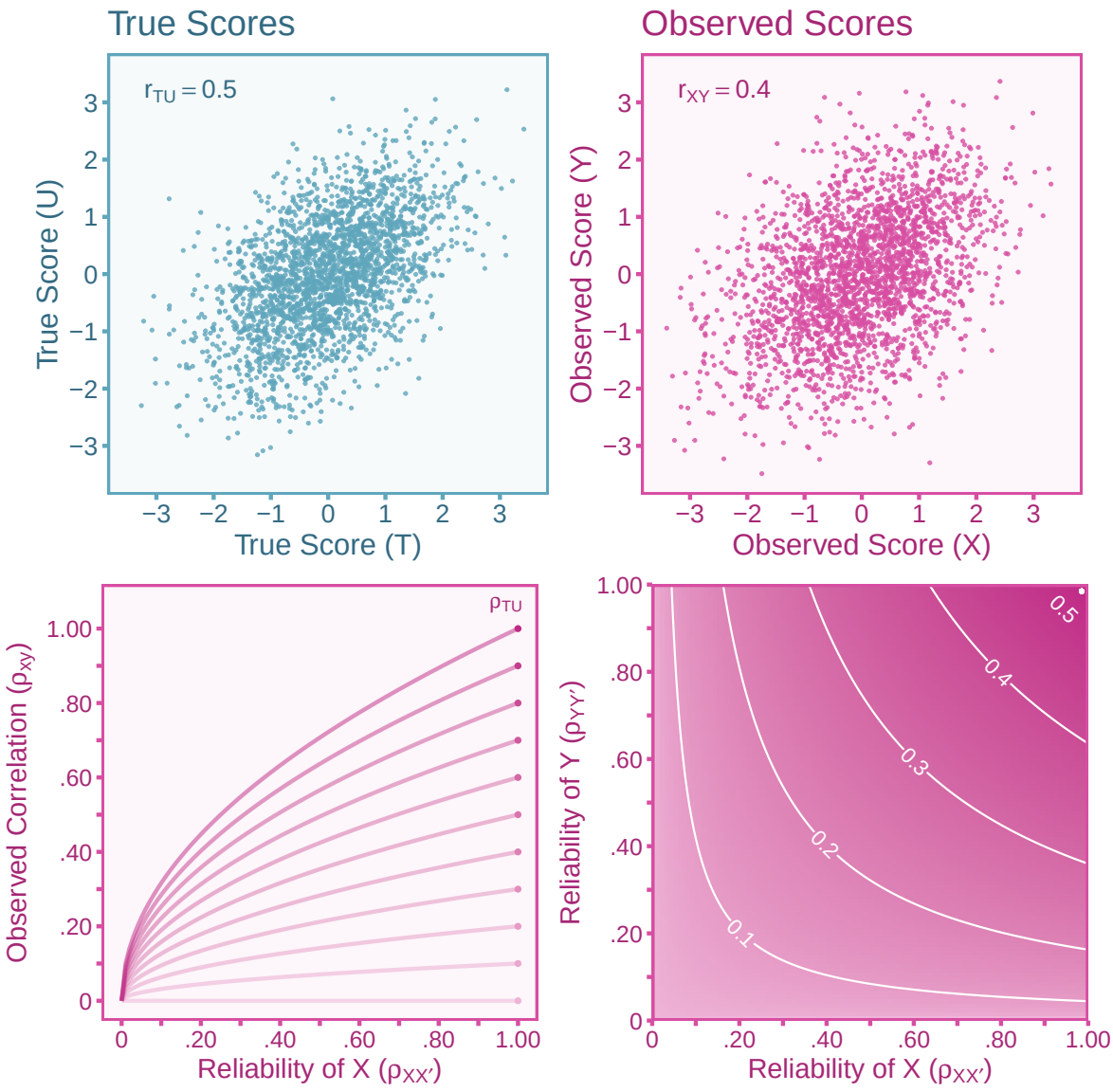


Figure 5.6: Visualizing the attenuation of an observed score correlation (ρ_{XY}) due to measurement error. Top left panel shows a scatter plot of the correlation between the true scores of the independent (T) and dependent variable (U). Top right panel shows a scatter plot of the correlation between the observed scores of the independent (X) and dependent variable (Y). The bottom left panel shows the reliability of a single variable (X) and it's relationship with the observed score correlation (ρ_{XY}) while varying the true score correlation (true score correlations are represented as the red dots, i.e., when reliability is perfect; $\rho_{TU} = \{0, .1, .2, .3, .4, .5, .6, .7, .8, .9, 1.0\}$). The bottom right panel shows attenuation of a a true score correlation of $\rho_{TU} = .5$, when both X and Y are affected by measurement error. The darker red indicates a larger observed correlation, where it peaks at an observed score correlation of .50 (when reliability of X and Y is 1) which is equivalent to the true score correlation

Let us recall that we can express the observed score correlation as a function of the true score population correlation (i.e., the target quantity, ρ_{TU}),

$$r_{XY} = a\rho_{TU} + e.$$

Where e is the sampling error and a is the artifact attenuation factor. a can now be replaced with the square root of the reliabilities of X and Y.

$$r_{XY} = \left(\sqrt{\rho_{XX'}}\sqrt{\rho_{YY'}}\right)\rho_{TU} + e.$$

Therefore we can correct the observed score study correlation by dividing by an estimate the attenuation factor using,

$$r_{TU} = \frac{r_{XY}}{\hat{a}} = \frac{r_{XY}}{\sqrt{r_{XX'}}\sqrt{r_{YY'}}}$$

The corrected correlation coefficient is an unbiased estimator of the target quantity, ρ_{TU}. We also need to obtain the standard error of the corrected correlation coefficient. The standard error of the uncorrected correlation, se, can be adjusted directly with one of two formulas:

1. If the reliability coefficients are estimated from the same sample as the correlation, then the standard error formula is,

$$se_c = \frac{se}{\sqrt{r_{XX'}}\sqrt{r_{YY'}}}$$

2. If the reliability coefficients and the correlation are all estimated from different samples (n will denote the sample size of the standardized mean difference, $n_{r_{XX'}}$ and $n_{r_{YY'}}$ and will denote the sample size of the respective reliability coefficients) then the standard error of the corrected correlation is approximately (for the derivation, see the appendix of this section),

$$se(r_{TU}) = \sqrt{\frac{se(r_{XY})^2}{r_{XX'}r_{YY'}} + \frac{r_{XY}^2\left(1 - r_{XX'}^2\right)^2}{4(n_{r_{XX'}} - 1)r_{XX'}^3 r_{YY'}} + \frac{r_{XY}^2\left(1 - r_{YY'}^2\right)^2}{4(n_{r_{YY'}} - 1)r_{XX'}r_{YY'}^3}} \tag{5.8}$$

The two ways of calculating standard error are due to the fact that when the reliability and correlation coefficient are computed from the same sample, their sampling errors will be correlated (Bobko and Rieck 1980). Note that if you only want to correct for unreliability in one variable, than you can set the reliability of the other variable to 1 in all the equations above.

Applied Example in R

Let's say we wanted to estimate the correlation between academic motivation and test performance among a sample of $n = 100$ students. Motivation scores are obtained using a self-report measure with a reliability of $r_{XX'} = .80$. Test performance is the sum of correctly answered multiple choice questions and the total scores were also fairly reliable ($r_{YY'} = .80$). The correlation between observed motivation and performance scores is $r_{XY} = .40$. To correct the correlation for measurement error, we can use the `correct_r` function in the `psychmeta` package (Dahlke and Wiernik 2019).

```
library(psychmeta)

rXY <- .40  # correlation between X and Y
rXX <- .80  # reliability of X
rYY <- .80  # reliability of Y
n <- 100    # sample size

# correct correlation
correct_r(correction = 'meas',
          rxyi = rXY,
          rxx = rXX,
          ryy = rYY,
          n = n)
```

```
Correlations Corrected for Measurement Error:
--------------------------------------------------------------------------------
  value CI_LL_95 CI_UL_95   n n_effective
1   0.5    0.276    0.691 100        51.5
```

The estimate of the true score correlation is $r_{TU} = .50$ $[.28, .69]$.

5.6 Correction for Bias in Standardized Mean Differences (*d*)

5.6.1 Defining the Target Quantity

Prior to correcting for measurement error let us define our target. Our target is the difference in the means of group A and B with respect to the true scores, U, of our dependent variable, Y,

$$Y_A = U_A + E_A$$

$$Y_B = U_B + E_B.$$

Where U_A and U_B are the true scores for group A and group B, respectively. The true score standardized mean difference can thus be be denoted by, δ_{gU} (g indicates the grouping variable, and U denotes the continuous true score outcome), and can be defined as the mean difference divided by the within-group standard deviation,

$$\delta_{gU} = \frac{\mu_{U_A} - \mu_{U_B}}{\sigma_U}.$$

Where μ_U indicates the population mean of true scores. The relationship between the population true score standardized mean difference (δ_{gU}) can be related to the observed study standardized mean difference with the following formulation:

$$d_{gY} = a\delta_{gU} + e.$$

Where a is the attenuation factor induced by measurement error and e denotes sampling error. To obtain an unbiased estimate of true score standardized mean difference, δ_{gU}, we can correct the observed standardized mean difference by dividing by an estimate of a,

$$d_{gU} = \frac{d_{gY}}{\hat{a}}.$$

5.6.2 Artifact Correction for Unreliability

The population mean of true scores and observed scores are identical since measurement error only affects the variance in scores. Therefore the population mean difference is equal between true and observed scores:

$$\mu_{U_A} - \mu_{U_B} = \mu_{Y_A} - \mu_{Y_B}$$

Note that this means that an unstandardized mean difference is *not* biased by measurement error. Since the mean difference in true scores and observed scores are the same, the observed score population standardized mean difference can also be expressed as the mean difference of true scores standardized by the observed score standard deviation,

$$\delta_{gY} = \frac{\mu_{Y_A} - \mu_{Y_B}}{\sigma_Y} = \frac{\mu_{U_A} - \mu_{U_B}}{\sigma_Y}$$

We know from Equation 5.2 that the standard deviation of true scores, σ_U, can be expressed as a function of reliability and the observed score standard deviation,

$$\sigma_U = \sigma_Y \sqrt{\rho_{YY'}} \tag{5.9}$$

Therefore the true score standardized mean difference could be corrected for by dividing the observed score standardized mean difference by the reliability,

$$\delta_{gU} = \frac{\mu_{U_A} - \mu_{U_B}}{\sigma_Y \sqrt{\rho_{YY'}}} = \frac{\delta_{gY}}{\sqrt{\rho_{YY'}}}$$

The reliability in this case is the within-group reliability ($\rho_{YY'} = \rho_{YY'_A} = \rho_{YY'_B}$) which is assumed to be equal between groups. Similarly, a sample estimator of the observed score standardized mean difference can be expressed as,

$$d_{gY} = \frac{\overline{Y}_A - \overline{Y}_B}{S_{Y_P}} = \frac{\overline{U}_A - \overline{U}_B}{S_{Y_P}}$$

Where $\overline{Y}$ and $\overline{U}$ are sample means of observed scores and true scores, respectively. The within-group standard deviation of observed scores can be estimated by pooling the standard deviation from both groups,

$$S_{Y_P} = \sqrt{\frac{(n_A + 1)S^2_{Y_A} + (n_B + 1)S^2_{Y_B}}{n_A + n_B - 2}}.$$

Where n_A and n_B are the sample sizes within group A and group B, respectively.

Pooled Reliability

To obtain the pooled standard deviation of true scores, we can correct the observed score standard deviations for measurement error such that, $S^2_Y = r_{YY'} S^2_U$ (similar to what we see in Equation 5.9). Therefore we can compute the pooled true score standard deviation (S_{U_P}) with,

$$S_{U_P} = \sqrt{\frac{(n_A + 1)\left(S_{Y_A}\sqrt{r_{YY'_A}}\right)^2 + (n_B + 1)\left(S_{Y_B}\sqrt{r_{YY'_B}}\right)^2}{n_A + n_B - 2}}.$$

Alternatively, we can pool the reliability and the true score standard deviations separately,

$$S_{Y_P} = \sqrt{\frac{(n_A+1)S^2_{Y_A} + (n_B+1)S^2_{Y_B}}{n_A+n_B-2}}$$

$$r_{YY'_P} = \sqrt{\frac{(n_A+1)r^2_{YY'_A} + (n_B+1)r^2_{YY'_B}}{n_A+n_B-2}}.$$

Then we can express S_{U_P} as the pooled standard deviation and the pooled reliability,

$$S_{U_P} = S_{Y_P}\sqrt{r_{YY'_P}}.$$

Now we can calculate the corrected sample standardized mean difference, so that it estimates the standardized mean difference in true scores rather than observed scores:

$$d_{gU} = \frac{\overline{U}_A - \overline{U}_B}{S_{U_P}} = \frac{\overline{U}_A - \overline{U}_B}{S_{Y_P}\sqrt{r_{YY'_P}}} = \frac{d_{gY}}{\sqrt{r_{YY'_P}}}$$

The corrected (true score) standardized mean difference, d_{gU}, will also need it's standard error to be adjusted from the observed score estimate. The standard error can be computed one of two ways:

1. If the reliability coefficient and the standardized mean difference are computed from the same sample then the standard error can be estimated by,

$$se(d_{gU}) = \frac{se(d_{gY})}{\sqrt{r_{YY'_P}}}.$$

2. If the reliability coefficient and the standardized mean difference are computed from separate samples (n will denote the sample size of the standardized mean difference, $n_{r_{YY'}}$ and will denote the sample size of the reliability coefficient), we can use the following formulation (see appendix for derivation),

$$se(d_{gU}) = \sqrt{\frac{se(d_{gY})^2}{r_{YY'_P}} + \frac{d^2_{gY}\left(1 - r^2_{YY'_P}\right)^2}{4\left(1 - n_{r_{YY'_P}}\right)r^3_{YY'_P}}} \tag{5.10}$$

Total Sample Reliability

It is common that studies will only report the full sample reliability and not the reliability within each group. If the groups differ substantially on the dependent variable, Y, then the total sample reliability will over-estimate the within-group reliability. When the total sample reliability is all that is available, we can correct the standardized mean difference by first converting d_{gY} to a point-biserial correlation, r_{gY}. To do this we also need the observed proportion of subjects in group A *or* B (p_g ; it does not matter which group is chosen, as long as it is consistent throughout).

$$r_{gY} = \frac{d_{gY}}{\sqrt{\frac{1}{p_g(1-p_g)} + d_{gY}^2}}.$$

Then we can the correlation coefficient as we did in Section 5.5.2. Taking the reliability of Y, the r_{gY} for the total sample reliability (see Section 5.5 for details),

$$r_{gU} = \frac{r_{gY}}{\sqrt{r_{YY'}}}$$

Then we can obtain the corrected (true score) standardized mean difference by converting r_{gU} back into d_{gU},

$$d_{gU} = \frac{r_{gU}}{\sqrt{p_g(1-p_g)(1-r_{gU}^2)}}$$

The standard error of the corrected value will also need to be adjusted. The same process can be done for the sampling variance as well, but instead we can put it all into one equation,

$$se_c = \frac{se\left(\frac{r_{gU}}{r_{gY}}\right)}{\sqrt{\left(1 + d_{gY}^2 p_g[1-p_g]\right)^3 \left(1 - r_{gU}^2\right)^3}}$$

Note that this formula is specifically for the case where the standardized mean difference and the reliability coefficient are estimated from the same sample. If they are estimated from separate samples, then we can convert the standard error of the standardized mean difference to a standard error of a point-biserial correlation and then use Equation 5.8. Once the point-biserial standard error is corrected, then it can be converted back to a correlation coefficient.

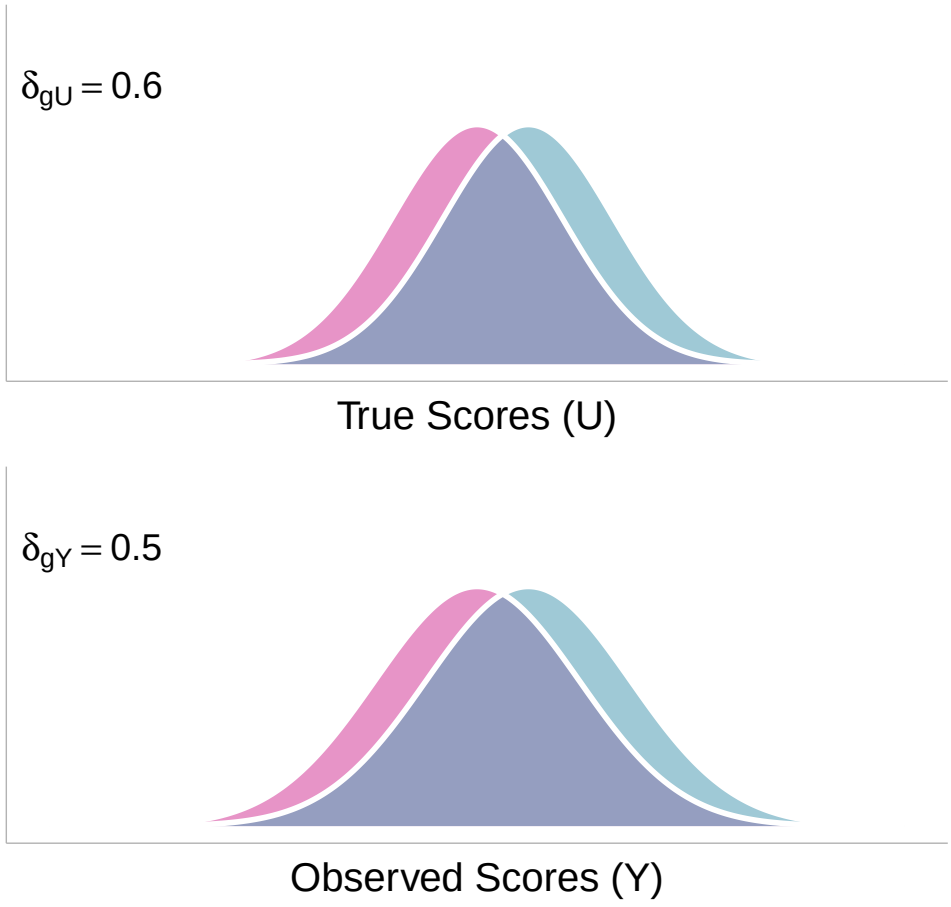


Figure 5.7: Visualizing the attenuation of the standardized mean differences in observed scores (bottom panel) from true scores (top panel). The true score standardized mean difference is $\delta_{gU} = 0.60$, but in the presense of measurement error ($\rho_{YY'} = .70$) the observed score standardized mean difference is $\delta_{gY} = 0.50$. This occurs due to the increased standard deviation in observed scores relative to true scores (the mean difference does not change).

Applied Example in R

Imagine we want to test the effectiveness of a drug that is meant to alleviate stress. Therefore we conduct a study and randomize a sample of 150 people into a control group ($n = 75$) and a treatment group ($n = 75$). After the intervention, the subjects report their stress levels on a self-report scale. The scale was shown to have a reliability of $r_{YY'} = .80$. We can correct the standardized mean differences on observed scores with the `correct_d` function in the `psychmeta` package (Dahlke and Wiernik 2019).

```
library(psychmeta)

dgY <- .40
rYY <- .80
n1 <- n2 <- 75

correct_d(correction = 'meas',
          d = dgY,
          ryy = rYY,
          n1 = n1,
          n2 = n2)
```

```
d Values Corrected for Measurement Error:
-------------------------------------------------------------------------------------------
  value CI_LL_95 CI_UL_95   n n_effective
1 0.449   0.0828    0.838 150         118
```

The output of psychmeta shows a slightly larger ($d_{gU} = 0.45\,[0.08, 0.84]$).

5.7 Estimating Reliability with Limited Information

Reliability estimates should preferably be calculated from within the study's sample, however there are a couple of ways to estimate reliability when this information is not provided. A common way to obtain an estimate of the reliability is to look in meta-analyses or a test manuals. If the number of items in a study differs from the test manual, you can approximate the reliability of the study's test, with a re-arrangement of the Spearman-Brown prophecy formula,

$$r_{xx'_{study}} \approx \frac{1}{\frac{\kappa_{\text{ref}}}{\kappa_{\text{study}}}\left(\frac{1}{r_{xx'_{study}}} - 1\right) + 1}.$$

Where κ_{ref} and κ_{study} denote the number of items in the reference test and the test used in the study, respectively.

5.8 Appendix

5.8.0.0.1 * Deriving the standard error of a corrected correlation with independent sampling errors

Equation 5.8 - This Taylor series approximation derivation can also be found in the Mathematica script, `Derivations.nb`, denoted as (SE.1).

$$
\begin{aligned}
se_c \approx & \sqrt{\frac{\partial}{\partial r_{XY}}\left[\frac{r_{XY}}{\sqrt{r_{XX'}}\sqrt{r_{YY'}}}\right]^2 \widehat{var}(r_{XY}) + \frac{\partial}{\partial r_{XX'}}\left[\frac{r_{XY}}{\sqrt{r_{XX'}}\sqrt{r_{YY'}}}\right]^2 \widehat{var}(r_{YY'}) + \frac{\partial}{\partial r_{XX'}}\left[\frac{r_{XY}}{\sqrt{r_{XX'}}\sqrt{r_{YY'}}}\right]^2} \\
\approx & \sqrt{\left(\frac{1}{r_{XX'}r_{YY'}}\right) se^2 + \left(-\frac{r_{XY}}{2\sqrt{r_{XX'}^3 r_{YY'}}}\right)^2 \frac{\left(1-r_{XX'}^2\right)^2}{n_{r_{XX'}}-1} + \left(-\frac{r_{XY}}{2\sqrt{r_{XX'}r_{YY'}^3}}\right)^2 \frac{\left(1-r_{YY'}^2\right)^2}{n_{r_{YY'}}-1}} \\
\approx & \sqrt{\frac{se^2}{r_{XX'}r_{YY'}} + \frac{r_{XY}^2\left(1-r_{XX'}^2\right)^2}{4(n_{r_{XX'}}-1)r_{XX'}^3 r_{YY'}} + \frac{r_{XY}^2\left(1-r_{YY'}^2\right)^2}{4(n_{r_{YY'}}-1)r_{XX'}r_{YY'}^3}}
\end{aligned}
$$

5.8.0.0.2 * Deriving the standard error of a corrected correlation with independent sampling errors

Equation 5.10 - Found in `Derivations.nb`, denoted as (SE.2).

6 Group Misclassification

6.1 Introduction

Group misclassification describes a situation where true group membership (e.g., people with a disorder) does not perfectly match the observed group membership (e.g., people *diagnosed* with a disorder). Group misclassification can be considered a type of measurement error where instead of accounting for errors in continuous variables (i.e., unreliability), group misclassification accounts for errors in categorical variables.

6.2 Defining Group Misclassification

Misclassification can be defined as any deviations between true group membership and observed group membership. Let us imagine two groups, group A and group B. In order to identify members of group A and group B, we have to use some measurement procedure. We can also suppose that this measurement instrument produces imperfect group assignments, that is, people who are actually in group A are sometimes assigned group B and vice versa. We can visualize the performance of the classification procedure with a contingency table (see Table 6.1) between actual group membership (G) and observed group membership (g):

Table 6.1: Contingency table between assigned group membership and actual group membership.

	$G = A$	$G = B$
$g = A$	n_{AA}	nBA
$g = B$	n_{AB}	n_{BB}

We can see from the contingency table that individual's who were correctly classified, would be labeled in the cell block AA or BB and those who were misclassified would belong to cells BA and AB. Therefore we can define the proportion of individuals that are accurately classified as $p_{\text{acc}} = \frac{n_{AA}+n_{BB}}{n_{AA}+n_{BB}+n_{AB}+n_{BA}}$ whereas the proportion of people misclassified can be defined as $p_{\text{mis}} = \frac{n_{AB}+n_{BA}}{n_{AA}+n_{BB}+n_{AB}+n_{BA}}$. A high-quality classifier would would minimize p_{mis} and maximize p_{acc}. Note that the proportion of people misclassified is inversely proportional to the proportion of people accurately classified such that, $p_{\text{mis}} = 1 - p_{\text{acc}}$.

6.3 Classification Reliability

Similar to quantifying reliability in continuous variables by calculating the correlation in parallel sets of observed scores, the same can be done in categorical variables. Instead of a contingency table between observed (g) and true (G) group membership, we will instead create a contingency table of two measurements producing two sets of observed group assignments (g and g'). Measurements often will take the form of inter-rater assessments, for example, two clinician's diagnosis of Major Depressive Disorder (MDD) in the same sample of patients.

	$g = A$	$g = B$
$g^{'} = A$	n_{AA}	n_{BA}
$g^{'} = B$	n_{AB}	n_{BB}

To obtain the reliability of the group assignments, we can calculate the correlation coefficient between g and g'. Since both variables are categorical, a Pearson correlation coefficient would be equivalent to a phi coefficient. The phi coefficient is often referred to as Matthew's correlation coefficient and is most frequently used as an index of performance of a binary classifier in machine learning. Let's denote the reliability (i.e., the correlation between g and g') as $\rho_{gg'}$. Remember that reliability from the chapter on unreliability can be defined as the square of the correlation between true scores and observed scores. As is the case here, we can define classification reliability as the square of the correlation between assigned group membership and actual group membership,

$$\rho_{gg'} = \rho_{gG}^2$$

There are a few ways to obtain a sample estimate of $\rho_{gg'}$ ($r_{gg'}$). The first way is to calculate the sample estimate directly from a contingency table,

$$r_{gg'} = \frac{n_{AA}n_{BB} - n_{AB}n_{BA}}{\sqrt{(n_{AA} + n_{BA})(n_{AB} + n_{BB})(n_{AA} + n_{AB})(n_{BA} + n_{BB})}}.$$

Where n_{AA}, n_{BB}, n_{AB}, and n_{BA} are the number of subjects within their respective cells of the contingency table. If the values of the contingency table are not available, we can calculate the reliability from the χ^2-statistic,

$$r_{gg'} = \sqrt{\frac{\chi^2}{n}}.$$

Where n is the total sample size (sum of all cells). If the χ^2-statistic is unavailable, we can approximate the reliability from the accuracy (p_{acc}) or the proportion of people misclassified (p_{mis}),

$$r_{gg'} = (2p_{\text{acc}} - 1)^2 = (1 - 2p_{\text{mis}})^2.$$

This approximation assumes that the group sizes are approximately equal *and* the misclassification rates are approximately equal between groups. Otherwise, $r_{gg'}$ will be overestimated (Wiernik and Dahlke 2020).

Applied Example in R

Let's say that a researcher wants to calculate the reliability of a clinician's diagnoses of major depressive disorder. They hire two clinicians to assign a sample of 100 patients to a control group or a major depressive disorder group. The researcher runs a χ^2-test to test the association between the clinicians group assignments and it returns $\chi^2 = 54.5$. We can then calculate the reliability of classification using base R.

```
chi2 <- 54.5
n <- 100

# calculate reliability from chi squared statistic
rgg <- sqrt(chi2/n)

# print reliability
rgg
```

```
[1] 0.7382412
```

The clinicians show a fair level of agreement with a classification reliability of $r_{gg'} = .74$.

6.4 Correcting for Group Misclassification in Standardized Mean Difference

6.4.1 Defining our Target Quantity

Our quantity of interest is the true score population standardized mean difference, δ_{GU}, between actual members of group A and group B on the true scores of the dependent variable, U. However, the observed sample standardized mean difference (d_{gY}) is estimating the difference in observed scores between individuals who are assigned group to A and group B. Non-differential error in the assignment of groups (i.e., group misclassification) will bias the observed correlation. We can model the observed standardized mean difference as a function of the target quantity, δ_{GU},

$$d_{gY} = a\delta_{GU} + e.$$

Where a is the artifact attenuation factor and e denotes the sampling error. Therefore an unbiased estimate of the target standardized mean difference can be obtained by dividing the observed standardized mean difference by an estimate of the artifact attenuation factor,

$$d_{GU} = \frac{d_{gY}}{\hat{a}}.$$

6.4.2 Artifact Correction for Standardized Mean Difference

The standardized mean differences will become biased when subject's assigned groups differ from their actual group. This is partially due to the fact that the means of each group are driven closer to one another. Let us suppose that, on average, group A and group B score differently on some outcome, Y. When some subjects are erroneously assigned to the incorrect group, the observed mean within each group will reflect a weighted average true means of both groups. This is due to the fact that the misclassified individuals are being drawn from a population with a different mean. To calculate the mean of the observed groups we must incorporate the true mean of the correctly classified subjects and the misclassified subjects,

$$\overline{Y}_A^{\text{obs}} = \left(\frac{n_{AA}}{n_{AA} + n_{BA}}\right)\overline{Y}_A^{\text{true}} + \left(\frac{n_{BA}}{n_{AA} + n_{BA}}\right)\overline{Y}_B^{\text{true}}$$

$$\overline{Y}_A^{\text{obs}} = \left(\frac{n_{BB}}{n_{BB} + n_{AB}}\right)\overline{Y}_B^{\text{true}} + \left(\frac{n_{AB}}{n_{BB} + n_{AB}}\right)\overline{Y}_A^{\text{true}}.$$

From the above equations, it becomes evident that as the number of misclassified individuals increases (n_{AB} and n_{BA}), the observed means of each group gradually converge towards each other. As the means converge, the standardized mean difference will correspondingly shift toward zero. To illustrate this phenomenon, Figure 6.1 shows the distributions for groups A and B without any misclassification. In this case, there is no attenuation of the standardized mean difference.

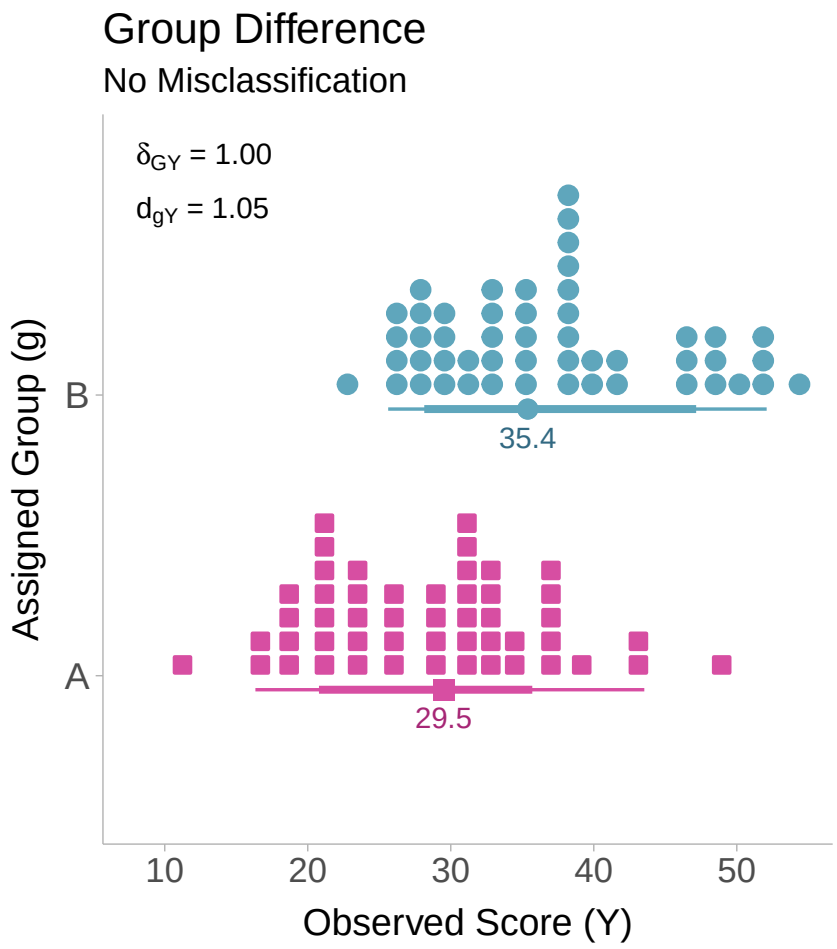


Figure 6.1: Distributions of scores without misclassification. True mean difference and observed mean differ only due to sampling error. Red squares denote actual group A members, blue circles denote actual group B members.

If some individuals are assigned to the incorrect group, then we will see attenuation in the standardized mean difference as the means converge. Figure 6.2 is showing what happens when the group misclassification rate is 10%. A group misclassification rate of 10% is equivalent to a classification reliability of $r_{gg'} = .64$.

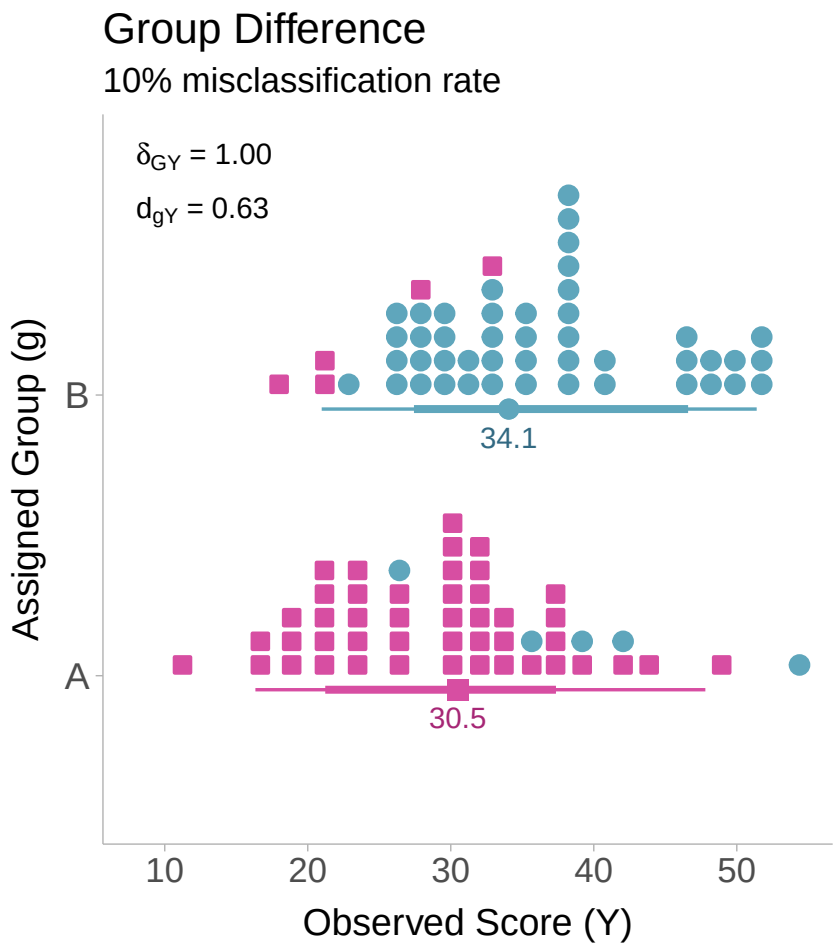


Figure 6.2: Distributions of scores with a 10% misclassification rate. Observed standardized mean differences are biased toward the null (i.e., $\delta = 0$). Red squares denote actual group A members, blue circles denote actual group B members. Note that a few members of group A (red squares) are within assigned group B and vice versa (indicative of misclassification).

It is important to note that for many of the corrections converting the standardized mean difference to a point-biserial correlation is often a necessary step. However, once the corrected point-biserial correlation is obtained, the correlation can then be converted back into a standardized mean difference. To correct for bias induced by misclassification we first need to convert the observed standardized mean difference to a point-biserial correlation coefficient by using the observed proportion of the sample that has been assigned to either group A or group B (p_g). The group proportion p_g in the following equations will only show up in the term $p_g(1-p_g)$ so it will not matter which group is used to calculate the proportion. Converting d_{gY} to r_{gY}:

$$r_{gY} = \frac{d_{gY}}{\sqrt{\frac{1}{p_g(1-p_g)} - d_{gY}^2}}.$$

We can then correct the point-biserial correlation for group misclassification by dividing by the square root of the classification reliability. Since we also want to correct for measurement error in the continuous dependent variable, Y, we can simultaneously apply the correction for unreliability:

$$r_{GU} = \frac{r_{gY}}{\sqrt{r_{gg'}}\sqrt{r_{YY'}}}.$$

Now we can convert the corrected point-biserial correlation into a corrected standardized mean difference (d_{GU}). When converting back to a standardized mean difference, we need to use the true group proportions, p_G. Although if we are to assume equal misclassification rates between groups, then the observed proportion can be used p_g:

$$d_{GU} = \frac{r_{GU}}{\sqrt{p_G\left(1-p_G\right)\left(1-r_{GU}^2\right)}}.$$

This process of converting, correcting, and then back-converting must also be done for the standard error. To avoid redundancy, we can incorporate each step into a single equation:

$$se(d_{GU}) = \frac{se(d_{gY}) \times r_{GU}}{r_{gY}\sqrt{\left(1 + d_{gY}^2 p[1-p]\right)^2 \left(d_{gY}^2 + \frac{1}{p_g(1-p_g)}\right) p_G(1-p_G)(1-r_{GU}^2)^3}}.$$

Applied Example in R

A researcher wants to compare the academic performance (measured by a standardized test) of people with and without Major Depressive Disorder (MDD). The researcher reports a classification reliability of $r_{gg'} = .80$ and a reliability of the standardized test as $r_{YY'} = .85$. The researcher than finds a standardized mean difference of $d_{gY} = .30$ favoring controls (i.e., controls had a higher average on the test). Using the `correct_d` function in the `psychmeta` package, we can obtain an unbiased estimate of the target standardized mean difference.

```r
library(psychmeta)

correct_d(correction = "meas",
          d = .30,   # observed standardized mean difference
          ryy = .85, # reliability of dependent variable
          rGg = sqrt(.80), # sqrt of classification reliability
          n1 = 100, # sample size in controls
          n2 = 100) # sample size in people with MDD
```

```
d Values Corrected for Measurement Error:
--------------------------------------------------------------------------------
  value CI_LL_95 CI_UL_95   n n_effective
1 0.366   0.0238    0.726 200         133
```

The corrected standardized mean difference is $d_{GU} = 0.37\,[0.02,\ .73]$.

7 Artificial Dichotomization

7.1 Introduction

Researchers occasionally split naturally continuous variables into two discrete groups to increase interpretability or conduct specific analyses (e.g., t-tests). However, artificially dichotomizing variables introduces measurement error variance thus attenuating effect size estimates Maxwell and Delaney (1993). The obvious solution to this problem is to simply not dichotomize variables, however if only summary data is available to us, then we may not have this luxury. Dichotomization can also be practical in some instances. For example, clinical disorder diagnoses such as generalized anxiety disorder, are examples of dichotomization where individuals are separated into either having the disorder or not even though individual differences in anxiety exist as a continuum.

7.2 Artificial Dichotomization Induced Measurement Error

Variables that are dichotomized contain measurement error. This can be demonstrated by the simple fact that dichotomized scores are not perfectly correlated with their underlying continuous scores. To demonstrate this, we can draw a sample of scores and then split the data into high and low scorers and then calculate the correlation coefficient between the two (see Figure 7.1).

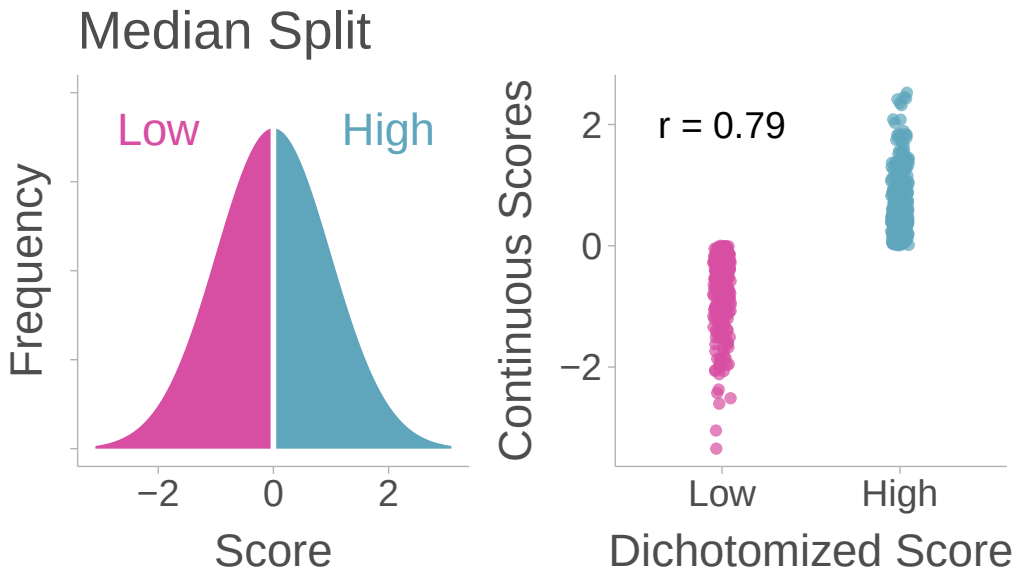


Figure 7.1: Visualizing the loss of precision when artificially dichotomizing. Left panel shows a normally distributed variable split (at the median/mean) into a high scoring group and a low scoring group.

Even with a perfectly reliable measure, dichotomization will introduce measurement error. Dichotomization occurs when data is split into two groups (low and high groups will be denoted as L and H, respectively) depending on whether they are above or below some cut-point C_X. We can define artificially dichotomized scores as,

$$X_D = \begin{cases} H, & \text{if } X \geq C_X \\ L, & \text{if } X < C_X \end{cases}$$

7.3 Correcting Correlations for Artificial Dichotomization

7.3.1 Defining the Target Quantity

We want to obtain the population correlation between continuous scores of the independent (X) and dependent variable (Y), ρ_{XY}.

There are two cases of dichotomization that may occur in a given study: the *univariate* case where only one variable (either X *or* Y) is dichotomized and the *bivariate* case where both variables are dichotomized. In either case, dichotomization will have a biasing effect on the study correlation coefficient. The study correlation on dichotomized data (using the bivariate case, $r_{X_D Y_D}$) can be modeled as a function of the population correlation on continuous scores (ρ_{XY}; i.e., the target), an artifact attenuation factor a, and sampling error, e,

$$r_{X_D Y_D} = a\rho_{XY} + e.$$

An unbiased estimate of ρ_{XY} can be calculated by dividing the study correlation by an estimate of the artifact attenuation factor, a,

$$r_{XY} = \frac{r_{X_D Y_D}}{a}.$$

7.3.2 Artifact Correction for Correlations

Correlations can suffer from dichotomization in one variable (univariate case) or both variables (bivariate case). This section will discuss the procedure for obtaining an unbiased estimate of the correlation coefficient for both cases. For a comparative visualization of a correlation on with no dichotomization, univariate dichotomization, and bivariate dichotomization, see Figure 7.2.

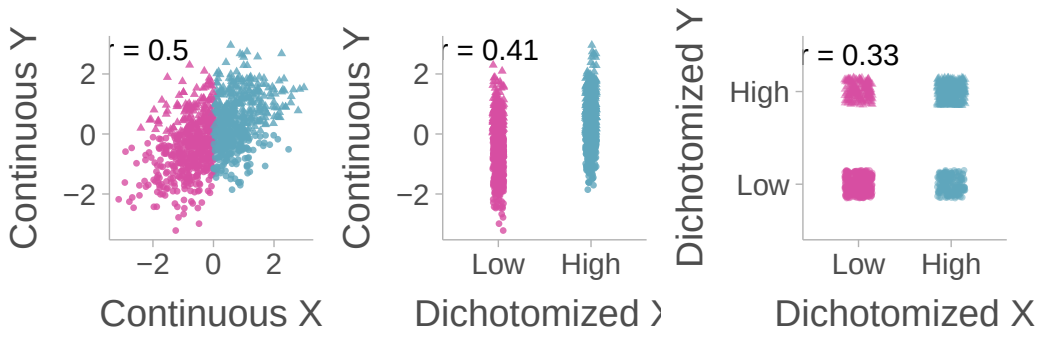


Figure 7.2: Scatter plots showing the correlation under dichotomization. The first panel (left to right) shows the correlation with no dichotomization (color and shapes of points denote where the split in the dichotomized cases will occur). The middle panel shows the univariate case where only the independent variable, X, is dichotomized. The last panel shows the bivariate case where both independent (X) and dependent (Y) variables are dichotomized.

The Univariate Case

In the simplest case of dichotomization, only one variable is dichotomized and the other is left continuous. A Pearson correlation computed between a dichotomous variable and a continuous variable is known as *point-biserial* correlation. However, if the variable is naturally continuous, we can estimate the correlation of the underlying continuous scores by computing a *biserial* correlation. If all we have access to is the dichotomized data, then we need to assume the shape of the underlying distribution, in this case, the biserial correlation assumes normality.

In the population, the study correlation $\rho_{X_D Y}$ is biased by some artifact biasing factor, a,

$$\rho_{X_D Y} = a\rho_{XY}.$$

To estimate the attenuation factor a, we must first figure out where the split of the data occured. To do this, we must first calculate the proportion of the sample in the assigned to the low or high scoring group:

$$p_X := p(X_i \geq C_X) = \frac{n_{\mathrm{H}}}{n_{\mathrm{H}} + n_{\mathrm{L}}}$$

or

$$p_X := p(X_i < C_X) = \frac{n_{\text{L}}}{n_{\text{H}} + n_{\text{L}}}.$$

Where n indicates the sample size within the low, L, and high, H, scoring groups.

> **Note**
>
> It will not matter whether you calculate the proportion of the sample in the high scoring group or the low scoring group for p_X. Once you decide on one, do not change it.

We can use the quantile function ($\phi^{-1}[\cdot]$, i.e., the inverse of the cumulative density of the standard normal distribution) to find where the split would have occured on a standard normal distribution, $s_X = \phi^{-1}[p_X]$. Using the location of the split on the standard normal, we can compute the artifact attenuation factor [an adaptation of equation 2, J. Hunter and Schmidt (1990)],

$$\hat{a} = \frac{\varphi\left(s_X\right)}{\sqrt{p_X(1 - p_X)}}. \tag{7.1}$$

Where $\varphi(\cdot)$ is the normal ordinate function (i.e., probability density function of a standard normal distribution). Figure 7.3 visually demonstrates how each of these relate to a standard normal distribution.

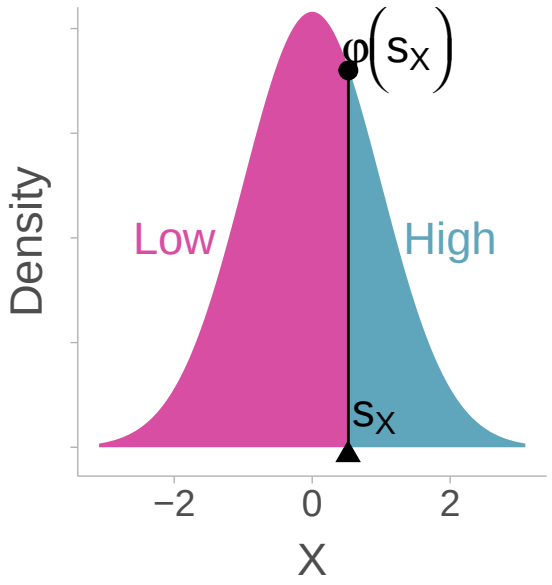


Figure 7.3: This figure shows a normal distribution of scores split into a high scoring and low scoring group. The cut-point of the standard normal distribution is computed with the quantile function, $\phi^{-1}[p_X]$. The ordinate of the normal distribution at that cut-point is calculated with the normal ordinate function, $\varphi\left(\phi^{-1}[p_X]\right)$.

In the case of a median split, where the cut-point would be placed at zero of a standard normal (splitting the distribution in equal halves), the attenuation factor would simplify to $\hat{a} = \frac{\varphi(0)}{\sqrt{.5(.5)}} = \frac{2}{\sqrt{2\pi}}$.

We can correct the study correlation using the estimated artifact factor, $\hat{a}$, therefore the full correction equation is,

$$r_{XY} = \frac{r_{X_DY}}{\hat{a}} = \frac{r_{X_DY}}{\left[\frac{\varphi(s_X)}{\sqrt{p_X(1-p_X)}}\right]}. \tag{7.2}$$

Converting to $\mathbf{r_{X_DY}}$ from means and standard deviations

It is common that studies do not report the correlation between dichotomized and continuous scores, r_{X_DY} (i.e., the point-biserial correlation), instead they may report the means and standard deviations instead. To obtain the point-biserial correlation, r_{X_DY}, we need the the mean of the high scoring group ($\overline{Y}_H$), mean of the low scoring groups ($\overline{Y}_L$), the standard deviation across all individuals (S_Y; not to be confused with the pooled/average standard deviation within each

group), and the sample sizes within each group (n_H and n_L) to calculate r_{X_DY},

$$r_{X_DY} = \frac{\overline{Y}_H - \overline{Y}_L}{S_Y}\sqrt{p_X(1-p_X)}.$$

If the standard deviations are reported only for within the high and low groups, you can estimate the total standard deviation with the following formula,

$$S_X = \sqrt{\frac{n_H S_{X_H} + n_L S_{X_L}}{n_H + n_L} + \frac{n_H n_L}{n_H + n_L}(\overline{X}_H - \overline{X}_L)^2}.$$

J. Hunter and Schmidt (1990) suggested that one should correct the standard error by dividing the uncorrected standard error by the artifact attenuation factor (see equation 6, J. Hunter and Schmidt 1990). However simulations have found that this computation does not work as well as Soper's exact method (Soper 1914; Jacobs and Viechtbauer 2017). Therefore the standard error of the corrected (biserial) correlation can be estimated with the following formula (equation 12, Jacobs and Viechtbauer 2017),

$$se(r_{XY}) = \sqrt{\frac{1}{n-1}\left(r_{XY}^4 + r_{XY}^2\left(\frac{p_X(1-p_X)s_X^2}{\varphi\left(s_X\right)^2} + \frac{2p_X - 1}{\varphi\left(s_X\right)} - \frac{5}{2}\right) + \frac{p_X(1-p_X)}{\varphi\left(s_X\right)^2}\right)}.$$

Soper (1914) also developed an approximation of the above formula,

$$se(r_{XY}) \approx \sqrt{\frac{1}{n-1}}\left(\frac{\sqrt{p_X(1-p_X)}}{\varphi\left(s_X\right)} - r_{XY}^2\right).$$

Applied Example in R

Let's say we want to assess the relationship between a sales person's score on a job knowledge test and their job performance estimated as the number of sales made per week. However the researchers of the study chose to dichotomize sales people into high sales performers and low sales performers by splitting the sample into two equally sized groups. They then reported the means and standard deviations of job knowledge test scores of both groups:

1. Low sales performers: Mean = 22 (SD = 4, n = 50)
2. High sales performers: Mean = 24 (SD = 4, n = 50)

To calculate the corrected correlation we can use `escalc()` function from the `metafor` package (Viechtbauer 2010). Using the argument `measure='RBIS'` will return the biserial correlation coefficient which is equivalent to the corrected correlation.

```r
library(metafor)

escalc(measure = 'RBIS',
       m1i = 24, # High performer mean
       m2i = 22, # Low performer mean
       sd1i = 4, # High performer SD
       sd2i = 4, # Low performer SD
       n1i = 50, # High performer sample size
       n2i = 50, # Low performer sample size
       var.names = c('rXY','se.2'),
       digits = 3)
```

```
    rXY  se.2
1 0.307 0.014
```

Therefore the estimated correlation on continuous scores is $r_{XY} = .31$.
If the study reported a Pearson correlation (point-biserial) between the dichotomized variable and the continuous variable of $r_{X_DY} = .245$, then

```r
library(psychmeta)

correct_r_dich(r = .245, # study point-biserial correlation
               px = .50, # proportion of sample in low or high group
               n = 100) # total sample size
```

```
  r_corrected var_e_corrected    n_adj
1    0.307062      0.01401901 59.51455
```

The corrected correlation for continuous scores is equal to the calculation using the means and standard deviations. Note that the estimated sampling variance of the corrected correlation, `se.2` (in the `psychmeta` example) and `var_e_corrected` (in the `metafor` example), are both equal to the squared corrected standard error $(se(r_{XY})^2)$.

The Bivariate Case

In some cases, both independent and dependent variables are dichotomized. A Pearson correlation calculated on these two dichotomized (binary) variables would be equal to the phi coefficient (or also known as Matthew's correlation coefficient) and we can denote it with our notation for dichotomized variables, $\rho_{X_DX_D}$. Dichotomized data can be structured in a contingency table (see Table 7.1).

Table 7.1: Contingency table.

	$X_D = \text{Low}$	$X_D = \text{High}$
$Y_D = \text{Low}$	n_{LL}	n_{HL}
$Y_D = \text{High}$	n_{LH}	n_{HH}

Figure 7.4 illustrates how this contingency table relates to an underlying continuous bivariate normal distribution.

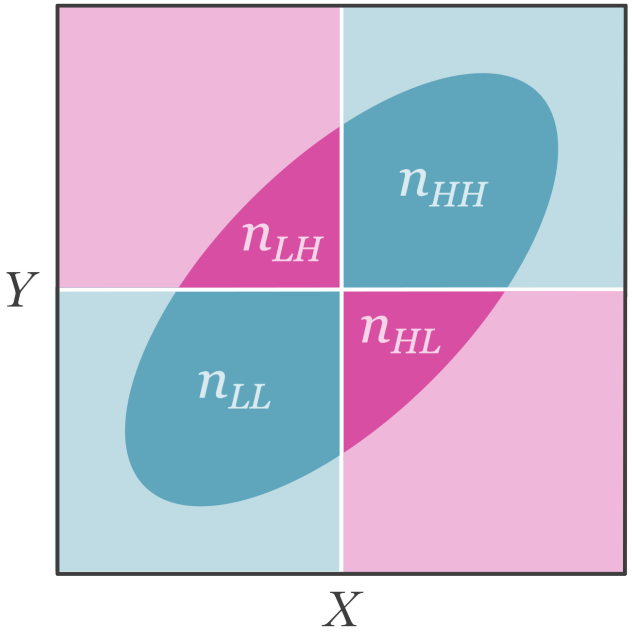


Figure 7.4: The ellipse indicates the bivariate normal distribution of X and Y with a strong positive correlation. If X and Y are positively correlated then we should more individuals populating the high-high and low-low cells rather than the high-low and low-high cells which can be seen by the area of the ellipse located in each quadrant.

The corrected correlation coefficient for two dichotomized variables is commonly referred to as the tetrachoric correlation coefficient. The tetrachoric correlation estimates the correlation on continuous scores assuming a bivariate normal distribution.

One of the difficulties of computing a dichotomization corrected (tetrachoric) correlation (r_{XY}) is that the relationship between binary variables are reported in different ways across studies, we will describe how to obtain a dichtomization corrected correlation in four different cases:

1. The full contingency table is provided, including the sample sizes for each cell.

2. The odds ratio is reported as well as the marginal proportions (proportions in low and high groups for each variable).
3. The Phi coefficient is reported.

7.3.2.0.1 * Case 1: Full contingency table is reported

If the full contingency table is reported, then the tetrachoric correlation can be calculated directly. Due to the complexity of the calculation, we will use R.

The `escalc()` function in the `metafor` package (Viechtbauer 2010) can take on values from a contingency table and compute a tetrachoric correlation using the `measure='RTET'` argument. The function uses the method described by Kirk (1973).

```
# Example Contingency Table
#    XL   XH
# YL 43   23
# YH 27   38

library(metafor)

escalc(measure = 'RTET',
       ai = 43,
       bi = 23,
       ci = 27,
       di = 38,
       var.names = c('rXY','se.2'))
```

```
     rXY   se.2
1 0.3637 0.0155
```

The results show a dichotomization corrected correlation of $r_{XY} = .36$ and an estimated sampling variance of $se_c^2 = .016$.

7.3.2.0.2 * Case 2: odds ratio is reported

If the odds ratio is all that is available, then we can use the tetrachoric correlation approximation described by Bonett and Price (2005). Using the estimated odds ratio ($OR = (n_{HH}n_{LL})/(n_{HL}n_{LH})$) and the marginal proportions (p_X and p_Y; for this case, both proportions should be with respect to the high scoring group) we can approximate the dichotomization corrected (tetrachoric) correlation (see equation 4, Bonett and Price 2005),

$$r_{XY} \approx \cos\left(\frac{\pi}{1 + OR^{\Omega}}\right)$$

Where Ω is

$$\Omega = 1 - \frac{|p_X - p_Y|}{5} - \left[\frac{1}{2} - \min(p_X, p_Y)\right]^2.$$

Note that $\min(p_X, p_Y)$ is the smallest marginal proportion. The standard error of the estimated correlation can be computed with the following formula [see equation 9, Bonett and Price (2005); note this equation is slightly changed in order to account for the fact that we only have access to the marginal sample sizes and proportions]

$$se(r_{XY}) = \frac{\pi \times \Omega \times OR^{\Omega} \times \sin\left(\frac{\pi}{1+OR^{\Omega}}\right)}{\left(1 + OR^{\Omega}\right)^2} \times \sqrt{\frac{4}{n}}$$

Where n is the total sample size. Using base R, we can convert the odds ratio to a tetrachoric correlation.

```
pX <- .4 # proportion of individuals in high group in XD
pY <- .5 # proportion of individuals in high group in YD
n <- 150 # total sample size
OR <- 1.43 # odds ratio
omega <- 1 - abs(pX-pY) / 5 - (1/2 - min(c(pX,pY)))^2

# calculate corrected correlation
rXY <- cos(pi/(1+OR^omega))

# calculate standard error
se <- pi*omega*OR^omega*sin(pi/(1+OR^omega)) / (1+OR^omega)^2 * sqrt(4/n)

cbind(rXY, se)
```

```
          rXY      se
[1,] 0.2665276 0.11637
```

7.3.2.0.3 * Case 3: Correlation between dichotomized variables

A Pearson correlation coefficient calculated between two binary variables is most commonly referred to as a Phi coefficient or Matthew's correlation coefficient. However, the Phi coefficient underestimates the correlation on the underlying continuous scores (assuming bivariate normality). Therefore we can

approximate the correlation on continuous scores by using a similar correction to the univariate case. We can define an artifact attenuation factor that is similar to Equation 7.1, but with the added attenuation of Y_D,

$$\hat{a} = \left[\frac{\varphi\left(\phi^{-1}[p_X]\right)}{\sqrt{p_X(1-p_X)}}\right] \times \left[\frac{\varphi\left(\phi^{-1}[p_Y]\right)}{\sqrt{p_Y(1-p_Y)}}\right].$$

Therefore we can correct $r_{X_D Y_D}$ for dichotomization in both variables by dividing by the attenuation factor,

$$r_{XY} = \frac{r_{X_D Y_D}}{\hat{a}} = \frac{r_{X_D Y_D}}{\left[\frac{\varphi(\phi^{-1}[p_X])}{\sqrt{p_X(1-p_X)}}\right] \times \left[\frac{\varphi(\phi^{-1}[p_Y])}{\sqrt{p_Y(1-p_Y)}}\right]}$$

This correction is only an approximation, however it performs fairly well when the correlation is below .8 (J. Hunter and Schmidt 1990). The standard error of the corrected correlation can be calculated similarly (adaptation of equation 9, J. Hunter and Schmidt 1990),

$$se(r_{XY}) = \frac{se(r_{X_D Y_D})}{\left[\frac{\varphi(\phi^{-1}[p_X])}{\sqrt{p_X(1-p_X)}}\right] \times \left[\frac{\varphi(\phi^{-1}[p_Y])}{\sqrt{p_Y(1-p_Y)}}\right]}$$

Using the `correct_r_dich()` function in the `psychmeta` package (Dahlke and Wiernik 2019), we can correct the observed study correlation for dichotomization.

```
library(psychmeta)

correct_r_dich(r = .20,  # observed study correlation
               px = .50, # proportion of sample in high group of XD
               py = .50, # proportion of sample in high group of YD
               n = 100)  # sample size
```

```
  r_corrected var_e_corrected    n_adj
1   0.3141593      0.02296926 36.36678
```

The output shows the corrected correlation to be $r_{XY} = .31$ and it's estimated sampling variance $se_c^2 = .023$.

7.4 Correcting Standardized Mean Differences for Artificial Dichotomization

7.4.1 Defining our Target Quantity

The target quantity is the standardized mean difference between groups of a naturally continuous variable. Our target can thus be defined as the population standardized mean difference between groups A and B on continuous scores of the dependent variable, δ_{gY}. For a given study the dichotomized standardized mean difference (d_{gY_D}) can be defined as

$$d_{gY_D} = a\delta_{gY} + e$$

Therefore to obtain an unbiased estimate of the target quantity, we can correct the observed study standardized mean difference by dividing by the attenuation factor,

$$d_{gY} = \frac{d_{gY_D}}{\hat{a}}$$

7.4.2 Artifact Correction for Standardized Mean Differences

The simplest way to correct for dichotomization in a standardized mean difference is to first convert the observed d value of the dichotomized dependent variable to a correlation coefficient. When converting to a correlation coefficient, it's important to note the binary nature of both variables, leading us to estimate the phi coefficient rather than the point-biserial correlation that we would be estimating if the dependent variable was continuous. To calculate the phi coefficient from a d value we can use the proportion of group membership in group A or group B (p_g; it does not matter which one is chosen, as long as it is consistent for every instance of p_g),

$$r_{gY_D} = \frac{d_{gY_D}}{\sqrt{d^2_{gY_D} + \frac{1}{p(1-p)}}}$$

We can then correct the correlation similar to how we did in Section 7.3.2,

$$r_{gY} = \frac{r_{gY_D}}{\left[\frac{\varphi(\phi^{-1}[p_Y])}{\sqrt{p_Y(1-p_Y)}}\right]}.$$

Then we can convert the corrected correlation back into a standardized mean difference,

$$d_{gY} = \frac{r_{gY}}{\sqrt{p_g\left(1-p_g\right)\left(1-r_{gY}^2\right)}}.$$

Where d_{gY} is our corrected correlation (i.e., the estimated correlation on continuous variables). The observed sampling variance must also be corrected using the same three step procedure. For simplicity, we will consolidate this into one formula,

$$se(d_{gY}) = \frac{se(d_{gY_D}) \times r_{XY}}{r_{x_D Y_D}\sqrt{\left(1 + d_{gY_D}^2 p_g[1-p_g]\right)^3 (1-r_{gY}^2)^3}}.$$

Obtaining Standardized Mean Difference from Odds Ratio

In most cases, difference in dichotomized outcomes between two groups is unlikely to be reported as a standardized mean difference, instead it will be more commonly reported as an odds ratio (OR). The odds ratio is asymmetric about 1 (i.e., the null), but we can make it symmetric by log transforming it ($\log(OR)$). The standard error of the log odds ratio can be defined as,

$$se(\log(OR)) = \sqrt{\frac{1}{n_{AH}} + \frac{1}{n_{AL}} + \frac{1}{n_{BH}} + \frac{1}{n_{BL}}}$$

The equation above requires the full contingency table to compute. From there we can use the cox-logit method to convert the odds ratio to a standardized mean difference (Cox 1989; Haddock, Rindskopf, and Shadish 1998). Them method is quite simple as it just divides the log odds ratio by 1.65,

$$d_{gY} = \frac{\log(OR)}{1.65} \tag{7.3}$$

and the corresponding sampling variance of the d value is,

$$se(d_{gY}) = \frac{se(\log(OR))}{1.65}.$$

Applied Example in R

Let's consider a hypothetical scenario where we want to examine the relationship between caffeine consumption and the occurrence of heart palpitations in a population. Our target quantity in this case is the standardized mean difference of coffee consuption between people with and without

heart palpitations. The variable of interest, caffeine consumption, is continuous (measured in milligrams per day). However, the researcher decides to dichotomize this variable into two groups: "High Caffeine Consumers" and "Low Caffeine Consumers."
Suppose we have a sample of 500 individuals, and we dichotomize their caffeine consumption into "High Caffeine Consumers" (more than 250mg per day) and "Low Caffeine Consumers" (less than or equal to 250mg per day). We also observe the occurrence of heart palpitations in these individuals.

	Heart Palpitations: Yes	Heart Palpitations: No
High Consumers	60	140
Low Consumers	20	280

We can calculate the dichotomization corrected standardized mean difference by calculating the log odds ratio with `escalc()` function and then applying Equation 7.3 to estimate d_{gY}.

```
# calculate log-odds ratio
OR <- escalc(measure = 'OR',
             ai = 60,
             bi = 140,
             ci = 20,
             di = 280,
             var.names = c('logOR','se.2'))

# convert to standardized mean difference
dgY <- OR$logOR / 1.65
se <- sqrt(OR$se.2) / 1.65

# print results
cbind(dgY,se)
```

```
         dgY        se
[1,] 1.085915 0.1685905
```

We can see that the standardized mean difference is estimated to be $d_{gY} = 1.09$ and the corrected standard error is $se_c = .169$

8 Scale Coarseness

8.1 Introduction

Scale coarseness describes a situation where a variable that is naturally continuous (e.g., happiness) is binned into discrete values (e.g., happiness measured on a scale of 1-10). This situation is quite common in the social and psychological sciences where Likert items or dichotomous yes/no responses are aggregated to form a coarse total score for a naturally continuous construct. When coarseness is present, measurement error is introduced into the observed scores and those scores lose information. Unlike dichotomization, coarseness is an artifact that occurs due to the design of the study rather than during the analysis phase (Aguinis, Pierce, and Culpepper 2009). Particularly, dichotomization occurs after scores are obtained (e.g., splitting a group into high scorers and low scorers), whereas coarseness occurs as an artifact of the measurement procedure itself. The primary issue with coarseness is that it limits the set of possible values a score can be which introduces error when the variable is naturally continuous. This can be visualized by correlating coarse scores with their underlying continuous scores (see Figure 8.1). You will notice that the correlation between coarse and continuous scores is not perfect, indicating that the coarse scores do not perfectly capture the underlying continuous scores.

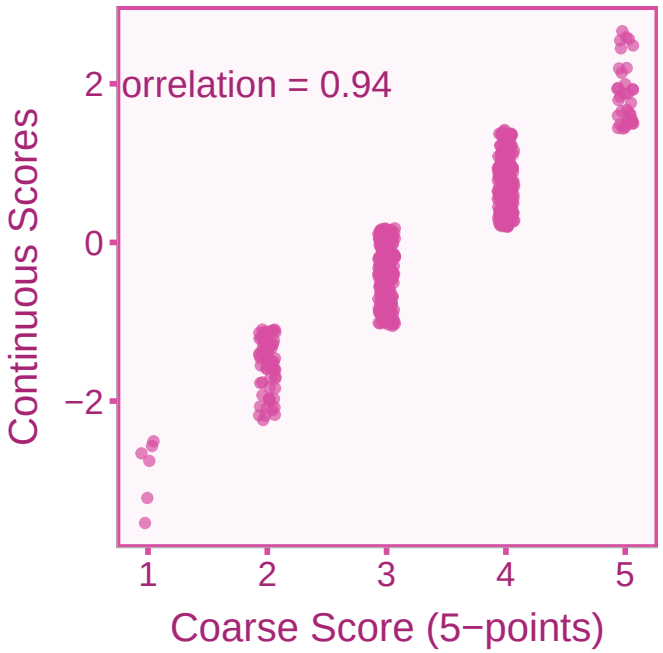


Figure 8.1: Scatterplot showing the correlation between coarse scores (on a 5-point scale) and the underlying continuous scores.

8.2 Correcting for Coarseness in Correlations

8.2.1 Defining our Target Quantity

Our quantity of interest is the population correlation, ρ, between continuous independent variable, X, and continuous dependent variable, Y. However, in a given study the measurement procedure may produce coarse scores for X and Y. We will denote coarse scores with the subscript C We can model the relationship between the observed sample correlation on coarse scores and the true population correlation,

$$r_{X_C Y_C} = a\rho_{XY} + e.$$

Where a is our artifact attenuation factor and e is our sampling error term. We can obtain an unbiased estimate of the continuous population correlation, ρ_{XY}, by dividing the observed correlation by an estimate of the attenuation factor,

$$r_{XY} = \frac{r_{X_C Y_C}}{\hat{a}}.$$

8.2.2 Artifact Correction for Coarseness

Correlating a coarse score with another variable will cause attenuation of the correlation (MacCallum et al. 2002). Furthermore if we correlate a coarse score with another coarse score than we will observe even more attenuation (see Figure 8.2). There are two cases that we can run into: 1) the univariate case where only one variable is coarse and 2) the bivariate case where both variables are coarse.

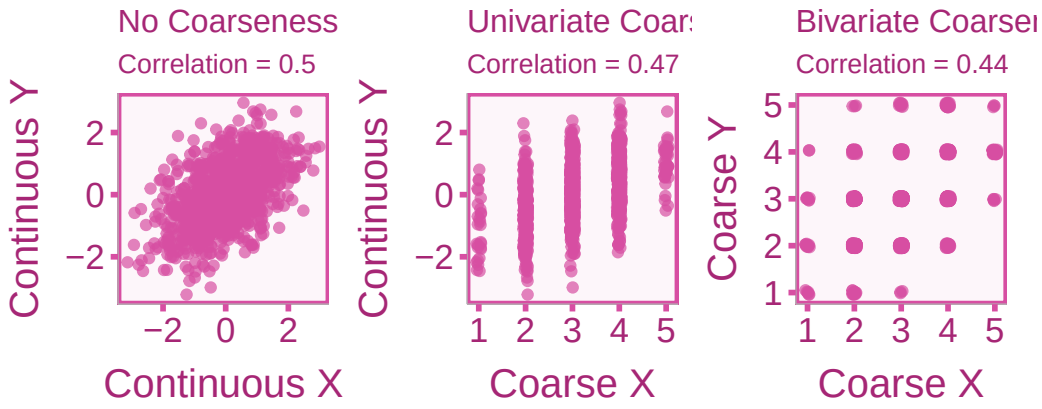


Figure 8.2: First plot (left to right) shows both variables as continuous and normal. The second plot shows coarseness (5-point scale) only on X, leaving Y continuous ($r_{X_C Y} = .47$). Last plot shows coarseness on both variables ($r_{X_C Y_C} = .47$).

To correct the correlation between coarse scores for X and Y, we need to know the correlation between coarse scores and their underlying continuous scores (r_{XX_C} or r_{YY_C}). The calculation of the correlation will require us two important assumptions:

1) The shape of the underlying distribution (i.e., normal or uniform).
2) The intervals between scale-points are equal.

Based on these assumptions, Peters and Voorhis (1940) constructed a table of correlations between coarse and continuous scores that is also reported more recently by Aguinis, Pierce, and Culpepper (2009). Table 8.1 is adapted from Peters and Voorhis (1940) and displays the correlation values for uniform and normal distributions for a given number of scale points. For the normal distribution correction, its been shown that even in cases of extreme skew, these correction factors perform well (Wylie 1976).

Table 8.1: Correlations between continuous and coarse scores (r_{XX_C}) from Peters and Voorhis (1940)

Scale Points	Continuous-Coarse score Correlation (normal)	Continuous-Coarse score Correlation (uniform)
2	.816	.866
3	.859	.943
4	.916	.968
5	.943	.980
6	.960	.986
7	.970	.990
8	.977	.992
9	.982	.994
10	.985	.995
11	.988	.996
12	.990	.997
13	.991	.997
14	.992	.997
15	.994	.998

The correlations between coarse and continuous scores (r_{XX_C} and r_{YY_C}) can be used to correct the correlation coefficient

$$r_{XY} = \frac{r_{X_C Y_C}}{r_{XX_C} r_{YY_C}}.$$

Where $r_{XX'}$ and $r_{YY'}$ are the appropriate correction factors from the table for x and y. We must also adjust the sampling variance as well,

$$\sigma^2_{\varepsilon_c} = \frac{\sigma^2_{\varepsilon_o}}{a_x^2 a_y^2}.$$

8.2.3 Correcting Correlations in R

To correct scale coarseness in R, we can use the table in Section 8.2.2. Lets start by simulating a coarse data (5 scale-points for x and 7 scale-points for y) set of 500 individuals with a true population correlation of $\rho = .50$.

Applied Example in R

Imagine that a researcher wants to relate depression and age. They collect a sample of 1,000 people from the general population and administer a very quick survey. Depression is measured based on a single item from the patient health questionnaire (PHQ, Kroenke, Spitzer, and Williams 2003) and age is measured in 5 age ranges:
Over the last 2 weeks, how often have you been bothered by the following problems?
What is your age?

1. Not at all
2. Several days
3. More than half the days
4. Nearly every day

1. 1-20 years
2. 21-40 years
3. 41-60 years
4. 61-80 years
5. 81+ years

Let's say we obtain a correlation of $r_{X_C Y_C} = -.20$. Since the correlation is computed on coarse scores, it is likely attenuated relative to the correlation on each variables continuous underlying scores. Therefore we can use the `correct_r_coarseness()` function in the `psychmeta` package (Dahlke and Wiernik 2019) to correct the correlation.

```
library(psychmeta)

correct_r_coarseness(r = -.20, # observed correlation
                     kx = 5, # 5 age range bins
                     ky = 4,  # 4 PHQ item options
                     n = 1000, # sample size
                     dist_x = "unif", # assumed X distribution
                     dist_y = "norm") # assumed Y distribution
```

```
  r_corrected var_e_corrected    n_adj
1  -0.2230339     0.001147251 788.0867
```

We see a slight increase in the magnitude of the correlation with the estimated correlation on continuous scores being $r_{XY} = -.22$ and a standard error of $se(r_{XY}) = \sqrt{.0011} = 0.034$.

8.3 Correcting for Coarseness in *d* values

8.3.1 Defining our Target Quantity

Our quantity of interest is the population standardized mean difference, δ_{gY}, between groups A and B on continuous variable, Y. We can define the standardized mean difference on coarse scores (d_{gY_C}) as,

$$d_{gY_C} = a\delta_{gY} + \varepsilon.$$

Where a is our coarseness biasing factor and ε is our sampling error term. Ultimately, we can obtain an unbiased estimate of the population standardized mean difference on continuous scores by correcting the observed standardized mean difference as follows,

$$d_{gY} = \frac{d_{gY_C}}{\hat{a}}.$$

8.3.2 Artifact Correction for Coarseness

To correct a standardized mean difference for coarseness in dependent variable, we can use the correlation between coarse scores and continuous scores from Table 8.1,

$$d_{gY} = \frac{d_{gY_C}}{r_{YY_C}}.$$

Where r_{YY_C} is the appropriate correction factor from the table. We must also adjust the standard error,

$$se(d_{gY}) = \frac{se(d_{gY_C})}{r_{YY_C}}.$$

Applied Example in R

Let's say that a researcher wants to investigate gender differences in depression. The researcher administers a survey to a sample of men and women from the general population. Depression is measured based on a single item from the patient health questionnaire (PHQ, Kroenke, Spitzer, and Williams 2003):
Over the last 2 weeks, how often have you been bothered by the following problems?

1. Not at all

2. Several days
3. More than half the days
4. Nearly every day

Let's say we obtain a standardized mean difference of $d_{gY_C} = .25$, slightly favoring women. Since there is currently no `correct_d_coarseness()` function in `psychmeta`, we can simply correct the correlation with base R. again however we will need t in the `psychmeta` package (Dahlke and Wiernik 2019) to correct the correlation.

```
library(psychmeta)

dgYc <- .25 # standardized mean difference on coarse scores
se.dgYc <- .10 # standardized mean difference on coarse scores
rYYc <- .916

dgY <- dgYc / rYYc # correct d
se.dgY <- se.dgYc / rYYc # correct se

# print results
cbind(dgY, se.dgY)
```

```
          dgY    se.dgY
[1,] 0.2729258 0.1091703
```

We see a slight increase in the magnitude of the correlation with the estimated correlation on continuous scores being $d_{gY} = .27$ and a standard error of $se(d_{gY}) = .109$.

9 Direct Selection

9.1 Introduction

Direct selection occurs when subjects are explicitly selected based on some eligibility criterion on the variables of interest (rather than a third variable). Range restriction is a form of selection bias that describes a situation where there is less variation in our sample then there is in the population. Whereas range enhancement indicates that there is *more* variation in a sample then there is in the population. Direct range restriction/enhancement biases the variances and effect size estimates.

9.2 An Example of Direct Range Restriction

Imagine a tech company that wants to assess the correlation between years of experience and programming proficiency for their software engineers. They have two primary divisions: Division A and Division B. Division A primarily hires entry-level software engineers, with less than 3 years of experience. Division B, on the other hand, hires experienced software engineers with more than 3 years of experience. The company decides to conduct a study to assess the correlation between years of experience and programming proficiency. However, they only collect data from Division A due to logistical reasons, assuming that the relationship found there would be represent the entire company. In this scenario, direct range restriction occurs because the sample used for the study (Division A) represents a narrow range of years of experience (0-3 years) compared to the broader range present in the entire company (0+ years). Consequently, the standard deviation will be smaller in the sample then it would if we had sampled from the entire company. As we will see in later sections of this chapter, the observed correlation between years of experience and programming proficiency would be attenuated, underestimating the true correlation.

9.3 A Direct Selection Function

A selection function described here is a type of indicator function that represents the mechanism of which observations are selected into a given sample. In the case of *direct* selection, the selection will be a function of the of the variable of interest, X. We will denote a selection function as $\mathcal{S}(X)$. The output of the selection function will be a binary: either the individual is selected $\mathcal{S}(X) = 1$ or rejected $\mathcal{S}(X) = 0$ (see Figure 9.1).

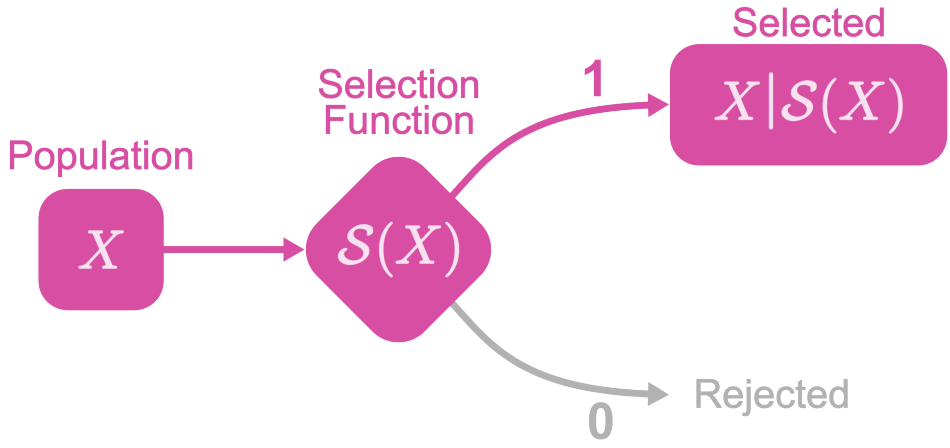


Figure 9.1: Diagram illustrating the direct selection function. The variable, X, in the population is selected based on a function of itself. Therefore the variable under selection is conditional on the selection procedure $X|\mathcal{S}(X)$.

The functional form of the selection function is going to be described by some inequality of X. The inequality will be dependent on the research context (e.g., the sampling procedure).

A Simple Example

Let's define the selection function as

$$\mathcal{S}(X) = \begin{cases} 1 & \text{if } X > 3 \text{ and } X < 8 \\ 0 & \text{if } X \leq 3 \text{ and } X \geq 8 \end{cases}$$

then let's say X contains the following observations,

$$X = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]$$

the selection function would then return,

$$\mathcal{S}(X) = [0, 0, 0, 1, 1, 1, 1, 0, 0, 0]$$

therefore X given selection on X would be,

$$X|\mathcal{S}(X) = [4, 5, 6, 7]$$

Direct selection on X will have a fairly straight forward effect on the distribution of X. The probability

density function of X ($f(X)$) under direct selection can be formally defined as,

$$f\left(X|\mathcal{S}(X)\right) = f(X)\mathcal{S}(X)$$

We can visualize the distribution under direct selection in Figure 9.2. Note that this is a simple case where selection is done on a single variable (i.e., univariate direct selection). In practice, selection is often a function of multiple variables. For instance, bivariate direct selection occurs when the selection is a function of both variables of interest, $\mathcal{S}(X, Y)$.

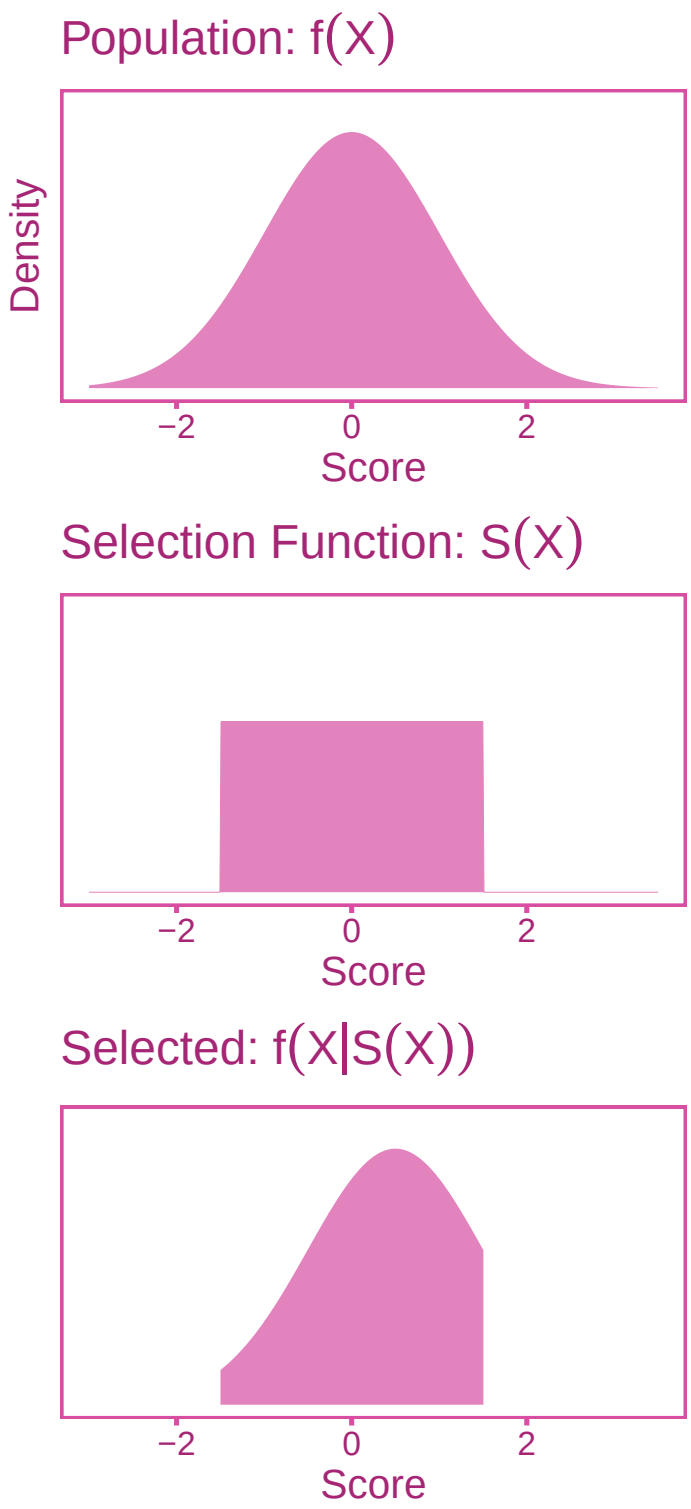


Figure 9.2: Visualizing direct selection with an example probability density and selection function. Top panel shows a normally distribution of X within the population. The middle panel shows the selection function. The selection function is based on the inequality, $-1.0 < X < 1.5$. The bottom panel shows the distribution of X after selecting on X (i.e., $X|\mathcal{S}(X)$).

9.4 Quantifying Direct Selection-Induced Restriction/Enhancement with the *u*-ratio

The distribution of scores in the population of individuals will exhibit a greater (or lesser) degree of variability compared to the population that meets the selection criterion. Therefore the standard deviation of scores in the target population (σ_X) will differ from that of the population under direct selection on X ($\sigma_{X|\mathcal{S}(X)}$). To index the difference between the two standard deviations, we can calculate the u-ratio Wiernik and Dahlke (2020). The u-ratio is defined as the ratio between the standard deviations of the population ($\sigma_{X|\mathcal{S}(X)}$) under selection and the target population (σ_X) such that,

$$\upsilon_X = \frac{\sigma_{X|\mathcal{S}(X)}}{\sigma_X}$$

Where υ_X denotes the population u-ratio. The u-ratio in cases of range restriction will exist in the interval (0–1). Conversely, when the u-ratio is greater than 1 it is indicative of range enhancement. For a sample, the u-ratio is calculated from sample standard deviations,

$$u_X = \frac{S_{X|\mathcal{S}(X)}}{S_X}$$

Where u_X denotes the sample u-ratio. The target population standard deviation is often quite difficult to estimate since generally we do not have access to an estimate of the population of interest. However, the unrestricted standard deviation can be estimated from some reference or norm study that is representative of the population of interest. This often comes in the form of standardization samples or norm samples (obtained from test manuals) if the population of interest is the general population. For example, the distribution full-scale IQ scores derived from the Wechsler Adult Intelligence Test has a standard deviation of 15 in the US population (Wechsler 2008). We can use this estimate as the standard deviation for the target population. Lets say we select a sample from members of Mensa, a high IQ society, where members are specifically selected on the basis high IQ scores. If the standard deviation of Mensa members is 5, then the u-ratio would be,

$$u_x = \frac{S_{X|\mathcal{S}(X)}}{S_X} = \frac{5}{15} = .33\,.$$

If an estimate of the population standard deviation is not readily available, then a reliability coefficient from the reference sample and the sample under selection can be used to estimate the u-ratio,

$$u_X = \sqrt{\frac{1 - r_{XX'}}{1 - r_{XX'|\mathcal{S}(X)}}}.$$

Where $r_{XX'|\mathcal{S}(X)}$ and $r_{XX'}$ are the reliability estimates within the selected and reference sample respectively.

9.5 Correcting Correlations for Direct Range Restriction

9.5.1 Defining our Target Quantity

We want to estimate the correlation in the target population between true scores of the independent (T) and dependent variable (U). Within a study that suffers from direct selection and measurement error, the observed score correlation will be biased relative to our target true score population correlation, ρ_{TU}. We can model observed score correlations under univariate direct selection as,

$$r_{XY|\mathcal{S}(X)} = a\rho_{TU} + e$$

or under bivariate direct selection,

$$r_{XY|\mathcal{S}(X,Y)} = a\rho_{TU} + e.$$

Where a is the artifact attenuation/inflation factor and e is the sampling error term. In either case, an unbiased estimate of the total population true score correlation can be estimated by dividing by an estimate of a,

$$r_{XY|\mathcal{S}(X)} = a\rho_{TU} + e$$

or

$$r_{XY|\mathcal{S}(X,Y)} = a\rho_{TU} + e.$$

9.5.2 Artifact Correction for Correlations

The Univariate Case

Range restriction (or enhancement) in either the independent or dependent variable will induce bias into the correlation coefficient. Let us consider a case where a study directly selects on X and not Y. It is important to note, that if there is direct selection one of the two variables, then there will be indirect selection in the other variable too if the two are correlated. This would suggest that if $u_X \neq 1$ and $\rho_{XY} \neq 0$ then $u_Y \neq 1$. Lets visualize the correlation between independent (X) and dependent (Y) variables under univariate direct range restriction (see Figure 9.3) by only selecting individuals above some cut off such that,

$$\mathcal{S}(X) = \begin{cases} 1 & \text{if } X \geq -.50 \\ 0 & \text{if } X < -.50 \end{cases}$$

The scores of individuals that have been selected will show less variance than the population pool. Specifically, the scenario below shows a u-ratio of $u_X = 0.69$ in the independent variable. We see in Figure 9.3 that the correlation in the restricted scores ($r_{XY|\mathcal{S}(X)}$) are attenuated relative to the unrestricted correlation (r_{XY}, indicative of $a < 1$).

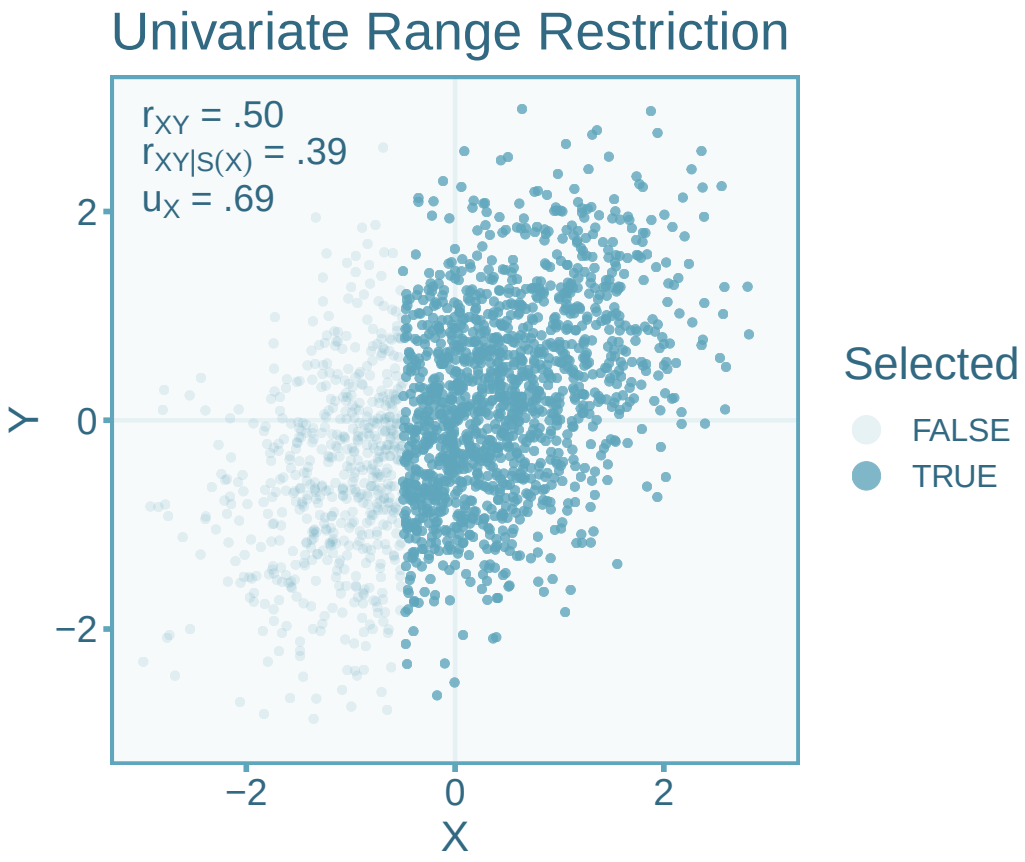


Figure 9.3: Scatterplot showing a correlation between X and Y under univariate direct range restriction. Dark blue dots indicate the selected sample and the transparent dots indicate the rejected sample.

We can also visualize what happens to the correlation when the range is enhanced. Enhancement can be accomplished by selecting individuals at the ends of the distribution (Taylor and Griess 1976). Therefore we can define the selection function for Figure 9.4 as,

$$\mathcal{S}(X) = \begin{cases} 1 & \text{if } X \leq -.50 \text{ or } X \geq .50 \\ 0 & \text{if } X > -.50 \text{ or } X < .50 \end{cases}$$

In the visualization below, we see an opposite effect on the correlation, that is, an inflation of the unrestricted correlation rather than an attenuation (indicating $a > 1$) like we see under range restriction. The scenario below has a u-ratio $u_X = 1.26$ in the independent variable.

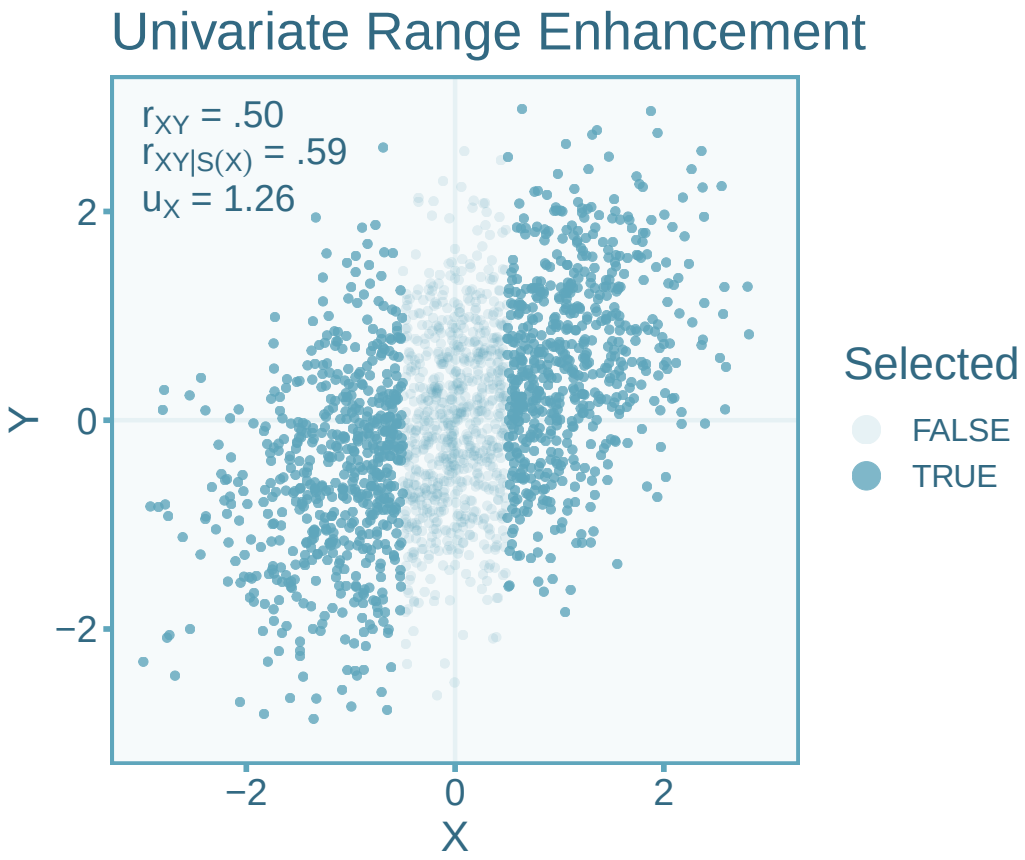


Figure 9.4: Scatterplot showing a correlation between X and Y under univariate direct range enhancement. Dark blue dots indicate the selected sample and the transparent dots indicate the rejected sample.

It starts to become apparent that if $u_X > 1$ (i.e., $S_X > S_{X|\mathcal{S}(X)}$) the observed correlation is inflated and when $u_X < 1$ it becomes attenuated (Sackett and Yang 2000).

The attenuation/inflation of the correlation is dependent on the magnitude of the correlation, this is due to the fact that selection occurs on X *and* X is correlated with Y, there will also be indirect range restriction in Y. Therefore unlike other artifacts that have been discussed so far, range restriction/enhancement depends not only on the artifact value (i.e., the u-ratio), but also on the restricted correlation (J. E. Hunter, Schmidt, and Le 2006). The attenuation/inflation factor can be estimated as (adapted from equation 4, J. E. Hunter, Schmidt, and Le 2006),

$$\hat{a} = u_X \sqrt{1 + r^2_{XY|\mathcal{S}(X)} \left(\frac{1}{u_X} - 1 \right)}$$

A bias correction formula for univariate direct selection was first developed by Pearson (1903) and also provided more recently by J. E. Hunter, Schmidt, and Le (2006) and Wiernik and Dahlke (2020). To correct for the systematic bias in correlations, we can divide the correlation under selection by the attenuation factor,

$$r_{XY} = \frac{r_{XY|\mathcal{S}(X)}}{\hat{a}} = \frac{r_{XY|\mathcal{S}(X)}}{u_X \sqrt{1 + r^2_{XY|\mathcal{S}(X)} \left(\frac{1}{u^2_X} - 1 \right)}}. \tag{9.1}$$

Where the standard error of the corrected correlation is,

$$se(r_{XY}) = \frac{se\left(r_{XY|\mathcal{S}(X)}\right)}{\hat{a}}. \tag{9.2}$$

If we want to also correct for measurement error in both samples, then we can also incorporate the reliability into these equations. Note that the following equations will incorporate the reliability within the selected sample ($r_{XX'|\mathcal{S}(X)}$) rather than the reference sample ($r_{XX'}$). If the reliability coefficient comes from the reference sample, then we can estimate the selected (restricted or enhanced) sample reliability with the corresponding u-ratio,

$$r_{XX'|\mathcal{S}(X)} = 1 - \frac{1 - r_{XX'}}{u^2_X}. \tag{9.3}$$

Then we can use the reliability and u-ratios simultaneously to calculate a new attenuation/inflation factor that accounts for both bias in measurement error and direct selection,

$$\hat{a} = u_X \sqrt{1 - u^2_X (1 - r_{XX'|\mathcal{S}(X)})} \sqrt{r_{YY'|\mathcal{S}(X)} + r^2_{XY|\mathcal{S}(X)} \left(\frac{1}{u^2_X} - 1 \right)}.$$

In the following equation to obtain an unbiased estimate of the true score unrestricted population correlation (adapted from table 3, Wiernik and Dahlke 2020),

$$r_{TU} = \frac{r_{XY|\mathcal{S}(X)}}{\hat{a}} = \frac{r_{XY|\mathcal{S}(X)}}{u_X\sqrt{1-u_X^2(1-r_{XX'|\mathcal{S}(X)})}\sqrt{r_{YY'|\mathcal{S}(X)} + r^2_{XY|\mathcal{S}(X)}\left(\frac{1}{u_X^2}-1\right)}}$$

Where the standard error can be corrected similarly,

$$se(r_{TU}) = \frac{se(r_{XY|\mathcal{S}(X)})}{\hat{a}}$$

There are three important things to note about the equations in this section: 1) these corrections assume linearity and homoskedasticity in the target population population, 2) all these corrections show selection on the independent variable, X, however it does not matter whether univariate selection is on X or Y, the corrections can be applied in the same fashion (just remember to flip X and Y and vice versa in the equations), 3) The corrections assume that any range restriction/enhancement observed in the non-selection variable (in our example this would be Y) is mediated by the range restriction/enhancement in the variable under selection (i.e., X).

Applied Example in R

To continue with our example of Mensa members (a high IQ society), let's say a researcher wants to know the correlation between IQ and test anxiety. The researcher conduct's a study on a sample of 100 Mensa members and finds a sample mean IQ of 135 and a standard deviation of $S_{X|\mathcal{S}(X)} = 5$. Since the mean IQ in the general population is 15, the u-ratio is computed as $u_X = .333$. The researcher then finds a correlation of $r_{XY|\mathcal{S}(X)} = -.15$ between IQ test scores and test anxiety. The reliability of the IQ test and the anxiety measure is $r_{XX'|\mathcal{S}(X)} = .85$ and $r_{YY'|\mathcal{S}(X)} = .80$, respectively.
In R, we can correct the correlation by using the `correct_r` function in the `psychmeta` package (Dahlke and Wiernik 2019). The `correction = 'uvdrr_x'`

```
library(psychmeta)

correct_r(correction = 'uvdrr_x',
          rxyi = -.15,  # restricted correlation
          ux = .333,    # u ratio of IQ scores
          rxx = .85,    # reliability of IQ scores
          ryy = .80,    # reliability of test anxiety measure
          n = 100)    # sample size
```

```
Correlations Corrected for Measurement Error and Univariate Direct Range Restriction:
--------------------------------------------------------------------------------
   value CI_LL_95 CI_UL_95   n n_effective
1 -0.459    -0.78     0.16 100        10.6
```

The true score correlation in the target population is estimated to be $r_{TU} = .46\,[-.16,\ .78]$.

The Bivariate Case

Bivariate direct range restriction/enhancement occurs when selection occurs on both variables of interest, therefore the selection function will be a function of X *and* Y. Let's visualize the correlation between independent (X) and dependent (Y) variables under bivariate range restriction by only selecting individuals above some cut off point for both X and Y (see Figure 9.5). For this example the selection function is

$$\mathcal{S}(X, Y) = \begin{cases} 1 & \text{if } X \geq -.5 \text{ and } Y \geq -.5 \\ 0 & \text{if } X < -.5 \text{ and } Y < -.5 \end{cases}$$

The scores of individuals that have been selected will show less variance than the entire pool of individuals. Specifically, the scenario below shows a u-ratio of about 0.70 in the independent variable and dependent variables. We see in the figure that the correlation in the restricted sample ($r_{XY|\mathcal{S}(X,Y)}$) is attenuated relative to the unrestricted correlation (r_{XY}).

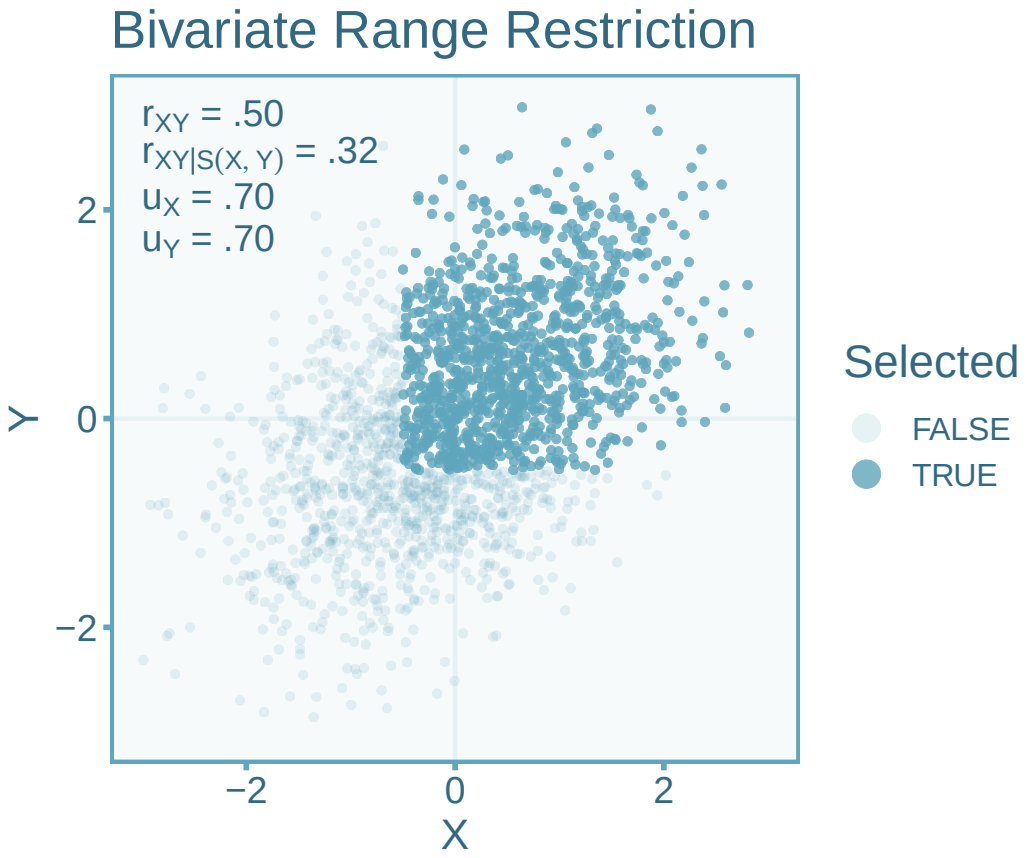


Figure 9.5: Scatterplot showing a correlation between X and Y under bivariate direct range restriction. Dark blue dots indicate the selected sample and the transparent dots indicate the rejected sample.

Likewise let's visualize what happens to the correlation when the range is enhanced. Enhancement in both variables can be accomplished by selecting individuals at the ends of the distribution of X and Y. Therefore we can define the selection function as,

$$\mathcal{S}(X, Y) = \begin{cases} 1 & \text{if } (X \leq -.5 \text{ or } X \geq .5) \text{ and } (Y \leq -.5 \text{ or } Y \geq .5) \\ 0 & \text{if } (X > -.5 \text{ or } X < .5) \text{ and } (Y > -.5 \text{ or } Y < .5) \end{cases}$$

Note that this type of selection would be exceedingly rare to see in practice. In Figure 9.6, we see inflation of the enhanced correlation relative to the target correlation. The scenario below has a u-ratio of about 1.32 in both the independent variable and dependent variable.

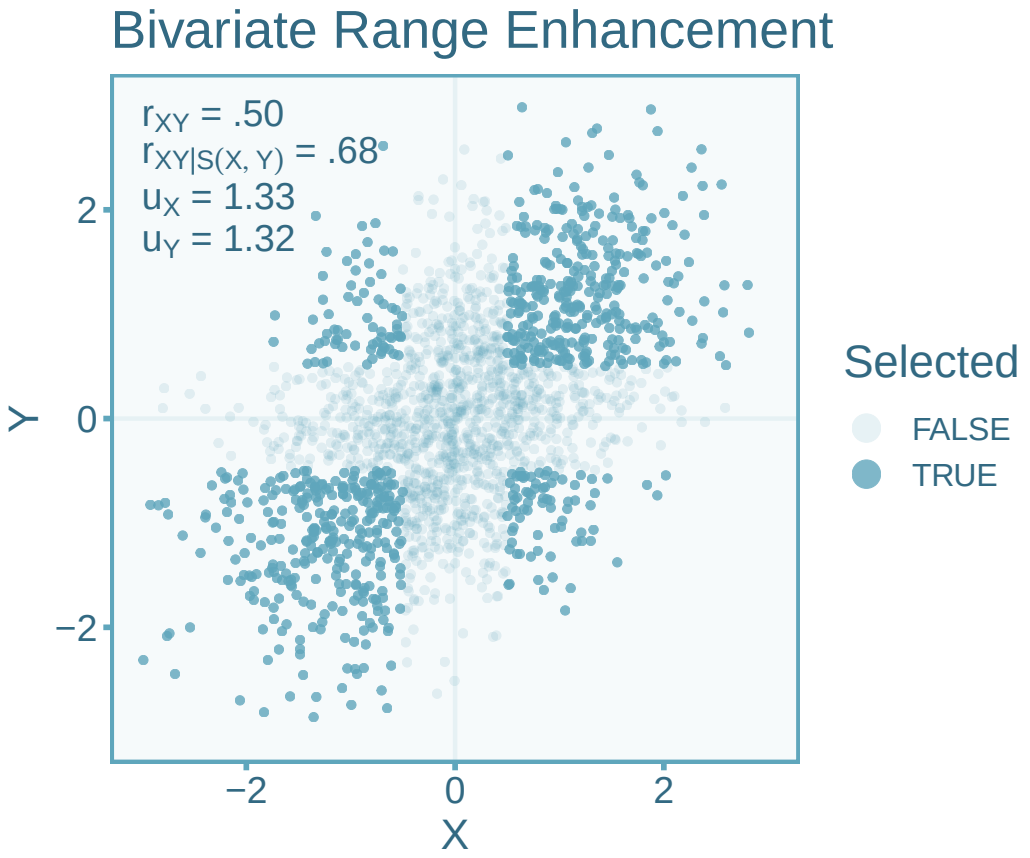


Figure 9.6: Scatterplot showing a correlation between X and Y under bivariate direct range enhancement. Dark blue dots indicate the selected sample and the transparent dots indicate the rejected sample.

A bias correction formula for bivariate range restriction is much more complicated than the univariate formulation. To break down the correction formula into simpler parts, let us first define a factor we will denote with the Greek letter ψ,

$$\psi = \frac{u_X u_Y \left(r^2_{XY|\mathcal{S}(X,Y)} - 1\right)}{2r_{XY|\mathcal{S}(X,Y)}} \tag{9.4}$$

This factor contains all the parameters needed to correct the correlation coefficient under direct selection. An unbiased estimate of the target population correlation can obtained by the following correction formula (adapted from table 3 Wiernik and Dahlke 2020),

$$r_{XY} = \psi + \text{sign}\left[r_{XY|\mathcal{S}(X,Y)}\right]\sqrt{\psi^2 + 1} \tag{9.5}$$

Where the standard error can be computed from calculating the artifact factor (a) from the corrected and restricted/enhanced correlation,

$$se(r_{XY}) = \frac{se\left(r_{XY|\mathcal{S}(X,Y)}\right)}{\hat{a}} = \frac{se\left(r_{XY|\mathcal{S}(X,Y)}\right)}{\left[\frac{r_{XY|\mathcal{S}(X,Y)}}{r_{XY}}\right]}. \tag{9.6}$$

Now we can also incorporate measurement error into the correction formula. Note that the following equations will incorporate the reliability within the selected sample ($r_{XX'|\mathcal{S}(X,Y)}$) rather than the unrestricted population ($r_{XX'}$; see Equation 9.3 on converting to the selected sample). Then we can use the restricted/enhanced (selected) sample reliability and the u-ratios in the following equation to obtain an unbiased estimate of the target true score population correlation,

$$r_{TU} = \frac{\psi + \text{sign}\left[r_{XY|\mathcal{S}(X,Y)}\right]\sqrt{\psi^2 + 1}}{\sqrt{1 - u_X^2\left(1 - r_{XX'|\mathcal{S}(X,Y)}\right)}\sqrt{1 - u_Y^2\left(1 - r_{YY'|\mathcal{S}(X,Y)}\right)}}.$$

If the reliability coefficient comes from the unrestricted population, the formula simplifies to,

$$r_{TU} = \frac{\psi + \text{sign}\left[r_{XY|\mathcal{S}(X,Y)}\right]\sqrt{\psi^2 + 1}}{\sqrt{r_{XX'}}\sqrt{r_{YY'}}}.$$

We can use the same equation as Equation 9.6 to calculate the corrected standard error. The standard error can then be calculated as,

$$se(r_{TU}) = \frac{se\left(r_{XY|\mathcal{S}(X,Y)}\right)}{\hat{a}} = \frac{se\left(r_{XY|\mathcal{S}(X,Y)}\right)}{\left[\frac{r_{XY|\mathcal{S}(X,Y)}}{r_{TU}}\right]}.$$

Applied Example in R

Let's say we are the admissions department at an top-tier university. This university only admits students that have a high school GPA over 3.5 *and* a standardized test score over 650. We decide to conduct a study on the predictive validity (i.e., the correlation) between standardized test performance and college grade-point average (GPA). However, the admitted students have a very narrow range of both GPA and test scores wheras the total pool of college applicants have a much wider range. We calculate u-ratios of test scores ($u_X = 0.70$) and GPA ($u_Y = .80$).
In R, we can correct the correlation by using the `correct_r` function in the `psychmeta` package (Dahlke and Wiernik 2019). The `correction = 'bvdrr'`

```r
library(psychmeta)

correct_r(correction = 'bvdrr',
          rxyi = .25,  # restricted correlation
          ux = .70,    # u ratio of IQ scores
          uy = .80,    # u ratio of GPA
          n = 100)     # sample size
```

```
Correlations Corrected for Measurement Error and Bivariate Direct Range Restriction:
--------------------------------------------------------------------------------
  value CI_LL_95 CI_UL_95   n n_effective
1   0.4   0.0999    0.597 100          32
```

The correlation in the target population is estimated to be $r_{XY} = .40\,[.10,\ .60]$.

9.6 Correcting Standardized Mean Differences for Direct Range Restriction

9.6.1 Defining our Target Quantity

The quantity of interest is the target population standardized mean difference between actual members of groups A and B on true scores, U. We can denote this standardized mean difference as δ_{GU}. Within a study that suffers from direct selection, the observed standardized mean difference ($d_{gY|\mathcal{S}(Y)}$) will be biased relative to the target, δ_{GU}. We can model the observed standardized mean difference as,

$$d_{gY|\mathcal{S}(Y)} = a\delta_{GU} + e.$$

Where a is the attenuation/inflation factor and e is the sampling error term. Therefore an unbiased estimate of the target population true score standardized mean difference would be computed from

$$d_{GU} = \frac{d_{gY|\mathcal{S}(Y)}}{\hat{a}}.$$

9.6.2 Artifact Correction for Standardized Mean Difference

9.6.2.1 Selection on the Continuous Variable

To correct for direct selection on the continuous variable, we can first convert the observed standardized mean difference ($d_{gY|\mathcal{S}(X)}$) to a point-biserial correlation ($r_{gY|\mathcal{S}(X)}$). Converting $d_{gY|\mathcal{S}(X)}$ to $r_{gY|\mathcal{S}(X)}$ can be done by using the observed proportion of individuals in group A (or B), p_g,

$$r_{gY|\mathcal{S}(X)} = \frac{d_{gY|\mathcal{S}(X)}}{\sqrt{\frac{1}{p_g(1-p_g)} - d^2_{gY|\mathcal{S}(X)}}}.$$

We can then correct the point-biserial correlation for univariate direct selection using the formulas in Section 9.5.2. Note that if you want to correct for measurement error as well, replace $r_{XX'}$ with $r_{gg'}$ (i.e., group classification reliability; see chapter on *group misclassification*) whenever you are working with standardized mean differences. Once we obtained the corrected correlation, r_{GU}, we can convert back to a standardized mean difference, we need to use an adjusted group proportions, p^*_g:

$$d_{GU} = \frac{r_{GU}}{\sqrt{p^*_g\left(1-p^*_g\right)\left(1-r^2_{GU}\right)}}.$$

Where p^*_g is

$$p^*_g = \frac{1}{2} - \frac{1}{2}\sqrt{1 - 4p_g(1-p_g)\left[1 + r^2_{gY|\mathcal{S}(Y)}\left(\frac{1}{u^2_X} - 1\right)\right]}$$

The adjusted proportion, p^*_g, can also be estimated from the proportion of individuals in the target population (e.g., the proportion of men vs women in the general population). This adjustment is necessary in order to account for indirect selection in the grouping variable when $d \neq 0$. This is similar to the situation described in Section 9.5.2, where one variable suffers from direct range restriction and any variable that is correlated with it, will suffer from indirect selection. The corresponding corrected sampling error can also be computed with the observed and adjusted proportions such that,

$$se(d_{GU}) = \frac{se\left(d_{gY|\mathcal{S}(Y)}\right)\left(\frac{r_{GU}}{r_{gY|\mathcal{S}(Y)}}\right)^2}{\left(1 + d^2_{gY|\mathcal{S}(Y)}p_g[1-p_g]\right)^2\left(d^2_o + \frac{1}{p_g(1-p_g)}\right)p^*(1-p^*_g)(1-r^2_{GU})^3}.$$

10 Indirect Selection

10.1 Introduction

Indirect selection occurs when the selection process is not directly on the variable of interest, but rather on another related variable. Similar to direct range restriction, this will cause restriction (or enhancement) in the variable of interest.

10.2 An Example of Indirect Range Restriction

Imagine a research team is conducting a study on academic motivation among college students using a survey that includes various questions related to academic engagement, goal orientation, and effort investment. The researchers administer the survey to a large sample of students across different universities. However, during the data cleaning process, the researchers identify a subset of respondents who exhibited signs of inattentiveness and carelessness in their responses. These signs include straightlining questions (e.g., consistently selecting the same response option without reading the questions) or responding randomly without considering the content of the questions. Recognizing that inattentive or careless responding can distort the measurement of academic motivation, the researchers decide to exclude these individuals from the analysis. The rationale is to ensure that the data collected represents genuine responses and validly measures academic motivation. The unintended consequence of this decision is indirect range restriction. By removing inattentive and careless responders, who likely also have lower academic motivation and engagement, from the dataset, the observed range of academic motivation scores is reduced. The excluded individuals, who may have had lower academic motivation scores, are not accounted for in the analysis, resulting in an underestimation of the variability of academic motivation relative to the population.

10.3 Selection Functions

In the previous chapter we introduced the concept of a selection function for direct range restriction. Here we will expand on that section for the case of indirect range restriction, where we must introduce a new variable, Z. Indirect selection by definition means that the variable of interest is not used in the selection process, rather, the selection directly on another variable such that, $\mathcal{S}(Z)$ (Figure 10.1 illustrates the selection process).

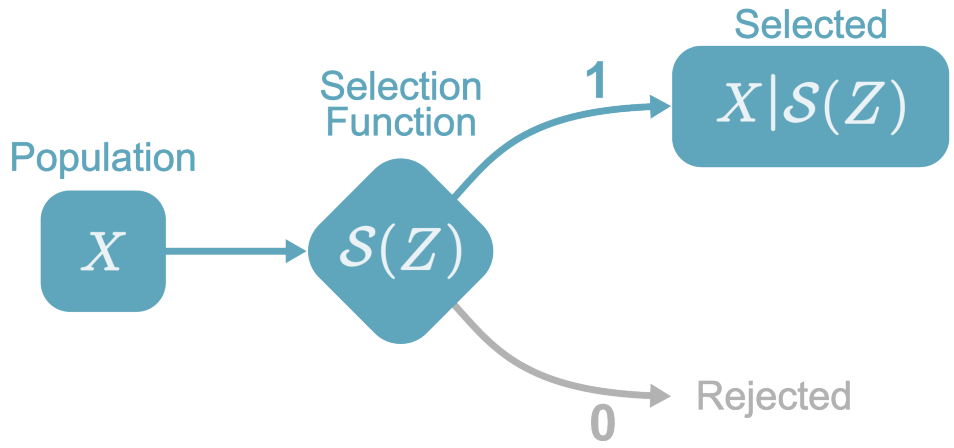


Figure 10.1: Diagram illustrating the selection process. The variable, X, is selected based on a function of itself. Therefore the variable under selection is conditional on the selection procedure $X|\mathcal{S}(X)$.

However, selecting on Z may affect the distribution of X depending on the relationship between X and Z. Particularly, if we assume normally distributed variables the correlation between X and Z can entirely describe the affect that $\mathcal{S}(Z)$ has on the distribution of X. If X and Z are independent ($\rho_{XZ} = 0$), then selection on Z would not affect the distribution of X such that,

$$f(X|\mathcal{S}(Z)) = f(X), \quad \rho_{XZ} = 0$$

For example, consider that we select individual's in the top half of Z (i.e., above the mean, μ_Z), such that our selection function can be defined as,

$$\mathcal{S}(Z) = \begin{cases} 1 & \text{if } Z \geq \mu_Z \\ 0 & \text{if } Z < \mu_Z \end{cases}$$

If the correlation, ρ_{XZ} is positive, then the distribution of $X|\mathcal{S}(Z)$ (i.e., X given selection on Z) will have a higher mean and less variability (see Figure 10.1)

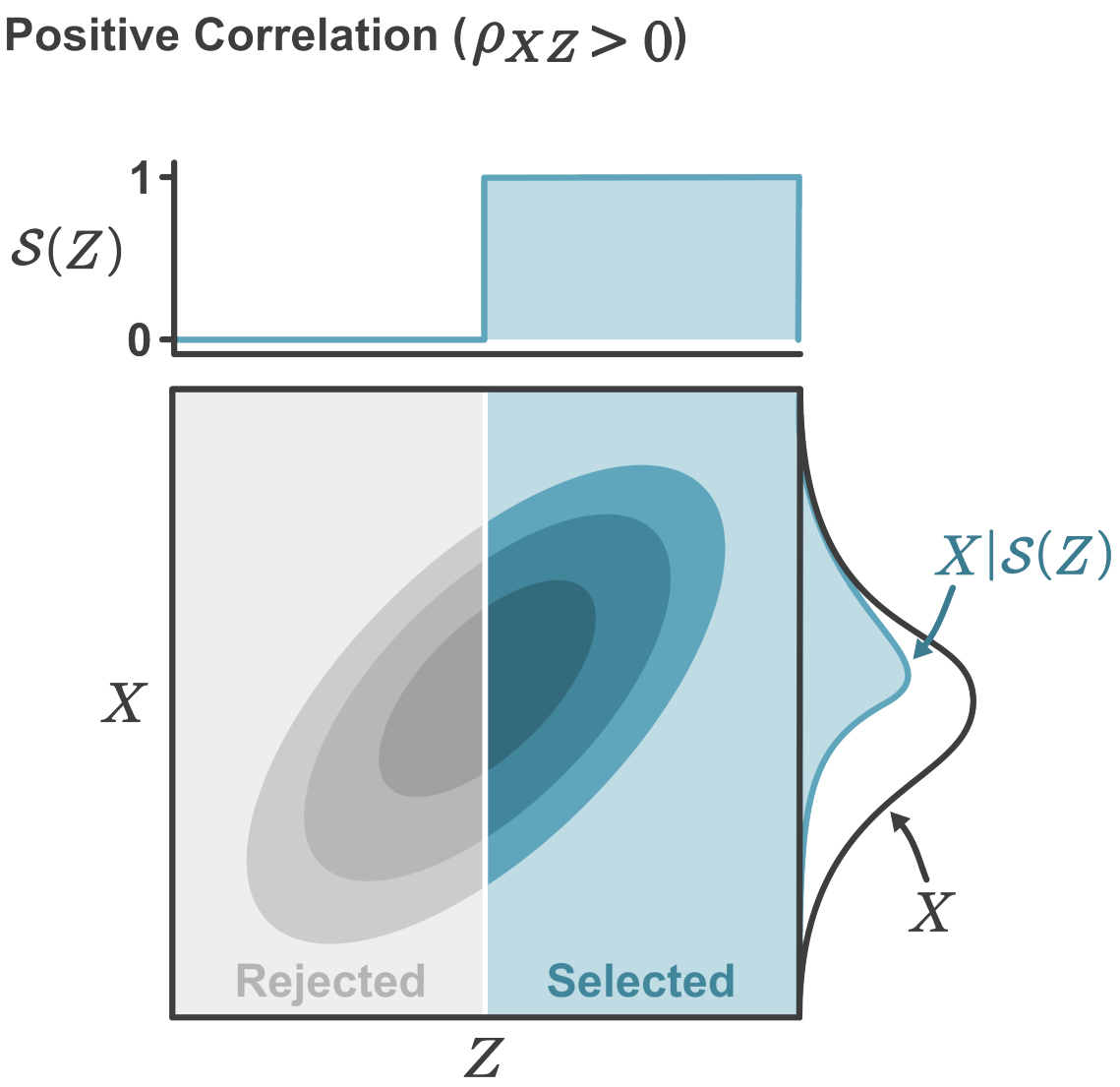


Figure 10.2: Diagram illustrating the impact of selecting on Z has on X given a positive correlation. The selection function on the top of the plot denotes what values of Z are selected (i.e., $\mathcal{S}(Z)$). The distributions on the right indicate the distribution of X (in black) and the distribution X after

If there is no correlation between X and Z, then the distribution of $X|\mathcal{Z}$ would be left unchanged (see Figure 10.3).

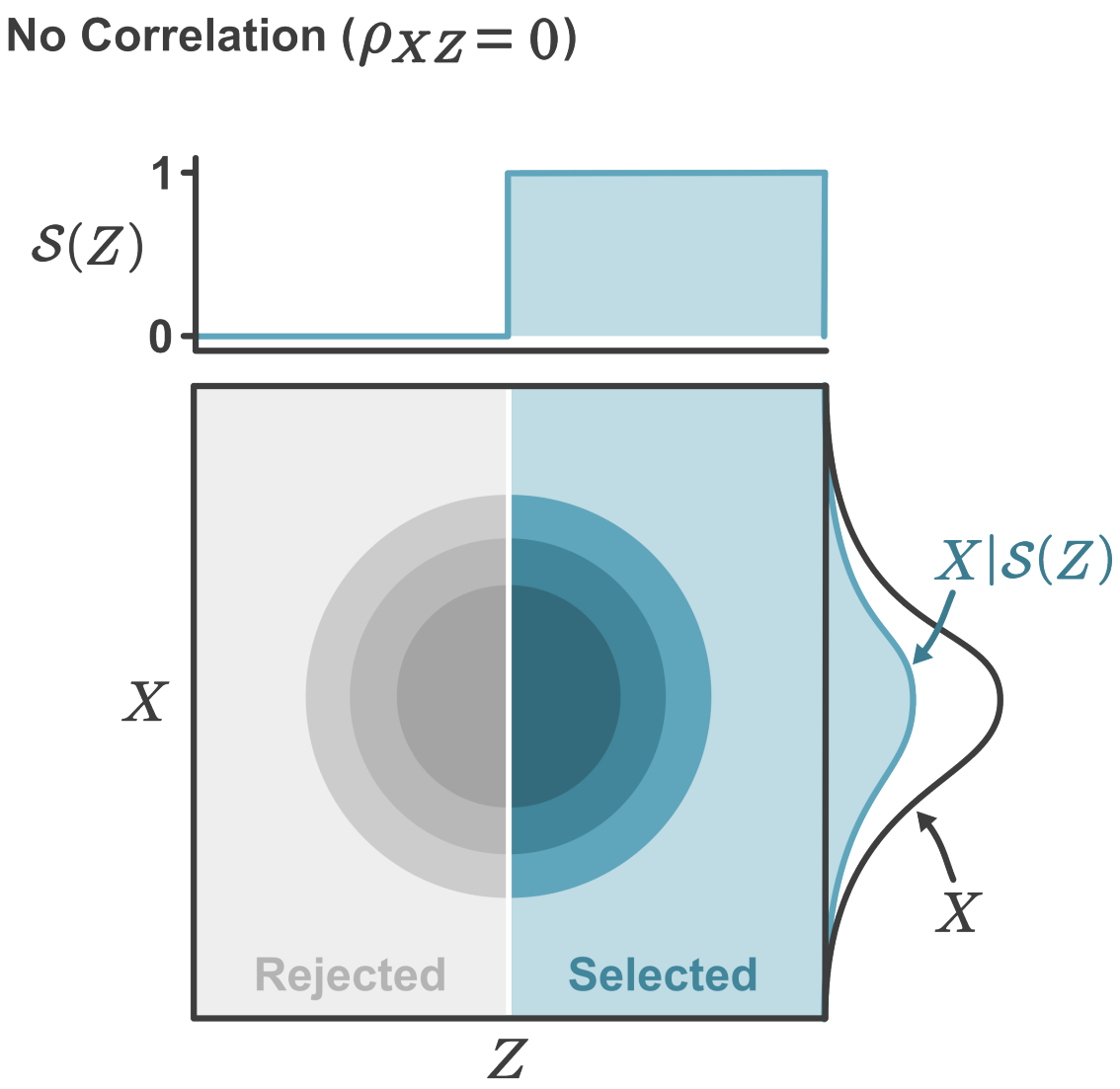


Figure 10.3: Diagram illustrating the direct selection function. The variable, X, in the population is selected based on a function of itself. Therefore the variable under selection is conditional on the selection procedure $X|\mathcal{S}(X)$.

10.4 Quantifying Selection-Induced Restriction/enhancement

The distribution of scores in the target population may exhibit a greater (or lesser) degree of variability compared to the sample that has been selected into the study. Therefore the standard deviation of scores in the target population (σ_X) may differ from the population under selection. ($\sigma_{X|mathcalS(Z)}$).

To index the difference between the two standard deviations, we can calculate the u-ratio Wiernik and Dahlke (2020). The u-ratio is the ratio between the standard deviations of the population under selection and the target population such that (υ denotes the population u-ratio),

$$\upsilon_X = \frac{\sigma_{X|\mathcal{S}(Z)}}{\sigma_X}$$

The u-ratio in cases of range *restriction* will exist in the interval (0–1). Conversely, when the u-ratio is greater than 1 it is indicative of range *enhancement*. The target population standard deviation is often quite difficult to acquire since we do not usually have access to a random sample from that population. However, the target population standard deviation can be estimated from a reference sample that is representative of the target population. This often comes in the form of standardization samples or norm samples (obtained from test manuals) if the unrestricted group is the general population. For example, the distribution full-scale IQ scores derived from the Wechsler Adult Intelligence Test has a standard deviation of 15 in the US population (Wechsler 2008). We can use this estimate as the standard deviation for the unrestricted population. Lets say we select a sample from members of Harvard students, who tend to have higher IQs than the general population (this is due to the fact that selection criterion, such as GPA and SAT scores are positively correlated with IQ). If the standard deviation of IQ in Harvard students is 10, then the u-ratio would be,

$$u_X = \frac{S_{X|\mathcal{S}(Z)}}{S_X} = \frac{10}{15} = .67$$

However it is not always the case that an estimate of the unrestricted standard deviation is readily available. Therefore if the reliability coefficient from the reference sample and the sample under selection can be used to estimate the u-ratio,

$$u_X = \sqrt{\frac{1 - r_{XX'}}{1 - r_{XX'|\mathcal{S}(Z)}}}$$

Where $r_{XX'|\mathcal{S}(Z)}$ and $r_{XX'}$ are the reliability estimates from the sample under selection and the reference (target population) sample, respectively. In the context of indirect range restriction, the selection does not occur directly on X (or Y), instead it occurs on a third variable, Z. The affect that selection on Z has on X is dependent on the correlation between them, ρ_{XZ}. Therefore we can see how the u-ratio of Z (u_Z) related to the u-ratio of X,

$$u_X = \sqrt{\rho_{XZ}^2 u_Z^2 - \rho_{XZ}^2 + 1}$$

If $\rho_{XZ} = 0$, then you will notice that $u_X = 1$, effectively having no selection effect on X. Also, notice that a correlation of $\rho_{XZ} = 1$ will return $u_X = u_Z$, indicating that selecting on Z would affect

the variation of Z similarly to the variation in X. This relationship between u_X, u_Z, and ρ_{XZ} can be visualized in Figure 10.4

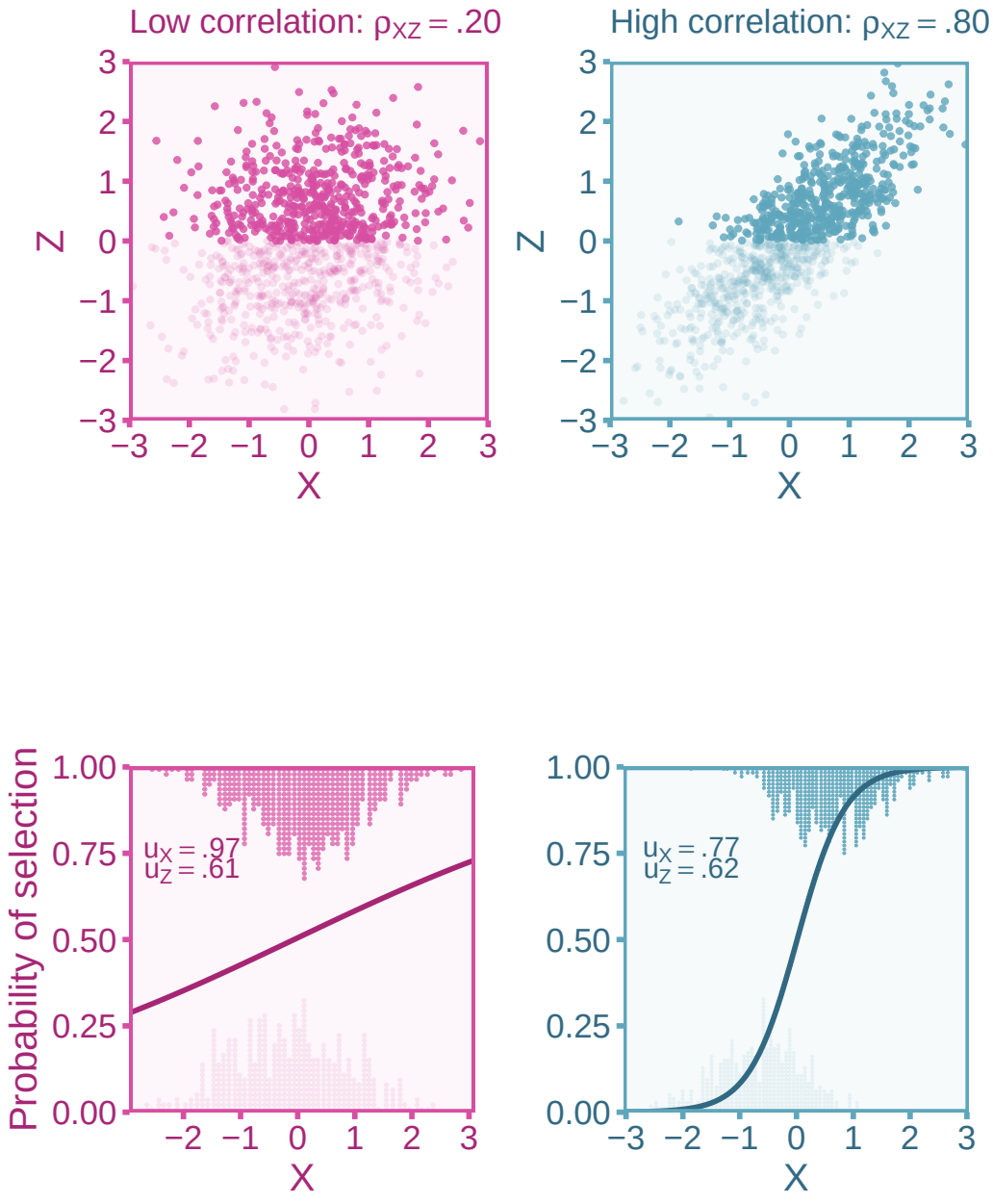


Figure 10.4: The impact of the association between X and Z. The data consists of $N = 1,000$ simulated observations, where the red plots show a low correlation case ($\rho_{XZ} = .20$) and the blue plots show a high correlation case ($\rho_{XZ} = .80$). The top plots show the relationship between X and Z with the darker points indicating individuals that have been selected into the sample. The bottom plots shows the probability of selection into the sample as a function of X. The dark distribution on the top of the plot show the distribution of individuals in the selected sample. Notice that the distribution of the selected individual's in the high correlation case is a narrower distribution than the low correlation case.

10.5 Correcting Correlations for Indirect Range Restriction

10.5.1 Defining our Target Quantity

We want to estimate the population correlation of the true scores of the independent variable (T) and dependent variable (U). We can denote this correlation as ρ_{TU}. Within a study sample that suffers from indirect selection (and sampling error), the study correlation is under selection (r_{xy_S}) will be biased relative to our target quantity, ρ_{xy}. This bias is captured by an artifact attenuation/inflation factor, a, such that,

$$r_{XY} = a\rho_{XY|\mathcal{S}(Z)} + e$$

Therefore an unbiased estimate of the true score target population correlation can be estimated by dividing the observed score correlation under selection by an estimate of a,

$$r_{TU} = \frac{r_{XY|\mathcal{S}(Z)}}{\hat{a}}.$$

10.5.2 Artifact Correction for Correlations

The Univariate Case

Range restriction (or enhancement) in either the independent or dependent variable will induce bias into the correlation coefficient. Let us consider a case where we select individuals based on meeting some criterion of some third variable, Z, such that $\mathcal{S}(Z)$. In the univariate case, we assume that selection on Z only directly affects restriction/enhancement in X while any restriction/enhancement in Y is mediated by the effect on X (see Figure 10.5).

Now consider a study where we want to calculate correlation in the target population between an independent variable, X, and a dependent variable, Y. However, the individual's are selected whether they are above the mean of Z (Mean = 0). We can thus define the selection function such that,

$$\mathcal{S}(Z) = \begin{cases} 1 & \text{if } Z \geq 0 \\ 0 & \text{if } Z < 0 \end{cases}$$

In the following examples, we will simulate a correlation of $\rho_{XZ} = .80$.

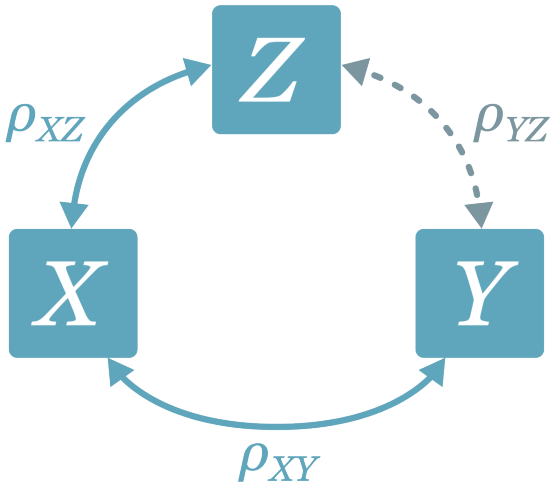


Figure 10.5: The assumed correlation structure of univariate indirect selection. The correlation between Z and Y is completely mediated by X (i.e., the partial correlation between Y and Z, controlling for X is zero such that, $\rho_{YZ.X} = 0$).

Figure 10.6 shows a u-ratio of about $u_X = 0.75$ in the independent variable. We see that the sample correlation in the restricted scores ($r_{XY|\mathcal{S}(Z)} = .42$) is attenuated relative to the unrestricted correlation ($r_{XY} = .50$).

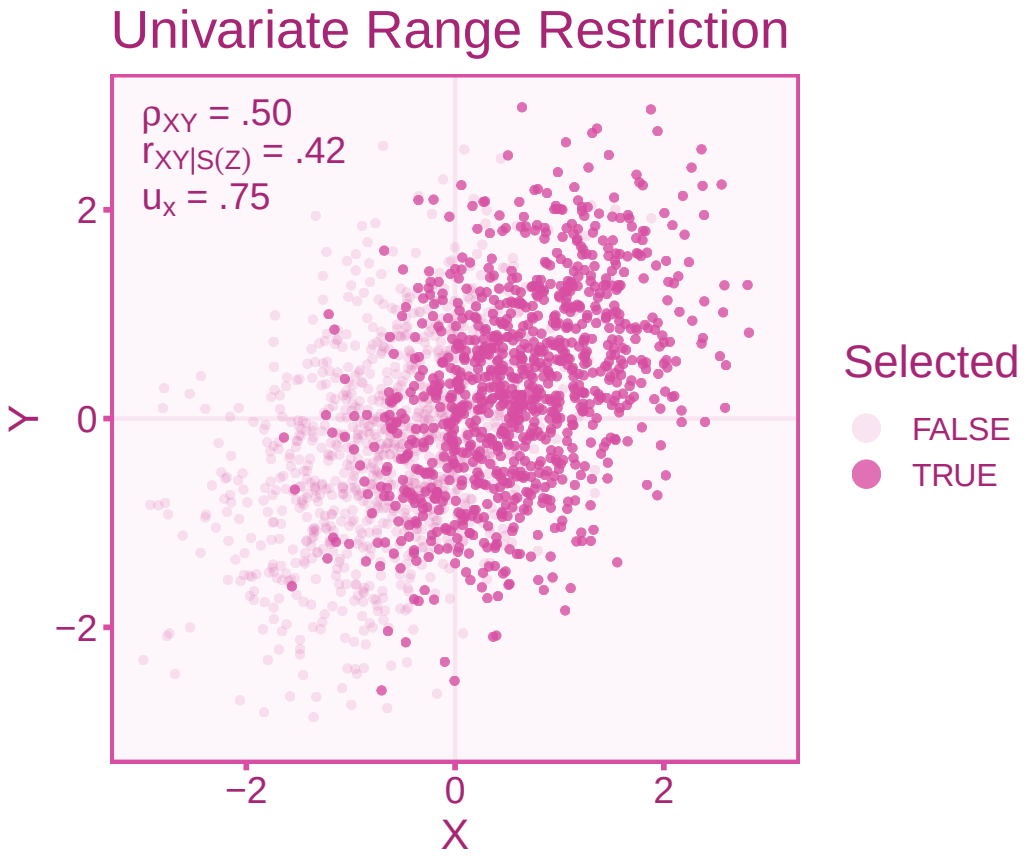


Figure 10.6: Scatterplot showing a correlation between X and Y under univariate indirect range restriction. Dark red dots indicate the selected sample and the transparent dots indicate the rejected sample.

We can also visualize what happens to the correlation when the range is enhanced. Enhancement can be accomplished by selecting individuals at the ends of the distribution (Taylor and Griess 1976). For indirect selection, individuals are selected at the ends of the distribution of Z such that the selection function can be defined as,

$$\mathcal{S}(Z) = \begin{cases} 1 & \text{if } Z \leq -1 \text{ or } Z \geq 1 \\ 0 & \text{if } Z < -1 \text{ or } Z > 1 \end{cases}$$

In Figure 10.7, we see an opposite effect on the correlation, that is, an inflation of the correlation rather than an attenuation like we see under range restriction. The scenario below has a u-ratio of about

$u_X = 1.32$ in the independent variable.

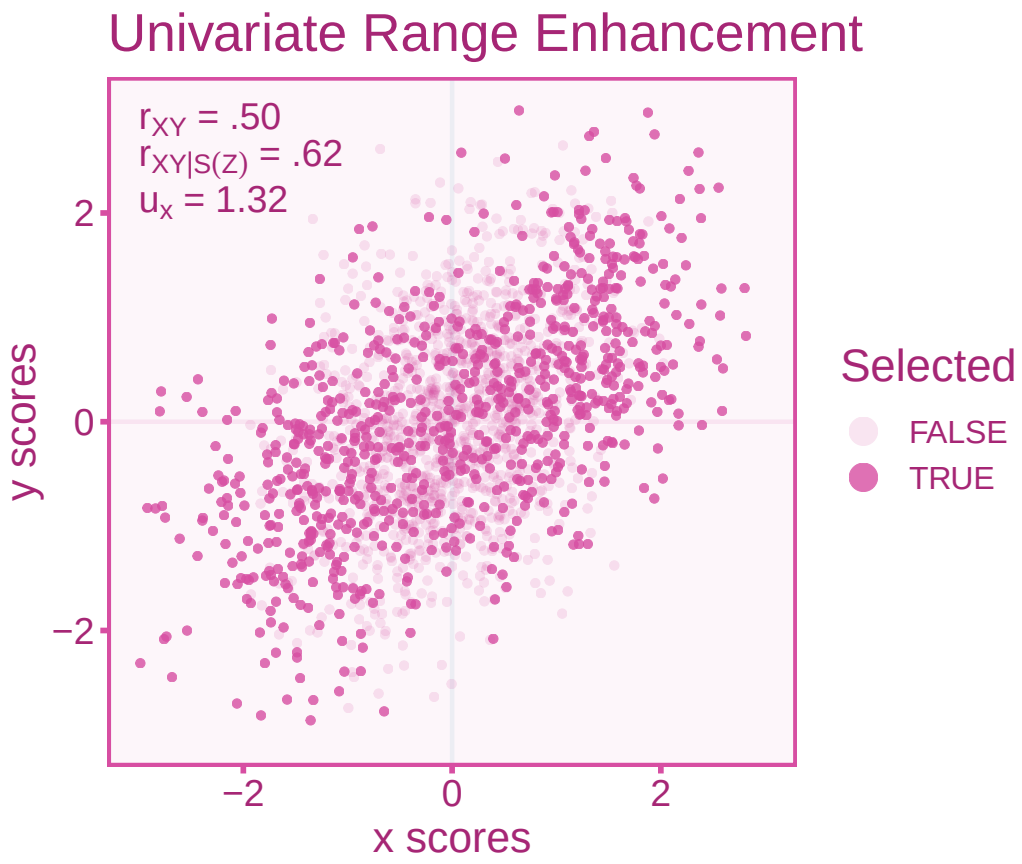


Figure 10.7: Scatterplot showing a correlation between X and Y under univariate indirect range enhancement. Dark red dots indicate the selected sample and the transparent dots indicate the rejected sample.

In summary, if $u_X > 1$ (i.e., $S_{X|\mathcal{S}(Z)} > S_X$) the observed correlation is inflated relative to the correlation in the target population. Whereas the correlation is attenuated when $u_X < 1$ (i.e., $S_{X|\mathcal{S}(Z)} < S_X$, Sackett and Yang 2000). An estimate of the attenuation/inflation factor, $\hat{a}$, can account for the bias in the observed correlation induced by range restriction/enhancement (equation 5, Le and Schmidt 2006).

$$\hat{a} = \sqrt{r^2_{XY|\mathcal{S}(Z)} + u^2_X(1 - r^2_{XY|\mathcal{S}(Z)})}$$

Using the estimated attenuation/inflation factor, we can correct the observed correlation for bias induced by indirect selection

$$r_{XY} = \frac{r_{XY|\mathcal{S}(Z)}}{\hat{a}} = \frac{r_{XY|\mathcal{S}(Z)}}{\sqrt{r^2_{XY|\mathcal{S}(Z)} + u^2_X(1 - r^2_{XY|\mathcal{S}(Z)})}} \tag{10.1}$$

If we wants to correct for range restriction/enhancement *and* measurement error, we can incorporate the reliability coefficients (under selection) of X ($r_{XX'|\mathcal{S}(Z)}$) and Y ($r_{YY'|\mathcal{S}(Z)}$) into the formula for $\hat{a}$,

$$\hat{a} = \sqrt{r^2_{XY|\mathcal{S}(Z)} + \frac{u^2_X r_{XX'|\mathcal{S}(Z)} \left(r_{XX'|\mathcal{S}(Z)} r_{YY'|\mathcal{S}(Z)} - r^2_{XY|\mathcal{S}(Z)}\right)}{1 - u^2_X \left(1 - r_{XX'|\mathcal{S}(Z)}\right)}}$$

Now correcting the observed correlation with this modified estimate of a to correct the observed correlation will yield the true score correlation in the target population,

$$r_{TU} = \frac{r_{XY|\mathcal{S}(Z)}}{\hat{a}} = \frac{r_{XY|\mathcal{S}(Z)}}{\sqrt{r^2_{XY|\mathcal{S}(Z)} + \frac{u^2_X r_{XX'|\mathcal{S}(Z)} \left(r_{XX'|\mathcal{S}(Z)} r_{YY'|\mathcal{S}(Z)} - r^2_{XY|\mathcal{S}(Z)}\right)}{1 - u^2_X \left(1 - r_{XX'|\mathcal{S}(Z)}\right)}}}$$

If the reliability coefficients come from the target population and do not suffer selection effects, we can estimate the reliability under selection using the following formulas (equation 11 and 12 Le and Schmidt 2006):

$$r_{XX'|\mathcal{S}(Z)} = 1 - \frac{1 - r_{XX'}}{u^2_X} \tag{10.2}$$

$$r_{YY'|\mathcal{S}(Z)} = 1 - \frac{1 - r_{YY'}}{u^2_Y} \tag{10.3}$$

We now need to adjust the standard error for the corrected correlation coefficient. To do this, we can either divide the observed standard error by the attenuation/inflation factor (or equivalently, the observed correlation divided by the corrected correlation),

$$se(r_{TU}) = \frac{se\left(r_{XY|\mathcal{S}(Z)}\right)}{\hat{a}} = \frac{se\left(r_{XY|\mathcal{S}(Z)}\right)}{\left[\frac{r_{XY|\mathcal{S}(Z)}}{r_{TU}}\right]}.$$

Applied Example in R

Let's say a university admits students based on some criterion upon a composite of multiple measures of performance (Z). One of those measures is a standardized test (X). A researcher then wants to see how test performance predicts stress (Y) in college. Since test performance is corre-

lated with the composite used in the college admissions process, it is likely that we will observe substantial range restriction in the college sample. To index this restriction in range, we calculate the u-ratio of test scores as $u_X = 0.70$ (relative to the total pool of applicants). We also want to correct for measurement error in test performance and the self-report questionnaire we use to measure stress, therefore we obtain reliability estimates within our sample of $r_{XX'|\mathcal{S}(Z)} = .90$ and $r_{XX'|\mathcal{S}(Z)} = .80$, respectively. The researcher then conducts the study and finds a sample correlation of $r_{XY|\mathcal{S}(Z)} = .25$, but he wants to know that true score correlation.
In R, we can correct the correlation by using the `correct_r` function in the `psychmeta` package (Dahlke and Wiernik 2019). The `correction = 'uvirr_x'`

```
library(psychmeta)

correct_r(correction = 'uvirr_x',
          rxyi = .25,  # restricted correlation
          rxx = .90,  # reliability of test scores
          ryy = .80,  # reliability of stress
          ux = .70,   # u ratio of SAT scores
          n = 100)    # sample size
```

```
Correlations Corrected for Measurement Error and Univariate Indirect Range Restriction:
-----------------------------------------------------------------------------------------
  value CI_LL_95 CI_UL_95   n n_effective
1 0.412   0.0973    0.648 100        33.8
```

The true score correlation in the target population is estimated to be $r_{TU} = .41\,[.10,\ .65]$.

The Bivariate Case

Bivariate indirect range restriction/enhancement occurs when the selection variable has independent relationships with both the independent and dependent variable. Like we did for the univariate case, let's visualize the correlation between independent (X) and dependent (Y) variables under range restriction by only selecting individuals above a score of -0.50 in our selector variable, Z. Therefore the selection function can be defined as,

$$\mathcal{S}(Z) = \begin{cases} 1 & \text{if } Z \geq -.50 \\ 0 & \text{if } Z < -.50 \end{cases}$$

We will also fix the correlations between the selector and independent variable (ρ_{XZ}), as well as the selector and dependent variable (ρ_{YZ}) to be .80. Unlike the univariate case, in the bivariate case X *and* Y have direct relationships with Z (see Figure 10.8).

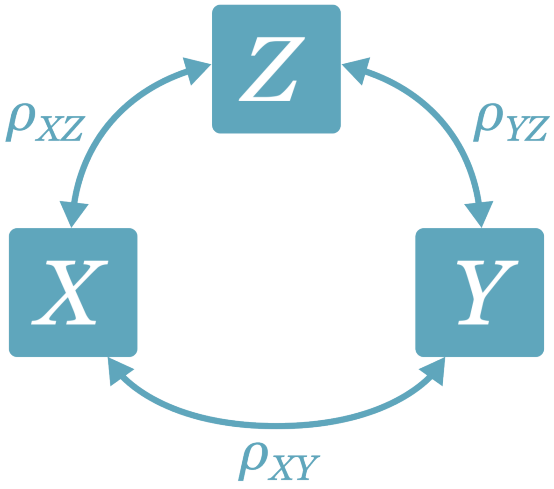


Figure 10.8: The assumed correlation structure of univariate indirect selection. The correlation between Z and Y is completely mediated by X (i.e., the partial correlation between Y and Z, controlling for X is zero such that, $\rho_{YZ.X} = 0$).

The scenario displayed in Figure 10.9, shows a u-ratio of about $u_X = u_Y = 0.81$ in the independent variable and dependent variables. We see in the figure below that the correlation in the restricted sample ($\rho_{XY|\mathcal{S}(Z)} = .25$) is attenuated relative to the target population correlation ($\rho_{XY} = .50$).

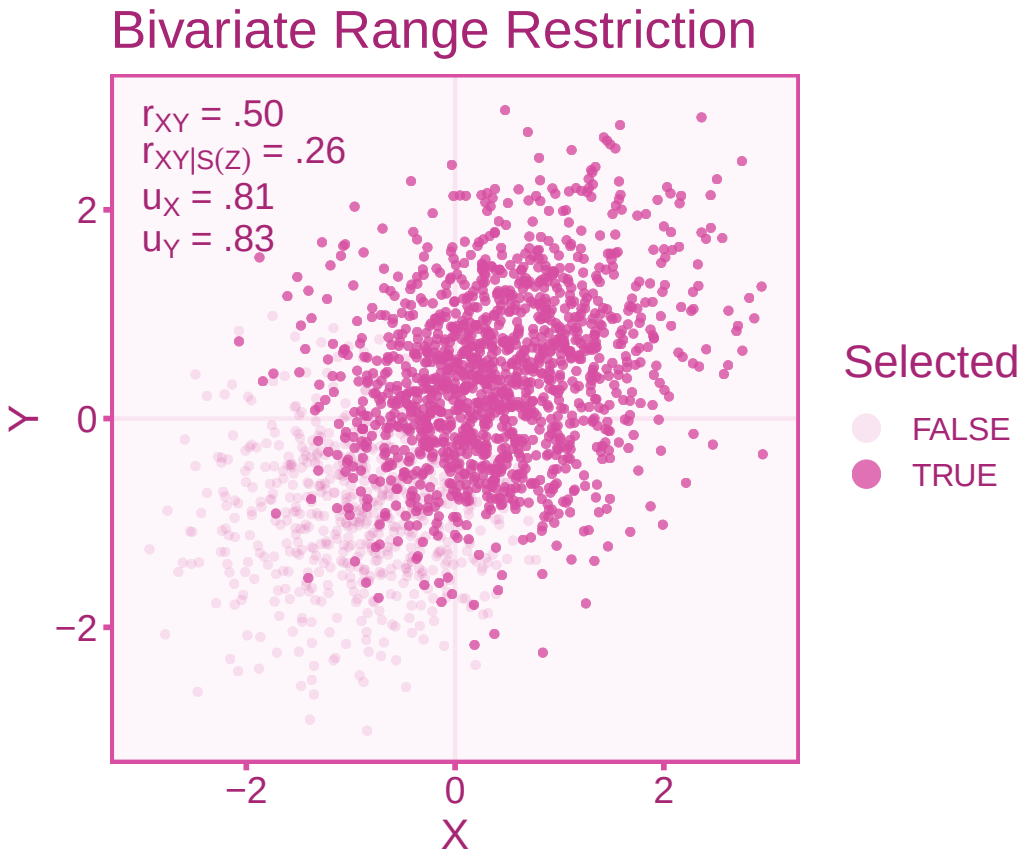


Figure 10.9: Scatterplot showing a correlation between X and Y under bivariate indirect range restriction. Dark red dots indicate the selected sample and the transparent dots indicate the rejected sample.

Likewise let's visualize what happens to the correlation when the range is enhanced. Enhancement in both variables can be accomplished by selecting individuals at the ends of the distribution of Z (for this case we will select individuals below a score of -1 and above a score of 1). We can thus define the selection function as,

$$\mathcal{S}(Z) = \begin{cases} 1 & \text{if } Z \leq -1 \text{ or } Z \geq 1 \\ 0 & \text{if } Z > -1 \text{ or } Z < 1 \end{cases}$$

In Figure 10.10, we observe an inflation of observed correlation ($\rho_{XY|\mathcal{S}(Z)} = .74$) relative to the target correlation ($\rho_{XY} = .50$). Figure 10.10 has a u-ratio of about $u_X = u_Y = 1.38$ in both the independent

variable and dependent variable.

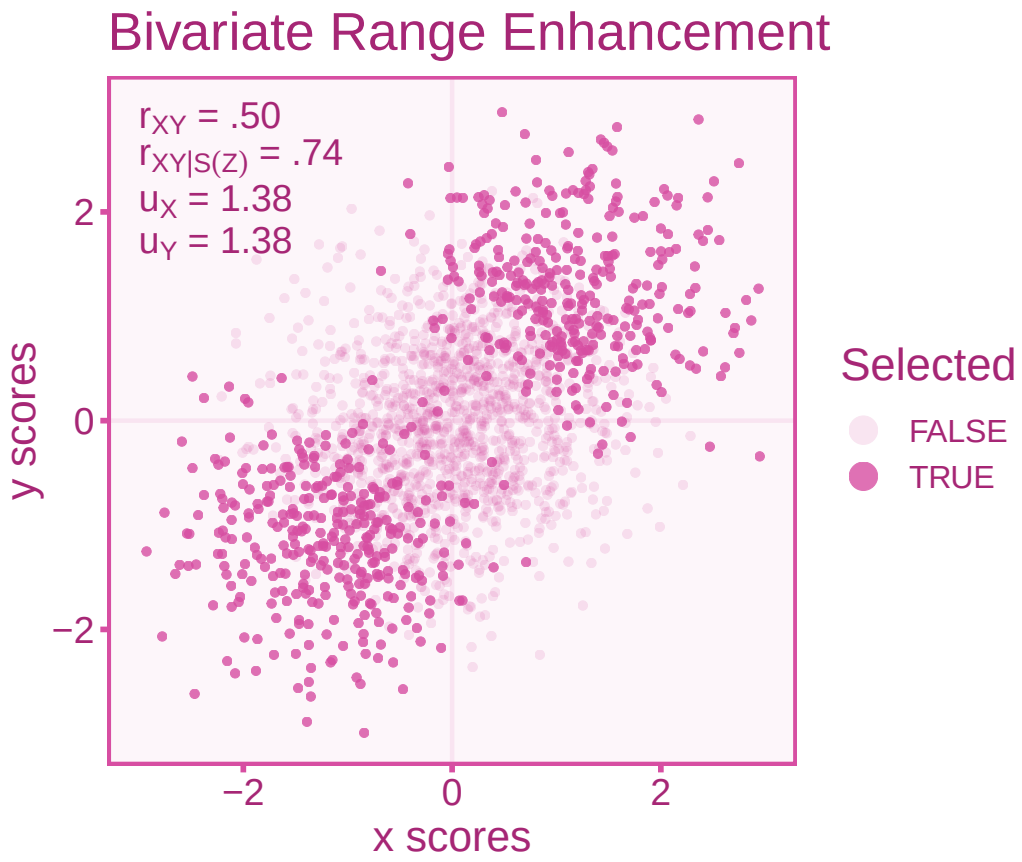


Figure 10.10: Scatterplot showing a correlation between X and Y under bivariate indirect range enhancement. Dark red dots indicate the selected sample and the transparent dots indicate the rejected sample.

A bias correction formula for bivariate range restriction is much more complicated than the univariate formulation. In the univariate case, we did not need any more information about the selection process beyond what we could infer from u_X. However in the bivariate case, we need to have a basic idea of the selection mechanism at play (Dahlke and Wiernik 2020). Particularly we at least know the direction of the correlation between the selector variable, Z, and the independent (ρ_{XZ}) and dependent variable (ρ_{YZ}). This will require a little bit of knowledge about the selection process within a given study. Let us first define a factor we will denote with λ (Dahlke and Wiernik 2020). This factor takes into account the direction of the correlation of ρ_{XZ} (if positive, we can set $\rho_{XZ} = 1$, if negative, $\rho_{XZ} = -1$, if zero, $\rho_{XZ} = 0$) and ρ_{YZ} (repeat the same procedure as ρ_{XZ}). Therefore λ can be defined as,

$$\lambda = \text{sign}\left(\rho_{XZ}\rho_{YZ}[1-u_X][1-u_Y]\right) \times \\ \frac{\text{sign}\left(1-u_X\right)\min\left(u_X, \frac{1}{u_X}\right) + \text{sign}\left(1-u_Y\right)\min\left(u_Y, \frac{1}{u_Y}\right)}{\min\left(u_X, \frac{1}{u_X}\right) + \min\left(u_Y, \frac{1}{u_Y}\right)}.$$

Although complex, the output of λ will be either -1, 0, or 1. We can then plug this factor into the full correction equation that provides us with an unbiased estimate of the correlation in the unrestricted population,

$$r_{XY} = r_{XY|\mathcal{S}(Z)}u_X u_Y + \lambda\sqrt{|1-u_X^2||1-u_Y^2|}$$

Similar to the univariate formula, we can also incorporate measurement error into the correction. Measurement error will bias the correlation on top of the bias induced by range restriction/enhancement. Therefore we can incorporate the reliabilities estimated within the restricted sample ($r_{XX'|\mathcal{S}(Z)}$ and $r_{YY'|\mathcal{S}(Z)}$), into our correction formula:

$$r_{TU} = \frac{r_{XY|\mathcal{S}(Z)}u_X u_Y + \lambda\sqrt{|1-u_X^2||1-u_Y^2|}}{\sqrt{1-u_X^2\left(1-r_{XX'|\mathcal{S}(Z)}\right)}\sqrt{1-u_Y^2\left(1-r_{YY'|\mathcal{S}(Z)}\right)}}$$

If the reliability estimates come from an target population reference sample, we can get estimates of the reliability coefficients in the selected sample using Equation 10.2 and Equation 10.3. We then can correct the observed sampling variance ($\sigma^2_{\varepsilon_o}$),

$$se(r_{TU}) = \frac{se\left(r_{XY|\mathcal{S}(Z)}\right)}{\hat{a}} = \frac{se\left(r_{XY|\mathcal{S}(Z)}\right)}{\left[\frac{r_{XY|\mathcal{S}(Z)}}{r_{TU}}\right]}.$$

Applied Example in R

Continuing with the example from the univariate case, a university admits applicants based on some criterion upon a composite of multiple measures of performance (Z). Two of those measures is a standardized test (X) and high school grade-point average (GPA). A researcher then wants to see how test performance predicts college GPA (Y). Since *both* test performance is correlated with the composite used in the college admissions process, it is likely that we will observe substantial range restriction in the college sample. The u-ratio in test performance and GPA was $u_X = u_Y = 0.80$ (applicants are the target population). We also want to correct for measurement error in the dependent variable (grade-point average) only we use to measure stress, therefore we obtain a reliability estimate within our sample of $r_{YY'|\mathcal{S}(Z)} = .90$. The researcher then conducts the study and finds a sample correlation of $r_{XY|\mathcal{S}(Z)} = .25$.
In R, we can correct the correlation by using the `correct_r` function in the `psychmeta` package

(Dahlke and Wiernik 2019). The `correction = 'uvirr_x'`

```
library(psychmeta)

correct_r(correction = 'uvirr_x',
          rxyi = .25,  # restricted correlation
          ryy = .90,  # reliability of GPA
          ux = .80,   # u ratio of SAT scores
          n = 100)    # sample size
```

```
Correlations Corrected for Measurement Error and Univariate Indirect Range Restriction:
--------------------------------------------------------------------------------------
  value CI_LL_95 CI_UL_95   n n_effective
1 0.323   0.0742    0.531 100        58.9
```

The true score correlation in the target population is estimated to be $r_{TU} = .32\,[.07,\ .53]$.

Beware of assumptions

Note that these corrections require the following assumptions: 1) linearity between X and Y, 2) homoskedasticity, that is, equal variance in Y at every level of X.

Part II

Application to Meta-Analysis

11 Introduction to Meta-Analysis Methods

11.1 Introduction

Meta-analysis is an analytic tool to synthesize quantitative evidence from multiple studies. By systematically combining and analyzing the results of multiple studies, meta-analysis provides a comprehensive overview, unveiling patterns, trends, and insights that individual studies might not be able to capture. Combining research findings also has the added benefit of increasing the precision of our results (i.e., greater statistical power). In this section we will cover the method described by (J. E. Hunter and Schmidt 2015) since it is readily compatible with artifact corrections (see next chapter). For the random-effects model however, we use an integrated approach that incorporates methods from J. E. Hunter and Schmidt (2015) and Hedges and Vevea (1998) that was first introduced by Morris et al. (2014). However it is important to note that there are other common methods to conduct meta-analyses that have their strengths and weaknesses (**hedges2014?**; Callender and Osburn 1980; Johnson, Mullen, and Salas 1995).

11.2 Common-Effect Model

A common effect model is the simplest form of meta-analysis. It assumes that all the variation in observed effect sizes is attributable to sampling error (see Figure 11.1). In other words, all the observed effect sizes are estimates of the same population effect size. Note that there is a distinction between *fixed-effects* models and a *common effect* model (Viechtbauer, n.d.; Laird and Mosteller 1990). The common effect model assumes that the true effect size is identical for each study while the fixed effects model does not assume this. Instead, the fixed effects model can be interpreted as the weighted average of true effects. Computationally, they are the same and provide the same parameter estimates, yet the interpretation differs (Viechtbauer, n.d.).

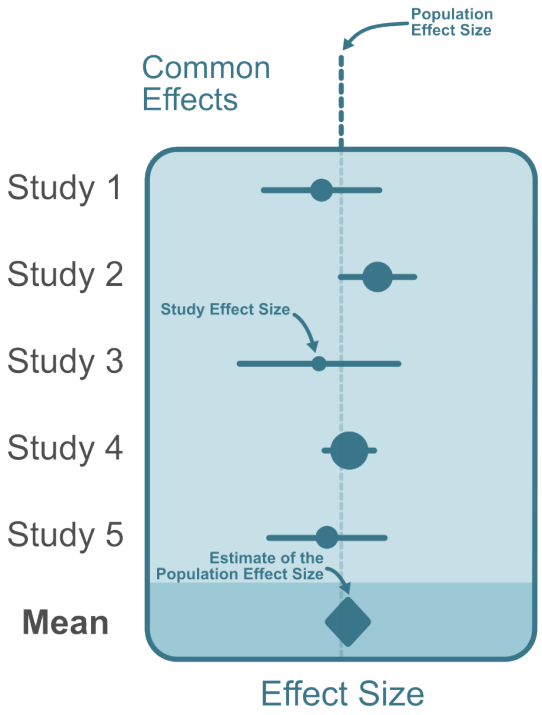


Figure 11.1: The diagram above depicts a common effect meta-analysis of five studies. The study effect sizes are homogenous and all estimate a single true population effect size.

The common effect model can be modeled such that population effect size θ is held constant each study effect size estimate (h_i), such that,

$$h_i = \theta + e_i \tag{11.1}$$

Where e_i indicates sampling error and the subscript i denotes each study. Similar to the true score theory model that we discussed in chapter 4, the variance components of each term can similarly be written out as,

$$\sigma^2_h = \sigma^2_\theta + \sigma^2_e$$

However in our common effect model, the population effect size is fixed across studies and will not vary, simplifying the formula to,

$$\sigma^2_h = \sigma^2_e \tag{11.2}$$

Therefore the only source of variation in the observed effect sizes, is sampling error.

Ultimately, our goal is to obtain a precise estimate of the population effect size. To obtain an estimate of the population effect size, θ, we can calculate the average observed effect size, $\bar{h}_i$ from k studies. However, in practice, effect sizes from different studies have varying levels of precision (i.e., different standard errors). A simple arithmetic average will not account for the differences between studies in their precision. Instead, we can calculate a weighted average where the weights each study can be calculated by the inverse variance (i.e., squared standard error) of each study such that,

$$w_i = \frac{1}{se(h_i)^2}.$$

Then we can calculate a weighted average,

$$\hat{\theta} = \frac{\sum_{i=1}^{k} w_i h_i}{\sum_{i=1}^{k} w_i}.$$

Where $\sum_{i=1}^{k}$ is the sum across all k studies. This weighted average will be an unbiased estimate of the population effect size. However, even though this mean effect size is more precise compared to single-study estimates, it is not exempt from error itself. We we can compute the standard error for $\hat{\theta}$ as,

$$se(\hat{\theta}) = \frac{1}{\sqrt{\sum_{i=1}^{k} w_i}}$$

The standard error can be used to compute the 95% confidence intervals (if the sampling distribution is approximately normal) of the meta-analytic point estimate:

$$CI_{95} = \hat{\theta} \pm 1.96 \cdot se(\hat{\theta})$$

Applied Example in R

Lets use a meta-analytic data set investigating the the effectiveness of a writing-to-learn intervention on academic achievement from Bangert-Drowns, Hurley, and Wilkinson (2004). This data set has standardized mean differences between the treatment group and a control group from $k = 48$ studies (total sample size: $n = 5,576$) and is available within the developmental version of the `metadat` package (White et al. 2022). Lets conduct a common effect meta-analysis using the equations from the previous section. We can use the `rma` function in the `metafor` package (Viechtbauer 2010) to conduct a common effect (`method = 'EE'`) meta-analysis without having to write each equation by hand.

```r
library(metadat)
library(metafor)

# display first 6 studies
head(dat.bangertdrowns2004[,c('author','year','ni','yi','vi')])
```

```
    author year  ni     yi    vi
1 Ashworth 1992  60  0.650 0.070
2    Ayers 1993  34 -0.750 0.126
3   Baisch 1990  95 -0.210 0.042
4    Baker 1994 209 -0.040 0.019
5   Bauman 1992 182  0.230 0.022
6   Becker 1996 462  0.030 0.009
```

```r
# fixed effects model
mdl <- rma(data = dat.bangertdrowns2004,
           yi = yi,
           vi = vi,
           method = 'EE') # Equal-effects = Common Effect

# print results
data.frame(theta_hat = mdl$b[1],
           se = mdl$se[1],
           CI_ll = mdl$ci.lb[1],
           CI_ul = mdl$ci.ub[1])
```

```
  theta_hat         se     CI_ll     CI_ul
1 0.1656264 0.02693136 0.1128419 0.2184109
```

The results show an estimated population effect of $\hat{\theta} = 0.17\,[0.11,\ 0.22]$.

11.3 Random Effects Model

The random-effects model refers to a model that allows for the population effect size to vary from study to study (see Figure 11.2). Random-effects differs from the fixed effects model in an important way: it does not assume that all observed effect sizes come from a single (fixed) population effect size (Borenstein et al. 2010). This variation in population effect sizes is called heterogeneity. In the traditional J. E. Hunter and Schmidt (2015) the weights utilized in the random effects meta-analysis are

identical to the common effect model. In other conventional meta-analysis methods (Hedges and Vevea 1998), random-effect weights usually include a random effect component containing the variation in population effect sizes (this has the effect of making study weights more similar to each other with more variation in population effects). A modern approach introduced by Morris et al. (2014) and later tested by Brannick et al. (2019), added this random effect component to the Hunter-Schmidt method. The simulation study by Brannick et al. (2019), concluded that weights incorporating random effect components improved the Hunter-Schmidt estimates. Here we will discuss Hedges-Vevea's method with some elements taken from Hunter-Schmidt.

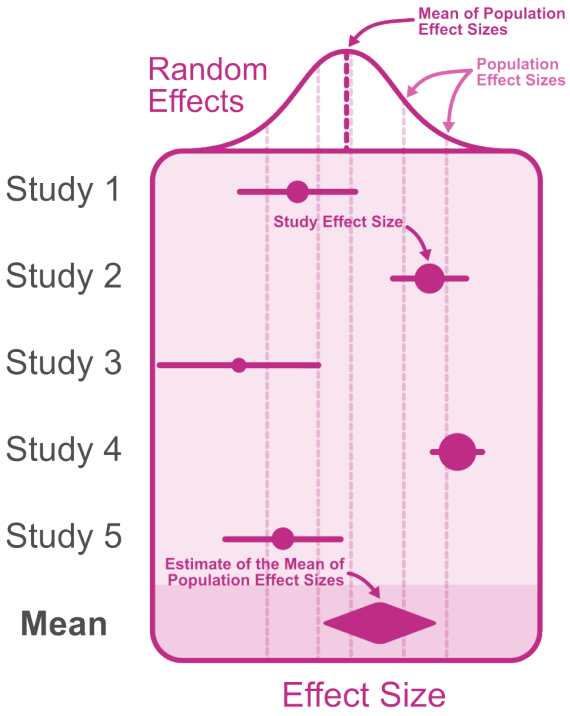


Figure 11.2: The diagram above depicts a random-effects meta-analysis of five studies. Effect sizes are more variable than the common effect meta-analysis since effect sizes vary due to sampling error *and* population effect sizes.

The model from Equation 11.1 can be changed slightly to encompass variation of the population effect size from study to study:

$$h_i = \theta_i + e_i.$$

In the common effect model, we assumed that all the variation in study effect sizes is accounted for by variation in sampling error (see Equation 11.2). However in the random-effects model the variance in

population effect sizes (σ^2_θ) is allowed to be greater than zero. The variance components can be written out as,

$$\sigma^2_h = \sigma^2_\theta + \sigma^2_e. \tag{11.3}$$

Estimating variance components can be done computationally through an iterative estimation procedure called REstricted Maximum Likelihood (REML) estimation. The estimated variance of population effect sizes, $\hat{\sigma}_\theta$, can now be incorporated into the inverse variance weights alongside

$$w_i = \frac{1}{se(h_i)^2 + \hat{\sigma}^2_\theta}$$

Calculating Standard Errors

The standard error for a given study often uses the sample estimate in the sampling variance equation. For example, a study correlation, r_i, will have a sampling variance of

$$se(r_i)^2 = \frac{(1-\rho_i^2)^2}{n}$$

Where ρ_i is the population effect size for that study. Since the population effect size is not known, ρ_i is often replaced with the study estimate, r_i (also the denominator will be adjusted to be $n-1$ for the decreased degrees of freedom). However this makes the sampling variances and estimates dependent on one another which can cause problems for other types analyses (e.g., funnel plot asymmetry). Instead, we will use the sample size weighted mean correlation for each study to replace ρ_i (J. E. Hunter and Schmidt 2015),

$$se(r_i)^2 = \frac{\left(1-\bar{r}^2\right)^2}{n-1}$$

Where,

$$\bar{r} = \frac{\sum_{i=1}^{k} n_i r_i}{\sum_{i=1}^{k} n_i}$$

In studies with no heterogeneity ($\sigma_\theta = 0$) this is nearly optimal, but it still works well in cases of substantial heterogeneity.

Now we can estimate the mean of population effects by taking the weighted average effect size (equation 16.5, Cooper, Hedges, and Valentine 2009),

$$\hat{\bar{\theta}} = \frac{\sum_{i=1}^{k} w_i h_i}{\sum_{i=1}^{k} w_i}.$$

Where $i = 1...k$ studies. The standard error of the mean of population effects can calculated from the summation of inverse weights (equation 16.6, Cooper, Hedges, and Valentine 2009),

$$se(\hat{\bar{\theta}}) = \frac{1}{\sqrt{\sum_{i=1}^{k} w_i}}.$$

The 95% confidence interval can then be calculated as,

$$CI = \hat{\bar{\theta}} \pm 1.96 \cdot se(\hat{\bar{\theta}})$$

In other conventions, the variance in population effects (σ^2_θ) is denoted as τ^2 (Borenstein et al. 2010; DerSimonian and Kacker 2007; Hedges and Vevea 1998), but conceptually σ^2_θ and τ^2 these are identical. Taking the root of σ^2_θ is the standard deviation of population effect sizes which can be a useful measure of heterogeneity. Furthermore, we can use $\hat{\sigma}_\theta$ to calculate credibility (prediction) intervals which allows us to draw inferences about the range of plausible population effect sizes. For example, the 90% credibility interval can be calculated with the following equations:

$$CR = \hat{\bar{\theta}} \pm 1.645\hat{\sigma}_\theta$$

The confidence interval and credibility interval have fundamentally different interpretations that are often misinterpreted in published work (Whitener 1990). When we are interpreting a single realized interval (i.e., our estimate-in-hand), the 90% *credibility* interval can be interpreted as the region in which 90% of population effect sizes exist, however, a 95% *confidence* interval describes the interval in which there is a 95% probability of containing the true *mean* of population effect sizes. It is important to note that the confidence interval interpretation here is only valid in the case of a single realized interval (Vos and Holbert 2022), if there is more than one interval obtained from the same population of studies, then the interpretation does not hold (this would be a rare in a meta-analysis).

Applied Example in R

Let's continue looking at the meta-analysis from Bangert-Drowns, Hurley, and Wilkinson (2004). This data set has standardized mean differences between the treatment group and a control group from $k = 48$ studies (total sample size: $n = 5,576$) and is available within the developmental version of the `metadat` package (White et al. 2022). Lets conduct a common effect meta-analysis using the equations from the previous section. We can use the `rma` function in the `metafor` package (Viechtbauer 2010) to conduct a random effect meta-analysis with REML estimation we can use the `method = 'REML'` argument.

```r
library(metadat)
library(metafor)

# display first 6 studies
head(dat.bangertdrowns2004[,c('author','year','ni','yi','vi')])
```

```
    author year  ni     yi    vi
1 Ashworth 1992  60  0.650 0.070
2    Ayers 1993  34 -0.750 0.126
3   Baisch 1990  95 -0.210 0.042
4    Baker 1994 209 -0.040 0.019
5   Bauman 1992 182  0.230 0.022
6   Becker 1996 462  0.030 0.009
```

```r
# fixed effects model
mdl <- rma(data = dat.bangertdrowns2004,
           yi = yi,
           vi = vi,
           method = 'REML') # Equal-effects = Common Effect

# print results
data.frame(theta_hat = mdl$b[1],
           se = mdl$se[1],
           CI_ll = mdl$ci.lb[1],
           CI_ul = mdl$ci.ub[1])
```

```
  theta_hat         se     CI_ll     CI_ul
1 0.2219296 0.04603453 0.1317036 0.3121556
```

The results show an estimated population effect of $\hat{\theta} = 0.22\,[0.13,\, 0.31]$.

12 Artifact Correction Meta-Analysis

12.1 Introduction

Artifact correction meta-analysis, also referred to as psychometric meta-analysis, is a form of meta-analysis where effect sizes are systematically corrected for sources of bias. These sources of bias have been discussed in previous chapters 4-10. Methodology for conducting artifact correction style meta-analyses were originally pioneered by Frank Schmidt and John Hunter (2015; 1977) and then reviewed more recently by Brenton Wiernik and Jeffrey Dahlke (2020). There has also been powerful R packages developed to aide in the application of artifact correction meta-analyses that we have used in previous chapters (Dahlke and Wiernik 2019). You will notice that in this section, we do not discuss standardized mean differences. This is due to the fact that the artifact correction model is designed for pearson correlations, in order to use this method for standardized mean differences, convert to pearson correlations using the methods described in chapter 11, and then use the correction methods used below. Once you apply the corrections to the converted correlations they can then be converted back to a standardized mean difference.

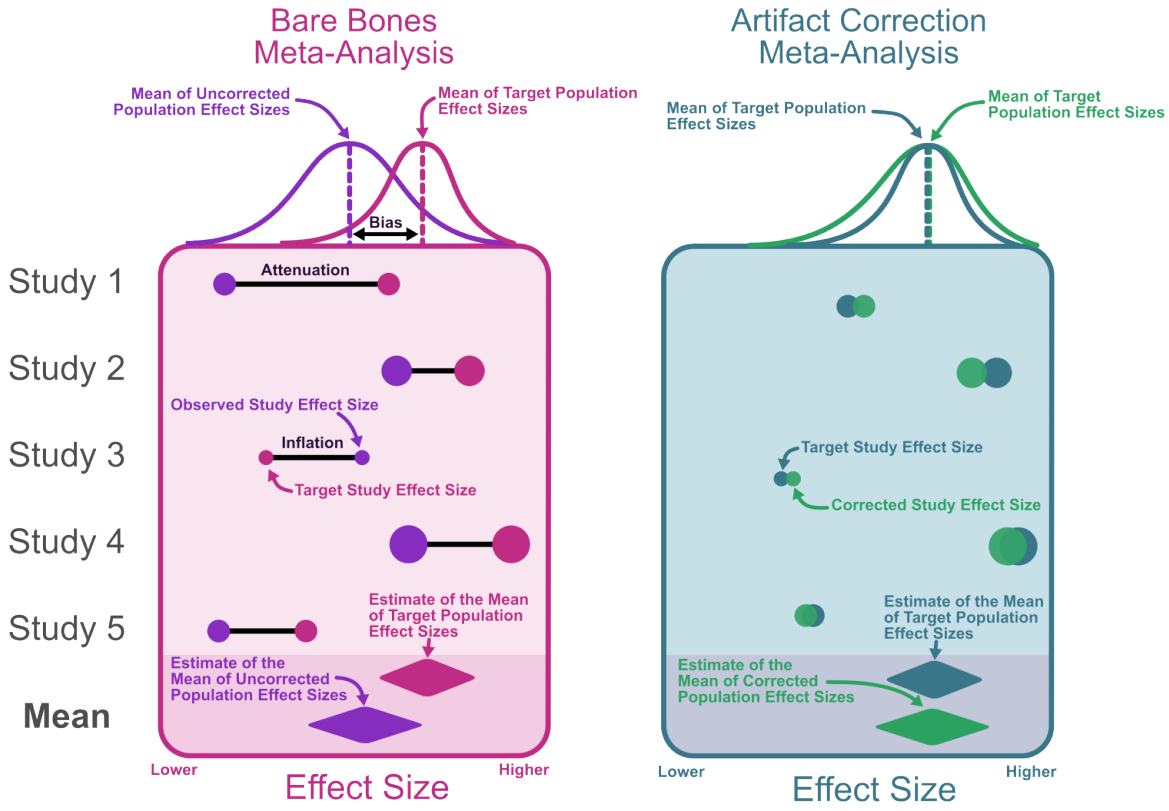
12.2 Bare Bones vs Artifact Correction Meta-Analysis

Chapter 11 focused on bare-bones meta-analysis, that is, meta-analyses that do not correct for biases in effect size estimates. This section will be dedicated to the artifact correction approach to meta-analysis that aims to correct for statistical artifacts. The choice between these two types of meta-analyses depends on the research question, the available data, and the assumptions researchers are willing to make. If the goal is to investigate effect sizes as they are reported while avoiding additional assumptions about the data, then a bare-bones meta-analysis might be the way to go. On the other hand, if the goal is to obtain a more accurate estimate of the target effect size by accounting for biases induced by statistical artifacts, an artifact correction meta-analysis is preferable.

1. Bare-Bones Meta-Analysis: In a bare-bones meta-analysis, the focus is on aggregating effect sizes from various studies without explicitly correcting for potential biases in these effect size estimates (see first panel of **?@fig-art-corr-mdl**).
2. Artifact Correction Meta-Analysis: In contrast, an artifact correction meta-analysis takes into account and attempts to correct for biases that may be present in the effect size estimates from individual studies. This involves addressing potential sources of bias, such as measurement error

or selection effects, through statistical techniques or adjustments. By doing so, the meta-analysis aims to provide a more accurate and unbiased estimate of the true effect size. Although it is important to note that this method will require additional assumptions about the nature of the data (see second panel of **?@fig-art-corr-mdl**).

Note that the bare-bones model does not assume that there is no bias, rather, the bare-bones model is estimating something else entirely, that is, the population effect size.



12.3 Individual Artifact Correction Model

The individual artifact correction model corrects each effect size individually prior to conducting the meta-analysis. Let us recall the random effects model in chapter 11, that an effect size h_i from study i can be modeled such that, $h_i = \theta_i + e_i$. This model would be considered a *bare-bones* meta-analytic model. In practice, observed (study) effect sizes (h_i) tend to be biased relative to our target quantity due to many artifacts, some that we can account for and some we can not. If we decide that corrections to observed effect sizes are necessary to answer our research question, then we can construct an artifact correction model. In the artifact correction model, we can incorporate an artifact attenuation/inflation factor, a, to the bare-bones formula such that,

$$h_i = a_i\theta_i + e_i \tag{12.1}$$

The attenuation/inflation factor, a_i, must be estimated for each study, i. Using estimates of a, Equation 12.1 can be re-arranged to obtain unbiased estimates of the target population effect size:

$$\frac{h_i}{\hat{a}_i} = \theta_i + \frac{e_i}{\hat{a}_i} \tag{12.2}$$

This division of $\hat{a}_i$ will provide us with our corrected effect size estimates that we can denote with the subscript, c,

$$h_{c_i} = \frac{h_i}{\hat{a}_i}$$

and of course we must also correct the sampling error term,

$$e_{c_i} = \frac{e_i}{\hat{a}_i}. \tag{12.3}$$

Therefore Equation 12.2 can be expressed as,

$$h_{c_i} = \theta_i + e_{c_i} \tag{12.4}$$

Like we saw in chapter 11, we can breakdown the variance components of the model,

$$\sigma^2_{h_c} = \sigma^2_\theta + \sigma^2_{e_c} \tag{12.5}$$

To obtain these variance components, we can start by correcting the standard errors from each study. We can calculate the corrected standard error of the corrected correlations,

$$se(h_{c_i}) = \frac{se(h_i)}{\hat{a}_i}$$

Note that correcting the standard error by dividing by $\hat{a}$ ignores any error in the estimation of a This may also be done by using the corrected effect size and the observed effect size to correct the sampling variance:

$$\sigma^2_{\varepsilon_c i} = \sigma^2_{\varepsilon_o i}\left(\frac{\theta_{c_i}}{\theta_{o_i}}\right)^2$$

The next step is to obtain the random effects weights of the study, we can do this with the inverse corrected variance for each study, $w_i = 1/(\sigma^2_{\varepsilon_c i} + \sigma^2_\rho)$. From here we can calculate our estimate of the mean of true population correlations,

$$\hat{\bar{\vartheta}} = \frac{\sum_{i=1}^{k} n_i \theta_{c_i}}{\sum_{i=1}^{k} n_i}$$

Remember that because this is a random effects model, $\hat{\bar{\vartheta}}$ is not an estimate of the true population effect size, instead it is an estimate of the mean of a distribution of true population effect sizes. Now that we have an estimate of the mean and the corrected sampling variances, the variance components from Equation 12.5 can be easily calculated as follows:

$$\sigma^2_{\theta_c} = \frac{\sum_{i=1}^{k} n_i (\theta_{c_i} - \hat{\bar{\vartheta}})^2}{\sum_{i=1}^{k} n_i}$$

$$\sigma^2_{\varepsilon_c} = \frac{\sum_{i=1}^{k} n_i \sigma^2_{\varepsilon_c i}}{\sum_{i=1}^{k} n_i}$$

$$\sigma^2_\vartheta = \sigma^2_{\theta_c} - \sigma^2_{\varepsilon_c}$$

The standard deviation of true effects is a useful measure of heterogeneity and is simply the square root of the variance of true population effect sizes (σ_ϑ). From the standard deviation in true effects, we can also calculate a credibility (prediction) interval that shows the range of plausible values for which a true effect size is likely to fall,

$$\vartheta_{\text{Upper}} = \hat{\bar{\vartheta}} + 1.645\sigma_\vartheta$$

$$\vartheta_{\text{Lower}} = \hat{\bar{\vartheta}} - 1.645\sigma_\vartheta \, .$$

Note that this is not to be confused with *confidence* intervals which denotes the range of plausible values that the *mean* of true effects can take on. This differentiation is akin to understanding the disparity between the standard error of the mean and the standard deviation in the context of a normal distribution. We can also see how the corrections reduced the heterogeneity in the effect size estimates by comparing variance in true effect sizes ($\sigma^2_{\vartheta_o}$) to the variance in observed effect sizes ($\sigma^2_{\vartheta_o}$; this can be calculated by conducting a bare-bones random effects meta-analysis described in chapter 11). The percent reduction in heterogeneity can be computed by taking the ratio of the two, $\sigma^2_\vartheta / \sigma^2_{\vartheta_o}$. J. E. Hunter and Schmidt (2015) suggest that if 75% of the heterogeneity is accounted for by artifact corrections, then we can assume that the remaining heterogeneity is attributable to remaining artifacts that have not

been addressed in the current meta-analysis. Although it is important to point out that this is simply a rule of thumb rather than a mathematical property.

12.3.1 Individual Corrections in Correlations

For correlation coefficients we can define the model similarly to Equation 12.1, with the only difference being that we will use the notation for pearson correlations,

$$r_{o_i} = A_i \rho_i + \varepsilon_{o_i}$$

The artifact correction formulation of this, corresponding to Equation 12.4, would be

$$r_{c_i} = \rho_i + \varepsilon_{c_i}$$

The corresponding variance components would then be,

$$\sigma^2_{r_c} = \sigma^2_{\rho} + \sigma^2_{\varepsilon_c}$$

In order to compute the variance components as well as the mean true population correlation, we first need to calculate the study weights. We will follow a similar procedure for calculating random effects weights in chapter 11. Lets define the corrected random effects weights as,

$$w_i = \frac{1}{\sigma^2_{\varepsilon_c i} + \sigma^2_{\rho}}.$$

However the variance components, $\sigma^2_{\varepsilon_c i}$ and σ^2_{ρ}, require the weights themselves to actually estimate them, so instead we can approximate the variance components using the sample size as the weights such that,

$$w_i = \frac{1}{\sigma^2_{\varepsilon_c i} + \sigma^2_{\rho}} = \frac{1}{\sigma^2_{\varepsilon_c i} + (\sigma^2_{r_c} - \sigma^2_{\bar{\varepsilon}})} \approx \frac{1}{\sigma^2_{\varepsilon_c i} + \left(\frac{\sum_{i=1}^k n_i (r_{c_i} - \bar{r}_c)^2}{\sum_{i=1}^k n_i} - \frac{\sum_{i=1}^k n_i \sigma^2_{\varepsilon_c i}}{\sum_{i=1}^k n_i} \right)}$$

Where $\bar{r}_c$ is the sample size weighted average corrected correlation. These weights can then be used to obtain a more precise estimate of the true population correlation,

$$\hat{\bar{\rho}} = \frac{\sum_{i=1}^k w_i r_{c_i}}{\sum_{i=1}^k w_i}$$

Now we can compute each of the three variance components:

1) Variance in corrected correlations:

$$\sigma^2_{r_c} = \frac{\sum_{i=1}^{k} w_i (r_{c_i} - \hat{\bar{\rho}})}{\sum_{i=1}^{k} w_i}.$$

2) Sampling error variance:

$$\sigma^2_{\varepsilon_c} = \frac{\sum_{i=1}^{k} w_i \sigma^2_{\varepsilon_c i}}{\sum_{i=1}^{k} w_i}.$$

3) Variance in population correlations:

$$\sigma^2_{\rho} = \sigma^2_{r} - \sigma^2_{\varepsilon}.$$

Now lets use these variance components to calculate the 90% credibility (prediction) interval and the 95% confidence interval. The 90% credibility interval can be calculated with the following equations:

$$\rho_{\text{Upper}} = \hat{\bar{\rho}} + 1.645\sigma_{\rho}$$

$$\rho_{\text{Lower}} = \hat{\bar{\rho}} - 1.645\sigma_{\rho}$$

We can also calculate the standard error of the mean of true population effect sizes ($SE_{\hat{\bar{\rho}}}$) by dividing the sampling error variance component by the number of studies, k,

$$SE_{\hat{\bar{\rho}}} = \sqrt{\frac{\sigma^2_{r_c}}{k}}$$

Which can then be used to calculate 95% confidence intervals:

$$\bar{\rho}_{\text{Upper}} = \hat{\bar{\rho}} + 1.96 \cdot SE_{\hat{\bar{\rho}}}$$

$$\bar{\rho}_{\text{Lower}} = \hat{\bar{\rho}} - 1.96 \cdot SE_{\hat{\bar{\rho}}}$$

12.3.2 Applied Example in R

Lets conduct an individual correction meta-analysis in r using the data set by Roth (2015). This data set consists of correlations between school grades and intelligence test scores. It also contains information on the reliability of the intelligence test scores and the extent of range restriction in test scores. We can conduct a meta-analysis correcting for univariate indirect range restriction and measurement error in test scores. The compound artifact biasing factor for the correlation would be:

$$A_i = \sqrt{r_{o_i}^2 + \frac{u_{x_i}^2 r_{xx'_i}(r_{xx'_i} - r_{o_i}^2)}{1 - u_{x_i}^2(1 - r_{xx'_i})}}$$

Sticking with our theme of doing everything in base R first, lets use the equations from the previous section to conduct the meta-analysis.

```
# Load in packages (we need the development version of psychmeta)
# install.packages("devtools")
# devtools::install_github("psychmeta/psychmeta")
library(psychmeta)

# obtain artifact values
rxx <- data_r_roth_2015$rxxi
ux <- data_r_roth_2015$ux
ro <- data_r_roth_2015$rxyi
n <- data_r_roth_2015$n
k <- length(ro)

# fill in missing artifact values with mean
rxx[is.na(rxx)] <- mean(rxx,na.rm=TRUE)
ux[is.na(ux)] <- mean(ux,na.rm=TRUE)

# calculate compound artifact biasing factor for univariate direct range restriction with
A <- sqrt(ro^2 + (ux^2*rxx*(rxx - ro^2)) / (1 - ux^2*(1-rxx)))

# calculate the sample size weighted average of r
ro_bar <- sum(ro*n) / sum(n)

# calculate the observed sampling variance for each study
var_eoi <- (1-ro_bar^2)^2 / (n-1)

# correct sampling variance
var_eci <- var_eoi / A^2
```

```r
# calculate corrected correlations
rc <- ro / A

# calculate weights
w <- 1/var_eci

# calculate population effect size estimate
mean_rho_hat <- sum(rc*w) / sum(w)

# calculate the variance in corrected correlations (rc)
var_rc <- sum(w*(rc - mean_rho_hat)^2) / sum(w)

# calculate average corrected sampling variance
var_ec <- sum(var_eci*w) / sum(w)

# calculate the variance in true population correlations
var_rho <- var_rc - var_ec

# calculate standard error of rho estimate
SE_rho = sqrt(var_rc/k)

# print results
data.frame(k = k,
           n = sum(n),
           mean_rho_hat,
           SE = SE_rho,
           SD_rho = sqrt(var_rho))
```

```
    k      n mean_rho_hat         SE    SD_rho
1 240 105151    0.5398838 0.01339916 0.2022865
```

The estimated mean correlation of .540 is precisely what is precisely what the original paper reported (Roth 2015). Lets conduct the meta-analysis using the the `psychmeta` package (Dahlke and Wiernik 2019). The function `ma_r_ic` is designed to conduct an individual correction meta-analysis on correlation coefficients.

```r
# install.packages('psychmeta')
library(psychmeta)

# conduct individual correction meta-analysis
```

```
mdl_ic <- ma_r_ic(rxyi = ro, n = n,
            correction_method = "uvirr",
            rxx = rxx,
            ux = ux,
            ux_observed = TRUE,
            rxx_restricted = TRUE)

summary_stats <- data.frame(k = mdl_ic$meta_tables$`analysis_id: 1`$individual_correction$
                            n = mdl_ic$meta_tables$`analysis_id: 1`$individual_correction$
                            mean_rho = mdl_ic$meta_tables$`analysis_id: 1`$individual_corr
                            SE = mdl_ic$meta_tables$`analysis_id: 1`$individual_correction
                            SD_rho = mdl_ic$meta_tables$`analysis_id: 1`$individual_correc
summary_stats
```

```
  k      n  mean_rho        SE    SD_rho
1 240 105151 0.5404134 0.0134356 0.2036946
```

We can also obtain credibility intervals by using the `credibility` function in the `psychmeta` package. The interval defaults to 80% intervals, however we can change that to 90% by inputting .90 into the `cred_level` argument.

```
credibility(mean = summary_stats$mean_rho_hat,
            sd = summary_stats$SD_rho,
            cred_method = "norm",
            cred_level = .90)
```

```
   CR_LL_90 CR_UL_90
```

Lets compare these results to the bare-bones model. In `psychmeta` the bare-bones model can be conduced using `ma_r_bb`. However, the `ma_r_ic` function also reports the bare-bones results as well. Therefore we can just extract the necessary statistics from the model.

```
data.frame(
  k = mdl_ic$meta_tables$`analysis_id: 1`$barebones$k,
  n = mdl_ic$meta_tables$`analysis_id: 1`$barebones$N,
  mean_rho_obs = mdl_ic$meta_tables$`analysis_id: 1`$barebones$mean_r,
  SE = mdl_ic$meta_tables$`analysis_id: 1`$barebones$se_r,
  SD_rho_obs = mdl_ic$meta_tables$`analysis_id: 1`$barebones$sd_r)
```

```
    k      n mean_rho_obs         SE SD_rho_obs
1 240 105151    0.4418789 0.01191933  0.1846534
```

We can see that the estimate of the population correlation is largely attenuated in the observed values. This is due to the fact tests of intelligence are not perfectly reliable and the scores were restricted in their range.

12.4 Artifact Distribution Model

When we observe a lot of missingness in artifact values (e.g., studies not reporting reliability), we may choose to use an artifact distribution model. The artifact distribution model conducts a meta-analysis on the observed effect sizes and artifact values separately, and then uses the aggregate artifact values to correct for the observed mean effect size. Since the artifact distribution method uses Taylor series approximations (Dahlke and Wiernik 2020) that are custom-tailored to estimate the sampling variance of corrected correlations, we will skip the general case to focus on its application to correlations.

12.4.1 The Correlational Case

The model here can be broken down into two parts, the first part aggregates the observed effect sizes and the second part aggregates the artifact values. The artifact values we will focus on here are the reliability coefficients (see chapter 5 and 6), however other artifact values like u-ratios will follow similar procedures. We can start with the bare-bones meta-analysis model: $r_{o_i} = \rho_{o_i} + \varepsilon_{o_i}$. We can estimate the observed population correlation (ϑ_{o_i}) by first calculating the weights (using the n-weighted mean correlation in the formula for sampling variance, $\bar{r}$):

$$\sigma^2_{\varepsilon_o i} \approx \frac{(1-\bar{r}^2)^2}{n_i - 1}$$

$$w_i = \frac{1}{\sigma^2_{\varepsilon_i} + \sigma^2_{\vartheta}} = \frac{1}{\sigma^2_{\varepsilon_i} + (\sigma^2_{\theta} - \sigma^2_{\varepsilon})} \approx \frac{1}{\sigma^2_{\varepsilon_i} + \left(\frac{\sum_{i=1}^k n_i(\theta_i - \bar{\theta})^2}{\sum_{i=1}^k n_i} - \frac{\sum_{i=1}^k n_i \sigma^2_{\varepsilon_o i}}{\sum_{i=1}^k n_i} \right)}$$

Taking the mean of the observed study correlations weighted by the inverse sampling variance,

$$\hat{\bar{\rho}}_o = \frac{\sum_{i=1}^k w_i r_{o_i}}{\sum_{i=1}^k w_i}$$

Then lets get the variance in observed population correlations, in order to do this we need the v

$$\sigma^2_{\rho_o} = \sigma^2_{r_o} - \sigma^2_{\varepsilon_o} = \frac{\sum_{i=1}^{k} w_i (r_{o_i} - \hat{\bar{\rho}}_o)^2}{\sum_{i=1}^{k} w_i} - \frac{\sum_{i=1}^{k} w_i \sigma^2_{\varepsilon_o i}}{\sum_{i=1}^{k} w_i}$$

With the weights we can also take the weighted average of the artifact values (such as u-ratios or reliabilities) that are available. For our example here, we will correct only for measurement error, therefore the weighted means for reliability in x and y will be:

$$\bar{r}_{xx'} = \frac{\sum_{i=1}^{k} w_i r_{xx'_i}}{\sum_{i=1}^{k} w_i}$$

$$\bar{r}_{yy'} = \frac{\sum_{i=1}^{k} w_i r_{yy'_i}}{\sum_{i=1}^{k} w_i}$$

Now recall from chapter 5 that the square root of the reliability is equal to the correlation between observed scores and true scores. We can denote the mean correlation as follows: $\bar{r}_{xT} = \sqrt{\bar{r}_{xx'}}$ and $\bar{r}_{yU} = \sqrt{\bar{r}_{xx'}}$. We then must also compute the average sampling variances of r_{xT_i} and r_{yU_i} between studies. These sampling variance of these correlations can be computed the same way as a pearson correlation:

$$\sigma^2_{r_{xT}i} \approx \frac{(1 - \bar{r}^2_{xT})^2}{n_i - 1}$$

$$\sigma^2_{r_{yU}i} \approx \frac{(1 - \bar{r}^2_{yU})^2}{n_i - 1}$$

Then weighted average of these sampling variances is

$$\sigma^2_{r_{xT}} = \frac{\sum_{i=1}^{k} w_i r_{xT_i}}{\sum_{i=1}^{k} w_i}$$

$$\sigma^2_{r_{yU}} = \frac{\sum_{i=1}^{k} w_i r_{yU_i}}{\sum_{i=1}^{k} w_i}$$

Now that we have the point-estimate of the population observed correlation, the variance of observed population correlations, the sampling variance of observed correlations, and the sampling variance of the square root of the reliability for x and y, we can now attempt to correct the point-estimate and the variance of population correlations.

12.4.1.0.1 * Correcting Using Summary Values

First, we can start by correcting the overall point-estimate for the observed population correlation in order to remove bias due to measurement error. Recall from chapter 5 the correction formula:

$$\hat{\bar{\rho}} = \frac{\hat{\bar{\rho}}_o}{\bar{A}} = \frac{\hat{\bar{\rho}}_o}{\bar{r}_{xT}\bar{r}_{yU}} = \frac{\hat{\bar{\rho}}_o}{\sqrt{\bar{r}_{xx'}}\sqrt{\bar{r}_{yy'}}}$$

Note that the artifact biasing factor, A, is the product of the two sources of attenuation. Correcting the variance in observed population correlations ($\sigma^2_{\rho_o}$), so that it is accurately estimating the variance of true population effect sizes (σ^2_ρ), we must use a Taylor series approximation. This formula can become fairly complex with more types of artifacts involved. The taylor series approximation is for estimating specifically the amount of sampling variance within the correction factor we apply to the observed correlation. The first step is lay out our attenuation formula (the equation where observed effect size is on the left side of the equation and the artifact values and true effect size is on the right hand side of the equation). In the case of correcting only for measurement error, the attenuation formula is relatively simple

$$\hat{\bar{\rho}}_o = \hat{\bar{\rho}} \cdot \bar{r}_{xT} \cdot \bar{r}_{yU}$$

For the taylor series approximation, we want to first find the partial derivitive with respect to each artifact component:

$$B_{r_{xT}} = \frac{\partial}{\partial r_{xT}}(\hat{\bar{\rho}} \cdot \bar{r}_{xT} \cdot \bar{r}_{yU}) = \hat{\bar{\rho}} \cdot \bar{r}_{yU}$$

$$B_{r_{yU}} = \frac{\partial}{\partial r_{yU}}(\hat{\bar{\rho}} \cdot \bar{r}_{xT} \cdot \bar{r}_{yU}) = \hat{\bar{\rho}} \cdot \bar{r}_{xT}$$

The variance due to artifacts is then approximately,

$$\sigma^2_A \approx B^2_{r_{xT}}\sigma^2_{r_{xT}} + B^2_{r_{yU}}\sigma^2_{r_{yU}}$$

Now we can approximate the variance in true population correlations,

$$\sigma^2_\rho = \frac{\sigma^2_{\rho_o} - \sigma^2_A}{\bar{A}^2}$$

Where the artifact biasing factor is: $\bar{A} = \bar{r}_{xT} \cdot \bar{r}_{yU}$. See the supplementary materials of Dahlke and Wiernik (2020) for detailed Taylor series approximation derivations for the immensely more complicated bivariate indirect range restriction plus measurement error correction.

12.4.2 Applied Example in R

Lets conduct an artifact distribution correction meta-analysis in R, instead using data from the meta-analysis by McDaniel et al. (1994). This dataset contains correlations between employment interviews and job performance. This data set has a lot of missing values for reliability coefficients and u-ratios which might suggest that the artifact distribution approach is a better choice compared to the individual correction approach. We can conduct a meta-analysis correcting for univariate indirect range restriction and measurement error in both job performance and employment interviews. The attenuation formula will be important for calculating the Taylro series approximation can be defined as

$$\bar{\rho}_o = \bar{\rho}\sqrt{\bar{r}^2_{o_i} + \frac{\bar{u}^2_{x_i}\bar{r}_{xx'_i}(\bar{r}_{xx'_i}\bar{r}_{yy'_i} - \bar{r}^2_{o_i})}{1 - \bar{u}^2_{x_i}(1 - \bar{r}_{xx'_i})}}$$

Instead of conducting a taylor series approximation by hand, we will simply use the `psychmeta` package to perform the artifact distribution meta-analysis. The function `ma_r_ad` is designed to conduct an artifact distribution meta-analysis on correlation coefficients. The function also reports the bare-bones model allowing us to compare the corrected estimates to the uncorrected.

```
# Load in packages (we need the development version of psychmeta)
# install.packages("devtools")
# devtools::install_github("psychmeta/psychmeta")
library(psychmeta)

# obtain artifact values
rxx <- data_r_roth_2015$rxxi
ux <- data_r_roth_2015$ux
ro <- data_r_roth_2015$rxyi
n <- data_r_roth_2015$n
k <- length(ro)

# compute barebones meta-analysis
ma_obj <- ma_r_bb(r = rxyi,
                  n = n,
                  correct_bias = FALSE,
                  wt_type = "REML",
                  data = data_r_mcdaniel_1994)

# construct artifact distribution for x
ad_obj_x <- create_ad(ad_type = "tsa",
                      mean_rxxi = data_r_mcdaniel_1994$Mrxxi[1],
                      var_rxxi = data_r_mcdaniel_1994$SDrxxi[1]^2,
                      ux = data_r_mcdaniel_1994$ux,
```

```
                      wt_ux = data_r_mcdaniel_1994$`ux frequency`)

# construct artifact distribution for y
ad_obj_y <- create_ad(ad_type = "tsa",
                      rxxi = data_r_mcdaniel_1994$ryyi,
                      wt_rxxi = data_r_mcdaniel_1994$`ryyi frequency`)

# compute artifact-distribution meta-analysis, correcting for measurement error only
mdl_ad <- ma_r_ad(ma_obj = ma_obj,
                  ad_obj_x = ad_obj_x,
                  ad_obj_y = ad_obj_y,
                  correction_method = "meas")


# summary table of meta-analysis
summary_stats <- data.frame(
  type = c('Artifact Distribution', 'Bare-Bones'),
  k = c(mdl_ad$meta_tables$`analysis_id: 1`$artifact_distribution$true_score$k,mdl_ad$meta
  n = c(mdl_ad$meta_tables$`analysis_id: 1`$artifact_distribution$true_score$N,mdl_ad$meta
  mean_rho = c(mdl_ad$meta_tables$`analysis_id: 1`$artifact_distribution$true_score$mean_r
  SE = c(mdl_ad$meta_tables$`analysis_id: 1`$artifact_distribution$true_score$se_r_c,mdl_a
  SD_rho = c(mdl_ad$meta_tables$`analysis_id: 1`$artifact_distribution$true_score$sd_rho,0

summary_stats
```

```
                  type   k     n  mean_rho         SE    SD_rho
1 Artifact Distribution 160 25244 0.3201828 0.02108407 0.1985571
2            Bare-Bones 160 25244 0.2205043 0.01452023 0.0000000
```

We can also obtain credibility intervals by using the `credibility` function in the `psychmeta` package. The interval defaults to 80% intervals, however we can change that to 90% by inputting .90 into the `cred_level` argument.

```
credibility(mean = summary_stats$mean_rho[1],
            sd = summary_stats$SD_rho[1],
            cred_method = "norm",
            cred_level = .90)
```

```
        CR_LL_90  CR_UL_90
[1,] -0.006414571 0.6467802
```

Lets compare these results to the bare-bones model. In `psychmeta` the bare-bones model can be conduced using `ma_r_bb`. However, the `ma_r_ic` function also reports the bare-bones results as well. Therefore we can just extract the necessary statistics from the model.

Part III

Conclusion

13 Conclusion

13.0.0.0.1 * Concluding Remarks

Artifacts contaminate research findings, obscure our view of reality, inhibit us from making sense of our observations, and they mislead us. But by addressing these biases at the source, we clear a path toward seeing things as they really are.

This book could not have been accomplished without the pioneering work of John Hunter and Frank Schmidt. This book stands as a testament to their foundational contributions to artifact corrections and meta-analysis. Notably, Hunter and Schmidt's (2015) influential book, laid the bedrock upon which subsequent advancements in artifact corrections were built.

The goal of this current book was to not only honor past work, but also to expand, refine, and augment the methodologies. This work aims to build upon the work of Hunter and Schmidt by offering an extended repertoire of techniques and insights to further enhance the understanding of statistical artifacts. It also aims to provide a comprehensive guide, drawing upon previous scholarship while adapting to the evolving landscape of research methodologies.

I hope reading this book will provide invaluable insights and tools that empower people to address the pervasive bias in our studies. May it encourage a deeper awareness of these issues when confronted with them.

We want our research findings to accurately describe reality. Artifact corrections help us get closer to that goal.

13.0.0.0.2 * Acknowledgements

Thank you to Dr. Blair T. Johnson, Dr. Christopher Rhoads, and Dr. Elizabeth Schifano for taking the time to review and provide extremely valuable feedback over the course of writing this book.

Matthew B. Jané

@MatthewBJane

Part IV

References

References

Aguinis, Herman, Charles A Pierce, and Steven A Culpepper. 2009. "Scale Coarseness as a Methodological Artifact," September.

Bangert-Drowns, Robert L, Marlene M Hurley, and Barbara Wilkinson. 2004. "The Effects of School-Based Writing-to-Learn Interventions on Academic Achievement: A Meta-Analysis." *Review of Educational Research* 74 (1): 29–58. https://doi.org/10.3102/00346543074001029.

Bishara, Anthony J., and James B. Hittner. 2017. "Confidence Intervals for Correlations When Data Are Not Normal." *Behavior Research Methods* 49 (1): 294–309. https://doi.org/10.3758/s13428-016-0702-8.

Bobko, Philip, and Angela Rieck. 1980. "Large Sample Estimators for Standard Errors of Functions of Correlation Coefficients." *Applied Psychological Measurement* 4 (3): 385–98. https://doi.org/10.1177/014662168000400309.

Bonett, Douglas G., and Robert M. Price. 2005. "Inferential Methods for the Tetrachoric Correlation Coefficient." *Journal of Educational and Behavioral Statistics* 30 (2): 213–25. https://www.jstor.org/stable/3701350.

Borenstein, Michael, Larry V. Hedges, Julian P. T. Higgins, and Hannah R. Rothstein. 2010. "A Basic Introduction to Fixed-Effect and Random-Effects Models for Meta-Analysis." *Research Synthesis Methods* 1 (2): 97–111. https://doi.org/10.1002/jrsm.12.

Borsboom, Denny, and Gideon J Mellenbergh. 2002. "True Scores, Latent Variables, and Constructs: A Comment on Schmidt and Hunter."

Borsboom, Denny, Gideon J. Mellenbergh, and Jaap Van Heerden. 2004. "The Concept of Validity." *Psychological Review* 111 (4): 1061–71. https://doi.org/10.1037/0033-295X.111.4.1061.

Brannick, Michael T., Sean M. Potter, Bryan Benitez, and Scott B. Morris. 2019. "Bias and Precision of Alternate Estimators in Meta-Analysis: Benefits of Blending Schmidt-Hunter and Hedges Approaches." *Organizational Research Methods* 22 (2): 490–514. https://doi.org/10.1177/1094428117741966.

Bravais, A. 1844. *Analyse mathématique sur les probabilités des erreurs de situation d'un point*. Impr. Royale.

Brown, William. 1910. "Some Experimental Results in the Correlation of Mental Abilities1." *British Journal of Psychology, 1904-1920* 3 (3): 296–322. https://doi.org/10.1111/j.2044-8295.1910.tb00207.x.

Callender, John C., and H. G. Osburn. 1980. "Development and Test of a New Model for Validity Generalization." *Journal of Applied Psychology* 65 (5): 543–58. https://doi.org/10.1037/0021-9010.65.5.543.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Academic Press.

———. 2013. *Statistical Power Analysis for the Behavioral Sciences*. Academic Press.

Cooper, Harris M., Larry V. Hedges, and Jeff C. Valentine, eds. 2009. *The Handbook of Research Synthesis and Meta-Analysis*. 2nd ed. New York: Russell Sage Foundation.

Cox, D. R. 1989. *Analysis of Binary Data*. 2nd ed. New York: Routledge. https://doi.org/10.1201/9781315137391.

Cronbach, Lee J. 1951. "Coefficient Alpha and the Internal Structure of Tests." *Psychometrika* 16 (3): 297–334. https://doi.org/10.1007/BF02310555.

Cronbach, Lee J., and Paul E. Meehl. 1955. "Construct Validity in Psychological Tests." *Psychological Bulletin* 52 (4): 281–302. https://doi.org/10.1037/h0040957.

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.

———. 2020. "Not Restricted to Selection Research: Accounting for Indirect Range Restriction in Organizational Research." *Organizational Research Methods* 23 (4): 717–49. https://doi.org/10.1177/1094428119859398.

DerSimonian, Rebecca, and Raghu N. Kacker. 2007. "Random-Effects Model for Meta-Analysis of Clinical Trials: An Update." *NIST* 28 (January): 105–14. https://www.nist.gov/publications/random-effects-model-meta-analysis-clinical-trials-update.

Galton, Francis. 1907. "Vox Populi." *Nature* 75 (1949): 450–51. https://doi.org/10.1038/075450a0.

Gamer, Matthias, Jim Lemon, Ian Fellows, and Puspendra Singh. 2019. *Irr: Various Coefficients of Interrater Reliability and Agreement*. https://CRAN.R-project.org/package=irr.

Goulet-Pelletier, Jean-Christophe, and Denis Cousineau. 2018. "A Review of Effect Sizes and Their Confidence Intervals, Part i: The Cohen's d Family." *The Quantitative Methods for Psychology* 14 (4): 242–65. https://doi.org/10.20982/tqmp.14.4.p242.

Haddock, C. Keith, David Rindskopf, and William R. Shadish. 1998. "Using Odds Ratios as Effect Sizes for Meta-Analysis of Dichotomous Data: A Primer on Methods and Issues." *Psychological Methods* 3 (3): 339–53. https://doi.org/10.1037/1082-989X.3.3.339.

Haertel, Edward H. 2006. "3. Reliability." In, 4th ed.

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.

———. 1989. "An Unbiased Correction for Sampling Error in Validity Generalization Studies." *Journal of Applied Psychology* 74 (3): 469–77. https://doi.org/10.1037/0021-9010.74.3.469.

Hedges, Larry V., and Ingram Olkin. 1985. *Statistical Methods for Meta-Analysis*. Academic press. https://books.google.com/books?hl=en&lr=&id=7GviBQAAQBAJ&oi=fnd&pg=PP1&dq=info:e6P1zfh2T6QJ:scholar.google.com&ots=Dx-YqN6_9B&sig=-39HgbYdWPp_BwSTzA9cRODs2Q0.

Hedges, Larry V., and Jack L. Vevea. 1998. "Fixed- and Random-Effects Models in Meta-Analysis." *Psychological Methods* 3 (4): 486–504. https://doi.org/10.1037/1082-989X.3.4.486.

Holtzman, Wayne H. 1950. "The Unbiased Estimate of the Population Variance and Standard Deviation." *The American Journal of Psychology* 63 (4): 615–17. https://doi.org/10.2307/1418879.

Hunter, John E., and Frank L. Schmidt. 2015. *Methods of meta-analysis: correcting error and bias in research findings (third)*. Third. Thousand Oaks, California: Sage Publications.

Hunter, John E., Frank L. Schmidt, and Huy Le. 2006. "Implications of Direct and Indirect Range

Restriction for Meta-Analysis Methods and Findings." *Journal of Applied Psychology* 91 (3): 594–612. https://doi.org/10.1037/0021-9010.91.3.594.

Hunter, John, and Frank Schmidt. 1990. "Dichotomization of Continuous Variables: The Implications for Meta-Analysis." *Journal of Applied Psychology* 75 (June): 334–49. https://doi.org/10.1037/0021-9010.75.3.334.

Hyndman, Rob J. 2010. Cross Validated. https://stats.stackexchange.com/q/3733.

Jacobs, Perke, and Wolfgang Viechtbauer. 2017. "Estimation of the Biserial Correlation and Its Sampling Variance for Use in Meta-Analysis." *Research Synthesis Methods* 8 (2): 161–80. https://doi.org/10.1002/jrsm.1218.

Johnson, Blair T., Brian Mullen, and Eduardo Salas. 1995. "Comparison of Three Major Meta-Analytic Approaches." *Journal of Applied Psychology* 80 (1): 94–106. https://doi.org/10.1037/0021-9010.80.1.94.

Kelley, Truman Lee. 1927. *Interpretation of Educational Measurements*. World Book Company.

Kempen, G.m.p. van, and L.j. van Vliet. 2000. "Mean and Variance of Ratio Estimators Used in Fluorescence Ratio Imaging." *Cytometry* 39 (4): 300–305. https://doi.org/10.1002/(SICI)1097-0320(20000401)39:4%3C300::AID-CYTO8%3E3.0.CO;2-O.

Kirk, David B. 1973. "On the Numerical Approximation of the Bivariate Normal (Tetrachoric) Correlation Coefficient." *Psychometrika* 38 (2): 259–68. https://doi.org/10.1007/BF02291118.

Kroenke, Kurt, Robert L. Spitzer, and Janet B. W. Williams. 2003. "The Patient Health Questionnaire-2: Validity of a Two-Item Depression Screener." *Medical Care* 41 (11): 1284–92. https://www.jstor.org/stable/3768417.

Laird, Nan M., and Frederick Mosteller. 1990. "Some Statistical Methods for Combining Experimental Results." *International Journal of Technology Assessment in Health Care* 6 (1): 5–30. https://doi.org/10.1017/S0266462300008916.

Le, Huy, and Frank L. Schmidt. 2006. "Correcting for Indirect Range Restriction in Meta-Analysis: Testing a New Meta-Analytic Procedure." *Psychological Methods* 11 (4): 416–38. https://doi.org/10.1037/1082-989X.11.4.416.

Lundberg, Ian, Rebecca Johnson, and Brandon M. Stewart. 2021. "What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory." *American Sociological Review* 86 (3): 532–65. https://doi.org/10.1177/00031224211004187.

MacCallum, Robert C., Shaobo Zhang, Kristopher J. Preacher, and Derek D. Rucker. 2002. "On the Practice of Dichotomization of Quantitative Variables." *Psychological Methods* 7: 19–40. https://doi.org/10.1037/1082-989X.7.1.19.

Maxwell, Scott, and Harold Delaney. 1993. "Bivariate Median Splits and Spurious Statistical Significance." *Psychological Bulletin* 113 (January): 181–90. https://doi.org/10.1037/0033-2909.113.1.181.

McDaniel, Michael A., Deborah L. Whetzel, Frank L. Schmidt, and Steven D. Maurer. 1994. "The Validity of Employment Interviews: A Comprehensive Review and Meta-Analysis." *Journal of Applied Psychology* 79 (4): 599–616. https://doi.org/10.1037/0021-9010.79.4.599.

Mendoza, Jorge L., and Michael Mumford. 1987. "Corrections for Attenuation and Range Restriction on the Predictor." *Journal of Educational Statistics* 12 (3): 282–93. https://doi.org/10.3102/10769986012003282.

Morris, Scott, Rebecca Daisley, Megan Wheeler, and Peggy Boyer. 2014. "A Meta-Analysis of the

Relationship Between Individual Assessments and Job Performance." *The Journal of Applied Psychology* 100 (May). https://doi.org/10.1037/a0036938.

Olkin, Ingram, and John W. Pratt. 1958. "Unbiased Estimation of Certain Correlation Coefficients." *The Annals of Mathematical Statistics* 29 (1): 201–11. https://www.jstor.org/stable/2237306.

Pearson, Karl. 1895. "Notes on the History of Correlation." *Society of Biometricians and Mathematical Statisticians*. https://doi.org/10.2307/2331722.

———. 1903. "I. Mathematical Contributions to the Theory of Evolution. —XI. On the Influence of Natural Selection on the Variability and Correlation of Organs." *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 200 (321-330): 1–66. https://doi.org/10.1098/rsta.1903.0001.

Pearson, Karl, and L. N. G. Filon. 1898. "Mathematical Contributions to the Theory of Evolution. IV. On the Probable Errors of Frequency Constants and on the Influence of Random Selection on Variation and Correlation. [Abstract]." *Proceedings of the Royal Society of London* 62: 173–76. https://www.jstor.org/stable/115709.

Peters, Charles C., and Walter R. Van Voorhis. 1940. "Further Methods of Correlation." In, 362–403. New York, NY, US: McGraw-Hill Book Company. https://doi.org/10.1037/13596-013.

Roth, Bettina. 2015. "Intelligence and School Grades: A Meta-Analysis."

Sackett, Paul R., and Hyuckseung Yang. 2000. "Correction for Range Restriction: An Expanded Typology." *Journal of Applied Psychology* 85 (1): 112–18. https://doi.org/10.1037/0021-9010.85.1.112.

Schmidt, Frank, and John Hunter. 1977. "Development of a General Solution to the Problem of Validity Generalization." *Journal of Applied Psychology* 62 (October): 529–40. https://doi.org/10.1037/0021-9010.62.5.529.

Soper, H. E. 1914. "On the Probable Error of the Bi-Serial Expression for the Correlation Coefficient." *Biometrika* 10 (2/3): 384–90. https://doi.org/10.2307/2331789.

Spearman, C. 1904. "The Proof and Measurement of Association Between Two Things." *International Journal of Epidemiology* 39 (5): 1137–50. https://doi.org/10.1093/ije/dyq191.

Spearman, Charles. 1910. "Correlation Calculated from Faulty Data." *British Journal of Psychology* 3 (3): 271295. https://www.proquest.com/docview/1293688112/citation/7E133DC1091D4E47PQ/1.

Taboga, Marco. 2021. "Gamma Function." https://www.statlect.com/mathematical-tools/gamma-function.

Taylor, Erwin K., and Thomas Griess. 1976. "The Missing Middle in Validation Research." *Personnel Psychology* 29 (1): 5–11. https://doi.org/10.1111/j.1744-6570.1976.tb00397.x.

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.

———. n.d. "Fixed-Effects and Random-Effects Models in Meta-Analysis." https://wviechtb.github.io/metafor/index.html.

Viswesvaran, Chockalingam, Deniz S. Ones, Frank L. Schmidt, Huy Le, and In-Sue Oh. 2014. "Measurement Error Obfuscates Scientific Knowledge: Path to Cumulative Knowledge Requires Corrections for Unreliability and Psychometric Meta-Analyses." *Industrial and Organizational Psychology* 7 (4): 507–18. https://doi.org/10.1017/S1754942600006799.

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.

Wechsler, David. 2008. *Wechsler Adult Intelligence Scale–Fourth Edition*. 4th ed. https://doi.org/10.1037/t15169-000.

White, Thomas, Daniel Noble, Alistair Senior, W. Kyle Hamilton, and Wolfgang Viechtbauer. 2022. *Metadat: Meta-Analysis Datasets*. https://CRAN.R-project.org/package=metadat.

Whitener, Ellen M. 1990. "Confusion of Confidence Intervals and Credibility Intervals in Meta-Analysis." *Journal of Applied Psychology* 75 (3): 315–21. https://doi.org/10.1037/0021-9010.75.3.315.

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.

Wylie, Peter B. 1976. "Effects of Coarse Grouping and Skewed Marginal Distributions on the Pearson Product Moment Correlation Coefficient." *Educational and Psychological Measurement* 36 (1): 1–7. https://doi.org/10.1177/001316447603600101.