

EXERCISES WITH ANSWERS

These problems concern chapters 2-7 in Kline (2004)¹ about basic statistical concepts, confidence intervals, and effect size estimation in univariate designs.

CHAPTER 2

1. What is the difference between s and s_M ?

Answer. Both s and s_M are standard deviations, but they differ in their level of analysis. The term s is the standard deviation among individual cases measured on the same continuous variable, and it estimates σ , the population standard deviation. Assuming that the group size is not extremely small or so large that it includes a relatively high proportion of the whole population, s should not vary systematically with group size. The term s_M is a standard deviation among sample means all based on the same number of cases, and it estimates σ_M , the population standard error of the mean. The relation between these two standard deviations is $s_M = s/N^{1/2}$. Thus, the standard error of

¹Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

Please reference this work as:

Kline, R. B. (2004). *Exercises with answers*. Retrieved [date, e.g., June 20, 2004] from <http://www.apa.org/books/kline>

the mean is affected by the variability of a sample and its size.

2. Given: $M = 50.00$ and $s_M = 5.00$. Explain why this statement is probably false: "The sample mean is 5 points away from the population mean."

Answer. The above statement implies that the population mean is either $\mu = 45.00$ or $\mu = 55.00$. The standard error is like an average distance between μ and the means of all random samples of the same size. Like any average, though, it may not describe any individual case (e.g., no family has 2.2 children). Also, the statistic s_M is subject to sampling error, so its value will probably be different in another sample. Given no other information, we cannot say just exactly how far $M = 50.00$ is away from μ .

3. Given: The interval 100.00 ± 3.72 , or 96.29 to 103.72, is a 95% confidence interval for μ . Explain why this statement is probably wrong: "A total of 95% of all sample means are between 96.29 and 103.72."

Answer. Assuming normality, the interval that includes 95% of means from all random samples is

$$\mu \pm \sigma_M (z_{2\text{-tail}, .05})$$

where $\sigma_M = \sigma/N^{1/2}$ and $z_{2\text{-tail}, .05} = 1.96$. However, we do not ordinarily know the value of either μ or σ , the population standard deviation. A 95% confidence interval for μ is

constructed around a sample mean as

$$M \pm s_M [t_{2\text{-tail}, .05} (N - 1)]$$

where $s_M = s/N^{1/2}$. It would be extraordinarily unlikely that both $\mu = M = 100.00$ and $\sigma_M (1.96) = s_M [t_{2\text{-tail}, .05} (N - 1)] = 3.72$. In general, we cannot say that the lower and upper bounds of a 95% confidence interval for μ include 95% of all sample means.

4. What is a possible problem with confidence intervals in large samples?

Answer. When the sample size is large, the width of the confidence interval may be quite narrow. For example, given

$$M = 100.00, N = 10,000, \text{ and } s^2 = 81.00$$

the 95% confidence interval for μ is

$$100 \pm (81.00/10,000)^{1/2} (1.96), \text{ or } 100.00 \pm .18$$

The apparent narrowness of this interval (i.e., 99.82 to 100.18) may give the impression that it almost certainly includes μ . From a frequentist view of probability, this interval either includes μ or it does not, and we also generally do not know whether this interval is one of the 95/100 of all possible 95% confidence intervals that contain μ or one of the 5/100 that do not. Even confidence intervals that appear narrow may not include the target population parameter.

5. Given: $M_1 = 35.00$, $s_1 = 5.00$, $n_1 = 20$, $M_2 = 30.00$, $s_2 = 7.00$, and $n_2 = 25$. Construct and interpret the 95% confidence interval for $\mu_1 - \mu_2$.

Answer. Using an exact probability calculator, the positive two-tailed critical value of t (43) at the .05 level is 2.017. The pooled within-groups variance is calculated as

$$s_p^2 = [19 (5.00^2) + 25 (7.00^2)] / (20 + 25 - 2) = 38.40$$

and the standard error is calculated as:

$$s_{M_1 - M_2} = [38.40 (1/20 + 1/25)]^{1/2} = 1.86$$

The observed mean difference is $(35.00 - 30.00) = 5.00$, so the 95% confidence interval for $\mu_1 - \mu_2$ is:

$$5.00 \pm 1.86 (2.107), \text{ or } 5.00 \pm 3.92$$

which defines the interval 1.08 to 8.92. Thus, a population mean difference as small as 1.08 is just as consistent with the data as a population mean difference as large as 8.92, with 95% confidence.

6. What are the statistical assumptions of the 95% confidence interval for $\mu_1 - \mu_2$ calculated for the previous question?

Answer. The confidence interval 1.08 to 8.92 has basically the same statistical assumptions as the t test for

independent samples on which it is based. One of these is homogeneity of variance. This assumption is necessary because the pooled within-groups variance, s_p^2 , is part of the standard error for the mean difference.

7. A researcher conducts the t test for independent samples and finds that $t = 25.00$. Explain what the “25.00” means. The researcher sees that this t statistic is associated with a very small probability (e.g., $p < .00001$) and concludes that a large mean difference must be indicated. Explain why this conclusion may be wrong.

Answer. The observed mean difference is 25 times as large as the estimated standard error of that difference. Both the t test and its p value measure group size and effect size. The value of t can be large and its probability very low if the group sizes are large but the magnitude of the mean difference is small. Without more information, the researcher should not be at all impressed with these results.

8. A researcher finds in a one-way design with independent samples that $F = 2,000.00$. Explain what the “2,000.00” means. Because the F statistic seems so large, the researcher concludes that a large omnibus effect must be present. Explain why this conclusion may be wrong.

Answer. The between-groups variance is 2,000 times larger than the pooled within-groups variance. However, the former measures group size plus the squared distances between the group means and the grand mean. Thus, if the group sizes are large, the value

of the between-groups variance can be large even though the group means are quite close to the grand mean. Again, a test statistic like F does not measure effect size apart from group size.

9. Conduct the dependent samples t test and construct the 95% confidence interval for $\mu_1 - \mu_2$ assuming $r_{12} = -.735$ for the means and variances in Table 2.3. Compare your results to those in Table 2.4 for the same means and variances, but where $r_{12} = .735$.

Answer. These calculations are pertinent to the answer:

$$cov_{12} = -.735 (7.50^{1/2}) (5.00^{1/2}) = -4.50$$

$$s_D^2 = 7.50 + 5.00 - 2(-4.50) = 21.50$$

$$s_{M_D} = (21.50/5)^{1/2} = 2.07$$

$$t(4) = 2.00/2.07 = .96, p = .39$$

$$t_{2\text{-tail}, .05}(4) = 2.776$$

The 95% confidence interval is $2.00 \pm 2.07 (2.776)$, or 2.00 ± 5.75 (i.e., -3.75 to 7.75).

The negative cross-conditions correlation increases the variance of the difference scores compared to the case where r_{12} is positive. This decreases the value of t , increases its p value, and increases the width of the confidence interval even though the condition means and variances are the same.

CHAPTER 3

1. The following are statements about p values in published or online works. Explain why each is wrong and offer a correct version.

(a) “If $p < .05$, the researcher knows that there is less than 1 chance in 20 that the observed difference between the experimental and control groups occurred by chance.”²

Answer. This is the odds-against-chance fantasy. A p value measures only the probability of the result under H_0 , not the likelihood that it is due to any specific cause such as sampling error. Correct version: “If $p < .05$, the researcher knows that there is less than 1 chance in 20 that the observed difference (or one even greater) between the experimental and control groups would be found assuming H_0 is true.”

(b) “A hypothesis is rejected if the probability that it is a true statement is very low—lower than some predetermined probability ... the level of significance.”³

Answer. This is the inverse probability error, and it confuses the conditional probability of the data with the conditional probability of some hypothesis, presumably H_0 . Correct version: “A null hypothesis is rejected if the probability of the data under that hypothesis is very

low—lower than the prior probability of a Type I error, or α .”

(c) “The message here is that in judging a study’s results, there are two questions. First, is the result statistically significant? If it is, you can consider there to be a real effect. The next question is then, is the effect size large enough for the result to be useful or interesting?”⁴

Answer: The second sentence seems to say that statistical significance implies that the research hypothesis is true. This is the valid research hypothesis fallacy. Correct version: “...First, is the result statistically significant? If it is, you can consider the estimated probability of the result under the null hypothesis to be lower than the value of α specified by the researcher.” The last sentence is reasonable. Indeed, whether the effect size magnitude is large enough to meet a standard of practical or theoretical importance may be the most important question of all.

(d) “... if he is restricted to a total N of 80 cases, with this plan he must reject the null hypothesis with a 17% α risk. Formally, no problem exists; an eventual positive conclusion must carry a one-in-six risk of being spurious.”⁵

Answer. The first sentence describes a situation where

²Hunt, M. (1997). *How science takes stock*. New York: Russell Sage Foundation, p. 24.

³Roscoe, J. T. (1969). *Fundamental research statistics for the behavioral sciences*. New York: Holt, Rinehart, Winston, p. 145.

⁴Aron, A., & Aron, E. N. (2002). *Statistics for the behavioral and social sciences* (2nd ed.). Upper Saddle River, NJ: Prentice Hall, p. 147.

⁵Cohen, J. (1965). *Some statistical issues in psychological research*. In B. B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 95–121). New York: McGraw-Hill, p. 99.

the conditional prior probability of a Type I error is $\alpha = .17$, or $1/6$. The second sentence is wrong because it confuses α with the conditional posterior probability of a Type I error given that the null hypothesis has been rejected. Correct version: "...an eventual positive conclusion [H_0 is rejected] carries a risk of being spurious [a Type I error] that is generally unknown."

(e) "Significance levels show you how likely a result is due to chance. The most common level, used to mean something is good enough to be believed, is .95. This means that the finding has a 95% chance of being true."⁶

Answer. Both the odds-against-chance fantasy and the validity fallacy are expressed in this statement. It also is not clear above whether "significance levels" refers to α or p , but let's assume the latter in this version: "Exact significance levels (p) show you the probability of a result or one even more extreme assuming the null hypothesis is true. The complement of the p value, $1 - p$, is just the probability of result even more consistent with the null hypothesis than the one actually found."

(f) "Caution: Do not infer trends from data that fail by a small margin to meet the usual levels of significance. Such results are best interpreted as caused by chance and are best reported as such."⁷

⁶Creative Research Systems. (1998). *Statistical significance*. Retrieved June 20, 2002, from <http://www.surveysystem.com/signif.htm>, ¶ 5.

⁷American Psychological Association. (1974). *Publication manual of the American Psychological Association* (2nd ed.). Washington, DC: Author.

Answer. This is the odds-against-chance fallacy. Statistical tests do not identify specific "causes" of observed results, including sampling error. Please note that the above sentence appeared in the second edition of the APA's *Publication Manual*, but it was removed in the third edition.⁸

2. Use the simplified form of Bayes's theorem given in Equation 3.1 to show the relation between the conditional prior probability of a Type I error, $\alpha = p(\text{reject } H_0 \mid H_0)$, and the conditional posterior probability $p(H_0 \mid \text{reject } H_0)$. Explain why it would be difficult to actually estimate the latter.

This derivation is similar to one described by Pollard:⁹

$$p(H_0 \mid \text{reject } H_0) = \alpha [p(H_0) / p(\text{reject } H_0)]$$

where $p(H_0)$ is the priority probability of the null hypothesis and $p(\text{reject } H_0)$ is the prior probability of rejecting the null hypothesis irrespective of the data. Although Pollard notes that the latter probability may be estimated as the base rate of null hypothesis rejections in a set of studies (published or unpublished) in some area, it may be difficult to estimate $p(H_0)$ unless one is familiar with Bayesian statistics (Kline, 2004, chap. 9).

⁸American Psychological Association. (1983). *Publication manual of the American Psychological Association* (3rd ed.). Washington, DC: Author.

⁹Pollard, P. (1993). *How significant is "significance"?* In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Vol. 1. Methodological issues* (pp. 449–460). Hillsdale, NJ: Erlbaum.

CHAPTER 4

1. A researcher compares the mean number of years that patients with a serious illness survive after given one of two different treatments. The results are summarized with a d statistic. Comment.

Answer. Because outcome is measured in meaningful units, there is little need for a standardized, metric-free effect size index, such as d . The unstandardized mean difference with a confidence interval for $\mu_1 - \mu_2$ would be a more informative summary of the results.

2. What is measured by the independent samples t test for a nil hypothesis versus Hedges's g ?

Answer. The t test expresses the mean difference as the proportion of its standard error. The standard error measures the pooled within-groups variability and the group sizes. The latter explains the sensitivity of the t test to sample size. Hedges's g expresses the mean contrast as a proportion of the pooled within-groups standard deviation. The standard deviation metric of g is unaffected by group size holding all else constant, including n_1/n_2 . Hedges's g is also the part of t that measures effect size, but not also sample size.

3. Compute and interpret $\hat{\eta}$ and partial $\hat{\eta}$ for the data in Table 4.13.

Answer. We need to calculate the sums of

squares from the dependent samples ANOVA for these data:

$$M_T = (415.39 + 411.01)/2 = 413.20$$

$$SS_A = 8 (415.39 - 413.20)^2 + 8 (411.01 - 413.20)^2 = 76.74$$

$$SS_W = 7 (211.12) + 7 (146.47) = 2,503.13$$

$$MS_W = SS_W/df_W = 2,503.13/14 = 178.80$$

$$MS_{A \times S} = MS_W - M_{cov} = 178.80 - .95 (14.53) (12.10) = 11.78$$

$$SS_{A \times S} = MS_{A \times S} \times df_{A \times S} = 11.78 (7) = 82.44$$

$$SS_S = SS_W - SS_{A \times S} = 2503.13 - 82.44 = 2,420.69$$

$$SS_T = SS_A + SS_W = 76.74 + 2,503.13 = 2,579.87$$

From these results:

$$\hat{\eta} = (76.74/2,579.87)^{1/2} = .17$$

$$\text{partial } \hat{\eta} = [76.74/(76.74 + 82.44)]^{1/2} = .69$$

The simple correlation between the administration of caffeine versus placebo and performance time is .17. After removing the subjects effect, this correlation increases to .69.

4. A researcher finds $d = 0$ and concludes that there is no difference between the groups. Comment.

Answer. The result $d = 0$ says only that there is no difference in the group means. The group frequency

distributions could still differ in their shapes. The researcher should consider whether these kinds of differences are important given the research context.

5. Suppose that a treatment increases both the central tendency and variability of treated cases compared with untreated cases. Describe the expected relation between Hedges's g and Glass's Δ . What other issue should be considered?

Answer. It is expected that $\Delta > g$ in absolute value. This is because the denominator of g includes the variance of the treatment group, which is larger than that of the control group. We should consider the possibility that treatment effect is not additive, that is, it may not have the same relative impact on all cases. This suggests the presence of an interaction effect, specifically, the success of treatment may depend on some unmeasured characteristic of cases. A multifactor design where this characteristic is represented as a distinct factor is recommended.

6. Explain why this statement is incorrect: A standardized mean difference is appropriate only for experimental designs, and a measure of association is suitable only for nonexperimental designs.

Answer. There is no logical connection between the type of effect size index and whether the design is experimental or not. The d and r families of indexes are essentially rescaled versions of each other. Accordingly, a correlation effect size can be used in an experimental

design just as d can be used in a nonexperimental design as an effect size index. The choice between d or a measure of association should be based on their relative statistical strengths and weaknesses. Another consideration is the kind of effect size index reported most often in a particular research area.

7. In a dependent-samples analysis of the data in Table 4.4, a researcher mistakenly uses Equation 4.6 to compute Hedges's g from the dependent samples t statistic. Show that the result is higher than the correct value, $g = .80$.

Answer. The dependent samples t statistic for the data in Table 4.4 is 2.38. Using Equation 4.6:

$$g_{\text{wrong!}} = 2.38 (1/5 + 1/5)^{1/2} = 1.51$$

which is almost twice as large as the correct result.

CHAPTER 5

1. Show that the phi coefficient ($\hat{\phi}$) is a margin-bound measure of association for the following 2×2 contingency table:

| | Relapsed | Not relapsed |
|-----------|----------|--------------|
| Control | 60 | 40 |
| Treatment | 40 | 60 |

Answer. For this table, $\hat{\phi} = .20$. Let's multiply the frequencies in the first row by 4, which quadruples the size of the control group only and then recalculate $\hat{\phi}$:

| | Relapsed | Not relapsed |
|-----------|----------|--------------|
| Control | 240 | 160 |
| Treatment | 40 | 60 |

For this next table, $\chi^2(1) = 12.99$, so $\hat{\phi} = (12.90/500)^{1/2} = .16$. For both contingency tables, however, $p_C = .60$, $p_T = .40$, $RD = .20$, $RR = 1.50$, and $OR = 2.25$ (i.e., these statistics are not margin-bound).

2. Calculate logit d for the data in Table 5.9 for the outcome variable injury-related accident. Also calculate an approximate 99% confidence interval based on OR for the same variable.

Answer. For this outcome variable, $OR = .58$, so

$$\text{logit } d = \ln(.58)/1.8138 = -.5447/1.8138 = -.30$$

which says that the difference in the odds for being involved in an injury-related accident are about one third of a standard deviation higher in logistic units for young adult drivers rated as exhibiting higher levels of attention problems as teenagers compared with other young adult drivers. Confidence interval computations are summarized below:

$$z_{2\text{-tail}, .01} = 2.58$$

$$\ln(.58) = -.5447$$

$$\begin{aligned} SE_{\ln(OR)} &= [1/(475 \times .036 \times .964) + 1/(466 \times .064 \times .936)]^{1/2} \\ &= (.0607 + .0358)^{1/2} = .3106 \end{aligned}$$

$$\begin{aligned} 99\% \text{ confidence interval: } &-.5447 \pm .3106 (2.58), \text{ or} \\ &-.5447 \pm .8013 \end{aligned}$$

which is -1.3460 to $.2566$ in natural log units. The antilogs of the lower and upper bounds are as follows:

$$\ln^{-1}(-1.3460) = e^{-1.3460} = .2603$$

$$\ln^{-1}(.2566) = e^{.2566} = 1.2925$$

The approximate 99% confidence interval based on $OR = .58$ in this example is .26 to 1.29. Thus, the population odds ratio may be as low as .26 or as high as 1.29, with 99% confidence. Note that this interval includes 1.0 as a reasonable estimate of the population odds ratio.

3. Construct the approximate 95% confidence interval for $\pi_1 - \pi_2$ based on $RD = .16$ for the data in Table 5.7.

Answer. The estimated standard error of RD for these data is:

$$\begin{aligned} SE_{RD} &= [(.57 \times .43)/137 + (.41 \times .59)/103]^{1/2} \\ &= (.0018 + .0023)^{1/2} = .0640 \end{aligned}$$

Given $z_{2\text{-tail}, .05} = 1.96$, the approximate 95% confidence interval is:

$$.16 \pm .0640 (1.96), \text{ or } .16 \pm .13$$

which defines the interval .03 to .29. Thus, the observed risk difference of .16 in rates of coronary heart disease for smokers versus nonsmokers is just as consistent with a population risk difference as low as $\pi_1 - \pi_2 = .03$ as it is with a population risk difference as large as $\pi_1 - \pi_2 = .29$, with 95% confidence.

4. A screening test for a disorder is 90% sensitive but only 50% specific. Explain what these results mean.

Answer. The test will correctly identify 90% of cases with the disorder but only 50% of those without the disorder.

5. For the screening test described in the previous problem, estimate predictive value for a base rate of 10%.

Answer. For this problem, we can arbitrarily assume a population size of 10,000 cases. (Selection of any other number will have no effect on calculations of predictive value.) Because the base rate is 10%, we can assume that 1,000 cases have the disorder and 9,000 cases do not. Of the 1,000 cases with the disorder, the test will correctly identify 90%, or 900 of them. Similarly, of the 9,000 cases without the disorder, a total of 4,500 (50%) will obtain nonclinical results. These observations are

summarized as follows:

| | Disorder | No disorder | |
|---|----------|-------------|-------|
| + | 900 | 4,500 | 5,400 |
| - | 100 | 4,500 | 4,600 |
| | 1,000 | 9,000 | |

Of all 5,400 positive test results, a total of 900 are correct, so $+PV = .17$. That is, just less than 20% of positive results are accurate. Of the 4,600 negative results, a total of 4,500 are correct, so $-PV = .98$. Because the large majority of the positive test results are false positives, this screening test is of limited value in detecting the disorder in the general population even though its sensitivity is 90%. It is the combination of the relatively low base rate (10%) and specificity (50%) that account for the low overall accuracy of positive test results.

6. For the data in Table 5.10, estimate the predictive value of $PO - SCAD \geq 12$ for a 10% base rate of learning-attention problems.

Answer. The sensitivity of $PO - SCAD \geq 12$ is .37 and the specificity is .91. We expect to find the following cross-tabulation in a hypothetical population of 1,000 children:

| | Learning-attention problem | No problem | |
|-----------|-------------------------------|------------|-----|
| ≥ 12 | 37 | 81 | 118 |
| < 12 | 63 | 819 | 882 |
| | 100 | 900 | |

Of the 118 test results where $PO - SCAD \geq 12$, a total of 37 are from children with learning-attention problems, so $+PV = .31$. Of the 882 results where the difference score is less than 12, a total of 819 are from children without a learning-attention problem, so $-PV = .93$. Compare the above results with those in Table 5.11: Positive predictive value is only slightly higher for a cutting score of 12 ($+PV = .31$) compared with that for a cutting score of 9 ($+PV = .27$) given a realistic base rate of learning-attention problems (10%).

CHAPTER 6

1. Specify a standard set of contrast weights for each of the following comparisons in a one-way design with five samples:
 - a. M_1 versus the average of the other means
 - b. the average of M_1 and M_2 versus the average of the other means
 - c. the average of the first two means versus the last two means

Answer. Coefficients for each comparison are listed as follows, but note that the signs are arbitrary:

| Comparison | M_1 | M_2 | M_3 | M_4 | M_5 | $\sum c_i $ |
|----------------|---------------|----------------|----------------|----------------|----------------|--------------|
| $\hat{\Psi}_a$ | 1 | $-\frac{1}{4}$ | $-\frac{1}{4}$ | $-\frac{1}{4}$ | $-\frac{1}{4}$ | 2.0 |
| $\hat{\Psi}_b$ | $\frac{1}{2}$ | $\frac{1}{2}$ | $-\frac{1}{3}$ | $-\frac{1}{3}$ | $-\frac{1}{3}$ | 2.0 |
| $\hat{\Psi}_c$ | $\frac{1}{2}$ | $\frac{1}{2}$ | 0 | $-\frac{1}{2}$ | $-\frac{1}{2}$ | 2.0 |

2. Specify for a factor with four levels a set of Helmert contrasts that compares each level against the average of the remaining levels except the last. Select the weights for each comparison so that they are a standard set. Show that the whole set of contrasts is orthogonal in a balanced design.

Answer. The three possible Helmert contrasts are specified below:

| Comparison | M_1 | M_2 | M_3 | M_4 | $\sum c_i $ |
|-------------------|-------|----------------|----------------|----------------|--------------|
| $\hat{\Psi}_{H1}$ | 1 | $-\frac{1}{3}$ | $-\frac{1}{3}$ | $-\frac{1}{3}$ | 2.0 |
| $\hat{\Psi}_{H2}$ | 0 | 1 | $-\frac{1}{2}$ | $-\frac{1}{2}$ | 2.0 |
| $\hat{\Psi}_{H3}$ | 0 | 0 | 1 | -1 | 2.0 |

3. What is the relation between Hedges's g and $g_{\hat{\psi}}$?

Answer. The denominators (standardizers) of both

statistics are square roots of a pooled within-groups variance computed as $s^2 = SS_w/df_w = MS_w$. When there are only two groups:

$$g = (M_1 - M_2)/s_p = g_{\hat{\psi}} = (M_1 - M_2)/(MS_w)^{1/2}$$

4. When there are only two independent samples, what is the relation between $\hat{\eta}_A$, $\hat{\eta}_{\hat{\psi}}$, partial $\hat{\eta}_{\hat{\psi}}$, and r_{pb} ?

Answer. $\hat{\eta}_A = \hat{\eta}_{\hat{\psi}} = \text{partial } \hat{\eta}_{\hat{\psi}} = |r_{pb}|.$

5. For the data in Table 6.13, show that $\hat{\psi} = 1.54$ and $g_{\hat{\psi}} = .50$ for the contrast between students with satisfactory versus unsatisfactory outcomes on the math skills test.

Answer. Note these calculations:

$$\hat{\psi} = (12.02 + 10.95 + 10.18)/3 - (9.58 + 9.89 + 9.05)/3 = 1.54$$

$$g_{\hat{\psi}} = 1.54/9.34^{1/2} = 27.89^{1/2} [(1/3)^2 (1/129 + 1/211 + 1/171 + 1/78 + 1/38 + 1/40)]^{1/2} = .50$$

6. In a one-way design with more than two groups, a researcher finds partial $\hat{\eta}_{\hat{\psi}} = .50$ and claims that 25% of the variance is explained by the contrast. Comment.

Answer. This statement is potentially misleading. The effect represented by $\hat{\psi}$ explains 25% of the variance

after all other noncontrast effects have been subtracted out of the total variance. The proportion of total explained variance is given by $\hat{\eta}_{\hat{\psi}}^2$, and this proportion could be much lower than .25.

7. A researcher finds $\hat{\eta}_{\hat{\psi}} = .10$ and concludes that the contrast is unimportant because it explains only 1% of the total variance. Comment.

Answer. Evaluating the importance of a contrast—or any effect—based solely on a variance-accounted-for effect size is potentially misleading. Doing so ignores the research context and implications of this effect at the case level.

8. Is it wrong to compute $\hat{\eta}^2 = SS_A/SS_T$ in a design with a random factor?

Answer. No. However, the squared correlation $\hat{\eta}^2$ is a descriptive statistic that applies only in the current sample. That is, it is not intended for inferences beyond the particular levels of the independent variable included in the study. It also does not take account of the different distributional theories that underlie a fixed-effects model versus a random-effects models. The intraclass correlation $\hat{\rho}_1$ is generally a better measure of association for a random-effects model.

CHAPTER 7

1. Specify the weights for an interaction contrast in a 3×3

design that compares A_1 with A_2 and A_3 combined across just the first two levels of B . Use mean difference scaling.

Answer. A standard set of weights for the comparison on factor A of the first level with the other two is $(1, -\frac{1}{2}, -\frac{1}{2})$, and a standard set for the comparison of just the first two levels of factor B is $(1, -1, 0)$. The cross-products of these two sets of weights specify the interaction contrast:

| | B_1 | B_2 | B_3 |
|-------|----------------|---------------|-------|
| A_1 | 1 | -1 | 0 |
| A_2 | $-\frac{1}{2}$ | $\frac{1}{2}$ | 0 |
| A_3 | $-\frac{1}{2}$ | $\frac{1}{2}$ | 0 |

Note that these weights are doubly centered and that the sum of their absolute values is 4.0.

2. Show for the data in Table 7.3 for a completely between-subjects 2×3 design that the sums of squares for the omnibus interaction can be uniquely partitioned into the sums of squares for the pairwise interaction contrast and the complex interaction contrast specified by the following sets of coefficients. Use Equation 7.10 for your work:

| | Simple | | | Complex | | |
|-------|--------|-------|-------|----------------|-------|----------------|
| | B_1 | B_2 | B_3 | B_1 | B_2 | B_3 |
| A_1 | 1 | 0 | -1 | $\frac{1}{2}$ | -1 | $\frac{1}{2}$ |
| A_2 | -1 | 0 | 1 | $-\frac{1}{2}$ | 1 | $-\frac{1}{2}$ |

Answer. From Table 7.4, we know that $SS_{AB} = 84.00$ and $df_{AB} = 2$. For the pairwise interaction contrast:

$$\hat{\psi}_{AB_1} = 9.00 - 9.00 - 5.00 + 13.00 = 8.00$$

$$SS_{\hat{\psi}_{AB_1}} = [3 (8.00^2)] / [(1^2 + 1^2) (1^2 + 0^2 + 1^2)] = 48.00$$

and for the complex interaction contrast:

$$\begin{aligned} \hat{\psi}_{AB_2} &= .5 (9.00) - 12.00 + .5 (9.00) - .5 (5.00) + 6.00 - .5 (13.00) \\ &= -6.00 \end{aligned}$$

$$SS_{\hat{\psi}_{AB_2}} = [3 (-6.00^2)] / [(1^2 + 1^2) (.5^2 + 1^2 + .5^2)] = 36.00$$

Based on these results:

$$SS_{AB} = SS_{\hat{\psi}_{AB_1}} + SS_{\hat{\psi}_{AB_2}} = 48.00 + 36.00 = 84.00$$

3. Cell means for a $2 \times 2 \times 2$ design are presented as follows. Show that Equation 7.14 holds for these data:

| | C ₁ | | C ₂ | |
|----------------|----------------|----------------|----------------|----------------|
| | B ₁ | B ₂ | B ₁ | B ₂ |
| A ₁ | 15.0 | 14.0 | 17.0 | 10.0 |
| A ₂ | 10.0 | 8.0 | 18.0 | 10.0 |

Answer. A set of weights that specify a three-way interaction contrast is as follows:

| | C ₁ | | C ₂ | |
|----------------|----------------|----------------|----------------|----------------|
| | B ₁ | B ₂ | B ₁ | B ₂ |
| A ₁ | 1 | -1 | -1 | 1 |
| A ₂ | -1 | 1 | 1 | -1 |

Applying these weights to the cell means gives us

$$\hat{\Psi}_{ABC} = 15.0 - 14.0 - 10.0 + 8.0 - 17.0 + 10.0 + 18.0 - 10.0 = 0$$

which says that there is no three-way interaction. The differences between all pairs of simple interactions are as follows:

$$\begin{aligned} \hat{\Psi}_{AB \text{ at } C_1} - \hat{\Psi}_{AB \text{ at } C_2} &= [(15.0 - 14.0) - (10.0 - 8.0)] \\ &\quad - [(17.0 - 10.0) - (18.0 - 10.0)] \\ &= -1 - (-1) = 0 \end{aligned}$$

$$\begin{aligned} \hat{\Psi}_{AC \text{ at } B_1} - \hat{\Psi}_{AC \text{ at } B_2} &= [(15.0 - 17.0) - (10.0 - 18.0)] \\ &\quad - [(14.0 - 10.0) - (8.0 - 10.0)] \\ &= 6 - 6 = 0 \end{aligned}$$

$$\begin{aligned} \hat{\Psi}_{BC \text{ at } A_1} - \hat{\Psi}_{BC \text{ at } A_2} &= [(15.0 - 17.0) - (14.0 - 10.0)] \\ &\quad - [(10.0 - 18.0) - (8.0 - 10.0)] \\ &= -6 - (-6) = 0 \end{aligned}$$

Because there are no differences between any pair of simple interactions, there is no three-way interaction. Thus, Equation 7.14 holds for these data.

4. Compute $MS_{W, A, AB}$ for the data in Table 7.3 for a completely between-subjects design using the reduced cross-classification method with (a) the raw data and (b) the descriptive statistics. Show that both results equal 11.07 as computed with the orthogonal sums of squares method in Kline (2004, chap. 7) for the same data.

Answer. (a) Collapsing the data matrix across the levels of factor A generates a one-way design where B is the sole factor and $n = 6$ in each of the cells. The raw data layout is:

| B_1 | B_2 | B_3 |
|-------|-------|-------|
| 8 | 10 | 9 |
| 7 | 11 | 7 |
| 12 | 15 | 11 |
| 3 | 5 | 10 |
| 5 | 5 | 14 |
| 7 | 8 | 15 |

From the one-way ANOVA for these data, $SS_B = 48.00$, $MS_B = 24.00$, $SS_W = 166.00$, and $MS_W = 11.07$. The latter equals $MS_{W,A,AB}$ for the original two-way design calculated with the orthogonal sum of squares method.

(b) From the descriptive statistics for the original two-way cross-classification in Table 7.3, we apply Equation 7.18 to calculate:

$$\begin{aligned}
 MS_{W,A,AB} &= [2(7.00) + 2(4.00) + 3(9.00 - 7.00)^2 + 3(5.00 - 7.00)^2 + \\
 &\quad 2(7.00) + 2(3.00) + 3(12.00 - 9.00)^2 + 3(6.00 - 9.00)^2 + \\
 &\quad 2(4.00) + 2(7.00) + 3(9.00 - 11.00)^2 + \\
 &\quad 3(13.00 - 11.00)^2] / (18 - 3) \\
 &= 166.00 / 15 = 11.07
 \end{aligned}$$

5. Show for the data in Table 7.10 that the difference between the standardized simple effects of aftercare program (P) does not equal the difference between the standardized simple effects of gender (G), assuming that both factors vary naturally in the research population.

Answer. The standardizer for single-factor effects of gender is the square root of $MS_{W,P,GP} = 466.31$. The difference between the standardized simple effects of gender is:

$$\begin{aligned}
 d_{\hat{\psi}_G \text{ at TSF}} - d_{\hat{\psi}_G \text{ at SRP}} &= [(10.54 - 17.90) - (27.91 - \\
 &\quad 16.95)] / 466.31^{1/2} = -18.32 / 21.59 = -.85
 \end{aligned}$$

The standardizer for single-factor comparisons of aftercare program is the square root of $MS_{W,G,GP} = 415.27$. Values of the standardized simple effects of program are reported in Table 7.10, and the difference between them is:

$$d_{\hat{\psi}_P \text{ at Women}} - d_{\hat{\psi}_P \text{ at Men}} = -.85 - .05 = -.90$$

6. Show how the results in Table 7.9 were computed for the data in Table 7.3 for a completely between-subjects design with two fixed factors. Use the ANOVA source table in the top part of Table 7.4.

Answer. Note these calculations:

$$\begin{aligned}
 \hat{\eta}_{A,B,AB}^2 &= (18.00 + 48.00 + 84.00) / 214.00 = .70 \\
 \text{partial } \hat{\eta}_A^2 &= 18.00 / (18.00 + 64.00) = .22 \\
 \text{partial } \hat{\eta}_B^2 &= 48.00 / (48.00 + 64.00) = .43 \\
 \text{partial } \hat{\eta}_{AB}^2 &= 84.00 / (84.00 + 64.00) = .57
 \end{aligned}$$

$$\hat{\omega}_{A,B,AB}^2 = 5[(18.00 + 48.00 + 84.00)/5 - 5.55]/(214.00 + 5.33) = .56$$

$$\text{partial } \hat{\omega}_A^2 = (3.38 - 1)/(3.38 - 1 + 18) = .12$$

$$\text{partial } \hat{\omega}_B^2 = 2(4.50 - 1)/[2(4.50 - 1) + 18] = .28$$

$$\text{partial } \hat{\omega}_{AB}^2 = 2(7.88 - 1)/[2(7.88 - 1) + 18] = .43$$

7. A source table for a completely between-subjects design with two dichotomous factors where $n = 5$ is as follows:

| Source | SS | df | MS |
|--------|--------|----|--------|
| A | 211.25 | 1 | 211.65 |
| B | 661.25 | 1 | 661.25 |
| AB | 61.25 | 1 | 61.25 |
| Within | 40.00 | 16 | 2.50 |
| Total | 973.75 | 19 | |

Compute estimates of ω^2 or ρ_I for the main and interaction effects assuming (a) both factors are fixed, (b) both factors are random, and (c) factor A is random and factor B is fixed.

Answer. Values of variance component estimators for each case are calculated as follows:

| Estimator | (a) Both factors fixed | (b) Both factors random | (c) A random, B fixed |
|--------------------------------|------------------------------|-------------------------------|-------------------------------|
| $\hat{\sigma}_\alpha^2$ | $(211.25 - 2.50)/20 = 10.44$ | $(211.25 - 61.25)/10 = 15.00$ | $(211.25 - 2.50)/10 = 20.88$ |
| $\hat{\sigma}_\beta^2$ | $(661.25 - 2.50)/20 = 32.94$ | $(661.25 - 61.25)/10 = 60.00$ | $(661.25 - 61.25)/20 = 30.00$ |
| $\hat{\sigma}_{\alpha\beta}^2$ | $(61.25 - 2.50)/20 = 2.94$ | $(61.25 - 2.50)/5 = 11.75$ | $(61.25 - 2.50)/5 = 11.75$ |
| $\hat{\sigma}_\varepsilon^2$ | 2.50 | 2.50 | 2.50 |
| $\hat{\sigma}_{\text{tot}}^2$ | 48.82 | 89.25 | 65.13 |

These results above generate the following estimates of ω^2 or ρ_I for each case:

(a) Both factors fixed:

$$\hat{\omega}_A^2 = 10.44/48.82 = (211.25 - 2.50)/(973.75 + 2.50) = .21$$

$$\hat{\omega}_B^2 = 32.94/48.82 = (661.25 - 2.50)/(973.75 + 2.50) = .67$$

$$\hat{\omega}_{AB}^2 = 2.94/48.21 = (61.25 - 2.50)/(973.75 + 2.50) = .06$$

(b) Both factors random:

$$\begin{aligned}\hat{\rho}_{I_A} &= 15.00/89.25 \\ &= 2(211.25 - 61.25)/(973.75 + 211.25 + 661.25 - 61.25) \\ &= .17\end{aligned}$$

$$\begin{aligned}\hat{\rho}_{I_B} &= 60.00/89.25 \\ &= 2(661.25 - 61.25)/(973.75 + 211.25 + 661.25 - 61.25) \\ &= .67\end{aligned}$$

$$\begin{aligned}\hat{\rho}_{I_{AB}} &= 11.75/89.25 \\ &= 4(61.25 - 2.50)/(973.75 + 211.25 + 661.25 - 61.25) \\ &= .13\end{aligned}$$

(c) A random, B fixed:

$$\hat{\rho}_{I_A} = 20.88/65.13 = .32$$

$$\hat{\omega}_B^2 = 30.00/65.13 = .46$$

$$\hat{\rho}_{I_{AB}} = 11.75/65.13 = .18$$

8. Compute and interpret $\hat{\eta}^2$ for the total effects and partial $\hat{\eta}^2$ for the interactive and main effects of gender and aftercare program for the data in Table 7.10.

Answer. For the total effects:

$$\hat{\eta}_{G,P,GP}^2 = 3,183.33/53,548.55 = .0594$$

which says that the main and interactive effects together explain about 6% of the total variance, and

their correlation with outcome is $.0594^{1/2} = .24$.

For the individual effects:

$$\text{partial } \hat{\eta}_G^2 = 24.37/(24.37 + 50,367.22) = .00048$$

$$\text{partial } \hat{\eta}_P = .0220$$

$$\text{partial } \hat{\eta}_P^2 = 804.28/(804.28 + 50,367.22) = .0157$$

$$\text{partial } \hat{\eta}_P = .1254$$

$$\text{partial } \hat{\eta}_{GP}^2 = 2,352.69/(2,352.69 + 50,367.22) = .0446$$

$$\text{partial } \hat{\eta}_{GP} = .2112$$

The magnitude of the gender main effect is quite small: It explains less than one-half of one-tenth of a percent of the variance controlling for other effects, and the partial correlation between gender and outcome is only about .02. The relative size of the aftercare program main effect is larger: It explains about 2% of the residual variance, and its partial correlation with outcome is about .13. The relative magnitude of the interaction effect is the largest of all: It accounts for about 5% of the residual variance, and its partial correlation with outcome is about .22.

9. For the data in Table 7.11, compute d for the comparison of students classified as heavy drinkers versus all other students separately for subsamples of young men and young women.

Answer. Values of the unstandardized contrasts are calculated as follows:

$$\hat{\Psi}_{\text{Heavy vs. all others}}$$

| | |
|-------------|--|
| Young women | $(64.91 + 65.38 + 61.16)/3 - 54.47$ $= 9.35$ |
| Young men | $(60.56 + 59.98 + 58.38)/3 - 45.59$ $= 19.88$ |

Because gender (G) varies naturally in the population, the appropriate standardizer for comparisons on the drinking level (D) factor is the square root of $MS_{W, G, GD}$. The sums of squares in Table 7.11 are not orthogonal, so we must use Equation 7.18 to calculate this adjusted mean square from the descriptive statistics:

$$\begin{aligned} MS_{W, G, GD} &= [45 (8.01^2) + 38 (8.40^2) + 46 (64.91 - 62.92)^2 \\ &\quad + 39 (60.56 - 62.92)^2 + \\ &\quad 51 (6.84^2) + 44 (9.95^2) + 52 (65.38 - 62.88)^2 \\ &\quad + 45 (59.98 - 62.88)^2 + \\ &\quad 99 (7.50^2) + 83 (9.36^2) + 100 (61.16 - 59.88)^2 \\ &\quad + 84 (58.38 - 59.88)^2 + \\ &\quad 54 (9.03^2) + 77 (11.90)^2 + 55 (54.47 - 49.26)^2 \\ &\quad + 78 (45.59 - 49.26)^2] / (499 - 4) \\ &= 44457.35 / 495 = 89.81 \end{aligned}$$

The standardized contrasts for each gender are:

$$d_{\hat{\Psi}_{\text{Heavy vs. all others}}}$$

| | |
|-------------|------------------------------|
| Young women | $9.35 / 89.81^{1/2} = .99$ |
| Young men | $19.88 / 89.81^{1/2} = 2.10$ |

In words, the relative social skills reasoning deficit of young men who are heavy drinkers compared to their same-gender peers who do not drink as much is about two standard deviations. The magnitude of this relative deficit is just over twice as large as it is among young women.