Notes from Kline, Rex. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association. 325 pp. {First edition. There is a second edition out now.}

Bookmark

Chapter 4. Parametric Effect Size Indexes.

Chapter 1. Changing Times.

Webpage at http://www.apa.org/books/resources/kline/ .

Presents a short history of NHST.

Chapter 2. Fundamental Concepts.

Introduces the basics of confidence interval estimation.

Notes that many statistics have more complex sampling distributions than do means, for example, δ, the standardized difference between means. Briefly discusses ways to construct confidence intervals about such statistics. Mentions **confidence interval transformation** (Steiger & Fouladi, 1997) – the statistic is transformed into normally distributed units, lower/upper limits of the interval are obtained by subtracting/adding $Z_{cc}SE$ to the transformed statistic, resulting in a confidence interval for the transformed statistic. The limits are then transformed back to the original metric. Construction of a CI about Pearson *r* is given as a well-known example of this approach.

A second approach briefly mentioned involves placing the CI limits -/+ the product of the two-tailed critical value of a central test statistic and an estimate of the **asymptotic standard error**. This generally requires large sample sizes and for some statistics (such as multiple *R*) requires a computer to conduct the calculations.

The third approach mentioned is **noncentrality interval estimation** (Steiger & Fouladi, 1997). It is noted that noncentral distributions include an additional parameter, the noncentrality parameter. A noncentral *t*, for example, has *df* and the noncentrality parameter. If the value of the noncentrality parameter is 0, we have the central *t*. As the |noncentrality parameter| becomes increasingly large, the noncentral *t* becomes increasingly skewed in one or the other direction. It is noted that one needs specialized software to compute noncentral confidence intervals.

The fourth approach very briefly mentioned is bootstrapping/resampling.

The rest of this chapter is a quick review of the basic of several common NHST.

Chapter 3. What's Wrong With Statistical Tests --- Where Do We Go From Here?

Starts out with a listing of common fallacies about NHST.

Misinterpretation of *p*.

Fallacy 1. *p* is the probability that the results are due to sampling error. Duh, unless you have the entire population, the probability that the results include sampling error is 1.

Fallacy 2: *p* is the probability of the null being true given the data – the "inverse probability error (Cohen, 1994) or the "Bayesian Id's wishful thinking error" (Gigerenzer, 1993). Duh, it is the probability of the data given the null. Bayes is briefly noted:

$p(H_0 \mid D) = \dfrac{p(H_0 \cap D)}{p(D)}$, that is,

$$\text{posterior probability of null given data} = \frac{(\text{prior prob of null})(\text{likelihood of data given null})}{\text{prior probability of the data}}.$$

Fallacy 3: *p* is the probability that the null is true given that we have rejected it.

Fallacy 4: 1-*p* is the probability that the alternative hypothesis is true given the data.

Fallacy 5: 1-*p* is the probability that a replication attempt will also reject the null.

Mistaken Conclusions.

Fallacy 1. The smaller *p*, the larger the effect.

Fallacy 2. "Rejection of the null hypothesis confirms the alternative hypothesis and the research hypothesis behind it." Kline argues that there are two conceptual errors here. First, the posterior probability of the statistical alternative hypothesis is not brought to absolutely one following rejection of the null hypothesis. Clearly Kline is using "confirm" to mean "establish veracity beyond all doubt," rather than to mean "increase belief in the truth of." Second, the truth of the statistical alternative hypothesis does not require the truth of the substantive research hypothesis. Kline uses research by John Arbuthnot to illustrate this error. Arbuthnot noted that in 82 consecutive years more boys than girls were born in London. This led him to reject the null hypothesis that 50% of births will be boys. Our accepting of the statistical alternative hypothesis does not, however, automatically lead to our accepting of his substantive research hypothesis, which was that God arranged to have more boys than girls born so that every woman will have a man, even though wars and the like lead to boys and men leaving this life earlier than do girls and women.

Of course, we all know that the truth is that more boys are born because Y bearing sperm swim more quickly than do X bearing sperm (but the X bearing sperm live longer in the acidic envrionment in which they are thrust). God has blessed us with wars and adolescent male automobile drivers to get rid of the troublesome extra men. ;-)

Fallacy 3. If you have not rejected the null, then it must be true.

Fallacy 4. Same as 3 in different words.

Fallacy 5: Rejecting the null means your results are of value.

Fallacy 6. Failing to reject the null means that your research is of no value.

Fallacy 7. Rejecting the null means that you have identified a causal mechanism producing the observed correlation.

Fallacy 8. If A rejects the null today, and B tests the same null tomorrow but does not reject it, the B's results cast doubt on A's conclusion.

At this point I found myself thinking "there can't be many people who subscribe to these fallacies, can there?" Kline goes on to provide evidence that there are.

Kline next argues that null hypotheses are almost always false, that one need not worry about making Type I errors, but that one should worry about making Type II errors.

The next several pages are spent on detailing how NHST have impeded scientific progress in those disciplines where it flourishes.

Kline then briefly presents some variations on NHST, including the testing of range null hypotheses (see Serlin & Zumbo, 2001), equivalence testing, inferential confidence intervals (Tyron, 2001), and three-valued logic (Kaiser, 1960 -- see Serlin & Zumbo, 2001). It was Kaiser who coined the term "Type III error" to mean "deciding upon a difference in the wrong direction." Tyron's inferential confidence intervals are constructed such that one can compare different groups' confidence intervals and conclude that the group means differ significantly if the confidence intervals overlap. The intervals are adjusted so that the they will always lead to the same conclusion about the value of μ_1-μ_2 that would be made using a traditional *t* test.

Kline wraps up this chapter with some suggestions regarding what changes we should make in the way we analyze our research data. One of those is dropping the word "significant" from our research discourse.

Chapter 4. Parametric Effect Size Indexes.

Cohen's *d* and Glass' Δ are introduced as estimates of the standardized difference between two population means. When variances are heterogeneous, it is recommended that one compute and report two values of Δ, one using, in the denominator, the standard deviation of the one group, and the other using the standard deviation of the other group.

With correlated samples, one can compute *d* or Δ just as one would with independent samples, or one can compute the ratio of the mean difference score divided by the standard deviation of the difference scores. The statistic on the difference scores does not seem very useful to me, as it factors in the variance-reducing effect of the correlation between conditions.

Approximate confidence intervals can be computed by $d \pm Z_{cc}*SE$ where *SE* is $\sqrt{\frac{d^2}{2df} + \frac{N}{n_1 n_2}}$.

For Δ, use $\sqrt{\frac{\Delta^2}{2(n_2 - 1)} + \frac{N}{n_1 n_2}}$, where n_1 is for the group whose standard deviation was used to compute Δ. For correlated samples, use $\sqrt{\frac{d^2}{2(n-1)} + \frac{2(1-r_{12})}{n}}$, where *d* is computed as for independent samples. Of course, this interval is narrowed by the correlation between conditions.

Wuensch's example of an approximate confidence interval. Suppose M_1 = 95, M_2 = 105, SD_1 = 20, SD_2 = 20, n_1 = 200, and n_2 = 200. Cohen's *d* = 10/20 = .5, $SE_{M1-M2} = \sqrt{\frac{20^2 + 20^2}{200}} = 2$, *t* = 10/2 = 5, $p < .001$. The unstandardized $CI_{.95}$ for $(\mu_1 - \mu_2) = 10 \pm 1.96(2) = 6.08, 13.92$. An approximate $CI_{.95}$ for *d* is $.5 \pm 1.96\sqrt{\frac{.5^2}{2(398)} + \frac{400}{200(200)}} = .5 \pm .20$ = 0.3, .7. Notice that this is nothing more than the unstandardized confidence interval after dividing each end of the interval by the pooled standard deviation, that is, 6.08/20, 13.92/20.

Now reduce the sample sizes to 10. $SE_{M1-M2} = \sqrt{\frac{20^2 + 20^2}{10}} = 8.944$, *t* = 10/8.944 = 1.118. The unstandardized $CI_{.95}$ for $(\mu_1 - \mu_2) = 10 \pm 2.101(8.944) = -8.79, 28.79$. Divide the ends by 20 and the approximate $CI_{.95}$ for *d* is −.44, 1.44.

Exact confidence intervals can be computed using noncentral *t*. One first obtains a *CI* for the noncentrality parameter, *ncp*. Each endpoint of the *ncp* is then multiplied by $\sqrt{\frac{n_1 + n_2}{n_1 n_2}}$ to get the *CI* for ***d***. Wuensch's SAS program Conf_Interval-d2.sas will do this.

Wuensch's example of exact confidence interval. Using my SAS program and the same statistics used for the approximate confidence interval above, we get the same answer, .3, .7. The approximation is very good with large sample sizes. Reduce the sample sizes to 10 and the exact confidence interval computes to −.40, 1.39. Clearly the approximation is off a bit with small sample sizes, but, at least for my example, not by that much.

I found, at http://psychology3.anu.edu.au/people/smithson/details/CIstuff/CI.html , SAS and SPSS scripts for doing all this and more. Joy!

Wuensch's SAS program Conf_Interval-d1.sas will construct an exact confidence interval for *d* for a single sample mean or the difference between means from correlated samples. With correlated samples one need keep in mind that *g* will be made larger and its confidence interval more narrow by the correlation between scores in the two samples. If you want to use the standard deviation of one

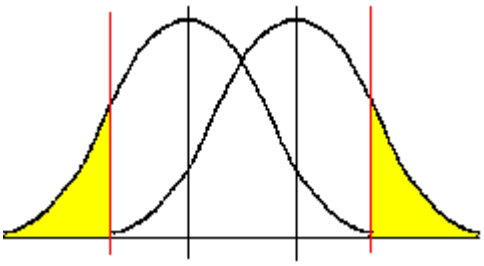
or both samples not corrected for that correlation, you need to use the approximation method – with correlated samples the distributions here are very complex, not following the noncentral *t*.

One can also estimate the magnitude of differences between/among group means with *r* or η, often squared so as to represent a proportion of variance. For within-subjects effects, one can compute a partial η^2 by dividing the effect SS by ($SS_{total} - SS_{subjects}$).

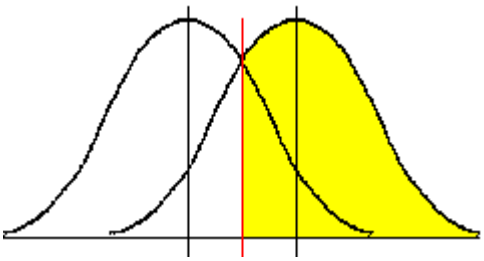
One can use Fisher's transformation of *r* to produce a confidence interval for the point biserial *r*, but this approximate method may not be very accurate, especially with uneven sample sizes. Better is to construct a confidence interval with a noncentral *F* distribution. First construct a CI for the noncentrality parameter and then convert it to η^2 units. My program Conf-Interval-R2-Regr.sas does this. This program is also appropriate for constructing a CI about R^2 from multiple regression (fixed effects). At the present time, there are no easy ways to get confidence intervals for eta-squared for correlated samples.

Jim Steiger's R2 program can be used to construct a CI about R^2 from multiple correlation (random effects).

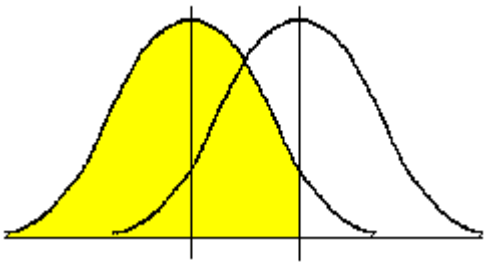
Measures of overlap. Cohen described three of these. Look at the overlapping distributions. U1 is the proportion of all scores that are in the shaded areas where the distributions do not overlap. U1 ranges from 0 to 1.



U2 is the proportion of scores in the lower group exceeded by the same proportion of scores in the upper group. U2 ranges from .5 to 1.



U3 is the proportion of scores in the lower group that are below the median of the upper group. U3 ranges from .5 to 1.



Schematic plots are mentioned as a graphical means of showing overlap between distributions.

Tail ratios tell one how much more likely a case from one group is to fall in the upper (or lower) tail of the combined distribution than is a case from the other group. An example is given comparing the verbal ability of men and women. Using a cutting point of 1 SD above the mean of the combined

groups, 18.67% of women were in the upper tail and 12.92% of men were, yielding a right tail ratio of .1867/.1292 = 1.45.

The Common Language Effect Size statistic is presented.

When group membership is being predicted (DFA or logistic regression), one can construct a statistic that describes effect size in terms of how much better classification is compared to what it would be by chance. Huberty and Lowman (2000) defined such a statistic, $I = \frac{H_o - H_e}{1 - H_e}$, where H_o is the observed hit rate (proportion of correct classifications) and H_e is the proportion that would be expected by change. $H_e = \frac{pr_1 n_1 + pr_2 n_2}{N}$, where pr_i the prior probability of being in group i, n_i is the sample size for group i, and N is the total number of cases. Of course, one could get classification better than H_e simply by classifying all cases into whichever group has the higher pr. Huberty and Lowman suggested that values of I less than .1 represent small effects and those greater than .35 represent large effects.

Page 131 has a table showing correspondence between these various measures, assuming we are comparing two normally distributed distributions with equal sample sizes and variances.

Chapter 5. Nonparametric Effect Size Indexes

Despite the title, this chapter is about contingency table analysis.

Consider a 2 x 2 table where one variable is type of therapy (some vs none) and outcome (success, failure – such as death of patient). Let p_c represent the sample probability of failure in the control group and p_t that in the treatment group. The sample risk difference, RD, is p_c-p_t. The sample risk ratio, RR, is $p_c \div p_t$. The odds are $O_c = \frac{p_c}{1 - p_c}$ and $O_t = \frac{p_t}{1 - p_t}$. The odds ratio, $OR = \frac{O_c}{O_t}$. The logit is the natural log of the odds ratio. Kline describes the logistic distribution as being approximately normal with a SD of $\frac{\pi}{\sqrt{3}}$ where π is 3.14…….. Kline then suggests the logit d as an effect estimate comparable to the Hedges g. Logit d is simply the logit divided by the SD of the logistic distribution, that is, logit d = $\frac{\ln(OR)\sqrt{3}}{\pi}$. The phi coefficient is also discussed.

Kline shows how to construct confidence intervals about proportions, RD, RR, and OR. For the logit, the SE is $\sqrt{\frac{1}{n_c p_c (1 - p_c)} + \frac{1}{n_t p_t (1 - p_t)}}$. For the Gender x Decision data from my lesson on binary logistic regression, the SE for the logit is, using this formula and with some rounding error, .2455. The OR is 3.3759, so the logit is 1.2166. Going out 1.96(.2455) from the logit gives a 95% CI of .73542 to 1.69778. Taking the antilog of each endpoint give the CI for the OR, 2.086 to 5.462. Of course, it is a hell of a lot easier to get this CI by using the binary logistic regression procedure in SPSS. It gave the CI as 2.090 to 5.452.

Kline describes the computations for a CI on phi as “quite complicated” and refers those interested to Fleiss (1994).

Kline briefly describes Cramér’s V (also known as Cramér’s phi) for contingency tables with more than two rows and/or more than two columns.

Sensitivity and specificity are described using as an example a screening test for some medical disorder. Sensitivity is, of those who have the disorder, the percentage that are correctly identified. Specificity is, of those who do not have the disorder, the percentage that are correctly identified. He notes that sensitivity and specificity are affected by the cutting point that is used. Kline defines “predictive value” as the overall percentage of cases which are correctly classified. “Positive

predictive value" is the percentage of all predicted occurrences of the disorder that are correctly identified – in other words, 100% minus the false positive rate. "Negative predictive value" is the percentage of all predicted nonoccurrences that are correctly identified – in other words, 100% less the false negative rate. Kline then demonstrates that the base rate of the disorder can greatly affect predictive values (but does not affect sensitivity and specificity). As the base rate increases, positive predictive value rises and negative predictive value falls.

Chapter 6. Effect Size Estimation in One-Way Designs

Defines a contrast in the usual way, but adds an additional stipulation for a "**standard set of weights**": the coefficients for the one set must equal +1 divided by the number of conditions in that set while those for the other set must equal -1 divided by the number of conditions in that other set. For three means, -2, 1, 1 codes a contrast between the first mean and the next two means, but are not standard. -1, ½, ½ codes the same contrast but with standard weights. $\hat{\psi} = \Sigma c_i M_i$. The use of standard weights is assumed for all that follows on standardized contrasts.

A contrast SS is computed as $SS_{\hat{\Psi}} = \frac{\hat{\psi}^2}{\sum \frac{c_j^2}{n_j}}$. When the sample sizes are equal, this simplifies to

$SS = \frac{n\hat{\psi}^2}{\sum c_j^2}$.

The contrast *SS* is divided by an appropriate error *MS* to yield the contrast *F*. Of course, the contrast *F* is the square of the contrast $t = \frac{\hat{\psi}}{s_{\hat{\psi}}}$, where $s_{\hat{\psi}} = \sqrt{MSE \sum \frac{c_j^2}{n_j}}$.

With correlated samples one may use the same error term used to test the omnibus main effect of the within-subjects factor, or one may use an error term based on only the conditions involved in the particular contrast being tested. The latter would be advised when there is a problem with the sphericity assumption. For *t*, the denominator would be the familiar $\frac{s_D}{\sqrt{n}}$, where the difference scores are computed, for each subject, from the contrast coefficients. Suppose we had four conditions and wanted to contrast the first condition with the (second and third) conditions. For each subject we would compute the mean of e's scores in the second and third conditions and use a standard correlated *t* test to compare those means with e's scores in the first condition.

To construct a confidence interval about $\hat{\psi}$, simply go out in each direction $t_{crit} s_{\hat{\psi}}$. When one is constructing multiple confidence intervals, one can use Bonferroni to adjust the per contrast alpha. Such intervals have been called simultaneous or joint confidence intervals.

A population standardized contrast, $\delta_{\psi} = \psi / \sigma$, can be estimated by $\hat{\psi} / s$, where *s* is the standard deviation of just one of the groups being compared (Glass' Δ), the pooled standard deviation of the two groups being compared (Hedges' g), or the pooled standard deviation of all of the groups (the square root of the *MSE*). Having obtained a contrast *F* from your computer program, you can take the square root to obtain the contrast *t* and then compute $g_{\hat{\psi}} = t_{\hat{\psi}} \sqrt{\sum \frac{c_j^2}{n_j}}$. The contrast *g* is computed in exactly the same way with correlated samples designs, but if computing from *t* you need to be sure to use the independent samples *t*, not the correlated samples *t*. One could standardize the contrast with the standard deviation of the difference scores, but I think that not usually appropriate.

An **approximate confidence interva**l for a contrast *g* can be computed simply by taking the confidence interval for the contrast and dividing its endpoints by the pooled standard deviation

(square root of *MSE*). In this case the confidence interval amounts to $g_{\hat{\psi}} \pm t_{crit} s_{g_{\hat{\psi}}}$, where $s_{g_{\hat{\psi}}} = \sqrt{\sum \frac{c_j^2}{n_j}}$.

When the samples are correlated, the same method is employed – that is, construct a confidence interval for the contrast as $\hat{\Psi} \pm t_{crit} s_{\hat{\Psi}}$, where the standard error is based on the difference scores, that is, $\frac{s_D}{\sqrt{n}}$. The width of this confidence interval is appropriately affected by the correlations between conditions. The endpoints of the unstandardized interval are then divided by the pooled standard deviation across conditions (computing this standardizer as if the samples were independent).

At http://www.psy.unsw.edu.au/research/research-tools/psy-statistical-program one can obtain PSY: A program for contrast analysis, by Kevin Bird, Dusan Hadzi-Pavlovic, and Andrew Isaac. This program computes unstandardized and approximate standardized confidence intervals for contrasts with between-subjects and/or within/subjects factors. It will also compute simultaneous confidence intervals. Contrast coefficients are provided as integers, and the program converts them to standard weights.

An exact confidence interval for a standardized contrast involving independent samples can be computed with my SAS program Conf_Interval-Contrast.sas.

Eta-squared is introduced as an estimator of the proportion of variance accounted for by a given effect. The intraclass correlation coefficient is briefly mentioned as the corresponding estimator when the effect is random rather than fixed. Partial eta-squared is introduced as $\frac{SS_{effect}}{SS_{effect} + SS_{error}}$. I have yet to be convinced that partial eta-squared estimates a parameter of as much interest as that estimated by eta-squared.

My program Conf-Interval-R2.sas will compute an exact confidence interval about eta-squared. This could be done for contrast eta-squared too, but I think a confidence interval about the standardized contrast probably more useful.

Effect-size estimation in ANCOV is briefly discussed. For a standardized contrast one can use in the denominator the difference between the original means or the difference between the adjusted means, and in the denominator one can use the square root of the error variance from an ANOVA (not removing the effect of the covariates) or the square root of the error variance from the ANCOV (with the effect of the covariate removed). Likewise, one could compute eta-squared using unadjusted scores (SS_{effect} divided by SS_{total} from ANOVA, ignoring covariate) or the adjusted scores (SS_{effect} from ANCOV divided by SS_{total} with effect of covariates removed). I prefer computing the increase in R^2 that accompanies adding the effect to a model that already has the covariates.

Chapter 7. Effect Size Estimation in Multifactor Designs

Kline presents a terse introduction to the various complexities of multifactor designs, including interaction contrasts. I have not often found it useful to employ interaction contrasts, but do find them of some interest.

The coefficients for an interaction contrast must be doubly centered in the sense that the coefficients must sum to zero in every row and every column of the a x b matrix. For example, consider a 2 x 2 ANOVA. The interaction has only one *df*, so there is only one contrast available.

	Coefficients			Means	
	B_1	B_2		B_1	B_2
A_1	1	-1		M_{11}	M_{12}
A_2	-1	1		M_{21}	M_{22}

This contrast is $M_{11} – M_{12} – M_{21} + M_{22}$. From one perspective, this contrast is the combined cells on one diagonal ($M_{11} + M_{22}$) versus the combined cells on the other diagonal ($M_{21} + M_{12}$). From another perspective, it is ($M_{11} - M_{12}$) – ($M_{21} – M_{22}$), that is, the simple main effect of B at A_1 versus the simple main effect of B at A_2. From another perspective it is ($M_{11} – M_{21}$) – ($M_{12} – M_{22}$), that is, the simple main effect of A at B_1 versus the simple main effect of A at B_2. I recall that Dave Howell had an exercise in his Methods text that led the student into realizing that a one *df* interaction does make this contrast. All of this is illustrated in my program Interact2x2.sas.

Now consider a 2 x 3 design. The interaction has two *df* and can be broken down into two orthogonal interaction contrasts. For example, consider the contrast coefficients in the table below:

	A x $B_{12 \text{ vs } 3}$				A x $B_{1 \text{ vs } 2}$		
	B_1	B_2	B_3		B_1	B_2	B_3
A_1	1	1	-2		1	-1	0
A_2	-1	-1	2		-1	1	0

The contrast on the left side of the table compares the simple main effect of A at combined levels 1 and 2 of B with the simple main effect of A at level 3 of B. From another perspective, it compares the simple main effect of (combined B_1 and B_2) versus B_3 at A_1 with that same effect at A_2. Put another way, it is the AxB interaction with levels 1 and 2 of B combined.

The contrast on the right side of the table compares the simple main effect of A at level 1 of B with the simple main effect of A at level 2 of B. From another perspective, it compares the simple main effect of B12 (excluding level 3 of B) at A1 with that same effect at A2. Put another way, it is the AxB interaction with level 3 of B excluded.

If we had reason to want the coefficients on the left side of the table above to be a standard set of weights, we would divide each by 2.

	A x $B_{12 \text{ vs } 3}$		
	B_1	B_2	B_3
A_1	.5	.5	-1
A_2	-.5	-.5	1

My program Interact2x3.sas illustrates the computation of these interaction contrasts and more.

Kline also briefly discusses the use coefficients to make trend contrasts. I prefer to handle such contrasts with a polynomial regression rather than with coefficients.

In introducing the topic of standardized mean differences, Kline notes that there is much disagreement regarding how to compute them with data from a multifactorial design, and opines that 1.) such estimates should be comparable to those that would be obtained from a one-way design, and 2.) changing the number of factors in the design should not necessarily change the effect size

estimates. Adding factors to a design is, IMHO, not different from adding covariates. Should the additional variance explained by added factors be excluded from the denominator of *g*? Imagine a 2 x 2 design, where A is type of therapy, B is sex of patient, and Y is post-treatment wellness. The *MSE* excludes variance due to sex, but in the population of interest sex may naturally account for some of the variance in wellness, so using the root mean square error as the standardizer will underestimate the population standard deviation. It may be desirable to pool the $SS_{within\text{-}cells}$, SS_B, and SS_{AxB} to form an appropriate standardizer in a case like this. I'd just drop B and AxB from the model, run a one-way ANOVA, and use the root mean square error from that as the standardizer.

Kline argues that when a factor like sex is naturally variable in both the population of interest and the sample then variance associated with it should be included in the denominator of *g*. While I agree with this basic idea, I am not entirely satisfied with it. Such a factor may be associated with more or less of the variance in the sample than it is in the population of interest. In experimental research it is often the case that the distribution of such a factor can be quite different in the experiment than it is in the population of interest. For example, in the experiment there may be approximately equal numbers of clients assigned to each of three therapies, but in the natural world patients may be given the one therapy much more often than the others.

Now suppose that you are looking at the simple main effects of A (therapy) at levels of B (sex). Should the standardizer be computed within-sex, in which case the standardizer for men would differ from that for women, or should the standardizer be pooled across sexes? Do you want each *g* to estimate *d* in a single-sex population, or do you want a *g* for men that can be compared with the *g* for women without having to consider the effect of the two estimators having different denominators?

Kline shows how to compute eta-squared and omega-squared in factorial designs and notes that confidence intervals can be constructed about such estimates using the same methods presented in the previous chapter. He then concludes by presenting results of factorial analyses with effect size estimates included. Added to one source table is a column which gives the effect size estimate and, in parentheses, a confidence interval.

Chapter 8. Replication and Meta-Analysis

The importance of replication is stressed, and Kline notes that editorial policies discourage such replication.

Meta-analytic methods are reviewed, not in sufficient detail to teach one how to conduct meta-analysis, but sufficiently to allow one to be an educated consumer of such analyses.

With respect to how many studies are necessary to conduct a meta-analysis, Kline says "A researcher can use meta-analytic methods to synthesize as few as two results, but more are typically needed. Although there is no absolute minimum number, it seems to me that at least 20 different studies would be required before a meta-analysis is really viable. This assumes that the studies are relatively homogeneous, and that only a small number of moderator variables are associated with study outcome."

In addition to estimating effect sizes, the modern meta-analysis attempts to explain observed variability in effect sizes. Commonly used predictors (moderators) of effect size include: Substantive factors such as subject characteristics, setting in which the data were collected, the date the data were collected (societal norms may change across time), and intensity of the treatment (for example, 10 mg dose or 50 mg dose); Methodological factors such as method of manipulation of independent variable, method of measuring the dependent variable, and the quality of the research design; and Extrinsic factors such as the gender or professional background of the author, whether the research was published or not, and who funded the research.

When computing an effect size across studies, one can weight the individual study effect sizes by factor such as sample size and quality of the study.

There is available a statistic that can be employed to test the null hypothesis that there is a single universe of studies with a single true effect size. If the variability among individual study effect sizes is sufficiently great, then this null hypothesis will be rejected. Following rejection of this hypothesis, one can segregate the studies on the basis of suspected moderators and retest within each segregated group of studies. Alternatively, one adopt a random-effects model for the meta-analysis (the sources of differences between studies is random, or cannot be identified) or a mixed-effects model (there are both random sources of differences and moderators that can be identified).

When the hypothesis that the overall effect size is zero has been rejected, one can estimate the "fail safe number" of studies with effect sizes of zero that would have to out there in file drawers (unavailable to the meta-analyst) to reduce to nonsignificance the test of the overall effect size.

Chapter 9. Resampling and Bayesian Estimation.

Resampling techniques are briefly described, and David Howell's program mentioned. Kline is not very enthusiastic about resampling techniques.

Reference

Serlin, R. C., & Zumbo, B. D. (2001). Confidence intervals for directional decisions. Retrieved from http://edtech.connect.msu.edu/searchaera2002/viewproposaltext.asp?propID=2678 on 20. February 2005.

Return to the Stat Help Page

Karl L. Wuensch
East Carolina University
Dept. of Psychology
Greenville, NC 27858