2

SAMPLING AND ESTIMATION

> In times of change, learners inherit the Earth, while the learned find themselves beautifully equipped to deal with a world that no longer exists.
>
> —Eric Hoffer (1973, p. 22)

Fundamental concepts of sampling and estimation are the subject of this chapter. You will learn that (a) sampling error affects virtually all sample statistics, (b) interval estimation approximates margins of error associated with statistics, but (c) there are other sources of error variance that should not be ignored. You will also learn about central versus noncentral test statistics, the role of bootstrapping in interval estimation, and the basics of robust estimation. Entire books are devoted to some of these topics, so it is impossible in a single chapter to describe all of them in detail. Instead, the goal is to make you aware of concepts that underlie key aspects of statistics reform.

SAMPLING AND ERROR

A basic distinction in the behavioral sciences is that between populations and samples. It is rare that entire populations are studied. If a population is large, vast resources may be needed. For instance, the budget for

DOI: 10.1037/14136-002
Beyond Significance Testing: Statistics Reform in the Behavioral Sciences, Second Edition, by R. B. Kline

the 2010 Census in the United States was $13 billion, and about 635,000 temporary workers were hired for it (U.S. Census Bureau, 2010). It may be practically impossible to study even much smaller populations. The base rate of schizophrenia, for example, is about 1%. But if persons with schizophrenia are dispersed over a large geographic area, studying all of them is probably impracticable.

Types of Samples

Behavioral scientists usually study samples, of which there are four basic kinds: random, systematic, ad hoc, and purposive. **Random (probability) samples** are selected by a chance-based method that gives all observations an equal likelihood of appearing in the sample. Variations on simple random sampling include stratified sampling and cluster sampling. In both, the population is divided into smaller groups that are mutually exclusive and collectively exhaustive. In **stratified sampling**, these groups are referred to as strata, and they are formed on the basis of shared characteristics. Strata may have quite different means on variables of interest. A random sample is taken from each stratum in proportion to its relative size in the population, and these subsamples are then pooled to form the total sample. Normative samples of psychological tests are often stratified on the basis of combinations of variables such as age, gender, or other demographic characteristics.

Partitions of the population are called clusters in cluster sampling. Each cluster should be generally representative of the whole population, which implies that clusters should also be reasonably similar on average. That is, most of the variation should be within clusters, not between them. In **single-stage cluster sampling**, random sampling is used to select the particular clusters to study. Next, all elements from the selected clusters contribute to the total sample, but no observations from the unselected clusters are included. In **two-stage cluster sampling**, elements from within each selected cluster are randomly sampled. One benefit of cluster sampling is that costs are reduced by studying some but not all clusters. When clusters are geographic areas, cases in the final sample are from the selected regions only.

Random sampling implies independent observations, which means that the score of one case does not influence the score of any other. If couples complete a relationship satisfaction questionnaire in the presence of each other, their responses may not be independent. The independence assumption is critical in many types of statistical techniques. Scores from repeated measurement of the same case are probably not independent, but techniques for such data estimate the degree of dependence in the scores and thus control for it. If scores are really not independent, results of analyses that assume independence could be biased. There is no magic statistical fix for lack of independence. Therefore, the independence requirement is usually met through design, measurement, and use of statistical techniques that take explicit account of score dependence, such as designs with repeated measures.

The discussion that follows assumes that random samples are not extraordinarily small, such as $N = 2$. More sophisticated ways to estimate minimum sample sizes are considered later, but for now let us assume more reasonable sample sizes of, say, $N = 50$ or so. There are misconceptions about random sampling. Suppose that a simple random sample is selected. What can be said about the characteristics of the observations in that sample? A common but incorrect response is to say that the observations are representative of the population. But this may not be true, because there is no guarantee that the characteristics of any particular random sample will match those in the population. People in a random sample could be older, more likely to be women, or wealthier compared with the general population. A stratified random sample may be representative in terms of the strata on which it is based (e.g., gender), but results on other, nonstrata variables are not guaranteed to be representative. It is only across replications, or in the long run, that characteristics of observations in random samples reflect those in the population. That is, random sampling generates representative samples on average over replications. This property explains the role of random sampling in the **population inference model**, which is concerned with generalizability of sample results (external validity).

There is a related misunderstanding about **randomization**, or random assignment of cases to conditions (e.g., treatment vs. control). A particular randomization is not guaranteed to result in equivalent groups such that there are no initial group differences confounded with the treatment effect. Randomization results in equivalent groups only on average. Sometimes it happens that randomly formed groups are clearly not equal on some characteristic. The expression "failure of random assignment" is used to describe this situation, but it is a misnomer because it assumes that randomization should guarantee equivalence every time it is used. Random assignment is part of the **randomization model**, which deals with the correctness of causal inference that treatment is responsible for changes among treated cases (internal validity).

The use of random sampling and randomization together—the **statistician's two-step**—guarantees that the average effect observed over replications of treatment–control comparisons will converge on the value of the population treatment effect. But this ideal is almost never achieved in real studies. This is because random sampling requires a list of all observations in the population, but such lists rarely exist. Randomization is widely used in experimental studies but usually with nonrandom samples. Many more studies are based on the randomization model than on the population inference model, but it is the latter that is assumed by the probabilities, or *p* values, generated by statistical tests and used in confidence intervals.

Observations in **systematic samples** are selected according to an orderly sampling plan that may yield a representative sample, but this is not certain. Suppose that an alphabetical list of every household is available for some area. A random number between 10 and 20 is generated and turns out to be 17. Every 17th household on the list is contacted for an interview, which yields a 6% (1/17) sample in that area. Systematic samples are relatively rare in the behavioral sciences.

Most samples are neither random nor systematic but rather are **ad hoc samples**, also known as **convenience samples**, **accidental samples**, or **locally available samples**. Cases in such samples are selected because they happen to be available. Whether ad hoc samples are representative is often a concern. Volunteers differ from nonvolunteers, for example, and patients seen in one clinic may differ from those treated in others. One way to mitigate bias is to measure a posteriori a variety of sample characteristics and report them. This allows others to compare the sample with those in related studies. Another option is to compare the sample profile with that of the population (if such a profile exists) in order to show that an ad hoc sample is not grossly unrepresentative.

The cases in a **purposive sample** are intentionally selected from defined groups or dimensions in ways linked to hypotheses. A researcher who wishes to evaluate whether the effectiveness of a drug varies by gender would intentionally select both women and men. After the data are collected, gender would be represented as a factor in the analysis, which may facilitate generalization of the results to both genders. A purposive sample is usually a convenience sample, and dividing cases by gender or some other variable does not change this fact.

Sampling Error

This discussion assumes a population size that is very large and assumes that the size of each sample is a relatively small proportion of the total population size. There are some special corrections if the population size is small, such as less than 5,000 cases, or if the sample size exceeds 20% or so of the population size that are not covered here (see S. K. Thompson [2012] for more information).

Values of population parameters, such as means (μ) or variances (σ^2), are usually unknown. They are instead estimated with sample statistics, such as M (means) or s^2 (variances). Statistics are subject to **sampling error**, which refers to the difference between an estimator and the corresponding parameter (e.g., $\mu - M$). These differences arise because the values of statistics from random samples vary around that of the parameter. Some of these statistics will be too high and others too low (i.e., they over- or underestimate the parameter), and only a relatively small number will exactly equal the population value. This variability among estimators is a statistical phenomenon akin to background (natural) radiation: It is always there, sometimes more or less, fluctuating randomly from sample to sample.

The amount of sampling error is generally affected by the variability of population observations, how the samples are selected, and their size. If the population is heterogeneous, values of sample statistics may also be quite variable. Obviously, estimators from biased samples may differ substantially from those of the corresponding parameters. But assuming random sampling and constant variability in the population, sampling error varies inversely with sample size. This means that statistics in larger samples tend to be closer on average than those in smaller samples to the corresponding parameter. This property describes the **law of large numbers**, and it says that one is more likely to get more accurate estimates from larger samples than smaller samples with random sampling.

It is a myth that the larger the sample, the more closely it approximates a normal distribution. This idea probably stems from a misunderstanding of the **central limit theorem**, which applies to certain group statistics such as means. This theorem predicts that (a) distributions of random means, each based on the same number of scores, get closer to a normal distribution as the sample size increases, and (b) this happens regardless of whether the population distribution is normal or not normal. This theorem justifies approximating distributions of random means with normal curves, but it does not apply to distributions of scores in individual samples. Thus, larger samples do not generally have more normal distributions than smaller samples. If the population distribution is, say, positively skewed, this shape will tend to show up in the distributions of random samples that are either smaller or larger.

The sample mean describes the central tendency of a distribution of scores on a continuous variable. It is the balance point in a distribution, because the mean is the point from which (a) the sum of deviations from M equals zero and (b) the sum of squared deviations is as small as possible. The latter quantity is the **sum of squares** (*SS*). That is, if *X* represents individual observations, then

$$\sum(X - M) = 0 \text{ and the quantity } SS = \sum(X - M)^2 \tag{2.1}$$

takes on the lowest value possible in a particular sample. Due to these properties, sample means are described as **least squares estimators**. The statistic M is also an **unbiased estimator** because its expected value across random samples of the same size is the population mean μ.

The sample variance s^2 is another least squares estimator. It estimates the population variance σ^2 without bias if computed as

$$s^2 = \frac{SS}{df} \tag{2.2}$$

where $df = N - 1$. But the sample variance derived as

$$S^2 = \frac{SS}{N} \tag{2.3}$$

is a **negatively biased estimator** because its values are on average less than σ^2. The reason is that squared deviations are taken from M (Equation 2.1), which is not likely to equal μ. Therefore, sample sums of squares are generally too small compared with taking squared deviations from μ. The division of *SS* by *df* instead of *N*, which makes the whole ratio larger ($s^2 > S^2$), is sufficient to render s^2 an unbiased estimator. In larger samples, though, the values of s^2 and S^2 converge, and in very large samples they are asymptotically equal. Expected values of **positively biased estimators** exceed those of the corresponding parameter.

There are ways to correct other statistics for bias. For example, although s^2 is an unbiased estimator of σ^2, the sample standard deviation *s* is a negatively biased estimator of σ. Multiplication of *s* by the correction factor in parentheses that follows

$$\hat{\sigma} = \left(1 + \frac{1}{4df}\right)s \tag{2.4}$$

yields a numerical approximation to the unbiased estimator of σ. Because the value of the correction factor in Equation 2.4 is larger than 1.00, $\hat{\sigma} > s$. There is also greater correction for negative bias in smaller samples than in larger samples. If $N = 5$, for example, the value of the correction factor is 1.0625, but for $N = 50$ it is 1.0051, which shows relatively less adjustment for bias in the larger sample. For very large samples, the value of the correction factor is essentially 1.0. This is another instance of the law of large numbers: Averages of even biased statistics from large random samples tend to closely estimate the corresponding parameter.

A **standard error** is the standard deviation in a **sampling distribution**, the probability distribution of a statistic across all random samples drawn from the same population(s) and with each sample based on the same number of cases. It estimates the amount of sampling error in standard deviation units. The square of a standard error is the error variance. Standard errors of statistics with simple distributions can be estimated with formulas that have appeared in statistics textbooks for some time. By "simple" I mean that (a) the statistic estimates only a single parameter and (b) both the shape and variance of its sampling distribution are constant regardless of the value of that parameter. Distributions of M and s^2 are simple as just defined.

The standard error in a distribution of random means is

$$\sigma_M = \frac{\sigma}{\sqrt{N}} \tag{2.5}$$

Because σ is not generally known, this standard error is typically estimated as

$$s_M = \frac{s}{\sqrt{N}} \tag{2.6}$$

As either sample variability decreases or the sample size increases, the value of s_M decreases. For example, given $s = 10.00$, s_M equals $10.00/25^{1/2}$, or 2.00, for $N = 25$, but for $N = 100$ the value of s_M is $10.00/100^{1/2}$, or 1.00. That is, the standard error is twice as large for $N = 25$ as it is for $N = 100$. A graphical illustration is presented in Figure 2.1. An original normal distribution is shown along with three different sampling distributions of M based on $N = 4$, 16, or 64 cases. Variability of the sampling distributions in the figure decreases as the sample size increases.

The standard error s_M, which estimates variability of the group statistic M, is often confused with the standard deviation s, which measures variability at the case level. This confusion is a source of misinterpretation of both statistical tests and confidence intervals (Streiner, 1996). Note that

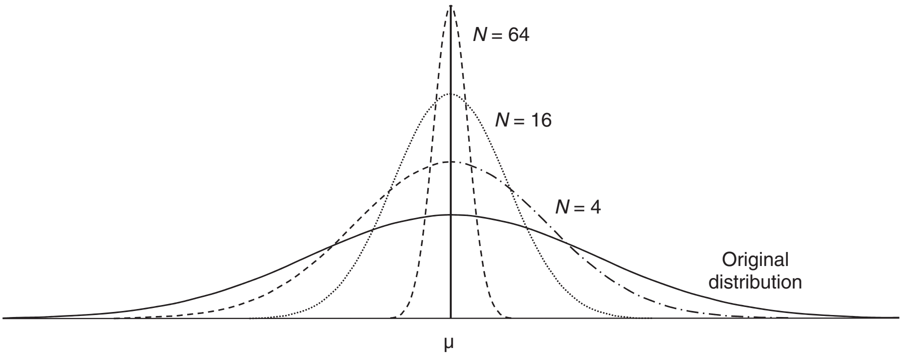

Figure 2.1. An original distribution of scores and three distributions of random sample means each based on different sample sizes, *N* = 4, *N* = 16, or *N* = 64.

the standard error s_M itself has a standard error (as do standard errors for all other kinds of statistics). This is because the value of s_M varies over random samples. This explains why one should not overinterpret a confidence interval or *p* value from a significance test based on a single sample. Exercises 1–2 concern the distinction between *s* and s_M.

Distributions of random means follow **central (Student's) *t* distributions** with degrees of freedom equal to $N - 1$ when σ is unknown. For very large samples, central *t* distributions approximate a normal curve. In **central test distributions**, the null hypothesis is assumed to be true. They are used to determine critical values of test statistics. Tables of critical values for distributions such as *t*, *F*, and χ^2 found in many statistics textbooks are based on central test distributions. There are also web calculating pages that generate critical values for central test statistics.[1] The *t* distribution originated from "Student's" (William Gosset's) attempt to approximate the distributions of means when the sample size is not large and σ is unknown. It was only later that central *t* distributions and other theoretical probability distributions were associated with the practice of significance testing.

The sample variance s^2 follows a **central χ^2 distribution** with $N - 1$ degrees of freedom. Listed next is the equation for the standard error of s^2 when the population variance is known:

$$\sigma_{s^2} = \sigma^2 \sqrt{\frac{2}{df}} \qquad (2.7)$$

If σ^2 is not known, the standard error of the sample variance is estimated as

$$s_{s^2} = s^2 \sqrt{\frac{2}{df}} \qquad (2.8)$$

As with M, the estimated standard error of s^2 becomes smaller as the sample size increases.

Other Kinds of Error

Standard errors estimate sampling error under random sampling. What they measure when sampling is not random may not be clear. The standard error in an ad hoc sample might reflect both sampling error and systematic

[1] This central *t* distributional calculator accepts either integer or noninteger *df* values: http://www.usablestats.com/calcs/tinv

selection bias that results in nonrepresentative samples. Standard errors also ignore the other sources of error described next:

1. **Measurement error** refers to the difference between an observed score X and the true score on the underlying construct. The reliability coefficient r_{XX} estimates the degree of measurement error in a particular sample. If $r_{XX} = .80$, for example, at least $1 - .80 = .20$, or 20%, of the observed variance in X is due to random error of the type estimated by that particular reliability coefficient. Measurement error reduces absolute effect sizes and the power of statistical tests. It is controlled by selecting measures that generally yield scores with good psychometric characteristics.
2. **Construct definition error** involves problems with how hypothetical constructs are defined or operationalized. Incorrect definition could include mislabeling a construct, such as when low IQ scores among minority children who do not speak English as a first language are attributed to low intelligence instead of to limited language familiarity. Error can also stem from **construct proliferation**, where a researcher postulates a new construct that is questionably different from existing constructs (F. L. Schmidt, 2010). Constructs that are theoretically distinct in the minds of researchers are not always empirically distinct.
3. **Specification error** refers to the omission from a regression equation of at least one predictor that covaries with the measured (included) predictors.[2] As covariances between omitted and included predictors increase, results based on the included predictors tend to become increasingly biased. Careful review of theory and research when planning a study is the main way to avoid a serious specification error by decreasing the potential number of left-out variables.
4. **Treatment implementation error** occurs when an intervention does not follow prescribed procedures. The failure to ensure that patients take an antibiotic medication for the prescribed duration of time is an example. Prevention includes thorough training of those who will administer the treatment and checking after the study begins whether implementation remains consistent and true.

[2] It can also refer to including irrelevant predictors, estimating linear relations only when the true relation is curvilinear, or estimating main effects only when there is true interaction.

Shadish, Cook, and Campbell (2001) described additional potential sources of error. Gosset used the term **real error** to refer all types of error besides sampling error (e.g., Student, 1927). In reasonably large samples, the impact of real error may be greater than that of sampling error. Thus, it is unwise to acknowledge sampling error only. This discussion implies that the probability that error of any kind affects sample results is virtually 1.00, and, therefore, practically all sample results are wrong (the parameter is not correctly estimated). This may be especially true when sample sizes are small, population effect sizes are not large, researchers chase statistical significance instead of substantive significance, a greater variety of methods is used across studies, and there is financial or other conflict of interest (Ioannidis, 2005).

INTERVAL ESTIMATION

Assumed next is the selection of a very large number of random samples from a very large population. The amount of sampling error associated with a statistic is explicitly indicated by a confidence interval, precisely defined by Steiger and Fouladi (1997) as follows:

1. A $1 - \alpha$ confidence interval for a parameter is a pair of statistics yielding an interval that, over many random samples, includes the parameter with the probability $1 - \alpha$. (The symbol α is the level of statistical significance.)
2. A $100\ (1 - \alpha)\%$ confidence interval for a parameter is a pair of statistics yielding an interval that, over many random samples, includes the parameter $100\ (1 - \alpha)\%$ of the time.

The value of $1 - \alpha$ is selected by the researcher to reflect the degree of statistical uncertainty due to sampling error. Because the conventional levels of statistical significance are .05 or .01, one usually sees either 95% or 99% confidence intervals, but it is possible to specify a different level, such as $\alpha = .10$ for a 90% confidence interval. Next we consider 95% confidence intervals only, but the same ideas apply to other confidence levels.

The lower bound of a confidence interval is the **lower confidence limit**, and the upper bound is the **upper confidence limit**. The *Publication Manual* (APA, 2010) recommends reporting a confidence interval in text with brackets. If 21.50 and 30.50 are, respectively, the lower and upper bounds for the 95% confidence interval based on a sample mean of 26.00, these results would be summarized as

$$M = 26.00,\ 95\%\ \mathrm{CI}\ [21.50,\ 30.50]$$

Confidence intervals are often shown in graphics as **error bars** represented as lines that extend above and below (or to the left and right, depending on orientation) around a point that corresponds to a statistic. When the length of each error bar is one standard error ($M \pm s_M$), the interval defined by those **standard error bars** corresponds roughly to $\alpha = .32$ and a 68% confidence interval. There are also **standard deviation bars**. For example, the interval $M \pm s$ says something about the variability of scores around the mean, but it conveys no direct information about the extent of sampling error associated with that mean. Researchers do not always state what error bars represent: About 30% of articles with such figures reviewed by Cumming, Fidler, and Vaux (2007) did not provide this information.

Traditional confidence intervals are based on central test distributions, and the statistic is usually exactly between the lower and upper bounds (the interval is symmetrical about the estimator). The interval is constructed by adding and subtracting from a statistic the product of its standard error and the positive two-tailed critical value at the α level of statistical significance in a relevant central test distribution. This product is the **margin of error**. In graphical displays of confidence intervals, each of the two error bars corresponds to a margin of error.

Confidence Intervals for μ

The relevant test statistic for means when σ is unknown is central t, so the general form of a 100 $(1 - \alpha)$% confidence interval for μ based on a single observed mean is

$$M \pm s_M \left[t_{2\text{-tail}, \alpha} (N - 1) \right] \tag{2.9}$$

where the term in brackets is the positive two-tailed critical value in a central t distribution with $N - 1$ degrees of freedom at the α level of statistical significance. Suppose that

$$M = 100.00,\ s = 9.00, \text{ and } N = 25$$

The standard error is

$$s_M = \frac{9.00}{\sqrt{25}} = 1.800$$

and $t_{2\text{-tail}, .05}(24) = 2.064$. The 95% confidence interval for μ is thus

$$100.00 \pm 1.800\,(2.064), \text{ or } 100.00 \pm 3.72$$

which defines the interval [96.28, 103.72]. Exercise 3 asks you to verify that the 99% confidence interval is wider than the 95% confidence interval based on the same data. Cumming (2012) described how to construct one-sided confidence intervals that are counterparts to statistical tests of null hypothesis versus directional (one-tailed) alternative hypotheses, such as H_1: $\mu > 130.00$.

Let us consider how to interpret the specific 95% confidence interval for μ just derived:

1. The interval [96.28, 103.72] defines a range of values considered equivalent within the limits of sampling error at the 95% confidence level. But equivalent within the bounds of sampling error does not imply equivalent in a scientific sense. This is especially true when the range of values included in the confidence interval indicates very different outcomes, such as when the upper confidence limit for the average blood concentration of a drug exceeds a lethal dosage.
2. It also provides a reasonable estimate of the population mean. That is, μ could be as low as 96.28 or μ could be as high as 103.72, again at the 95% confidence level.
3. There is no guarantee that μ is actually included in the confidence interval. We could construct the 95% confidence interval based on the mean in a different sample, but the center or endpoints of this new interval will probably be different. This is because confidence intervals are subject to sampling error, too.
4. If 95% confidence intervals are constructed around the means of very many random samples drawn from the same very large population, a total of 95% of them will contain μ.

The last point gives a more precise definition of "95% confident" from a **frequentist** or **long-run relative-frequency** view of probability as the likelihood of an outcome over repeatable events under constant conditions except for random error. A frequentist view assumes that probability is a property of nature that is independent of what the researcher believes. In contrast, a **subjectivist** or **subjective degree-of-belief** view defines probability as a personal belief that is independent of nature. The same view also does not distinguish between repeatable and unique events (Oakes, 1986). Although researchers in their daily lives probably take a subjective view of probabilities, it is the frequentist definition that generally underlies sampling theory.

A researcher is probably more interested in knowing the probability that a specific 95% confidence interval contains μ than in knowing that

95% of all such intervals do. From a frequentist perspective, this probability for any specific interval is either 0 or 1.00; that is, either the interval contains the parameter or it does not. Thus, it is generally incorrect to say that a specific 95% confidence interval has a 95% likelihood of including the corresponding parameter. Reichardt and Gollob (1997) noted that this kind of **specific probability inference** is permitted only in the circumstance that every possible value of the parameter is considered equally likely before the data are collected. In Bayesian estimation, the same circumstance is described by the **principle of indifference**, but it is rare when a researcher truly has absolutely no information about plausible values for a parameter.

There is language that splits the difference between frequentist and subjectivist perspectives. Applied to our example, it goes like this: The interval [96.28, 103.72] estimates μ, with 95% confidence. This statement is not quite a specific probability inference, and it also gives a nod to the subjectivist view because it associates a degree of belief with a unique interval. Like other compromises, however, it may not please purists who hold one view of probability or the other. But this wording does avoid the blatant error of claiming that a specific 95% confidence interval contains the parameter with the probability .95.

Another interpretation concerns the **capture percentage** of random means from replications that fall within the bounds of a specific 95% confidence interval for μ. Most researchers surveyed by Cumming, Williams, and Fidler (2004) mistakenly endorsed the **confidence-level misconception** that the capture percentage for a specific 95% confidence interval is also 95%. This fallacy for our example would be stated as follows: The interval [96.28, 103.72] contains 95% of all replication means. This statement would be true for this interval only if the values of $\mu - M$ and $\sigma - s$ were both about zero; otherwise, capture percentages drop off quickly as the absolute distance between μ and M increases. Cumming and Maillardert (2006) estimated that the average capture percentage across random 95% confidence intervals for μ is about 85% assuming normality and $N \geq 20$, but percentages for more extreme samples are much lower (e.g., < 50%).

These results suggest that researchers underestimate the impact of sampling error on means. Additional evidence described in the next chapter says that researchers fail to appreciate that sampling error affects *p* values from statistical tests, too. It seems that many researchers believe that results from small samples behave like those from large samples; that is, they believe that results from small samples are likely to replicate. Tversky and Kahneman (1971) labeled such errors **the law of small numbers**, an ironic twist on the law of large numbers, which (correctly) says that there is greater variation across results from small samples than from large samples.

Confidence Intervals for $\mu_1 - \mu_2$

Next we assume a design with two independent samples. The standard error in a distribution of contrasts between pairs of means randomly selected from different populations is

$$\sigma_{M_1-M_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{2.10}$$

where σ_1^2 and σ_2^2 are the population variances and n_1 and n_2 are the sizes of each group. If we assume homogeneity of population variance or **homoscedasticity** (i.e., $\sigma_1^2 = \sigma_2^2$), the expression for the standard error reduces to

$$\sigma_{M_1-M_2} = \sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{2.11}$$

where σ^2 is the common population variance. This parameter is usually unknown, so the standard error of mean differences is estimated by

$$s_{M_1-M_2} = \sqrt{s_{\text{pool}}^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{2.12}$$

where s_{pool}^2 is the weighted average of the within-groups variances. Its equation is

$$s_{\text{pool}}^2 = \frac{df_1(s_1^2) + df_2(s_2^2)}{df_1 + df_2} = \frac{SS_W}{df_W} \tag{2.13}$$

where s_1^2 and s_2^2 are the group variances, $df_1 = n_1 - 1$, $df_2 = n_2 - 1$, and SS_W and df_W are, respectively, the pooled within-groups sum of squares and the degrees of freedom. The latter can also be expressed as $df_W = N - 2$. Only when the group sizes are equal can s_{pool}^2 also be calculated as the simple average of the two group variances, or $(s_1^2 + s_2^2)/2$.

The general form of a 100 $(1 - \alpha)$% confidence interval for $\mu_1 - \mu_2$ based on the difference between two independent means is

$$(M_1 - M_2) \pm s_{M_1-M_2}\left[t_{2\text{-tail},\,\alpha}(N-2)\right] \tag{2.14}$$

Suppose in a design with $n = 10$ cases in each group we observe

$$M_1 = 13.00,\ s_1^2 = 7.50 \quad \text{and} \quad M_2 = 11.00,\ s_2^2 = 5.00$$

which implies $M_1 - M_2 = 2.00$ and $s^2_{\text{pool}} = (7.50 + 5.00)/2 = 6.25$. The estimated standard error is

$$s_{M_1-M_2} = \sqrt{6.25\left(\frac{1}{10}+\frac{1}{10}\right)} = 1.118$$

and $t_{\text{2-tail, .05}}\,(18) = 2.101$. The 95% confidence interval for $\mu_1 - \mu_2$ is

$$2.00 \pm 1.118\,(2.101)$$

which defines the interval [–.35, 4.35]. On the basis of these results, we can say that $\mu_1 - \mu_2$ could be as low as –.35 or as high as 4.35, with 95% confidence.

The specific interval [–.35, 4.35] includes zero as an estimate of $\mu_1 - \mu_2$. This fact is subject to misinterpretation. For example, it may be incorrectly concluded that $\mu_1 = \mu_2$ because zero falls within the interval. But zero is only one value within a range of estimates of $\mu_1 - \mu_2$, so it has no special status in interval estimation. Confidence intervals are subject to sampling error, so zero may not be included within the 95% confidence interval in a replication. Confidence intervals also assume that other sources of error are nil. All these caveats should reduce the temptation to fixate on a particular value (here, zero) in a confidence interval.

There is special relation between a confidence interval for $\mu_1 - \mu_2$ and the outcome of the independent samples *t* test based on the same data: Whether a 100 $(1 - \alpha)$% confidence interval for $\mu_1 - \mu_2$ includes zero yields an outcome equivalent to either rejecting or not rejecting the corresponding null hypothesis at the α level of statistical significance for a two-tailed test. For example, the specific 95% confidence interval [–.35, 4.35] includes zero; thus, the outcome of the *t* test for these data of H_0: $\mu_1 - \mu_2 = 0$ is not statistically significant at the .05 level, or

$$t\,(18) = \frac{2.00}{1.118} = 1.789,\ p = .091$$

But if zero is not contained within a particular 95% confidence interval for $\mu_1 - \mu_2$, the outcome of the independent samples *t* test will be statistically significant at the .05 level.

Be careful not to falsely believe that confidence intervals are just statistical tests in disguise (B. Thompson, 2006a). One reason is that null hypotheses are required for statistical tests but not for confidence intervals. Another is that many null hypotheses have little if any scientific value. For example, Anderson et al. (2000) reviewed null hypotheses tested in several hundred empirical studies published from 1978 to 1998 in two environmental sciences

TABLE 2.1
Results of Six Hypothetical Replications

Study	$M_1 - M_2$	s_1^2	s_2^2	t (38)	Reject H_0?	95% CI
1	2.50	17.50	16.50	1.92	No	−.14, 5.14
2	4.00	16.00	18.00	3.07	Yes	1.36, 6.64
3	2.50	14.00	17.25	2.00	No	−.03, 5.03
4	4.50	13.00	16.00	3.74	Yes	2.06, 6.94
5	5.00	12.50	16.50	4.15	Yes	2.56, 7.44
6	2.50	15.00	17.00	1.98	No	−.06, 5.06
Average:	3.54					2.53, 4.54

Note. Independent samples assumed. For all replications, the group size is $n = 20$, $\alpha = .05$, the null hypothesis is H_0: $\mu_1 - \mu_2 = 0$, and H_1 is two-tailed. Results for the average difference are from a meta-analysis assuming a fixed effects model. CI = confidence interval.

journals. They found many implausible null hypotheses that specified things such as equal survival probabilities for juvenile and adult members of a species or that growth rates did not differ across species, among other assumptions known to be false before collecting data. I am unaware of a similar survey of null hypotheses in the behavioral sciences, but I would be surprised if the results would be very different.

Confidence intervals over replications may be less susceptible to misinterpretation than results of statistical tests. Summarized in Table 2.1 are outcomes of six hypothetical replications where the same two conditions are compared on the same outcome variable. Results of the independent samples *t* test lead to rejection of the null hypothesis at $p < .05$ in three out of six studies, a "tie" concerning statistical significance (3 yeas, 3 nays). More informative than the number of null hypothesis replications is the average of $M_1 - M_2$ across all six studies, 3.54. This average is from a meta-analysis of all results in the table for a **fixed effects model**, where a single population effect size is presumed to underlie the observed contrasts. (I show you how to calculate this average in Chapter 9.) The overall average of 3.54 may be a better estimate of $\mu_1 - \mu_2$ than $M_1 - M_2$ in any individual study because it is based on all available data.

The 95% confidence intervals for $\mu_1 - \mu_2$ in Table 2.1 are shown in Figure 2.2 as error bars in a **forest plot**, which displays results from replications and a meta-analytic weighted average with confidence intervals (Cumming, 2012). The 95% confidence interval based on the overall average of 3.54, or [2.53, 4.54] (see Table 2.1), is narrower than any of the intervals from the six replications (see Figure 2.2). This is because more information contributes to the confidence interval based on results averaged over all replications. For these data, $\mu_1 - \mu_2$ may be as low as 2.53 or as high as 4.54, with 95% confidence based on all available data.

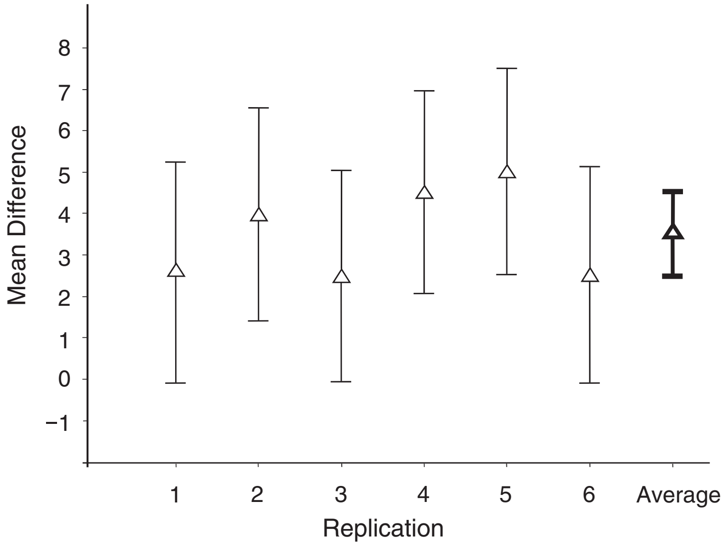

Figure 2.2. A forest plot of 95% confidence intervals for $\mu_1 - \mu_2$ based on mean differences from the six replications in Table 2.1 and the meta-analytic 95% confidence interval for $\mu_1 - \mu_2$ across all replications for a fixed effects model.

There is a widely accepted—but unfortunately incorrect—rule of thumb that the difference between two independent means is statistically significant at the α level if there is no overlap of the two 100 $(1 - \alpha)$% confidence intervals for μ (Belia, Fidler, Williams, & Cumming, 2005). It also maintains that the overlap of the two intervals indicates that the mean contrast is not statistically significant at the corresponding level of α. This rule is often applied to diagrams where confidence intervals for μ are represented as error bars that emanate outward from points that symbolize group means.

A more accurate heuristic is the **overlap rule for two independent means** (Cumming, 2012), which works best when $n \geq 10$ and the group sizes and variances are approximately equal. The overlap rule is stated next for $\alpha = .05$:

1. If there is a gap between the two 95% confidence intervals for μ (i.e., no overlap), the outcome of the independent samples *t* test of the mean difference is $p < .01$. But if the confidence intervals just touch end-to-end, *p* is approximately .01.
2. No more than moderate overlap of the 95% confidence intervals for μ implies that the *p* value for the *t* test is about .05, but less overlap indicates $p < .05$. *Moderate overlap* is about one half the length of each error bar in a graphical display.

Summarized next are the basic descriptive statistics for the example where $n_1 = n_2 = 10$:

$$M_1 = 13.00,\ s_1^2 = 7.50 \quad \text{and} \quad M_2 = 11.00,\ s_2^2 = 5.00$$

You should verify for these data the results presented next:

$$s_{M_1} = .866,\ 95\%\ \text{CI for}\ \mu_1\ [11.04, 14.96]$$

$$s_{M_2} = .707,\ 95\%\ \text{CI for}\ \mu_2\ [9.40, 12.60]$$

These confidence intervals for μ are plotted in Figure 2.3 along with the 95% confidence interval for $\mu_1 - \mu_2$ for these data [–.35, 4.35]. Group means are represented on the y-axis, and the mean contrast (2.00) is represented on the floating difference axis (Cumming, 2012) centered at the grand mean across both groups (12.00). The error bars of the 95% confidence intervals for μ overlap by clearly more than one half of their lengths. According to the overlap rule, this amount of overlap is more than moderate. So the mean difference should not be statistically significant at the .05 level, which is true for these data.

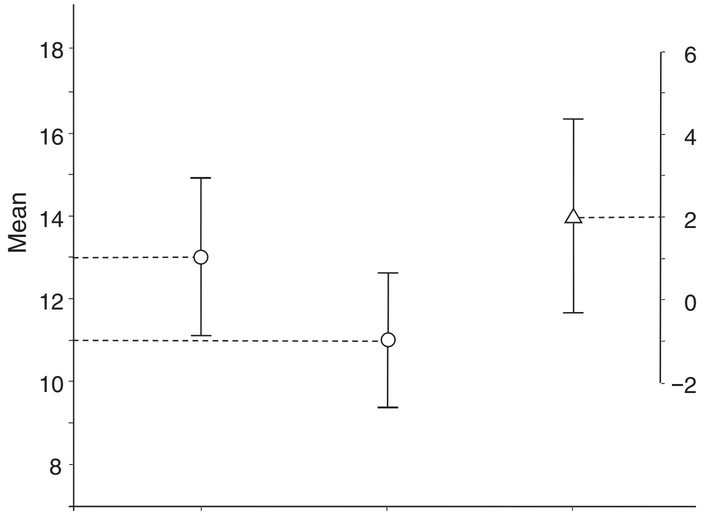

Figure 2.3. Plot of the 95% confidence interval for μ_1, 95% confidence interval for μ_2, and 95% confidence interval for $\mu_1 - \mu_2$, given $M_1 = 13.00$, $s_1^2 = 7.50$, $M_2 = 11.00$, $s_2^2 = 5.00$, and $n_1 = n_2 = 10$. Results for the mean difference are shown on a floating difference axis where zero is aligned at the grand mean across both samples (12.00).

Confidence intervals for $\mu_1 - \mu_2$ based on $s_{M_1 - M_2}$ assume homoscedasticity. In the **Welch procedure** (e.g., Welch, 1938), the standard error of a mean contrast is estimated as

$$s_{\text{Wel}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{2.15}$$

where s_1^2 estimates σ_1^2 and s_2^2 estimates σ_2^2 (i.e., heteroscedasticity is allowed). The degrees of freedom for the critical value of central *t* in the Welch procedure are estimated empirically as

$$df_{\text{Wel}} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2)^2}{n_1^2(n_1 - 1)} + \frac{(s_2^2)^2}{n_2^2(n_2 - 1)}} \tag{2.16}$$

Summarized next are descriptive statistics for two groups:

$$\text{M}_1 = 112.50, s_1^2 = 75.25, n_1 = 25$$

$$\text{M}_2 = 108.30, s_2^2 = 15.00, n_2 = 20$$

Variability among cases in the first group is obviously greater than that in the second group. A pooled within-groups variance would mask this discrepancy. The researcher elects to use the Welch procedure. The estimated standard error is

$$s_{\text{Wel}} = \sqrt{\frac{75.25}{25} + \frac{15.00}{20}} = 1.939$$

and the approximate degrees of freedom are

$$df_{\text{Wel}} = \frac{\left(\frac{75.25}{25} + \frac{15.00}{20}\right)^2}{\frac{75.25^2}{25^2(24)} + \frac{15.00^2}{20^2(19)}} = 34.727$$

The general form of a 100 (1 – α)% confidence interval for $\mu_1 - \mu_2$ in the Welch procedure is

$$(\text{M}_1 - \text{M}_2) \pm s_{\text{Wel}} \left[t_{2\text{-tail},\,\alpha}(df_{\text{Wel}})\right] \tag{2.17}$$

Tables for critical values of central *t* typically list integer *df* values only. An alternative is to use a web distributional calculator page that accepts noninteger *df* (see footnote 1). Another is to use a statistical density function built into widely available software. The statistical function TINV in Microsoft Excel returns critical values of central *t* given values of α and *df*. The function Idf.T (Inverse DF) in SPSS returns the two-tailed critical value of central *t* given *df* and $1 - \alpha/2$, which is .975 for a 95% confidence interval. For this example, SPSS returned

$$t_{\text{2-tail, .05}}\,(34.727) = 2.031$$

The 95% confidence interval for $\mu_1 - \mu_2$ is

$$(112.50 - 108.30) \pm 1.939\,(2.031)$$

which defines the interval [.26, 8.14]. Thus, the value of $\mu_1 - \mu_2$ could be as low as .26 or as high as 8.14, with 95% confidence and not assuming homoscedasticity. Widths of confidence intervals in the Welch procedure tend to be narrower than intervals based on $s_{M_1 - M_2}$ for the same data when group variances are unequal. Welch intervals may less accurate when the population distributions are severely and differently nonnormal or when the group sizes are unequal and small, such as $n < 30$ (Bonett & Price, 2002); see also Grissom and Kim (2011, Chapter 2).

Confidence Intervals for μ_D

I use the symbol M_D to refer to the mean **difference (change**, **gain) score** when two dependent samples are compared. A difference score is computed as $D = X_1 - X_2$ for each of the *n* cases in a repeated measures design or for each of the *n* pairs of cases in a matched groups design. If $D = 0$, there is no difference; any other value indicates a higher score in one condition than in the other. The average of all difference scores equals the dependent mean contrast, or $M_D = M_1 - M_2$. Its standard error is

$$\sigma_{M_D} = \frac{\sigma_D}{\sqrt{n}} \tag{2.18}$$

where σ_D is the population standard deviation of the difference scores. The variance of the difference scores can be expressed as

$$\sigma_D^2 = 2\sigma^2\,(1 - \rho_{12}) \tag{2.19}$$

where σ^2 is the common population variance assuming homoscedasticity and ρ_{12} is the population cross-conditions correlation of the original scores.

When there is a stronger **subjects effect**—cases maintain their relative positions across the conditions—ρ_{12} approaches 1.00. This reduces the variance of the difference scores, which in turn lowers the standard error of the mean contrast (Equation 2.18). It is the subtraction of consistent individual differences from the standard error that makes confidence intervals based on dependent mean contrasts generally narrower than confidence intervals based on contrasts between unrelated means. It also explains the power advantage of the *t* test for dependent samples over the *t* test for independent samples. But these advantages are realized only if $\rho_{12} > .50$ (Equation 2.19); otherwise, confidence intervals and statistical tests may be wider and less powerful (respectively) for dependent mean contrasts.

The standard deviation σ_D is usually unknown, so the standard error of M_D is estimated as

$$s_{\text{M}_D} = \frac{s_D}{\sqrt{n}} \tag{2.20}$$

where s_D is the sample standard deviation of the *D* scores. The corresponding variance is

$$s_D^2 = s_1^2 + s_2^2 - 2cov_{12} \tag{2.21}$$

where cov_{12} is the cross-conditions covariance of the original scores. The latter is

$$cov_{12} = r_{12}\ s_1\ s_2 \tag{2.22}$$

where r_{12} is the sample cross-conditions correlation. (The correlation r_{12} is presumed to be zero when the samples are independent.)

The general form of a 100 $(1 - \alpha)$% confidence interval for μ_D is

$$\text{M}_D \pm s_{\text{M}_D} \left[t_{\text{2-tail},\,\alpha}\left(n-1\right)\right] \tag{2.23}$$

Presented in Table 2.2 are raw scores and descriptive statistics for a small data set where the mean contrast is 2.00. In a dependent samples analysis of these data, $n = 5$ and $r_{12} = .735$. The cross-conditions covariance is

$$cov_{12} = .735\left(2.739\right)\left(2.236\right) = 4.50$$

and the variance of the difference scores is

$$s_D^2 = 7.50 + 5.00 - 2\left(4.50\right) = 3.50$$

which implies that $s_D = 3.50^{1/2}$, or 1.871. The standard error of $\text{M}_D = 2.00$ is estimated as

TABLE 2.2
Raw Scores and Descriptive Statistics for Two Samples

	Sample	
	1	2
	9	8
	12	12
	13	11
	15	10
	16	14
M	13.00	11.00
s^2	7.50	5.00
s	2.739	2.236

Note. In a dependent samples analysis, $r_{12} = .735$.

$$s_{M_D} = \frac{1.871}{\sqrt{5}} = .837$$

The value of $t_{\text{2-tail, .05}}$ (4) is 2.776, so the 95% confidence interval for μ_D is

$$2.00 \pm .837(2.776)$$

which defines the interval [–.32, 4.32]. Exercise 4 asks you to verify that the 95% confidence interval for μ_D assuming a correlated design is narrower than the 95% confidence interval for $\mu_1 - \mu_2$ assuming unrelated samples for the same data (see Table 2.2), which is [–1.65, 5.65].

Confidence Intervals Based on Other Kinds of Statistics

Many statistics other than means have complex distributions. For example, distributions of the Pearson correlation *r* are symmetrical only if the population correlation is $\rho = 0$, but they are negatively skewed when $\rho > 0$ and positively skewed when $\rho < 0$. Other statistics have complex distributions, including some widely used effect sizes introduced in Chapter 5, because they estimate more than one parameter.

Until recently, confidence intervals for statistics with complex distributions were estimated with approximate methods. One method involves **confidence interval transformation** (Steiger & Fouladi, 1997), where the statistic is mathematically transformed into normally distributed units. The confidence interval is built by adding and subtracting from the transformed statistic the product of the standard error in the transformed metric and the appropriate critical value of the normal deviate *z*. The lower and upper bounds of this interval are then transformed back into the original metric, and the resulting confidence interval may be asymmetric (unequal margins of error). **Fisher's transformation** is used to approximate construct intervals for ρ. It converts a sample correlation *r* with the function

$$Z_r = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right) \tag{2.24}$$

where ln is the natural log function to base e, which is about 2.7183. The sampling distribution of Z_r is approximately normal with the standard error

$$s_{Z_r} = \sqrt{\frac{1}{N-3}} \tag{2.25}$$

The lower and upper bounds of the 100 (1 – α)% confidence interval based on Z_r are defined by

$$Z_r \pm s_{Z_r} \left(z_{2\text{-tail}, \alpha} \right) \tag{2.26}$$

where $z_{2\text{-tail}, \alpha}$ is the positive two-tailed critical value of the normal deviate, which is 1.96 for $\alpha = .05$ and the 95% confidence level. Next, transform both the lower and upper bounds of the confidence interval in Z_r units back to *r* units by applying the inverse transformation

$$r_Z = \frac{e^{2Z_r} - 1}{e^{2Z_r} + 1} \tag{2.27}$$

There are calculating web pages that automatically generate approximate 95% or 99% confidence intervals for ρ, given values of *r* and the sample size.[3] Four-decimal accuracy is recommended for hand calculation.

In a sample of *N* = 20 cases, *r* = .6803. Fisher's transformation and its standard error are

$$Z_r = \frac{1}{2} \ln \left(\frac{1 + .6803}{1 - .6803} \right) = .8297 \quad \text{and} \quad s_{Z_r} = \sqrt{\frac{1}{20 - 3}} = .2425$$

The approximate 95% confidence interval in Z_r units is

$$.8297 \pm .2425 (1.96)$$

[3] http://faculty.vassar.edu/lowry/rho.html

which defines the interval [.3544, 1.3051]. To convert the lower and upper bounds of this interval to *r* units, I apply the inverse transformation to each:

$$\frac{e^{2(.3544)} - 1}{e^{2(.3544)} + 1} = .3403 \quad \text{and} \quad \frac{e^{2(1.3051)} - 1}{e^{2(1.3051)} + 1} = .8630$$

In *r* units, the approximate 95% confidence interval for ρ is [.34, .86] at two-place accuracy.

Another approximate method builds confidence intervals directly around the sample statistic; thus, they are symmetrical about it. The width of the interval on either side is a product of the two-tailed critical value of a central test statistic and an estimate of the **asymptotic standard error**, which estimates what the standard error would be in a large sample (e.g., > 500). If the researcher's sample is not large, though, this estimate may not be accurate. Another drawback is that some statistics, such as R^2 in multiple regression, have distributions so complex that a computer is needed to estimate standard error. Fortunately, there are increasing numbers of computer tools for calculating confidence intervals, some of which are mentioned later.

A more precise method is **noncentrality interval estimation** (Steiger & Fouladi, 1997). It also deals with situations that cannot be handled by approximate methods. This approach is based on **noncentral test distributions** that do not assume a true null hypothesis. Some perspective is in order. Families of central distributions of *t*, *F*, and χ^2 (in which H_0 is assumed to be true) are special cases of noncentral distributions of each test statistic just mentioned. Compared to central distributions, noncentral distributions have an extra parameter called the **noncentrality parameter** that indicates the degree to which the null hypothesis is false.

Central *t* distributions are defined by a single parameter, the degrees of freedom (*df*), but noncentral *t* distributions are described by both *df* and the noncentrality parameter Δ (Greek uppercase delta). In two-group designs, the value of Δ for noncentral *t* is related to (but not exactly equal to) the true difference between the population means μ_1 and μ_2. The larger that difference, the more the noncentral *t* distribution is skewed. That is, if $\mu_1 > \mu_2$, then $\Delta > 0$ and the resulting noncentral *t* distributions are positively skewed, and if $\mu_1 < \mu_2$, then $\Delta < 0$ and the corresponding resulting noncentral *t* distributions are negatively skewed. But if $\mu_1 = \mu_2$ (i.e., there is no difference), then $\Delta = 0$ and the resulting distributions are the familiar and symmetrical central *t* distributions. Presented in Figure 2.4 are two *t* distributions where $df = 10$. For the central *t* distribution in the left part of the figure, $\Delta = 0$, but for the noncentral *t* distribution in the right side of the figure, $\Delta = 4.00$. (The meaning of a particular value for Δ is defined in Chapter 5.) Note in the figure that the distribution for noncentral *t* (10, 4.00) is positively skewed.

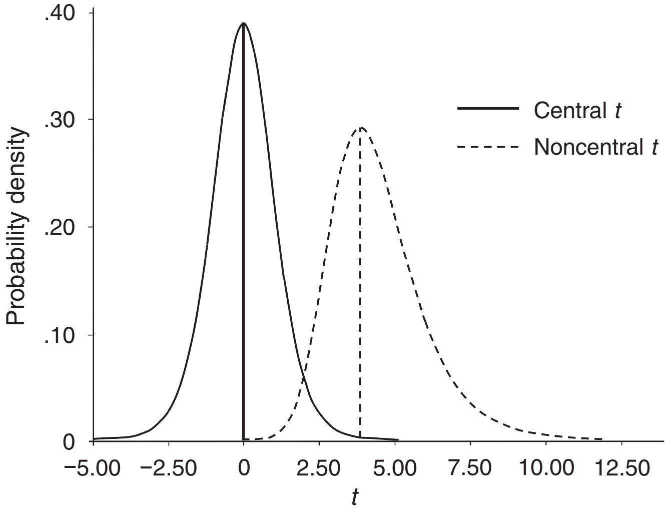

Figure 2.4. Distributions of central t and noncentral t where the degrees of freedom are $df = 10$ and where the noncentrality parameter is $\Delta = 4.00$ for noncentral t.

Noncentral test distributions play a role in estimating the power of statistical tests. This is because the concept of power assumes that the null hypothesis is false. Thus, computer tools for power analysis analyze noncentral test distributions. A population effect size that is not zero generally corresponds to a value of the noncentrality parameter that is also not zero. This is why some methods of interval estimation for effect sizes rely on noncentral test distributions. Noncentrality interval estimation for effect sizes is covered in Chapter 5.

Calculating noncentral confidence intervals is impractical without relatively sophisticated computer programs. Until recently, such programs were not widely available to applied researchers. An exception is Exploratory Software for Confidence Intervals (ESCI; Cumming, 2012), which runs under Microsoft Excel. It is structured as a tool for learning about confidence intervals, noncentral test distributions, power estimation, and meta-analysis. Demonstration modules for ESCI can be downloaded.[4] I used ESCI to create Figure 2.4.

Another computer tool for power estimation and noncentrality interval estimation is Steiger's Power Analysis procedure in STATISTICA 11 Advanced, an integrated program for general statistical analyses, data mining,

[4] http://www.thenewstatistics.com/

and quality control.[5] Power Analysis can automatically calculate noncentral confidence intervals based on several different types of effect sizes. Other computer tools or scripts for interval estimation with effect sizes are described in later chapters. The website for this book also has links to corresponding download pages. Considered next is bootstrapping, which can also be used for interval estimation.

BOOTSTRAPPED CONFIDENCE INTERVALS

The technique of **bootstrapping**, developed by the statistician Bradley Efron in the 1970s (e.g., 1979), is a computer-based method of **resampling** that recombines the cases in a data set in different ways to estimate statistical precision, with fewer assumptions than traditional methods about population distributions. Perhaps the best known form is **nonparametric bootstrapping**, which generally makes no assumptions other than that the distribution in the sample reflects the basic shape of that in the population. It treats your data file as a pseudo-population in that cases are randomly selected with replacement to generate other data sets, usually of the same size as the original. Because of sampling with replacement, (a) the same case can be selected in more than one generated data set or at least twice in the same generated sample, and (b) the composition of cases will vary slightly across the generated samples.

When repeated many times (e.g., 1,000) by the computer, bootstrapping simulates the drawing of many random samples. It also constructs an **empirical sampling distribution**, the frequency distribution of the values of a statistic across the generated samples. **Nonparametric percentile bootstrapped confidence intervals** for the parameter estimated by the statistic are calculated in the empirical distribution. The lower and upper bounds of a 95% bootstrapped confidence interval correspond to, respectively, the 2.5th and 97.5th percentiles in the empirical sampling distribution. These limits contain 95% of the bootstrapped values of the statistic.

Presented in Table 2.3 is a small data set where $N = 20$ and $r = .6803$. I used the nonparametric bootstrap procedure of SimStat for Windows (Provalis Research, 1995–2004) to resample from the data in Table 2.3 in order to generate a total of 1,000 bootstrapped samples each with 20 cases.[6] The empirical sampling distribution is presented in Figure 2.5. As expected, this distribution is negatively skewed. SimStat reported that the mean and median of the sampling distribution are, respectively, .6668 and .6837. The standard deviation in the distribution of Figure 2.5 is .1291, which is actually

[5] http://www.statsoft.com/#
[6] http://www.provalisresearch.com/

TABLE 2.3
Example Data Set for Nonparametric Bootstrapping

Case	*X*	*Y*	Case	*X*	*Y*
A	12	16	K	16	37
B	19	46	L	13	30
C	21	66	M	18	32
D	16	70	N	18	53
E	18	27	O	22	52
F	16	27	P	17	34
G	16	44	Q	22	54
H	20	69	R	12	5
I	16	22	S	14	38
J	18	61	T	14	38

the bootstrapped estimate of the standard error. The nonparametric bootstrapped 95% confidence interval for ρ is [.3615, .8626], and the bias-adjusted 95% confidence interval is [.3528, .8602]. The latter controls for lack of independence due to potential selection of the same case multiple times in the same generated sample.

The bias-adjusted bootstrapped 95% confidence interval for ρ, which is [.35, .86] at two-decimal accuracy, is similar to the approximate 95% confidence interval of [.34, .86] calculated earlier using Fisher's approximation for the same data. The bootstrapped estimate of the standard error in correlation units generated by SimStat is .129. Nonparametric bootstrapping is potentially

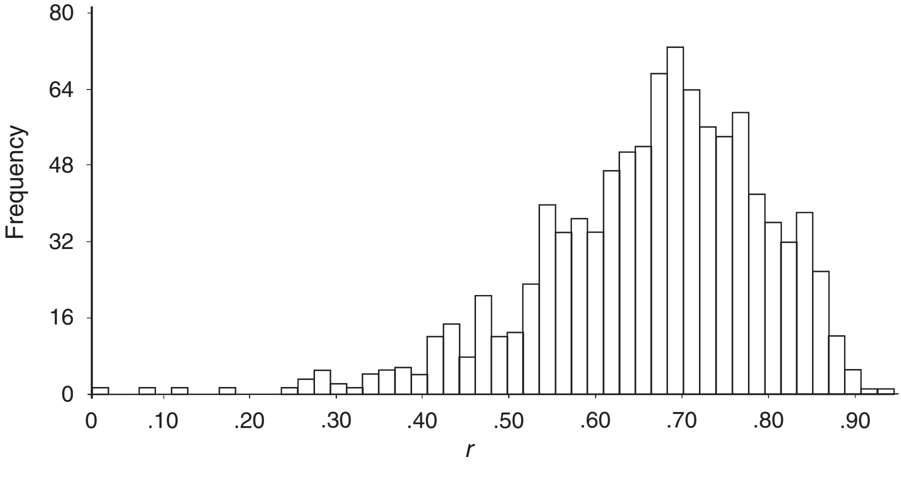

Figure 2.5. Empirical sampling distribution for the Pearson correlation *r* in 1,000 bootstrapped samples for the data in Table 2.3.

more useful when applied to statistics for which there is no approximate method for calculating standard errors and confidence intervals. This is also true when no computer tool for noncentral interval estimation is available for statistics with complex distributions.

The technique of nonparametric bootstrapping seems well suited for interval estimation when the researcher is either unwilling or unable to make a lot of assumptions about population distributions. Wood (2005) demonstrated the calculation of bootstrapped confidence intervals based on means, medians, differences between two means or proportions, correlations, and regression coefficients. His examples are implemented in an Excel spreadsheet[7] and a small stand-alone program.[8] Another computer tool is Resampling Stats (Statistics.com, 2009).[9] Bootstrapping capabilities were recently added to some procedures in SPSS and SAS/STAT.

Outlined next are potential limitations of nonparametric bootstrapping:

1. Nonparametric bootstrapping simulates random sampling, but true random sampling is rarely used in practice. This is another instance of the design–analysis mismatch.
2. It does not entirely free the researcher from having to make assumptions about population distributions. If the shape of the sample distribution is very different compared with that in the population, results of nonparametric bootstrapping may have poor external validity.
3. The "population" from which bootstrapped samples are drawn is merely the original data file. If this data set is small or the observations are not independent, resampling from it will not somehow fix these problems. In fact, resampling can magnify the effects of unusual features in a small data set (Rodgers, 2009).
4. Results of bootstrap analyses are probably quite biased in small samples, but this is true of many traditional methods, too.

The starting point for **parametric bootstrapping** is not a raw data file. Instead, the researcher specifies the numerical and distributional properties of a theoretical probability density function, and then the computer randomly samples from that distribution. When repeated many times by the computer, values of statistics in these synthesized samples vary randomly about the parameters specified by the researcher, which simulates sampling error. Bootstrapped estimation in parametric mode can also approximate standard

[7] http://woodm.myweb.port.ac.uk/nms/resample.xls
[8] http://woodm.myweb.port.ac.uk/nms/resample.exe
[9] Resampling Stats is available for a 10-day trial from http://www.resample.com/

errors for statistics where no textbook equation or approximate method is available, given certain assumptions about the population distribution. These assumptions can be added incrementally in parametric bootstrapping or successively relaxed over the generation of synthetic data sets.

ROBUST ESTIMATION

The least squares estimators M and s^2 are not robust against the effects of extreme scores. This is because their values can be severely distorted by even a single outlier in a smaller sample or by just a handful of outliers in a larger sample. Conventional methods to construct confidence intervals rely on sample standard deviations to estimate standard errors. These methods also rely on critical values in central test distributions, such as t and z, that assume normality or homoscedasticity (e.g., Equation 2.13).

Such distributional assumptions are not always plausible. For example, skew characterizes the distributions of certain variables such as reaction times. Many if not most distributions in actual studies are not even symmetrical, much less normal, and departures from normality are often strikingly large (Micceri, 1989). Geary (1947) suggested that this disclaimer should appear in all introductory statistics textbooks: "Normality is a myth; there never was, and never will be, a normal distribution" (p. 214). Keselman et al. (1998) reported that the ratios across different groups of largest to smallest variances as large as 8:1 were not uncommon in educational and psychological studies, so perhaps homoscedasticity is a myth, too.

One option to deal with outliers is to apply transformations, which convert original scores with a mathematical operation to new ones that may be more normally distributed. The effect of applying a **monotonic transformation** is to compress one part of the distribution more than another, thereby changing its shape but not the rank order of the scores. Examples of transformations that may remedy positive skew include $X^{1/2}$, $\log_{10} X$, and odd-root functions (e.g., $X^{1/3}$). There are many other kinds, and this is one of their potential problems: It can be difficult to find a transformation that works in a particular data set. Some distributions can be so severely nonnormal that basically no transformation will work. The scale of the original scores is lost when scores are transformed. If that scale is meaningful, the loss of the scaling metric creates no advantage but exacts the cost that the results may be difficult (or impossible) to interpret.

An alternative that also deals with departures from distributional assumptions is robust estimation. **Robust (resistant) estimators** are generally less affected than least squares estimators by outliers or nonnormality.

An estimator's quantitative robustness can be described by its **finite-sample breakdown point** (BP), or the smallest proportion of scores that when made arbitrarily very large or small renders the statistic meaningless. The lower the value of BP, the less robust the estimator. For both M and s^2, BP = 0, the lowest possible value. This is because the value of either statistic can be distorted by a single outlier, and the ratio 1/*N* approaches zero as sample size increases. In contrast, BP = .50 for the median because its value is not distorted by arbitrarily extreme scores unless they make up at least half the sample. But the median is not an optimal estimator because its value is determined by a single score, the one at the 50th percentile. In this sense, all the other scores are discarded by the median.

A compromise between the sample mean and the median is the **trimmed mean**. A trimmed mean M_{tr} is calculated by (a) ordering the scores from lowest to highest, (b) deleting the same proportion of the most extreme scores from each tail of the distribution, and then (c) calculating the average of the scores that remain. The proportion of scores removed from each tail is p_{tr}. If $p_{tr} = .20$, for example, the highest 20% of the scores are deleted as are the lowest 20% of the scores. This implies that

1. the total percentage of scores deleted from the distribution is $2p_{tr} = 2(.20)$, or 40%;
2. the number of deleted scores is $2np_{tr} = .40n$, where n is the original group size; and
3. the number of scores that remain is $n_{tr} = n - 2np_{tr} = n - .40n$, where n_{tr} is the trimmed group size.

For an odd number of scores, round the product np_{tr} down to the nearest integer and then delete that number of scores from each tail of the distribution. The statistics M_{tr} and M both estimate μ without bias when the population distribution is symmetrical. But if that distribution is skewed, M_{tr} estimates the trimmed population mean μ_{tr}, which is typically closer to more of the observations than μ.

A common practice is to trim 20% of the scores from each tail of the distribution when calculating trimmed estimators. This proportion tends to maintain the robustness of trimmed means while minimizing their standard errors when sampling from symmetrical distributions; it is also supported by the results of computer simulation studies (Wilcox, 2012). Note that researchers may specify $p_{tr} < .20$ if outliers constitute less than 20% of each tail in the distribution or $p_{tr} > .20$ if the opposite is true. For 20% trimmed means, BP = .20, which says they are robust against arbitrarily extreme scores unless such outliers make up at least 20% of the sample.

A variability estimator more robust than s^2 is the **interquartile range**, or the positive difference between the score that falls at the 75th percentile in a distribution and the score at the 25th percentile. Although BP = .25 for the interquartile range, it uses information from just two scores. An alternative that takes better advantage of the data is the **median absolute deviation** (MAD), the 50th percentile in the distribution of $|X - Mdn|$, the absolute differences between each score and the median. Because it is based on the median, BP = .50 for the MAD. This statistic does not estimate the population standard deviation σ, but the product of MAD and the scale factor 1.483 is an unbiased estimator of σ in a normal population distribution.

The estimator 1.483 (MAD) is part of a **robust method for outlier detection** described by Wilcox and Keselman (2003). The conventional method is to calculate for each score the normal deviate $z = (X - M)/s$, which measures the distance between each score and the mean in standard deviation units. Next, the researcher applies a rule of thumb for spotting potential outliers based on z (e.g., if $|z| > 3.00$, then X is a potential outlier). Masking, or the chance that outliers can so distort values of M or s that they cannot be detected, is a problem with this method. A more robust method is based on this decision rule applied to each score:

$$\frac{|X - Mdn|}{1.483\,(\text{MAD})} > 2.24 \qquad (2.28)$$

The value of the ratio in Equation 2.28 is the distance between a score and the median expressed in robust standard deviation units. The constant 2.24 in the equation is the square root of the approximate 97.5th percentile in a central χ^2 distribution with a single degree of freedom. A potential outlier thus has a score on the ratio in Equation 2.28 that exceeds 2.24. Wilcox (2012) described additional robust detection methods.

A robust variance estimator is the **Winsorized variance** s^2_{Win}. (The terms *Winsorized* and *Winsorization* are named after biostatistician Charles P. Winsor.) When scores are Winsorized, they are (a) ranked from lowest to highest. Next, (b) the p_{tr} most extreme scores in the lower tail of the distribution are all replaced by the next highest original score that was not replaced, and (c) the p_{tr} most extreme scores in the upper tail are all replaced by the next lowest original score that was not replaced. Finally, (d) s^2_{Win} is calculated among the Winsorized scores using the standard formula for s^2 (Equation 2.3) except that squared deviations are taken from the **Winsorized mean** M_{Win}, the average of the Winsorized scores, which may not equal M_{tr} in the same sample. The statistic s^2_{Win} estimates the Winsorized population variance σ^2_{Win}, which may not equal σ^2 if the population distribution is nonnormal.

Suppose that $N = 10$ scores ranked from lowest to highest are as follows:

15 16 19 20 22 24 24 29 90 95

The mean and variance of these scores are $M = 35.40$ and $s^2 = 923.60$, both of which are affected by the extreme scores 90 and 95. The 20% trimmed mean is calculated by first deleting the lower and upper .20 (10) = 2 most extreme scores from each end of the distribution, represented next by the strikethrough characters:

~~15~~ ~~16~~ 19 20 22 24 24 29 ~~90~~ ~~95~~

Next, calculate the average based on the remaining 6 scores (i.e., 19–29). The result is $M_{tr} = 23.00$, which as expected is less than the sample mean, $M = 35.40$.

When one Winsorizes the scores for the same trimming proportion (.20), the two lowest scores in the original distribution (15, 16) are each replaced by the next highest score (19), and the two highest scores (90, 95) are each replaced by the next lowest score (29). The 20% Winsorized scores are listed next:

19 19 19 20 22 24 24 29 29 29

The Winsorized mean is $M_{Win} = 23.40$. The total sum of squared deviations of the Winsorized scores from the Winsorized mean is $SS_{Win} = 166.40$, and the degrees of freedom are 10 – 1, or 9. These results imply that the 20% Winsorized variance for this example is $s^2_{Win} = 166.40/9$, or 18.49. The variance of the original scores is greater (923.60), again as expected.

Robust Interval Estimation

The **Tukey–McLaughlin method** (Tukey & McLaughlin, 1963) to calculate robust confidence intervals for μ_{tr} based on trimmed means and Winsorized variances is described next. The standard error of M_{tr} is estimated in this method as

$$s_{TM} = \frac{s_{Win}}{(1 - 2p_{tr})\sqrt{n}} \tag{2.29}$$

For the example where

$$n = 10, p_{tr} = .20, s_{Win} = 18.49^{1/2} = 4.30, \text{ and } M_{tr} = 23.00$$

the standard error of the trimmed mean (23.00) is

$$s_{\text{TM}} = \frac{4.30}{[1-2(.20)]\sqrt{10}} = 2.266$$

The general form of a robust 100 (1 – α)% confidence interval for μ_{tr} in this method is

$$M_{\text{tr}} \pm s_{\text{TM}}\left[t_{2\text{-tail},\,\alpha}\left(n_{\text{tr}} - 1\right)\right] \tag{2.30}$$

where n_{tr} is the number of scores that remain after trimming. For the example where $n = 10$ and $p_{\text{tr}} = .20$, the number of deleted scores is 4, so $n_{\text{tr}} = 6$. The degrees of freedom are thus 6 – 1 = 5. The value of $t_{2\text{-tail},\,.05}$ (5) is 2.571, so the robust 95% confidence interval for μ_{tr} is

$$23.00 \pm 2.266(2.571)$$

which defines the interval [17.17, 28.83]. It is not surprising that this robust interval is narrower than the conventional 95% confidence interval for μ calculated with the original scores, which is [13.66, 57.14]. (You should verify this result.)

A robust estimator of the standard error for the difference between independent trimmed means when not assuming homoscedasticity is part of the **Yuen–Welch procedure** (e.g., Yuen, 1974). Error variance of each trimmed mean is estimated as

$$w_i = \frac{s^2_{\text{Win}_i}\left(n_i - 1\right)}{n_{\text{tr}_i}\left(n_{\text{tr}_i} - 1\right)} \tag{2.31}$$

where $s^2_{\text{Win}i}$, n_i, and n_{tr_i} are, respectively, the Winsorized variance, original group size, and effective group size after trimming in the *i*th group. The Yuen–Welch estimate for the standard error of M_{tr} may be somewhat more accurate than the estimate in the Tukey–McLaughlin method (Equation 2.29), but the two methods usually give similar values (Wilcox, 2012).

The Yuen–Welch standard error of $M_{\text{tr}1} - M_{\text{tr}2}$ is

$$s_{\text{YW}} = \sqrt{w_1 - w_2} \tag{2.32}$$

and the adjusted degrees of freedom in a central *t* distribution are estimated as

$$df_{\text{YW}} = \frac{\left(w_1 + w_2\right)^2}{\dfrac{w_1^2}{n_{\text{tr}1} - 1} + \dfrac{w_2^2}{n_{\text{tr}2} - 1}} \tag{2.33}$$

TABLE 2.4
Raw Scores With Outliers and Descriptive Statistics for Two Groups

	Group	
	1	2
	15	3
	16	2
	19	21
	20	18
	22	16
	24	16
	24	13
	28	19
	90	20
	95	82
M	35.40	21.00
M_{tr}	23.00	17.00
M_{Win}	23.40	16.80
s^2	923.600	503.778
s^2_{Win}	18.489	9.067

Note. The trimming proportion is $p_{\text{tr}} = .20$.

The general form of a 100 (1 – σ)% confidence interval for $\mu_{\text{tr}1} - \mu_{\text{tr}2}$ in this method is

$$M_{\text{tr}1} - M_{\text{tr}2} \pm s_{\text{YW}} \left[t_{2\text{-tail},\,\alpha} \left(df_{\text{YW}} \right) \right] \tag{2.34}$$

Listed in Table 2.4 are raw scores with outliers and descriptive statistics for two groups where $n = 10$. The trimming proportion is $p_{\text{tr}} = .20$, so $n_{\text{tr}} = 6$ in each group. Outliers in both groups inflate variances relative to their robust counterparts (e.g., $s_2^2 = 503.78$, $s^2_{\text{Win}2} = 9.07$). Extreme scores in group 2 (2, 3, 82) fall in both tails of the distribution, so nonrobust versus robust estimates of central tendency are more similar ($M_2 = 21.00$, $M_{\text{tr}2} = 17.00$) than in group 1. Exercise 5 asks you to verify the robust estimators for group 2 in Table 2.4.

Summarized next are robust descriptive statistics for the data in Table 2.4:

$$M_{\text{tr}1} = 23.00,\ s^2_{\text{Win}1} = 18.489 \quad \text{and} \quad M_{\text{tr}2} = 17.00,\ s^2_{\text{Win}2} = 9.067$$

$$M_{\text{tr}1} - M_{\text{tr}2} = 6.00$$

The standard error of the trimmed mean contrast is estimated in the Yuen–Welch method as

$$w_1 = \frac{18.489\,(9)}{6\,(5)} = 5.547 \quad \text{and} \quad w_2 = \frac{9.067\,(9)}{6\,(5)} = 2.720$$

$$s_{\text{YW}} = \sqrt{5.547 + 2.720} = 2.875$$

and the degrees of freedom are calculated as

$$df_{\text{YW}} = \frac{(5.547 + 2.720)^2}{\dfrac{5.547^2}{5} + \dfrac{2.720^2}{5}} = 8.953$$

The value of $t_{2\text{-tail, .05}}$ (8.953) is 2.264. The robust 95% confidence interval for $\mu_{\text{tr1}} - \mu_{\text{tr2}}$ is

$$6.00 \pm 2.875(2.264)$$

which defines the interval [–.51, 12.51]. Thus, $\mu_{\text{tr1}} - \mu_{\text{tr2}}$ could be as low as –.51 or it could be as high as 12.51, with 95% confidence and not assuming homoscedasticity. Wilcox (2012) described a robust version of the Welch procedure that is an alternative to the Yuen–Welch method, and Keselman, Algina, Lix, Wilcox, and Deering (2008) outlined robust methods for dependent samples.

A modern alternative in robust estimation to relying on formulas to estimate standard errors and degrees of freedom in central test distributions that assume normality is bootstrapping. There are methods to construct robust nonparametric bootstrapped confidence intervals that protect against repeated selection of outliers in the same generated sample (Salibián-Barrera & Zamar, 2002). Otherwise, bootstrapping is applied in basically the same way as described in the previous section but to generate empirical sampling distributions for robust estimators.

Standard computer programs for general statistical analyses, such as SPSS and SAS/STAT, have limited capabilities for robust estimation. Wilcox (2012) described add-on modules (packages) for conducting robust estimation in R, a free, open source computing environment for statistical analyses, data mining, and graphics.[10] It runs on Unix, Microsoft Windows, and Apple Macintosh families of operating systems. A basic R installation has about the same capabilities as some commercial statistical programs, but there are now over 2,000 packages that further extend its capabilities. Wilcox's (2012) WRS package has routines for robust estimation, outlier detection, comparisons, and confidence interval

[10] http://www.r-project.org/

construction in a variety of univariate or multivariate designs.[11] Additional R packages for robust estimation are available from the Institut universitaire de médecine sociale et préventive (IUMSP).[12] See Erceg-Hurn and Mirosevich (2008) for more information about robust estimation.

CONCLUSION

The basic logic of sampling and estimation was described in this chapter. Confidence intervals based on statistics with simple distributions rely on central test statistics, but statistics with complex distributions may follow noncentral distributions. Special software tools are typically needed for noncentrality interval estimation. The lower and upper bounds of a confidence interval set reasonable limits for the value of the corresponding parameter, but there is no guarantee that a specific confidence interval contains the parameter. Literal interpretation of the percentages associated with a confidence interval assumes random sampling and that all other sources of imprecision besides sampling error are nil. Interval estimates are better than point estimates because they are, as the astronomer Carl Sagan (1996, pp. 27–28) described them, "a quiet but insistent reminder that no knowledge is complete or perfect." Methods for robust interval estimation based on trimmed means and Winsorized variances were introduced. The next chapter deals with the logic and illogic of significance testing.

LEARN MORE

Cumming (2012) gives clear introductions to interval estimation, effect size estimation, and meta-analysis. Chernick (2008) describes bootstrapping methods for estimation, forecasting, and simulation. The accessible book by Wilcox (2003) gives more detail about robust statistics.

Chernick, M. R. (2008). *Bootstrap methods: A guide for practitioners and researchers* (2nd ed.). Hoboken, NJ: Wiley.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.

Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. New York, NY: Academic Press.

[11] http://dornsife.usc.edu/labs/rwilcox/software/

[12] http://www.iumsp.ch/Unites/us/Alfio/msp_programmes.htm

EXERCISES

1. Explain the difference between the standard deviation s and the standard error s_M.
2. Interpret $s = 60.00$ and $s_M = 6.00$ for the same data set. What is the sample size?
3. For $M = 100.00$, $s = 9.00$, and $N = 25$, show that the 99% confidence interval for μ is wider than the corresponding 95% interval.
4. For the data in Table 2.2, calculate the 95% confidence interval for μ_D and the 95% confidence interval for $\mu_1 - \mu_2$.
5. For the data in Table 2.4, verify the values of the robust estimators for group 2.
6. What is the relation between M_{tr} and M_{Win} in the Tukey–McLaughlin method?