PART I

Basic Experimental Design and Analysis
CHAPTER 1

Review of Basic Statistical Methods

1.1 INTRODUCTION

Statistical methods are often subsumed under the general headings “descriptive” and “inferential.” The emphasis in behavioral and medical science statistics books is frequently on the inferential rather than the descriptive aspects of statistics. This emphasis sometimes occurs because descriptive statistics such as means, variances, standard deviations, correlation coefficients, regression coefficients, and odds ratios can be described and explained in relatively few chapters. The foundations of inferential statistics generally require longer explanations, a higher level of abstract thinking, and more complex formulas. Associated with the large amount of space devoted to inference is a failure on the part of many professionals to appreciate that description is usually the most informative aspect of a statistical analysis.

A perusal of many current journals reveals that the overemphasis on inference (especially tests of significance) and the underemphasis on simple description is widespread; the inferential tail is frequently allowed to wag the descriptive dog. Descriptive statistics are not only underemphasized, they are sometimes completely ignored. One frequently encounters research outcomes reported in terms of only probability values or statements of “significant” or “nonsignificant” with no mention of the size of the difference (e.g., a difference between sample means) associated with the inferential results. The sources of this perversity go beyond statistical training; editorial policies of journals, demands of funding agencies, and criteria established by governmental agencies are also involved. An exploration of the development of this problem is interesting but tangential to this review; it will not be pursued here.

The remaining sections of this chapter begin with a review of conventional hypothesis testing (including elementary statistical decision theory) and interval estimation procedures for the simple randomized two-group experiment. Issues associated with standardized effect sizes, measures of association, generalization of results, and the control of nuisance variation are also presented.
1.2 ELEMENTARY STATISTICAL INFERENCE

Research workers generally employ inferential statistics in dealing with the problem of generalizing results based on sample data to the populations from which the subjects were selected. Suppose we (1) randomly select \( N \) mentally retarded patients from a Michigan institution, (2) randomly assign these patients to treatments 1 and 2, (3) apply the treatments, and (4) obtain an outcome measure \( Y \) on each patient. A useful descriptive measure of the differential effectiveness of the two treatments is the difference between the two sample means. This difference is an unbiased point estimate of the difference between the corresponding population means.

If it turns out that the sample mean difference is large enough to be of clinical or practical importance, the investigator may want to make a statement about the difference between the unknown population means \( \mu_1 \) and \( \mu_2 \). Population mean \( \mu_1 \) is the mean score that would have been obtained if treatment 1 had been administered to all mentally retarded patients in the whole institutional population. Population mean \( \mu_2 \) is the mean score that would have been obtained if instead the second treatment had been administered to the whole institutional population. Inferential tests and confidence intervals are widely used to evaluate whether there are sufficient sample data to state that there is a difference between unknown population means and (in the case of the confidence interval) to provide an interval within which it can be argued that the population mean difference will be found. A summary of hypothesis testing and confidence interval methods for the two-group design is presented in Table 1.1. An understanding of the conceptual foundation for these methods requires that several crucial distinctions be made among different types of measures and distributions; these are reviewed next.

Recall that the purpose of statistical inference is to make statements about unknown population parameters on the basis of sample statistics. Hence, the purpose of inferential methods is clear only if the distinction between statistics and parameters is understood. A statistic is a measure of a characteristic of a sample distribution. A sample is a subset of a population and a population refers to the entire collection of individuals (or some other unit) about which one wishes to make a statement. When the population size is finite, it may be feasible to obtain data on an outcome measure \( Y \) for all members of the population. In this case, summary measures that describe characteristics of the entire population distribution can be computed. Recall that Greek symbols are generally used to denote measures of population distribution characteristics; these measures are known as population parameters. For example, the population mean (\( \mu \)) and the population standard deviation (\( \sigma \)) describe the distribution characteristics known as central tendency and variability. Just as population parameters are used to describe characteristics of population distributions, sample statistics are used to describe characteristics of sample distributions. Statistics are generally denoted using Roman symbols. Examples of statistics include the sample mean \( \bar{Y} \) and the sample standard deviation \( s \). The distinction between statistics and parameters is crucial to understanding statistical inference.

A third type of distribution (in addition to sample and population distributions) provides the basis for both hypothesis tests and confidence intervals. It is known
Table 1.1 Summary of the Computation and Interpretation of the Independent Samples t-Test and 95% Confidence Interval for the Case of a Randomized Two-Group Experiment

A. Hypothesis Test

Null hypothesis: \( H_0: \mu_1 = \mu_2 \).

Test: Independent samples t (assume homogeneous population variances):

\[
\bar{Y}_1 - \bar{Y}_2 = \frac{\bar{Y}_1 - \bar{Y}_2}{s_{\bar{Y}_1 - \bar{Y}_2}} = t,
\]

where

\[
\bar{Y}_1 \text{ and } \bar{Y}_2 \text{ are the within-group sum of squared deviation (centered) scores for groups 1 and 2, respectively;}
\]

\[
n_1 \text{ and } n_2 \text{ are sample sizes associated with groups 1 and 2; and}
\]

\[
s_{\bar{Y}_1 - \bar{Y}_2} \text{ is the estimate of the standard error of the difference.}
\]

When the t-ratio is obtained from sample data it is called the obtained value of t (denoted as \( t_{obt} \)). After it is computed a decision rule is invoked in order to decide whether to reject the null hypothesis. Two forms of decision rules are described below; before either one can be used it is necessary to specify the level of \( \alpha \). Recall that the level of significance \( \alpha \) is set before the experiment is carried out. If a directional (one-tailed) test is desired, the level of \( \alpha \) may be denoted as \( \alpha_1 \). If a nondirectional (two-tailed) test is involved, then the level of alpha may be denoted as \( \alpha_2 \).

Decision rule for a nondirectional test using t. If the absolute value of \( t_{obt} \) is \( \geq t_{cv} \) (where \( t_{cv} \) is the critical value), then reject \( H_0 \); otherwise retain. The critical value of \( t \) is often based on \( \alpha_2 \) set at .05; the degrees of freedom = \( N - 2 \) (where \( N = n_1 + n_2 \)).

Decision rule for a nondirectional test using the p-value. Because most current statistics computer programs provide p-values in addition to obtained t-values, the modern approach is to simply inspect the p-value and compare it with the level of \( \alpha \) that has been specified. The decision rule for a nondirectional test is as follows: if the p-value is \( \leq \alpha_2 \), reject \( H_0 \); otherwise retain.

B. Confidence Interval

In the case of a simple two-group independent samples experiment the relevant 95% confidence interval is computed as follows:

\[
(\bar{Y}_1 - \bar{Y}_2) \pm [s_{\bar{Y}_1 - \bar{Y}_2}(t_{cv})] = (L, U),
\]

where

\[
s_{\bar{Y}_1 - \bar{Y}_2} \text{ is the estimate of the standard error of the difference; and}
\]

\[
t_{cv} \text{ is the critical value of the } t \text{ statistic based on } \alpha_2 = .05 \text{ and } N - 2 \text{ degrees of freedom, and}
\]

\[
L \text{ and } U \text{ are the lower and upper limits of the 95% confidence interval.}
as the sampling distribution (not to be confused with the sample distribution). A sampling distribution is a theoretical probability distribution of some statistic. We can conceptualize the sampling distribution of any type of statistic of interest. Examples that you have almost certainly encountered before are the sampling distribution of the mean and the sampling distribution of the difference between two independent sample means.

If we are interested in carrying out statistical inference regarding a mean, we need to know something about the sampling distribution of sample means. This distribution can be conceptualized as follows. Suppose (unrealistically) that we know the mean value ($\mu$) for a specified population; let us say it is 30. We select a sample size, say, $n = 10$. Next, we randomly sample 10 observations from the specified population, compute the mean of these observations, and record this value. Now conceptualize repeating these steps an infinite number of times. The whole distribution of this infinite number of sample means is the sampling distribution of the mean, where each sample mean is based on $n = 10$. We know that the mean of this distribution of sample means will be equal to the mean of the raw scores in the population. The standard deviation of the sampling distribution of the mean is called the standard error of the mean; it is given this name because it can be conceptualized as the standard deviation of the sampling error associated with the sample means included in the sampling distribution. The notation used to describe the standard error of the mean is $\sigma_{\bar{Y}}$. You may recall from elementary statistics that estimating the standard error of the mean is a simple task and that such an estimate (denoted as $s_{\bar{Y}}$) is required when statistical inference is applied to make statements about the unknown population mean in the one-group case.

Similarly, when inferential procedures are needed in the comparison of two sample means (e.g., in the case of a two-group independent sample experiment) it is necessary to have information regarding the sampling error associated with the difference between the two means. In this case the relevant sampling distribution is known as “the sampling distribution of the difference” (which is the shortened term for “the sampling distribution of the difference between two independent sample means”). This type of distribution can be conceptualized as follows.

Imagine two populations, where each one has a mean of 40 (i.e., $\mu_1 = \mu_2 = 40$). Select a sample size to be drawn from each population, say, $n_1 = 15$ and $n_2 = 15$. After 15 observations are randomly selected from the first population and 15 observations are randomly selected from the second population, the sample means $\bar{Y}_1$ and $\bar{Y}_2$ are computed. Next, the second sample mean is subtracted from the first and the difference is recorded. This process is then repeated (hypothetically) an infinite number of times. The distribution of the mean differences that result from this process constitutes the sampling distribution of the difference. The standard deviation of the sampling distribution of the difference is called the standard error of the difference (denoted $\sigma_{\bar{Y}_1 - \bar{Y}_2}$); an estimate of this parameter is required to carry out hypothesis tests (and confidence intervals) on the difference between two sample means. The essential conceptual issues associated with hypothesis testing are subsumed under the topic of statistical decision theory; this topic is briefly reviewed in the next section in the context of testing the equality of two population means ($\mu_1$ and $\mu_2$) using the independent samples $t$-test.
ELEMENTARY STATISTICAL DECISION THEORY

1.3 ELEMENTARY STATISTICAL DECISION THEORY

The concepts of type I error, type II error, and power are central to elementary statistical decision theory. Recall that the decision to reject the null hypothesis regarding the population means is made when (a) the obtained test statistic equals or exceeds the critical (i.e., tabled) value of the statistic or (b) the obtained p-value is equal to or less than the specified level of alpha. If the obtained test statistic is less than the critical value (or if the p-value exceeds alpha), the null hypothesis is retained. This decision strategy will not always lead to the correct decision.

**Type I error.** If the null hypothesis (i.e., \( H_0: \mu_1 = \mu_2 \)) is true in the case of a two-group experiment, the two population means are equal but we can anticipate that each sample mean will deviate somewhat from the corresponding population mean and that the difference between the sample means will deviate somewhat from zero. Although we expect the difference of an infinite number of sample mean differences (where each mean is based on a random sample) to equal the population mean difference (i.e., \( \mu_1 - \mu_2 \)), we anticipate that, in general, differences between sample means will differ somewhat from the difference between the corresponding population means as a result of sampling fluctuation.

This suggests that differences between two sample means will sometimes be large even though the two populations from which the samples were selected have exactly the same mean. We often employ significance tests to help decide whether an obtained difference between two sample means reasonably can be explained by sampling fluctuation (sometimes called sampling error). If the difference cannot reasonably be attributed to sampling error, we conclude that the population means are different. But if we conclude that the population means are different (i.e., we reject the null hypothesis) when the null hypothesis is true, a type I error is committed. That is, because of sampling error the difference between sample means will sometimes be large enough to result in the decision to reject the null hypothesis even though there is no difference between the corresponding (but unknown) population means. When this type of event occurs (i.e., rejecting the null hypothesis when it is true), we label it as a type I error. A priori (i.e., before the study is carried out) the probability of this happening is known as alpha (\( \alpha \)). The researcher can control \( \alpha \) by simply selecting the desired \( \alpha \) level in either (1) a table of critical values of the test statistic or in (2) appropriate statistical software. For example, if it is decided before the experiment is carried out that it is acceptable for the probability of type I error to be no higher than 5%, the test will be carried out using the critical value associated with \( \alpha = .05 \). Equivalently, the null hypothesis will be rejected if the observed p-value provided by statistical software is .05 or less.

**Type II error.** If the difference between sample means is not large enough to yield an obtained \( t \) that exceeds the critical value, the null hypothesis is retained. This is a decision error of the second kind if the null hypothesis is actually false. That is, a type II error is committed when the experimenter fails to reject a false null hypothesis. The probability of making a type II error is known as beta (\( \beta \)).

**Power.** The power of a statistical test refers to the probability that the test will reject the null hypothesis when the null hypothesis is false. Power is equal to \( 1 - \beta \) (i.e., one minus the probability of making a type II error). Unlike \( \alpha \), which is simply set by the
experimenter before the data are collected, the power of a designed experiment must be computed. In the context of an independent samples two-group \( t \)-test, power is a function of (1) the size of the population effect (e.g., the population mean difference), (2) the size of the random error variance (i.e., the within-population variance), (3) the sample size \( n \), and (4) the level specified for \( \alpha \). Power goes up with increases in population effect size, sample size, and \( \alpha \); it goes down as the error variance increases. Charts, tables, and software for estimating power are widely available.

Relationships among the concepts of type I error, type II error, and power are relevant to the appropriate interpretation of hypothesis tests. The issue of power is especially important in understanding how to interpret a result that is statistically nonsignificant. Because there is a widespread misunderstanding of the terms “statistically significant” and “nonsignificant,” the next section considers some interpretation issues associated with each of these outcomes.

**Interpretation of the Results of Hypothesis Tests and Confidence Intervals**

Although the computation of hypothesis tests and confidence intervals of the type shown in Table 1.1 is a trivial issue with the availability of modern statistical software, the interpretation of the outcome of these tests and confidence intervals requires a little thought.

**Statistically significant result.** When the null hypothesis is rejected it is appropriate to state that there is sufficient information to conclude that the population means are not equal; this is equivalent to stating that the difference between sample means is statistically significant. The finding of statistical significance simply means that there is a strong support for the conclusion that sampling error is not the only explanation for the observed sample difference. Often a difference that is too large to be explained only as sampling error (and is therefore statistically significant) will still be too small to be of any substantive interest.

Recall that a result is declared statistically significant when the \( p \)-value is small; the \( p \)-value is simply a statement of conditional probability. It refers to the probability of obtaining a sample mean difference at least as large as the one obtained, under the condition that there is no difference between the population means. There is no aspect of this probability statement that says the sample difference is large or important. Hence, there is no justification for claiming that a difference sufficiently large to yield a small \( p \)-value (say .05) is necessarily important (unless one is working in an area in which any deviation whatsoever from zero is important).

One determines whether a difference is of practical, clinical, or theoretical importance through knowledge of substantive considerations beyond the inferential test. Credible statements regarding the importance of statistically significant results must rest on knowledge of many nonstatistical aspects of the experiment; these aspects include the type of subjects, the nature of the independent variable, the properties of the dependent variable, previous results, and various contextual variables surrounding the investigation. In short, it is usually necessary to be immersed in the content area of the study in order to understand whether a result is important; a statistical test does not provide this context.
**Statistically nonsignificant result.** A nonsignificant result is not proof that there is no difference between the population means. Rather, a nonsignificant result should be interpreted to mean that there is insufficient information to reject the null hypothesis. There is a big difference between having proof that the null hypothesis is true and having insufficient evidence to reject it.

There are two major interpretations to keep in mind when a nonsignificant outcome is obtained. First, the nonsignificant result may be obtained because there is very little or no true effect whatsoever. Second, there may be an important true effect but the analysis may have such low power that the effect is not detected. The second interpretation should not be ignored, especially if no information is available regarding the power of the analysis. On the other hand, if (a) the difference between the sample means is trivial and (b) the power of the analysis for detecting small effects is known to be very high, then the second interpretation is less credible.

**Interpretation of confidence intervals.** Confidence intervals are frequently recommended in place of hypothesis tests as the basic method of statistical inference. Unfortunately, an inspection of many current journals will reveal that this recommendation is not often followed. Perhaps the most important advantage is that a confidence interval is reported in the metric of the dependent variable rather than in a derived measure such as $t$, $F$, or a $p$-value. Because hypothesis tests are so frequently misinterpreted (often because they are not accompanied by appropriate descriptive statistics), the arguments in favor of confidence intervals have substantial weight. A major reason for the misinterpretation of significance tests is a failure to distinguish between point estimation (e.g., the size of the difference between means) and the size of $p$-values. This confusion is a continuing issue in published research.

I recently reviewed a research paper that was carried out and reported by a well-qualified professional associated with a major school of medicine. Three treatment methods were compared and an inferential statistical analysis based on hypothesis testing was reported. The results section focused exclusively on the outcome of hypothesis tests on three comparisons. There was not a single graphic display or descriptive statistical measure to be found in either the text or the table that accompanied the results section. The only statistics reported were $p$-values. It was impossible to discover how large the effects were, although it was stated that the effects were large because the $p$-values were small. Obviously, the author thought that the presentation of small $p$-values was tantamount to a demonstration of large effects. Unfortunately, this confusion of small $p$-values with large or important treatment effects is very common. Some journals continue to publish articles that present inferential results unaccompanied by any descriptive measures whatsoever. One should never present inferential test results without the associated descriptive statistics.

If the inferential results in the medical study mentioned above had consisted of confidence intervals rather than hypothesis tests, the failure to provide any descriptive information would have been impossible. This is true because the confidence interval is constructed around the observed sample mean difference. An inspection of the upper and lower limits of the interval provides both descriptive and inferential information because (a) the point at the center of the interval corresponds to the mean
difference, and (b) the width of the interval provides inferential information regarding how much sampling error is present.

When a 95% confidence interval is presented, it is often stated that the unknown population mean difference is contained in the interval and that 95% confidence is attached to this statement. This interpretation needs a little clarification. The interpretation that we are 95% confident that the population mean difference is contained in the computed interval requires that we conceptualize a large number of experiments, not just the one experiment we are analyzing. If the experiments were replicated an infinite number of times, there would be an infinite number of confidence intervals and we would find that 95% of them would contain the population mean difference. Hence, the confidence coefficient (.95 in this example) refers to how confident we are that the one interval computed in this single experiment is one of those that contains the population difference. The values contained in the interval are generally viewed as credible ones to entertain as the true population mean difference; values outside the interval are not.

Hence, a confidence interval provides much more information than does \( t \) or \( p \) because simple inspection of the interval provides both descriptive and inferential information not provided by a \( t \)-value or \( p \)-value alone. Although the mean difference and the associated confidence interval supply much useful information, several supplementary approaches have recently become popular. An overview of two of these methods and some of the reasons they are currently being heavily promoted are presented next.

### 1.4 EFFECT SIZE

The major reason for performing a statistical analysis is to describe and effectively communicate the research outcome. If the description and communication are to be effective, the methods of analysis should be as transparent as possible. Although transparency should be the data analyst’s credo, this is not the impression conveyed in many published articles; too often obfuscation appears to be the rule.

Suppose a two-group experiment has been carried out, the data have been plotted, and the means and variances have been reported along with the mean difference on the original metric. Before any hypothesis tests or confidence intervals are considered, it is useful to first ask whether the obtained mean difference is of any practical or theoretical importance. It is often possible for a researcher to state values of differences between means (or some other appropriate descriptive statistic) that fall into one of the following categories: (1) trivial or unimportant, (2) of questionable importance, or (3) of definite practical or theoretical importance. The consideration of what constitutes an important result should affect the emphasis one attaches to inferential results.

If the obtained difference is judged to be unimportant, there is little justification for emphasizing the results of significance tests. If the obtained difference is not trivial, significance tests provide information that may be of interest. This does not mean that significance tests are invalid when applied to trivial differences. Rather, in most
EFFECT SIZE

cases it is unnecessary to ask whether a trivial difference is statistically significant (as is often the case with large sample sizes). There is a limited interest in learning that a trivial difference is too large to be explained as sampling error.

If a nontrivial difference is obtained, both the difference between means and the associated p-value are likely to be of interest. The failure to distinguish between point estimates of treatment effects and p-values associated with point estimates of treatment effects has led to decades of confusion regarding tests of significance. Because researchers so often imply that p-values provide information on the size and/or importance of the difference between means there have been recent attempts on the part of research methodologists and some journal editors to encourage researchers to provide more adequate descriptions of results. Consequently, there are now several journals in the behavioral and medical sciences that require the reporting of so-called effect sizes or measures of association.

These editorial policies can be traced back to the *Publication Manual of the American Psychological Association* (APA). It states, “... it is almost always necessary to include some index of effect size or strength of relationship ...” (2001, p. 25). Because researchers in many areas have become disenchanted with tests of significance the APA recommendation is likely to spread to other areas as well. This movement to focus on the so-called “effect size” is a step forward only if it is understood that such measures should be used to supplement rather than supplant conventional descriptive statistics.

Although the term “effect size” is now well established in the behavioral sciences, it is actually a misnomer. It refers to neither the most natural measure of the size of the effect (viz., the difference between two means) nor to the “effect” parameter, as it is defined in the analysis of variance structural model (i.e., the difference between the population mean for a specific treatment and the combined average of all population treatment means in the experiment). Consequently, the term is often confusing to both statisticians and researchers outside the behavioral sciences. This confusion between the term “effect size” and the actual size of the effect (i.e., the mean difference in the original metric) would be eliminated if precise language were employed. That which is currently called the effect size should be called the “standardized effect size” because it is defined as the following parameter:

\[
\frac{\mu_1 - \mu_2}{\sigma_w} = \delta.
\]

It can be seen that the difference between the two population means (the effect) is standardized by the common within-population standard deviation \(\sigma_w\). There are several ways to estimate this standardized effect size parameter. Cohen’s \(d\) is defined as

\[
\frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{\sum_{i=1}^{n_1} y_i^2}{n_1} + \frac{\sum_{i=1}^{n_2} y_i^2}{n_2}}} = d.
\]
A convenient formula that I prefer is known as Hedges’ $g$; it is defined as

$$g = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{n_1 \sum y_1^2 + n_2 \sum y_2^2}{n_1 + n_2 - 2}}},$$

A frequently recommended modification of this formula that provides slightly less biased estimates of the population effect size is

$$g(\frac{N - 3}{N - 2}) = \frac{g(N - 3)}{N - 2.25}.$$

Although the latter formula provides less biased estimates than are provided by $g$, the difference is very small in the case of reasonable sample size and it is certainly less intuitive. So I would not bother with the less biased estimator. Regardless of the formula used, the result is interpreted as the difference between means in standard deviation units.

In many cases the standardized effect size can be a useful supplement to the analysis, but in some contexts these measures cloud rather than clarify the outcome of an experiment. Clarity may be lost if the outcome measure is easily conceptualized and is well understood by both the researcher and the audience for the study.

Suppose a researcher is interested in the differential effect of two drugs on litter size in rats and that she has carried out a very large randomized-groups experiment (assume very little sampling error); the number of rats in each litter is used as the dependent variable. The mean number produced under drug I is 4.8 and the mean number under drug II is 8.8; the pooled within-group standard deviation is 2.5. Because the dependent variable is easy to conceptualize, the four-point difference between the means needs little additional explanation. The four-point difference would be more meaningful with the knowledge of typical litter sizes in untreated rats, but even without this additional information the outcome is understandable. It is obvious that litter size is about twice as large under drug II as it is under drug I. On the other hand, if the results are presented in terms of the standardized effect size (i.e., $g = 4/2.5 = 1.6$) instead of the mean difference, the statistical description is likely to be far less transparent for many readers in the audience. The original metric (number of animals in a litter) is easily understood whereas the standardized metric is not. If the research outcome is to be described in a publication intended for the general public, it is a mistake to present only standardized effect sizes. This is not to say that the standardized effect size should never be reported in professional journals. But in most cases results are more effectively communicated in the original metric than in standard deviation units. The typical researcher finds it easier to think (and communicate) in terms of the difference between means on the original metric than in terms of the difference between means in standard deviation units. But there is an exception to the general recommendation to focus on the original metric.
EFFECT SIZE

Although the mean difference on the original metric often communicates results more clearly than does the standardized effect size, this is not always so. There are many cases in which the outcome measure is new, unfamiliar, or difficult to conceptualize. In these situations it may be appropriate to focus the description of results on the standardized effect size rather than the difference between group means on the original metric. Hence, mean differences are usually more appropriate for studies that use well-understood dependent variables whereas standardized effect sizes may be preferable for reporting results from studies that use poorly understood outcome measures. The degree to which an outcome measure is understood depends upon both the audience for the study and the measurement conventions of the content area. Both of these should be considered when choosing the metric upon which to focus.

If the original metric seems to be a poor choice because it is poorly understood but the standardized effect size is also poorly understood by the intended audience, there is another choice. I recommend that the standardized effect size be computed and then transformed to a more understandable form. It is possible to make the information contained in the standardized effect size understandable to any audience through the use of appropriate graphics and/or simpler statistics (described below). This is especially true if one of the groups in a two-group experiment is a control. Consider the following example.

Suppose that a randomized-groups experiment is carried out to compare (1) a new method of treating depression with (2) a control condition; an established measure of depression is the dependent variable. If $g = -0.75$, the conventional interpretation is that the mean depression score for the treatment group is three-quarters of a within-group standard deviation below the control group mean. This interpretation will not be meaningful to many who read such a statement. But it is possible to transform the information contained in $g$ to a form that is much more understandable by the general reader. It can be determined (under the assumptions of normality and homogeneity of population variances) that 50% of the treated patients have depression scores below that of the average treated patient whereas only 23% of the control patients have depression scores that low. Hence, the treatment increases the percentage from 23% to 50%. The presentation of the results in terms of these percentages is far more meaningful for most readers than is the statement that $g = -0.75$. These percentages can be computed by any researcher who is familiar with the process of computing the area of the unit normal distribution below a specified standard score. In this example, we simply determine (using tables or software) the area in the unit normal distribution that falls below a $z$-score of $-0.75$. Alternatively, we could look up the area above a $z$-score of $0.75$. In this case, we can report that 50% of the control group has depression scores at or above the control group mean whereas only 23% of the treated group has depression scores this high. A graph illustrating the two hypothetical population distributions (treatment and control) and the associated percentages below the treatment-group mean further facilitates an understanding of the outcome. When the distributions are appropriately graphed the advantage of the treatment group over the control group is immediately apparent.

Attempts have been made to provide meaning to standardized effect sizes by labeling them as small, medium, or large (Cohen, 1988). The conventions for absolute
obtained standardized effect sizes are as follows: small $= .20$, medium $= .50$, and large $= .80$. Hence, the example $g$ presented in the previous paragraph would be classified as a medium standardized effect size. As is the case with most conventions of this type, they have taken on a life of their own; much is made of having a result that is “large” whereas “small” results may be denigrated. It should be kept in mind that size and importance is not necessarily the same thing. Small effects can be very important and large effects in some studies may be unimportant. Importance is dictated by substantive issues rather than by the value of $g$. Unfortunately, the current emphasis on standardized effect sizes has led some researchers to omit the original metric means in published articles. Both the standardized and unstandardized results should be reported even though a choice is made to focus on one or the other in the text describing the experimental outcome.

Perhaps the main reason for reporting standardized effect sizes is to provide an outcome measure for other researchers who are carrying out meta-analytic research. In these situations there is interest in characterizing the effects of certain treatments found in a large collection of studies, where each study provides comparative evidence on the same treatments. If all of the studies that use the same method of measuring a single-outcome construct, the task of summarizing the outcome of the studies is simple. But different studies often use different ways of operationalizing a single outcome construct. In this case a method of converting all the outcome measures to the same metric is desirable; the standardized effect size provides such a metric.

### 1.5 MEASURES OF ASSOCIATION

A second type of derived descriptive measure of outcome is sometimes called a measure of association; often these measures are described as a type of effect size. Like standardized effect sizes, measures of association are useful when the dependent variable is not a familiar measure. Unlike standardized effect sizes, measures of association are not interpreted as the size of the mean difference on some alternative metric. Rather, measures of association describe the proportion of the variation on the dependent variable that is explained by the independent variable. In a sense these measures describe the extent to which variation in the observed sample behavior is under experimental control.

Although there are several different measures of association, the intent of all of them is essentially the same. The one called the correlation ratio will be reviewed here. The population correlation ratio is denoted as $\eta^2$; the sample estimate of this parameter is denoted as $\hat{\eta}^2$. The computation of $\hat{\eta}^2$ is easily accomplished using the output of conventional statistical software that provides $t$ statistics (for the two-group case) or the analysis of variance summary table (for two or more groups). In the case of a two-group design, the sample correlation ratio can be defined as

$$\frac{t^2_{\text{obs}}}{t^2_{\text{obs}} + (N - 2)} = \hat{\eta}^2,$$
MEASURES OF ASSOCIATION

where
\[ t_{obt} \] is the obtained value of the conventional two-group independent samples
\[ t \]-ratio; and
\[ N \] is the total number of observations in the experiment (i.e., \( n_1 + n_2 \)).

A more general expression that applies to independent sample designs with two
or more groups is
\[ \frac{SS_B}{SS_T} = \hat{\eta}^2, \]
where
\[ SS_B \] is the between groups sum of squares (from a one-factor analysis of variance); and
\[ SS_T \] is the total sum of squares.

When two groups are involved, the \( t \) approach and the sum of squares approach
for computing the correlation ratio are equivalent. If more than two groups are
involved, the \( t \) approach is irrelevant and only the sum of squares approach is
appropriate. Regardless of the number of groups in the experiment, the interpretation
of the correlation ratio is the same. That is, it describes the proportion of the to-
tal variation in the experiment that appears to be explained by the independent
variable.

It was mentioned earlier that there is a correspondence between standardized effect
sizes and measures of association. In the case of two groups, it turns out that
\[ \hat{\eta} = \frac{g}{\sqrt{g^2 + 4 \left( \frac{N - 2}{N} \right)}}. \]

An advantage of reporting correlation ratios in addition to mean differences,
significance tests, and/or confidence intervals is that they describe the effectiveness
of the treatments in relative rather than absolute units. If the correlation ratio is 0.40, it
can be stated that approximately 40% of the total variability in the experiment appears
to be under experimental control and 60% of the total variability is due to the sources
other than treatment group differences. If all variability in the experiment is due to
treatments, the correlation ratio is 1.0; in this unrealistic case, the experimenter has
complete control of the outcome through manipulating the levels of the independent
variable. If there is no difference between the sample means, the correlation ratio
is zero; in this case no variability on the dependent variable is explained by the
independent variable.

The proportion of variability explained on the dependent variable by the inde-
pendent variable may be a more meaningful way to describe the outcome of the
experiment than is the mean difference on a variable that is completely unknown
to the reader of a research report. Like the standardized effect size, the correlation
ratio can be useful in comparing results of different studies, especially if the different studies have used different methods of operationalizing a common outcome construct. Also, as is the case for the standardized effect size, there are conventions regarding labels that are attached to correlation ratios of various sizes. The conventions some methodologists use to label the size of computed values of $\hat{\eta}^2$ are: small = .01, medium = .09, and large = .25; others recommend using .01, .06, and .15, respectively.

Caveats regarding conventions of this type were mentioned previously for the standardized effect size; they also apply to the correlation ratio. Recall that the size of an effect and the importance of an effect are not the same thing. A study demonstrating the effects of aspirin on heart attack in men (Steering Committee of the Physicians’ Health Study Research Group, 1988) is frequently cited as an example of data illustrating how elusive the notion of importance can be.

Over 22,000 male physicians were randomly assigned to receive either aspirin or a placebo every other day. After 5 years of treatment, the proportion of heart attacks experienced by the aspirin group was .0094; the placebo group proportion was .0171. The estimate of $\eta^2$ was .0011 and the $p$-value was < .000001. Hence, there was no doubt that aspirin therapy decreased the proportion of heart attacks. There was also no doubt that the proportion of variation on the outcome measure explained by the independent variable was very small. Note that the estimate of $\eta^2$ is far below the cutoff for declaring that there is even a “small” effect. In spite of these results regarding the small size of the effect expressed as a proportion, the Committee decided that the experiment should be terminated early because the evidence was so convincing. Aspirin therapy to prevent heart attack is now widely recommended for a large proportion of the male population.

But the advantage of the aspirin therapy may not be as impressive as has often been claimed in the popular press. Many people do not consider changing one’s probability of heart attack from .0171 to .0094 a massive effect. That is, the change in the probability of heart attack attributed to aspirin therapy was only about three-quarters of a percent. This is the reduction in what is known as absolute risk; it is the reduction that is relevant to the entire population of male physicians. But it is unlikely that this is the measure the Committee had in mind when the statement was made that the effect was large. Instead, it appears that the reason the study was terminated was more influenced by the $p$-value and the relative reduction in risk. The latter is determined by computing the difference between control and treatment groups on the proportion having a heart attack (.0171 − .0094) by the proportion in the control group having one (.0171); this yields a very impressive looking .45. The implication of this is that those in the control condition who will have a heart attack will be 45% less likely to have one if they take aspirin. The main point of this example is that there are different ways of expressing the size of an effect and these different measures are often very inconsistent. A single study can be described as having a trivial effect or a massive effect, depending on the measure chosen.

A computational example of the major statistics described in this chapter is summarized in Table 1.2. Statistics similar to the correlation ratio and standardized effect
size measures reviewed in this section are described in subsequent chapters for more complex designs and for analysis of covariance problems.

1.6 A PRACTICAL ALTERNATIVE TO EFFECT SIZES AND MEASURES OF ASSOCIATION THAT IS RELEVANT TO THE INDIVIDUAL: \( p(Y_{TX} > Y_{CONTROL}) \)

Many clinicians and patients find it less than satisfactory that research results are virtually always reported in terms of group measures and associated \( p \)-values. Indeed, one often hears the complaint that group comparisons are essentially irrelevant to an individual patient. This discontent generalizes to effect sizes and measures of association. When attempting to convey the meaning of a two-group trial to patients and others outside the research community, I recommend abandoning conventional group comparison statistics. Instead, provide the probability that a subject randomly selected from the treatment group has a higher outcome score (or a lower outcome score when the treatment is intended to lower the score) than does a subject randomly selected from the control group. This is not a conventionally computed value; I am aware of no standard software that provides it, but the computation is straightforward.

Consider the outcome of the two-group experiment on pain reduction described in Table 1.2. It can be seen that the mean pain for the treatment group is lower than the mean for the control group and that the standardized effect size is \(-1.15\). The approximate probability that a subject randomly selected from the treatment group has a pain score that is higher than the score of a subject selected at random from the control group is obtained by first computing:

\[
\frac{g}{\sqrt{2}} = \frac{-1.15}{\sqrt{2}} = -0.8131728 = z.
\]

The area below a \( z \)-score of \(-0.8131728\) can be found in a table of the normal distribution to be .21; this value approximates the probability that a subject selected at random from the treatment group has a pain score that is higher than the score of a subject selected at random from the control group. If, instead, we are interested in the probability that a subject selected at random from the treatment group has a pain score that is lower than the score of a subject selected at random from the control group, we compute the area above the computed \( z \); in this example that area is equal to .79. When the study is one in which the treatment is expected to increase the score on the dependent variable, it will usually be of greater interest to find the area below the \( z \)-score.

Note that the probability value provided using this approach is likely to answer the question a typical patient wants to be answered. It is natural to seek information of this type when contemplating a treatment procedure. Those who think in terms of odds rather than probability values can easily convert from one to the other. Browne (2010a, 2010b) does a nice job of illustrating both.
Table 1.2 Example of Point Estimation, Hypothesis Testing, Interval Estimation, Standardized Effect Size, Correlation Ratio, and $p(Y_{Tx} < Y_{Control})$ for a Randomized Two-Group Experiment

Example Data (A Subset of Pain Score Data Analyzed in Wang et al., 2010)

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
</tr>
</tbody>
</table>

A. Point Estimate of Treatment Effect

The point estimate of the population difference between treatment and control means on the pain scale is the sample mean difference: $(\bar{Y}_{Tx} - \bar{Y}_{Control}) = -2.125$.

B. Hypothesis Test (Independent Samples $t$)

Null hypothesis: $H_0: \mu_1 = \mu_2$.

Computed value of $t$: $t_{obt} = \frac{\bar{Y}_{1} - \bar{Y}_{2}}{s_{\bar{Y}_{1} - \bar{Y}_{2}}} = \frac{5.500 - 7.625}{.92461} = -2.298$.

Critical value of $t$ (for $\alpha = .05$) based on 14 degrees of freedom = 2.145.

Decision. Reject $H_0$ because $| -2.298 | > 2.145$ and state that the sample mean difference is statistically significant; infer that $\mu_1 \neq \mu_2$. Equivalently, reject $H_0$ because the $p$-value (from Minitab) = .037, which is less than the value (.05) set for $\alpha$.

Conditional probability interpretation. The $p$-value .037 is the probability of obtaining an absolute difference between the sample means of at least 2.125 points under the condition that there is no difference between the treatment and control population means (i.e., $\mu_1 = \mu_2$). Because this is an unlikely event under the stated condition, it is reasonable to argue that a better explanation for the sample mean difference is that the population means are different.

Substantive conclusion. There is sufficient evidence to conclude that the mean pain in the treated population is lower than the mean pain in the control population.

C. 95% Confidence Interval for the Difference ($\mu_1 - \mu_2$)

Computation.

$$(\bar{Y}_{1} - \bar{Y}_{2}) \pm [s_{\bar{Y}_{1} - \bar{Y}_{2}}(t_{.05})] = (L, U) = (\bar{Y}_{1} - \bar{Y}_{2}) \pm [s_{\bar{Y}_{1} - \bar{Y}_{2}}(t_{.05})] = (L, U)$$

$$= (-2.125) \pm [.9246(2.145)] = (-4.11, -1.142).$$

Interpretation. The population mean difference ($\mu_1 - \mu_2$) between the treated and control groups on the pain scale is contained in the interval $(-4.11, -1.142)$; the confidence coefficient associated with this statement is .95.
GENERALIZATION OF RESULTS

Table 1.2 (Continued)

D. Standardized Effect Size $g$

Computation.

$$\frac{\bar{Y}_1 - \bar{Y}_2}{s_w} = \frac{5.500 - 7.625}{1.849} = -1.15 = g.$$  

Interpretation. The treatment group mean pain score is $-1.15$ standard deviation units below the mean pain score of the control group. In addition, because the area in the unit normal distribution below a standard score of $-1.15$ is .1251, it is estimated that approximately 13% of the control population has pain scores falling below the average pain score in the treatment population whereas 50% of those in the treatment population have scores that low.

E. Correlation Ratio

Computation.

$$\frac{t^2_{obt}}{t^2_{obt} + (N - 2)} = \frac{-2.298^2}{-2.298^2 + (N - 2)} = \frac{5.28}{5.28 + 14} = .27 = \hat{\eta}^2.$$  

Interpretation. Twenty-seven percent of the total variation among subjects on the pain score appears to be explained by the independent variable.

F. Probability That a Randomly Selected Treated Subject Has Lower Pain Than Does a Randomly Selected Control Subject

$$\hat{\eta}^2 = -1.15 = -.8131728 = z; \text{area below } z \text{ is .21. This is } p(\text{Y}_{\text{Tx}} > \text{Y}_{\text{Control}}), \text{but, because the } \text{Y}_{\text{Tx}} \text{ is designed to reduce pain, the interest is in } p(\text{Y}_{\text{Tx}} < \text{Y}_{\text{Control}}), \text{which is } (1 - .21) = .79.$$  

1.7 GENERALIZATION OF RESULTS

A great deal of methodological literature is devoted to the topic of generalizing experimental results (see, e.g., Shadish et al., 2002). Issues such as the methods used to operationalize the independent and dependent variables, the nature of the experimental setting, and the method of selecting subjects are relevant to this topic. Subject selection is a major statistical concern; it plays a major role in defining the population to which the results can be generalized. The selection procedures associated with two versions of the one-factor randomized-groups experimental design are described below.

Case I: Ideal Randomized-Groups Experiment: Random Selection and Random Assignment. An ideal experimental design involves both (a) random selection of $N$ subjects from a defined population and (b) random assignment of the selected subjects to treatment conditions. The results of the ideal experiment (if properly executed) can be generalized to the population from which the subjects were randomly selected.
Case II: Typical Randomized-Groups Experiment: Accessible Selection and Random Assignment. If the subjects for the experiment are not randomly selected from a defined population but are simply a collection of people accessible to the experimenter, it is still appropriate to randomly assign these subjects to treatment conditions and to apply conventional inferential procedures. Although the tests and/or confidence intervals are appropriate, the generalization of results is more ambiguous than with Case I selection. If little is known about the characteristics of the selected subjects, then little can be stated about the population to which the experimental results generalize. The experimenter can only state that the results can be generalized to a population of subjects who have characteristics similar to those who were included in the study. This may not be saying much if little is known about the accessible subjects. Generalization in Case I situations is based on principles of statistical inference; Case II generalization is based on logical considerations and speculation concerning the extent to which the accessible subjects are similar to those in the population to which the experimenter hopes to generalize. Although Case I selection is desirable from the point of view of generalizing results, such selection is frequently impractical; Case II selection is the rule rather than the exception in many areas of behavioral and medical science research.

Knowledge regarding the generality of results from experiments based on Case II selection can be improved by collecting thorough information regarding the characteristics of the accessible subjects. The form of this information may range from general measures of demographic characteristics to very specific medical diagnostic indices. After this information is collected, it is usually reported along with treatment results in order to characterize the type of subjects to which the estimated treatment results apply. The role of the information in this situation is to help provide a context for interpreting the results after the experiment is finished. But there are two other important roles for information regarding subject characteristics.

Subject information can play a role in both the design of the experiment and in the statistical analysis. If the information is appropriately used in the design or analysis, there are dual advantages of doing so. First, the power will be increased relative to what it would have been otherwise; second, the generality of the results may be more clearly defined. The relevance of subject information to these two issues is considered in the next section.

1.8 CONTROL OF NUISANCE VARIATION

The term “nuisance variable” is often applied to variables that are believed to affect scores on the dependent variable but are of no experimental interest. This is an important issue in experimental design because the size of the error term used in computing hypothesis tests and confidence intervals is a function of the amount of nuisance variation. Hence, the power of statistical tests and the width of confidence intervals is also a function of the amount of nuisance variation. The use of many
complex experimental designs and advanced statistical analyses is motivated by the desire to control nuisance variation.

Suppose we are interested in the effects of two methods of teaching reading. Students from a single classroom are randomly assigned to one of the two treatment conditions. If the students in the classroom are very heterogeneous with respect to reading ability (a subject characteristic), we can expect much within-group variation on measures of reading proficiency. The large within-group variation will be reflected in a large estimate for the standard error of the difference. Consequently, the large estimate for the standard error of the difference will lead to a small $t$-ratio and it will be necessary to conclude that the inferential evidence for an effect is insufficient to reject the null hypothesis. Even if the outcome of the experiment reveals a relatively large difference between the dependent variable means, the $t$-ratio may be too small to be declared statistically significant; correspondingly, the confidence interval on the mean difference is likely to be very wide.

Some method of controlling (at least partially) this nuisance variable should be employed to increase the power of the analysis and to reduce the width of the confidence interval. Several methods available for this purpose are described below.

**Select subjects who are homogeneous with respect to the nuisance variable.** One method of controlling nuisance variability involves (a) collecting data on the nuisance variable from a large collection of potential subjects, and (b) selecting only those subjects from the collection who have the same score (or very similar scores) on a measure of the nuisance variable. Obviously, if subjects included in the experiment have been selected in such a way that they all have the same preliminary reading score, the influence of reading skill as a source of within-group variation on the dependent variable has been reduced. If an acceptable number of subjects having the same nuisance variable score can be found, these subjects are randomly assigned and the experiment is carried out as a conventional randomized-groups experiment. The major problem with this method is that it is often very difficult (or expensive) to find a sufficient number of subjects with the same score on the nuisance measure.

**Use the nuisance variable as one of the factors in a two-factor design.** A frequently used method of dealing with nuisance variability is to employ the nuisance variable to form levels of one of the factors in a two-factor design. Hence, one factor would consist of levels of the nuisance factor and the other factor would consist of levels of the treatment factor.

**Use blocking or matching.** Matched pair designs and randomized block designs are effective design strategies for increasing power relative to what it would be using a randomized-groups design. Measures on a nuisance variable are used as the matching or blocking variable with these designs.

**Use a repeated measures design.** In many cases, power can be greatly increased if each subject is exposed to each treatment in a repeated measures design. This is an excellent strategy if there is little chance of carryover effects. Unfortunately, there are many cases in which this is not true.

**Use statistical control of nuisance variation: The analysis of covariance.** Each of the approaches mentioned above is a design strategy for contending with nuisance variation...
variation in true experiments. The appropriate analysis of data from these designs is likely to have higher power than is associated with the conventional analysis of the typical randomized-groups design. But these alternative designs are not the only path to higher power.

An alternative strategy for contending with nuisance variation (and therefore increasing power) is to use a form of statistical control known as the analysis of covariance (ANCOVA). Unlike the strategies mentioned above, ANCOVA requires no departure from the conventional randomized-groups design with respect to subject selection and assignment. This approach has certain practical and statistical advantages over the alternative procedures mentioned above. Because these advantages are best revealed through direct comparison with conventional alternative methods, a review of matched pair, randomized block, and repeated measures designs is presented in the next two chapters.

1.9 SOFTWARE

Minitab input and output for the analysis of the pain data example are presented below.

As soon as Minitab is opened follow this path: Menu bar > Editor > Enable Commands. The Minitab prompt will then appear. Type the dependent variable scores in c1 and the associated group number in column c2. Then follow the commands shown below Figures 1.1 and 1.2.

![Dotplot of pain data](image-url)
Figure 1.2 Individual value plot of pain data.

MTB > print c1 c2

Data Display
Row  DV  Group
  1   6    1
  2   7    1
  3   3    1
  4   7    1
  5   7    1
  6  10    2
  7   4    1
  8   6    1
  9   5    2
 10  10    2
 11   7    2
 12  10    2
 13   6    2
 14   7    2
 15   6    2
 16  10    2

MTB > Dotplot (‘DV’) * ‘Group’.

MTB > TwoT ‘DV’ ‘Group’;
SUBC > Pooled;
SUBC > GIndPlot.
Two-Sample T-Test and CI: DV, Group

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>5.50</td>
<td>1.60</td>
<td>0.57</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>7.63</td>
<td>2.07</td>
<td>0.73</td>
</tr>
</tbody>
</table>

Difference = mu (1) - mu (2)
Estimate for difference: -2.125

95% CI for difference: (-4.108, -0.142)

T-Test of difference = 0 (vs not =): T-Value = -2.30
P-Value = 0.037 DF = 14
Both use Pooled StDev = 1.8492

1.10 SUMMARY

Both descriptive and inferential statistics are relevant to the task of describing the outcome of a research study and to generalizing the outcome to the population from which the subjects were selected. The major emphasis in evaluating data should be on descriptive statistics, beginning with plots of the original data. Description in the form of easily understood summary statistics such as means, mean differences, and variances (in the original metric) is also essential. Standardized effect sizes and correlation ratios are sometimes helpful additional descriptors. Statistical inference is likely to be of interest when interpreting results, but it should be pursued only after thorough description. Although the reporting of results in terms of statistical significance is conventional practice, confidence intervals are usually more informative. The essentials of elementary statistical decision theory include the concepts of type I error, type II error, and power. Among the goals of a well-designed experiment are low probability of type I error ($\alpha$), low probability of type II error ($\beta$), and high power ($1 - \beta$). More important than these goals is the extent to which the experimental design provides clear results. Random assignment plays an important role in both providing unconfounded estimates of treatment effects and in justifying statistical inference. Complex experimental designs are frequently used to provide higher power than is associated with simple randomized-groups designs. An alternative approach for increasing power is known as the analysis of covariance.