1 Comparing Two Means

Introduction

Analysis of variance (ANOVA) is the standard method used to generate confident statistical inferences about systematic differences between means of normally distributed outcome measures in randomized experiments. In order to provide a context for what follows, we will begin by considering briefly what is implied by the previous sentence.

First, the reference to *randomized* experiments suggests that ANOVA is particularly appropriate for the analysis of data from experiments in which each subject (participant, experimental unit) is randomly assigned to one of two or more different treatment conditions (experimental groups). All subjects in a particular treatment condition receive the same treatment, and differences in the effects of the treatments are expected to produce differences between groups in post-treatment scores on a relevant outcome measure. Random assignment of subjects to experimental conditions usually ensures that systematic differences between means of groups given different treatments can be attributed to differences in the effects of the treatments, rather than to any other explanation such as the presence of pre-existing differences between the groups. If nothing goes wrong in the conduct of a randomized experiment, systematic differences between group means on a dependent variable can quite properly be attributed to differences in the effects of the treatments, and ANOVA procedures are specifically designed to produce inferences about differential treatment effects.

ANOVA procedures are not well suited for the analysis of data from quasi-experiments, where different treatments are given to pre-existing groups, such as different classes in a school. It is usually preferable to analyse quasi-experimental data by methods that attempt to take pre-existing differences into account (Reichardt, 1979).

The statement that ANOVA (or at least fixed-effects ANOVA, by far the most widely used version) is concerned with the pattern of differences between *means* implies that, despite the name of the procedure, the focus is on means rather than variances.[1] When the outcome of an experiment is not well summarized by a set of means, as is the case when the dependent variable is categorical rather than continuous, ANOVA is not appropriate.

The theory underlying inferential procedures in ANOVA assumes (among other things) that the distributions to be summarized in terms of means are *normal* distributions. ANOVA procedures are often used, sometimes with very little justification, when outcome measures are not even approximately normally distributed.

We should also consider what is meant by *systematic* differences between means on an outcome measure. According to the simplest ANOVA model, the difference between two sample means (the means calculated from the data produced by subjects given two different treatments in a randomized experiment) is influenced by two independent components, one of which is systematic, the other random. The systematic component is the difference between the effects of the two treatments; in a simple randomized experiment this difference between effects can be thought of as a difference between two *population* means, one for each treatment. The experimenter's problem, of course, is that the observed difference between sample means can also be influenced by a number of other unknown factors, some of which are associated with the particular subjects who happen to be assigned to a particular treatment, and some of which may be regarded as factors contributing to measurement error. In a randomized experiment these unknown factors contribute to the difference between sample means in a random or nonsystematic way. ANOVA methods allow for the influence of random as well as systematic influences on sample means.

Finally, it is necessary to consider what is meant by *confident* statistical inference. Suppose that the dependent variable in a two-group randomized experiment is a 40 item ability test, and the sample means on this test are $M_1 = 25.76$ and $M_2 = 19.03$, so that the difference between sample means is $M_1 - M_2 = 6.73$. Given this difference, which is specific to this particular sample (or pair of samples), what can we infer about the difference between the effects of the two treatments (that is, the difference between population means $\mu_1 - \mu_2$)? Because $M_1 - M_2$ is an unbiased estimator of $\mu_1 - \mu_2$ it might seem reasonable to infer that $\mu_1 - \mu_2 = 6.73$. This is indeed the best *point* estimate of the value of $\mu_1 - \mu_2$. The problem with a point estimate of a difference between two population means is that it is almost certain to be different from the actual value of the parameter being estimated, so that we are almost certain to be wrong if we assert that $\mu_1 - \mu_2 = 6.73$. Imagine *replicating* the experiment (repeating the experiment with different subjects) a very large number of times.[2] Values of $M_1 - M_2$ would vary across replications, almost never being exactly equal (given an unlimited number of decimal places) to the unknown value of $\mu_1 - \mu_2$. Because we cannot be confident in any sense that this (or any other) point estimation procedure is likely to produce a correct inference, point estimation is not a confident inference procedure. An *interval estimate* of the value of $\mu_1 - \mu_2$, however, does allow for the possibility of confident inference.

Suppose that the 95% confidence interval (CI) on $\mu_1 - \mu_2$ turns out to have a lower limit of 4.39 and an upper limit of 9.07, and as a consequence we assert that the value of $\mu_1 - \mu_2$ is somewhere between these limits. We cannot be sure that this particular inference is correct, but we know that the CI procedure controls the probability (at .95) that inferences of this kind are correct, meaning that 95% of the CIs constructed in this way (one interval per replication) would include the value of $\mu_1 - \mu_2$. We can express the same thing in terms of an *error rate* by saying that the probability of being in error when asserting that the interval includes $\mu_1 - \mu_2$ is .05. (This means that 5% of assertions of this kind, one per replication, are incorrect.) The difference between confident statistical inference and point estimation is that the former allows the researcher to control the probability of inferential error.

ANOVA provides for confident inference on differences between means via CIs and statistical significance tests. As we will see, inference from CIs is more informative than inference from tests.

Organization of this book

This book is intended for readers who have completed at least one course in statistics and data analysis including a reasonably substantial treatment of statistical inference. If you feel the need to revise some of the basic ideas underlying statistical inference before dealing with the material in this book, you will find that the accounts given by Lockhart (1998) and Smithson (2000) are compatible with the approach taken here.

Chapter 1 applies a hierarchy of levels of inference to the problem of producing and justifying a confident inference on a single comparison between two means. If your introduction to statistical inference emphasized statistical hypothesis testing you may be surprised to discover that CI inference occupies the highest level in this hierarchy.

Chapter 2 deals with the application of the ANOVA model to data from single-factor between-subject designs. CI inference on individual contrasts (generalized comparisons) is emphasized, and the hierarchy of inference levels is used to show how this approach is related to traditional approaches emphasizing test-based heterogeneity and directional inference.

In Chapter 3 we consider methods of controlling the precision of CI inferences on contrasts. The relationship between precision of interval estimates on contrasts and the analogous concept of the power of significance tests on contrasts is discussed.

Chapter 4 deals with simple between-subjects factorial designs in which the effects of varying the levels of at least two different experimental factors are examined within a single experiment. In this chapter we examine a number of different approaches to the problem of producing coherent inferences on contrasts on three types of factorial effects defined by ANOVA models: simple effects, main effects and interaction effects. Analyses of complex between-subjects factorial designs are considered in Chapter 5.

In Chapter 6 we consider within-subjects (repeated measures) designs, where each subject in a single group is subjected to all of the treatments examined in an experiment, or where each subject is examined on a number of trials or measurement occasions. ANOVA models for within-subjects designs must somehow deal with the fact that repeated measurements on a single individual are not independent of one another, so these models, and the data-analytic procedures that make use of them, differ in important ways from the models and methods discussed in earlier chapters.

Chapter 7 deals with mixed factorial designs: designs with at least one between-subjects factor and at least one within-subjects factor.

Many of the analyses recommended in this book are not currently supported by most of the popular statistical packages. You can carry out most of these analyses with a program called *PSY* (Bird, Hadzi-Pavlovic and Isaac, 2000). See Appendix A for a general overview of the program and the website from which it can be downloaded. *PSY* does not carry out some of the more traditional analyses based on significance tests that are supported by statistical packages, and it does not carry out analyses based on various extensions of ANOVA models. For these reasons (among others), Appendix B provides *SPSS* syntax required to carry out various analyses with *SPSS*, probably the most popular of all statistical packages. Finally, some of the more advanced analyses are most easily carried out with the *STATISTICA Power Analysis* program (Steiger, 1999). The use of this program is discussed where appropriate.

The data sets used for most of the examples and exercises can be downloaded from the Sage website at

http://www.sagepub.co.uk/resources/bird.htm

Every input file mentioned in this book can be downloaded from that website and can be opened by *PSY*.

Confident inference on a single comparison

Many of the basic ideas concerning inference in ANOVA can be developed in the context of a two-group randomized experiment, where the experimenter wishes to use the data to produce a confident inference on a single comparison between two means. In this chapter we will examine in some detail the logic of confident inference in the two-group case. Suppose that an experimenter wishes to examine the effect of a treatment (an experimental manipulation of some kind) by comparing the mean score of treated subjects on a relevant dependent variable (outcome measure) with the mean score obtained by a different group of subjects who did not receive the treatment (a control group). N subjects are randomly assigned to two groups (with $n = N/2$ subjects per group), and steps are taken to ensure that the problems that can sometimes arise in randomized experiments (Cook and Campbell, 1979, Shadish, Cook and Campbell, 2002) do not arise in this one.

The parameter of most interest to the experimenter is $\mu_1 - \mu_2$, the difference between the mean of all potential subjects who receive the treatment and the mean of all potential subjects in the control condition. This difference between the means of two hypothetical populations is unknown, but can be estimated from $M_1 - M_2$, the difference between the means of subjects in the two groups. In a randomized experiment the *expected value* of $M_1 - M_2$ is $\mu_1 - \mu_2$. This means that if the experiment were replicated an infinite number of times and each replication produced a value of $M_1 - M_2$, then the mean of the distribution of $M_1 - M_2$ values would be $\mu_1 - \mu_2$. $M_1 - M_2$ is therefore an *unbiased* estimator of $\mu_1 - \mu_2$.

Absence of bias is a desirable property of an estimator, but it does not imply that the estimate of an effect size obtained from a single experiment will be so close to the actual effect size that the discrepancy between the two can be safely ignored. A confident inference about the value of $\mu_1 - \mu_2$ can be obtained from an analysis of variance, but in the two-group case the same inference can be obtained from the familiar two-group t test or from a t-based CI. It will be convenient to deal with a number of issues in this familiar context before we discuss the more general ANOVA model. We need to remember that the standard two-group t (test or CI) procedure is based on the assumption that both populations of dependent variable values have a normal distribution, and that the two within-population standard deviations are identical.[3]

Strength of inference on a comparison

Following Hsu (1996), we will distinguish between three levels of confident inference on a comparison between two means: CI inference, confident direction inference and confident inequality inference. These three levels are ordered in terms of strength of inference, CI inference being stronger than the other two.

Confidence interval inference In Hsu's terminology, which we will adopt here, a CI inference on the comparison $\mu_1 - \mu_2$ can be expressed as

$$\mu_1 - \mu_2 \in (ll, ul)$$

where the symbol '$\in$' means 'is included in' or 'is covered by',

ll is the lower limit of the interval,

and ul is the upper limit of the interval.

Thus the inference $\mu_1 - \mu_2 \in$ (10.1, 12.3) asserts that

$$10.1 < (\mu_1 - \mu_2) < 12.3,$$

that is, that μ_1 is greater than μ_2 by at least 10.1 units, but by no more than 12.3 units. The CI (as distinct from the inference derived from it) is (10.1, 12.3).

If the CI covers the parameter (that is, if $\mu_1 - \mu_2$ is in fact somewhere between the upper and lower limits), then the inference is correct, and no error has been made. If the interval does not cover the parameter (that is, if $\mu_1 - \mu_2$ is lower than the interval's lower limit or higher than the upper limit), then the inference is false. We will use the term *noncoverage error* to refer to this type of inferential error. Given the assumptions required to justify the CI procedure, the *noncoverage error rate* α (the probability of a noncoverage error) can be controlled at a nominated low level, usually set at .05 or .10. It is customary to specify the noncoverage error rate indirectly by setting the *confidence level*, defined as $100(1 - \alpha)\%$, at a high level. If the confidence level is set at 95%, then the noncoverage error rate is $\alpha = .05$.

It is important to understand what the confidence level and the noncoverage error rate mean, and also what they do not mean. Imagine that a two-group experiment is replicated a very large number of times. Each replication produces a 95% CI on $\mu_1 - \mu_2$. Because sample means and other statistics vary across replications, CI limits and the inferences following from them will also vary across replications. Provided that the relevant assumptions are satisfied, 95% of these repetitions of the experiment will produce a CI covering the population mean difference, thereby producing a correct inference. The (unknown) value of the parameter of interest $(\mu_1 - \mu_2)$ does not vary across these replications. The parameter is fixed, but the CIs vary across replications.

The language sometimes used to describe CIs ('the probability that the parameter lies inside the interval is …') can encourage incorrect interpretations implying that the parameter is a variable rather than a constant. The important point is that the probability statement refers to the relative frequency with which variable intervals include (or exclude) the fixed parameter, given an indefinitely large number of replications of the experiment.

Confident direction inference A directional inference specifies the direction of the difference between means, but nothing more than that. Directional inference on the comparison $\mu_1 - \mu_2$ is an assertion of the form $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$.

CI inference implies directional inference, provided that the CI excludes the value zero so that all values in the interval have the same sign. Thus the CI inference $\mu_1 - \mu_2 \in$ (10.1, 12.3) implies that $\mu_1 > \mu_2$.

In practice, a directional inference usually follows from the rejection of a null hypothesis such as H_0: $\mu_1 - \mu_2 = 0$ by a statistical test. Many test procedures designed to test hypotheses about differences between means (including the two-group *t* test) are associated with (and derivable from) a CI procedure. If a 95% CI constructed with the *t* procedure (the standard procedure in the two-group case) turns out to be (10.1, 12.3), justifying the directional inference $\mu_1 > \mu_2$, then the associated .05-level two-tailed *t* test will necessarily reject the null hypothesis $\mu_1 - \mu_2 = 0$. Rejection of this hypothesis implies the inequality inference $\mu_1 \neq \mu_2$, but does not by itself imply anything about the direction of the difference between means. The justification for directional inference in this case depends on the relationship between the CI and the test: the test is able to reject the null hypothesis when $M_1 > M_2$ if and only if the associated CI justifies the directional inference $\mu_1 > \mu_2$. Similarly, a statistically significant outcome when $M_1 < M_2$ implies that the associated CI includes only negative values [such as (–8.6, –2.3)], thereby justifying the directional inference $\mu_1 < \mu_2$.

We may note in passing that a two-sided CI (with both an upper and lower limit) implies the outcome of a two-tailed test, but it does not imply the outcome of a one-tailed test. One-tailed tests are associated with single-sided CIs (see Hsu, 1996), which are rarely used in psychological research. One disadvantage of single-sided CIs is that they provide no information about precision of estimation.

Erroneous directional inferences can occur under two conditions: first, if a directional inference is made when no difference exists; second, if a difference exists in the direction opposite to that asserted. If $\mu_1 - \mu_2 = 0$ and it is asserted that $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$, then a Type I error has been made. If $\mu_1 - \mu_2 = 0$, then the Type I error rate from an α-level two-tailed test procedure is equal to the noncoverage error rate for confidence interval inference. The Type I error rate is hypothetical, in the sense that it refers to errors that cannot occur unless $\mu_1 = \mu_2$. If $\mu_1 > \mu_2$ and it is asserted that $\mu_1 < \mu_2$, or if $\mu_1 > \mu_2$ and it is asserted that $\mu_1 < \mu_2$, then a Type III error has been made. The Type III error rate cannot exceed $\alpha/2$, which it approaches for very small values of $|\mu_1 - \mu_2|$.

Confident inequality inference Inequality inference occurs when a data analyst asserts simply that $\mu_1 \neq \mu_2$, without specifying the direction of the difference between means. This is a particularly weak form of inference, because it denies only that the two means are absolutely identical. CI inference implies confident inequality inference if zero is excluded from the CI: the inference $\mu_1 - \mu_2 \in$ (10.1, 12.3) implies that $\mu_1 \neq \mu_2$. Confident direction inference also implies confident inequality inference, because the assertions $\mu_1 > \mu_2$ and $\mu_1 < \mu_2$ both imply that $\mu_1 \neq \mu_2$. A Type I error is the only type of error possible when inequality is asserted.

Equality inference? Following Hsu (1996), we do not consider Type II errors when defining error rates for directional inference or inequality inference. A Type II error (as defined in most discussions of significance tests) would occur if it were incorrectly asserted that $\mu_1 = \mu_2$. The approach we are considering makes no provision for such an inference. Indeed, a number of statisticians and data analysts argue that there is always some difference, perhaps an extremely small difference, between any two population means (Cohen, 1994, Schmidt, 1996). It is important, however, to allow for the possibility of inferring that two population means might be *practically equivalent*, meaning that the difference between them is in some sense trivially small. We will discuss practical equivalence inference after we consider effect size.

Interpreting effect size

CI inference is stronger (more informative) than confident direction or confident inequality inference, and is therefore emphasized in this book. Traditionally, however, the benefits of CI inference have been largely ignored in practice, and researchers have generally relied on significance tests that provide relatively weak inference, and no inference at all about the magnitude of the effects under investigation. Since the early 1960s, this traditional approach to inference has been severely criticized (see Harlow, Mulaik and Steiger, 1997, for a comprehensive review). No doubt the conservative approach taken by many textbook authors, journal editors and software producers has played a role in reinforcing the much criticized traditional approach. Researchers often respond to an obvious need to say something about the magnitude of an experimental effect by relying on a significance test to justify an assertion about the existence of the effect (via confident direction or confident inequality inference), then discussing the magnitude of the observed difference between means as though it is not subject to sampling error. As a consequence, a confident inference may (or may not) be made about the existence or direction of an effect, but no attempt is made to produce a confident inference about the magnitude of the effect. A much better approach is to incorporate the intention to make statements about effect size into the inferential analysis from the outset.

Before the magnitude of an experimental effect can be profitably estimated, it must be defined in a way that makes sense to the researcher. Although various approaches to effect size measurement have been proposed, a difference between two relevant population means has distinct advantages over alternative approaches, provided that the difference is expressed in an informative metric.

Dependent variable units The most obvious metric for an effect size is the metric used to scale the dependent variable. In well-developed areas of research, experimenters often have access to a substantial literature documenting effects of all sorts of experimental manipulations on dependent variables scaled in a common metric. Experiments in cognitive psychology, for example, often measure reaction time scaled in milliseconds. Researchers in this area should have little difficulty in deciding whether a difference between population means of a given magnitude (in units of milliseconds) should be regarded as trivially small, extremely large, or somewhere in between.

If a dependent variable is not scaled in informative units, then a CI scaled in dependent variable units will provide no useful inferences beyond confident direction inference. Suppose, for example, that a two-group randomized experiment is carried out to examine differences in the effects of two levels of blood alcohol on performance in a driving task. Subjects assigned to one group consume sufficient alcohol to raise their blood alcohol concentration (BAC) to 0.08%, while subjects in the second group have their BAC raised to 0.05%. Subjects are required to negotiate a slalom course designed specifically for the experiment, and the dependent variable is the number of cones (witches hats) hit by the car, a relatively low score indicating relatively good performance. It is expected, of course, that drivers with a BAC of 0.08 will hit more cones than will drivers with a BAC of 0.05. The experimenter is primarily interested in the size of that difference. The problem is that differences between means in the number of cones hit will be influenced by arbitrary features of this particular slalom, such as the difference between the width of the lane defined by the cones and the width of the car, the distance between adjacent cones, and the length of the slalom. The units of the dependent variable are essentially arbitrary, so an effect size expressed in these units may provide little or no useful information beyond the direction of the difference. Suppose that, unknown to (and unknowable by) the experimenter, the difference between population means is $\mu_{0.08} - \mu_{0.05}$ = 3.6 cones. By itself, this figure conveys almost no useful information beyond the directional inequality $\mu_{0.08} > \mu_{0.05}$. The parameter value implies that a BAC of 0.08 increases the average number of cones hit, relative to a BAC of 0.05, but this may be a small and trivial effect, or a substantial and important effect. This is not an issue about inference from statistics to parameters; it is an issue about the interpretation of parameters.

Standard deviation units The conventional way of dealing with an effect size scaled in arbitrary units of measurement is to divide it by a relevant standard deviation scaled in the same arbitrary units, thereby removing the influence of those units. The resulting quantity is scaled in standard deviation units, thereby making it informative to anyone familiar with these units. Following Cohen (1965), it has become standard practice to assume that both population standard deviations are equal $(\sigma_1 = \sigma_2 = \sigma)$, so the standardized difference between two means (usually called Cohen's *d*) is $(\mu_1 - \mu_2) / \sigma$.[4] Cohen (1969) suggested that in the absence of any better basis for interpreting effect size, standardized effect sizes of 0.2, 0.5 and 0.8 should be interpreted respectively as small, medium and large effects. These suggestions have been widely accepted, and now have the status of established conventions.

Returning now to the slalom example, assume for the moment that the standard deviation of slalom scores in both populations (drivers with a BAC of 0.05 and drivers with a BAC of 0.08) is $\sigma = 20.3$ cones. (Incidentally, this means that the slalom must have a very large number of cones.) The mean difference of 3.6 cones is equivalent to a difference of 3.6 cones/20.3 cones = 0.18 standard deviation units. Most people who are familiar with standard deviation units would interpret this as a small and possibly trivial difference.

If the common population standard deviation were 3.3 cones (rather than 20.3 cones), then the standardized mean difference would be would be 1.09 standard deviation units, a large difference in terms of Cohen's guidelines.

While standardized effect sizes can sometimes provide a basis for relatively informative CI inference when none might otherwise exist, it should not be assumed that standardization always provides more information than the original scaling. A difference between means scaled in an interpretable dependent variable metric can be more informative than the corresponding standardized effect size, because the standardizing transformation removes information in this case. Returning again to the driving experiment, consider another test that might be used to assess the difference between the effects of the two blood alcohol levels. This test requires subjects to perform an emergency braking task: after an appropriate signal is given, the car is to be brought from a particular speed to a complete stop in the shortest possible distance. The dependent variable is braking distance: the distance travelled by the car after the presentation of the signal. It is expected that the higher blood alcohol level will produce longer braking distances than the lower level, but the experimenter is primarily interested in estimating the magnitude of the difference, to be scaled in metres, a familiar and informative unit of distance. The practical implications of an effect size expressed in units of metres can be discussed without any reference to standard deviation units.

Practical equivalence inference

If it is not possible to justify a confident inference that two treatments have identical effects on an outcome measure (that is, the inference $\mu_1 = \mu_2$), what can we make of claims that a treatment has no effect, or that a new treatment is no better than an old treatment, or that the effect of a treatment on males is the same as the effect on females? The problem with assertions of equality is not that there is something peculiar about asserting that the value of a parameter is zero, but rather that it is not possible to justify a confident inference about any point estimate. This problem can be resolved by replacing point estimation with interval estimation.

The first step in practical equivalence inference is to define a range of values of $\mu_1 - \mu_2$ within which the two treatments would be interpreted as equivalent for practical (or theoretical) purposes. In standard deviation units, as assessed by Cohen's (1969) effect size guidelines, the required difference might be very small, small or small–medium. The important point is that the experimenter, a group of researchers, a regulatory body or someone else must be able to specify the maximum difference that can be regarded as trivially small. Call this maximum trivially small difference τ. Then the two treatments are deemed to be practically equivalent if $\mu_1 - \mu_2 \in (-\tau, +\tau)$.

This *practical equivalence interval* defines what is to be meant by a trivially small effect. It is highly desirable, of course, that the value of τ should be specified independently of the data used to justify a claim of practical equivalence, and that researchers in a particular domain should be able to agree on what is to be meant by practical equivalence in that domain.

Note that the practical equivalence interval is a definition, not a CI or any other kind of statistical inference. It is possible to know what is meant by a trivially small effect without being confident that the actual difference between two population means is, in fact, trivially small.

If it turns out when an experiment is run that the CI on $\mu_1 - \mu_2$ lies entirely within the practical equivalence interval, then the inference from the CI implies that the two treatments must have practically equivalent effects on the dependent variable. If the agreed-upon practical equivalence interval is $(-0.25\sigma, 0.25\sigma)$ and it turns out that the CI from an experiment is $(-0.05\sigma, 0.17\sigma)$, then all of the values inside the CI are trivially small. Therefore the CI $\mu_1 - \mu_2 \in (-0.05\sigma, 0.17\sigma)$ implies that $\mu_1 \approx \mu_2$, where the symbol '$\approx$' means 'is practically equivalent to'.

Practical equivalence inference, then, is a special case of CI inference. It is not possible to justify a confident assertion about practical equivalence from a single inference at any lower level (such as confident direction inference or confident inequality inference).[5] It is possible to be confident about practical equivalence and also be confident about direction. The CI $(0.01\sigma, 0.18\sigma)$, for example, justifies the inference $\mu_1 > \mu_2$, and it also justifies the inference $\mu_1 \approx \mu_2$ if the practical equivalence interval is defined as $(-0.20\sigma, 0.20\sigma)$.

Practical equivalence inference is possible only from experiments providing a high degree of precision in estimation, as evidenced by narrow CIs relative to the practical equivalence interval. In practice, it turns out that most randomized experiments in psychology and related disciplines are not capable of producing sufficiently precise estimates of effect size to justify practical equivalence inference, whatever the outcomes of those experiments may be. Within-subjects (repeated measures) designs usually produce more precise estimates of effects than fully randomized between-subjects designs. Practical equivalence inference is sometimes possible from within-subjects designs, even when the sample size is not large. Chapter 6 contains examples of practical equivalence inference from a within-subjects design.

Practical equivalence inference is taken very seriously in some areas of research where the consequences of errors in claims of equivalence can be important, as might be the case when a relatively inexpensive new drug treatment is being considered as a replacement for an expensive standard treatment, and it is important to discover whether the two treatments have practically equivalent effects. In other areas, confident practical equivalence inference is sometimes possible on the basis of a meta-analysis (a quantitative analysis of results from a set of similar studies with a very large total sample size). The requirements of confident practical equivalence inference are the same in the context of meta-analysis as in any other context: a relevant CI must be included in (covered by) a relevant practical equivalence interval.

It is often difficult to justify a precise value of τ (the maximum trivially small difference), so the limits of a practical equivalence interval are often somewhat fuzzy. If the fuzziness is extreme, then practical equivalence inference is not possible.

Constructing a confidence interval on a single comparison

Population standard deviation known

Some of the basic principles about CI construction can be illustrated most clearly if we assume not only that dependent variable scores are normally distributed with the same standard deviation in each population, but also that the experimenter knows the value of the population standard deviation.

Given these assumptions, the procedure for constructing a raw $100(1 - \alpha)\%$ CI on $\mu_1 - \mu_2$ is quite straightforward. First, an unbiased point estimate of the difference between population means (namely $M_1 - M_2$) is calculated, together with the standard error (SE) of that estimate. Second, the standard error is multiplied by a critical value (CV) from a relevant theoretical probability distribution to determine the half-width (w) of the CI.[6] Finally, the CI limits are obtained by adding (and subtracting) the half-width to (and from) the estimated parameter value. That is, the limits of the interval are obtained from

$$(M_1 - M_2) \pm \text{CV} \times \text{SE}.$$

If both groups have the same sample size ($n_1 = n_2 = n = N/2$), the standard error of the difference between means is

$$\text{SE} = \sigma_{M_1 - M_2} = \sigma\sqrt{\frac{2}{n}} \tag{1.1}$$

and the relevant critical value is CV = $z_{\alpha/2}$, the $100(1-\alpha/2)$th percentile point of the z (standard normal) distribution. The $100(1 - \alpha)$% CI is

$$\mu_1 - \mu_2 \in (M_1 - M_2) \pm z_{\alpha/2} \times \sigma_{M_1 - M_2}. \tag{1.2}$$

An alternative and popular way of writing the same CI is

$$(M_1 - M_2) - z_{\alpha/2} \times \sigma_{M_1-M_2} < \mu_1 - \mu_2 < (M_1 - M_2) + z_{\alpha/2} \times \sigma_{M_1-M_2}.$$

In this book we will use expressions like (1.2).

Consider an experiment with $n = 20$ subjects in each of two groups, where the experimenter somehow knows before running the experiment that the population standard deviation is $\sigma = 15$. It follows from (1.1) that the standard error of the difference between the two sample means, a parameter whose value can also be known before the experiment is run, will be $15\sqrt{2/20} = 4.743$. The critical value required for a 95% CI is the 97.5th percentile point of the standard normal distribution, namely $z_{.025} = 1.960$. Therefore the half-width of a 95% CI constructed with the z procedure will be $w = 1.960 \times 4.743 = 9.30$.

Suppose that the sample means from the experiment turn out to be $M_1 = 105.31$ and $M_2 = 98.54$, so that the unbiased point estimate of $\mu_1 - \mu_2$ is $M_1 - M_2 = 6.77$. The limits of the 95% CI are obtained from $(M_1 - M_2) \pm w$ $= 6.77 \pm 9.30$. The required raw CI, then, is $\mu_1 - \mu_2 \in (-2.53, 16.07)$.

If a standardized CI is required and σ is known to be 15, then any of the statistics of interest can be transformed into standard deviation (SD) units by dividing by 15. The difference between sample means is $(6.77/15)\sigma = 0.45\sigma$ (0.45 SD units) and the CI can be expressed as $\mu_1 - \mu_2 \in (-0.17\sigma, 1.07\sigma)$, or in SD units as $(\mu_1 - \mu_2)/\sigma \in (-0.17, 1.07)$. The best point estimate of the standardized difference between population means suggests a medium effect. This estimate is not particularly precise, and all that can be inferred at the 95% confidence level is that either μ_1 is greater than μ_2 by some unknown but nontrivial amount (somewhere between small and large, but not massively large), or the two population means are practically equivalent. The interval does, therefore, justify the inference that μ_2 is not nontrivially larger than μ_1.

Precision The precision of inferences from this CI procedure can be determined before the experiment is run. The standard error of the difference between means is $\sigma\sqrt{2/n} = 0.316\sigma$, and the half-width of the CI is $1.960 \times 0.316\sigma = 0.620\sigma$. The experimenter knows in advance that if the difference between sample means turns out be zero, so that the data contain no suggestion of a difference between population means, the 95% CI will nevertheless include non-negligible positive values (such as $\mu_1 - \mu_2 = 0.6\sigma$), as well as non-negligible negative values (such as $\mu_1 - \mu_2 = -0.6\sigma$). Prior information about precision of estimation can be very useful: an experimenter who expects a small effect may decide to increase the sample size to increase precision, or, if the required resources are not available, to change the experimental design or even abandon the experiment altogether. If a very large effect is expected (such as $\mu_1 - \mu_2 = 2\sigma$), a CI half-width of 0.620σ might indicate an acceptable level of precision.

Implications for directional inference The fact that zero is inside the obtained CI implies that a two-tailed .05 level z test would not have rejected the null hypothesis H_0: $\mu_1 - \mu_2 = 0$, so no inference would have been justified by the test. (The z test rather than the t test is relevant here because of the assumption that σ is known.) In general, it is difficult to know what to make of the failure of a test to produce a directional or inequality inference. Either the difference between $\mu_1 - \mu_2$ and zero is trivially small, or the statistical power of the test has been insufficient to enable the test to detect whatever nontrivial difference exists. It is more difficult to determine the power of the test than it is to determine the precision of the CI, because the power of the test depends on the (unknown) magnitude of $\mu_1 - \mu_2$. It turns out that if $\mu_1 - \mu_2 = 0.8\sigma$ (a large effect according to Cohen's effect size guidelines), then the power of the test is 0.72; if $\mu_1 - \mu_2 = 0.5\sigma$ (a medium effect), then the power is 0.35. Figures like these can provide some help in the task of interpreting a 'nonsignificant' difference, but it is much easier to make sense of the associated CI.

In practice, the z method of CI construction is rarely used, because experimenters are rarely (if ever) in a position to know the population standard deviation. A value of $\sigma = 15$ might perhaps be assumed if the dependent variable is an IQ test scaled so that the standard deviation in the standardization population was set at 15. Even this case is problematic, however, unless subjects in the experiment are randomly sampled from the same population used to standardize the test. In general, it is not necessary to assume that the population standard deviation is known, because methods making using of t distributions do not require this assumption.

Population standard deviation unknown

We now abandon the assumption that the population standard deviation is known, but retain the assumption of normally distributed dependent variable scores with the same standard deviation in each population. Because σ is unknown, it is not possible to calculate the standard error of the difference between means. It is necessary to estimate σ from the data in order to use an expression similar to (1.1) to estimate $\sigma_{M_1 - M_2}$. ANOVA procedures produce a statistic based on variation within groups called *mean square error* (MS_E), an unbiased estimate of the population variance σ^2. We will defer a detailed discussion of this statistic until we discuss the ANOVA model in Chapter 2. At this point we merely note that it is possible to obtain an appropriate estimate $\hat{\sigma}^2$ of the unknown parameter σ^2, and we can use $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ to estimate σ. Replacing σ with $\hat{\sigma}$ in (1.1) produces an expression for the estimated standard error of the difference between sample means (a statistic), rather than the standard error itself (a parameter). If each group has n subjects, the estimated standard error is

$$\hat{\sigma}_{M_1-M_2} = \hat{\sigma}\sqrt{\frac{2}{n}}. \qquad (1.3a)$$

If the two groups have different sample sizes $(n_1 \text{ and } n_2)$, the standard error is estimated from

$$\hat{\sigma}_{M_1-M_2} = \hat{\sigma}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}. \qquad (1.3b)$$

Because we must estimate σ from the data in order to use (1.3a) or (1.3b), the critical value required to construct a CI is a percentile point from a t distribution rather than the standard normal distribution. The half-width of a raw $100(1-\alpha)\%$ CI is $t_{\alpha/2;N-2} \times \hat{\sigma}_{M_1-M_2}$, where $t_{\alpha/2;N-2}$ is the value of the upper $100(1-\alpha/2)$th percentile of the *central t* distribution with $(N-2)$ degrees of freedom. (A central t distribution is an 'ordinary' t distribution. We give it its full title here because we will subsequently need to distinguish between central and noncentral t distributions.) Therefore the CI is

$$\mu_1 - \mu_2 \in (M_1 - M_2) \pm t_{\alpha/2;N-2} \times \hat{\sigma}_{M_1-M_2}. \qquad (1.4)$$

The t procedure produces exact 100(1 – α)% raw CIs, in the sense that if the relevant assumptions are satisfied, the noncoverage error rate produced by the procedure is exactly α.

Suppose that the experiment under discussion produces a variance estimate of $\hat{\sigma}^2 = 234.78$ (slightly larger than the actual population variance of $\sigma^2 = 225$). An experimenter who does not know the value of σ^2 would use the estimated value to produce an estimated standard error of $\hat{\sigma}_{M_1-M_2} = \sqrt{234.78 \times 2/20}$ = 4.845. The experimenter would be in no position to know that in this particular case the estimated standard error is slightly larger than the actual value of $\sigma_{M_1-M_2}$ = 4.743. The critical value required to construct a 95% CI is $t_{.025;38}$ = 2.024. (Note that this is slightly larger than the critical z value of 1.960 used previously when the experimenter supposedly knew the population standard deviation.) The half-width of the interval is 2.024 × 4.845 = 9.81, so the confidence limits are 6.77 ± 9.81, and the interval is (–3.04, 16.58).

This interval is wider than that calculated on the assumption that the experimenter knows the value of the population standard deviation. Two factors contribute to the difference in width. The first is the fact that when the standard deviation is unknown, CI width depends on a statistic (the estimated standard error), and therefore varies across samples. In this particular case the estimated standard error happens to be larger than the actual standard error, but the reverse could just as easily have been the case. The second factor contributing to the difference in CI width is the use of a critical value from a t distribution rather than the z distribution. A critical t value is always larger than the corresponding critical z value, due to the fact that central t distributions have thicker tails than the z distribution.

Standardized confidence intervals Unfortunately the principles used to construct exact raw CIs cannot be used to construct exact standardized CIs when the population standard deviation is unknown, because we cannot divide the raw CI limits by the population standard deviation. We can, of course, divide the raw limits by the sample standard deviation, thereby producing an approximation to an exact standardized CI. (It is worth repeating here that an 'exact' CI is one produced by a procedure controlling the noncoverage error rate exactly over an indefinitely large series of replications of the experiment when the relevant assumptions are satisfied. An inference from an exact interval is still subject to error.)

In this particular case the sample standard deviation is $\hat{\sigma} = 15.32$, slightly larger than the population standard deviation of $\sigma = 15.00$. Dividing the relevant raw statistics by 15.32 produces statistics scaled in sample standard deviation units: $M_1 - M_2 = 0.44\hat{\sigma}$ and the 95% CI is $(-0.20\hat{\sigma},\ 1.08\hat{\sigma})$. If we were to interpret –0.20 and 1.08 as the limits of a standardized CI, we would in effect be inferring that $\mu_1 - \mu_2 \in (-0.20\sigma, 1.08\sigma)$, thereby ignoring the distinction between the sample standard deviation and the population standard deviation. While we cannot claim that $(-0.20, 1.08)$ is an exact standardized CI, we can treat it as an approximate standardized CI. In this particular case the approximation is a good one: it turns out that the exact standardized interval produced by the noncentral t procedure developed by Steiger and Fouladi (1997) is $(-0.19, 1.07)$. In most cases where the total sample size is similar to or larger than that in the example ($N = 40$), interpretations of effect size inferences from approximate standardized intervals should be virtually indistinguishable from those derived from exact intervals. In some cases, however, particularly when the sample size is small and the estimated effect size is large, the approximation can be poor.

If zero is inside (outside) an approximate standardized CI, it will also be inside (outside) the corresponding exact standardized interval and the exact raw CI. That is, the three intervals have the same implication for directional inference.

The relatively complex noncentral t CI procedure is described in Appendix C. While this procedure should be preferred to the central t procedure for the construction of t-based standardized CIs, it cannot be used for the construction of standardized CIs in unrestricted ANOVA-model analyses. Many of the analyses recommended in this book make use of test statistics for which noncentral CI methods are not available.[7]

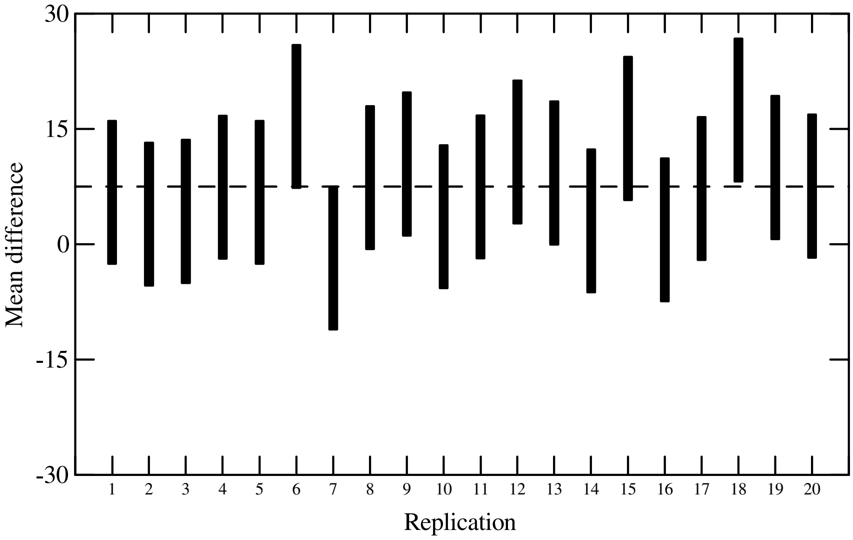
Replicating the experiment: a simulation

The time has come to reveal the source of the data set we have been analysing. The numbers were generated by a computer in a simulation of what might happen if an experiment was replicated 20 times, each replication including a different random sample of 40 subjects from the same population of potential subjects. The difference between population means was set at $\mu_1 - \mu_2 = 7.5$, so that the standardized mean difference is $(\mu_1 - \mu_2)/\sigma = 7.5/15 = 0.5$, a medium effect according to Cohen's guidelines. The data set from Replication 5 in this series of 20 replications was used for the analyses discussed earlier. (Data from any of the other 19 replications could have been used to demonstrate CI procedures. Replication 5 was chosen because that particular data set happens to illustrate certain principles better than some of the other data sets.)

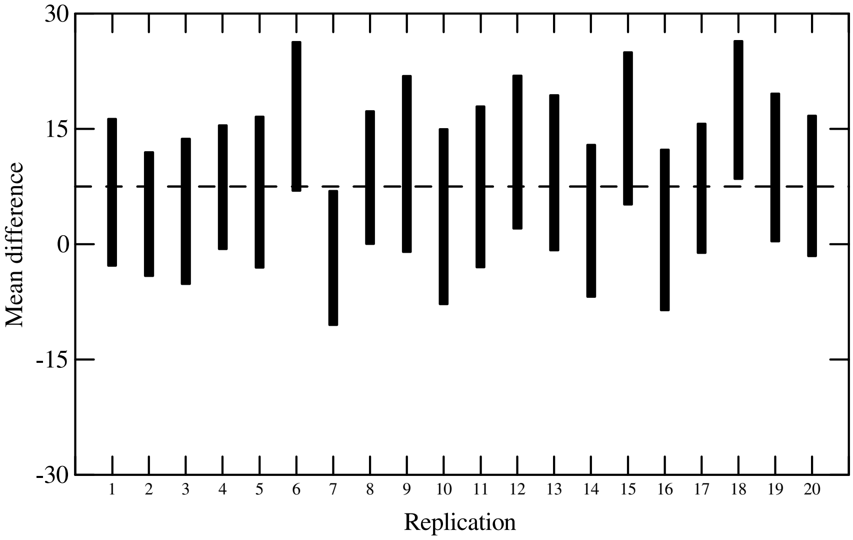
Raw confidence intervals We can see the operation of the most important principles underlying confident inference on mean differences by examining the variation between intervals across replications. The 95% raw CIs are shown in Figure 1.1, with intervals constructed with the *z* procedure (assuming known population variance) in the upper panel, and intervals constructed with the *t* procedure in the lower panel. The most striking feature of these graphs is their similarity – estimating the standard error from the data does not have a substantial impact on the outcomes of the analysis when the relevant *t* distribution has 38 degrees of freedom. The most important difference between the graphs is the absence of variability in the widths of the intervals produced by the *z* procedure; the width of all such intervals is 18.59, while the width of the *t* intervals varies between 16.07 and 22.88.

The broken line shows the value of the population mean difference, so any CI covering this line produces a correct inference, and any interval not covering the line produces a noncoverage error. Both procedures produce a noncoverage error from Experiment 18, and the *t* procedure also produces a noncoverage error from Experiment 7. We can be confident from statistical theory that if the number of replications in this simulation had been increased from 20 to (say) 10,000, then the percentage of replications with noncoverage errors would be extremely close to 5 for both the *z* and *t* procedures. (As Experiment 7 shows, this does not mean that both procedures always produce the same inference.)

Both procedures produce an interval containing only positive values (thereby justifying the correct directional inference $\mu_1 > \mu_2$) from 6 of the 20 replications. (Replications 8 and 9 produce a correct directional inference from



(a) Confidence intervals constructed with the z procedure



(b) Confidence intervals constructed with the t procedure

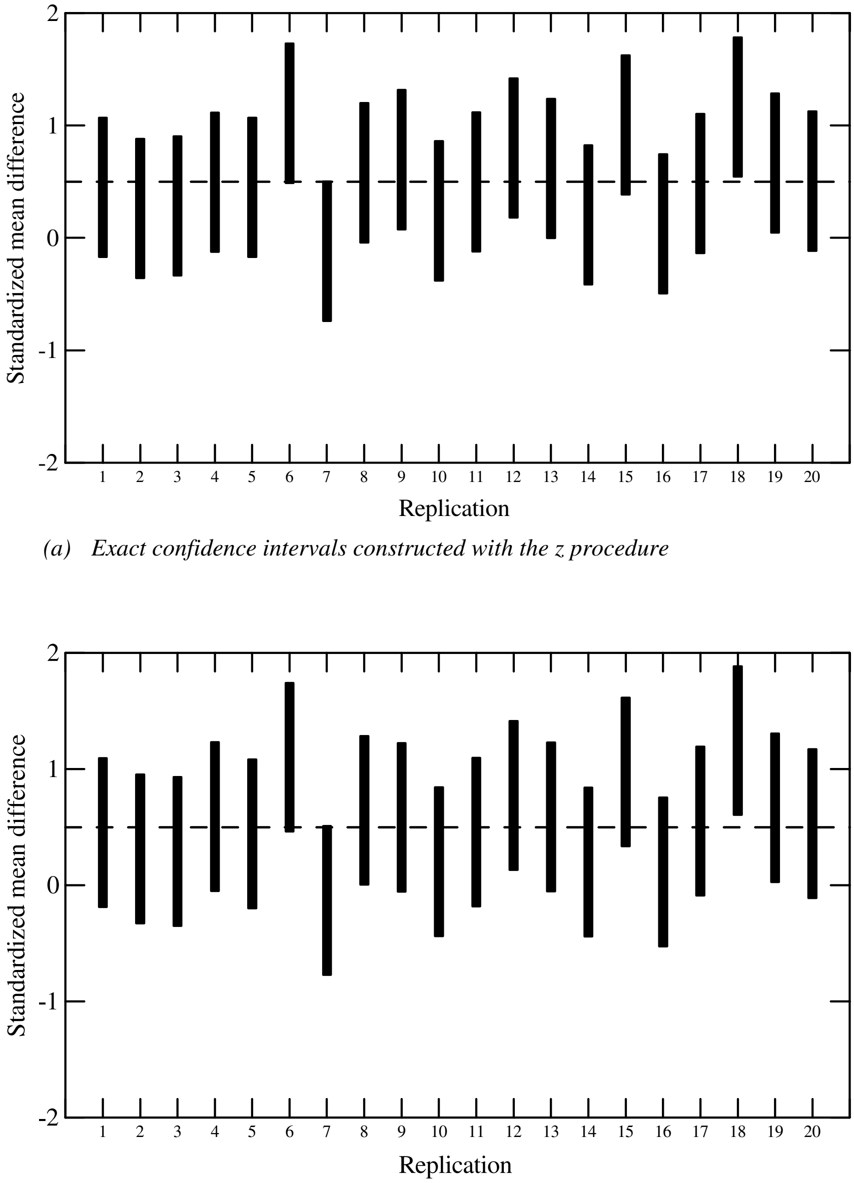
Figure 1.1 *Raw 95% confidence intervals on a difference between two means from 20 replications of one experiment*

one but not both of the two procedures.) This result is consistent with expectations from a statistical power analysis, which shows that the power of a two-tailed t test in this context is .34, while the power of a two-tailed z test is .35. Neither procedure produces a Type III error (a directional inference in the wrong direction). Note that a Type I error is not possible here, because the null hypothesis is false.

Standardized confidence intervals Standardized CIs are shown in Figure 1.2. The upper panel shows the exact standardized intervals constructed on the assumption that the population variance is known, and the lower panel shows the approximate standardized intervals obtained by dividing the limits of each raw t interval from Figure 1.1(b) by the relevant estimated standard deviation. Again, the two sets of intervals are similar, indicating that the approximation is probably a reasonable one, at least for this particular combination of sample size and effect size. Unlike the exact raw intervals in Figure 1.1(b), the approximate standardized intervals in Figure 1.2(b) do not vary in width: the width of all 20 intervals is 1.28 (estimated) standard deviation units. The exact standardized intervals from the z (known variance) procedure also have constant width, namely 1.24 population standard deviation units. Unlike the raw intervals, the two sets of standardized intervals differ slightly in their midpoints, because the raw midpoints are divided by slightly different quantities (the constant population standard deviation in the case of the z procedure, and the variable sample standard deviation in the case of the t procedure).

If standardized CIs are to be interpreted in accordance with Cohen's effect size guidelines, then a t-based approximate standardized interval emerging from a given experiment in this series is likely to produce much the same interpretation as a z-based exact interval. Consider, for example, the results from Replication 9, which produced the largest estimated standard deviation (17.88) and the greatest discrepancy (0.11) between the two estimates of the standardized difference between population means. Furthermore, this was the only replication producing a directional inference (a statistically significant difference) from the z procedure but not the t procedure. The z-based interval is (0.08, 1.32), while the t-based interval is (–0.06, 1.22). Both intervals assert that μ_1 is not nontrivially smaller than μ_2, and that μ_1 is practically equivalent to or greater than μ_2. Both intervals fail to provide a precise estimate of the magnitude of the difference between means, excluding only nontrivial negative and very large positive differences.

The sample size ($n = 20$, $N = 40$) used in this simulation is not small, relative to the sample sizes often used in experimental psychology. Two aspects of the simulation data would probably surprise many researchers, particularly those who expect relatively small samples to provide reasonably precise estimates of parameters of interest. First, the width of the standardized CIs (1.28 for t-based



(a) Exact confidence intervals constructed with the z procedure

(b) Approximate confidence intervals constructed with the t procedure

Figure 1.2 *Standardized 95% confidence intervals on a difference between two means from 20 replications of one experiment*

intervals) implies that a substantial range of possibly important differences between population means must be included in any given interval. An interval with a midpoint of zero, for example, also includes small and medium positive differences, as well as small and medium negative differences. If this degree of (lack of) precision is unacceptable in a particular context, then the problem is with the sample size, not with the method used to construct the CI. Indeed, an advantage of the CI approach is that it is a relatively simple matter to estimate the precision of estimation (standardized CI width) before the experiment is run. In any two-group experiment with 20 subjects per group, the width of an exact 95% standardized CI produced by the z method will be exactly 1.24, and the width of an approximate 95% interval produced by the t method will be exactly 1.28. If such an interval is deemed to be unacceptably imprecise, it is not difficult, as we will see in Chapter 3, to determine the sample size required to produce a standardized interval of any desired width. Those who are surprised by the width of the intervals in this simulation would probably also be surprised by the magnitude of the variability across replications in the locations (midpoints) of the intervals. The standard deviation of (raw) interval midpoints across an indefinitely large number of replications is simply the standard error used to construct the z-based intervals. (For raw z-based intervals, this figure is 4.74; the standard deviation of the midpoints of the sample of 20 such intervals shown in Figure 1.1(a) is 4.89.) It follows that an experimenter who has access only to data from a single replication is nevertheless able to estimate the variability across replications in CI midpoints from the same statistic (the estimated standard error, which is 4.84 in Replication 5) used to construct the t-based interval emerging from that single replication.

Implications for directional inference Because confident direction inference ($\mu_1 > \mu_2$ or $\mu_1 < \mu_2$) is the highest level of statistical inference aspired to by many researchers, it is of some interest to see what inferences at this level would be possible from the 20 replications in the simulation. Before running such an experiment, a researcher planning to carry out a .05 level two-tailed t test (equivalent to using a 95% CI for directional inference only) would know that if $\mu_1 - \mu_2 = 0$, then the probability of no inference is .95, while the probability of a (necessarily incorrect) directional inference is .05. These probabilities do not apply to the replications in the simulation, because, unknown to the experimenter, the difference between μ_1 and μ_2 is greater than zero by 0.5 standard deviation units. The probability of a correct directional inference ($\mu_1 > \mu_2$) is only .34, while the probability of no directional inference is .66. The probability of a Type III error (getting the direction wrong) in this case is too small to worry about. After the experiment is run, the experimenter knows that in that particular sample from a population of potential replications of the experiment, the test outcome either does or does not justify a confident directional inference. If the experiment happens to be Replication 5 from the simulation, then it turns out that no inference can be justified by the test. (The probability under the null hypothesis of a *t* at least as large as the obtained value of 1.40 is $p = .17$.) This 'nonsignificant' (no inference) outcome is, needless to say, absolutely uninformative, not because there is anything wrong with the test procedure, but simply because at this level of inference much of the information in a potentially informative 95% CI is ignored. As we saw earlier, the *t*-based standardized CI (–0.20, 1.08) excludes the possibility of a confident direction inference, but it does support the inference that μ_2 is not nontrivially larger than μ_1, among other things. Unlike the test, the interval also shows that the experiment is not sufficiently sensitive to permit a precise estimate of the magnitude of the effect.

The subjectivist critique of confidence interval inference

CI inference is part of the *classical* or *frequentist* approach to statistical inference, which treats parameters such as population means as fixed and statistics such as sample means as variable. As a consequence, probability statements refer to statistics rather than parameters. Thus, while it may be possible to justify a statement like: 'The probability that a CI constructed in a particular way will cover the parameter is .95', it is considered improper to state that 'the probability that the parameter is covered by this CI is .95'.

CIs are often misinterpreted because the ordinary-language meaning of the word 'probability' is more closely related to the interpretation of that term in the *subjectivist* approach to statistical inference than it is to the interpretation of the same term in the classical approach to inference. The subjectivist (or Bayesian) approach treats parameter values as values of random variables, thereby making it possible to define probability distributions referring to parameters like population means or differences between population means.

The Bayesian inferential framework requires the experimenter to specify a *prior* probability distribution of the parameter of interest. The prior probability distribution is generally interpreted as a distribution of *subjective* probabilities reflecting the researcher's beliefs about the relative credibility of various parameter values, prior to seeing the data. Given this prior distribution and the data (together with an additional probability distribution referring to data rather than parameters), a Bayesian analysis produces a *posterior* distribution of the parameter (a revised distribution taking the data into account). The posterior distribution can be used to construct an interval (usually called a credible interval) for which an interpretation like 'the probability that the parameter lies in this interval is .95' is appropriate. For further details on Bayesian inferential procedures, see Pruzek (1997).

From a Bayesian perspective the classical significance testing and CI approaches are flawed, because they do not allow for probability distributions on parameters (either prior or posterior), and therefore cannot justify the kinds of inferential statements researchers would like to make.

Bayesian inference is not without its critics (see, for example, Oakes, 1986). The most obvious target for criticism is the role played in Bayesian inference by the prior probability distribution, allowing the beliefs and prejudices of an individual researcher to influence the interpretation of the data. This criticism can be dealt with in a number of ways, one of which is to use an 'uninformative' prior distribution, thereby ensuring that prior beliefs have no influence on the outcome of the analysis. It turns out that the use of uninformative prior distributions in a Bayesian analysis produces credible intervals whose limits are similar (if not identical) to those of CIs from a classical analysis (Pruzek, 1997). It would appear, then, that the consequences of common misinterpretations of CIs are less profound in practice than they might appear in theory.

It would be a mistake, however, to ignore the implications of the Bayesian approach to inference, because it does help to make explicit some of the limitations of classical methods of analysis. If an experiment is one of a series, each of which adds something to an existing body of knowledge, a Bayesian analysis can, at least in principle, take that knowledge into account in the specification of the prior distribution. Inferences from a classical analysis, on the other hand, can be influenced only by the data in the current experiment. In a discipline such as psychology where statistical power is typically low (Sedlmeier and Gigerenzer, 1989), classical analyses at the level of individual experiments can be expected to produce imprecise inferences, relative to those sometimes possible from meta-analyses of sets of similar experiments (Schmidt, 1996), or, in a Bayesian framework, analyses of individual experiments that take prior research into account. It does not follow, however, that the rate of incorrect inferences from CIs is likely to be higher than the nominal error rate.

Further reading

Cumming and Finch (2001) and Smithson (2003) provide extensive discussions of CI procedures based on both central and noncentral *t* distributions. If you feel the need to consolidate your understanding of CI inference before proceeding to Chapter 2, these are good places to start.

Hsu (1996) discusses some of the most important ideas introduced in this chapter (particularly levels of inference and practical equivalence inference). Hsu's treatment does, however, assume a greater degree of mathematical sophistication than is assumed here. For examples of practical equivalence inference in psychology, see Rogers, Howard and Vessey (1993). Cohen (1988), Richardson (1996) and Rosenthal (1994) provide discussions of effect size measures.

If you would like to become acquainted with the recent history of the debate in psychology concerning the relative value of CIs and significance tests, you will find it worthwhile to consult Cohen (1994), Nickerson (2000), Schmidt (1996), or some of the relevant chapters in Harlow, Mulaik and Steiger (1997). For an indication of the approach to this issue recently adopted by the American Psychological Association, see Wilkinson and the Task Force on Statistical Inference (1999) and the 5th edition of the APA Publication Manual (American Psychological Association, 2001).

Oakes (1986) provides a very readable book-length discussion and critique of a number of approaches to statistical inference, including classical and Bayesian approaches.

Questions and exercises

At the end of each chapter you will find a set of questions and exercises designed to test your understanding of the material in the chapter, and, particularly in subsequent chapters, to provide you with opportunities to practise carrying out relevant analyses. You can check on your answers by consulting Appendix E.

1. In a study designed to investigate the effect of practice on performance on an aptitude test, participants are randomly assigned to one of two experimental conditions. Those in the first (treatment) condition are given practice on items similar to those in the aptitude test, while those in the second (control) condition spend the same amount of time answering questions in an interest inventory. The experimenters are primarily interested in knowing whether the magnitude of the practice effect is large enough to justify changes in a selection procedure that makes use of the test. A mean practice effect of 3 (items correct) is regarded as the smallest nontrivial effect.

The inference about the practice effect is to be based on a 95% CI on $\mu_T - \mu_C$.

What conclusion (if any), at each of the three levels of confident inference discussed in this chapter (interval, direction and inequality inference), would follow from each of the following CIs:

(a) $\mu_T - \mu_C \in (6.5, 8.7)$

(b) $\mu_T - \mu_C \in (0.9, 16.8)$

(c) $\mu_T - \mu_C \in (-0.6, 1.6)$

(d) $\mu_T - \mu_C \in (-7.4, 8.5)$?

2. Assume that the within-condition standard deviation for the experiment in Question 1 is known (independently of the data) to be $\sigma = 8.2$. Given the raw CIs in Question 1, construct and interpret standardized CIs.

3. Comment on the relative precision of the CIs you constructed in your answer to Question 2. Without doing any additional calculations, comment also on sample sizes in the four cases.

4. What types of inferential error (if any) follow at each level of confident inference (CI, confident direction, confident inequality) from the raw CIs produced by the *t* procedure from Replication 1 and Replication 18 of the simulated experiment discussed on page 17? To answer this question you will need to consult Figure 1.1(b).

5. Suppose that you were an experimenter running the simulated experiment, and you obtained the same data as that produced by Replication 18. Could you answer Question 4 if it referred to your data? If not, why not?

Notes

1. A fixed-effects ANOVA model (the type of model usually implied when the term ANOVA is used without qualification) is appropriate for the analysis of randomized experiments where the various experimental conditions (treatments) are selected by the experimenter. A *random-effects* ANOVA model, otherwise known as a variance components model, is appropriate when treatments are randomly sampled from a population of potential treatments. See Bird (2002) or Smithson (2003) for brief discussions of CI inference on parameters of random-effects models.

2. The term *replication* is used in at least three different senses by statisticians and experimenters. The term is used in this book to refer to a repetition of an experiment that makes use of a different random sample of subjects from the same population, but is otherwise identical to the original experiment (or another replication).

3. The usual derivation of the standard error used in a two-group *t* test (or CI) assumes that the subjects assigned to each treatment are randomly sampled from a population of infinite size. In practice, the 'sample' of $N = 2n$ subjects in a two-group experiment is usually a convenience sample, not a random sample from any population. Given random assignment from a convenience sample, the same standard error can be derived by replacing the random sampling assumption with the assumption that the size of the treatment effect does not vary across subjects (Reichardt and Gollub, 1999). If the size of the treatment effect does vary across subjects, the standard error used by the *t*-test procedure is likely to be too large, so that the inferences from the procedure are valid but conservative, with too few Type I errors from tests and too few noncoverage errors from CIs. Given random assignment from a convenience sample, the parameter $\mu_1 - \mu_2$ refers to the 'population' of all subjects in the experiment. Of course, random sampling

has one important advantage over random assignment: it provides a justification for a generalization beyond the $N = 2n$ subjects in the experiment to the population from which they were sampled. In the terminology of Cook and Campbell (1979), random assignment provides a basis for claims of internal validity, while random sampling provides a basis for claims of both internal and external validity.

4. Glass (1976) suggested that the standardized effect size should be expressed in units of variability in the control population. For various reasons, including the fact that the designation of one treatment as the 'control' is often arbitrary, it has become standard practice to assume that both treatment populations have the same standard deviation, and to use this common standard deviation as the unit of measurement when defining a standardized effect size.

5. It is possible to justify confident practical equivalent inference by carrying out two nonstandard tests, one allowing for the possibility of the inference $\mu_1 - \mu_2 < \tau$, the other allowing for the possibility of the inference $\mu_1 - \mu_2 > -\tau$ (Rogers, Howard and Vessey, 1993). The CI approach, however, is simpler and more informative.

6. Following Lockhart (1998) and Smithson (2000), the symbol w is used to refer to the half-width (rather than the width) of a CI. The width of a CI is therefore $2w$.

7. Many ANOVA-model analyses make use of an F test statistic, critical values of which can be used to construct CIs on various monotonic functions of the noncentrality parameter of a noncentral F distribution. (See Appendix C for details.) In most cases it is not possible to transform a CI on this noncentrality parameter into a CI on contrasts (generalized comparisons). A number of ANOVA-model analyses use test statistics for which noncentral interval estimation procedures have not been developed.