CHAPTER 1

Brief Review of Hypothesis Testing Concepts/Issues and Confidence Intervals

Statistical techniques are used for purposes such as estimating population parameters using either point estimates or interval estimates, developing models, and testing hypotheses. For each of these uses, a sample must be obtained from the population of interest. The immediate question is then "How large should the sample be?" That is the focus of this book. There are several types of sampling methods that are used, such as simple random sampling, stratified random sampling, and cluster sampling. Readers interested in learning about these methods are referred to books on sampling. Such books range from books with an applied emphasis such as Thompson (2012) to an advanced treatment with some theoretical emphasis as in Lohr (2010). Readers interested in an extensive coverage of sample survey methodology may be interested in Groves, Fowler, Couper, Lepkowski, Singer, and Tourangeau (2009).

1.1 BASIC CONCEPTS OF HYPOTHESIS TESTING

If sampling is very inexpensive in a particular application, we might be tempted to obtain a very large sample, but settle for a small sample in applications where sampling is expensive.

The cliché "the bigger the better" can cause problems that users of statistical methods might not anticipate, however. To illustrate, assume that there are two alternative methods that could be employed at some stage of a manufacturing

Sample Size Determination and Power, First Edition. Thomas P. Ryan.

process, and the plant manager would like to determine if one is better than the other one in terms of process yield. So an experiment is performed with one of the methods applied to thousands of units of production, and then the other method applied to the same number of units.

What is likely to happen if a hypothesis test (also called a significance test) is performed, testing the equality of the population means (i.e., the theoretical average process yield using each method), against the alternative hypothesis that those means are not equal? Almost certainly the test will lead to rejection of the (null) hypothesis of equal population means, but we should know that the means, recorded to, say, one decimal place are not likely to be equal before we even collect the data! What is the chance that any two U.S cities, randomly selected from two specified states, will have exactly the same population? What is the probability that a company's two plants will have exactly the same proportion of nonconforming units? And so on. The bottom line is that null hypotheses (i.e., hypotheses that are tested) are almost always false. This has been emphasized in the literature by various authors, including Nester (1996) and Loftus (2010).

Other authors have made similar statements, although being somewhat conservative and less blunt. For example, Hahn and Meeker (1991, p. 39) in pointing out that hypothesis tests are less useful than confidence intervals stated: "Thus, confidence intervals are usually more meaningful than statistical hypothesis tests. In fact, one can argue that in some practical situations, there is really no reason for the statistical hypothesis to hold exactly."

If null hypotheses are false, then why do we test them? [This is essentially the title of the paper by Murphy (1990).] Indeed, hypothesis testing has received much criticism in the literature; see, for example, Nester (1996) and Tukey (1991). In particular, Loftus (1993) stated "First, hypothesis testing is overrated, overused, and practically useless as a means of illuminating what the data in some experiment are trying to tell us." Provocative discussions of hypothesis testing can also be found in Loftus (1991) and Shrout (1997). Howard, Maxwell, and Fleming (2000) discuss and endorse a movement away from heavy reliance on hypothesis testing in the field of psychology. At the other extreme, Lazzeroni and Ray (2012) refer to millions of tests being performed with genomics data.

Despite these criticisms, a decision must be reached in some manner about the population parameter(s) of interest, and a hypothesis test does directly provide a result ("significant" or "not significant") upon which a decision can be based. One of the criticisms of hypothesis testing is that it is a "yes–no" mechanism. That is, the result is either significant or not, with the magnitude of an effect (such as the effect of implementing a new manufacturing process) hidden, which would not be the case if a confidence interval on the effect were constructed.

Such criticisms are not entirely valid, however, as the magnitude of an effect, such as the difference of two averages, is in the numerator of a test statistic. When we compute the value of a test statistic, we can view this as a linear transformation of an effect. For example, if we are testing the null hypothesis,

$H_0: \mu_1 = \mu_2$, which is equivalent to $\mu_1 - \mu_2 = 0$, the difference in the two parameters is estimated by the difference in the sample averages, $\bar{x}_1 - \bar{x}_2$, which is in the numerator of the test statistic,

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{S_{\bar{x}_1 - \bar{x}_2}} \tag{1.1}$$

with $S_{\bar{x}_1 - \bar{x}_2}$ denoting the standard error (i.e., estimator of the standard deviation) of $\bar{x}_1 - \bar{x}_2$, and 0 is the value of $\mu_1 - \mu_2$ under the null ypothesis. Thus, the "effect," which is estimated by $\bar{x}_1 - \bar{x}_2$, is used in computing the value of the test statistic, with every type of t-statistic having the general form: $t =$ estimator/standard error of estimator.

Many practitioners would prefer to have a confidence interval on the true effect so that they can judge how likely the true (unknown) effect, $\mu_1 - \mu_2$, is to be of practical significance. For example, Rhoads (1995) stated that many epidemiologists consider confidence intervals to be more useful than hypothesis tests. Confidence intervals are reviewed in Section 1.2.

In using the test statistic in Eq. (1.1) to test the null hypothesis of equal population means, we must have either a reference value in mind such that if the test statistic exceeds it in absolute value, we will conclude that the means differ, or, as is commonly done, a decision will be based on the "p-value," which is part of the computer output and is the probability of obtaining a value of the test statistic that is more extreme, relative to the alternative hypothesis, as the value that was observed, conditioned on the null hypothesis being true. As discussed earlier in this section, however, null hypotheses are almost always false, which implies that p-values are hardly ever valid. Therefore, the p-values contained in computer software output should not be followed slavishly, and some people believe that they shouldn't be used at all (see, e.g., Fidler and Loftus, 2009).

If we use the first approach, the reference value would be the value of the test statistic determined by the selected significance level, denoted by α, which is the probability of rejecting a (conceptually) true null hypothesis. This is also called the probability of a Type I error. If the test is two-sided, there will be two values that are equal in absolute value, such as ± 1.96, with the null hypothesis rejected if the test statistic exceeds 1.96 or is less than -1.96. If we adopt the second approach and, for example, $p = .038$, we may (or may not) conclude that the null hypothesis is false, whereas there would be no doubt if $p = .0038$, since that is a very small number and in particular is less than .01. (Recall the discussion about null hypotheses almost always being false, however.)

There are four possible outcomes of a hypothesis test, as the null hypothesis could be (1) correctly rejected, (2) incorrectly rejected, (3) correctly not rejected, or (4) incorrectly not rejected. The latter is called a Type II error and the probability of a Type II error occurring is denoted by β. Thus, $1 - \beta$ is the probability of correctly rejecting a false null hypothesis and this is termed "the power of

the test." An experimenter must consider the costs associated with each type of error and the cost of sampling in arriving at an appropriate sample size to be used in hypothesis tests, as well as to determine an appropriate sample size for other purposes.

Some practitioners believe that the experiments should be conducted with the probability of a Type I error set equal to the probability of a Type II error. Although the former can literally be "set" by simply selecting the value, the latter depends on a number of factors, including the difference between the hypothesized parameter value and the true parameter value α, the standard deviation of the estimator of the parameter, and the sample size. We cannot literally set the probability of a Type II error because, in particular, the standard deviation of the estimator of the parameter will be unknown. So even though we may think we are setting the power for detecting a certain value of the parameter with the software we use, we are not literally doing so since the value for the standard deviation that the user must enter in the software is almost certainly not the true value.

Since $\alpha \leq .10$, typically, and usually .05 or .01, this would mean having power $\geq .90$ since power $= 1 - \beta$, as stated previously. Although this rule-of-thumb may be useful in some applications, it would result in a very large required sample size in many applications since increased power means increased sample size and power of .95 or .99 will often require a much larger sample size than power $=$ .90, depending on the value of the standard error. Thus, in addition to being an uncommon choice for power, .95 or .99 could require a sample size that would be impractical. The increased sample size that results from using .95 or .99 is illustrated in Section 3.1.

Regarding the choice of, α one of my old professors said that we use .05 because we have five fingers on each hand, thus making the point that the selection of .05 is rather arbitrary. Mudge, Baker, Edge, and Houlahan (2012) suggested that α be chosen to either (a) minimizing the sum of the probability of a Type I error plus the probability of a Type II error at a critical effect size, or (b) "minimizing the overall cost associated with Type I and Type II errors given their respective probabilities."

There are various misinterpretations of hypothesis test results and p-values, such as concluding that the smaller the p-value, the larger the effect or, for example, the difference in the population means is greater if the equality of two means is being tested. A p-value has also been misinterpreted as the probability that the null hypothesis is true. These types of misinterpretations have been discussed in the literature, such as in Gunst (2002) and Hubbard and Bayarri (2003). There have also been articles about p-value misconceptions in which the author gives an incorrect or at least incomplete definition of a p-value. Goodman (2008) is one such example, while giving 12 p-value misconceptions. Hubbard and Bayarri (2003) stated: "The p-value is then mistakenly interpreted as a frequency-based Type I error rate." They went on to state that "confusion over the meaning and interpretation of p's and α's is almost total . . . this same confusion

exists among some statisticians." The confusion is indeed apparent in some introductory statistics textbooks, some of which have defined a *p*-value as "the smallest Type I error rate that an experimenter is willing to accept." Berk (2003), in discussing Hubbard and Bayarri (2003), quoted Boniface (1995, p. 21): "The *level of significance* is the probability that a difference in means has been erroneously declared to be significant. Another name for significance level is *p*-value." See also the discussion in Seaman and Allen (2011). Additionally, Casella and Berger (1987, p. 133) stated that "there are a great many statistically naive users who are interpreting *p*-values as probabilities of Type I error."

The bottom line is that *p*-values are completely different conceptually from the probability of a Type I error (i.e., significance level) and the two concepts should never be intermingled. There has obviously been a great deal of confusion about these concepts in the literature and undoubtedly also in practice.

There has also been confusion over what can be concluded regarding the null hypothesis. If the sample data do not result in rejection of it, that does not mean it is true (especially considering the earlier discussion of null hypotheses in this chapter), so we should not say that it is accepted. Indeed, the null hypothesis can never be proved to be true, and for that matter, it can never be proved that it isn't true (with absolute, 100% certainty), so we should say that it is "not rejected" rather than saying that it is "accepted." This is more than just a matter of semantics, as there is an important, fundamental difference. (The alternative hypothesis also cannot be "proved," nor can *anything* be proved whenever a sample is taken from a population.) The reader who wishes to do additional reading on this may wish to consult Cohen (1988, pp. 16–17).

A decision must be reached as to whether a two-sided test or a one-sided test will be performed. For the former, the alternative hypothesis is that the parameter or the difference of two parameters is not equal to the value specified in the null hypothesis. A one-sided test is a directional test, with the parameter or the difference of two parameters specified as either greater than or less than the value specified in the null hypothesis. Bland and Altman (1994) stated that a one-sided test is sometimes appropriate but further stated the following:

> In general a one sided test is appropriate when a large difference in one direction would lead to the same action as no difference at all. Expectation of a difference in a particular direction is not adequate justification. In medicine, things do not always work out as expected, and researchers may be surprised by their results . . . . Two sided tests should be used unless there is a very good reason for doing otherwise.

1.2 REVIEW OF CONFIDENCE INTERVALS AND THEIR RELATIONSHIP TO HYPOTHESIS TESTS

Many practitioners prefer confidence intervals to hypothesis tests, especially Smith and Bates (1992). Confidence intervals do provide an interval that will

contain the parameter value (or difference of parameter values) of interest with the stated probability, such as .95. Many types of confidence intervals are symmetric about the estimate of the parameter for which the interval is being constructed. Such intervals are of the form

$$\hat{\theta} \pm t(\text{or} Z)\hat{\sigma}_{\hat{\theta}}(\text{or}\, \sigma_{\hat{\theta}})$$

where θ is the parameter for which the confidence interval is being constructed, $\hat{\theta}$ is the estimator of that parameter, $\hat{\sigma}_{\hat{\theta}}$ is the estimator of the standard deviation of the estimator ($\sigma_{\hat{\theta}}$), and either t or Z is used in constructing the interval, depending on which should be used.

A confidence interval is constructed by taking a single sample, but, speaking hypothetically to add insight, if we were to take a very large number of samples and construct a 95% confidence interval using the data in each sample, approximately 95% of the intervals would contain the (unknown value) of the parameter since the probability that any one interval will contain the parameter is .95. (Such statements can of course be verified using simulation.) Such a probability statement must be made before a sample is obtained because after the interval has been computed the probability is either zero or one that the interval contains the parameter, and we don't know which it is because we don't know the value of the parameter.

A confidence interval does have the advantage of preserving the unit of measurement, whereas the value of a test statistic is a unitless number. There is a direct relationship between a hypothesis test and the corresponding confidence interval, as emphasized throughout Ryan (2007). In particular, we could use a confidence interval to test a hypothesis, as there is a direct relationship between a two-sided hypothesis test with significance level α and a $100(1 - \alpha)\%$ confidence interval using the same data. Similarly, there is a direct relationship between a one-sided hypothesis test and the corresponding one-sided confidence bound.

Specifically, if H_0: $\mu_1 = \mu_2$, equivalently H_0: $\mu_1 - \mu_2 = 0$, is not rejected using a two-sided test with significance level α, then the corresponding $100(1 - \alpha)\%$ confidence interval will contain zero. Similarly, if the hypothesis test had led to rejection of H_0, then the confidence interval would not have included zero. The same type of statements can be made regarding what will happen with the hypothesis test based on the confidence interval. This relationship holds true for almost all hypothesis tests. An argument could be made that it is better to test a hypothesis by constructing the confidence interval because the unit of measurement is not lost with the latter, but is lost with the former.

Although an alternative hypothesis value for the parameter of interest is not specified in confidence interval construction because power is not involved, since the form of a confidence interval is just a rearrangement of the components of

the corresponding hypothesis test, values of those components must be specified before the sample size for a confidence interval can be determined, just as is the case with hypothesis tests. So confidence intervals share this obstacle with hypothesis tests.

Software for sample size determination is primarily oriented toward hypothesis testing, however. For example, although Power and Precision provides a 95% confidence interval in addition to the necessary sample size for the specified power value, in addition to the capability for obtaining a tolerance interval for a future 95% confidence interval for the mean, there is no way to solve for the sample size such that a confidence interval will have a desired expected width, a topic that is usually presented in introductory statistics texts. This capability is also absent in some multipurpose statistical software that can be used for sample size determination, such as Stata. Sample size for confidence intervals can be determined using MINITAB, however.

Among software specifically for sample size determination and power, the capability for solving for sample size for specified confidence interval widths is available in PASS, as well as the capability to obtain a tolerance interval for a future confidence interval. nQuery also provides the capability for determining sample size for confidence intervals, with the user specifying the desired half-width of the interval.

Software capability for sample size determination and power is discussed in detail in subsequent chapters.

If a null hypothesis is false, the experimenter either correctly rejects it or makes what has been termed a Type II error in failing to reject it. (A Type I error occurs, conceptually at least, when an experimenter rejects a true null hypothesis.) If the true parameter value, or difference of two parameter values, is very close to the hypothesized value, there is a high probability of making a Type II error, but that is not of any consequence since it should be understood that the true value is almost certainly not equal to the hypothesized value. Thus, there is some true, unknown parameter value that is presumably close to the hypothesized value. What is important, however, is to detect a difference between the hypothesized and assumed parameter value that is of practical importance, recognizing the difference between statistical significance and practical significance.

Of course, experimenters are going to know what is of practical significance in their studies, and they might frown on tests that show a statistically significant result that they know to be not of practical significance. The first sentence of the article Thomas and Juanes (1996) states: “Statistical significance and biological significance are not the same thing.”

The probability of correctly detecting the difference between the true and hypothesized parameter values is called the *power of the test*, which is $1 - \beta$, with β representing the probability of a Type II error. The latter is computed by determining the probability that the value of the random variable that is the

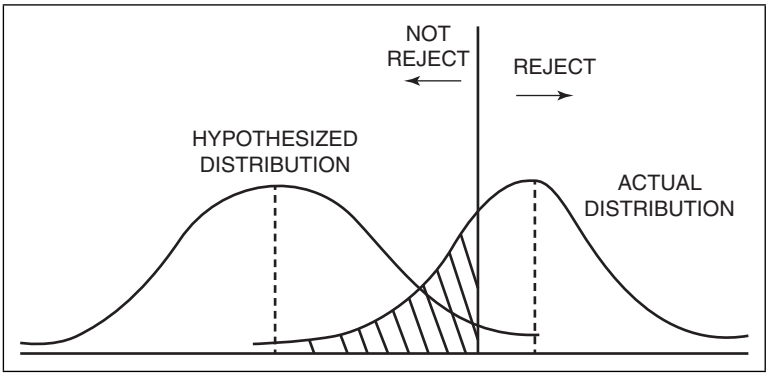

Figure 1.1 Illustration of Type II error probability.

estimator of the parameter being tested falls in the nonrejection region for the curve of the hypothesized distribution of the random variable, which contains the hypothesized parameter value. This is illustrated in Figure 1.1, for which a one-sided test is assumed.

The probability of a Type II error is represented by the shaded area. The smaller this area, the greater the power of the test. It should be apparent, however, that the shaded area will not be small when there is only a small difference between the true parameter value and the hypothesized value. The area can be decreased by increasing the sample, which will cause the spread of each curve to be less, but a very large sample size can make a test too sensitive to small differences between the true and hypothesized parameter values.

There is a trade-off between Type I and Type II errors because increasing one will decrease the other. An experimenter can decide which is the more important of the two in a particular application. Students in a statistics course might be exposed to this in a courtroom setting, where the null hypothesis of course is that the person on trial is innocent (until proved guilty). For example, see Feinberg (1971) and Friedman (1972). So a Type I error would be convicting an innocent person and a Type II error would be not convicting a guilty person. While either error could have dire consequences, in the United States, avoidance of a Type I error would be considered most important. In drug testing, the Food and Drug Administration naturally wants to see a small Type I error, whereas a drug developer of course wants to see a small Type II error, which means high power. Thomas and Juanes (1996) stated: "What constitutes 'high power' is best judged by the researcher . . . ." Another important statement is: "There are no agreed conventions as to what constitutes a biologically significant effect; this will depend upon the context of the experiment and the judgment of the researcher." That paper, which seems to have been intended to be a guide for researchers, contains some excellent practical advice.

1.3 SPORTS APPLICATIONS

Professional sports teams do a form of testing all of the time when they make decisions regarding individual players (making the team, being benched, etc.) Frequently, there are statements made in the media regarding the comparison of two players competing for a position based on a small sample size, with the latter discussed frequently in print. For example, the April 20, 2009 edition of the *The New York Times* had an article with the headline "Over the Wall and Under the Microscope in the Bronx," referring to the number of home runs (20) hit in the new Yankee Stadium during the first four games of the 2009 season, compared to only 10 home runs being hit in the first six games at Citi Field, the new home of the New York Mets. Regarding the latter, the chief operating officer of the Mets stated: "It's a small sample size . . . ."

Similarly, a player's batting average might be over .400 after the first week of a season, but he almost certainly won't hit over .400 for the entire season. The problem, of course, is the small sample size.

1.4 OBSERVED POWER, RETROSPECTIVE POWER, CONDITIONAL POWER, AND PREDICTIVE POWER

We should restrict our attention to thinking of power as being a concept that is applicable *before* the data have been collected. Unfortunately, the term "observed power" (see, e.g., Hoenig and Heisey, 2001) is used to represent the power *after* the data have been collected, acting as if parameter values are equal to the observed sample statistics. That is poor practice because such equality will rarely exist. Hoenig and Heisey (2001) stated that "observed power can never fulfill the goals of its advocates" and explained that this is because observed power is a 1:1 function of the p-value. Similarly, Thomas (1997) stated: "Therefore calculating power using the observed effect size and variance is simply a way of re-stating the statistical significance of the test." Since "power" is a probability and a p-value is a sample statistic, the latter cannot in any way be equated with the former. The practice of using retrospective power has also been debunked by Lenth (2001, 2012), who additionally used the term "retrospective power" in referring to observed power.

Various other prominent scholars have said the same thing, including Senn (2002), who stated that power is irrelevant in interpreting completed studies. The esteemed Sir David Cox also stated that power is irrelevant in the analysis of data (Cox, 1958). Zumbo and Hubley (1998) went further and stated (p. 387): "Finally, it is important to note that retrospective power cannot, generally, be computed in a research setting." Zumbo and Hubley (1998) explained that "we know that retrospective power can be written as a function of two unconditional probabilities. However, the unconditional probabilities are not attainable in a

research setting." They further stated: "We suggest that it is nonsensical to make power calculations *after* a study has been conducted and a decision has been made." Yuan and Maxwell (2005) found that observed power is almost always a biased estimator of the true power. Unfortunately, bad advice regarding this can be found in the literature. For example, in an editorial explaining what researchers should know about sample size, power, and effect size, Hudson (2009) stated: "If data is not available in the literature, then a pilot study is justified. If, for whatever reason, sufficient subjects are not recruited, then power can be conducted on a post hoc basis." Somewhat similarly, Thomas (1997) also presented retrospective power as an option and indicated that it can be useful for certain study goals.

Conditional power, as proposed by Lan and Wittes (1988), is prospective rather than retrospective in that it is essentially a conditional probability of rejecting the null hypothesis in favor of the alternative hypothesis at the end of a study period, conditional on the data that have been accumulated up to the point in time at which the conditional power is computed. It is applicable when data are slowly accruing from a nonsequentially designed trial. See also the discussion in Zhu, Ni, and Yao (2011) and Denne (2001), with the latter determining the number of additional observations required at the end of the main study to gain the prescribed power, conditional on the data that had been observed to that point. This is in the same general spirit as Proschan and Hunsberger (1995).

Predictive power, which takes all uncertainties into account, parts of which are ignored by standard sample size calculations and conditional power, might seem preferable, but Dallow and Fina (2011) pointed out that the use of predictive power can lead to much larger sample sizes than occur with the use of either conditional power or standard sample size calculations.

1.5 TESTING FOR EQUALITY, EQUIVALENCE, NONINFERIORITY, OR SUPERIORITY

In traditional hypothesis testing, as presented in introductory level textbooks and taught in introductory courses, the null hypothesis in comparing, say, two means or two proportions is that they are equal, which implies that the difference between them is zero. If there is no prior belief that one mean is larger than the other mean, the alternative hypothesis is that the means are unequal, with a directional alternative hypothesis used if there is prior information to suggest that one mean is larger than the other one.

The difference between the means in the null hypothesis need not be specified as zero, and in practice the difference almost certainly won't be zero, as was mentioned earlier in the chapter. Equivalence testing simply formalizes this approach, although it is really just a part—although an unorthodox part—of hypothesis testing. That is, with equivalence testing, the means from two

populations are considered to be "equivalent" if they differ only slightly, and of course this would have to be quantified, with the acceptable difference determined by the specific application.

Reeve and Giesbrecht (1998) stated: "Many questions that are answered with hypothesis testing could be better answered using an equivalence approach." The latter is used with dissolution tests and in other applications. Whereas equivalence tests are often presented as being different from hypothesis tests, they are really a form of a hypothesis test, as discussed later in this section. Bioequivalence testing, which is discussed in Chapter 10 in Chow, Shao, and Wang (2008), is an important part of hypothesis testing. That source also contains theoretical details and sample size computations for equivalence, noninferiority, and superiority in other chapters. There is also material on these topics in subsequent chapters of this book, including software capability and output.

Mathematically, if the absolute value of the difference of the two means is less than δ, then the means are considered to be equivalent. This is the alternative hypothesis, with the null hypothesis being that the absolute value of the difference between the two means is at least equal to δ. A few examples of equivalence testing, with the appropriate sample size formula, are given in later chapters.

As stated by Schumi and Wittes (2011), "non-inferiority trials test whether a new product is not unacceptably worse than a product already in use." Of course, the obvious question is: "Why would anyone be interested in a new treatment that is worse by any amount than the standard treatment?" The answer is that a new treatment that is only slightly worse than the standard treatment relative to the intended benefit of each may be less costly to produce and less costly to the consumer than the standard treatment, and might also have fewer side effects. See Pocock (2003) and Schumi and Wittes (2011) for additional information on noninferiority testing.

Since "noninferiority" thus essentially means "not much worse than" and the latter implies a one-sided test, noninferiority can be tested with either a one-sided hypothesis test or a one-sided confidence interval.

In superiority testing, the objective is to show that the new drug is superior, so the null hypothesis is $\mu_T - \mu_S \leq \delta$ and the alternative hypothesis is $\mu_T - \mu_S > \delta$. Of course, the objective is to reject the null hypothesis and accept the alternative hypothesis.

1.5.1 Software

Software for equivalence, noninferiority, or superiority testing is not widely available; nQuery Advisor has some capability, most of which is for *t*-tests. PASS, on the other hand, has over 20 routines for determining sample size for equivalence testing and about the same number for noninferiority tests.

REFERENCES

Berk, K. N. (2003). Discussion (of a paper by Hubbard and Bayarri). *The American Statistician*, **57**(3), 178–179.

Bland, J. M. and D. G. Altman (1994). Statistics Notes: One and two-sided tests of significance. *British Medical Journal*, **309**, 248.

Boniface, D. R. (1995). *Experiment Design and Statistical Method for Behavioural and Social Research*. Boca Raton, FL: CRC Press.

Casella, G. and R. L. Berger (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem (with comments). *Journal of the American Statistical Association*, **82**, 106–139.

Chow, S.-C., J. Shao, and H. Wang (2008). *Sample Size Calculations in Clinical Research*, 2nd ed. Boca Raton, FL: Chapman & Hall/CRC.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Philadelphia: Lawrence Erlbaum Associates.

Cox, D. R. (1958). *Planning of Experiments*. New York: Wiley.

Dallow, N. and P. Fina (2011). The perils with the misuse of predictive power. *Pharmaceutical Statistics*, **10**(4), 311–317.

Denne, J. S. (2001). Sample size recalculation using conditional power. *Statistics in Medicine*, **20**(17-18), 2645–2660.

Feinberg, W. E. (1971). Teaching the Type I and Type II errors: The judicial process. *The American Statistician*, **25**, 30–32.

Fidler, F. and G. R. Loftus (2009). Why figures with error bars should replace *p*-values: Some conceptual arguments and empirical demonstrations. *Zeitschrift für Psychologie [Journal of Psychology]*, **217**(1), 27–37.

Friedman, H. (1972). Trial by jury: Criteria for convictions, jury size, and Type I and Type II errors. *The American Statistician*, **26**, 21–23.

Goodman, S. (2008). A dirty dozen: Twelve *p*-value misconceptions. *Seminars in Hematology*, **45**(3), 135–140.

Groves, R. M., F. J. Fowler, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau (2009). *Survey Methodology*, 2nd ed. Hoboken, NJ: Wiley.

Gunst, R. F. (2002). Finding confidence in statistical significance. *Quality Progress*, **35** (10), 107–108.

Hahn, G. J. and W. O. Meeker (1991). *Statistical Intervals: A Guide for Practitioners*. New York: Wiley.

Hoenig, J. M. and D. M. Heisey (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician*, **55**(1), 19–24.

Howard, G. S., S. E. Maxwell, and K. J. Fleming (2000). The proof of the pudding: An illustration of the relative strengths of null hypothesis, meta-analysis, and Bayesian analysis. *Psychological Methods*, **5**(3), 315–332.

Hubbard, R. and M. J. Bayarri (2003). Confusion over measures of evidence (*p*'s) versus errors (α's) in classical testing. *The American Statistician* **57**(3), 171–178.

Hudson, Z. (2009). Sample size, power, and effect size—What all researchers need to know (editorial). *Physical Therapy in Sport*, **10**, 43–44.

Lan, K. K. G. and J. Wittes (1988). The B-value: A tool for monitoring data. *Biometrics*, **44**, 579–585.

Lazzeroni, L. C. and A. Ray (2012). The cost of large numbers of hypothesis tests on power, effect size, and sample size. *Molecular Psychiatry*, **17**, 108–114.

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician*, **55**(3), 187–193.

Lenth, R. (2012). Sample size determination using applets. Talk given at the American Statistical Association Conference on Statistical Practice. Orlando, FL. Feb. 18.

Loftus, G. R. (1991). On the tyranny of hypothesis testing in the social sciences. *Contemporary Psychology*, **36**(2), 102–105.

Loftus, G. R. (1993). A picture is worth a thousand p-values: On the irrelevance of hypothesis testing in the computer age. *Behavior Research Methods, Instrumentation and Computers*, **25**, 250–256.

Loftus, G. R. (2010). The null hypothesis. In *Encyclopedia of Research Design*, pp. 939–943. Thousand Oaks, CA: Sage Publications.

Lohr, S. L. (2010). *Sampling: Design and Analysis*, 2nd ed. Boston, MA:Brooks/Cole, Cengage Learning.

Mudge, J. F., L. F. Baker, C. B. Edge, and J. E. Houlahan (2012). Setting an optimal α that minimizes errors in null hypothesis significance tests. *PLoS One*, **7**(2), 1–7.

Murphy, K. R. (1990). If the null hypothesis is impossible, why test it? *American Psychologist*, **45**, 403–404.

Nester, M. R. (1996). An applied statistician's creed. *Applied Statistics*, **45**, 401–410.

Pocock, S. (2003). The pros and cons of noninferiority trials. *Fundamental and Clinical Pharmacology*, **17**, 483–490.

Proschan, M. A. and S. A. Hunsberger (1995). Designed extension of studies based on conditional power. *Biometrics*, **51**, 1315–1324.

Reeve, R. and F. Giesbrecht (1998). Dissolution method equivalence. In *Statistical Case Studies: A Collaboration Between Academe and Industry* (R. Peck, L. D. Haugh, and A. Goodman, eds.). Alexandria, VA: American Statistical Association and Philadelphia, PA: Society for Industrial and Applied Mathematics.

Rhoads, G. G. (1995). Reporting of power and sample size in randomized controlled trials. *Journal of the American Medical Association*, **273**(1), 22–23.

Ryan, T. P. (2007). *Modern Engineering Statistics*. Hoboken, NJ: Wiley.

Schumi, J. and J. T. Wittes (2011). Through the looking glass: Understanding noninferiority. *Trials*, **12**, 106.

Seaman, J. E. and E. Allen (2011). Not significant, but important? *Quality Progress*, 58–59 (August).

Sedgwick, P. (2011). Sample size and power. *British Medical Journal*, **343**, 5579.

Senn, S. J. (2002). Power is indeed irrelevant in interpreting completed studies. *British Medical Journal*, **325**, 1304–1304.

Shrout, P. E. (1997). Should significance tests be banned? Introduction to a special section exploring the pros and cons. *Psychological Science*, **8**, 1–2.

Smith, A. H. and M. N. Bates (1992). Confidence limit analyses should replace power calculations in the interpretation of epidemiologic studies. *Epidemiology*, **3**, 449–452.

Thomas, L. (1997). Retrospective power analysis. *Conservation Biology*, **11**(1), 276–280.

Thomas, L. and F. Juanes (1996). The importance of statistical power analysis: An example from *Animal Behaviour*. *Animal Behaviour*, **52**, 856–859.

Thomson, S. (2012). *Sampling*, 3rd ed. Hoboken, NJ: Wiley.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science*, **6**, 100–116.

Yuan, K. and S. E. Maxwell (2005). On the post hoc power in testing mean differences. *Journal of Educational and Behavioral Statistics*, **30**(2), 141–167.

Zhu, L., L. Ni, and B. Yao (2011). Group sequential methods and software applications. *The American Statistician*, **65**(2), 127–135.

Zumbo, B. D. and A. M. Hubley (1998). A note on misconceptions concerning prospective and retrospective power. *The Statistician*, **47**(2), 385–388.

EXERCISES

1.1. Explain what a p-value is. Why do you think there are so many misconceptions about what a p-value really is?

1.2. Why would you not want to have a hypothesis test that has high power (such as .90) for detecting a very small difference in two population means?

1.3. Why would a confidence interval be preferred over the corresponding hypothesis test when they can both be used to test a hypothesis and the results will agree?

1.4. Although a statistical null hypothesis is generally not true, assume that a particular null hypothesis *is* true. If $\alpha = .05$, what is the probability of a Type II error, or does it make any sense to talk about the probability of a Type II error for this scenario? What would be the approximate probability if the null hypothesis were for a single parameter whose mean differed from the hypothesized mean by approximately 10^{-5}, with the standard deviation of the estimator of the parameter being of the order 10^{-2}?

1.5. Perform the following exercise to verify what was stated in Section 1.2. Write appropriate computer code to generate 1000 samples of size 100 from the standard normal distribution (i.e., normal distribution with mean zero and standard deviation one) and construct a 95% confidence interval for each sample, using the form: $\bar{x} \pm 1.96\, s/\sqrt{100}$, with $\bar{x}$ denoting the sample mean and s the sample standard deviation. How many intervals would you expect to contain the population mean of zero? How many did contain zero? Comment.

1.6. An experimenter rejects a null hypothesis and explains that he was not surprised that happened since the probability of it happening was .05, the chosen significance level of the test. Comment on that statement.

1.7. Sedgwick (2011) provided a tutorial question in a journal column that involved power for a clinical setting. Four statements were given and the reader was asked to determine which statements, if any, were true. The one statement that was declared to be true was "Power is the probability of observing the smallest effect of clinical interest, if it exists in the population." One might view this as a "short answer," although it was adequate relative to the context. If you were trying to explain to someone what power is, in general, what must be added to this in order to have a full explanation?