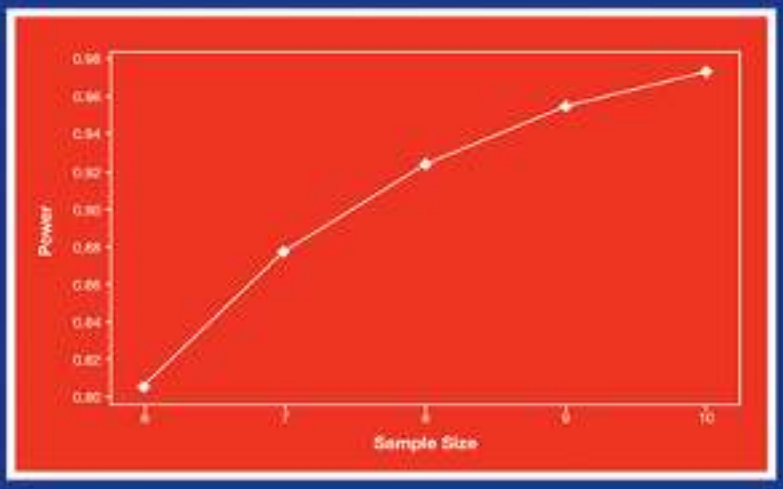
Wiley Series in Probability and Statistics

Sample Size Determination and Power



Thomas P. Ryan

WILEY

Sample Size Determination and Power

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

Editors: *David J. Balding, Noel A. C. Cressie, Garrett M. Fitzmaurice, Harvey Goldstein, Iain M. Johnstone, Geert Molenberghs, David W. Scott, Adrian F. M. Smith, Ruey S. Tsay, Sanford Weisberg*
Editors Emeriti: *Vic Barnett, J. Stuart Hunter, Joseph B. Kadane, Jozef L. Teugels*

A complete list of the titles in this series appears at the end of this volume.

Sample Size Determination and Power

THOMAS P. RYAN

Institute for Statistics Education, Arlington, Virginia and
Northwestern University, Evanston, Illinois

WILEY

Cover design: John Wiley & Sons, Inc.
Cover image: © Thomas P. Ryan

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

Ryan, Thomas P., 1945–
Sample size determination and power / Thomas P. Ryan.
p. ; cm.
Includes bibliographical references and index.
ISBN 978-1-118-43760-5 (cloth)
I. Title.
[DNLM: 1. Sample Size. 2. Clinical Trials as Topic. 3. Mathematical Computing. 4. Regression Analysis. 5. Sampling Studies. WA 950]
615.5072′4–dc23

2013000329

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Contents

Preface

Determining a good sample size to use in a scientific study is of utmost importance, especially in clinical studies with some participants receiving a placebo or nothing at all and others taking a drug whose efficacy has not been established. It is imperative that a large enough sample be used so that an effect that is large enough to be of practical significance has a high probability of being detected from the study. That is, the study should have sufficient *power*. It is also important that sample sizes not be larger than necessary so that the cost of a study not be any larger than necessary and to minimize risk to human subjects in drug studies.

Compared to other subjects in the field of statistics, there is a relative paucity of books on sample size determination and power, especially general purpose books. The classic book on the subject has for decades been Jacob Cohen's *Statistical Power Analysis for the Behavioral Sciences*, the second edition of which was published in 1988. That book is oriented, as the title indicates, toward the behavioral sciences, with the statistical methodology being quite useful in the behavioral sciences. The second edition has 567 numbered pages, 208 of which are tables, reflecting the "noncomputer" age in which the two editions of the book were written. In contrast, the relatively recent book by Patrick Dattalo, *Determining Sample Size*: *Balancing Power, Precision, and Practicality* (2008), which is part of the series in Pocket Guides to Social Work Research Methods, is 167 pages with more than 20% consisting of tables and screen displays reflecting the now heavy reliance on software for sample size determination. An even smaller book is *Sample Size Methodology* (1990) by Desu and Raghavarao at 135 pages, while *How Many Subjects*: *Statistical Power Analysis in Research* (1987) by Kraemer and Thiemann is just 120 pages and was stated in a review as being an extension of a 1985 journal article by Kraemer. *Sample-Size Determination* (1964) by Mace is larger at 226 pages and *Sample Size Choice*: *Charts for Experimenters*, 2nd ed. (1991) by Odeh and Fox is 216 pages. Thus, some rather small books have been published on the subject, with almost all of these books having been published over 20 years ago.

At the other extreme in terms of size, focus, and mathematical sophistication, there are books on sample determination for clinical studies, such as *Sample Size Calculations in Clinical Research*, 2nd ed. (2008) by Chow, Shao, and Wang, that are mathematically sophisticated, with the title of this book perhaps suggesting that. A similar recent book is *Sample Sizes for Clinical Trials* (2010) by Julious, whereas *Sample Size Calculations*: *Practical Methods for Engineers and Scientists* (2010) by Mathews is oriented toward engineering and industrial applications.

There are additional statistical methods that are useful in fields other than behavioral sciences, social sciences, and clinical trials, however, and during the past two decades new needs for sample size determination have arisen in fields that are part of the advancement of science, such as microarray experiments.

Although many formulas are given in Cohen's book, they are not derived in either the chapters or chapter appendices, so the inquisitive reader is left wondering how the formulas came about.

Software is also not covered in Cohen's book, nor is software discussed in the books by Mathews, Julious or Chow, Shao, and Wang. Software and Java applets for sample size determination are now fairly prevalent and, of course, are more useful than tables since theoretically there are an infinite number of values that could be entered for one or more parameter values. There was a need for a book that has a broader scope than Cohen's book and that gives some of the underlying math for interested readers, as well as having a strong software focus, along the lines of Dattalo's book, but is not too mathematical for a general readership. No such book met these requirements at the time of writing, which is why this book was written.

This book can be used as a reference book as well as a textbook in special topics courses. Software discussion and illustration is integrated with the subject matter, and there is also a summary section on software at the end of most chapters. Mixing software discussion with subject matter may seem unorthodox, but I believe this is the best way to cover the material since almost every experimenter faced with software determination will probably feel the need to use software and should know what is available in terms of various software and applets. So the book is to a significant extent a software guide, with considerable discussion about the capabilities of each software package. There is also a very large number of references, considerably more than in any other book on the subject.

THOMAS P. RYAN

Smyrna, Georgia
October 2012

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