6

Power and Sample Size Considerations in Molecular Biology

L. Jane Goldsmith

1. Introduction

Sample size is an important interest of researchers in laboratory and clinical settings. The number of cases to be investigated profoundly affects the cost and duration of a study. Sample size is estimated to achieve a certain statistical power and a careful power and sample size analysis can predetermine the success of a study or experiment.

1.1. What Is Statistical Power?

In the hypothesis test setting, statistical power is the probability of rejecting the null hypothesis when it is appropriate to do so, that is, when the null hypothesis is false. It is clear that it is desirable for statistical power to be high, representing a probability close to 1, because power is the probability of drawing the correct conclusion when the null hypothesis is false.

Statistical power is related to β, the probability of an error of type II. β is the probability that the hypothesis test in question will erroneously fail to reject H_0 when H_1, the alternative hypothesis, is true. It is easy to see that:

P[do not reject H_0 | H_1 is true] = β	Probability of an error
P[reject H_0 | H_1 is true] = statistical power = 1 – β	Probability of the correct conclusion

Statistical power is one of the central concepts in statistics. Many statistical methods are designed to increase power, and new statistical methods are often justified by claims that they are the most powerful or at least enhance power (***1***, pp. 60–63).

From: *Methods in Molecular Biology, vol. 184: Biostatistical Methods*
Edited by: S. W. Looney © Humana Press Inc., Totowa, NJ

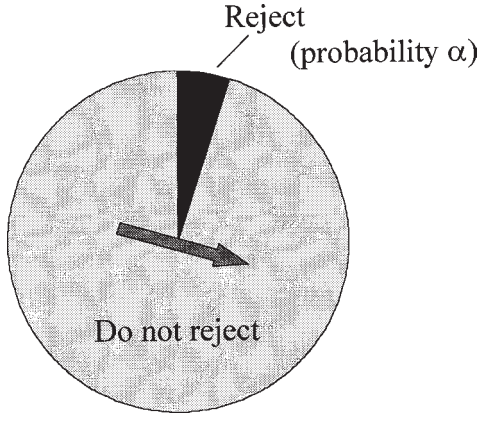

Fig. 1. The hypothesis test spinner.

Indeed, the reduction of the probability of error of type II, or, as outlined previously, the increase of statistical power, is the primary reason for using statistical methods and theory. This can be seen directly by observing that if one wants merely to control α, the probability of rejecting H_0 when it is true, one can merely use a uniform random number generator to generate a random number between 0 and 1. On occasions when the random number is ≤ α, reject the null hypothesis. Otherwise, do not reject. The probability of a spurious result (the type I error probability) is fixed at α. The statistical power is, unfortunately, also α. This method can also be criticized for the property that it is not based on measurements or data. *See* **Fig. 1** for a prototype random number generator, a hypothesis test spinner with $\alpha = 0.05$.

Using statistical theory, one can design a hypothesis test with the required α and high statistical power to detect whether or not the research hypothesis is true.

1.2. Power and Sample Size

Statistical power and sample size are often discussed in the same breath. Indeed, power increases with sample size in any "consistent" statistical test (***1***, p. 305). The desirable property of consistency in a statistical test is intuitively appealing: the more subjects or cases in the experiment or research study the more likely a correct conclusion. In such a consistent statistical test, "the more the merrier" slogan applies. That is, the more subjects, the more likely the statistical conclusion will be correct.

The hypothesis spinner in **Fig. 1** is an obvious example of a nonconsistent statistical test. No matter how many subjects are included in the experiment, the probability of rejection of the null hypothesis remains constant at α.

1.3. Other Uses for Sample Size Calculations

Sample size calculations are important in other areas of statistical inference, including determination of confidence interval widths and coverage probabilities. These are beyond the scope of the presentation in this chapter.

1.4. Importance in Research

Upon reflection, any scientist or experimenter can see the importance of correct or sufficient statistical power or sample size in an experiment. If the sample size is too small, and, consequently, the statistical power is too low, there is a high probability that the experiment will not detect the effect of interest. That is, even though the research hypothesis is true, the data collected under this experiment design will not yield a statistically significant test result. The researchers involved in such an experiment will be hard pressed to prove that the effect is *not* present, as their research design, because of its low statistical power, was not sensitive enough to detect an effect. In many cases journal editors will not accept a negative (not statistically significant) result for publication if the authors cannot demonstrate adequate sample size and statistical power.

It is easy to see that failure to plan for an experiment with adequate statistical power is risking that a negative conclusion will mean a waste of time and resources, resulting in research that is not meaningful and not publishable.

The convention for planning for statistical power is that an experiment should be designed to have power of 80% or 90%. A recent article uses Bayesian techniques to support different standards for α and β ***(2)***.

1.5. Ethics of Power and Sample Size

Many authors have written of the importance of sample size for efficient research ***(3–5)***. In biomedical experiments involving increased pain, discomfort, fear, or risk on the part of the subjects, it is apparent that it is entirely unethical to conduct an experiment with too small a sample size. If this is done, the subjects, human volunteers' or helpless animals, have suffered to no avail. Implicit in informed consent is the notion that the research is conducted efficiently. If the sample size and statistical power are wastefully high, then extra subjects have suffered needlessly. In studies involving suffering or sacrifice, the "Goldilocks Principle" applies for sample size: not too big and not too small, but just right.

Even in research where no suffering is involved, resources, money, and the valuable time of researchers and human subjects can be wasted with the wrong sample size.

2. Data Considerations

2.1. Types of Data

In molecular biology and other types of biological research, different types of data may be collected for research purposes. These different data types are summarized below:

1. *Categorical* data are data indicating group membership. Natural examples are allele type, race, country of birth, and religion.

   An important special class of categorical data arises when there are exactly two categories. These data, often characterized as "yes/no" data, are called *dichotomous* data. Natural examples are presence of a trait (yes/no), presence of an allele (yes/no), death (yes/no), evidence of infection (yes/no), and cure (yes/no).

   Categorical data are often coded as integer values, corresponding to group numbers that are associated with specific groups. Some statistical software packages, such as SAS ***(6)***, allow character data, or names, to represent categorical data. It is often advisable to code "yes/no," or dichotomous data, as the integers 1 or 0, with the value 1 denoting the occurrence of the event of interest ("yes") and 0 denoting the absence of the event of interest ("no"). This coding scheme sets things up for logistic regression, a statistical method sometimes used to search for predictors of dichotomous events.
2. *Ordinal* data are data reflecting a natural ordering. Ordinal data, however, do not have a "distance" measure associated with them. Examples of ordinal data in medical research include health index measures such as APGAR scores (used to indicate health of neonates), cancer stage (used to indicate the severity or extent of disease in cancer), and anesthesia class (used to indicate general overall health of a surgical patient). In each of these examples, subtraction (computing a "distance" between values) does not make sense. That is, a newborn with an APGAR of 9 is not "2 better" than an infant with an APGAR of 7. It is just known that the baby with APGAR 9 appears healthier at birth than the one with a score of 7, and that a baby with APGAR 8 would rank between them in terms of apparent health.
3. *Numerical* data reflect ordering and distance. Numerical data can be integer data, such as parity (number of live births), number of teeth, number of lesions, and so forth. Medical measurements, such as hematocrit, oxygen saturation, and systolic blood pressure, are examples of numerical data described as *interval-level* data.

2.2. Switching Between Data Types

In biomedical research it is common for data collected naturally or initially as one of the data types described previously to be transformed into data of a different type. Grouping subjects into age-group categories is an example of transformation of numerical data (age) into ordinal data (age group). Statistics programs such as SPSS make this transformation easy ***(7)***.

Often data are dichotomized, that is, changed from numerical data into high–low categories. "Cut-points" provide the boundaries used to change

numerical data into dichotomous data. Biomedical language recognizes this ubiquitous transformation with special nomenclature: for example, the terms *premature*, *hypoglycemic*, and *anemic* all represent dichotomizations of numerical measurements. The description "natural" dichotomous data was used above to indicate that those examples do not represent dichotomizations, but naturally occurring dichotomies of interest.

Another type of transformation involves a progression from dichotomous to numerical data. For example, a sequence of 1 or 0 (yes or no) answers to questions from a survey or checklist can be summed into a numerical score.

Categorical data can sometimes be found through experience and statistical testing to have a natural ordering, thus converting it to ordinal or even numerical type. An example here is the staging system of cancer. Developed as a systematic method of describing the extent of disease at diagnosis, the tumor-node-metastatis (TNM) system has a natural ordering related to the natural progression of the disease, and the stages have been shown to be negatively correlated with survival probability (***8***, pp. 3–5). Ongoing efforts to refine the staging system ascertain that the addition of new stage definitions explain survival meaningfully (***8***, p. 12). Cancer stage at diagnosis is often considered an ordinal or numerical variable.

2.3. Statistical Power, Data Type, and Strategies for Efficiency

Different kinds of statistical methods are appropriate for each data type. Contingency tables, chi-square (χ^2) tests, log-linear models, and logistic regression are among the methods used for categorical and dichotomous data. In many cases, categorical data and associated statistical methods require large sample sizes.

Ordinal data are often analyzed by nonparametric statistical methods. These methods can be very efficient in terms of statistical power and sample size. However, often the most powerful statistical methods are parametric tests on numerical data.

Numerical data represent the most informative data in biomedical research. Statistical theory and experience have indicated that, in general, research utilizing numerical data with parametric statistical methods affords the most efficient statistical power and sample size. Authors recognizing this relationship between the information in data and efficient research deplore the practice of dichotomization or categorization described above as a waste of resources ***(9–11)***.

In summary, the current best advice is to eschew dichotomization or categorization in data collection. Record and analyze numerical data when possible.

Further advice is to use strategies converting naturally occurring dichotomous or ordinal data to numerical data whenever large samples are not practicable. One such strategy is to sum dichotomous responses, as mentioned

previously. Another is to record survival time or time elapsed until the occurrence of an event of interest rather than the simple, relatively uninformative outcome of event/no event. An example here is the notation of age of onset of a disease. Earlier onset may be predicted by a genetic marker, indicating a genetic link to the disorder.

A recent method involves substituting a numerical outcome or "surrogate marker" for dichotomous data. An example is monitoring CEA levels to detect increase in tumor burden rather than monitoring cancer patients for a recurrence of their tumor, a yes/no variate. Recent statistical studies grapple with the problem of determining when surrogate markers are justified ***(12–14)***.

3. Experiment Design Considerations

An important concept in understanding statistical power and sample size calculations is that each statistical method has its own associated sample size formula or calculation method. In the early part of the 20th century, when inferential statistics was a new discipline, sample size calculation was rare. In the days of hand-cranked calculators, large samples were a computational burden. "Rule of thumb" methods were often used, with 30 as a popular number ***(15)***. As the importance of careful power and sample size determination as part of the research planning process became more recognized, researchers would try to use or adapt simple formulas for more complicated analyses. For example, a sample-size formula for 80% power for a two-sample independent-group *t*-test would be simply doubled for a four-group one-way analysis of variance (ANOVA). For an analysis of covariance (ANCOVA), the effect of the covariate would be ignored or assumed to increase power (an unwarranted assumption, in many cases). In a stratified design, the stratification would be ignored to use simple sample size methods. With the passage of time and the expansion of the statistical literature, more complicated sample size and power calculation methods have become known. In recent years, statistical software has become available that allows calculation for some complex designs.

Appropriate sample size and power methods for complicated analyses allow more precise accurate estimates of necessary sample size. Sometimes the proposed sample sizes are larger than those using simpler methods and sometimes they are smaller. The good news is that they allow for the most efficient research.

4. Effect Size

By definition, statistical power is the probability of rejecting H_0, the null hypothesis, in favor of H_1, the alternative hypothesis, when H_1 is true. It is intuitively obvious that if H_1 represents a large distance or difference from the situation under H_0, the sample size needed to detect a statistically significant difference at level α would be small. Conversely, if H_1 represents a small difference from H_0, the required sample size would be large.

For example, if the null hypothesis is that elephants weigh on average the same as mice, we would not need many specimens of elephants and mice to detect a statistically significant difference. Elephants and mice are far apart in terms of weight. However, to detect a difference in mean weight between two strains of mice, researchers would need a large sample of each type. The tiny difference in mean weight between two strains of mice would yield a very small effect size.

The term "effect size" has been used to signify the critical difference to be detected. In more formal usage, effect size defines a formula for the difference between H_0 and H_1 that is useful for sample size calculation. For example, for a two-group independent-sample *t*-test, the effect size is $(\mu_1 - \mu_2)/\sigma$, the difference in means in terms of the common standard deviation. This effect size, often called *d*, appears in power and sample size formulas. Obviously, different versions of sample size formulas for the same method may call for differing effect size formulas, but if differing effect sizes and their associated formulas are used correctly, they will yield identical sample sizes. Some common effect size formulas are given in **Table 1**.

From the formulas in **Table 1**, it is apparent that effect size, for each sample size calculation, represents a "distance" from the null value. The sample size for fixed power (say, 80%) will be larger for smaller effect sizes and smaller for larger effect sizes. This relationship is usually not linear, but of course depends on how effect size is represented in the sample size formula.

Figure 2 contains a graph of the relationship between effect size and sample size needed for 80% power in a one-sample *t*-test. As demonstrated in this plot, sample size varies extremely as the effect size increases from 0.1 to 1.0.

5. Steps in Sample Size Calculations

The following steps are a general guideline for sample size calculations.

1. Determine, in conjunction with other researchers, the research question. This may be very vague at first, but an attempt should be made to make it specific and quantitative. For example, the initial research question "Does hunger affect mood?" might be quantified to "How is blood glucose level related to mood, as measured by the Beck Depression Inventory?" ***(16)***.
2. Reword the research question into a research hypothesis. Our glucose example hypothesis might be: "Glucose level is negatively correlated with Beck Depression Score."
3. Reverse the research hypothesis to form a null hypothesis. For example, in our glucose study:

$$H_0: \rho = 0$$

where ρ is the correlation between blood glucose and BDI score. Note that we have used a two-sided hypothesis, rather than the one-sided hypothesis suggested by the research hypothesis. Many researchers and journal editors prefer a two-sided hypothesis, which is considered to be more objective and conservative than a one-sided hypothesis.

Table 1
Some Effect Size Formulas

Statistical method	H_0	H_1	Effect size
One-group *t*-test or *z*-test	$\mu = 0$	$\mu = \mu_1 \neq 0$	μ_1/σ
Two-group *t*-test or *z*-test	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$(\mu_1 - \mu_2)/\sigma$
Linear correlation	$\rho = 0$	$\rho = \rho_1 \neq 0$	ρ_1
K-group ANOVA	$\mu_1 = \cdots = \mu_k = \mu$	some difference among μ_i's	$\frac{\left(\sum_1^k (\mu_i - \mu)^2/k\right)}{\sigma^2}$
Contrast in *K*-group ANOVA	$\sum_1^k c_i \mu_i = 0$	$\sum_1^k c_i \mu_i \neq 0$	$\frac{\left\lvert\sum_1^k c_i \mu_i\right\rvert}{\sigma \sqrt{\sum_1^k c_1^2}}$

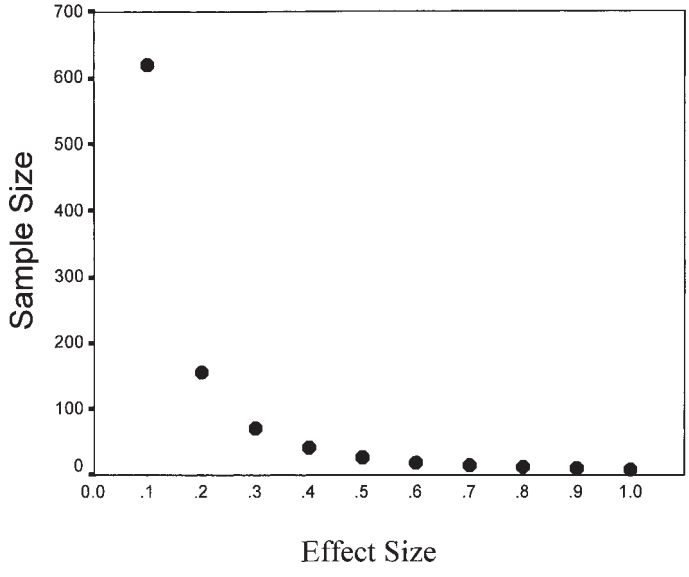

Fig. 2. Relationship between effect size and sample size needed for 80% power in a one-sample *t*-test.

4. Determine the significance level for the study. In most cases, $\alpha = 0.05$ will be the chosen level of significance, although some researchers choose $\alpha = 0.01$ or $\alpha = 0.02$ as a more conservative significance level.
5. Determine the statistical power to be used. As mentioned above, 80% or 90% is considered acceptable for most research studies.

6. Determine an important difference or distance for detection in the research study. A correlation of 0.3 or more, say, would be important to recognize in the glucose study.
7. Determine the statistical method to be used. Use the important difference proposed in **step 6**, and other estimates such as standard deviation, if necessary, to compute effect size. In the proposed glucose study, a correlation analysis is proposed, so the effect size is $\rho = 0.3$.
8. Calculate the sample size for the proposed study. The appropriate sample size for our correlation study is 85. Some methods of calculation are summarized in the paragraphs that follow.

At this point the initial sample size has been calculated. This critical number must be scrutinized.

5.1. Is the Calculated Sample Size Too Large to Be Practical?

Can we recruit this many subjects in a reasonable time frame? Can we afford to make this many expensive measurements?

If the sample size is unsatisfactorily large, an effort can be made to reduce it by alteration of the experiment plan. Researchers should investigate one or more of the following modifications, all of which are designed to decrease the required sample size:

1. Raise the level of significance (from, say, 0.01 to 0.05).
2. Decrease the desired power level (from, say, 90% to 80%).
3. Increase the effect size to be detected. A caveat here is that effect size should not be made *too* large to reduce the required sample size. For example, a correlation of 0.9 is almost never found in nature, although this effect size leads to a very small calculated sample size in correlation studies. An experiment designed to detect only unrealistically large differences will not be respected or valued.
4. If feasible, switch to a one-tailed test, which, in most cases, more powerful than its two-tailed counterpart. It is sometimes possible, on scientific grounds, to justify a one-sided test, although, as mentioned previously, a two-sided test is preferred by many.
5. Consider a redesign of the study, using more informative data, as discussed earlier. Reverting to original measurement data before cut-points are taken or substituting a research question involving a surrogate variable can lead to a substantial reduction in the required sample size, cutting the size in half or even reducing it 10-fold (***11***, pp. 275–276).
6. Consider a redesign of the experiment, using a different statistical method. For example, using repeated measures, crossover designs, or incorporating covariates for variance reduction can sometimes achieve significant savings in the required sample size ***(17)***.

All of these modifications must lead to a recalculation of the required sample size, which will, it is hoped, be reduced sufficiently for the experiment to be practicable.

Even if the initial stages of sample size determination yield sample sizes that are achievable and satisfactory to the researchers, it is often advisable to calculate sample sizes for several experiment designs. Each design, such as pre–post, crossover, repeated measures, and so on has advantages and disadvantages in research and sample sizes will vary. A careful, thoughtful sample size calculation will lead to the most efficient, informative research ***(18)***.

If all attempts to address the research question with an experiment design with a practicable sample size fail, then the experiment should be abandoned. To proceed with an experiment whose sample size is insufficient to achieve a respectable level of statistical power is wasteful of resources and, in some instances, unethical.

5.2. Is the Calculated Sample Size Too Small?

This may seem an unlikely problem, but sometimes, particularly with interval measures, repeated measures, and correlation studies, the sample size from the initial calculation would not impress the peer reviewers or readers of the scientific literature. If the statistical method is an asymptotic one, such as a χ^2 test based on the normal approximation, it may be that the calculated small sample size would not be sufficient to provide accurate statistics and provide accurate approximations. If the sample size is too low, researchers should consider or discuss sample size recomputation using a smaller effect size, lower α-level, or higher statistical power. This might seem wasteful, but if correct research results cannot be published or respected, then the experiment has been a waste in terms of contributing to the body of knowledge.

Finally, it is advisable to boost the sample size slightly to allow for failed experiments or loss-to-follow-up. A rule of thumb is to add 10% for loss-to-follow-up (***19***, p. 1), but each laboratory or researcher should estimate this number from past experience in the research setting. Some sample size software (for example, Power and Precision, discussed in the following section) allows for built-in attrition adjustments.

6. Methods for Sample Size Calculations

6.1. Formulas

In some cases, a closed-form formula, usually involving the appropriate effect size, is available. Lachin ***(20)*** presents numerous formulas for sample size calculations. Biostatistics textbooks, such as those by Zar ***(21)*** and Dawson-Saunders and Trapp ***(22),*** as well as clinical trial books ***(23,24)*** contain sample size formulas. Friedman et al. (***24***, pp. 125–129) contains an extensive bibliography of articles with sample size methods.

Matrix-based formulas for power calculations for linear models appear in ***(25)***. These formulas can be implemented using any computer-based matrix

language that also incorporates functions for the noncentral F-distribution. SAS IML® is an example of such a language.

If one is faced with a power analysis for a new or complicated method, sometimes a search through the Current Index to Statistics ***(26)*** will yield a reference to an article detailing the appropriate power analysis. In some cases, power and sample size tables will be a part of the original article describing a new method.

6.2. Software

In recent years, several excellent computer packages have been developed for sample size and power calculations. The early versions of these programs provided power and sample size calculations mainly for simple models. The programs are under continuous development, improving user friendliness and the range of statistical methods covered. Up-to-date descriptions of the programs, as well as new programs not mentioned here, can be found on the World Wide Web. Some sample size and power programs and the software companies are:

nQuery ***(27)***, an easy-to-use, versatile system	Statistical Solutions www.statsol.ie/nquery/nquery.htm
PASS (Power and Sample Size) ***(28)***, user-friendly system with good graphics. Includes group sequential clinical trials	NCSS, Number Cruncher Statistical System www.ncss.com/
SamplePower ***(29)***	SPSS, Inc. www.spss.com/spower/
Also marketed as Power and Precision ***(30)*** User-friendly, especially good for survival analysis, survival analysis, excellent graphics	Biostat www.powerandprecision.com

Other specialized programs that have some sample size calculation capabilities are:

EAST ***(31)***, group sequential clinical trials	Cytel Software Corporation www.cytel.com
Egret Siz ***(32)***, Cox regression and epidemiological models	Cytel Software Corporation www.cytel.com
Epi Info ***(33)***, free downloadable epidemiological software	Centers for Disease Control and Prevention www.cic.gov/epiinfo

For the latest information regarding software capability and availability, visit the appropriate website or contact the individual or institution distributing the software.

Two excellent freeware programs for complicated linear models, multivariate linear models including repeated measures, and other sample size problems are:

UnifyPow.sas	Ralph G. O'Brien, Cleveland Clinic Foundation www.bio.ri.cef.org/power.html
IML Power Program	Lynette L. Keyes and Keith E. Muller University of North Carolina ftp://ftp.uga.edu/pub/sas/contrib/cntb0014/

These programs, using SAS ***(6)*** macros, is are described in **ref. *34***.

Increasing software development is resulting in more programs for sample size determination. As of this writing, some sample size calculations are available on the World Wide Web. Using a search engine, it is possible to find websites that use Java Applets and other software to perform free sample-size calculations in real time.

A caveat with any computer program is to test it thoroughly if the methods and results upon which it is based have not been published. Testing can be accomplished by checking agreement with hand calculations, other software, or tables. All programs have disclaimers, absolving the authors and corporations involved in distribution of the software of any liability in case the results of the program are erroneous. The cost of an experiment with the wrong sample size will not be borne by software vendors or creators.

6.3. Tables

Books and journal articles with sample size and power tables are widely available. A list of books (and authors) with sample size tables appears below:

Statistical Power Analysis for the Behavioral Sciences ***(5)***, introduction and basic sample size calculations	Jacob Cohen
How Many Subjects? ***(9)***, basic sample size tables and explanations	Helena Chmura Kraemer and Sue Thiemann
CRC: Guide to Clinical Trials ***(19)***, concentrating on survival analysis	Jonathan J. Shuster

6.4. Nomograms

The following resources provide useful graphs for sample size determination:

Sample Size Choice ***(35)***, for ANOVA models	Robert E. Odeh and Martin Fox
"Nomograms for Calculating the Number of Patients Needed for a Clinical Trial with Survival as an Endpoint" ***(36)***	David A. Schoenfeld and Jane R. Richter

6.5. Simulation

Simulation is the method of last resort for sample size calculation. This labor-intensive method involves writing computer programs for empirical, estimated power calculations. Simulation is sometimes necessary when the statistical method is very complex or when sample size formulas, software, tables, or nomograms are not yet available for a particular statistical method. Statistics is an academic discipline, and academic statisticians who develop new statistical methods will often publish articles and papers describing the methods as soon as they are developed and proven. Quick publication is desirable to make the method available for use as soon as possible and also to establish priority and to obtain academic credit for publication. A later article, by the same author or perhaps by another author, may detail the sample size calculation. In the absence of formulas or tables, power and sample size estimates may come from simulation, using a computer package or language. The steps for simulation are outlined as follows:

1. Decide on the null hypothesis, a statistical method, the levels for α and power, the alternative hypothesis reflecting the important difference of interest, and an initial sample size "guestimate" n.
2. Write a computer program to generate data sets of size n according to the distribution described by the alternative hypothesis. A rule of thumb here is to generate at least 1000 datasets of the appropriate sample size, although more would mean more accurate estimates, of course. With this simulation we are estimating a proportion, $1 - \beta$, the statistical power under H_1.
3. For each dataset, compute the statistical test of H_0. Keep a tally of the number of tests that reject and the number that fail to reject.
4. Calculate the percentage of the datasets that lead to rejection. This is the power estimate for sample size n for this distance from the null hypothesis. If power is too low, adjust n upward and repeat the simulation steps above.
5. Continue this process until a satisfactory sample size and power are achieved. The Statistical statistical packages SAS and S-Plus have been used for simulation studies, as well as programming languages such as FORTRAN and BASIC (***37***, p. 139).

7. Special Topics in Power and Sample Size Analysis

7.1. Achieved Power

Achieved power is a power estimate based on the results of a study. That is, the data in the study are used to generate an estimate of effect size and statistical power. Often, achieved power is not very informative. It is usually very high when the experiment is statistically significant. Indeed, some researchers opine that achieved power must equal 1 if a study is statistically significant, observing that power is the probability of rejecting the null hypothesis and the null hypothesis has already been rejected in the study ***(38)***.

However, an achieved power not equal to 1, but usually quite high, can be estimated using the data. SPSS *(7)* prints achieved power for some analyses on request.

If the null hypothesis is not rejected, the achieved power may be very low (approaching α) or attain some middle value. This power estimate is a rough evaluation of the adequacy of the sample size of the study if the difference or distance from H_0 that has been observed is approximately equal to a difference of research interest. That is, if the data show an important research effect, but it is not statistically significant, the achieved power will give an indication of how far from the desirable power of 80% or 90% this sample size is. If the power is low even when an important effect is demonstrated in the descriptive statistics calculated from the data, the sample size is probably far from adequate.

Achieved power in general is a biased, inflated power estimate. If a new experiment is to be planned using the present results, it is best to adjust the sample size estimates to compute obtain an unbiased estimate of statistical power and sample size for the new study (***39***, pp. 405–416).

7.2. Post Hoc Power

Post hoc power is the term used to describe power computed after the completion of an experiment. Researchers use some of the experiment results (say, observed standard deviations, correlations, or variances) to compute the power to detect an important, conjectured difference. Achieved power, discussed previously, is thus a special case of *post hoc* power, one in which the effect size is also estimated from the completed experiment.

When an experiment is not statistically significant and power and sample size were not calculated carefully before execution of the study, power can be estimated after the fact to determine if the experiment was sensitive enough to detect a difference of value to the researchers. Power computed *post hoc* is late, at best, and the situation is tantamount to scientific fraud if researchers report that their experiment was carefully planned when it was not. Some researchers deplore *post hoc* power, while some believe it has value to salvage nonsignificant results for reporting purposes if the sample size of an unplanned study happens by chance to have been adequate for a reasonable alternative hypothesis.

Another interpretation is that the nonsignificant experiment now functions as a pilot study, and the *post hoc* power calculation is the first step toward designing a new study of appropriate sample size (***9***, p. 25).

7.3. Pilot Studies

When estimates needed for power and sample size calculations are not available from the literature, from existing databases, or from previous experience

in the laboratory or clinic, a pilot study may be in order to develop information needed for adequate estimation. The importance of pilot studies has been recognized in some research institutions by the provision of intramural grants to help new researchers and by special pilot study grant programs in United States federal granting agencies. For example, the National Institute of Mental Health has provided special grant support to develop effect sizes to improve future research.

These are the hallmarks of a good pilot study:

1. A pilot study should be small. A large pilot study is an oxymoron, as a large, expensive study should answer the research question, not just help to plan another study. Careful planning of a pilot study to provide a tight (narrow) confidence interval for a variance can lead to a larger sample size than would be feasible for the ultimate research question! Common sense must be the guide.
2. The pilot study should measure various outcomes, as researchers may find that a surrogate marker, described previously, may be more suitable and economical for their research question. Time to occurrence of an event of interest is sometimes estimated in pilot studies for planning purposes.
3. The pilot study should obtain all estimates needed for power and sample size analysis for any feasible research design. These estimates usually include sample means and standard deviations, but they also should include estimated correlations for repeated measurements if this might be a possible experimental design. Estimated correlation between measures or repeated measures has a large effect on sample size (***17***, p. 41).

7.4. Grant Applications

At the 1997 Joint Statistical Meetings of the American Statistical Association, Ralph O'Brien of the Cleveland Clinic led a group of discussants in a seminar entitled "Statistical Grantsmanship." Among the suggestions for applications seeking national funding were the following:

1. Use of sophisticated power and sample size calculations, appropriate for the ultimate planned statistical analysis.
2. Use of appendices in grant applications for long sample size formulas and theory.

Other suggestions for optimal power and sample size calculations for grant applications include:

1. Use of previous knowledge gained from intramural studies and pilot studies. Known colloquially as "sweat equity," the work invested in well-designed pilot studies can give grant applicants an advantage over others in the selection process.
2. Sensitivity analyses to demonstrate adequate statistical power if the assumptions in the primary power analysis are not met. For example, it would be valuable to demonstrate that the planned study affords reasonable power for a range of variance estimates. It is also wise to investigate several alternative designs to choose the best one (***18***, p. 1209).

3. A summary table at the end of the power analysis section of a grant application can summarize the value of n chosen, the power and sensitivity for primary hypotheses, and the estimated power for secondary hypotheses. Sometimes power for secondary hypotheses will not be high, owing to budget constraints.

7.5. Cost of Power Analysis

The length of this chapter, which merely outlines power and sample size considerations in planning a study, is an indication of the complexity of this subject. It is not unusual for the tasks of power and sample size estimation for complicated research plans or for designs using new statistical methods to require weeks of work for statisticians and researchers. It is important to allocate adequate time and resources for sample size calculation (***18***, p. 1224).

8. Examples

8.1. Segregation Analysis for Codominant Loci

This problem is suggested by an example in *Statistics in Human Genetics* by Pak Sham ***(40)***. In this genetic problem, individuals with heterozygous inheritance at the locus are phenotypically different from homozygous individuals. Mendelian inheritance implies that the three phenotypes will appear as offspring from two heterozygous parents in the proportions 1/4, 1/2, 1/4, where the 1/4 fractions represent homozygous offspring and the 1/2 fraction represents heterozygous inheritance. The research question is, "Do the proportions in offspring of heterozygous parents differ from those expected in Mendelian inheritance?" The research hypothesis is, "Proportions of the three phenotypes differ from Mendelian inheritance." We have H_0: p_1= 1/4, p_2 = 1/2, p_3 = 1/4. We wish to detect a "difference" reflected by p_1= 0.3, p_2 = 0.4, p_3 = 0.3. We use nQuery ***(27)*** and select the program "Chi-square test of specified proportions in C categories." We choose $\alpha = 0.05$ and power = 0.80 and type these numbers into the input table along with the number of categories, $C = 3$. nQuery has an effect size calculator, so we enter the null hypothesis proportions and the alternative hypothesis proportions and obtain the effect size Δ^2 = 0.04. The required sample size is $n = 241$.

8.2. Time to Onset of Disease in a Small Animal Model

In laboratory mice an allele has been identified that is associated with the onset of cancer. Researchers can induce a certain type of cancer by introduction of a virus. Animals will be tested for presence of the suspected allele. Then the virus will be injected into each mouse. Subjects will be observed daily for onset of cancer. We anticipate that the population of mice without the suspect allele will have a median age of onset of disease of 60 d. Mice with the allele are suspected to experience earlier onset, median = 45 d. Animals will

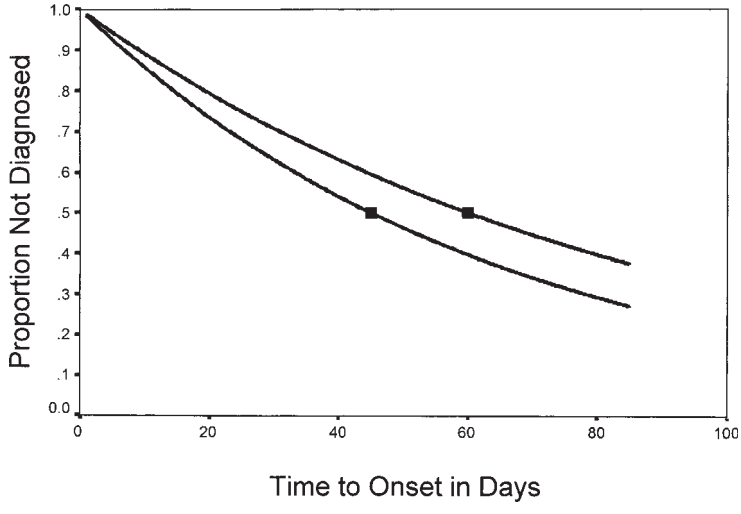

Fig. 3. Hypothesized survival curves and the median survival values for time until onset of disease.

be observed for 85 d. **Figure 3** presents the hypothesized survival curves and the median survival values.

The research question is, "Do subjects with the suspect allele tend to experience earlier onset of disease?" The research hypothesis is, "Animals with the suspect allele tend to experience earlier onset of disease." The null hypothesis is H_0: $S_1 = S_2$, where S_1 represents the survival curve for subjects with the allele and S_2 is the survival curve for subjects without the allele. We have chosen a two-sided hypothesis using the log-rank test.

We choose first to use nomograms for survival curves ***(36)***. We compute R = ratio of median survival times = 60/45 = 1.33. Using **Fig. 1**, p. 164, in **ref. *36***, for power = 0.8 and $\alpha = 0.05$, we draw lines according to the instructions for accrual = 0 (we will follow all animals for 85 d) and follow-up period = 1.6 ((85/.5 [(45 + 60])/2]). We obtain 140 per group for $R = 1.5$, using the dashed line for two-sided hypotheses. Turning to **Fig. 3** of the paper, p. 166, in **ref. *36***, we perform an adjustment to obtain a sample size for $R = 1.33$: 300 per group. Thus, using the nomograms, we obtain $n = 600$.

Using the *CRC Handbook* ***19***, we must first compute the proportion without disease at 85 d. We compute λ, the hazard rate for exponential survival, according to the formula $\lambda = -(\ln 0.5)/(\text{median survival})$. We obtain $\lambda_1 = 0.0154$ and $\lambda_2 = 0.01155$. Using the formula $e^{-\lambda t}$ for the proportion not diagnosed at time t, we obtain 0.27 for the allele group and 0.37 for the controls (no suspect alleles) at $t = 85$. On p. 612 of **ref. *19*** for ALPHA = 0.025 (one-tailed, implying

0.05 ALPHA for two-tailed), PCONT = 0.25, DEL = 0.37 – 0.27 = 0.10, using FACT = 0.00 for up-front accrual, we find $n = 579$. For PCONT = 0.30, we find $n = 665$. Using linear interpolation, we have $n = 613$.

Finally, using the nQuery ***(27)*** program "Log-rank test of survival in two groups followed for fixed time, constant hazard ratio," and using the built-in parameter calculators to compute the λ's, we obtain 309 per group, or $n = 618$.

We have obtained similar sample sizes using three methods. All sample size methods hypothesized exponential survival. The sample size program nQuery was the easiest to use. An adjustment upward for loss-to-follow-up would be at the discretion of the researchers. Further refinements should be made if the proportions with and without the suspect allele are far from 0.50.

8.3. Validation of New Assay Method

Laboratory researchers wish to use a new, simpler assay method. They wish to establish that the new method affords a satisfactory level of accuracy. They decide to use Lin's concordance coefficient, which is preferable to correlation or *t*-test comparisons ***(41)***. (*See* also Chapter 5 by Stephen W. Looney for further discussion of this issue.) Lin's original article defining the concordance coefficient was followed by another paper outlining sample size estimation and giving tables for sample size ***(42)***. Using the guidelines in Lin's second paper **ref.** ***42***, the researchers determine that they expect, under ideal conditions, that the new assay will explain 98% of the standard assay. They decide they can tolerate a 1% reduction in precision, a 12.5% location shift, and a 10% scale shift. Using the tables on p. 602 of ***(42)***, they determine the minimum acceptable concordance, $\rho_{c,a}$, to be 0.972 and that a sample of 41 paired assays will be required for 95% power.

Acknowledgments

Thanks to Ralph O'Brien of the Cleveland Clinic for review of this chapter and excellent suggestions. Grateful appreciation also to Stephen Looney, my colleague at the University of Louisville, for his leadership and encouragement. Any remaining mistakes are my own.

References

1. Lehman, E. L. (1959) *Testing Statistical Hypotheses*. John Wiley & Sons, New York.
2. Lee, S. E. and Zelen, M. (2000) Clinical trials and sample size considerations: another perspective. *Statist. Sci.* **15,** 95–100.
3. Altman, D. G. (1994) The scandal of poor medical research. *Br. Med. J.* **308,** 283–284.
4. Freiman, J. A., Chalmers, T. C., Smith, H., Jr., and Kuebler, R. R. (1978) The importance of beta, the Type II error and sample size in the design and interpretation of the randomized control trial. *N. Engl. J. Med.* **20,** 690–694.

5. Cohen, J. (1987) *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Lawrence Erlbaum, Hillsdale, NJ.
6. *SAS*, Statistical Analysis System, SAS Institute Inc., SAS Campus Drive, Cary, NC 27513.
7. *SPSS*, Statistical Package, for the Social Sciences, 233 South Wacker Drive, Chicago, IL 60606.
8. Beahrs, O. H., Henson, D. E., Hutter, R. V. P., Kennedy. (ed.) (1993) *Handbook for Staging of Cancer*. J. B. Lippincott, Philadelphia.
9. Kraemer, H. C. and Thiemann, S. (1987) *How Many Subjects? Statistical Power Analysis in Research*. SAGE, Newbury Park, CA, pp. 81–83.
10. Cohen, J. (1983) The cost of dichotomization. *Appl. Psychol. Meas.* 78, 240–253.
11. Goldsmith, L. J. (1995) Pros and cons of cutpoints, in *Proceedings of the Biometrics Section, Papers presented at the Annual Meeting of the American Statistical Association*, Orlando, FL, August 13–17, 1995, pp. 272–276.
12. Hughes, M D. (1999) The use and evaluation of suggogate endpoints in clinical trails, in *54th Deming Conference on Applied Statistic*s, Atlantic City, NJ, December 8, 1999.
13. Lefkopoulou, M. and Zelen, M. (1995) Intermediate clinical events, surrogate markers and survival. *Lifetime Data Anal.* **1,** 73–86.
14. Topol, E. J., Califf, R. M., VandeWerf, F., Simoons, M., et al. (1997) Perspectives on large-scale cardiovascular clinical trials for the new millennium. *Circulation* **95,** 1072–1082.
15. Cohen, J. (1990) Things I have learned (so far). *Am. Psychol.* **45,** 1304.
16. Beck, A. T. and Steer, R. A. (1993) *Beck Depression Inventory: Manual.* Psychological Corporation, San Antonio, TX.
17. Venter, A. and Maxwell, S. E. (1999) Maximizing power in randomized designs when N is small, in *Statistical Strategies for Small Sample Research* (Hoyle, R. H., ed.) SAGE, Thousand Oaks, CA, pp. 31–58.
18. Muller, K. E., LaVange, L. M., Ramey, S. L., and Ramey, C. T. (1992) Power calculations for general linear multivariate models including repeated measures applications. *J. Am. Statist. Assoc.* **87,** 1209–1226.
19. Shuster, J. J. (1990) *Handbook of Sample Size Guidelines for Clinical Trials*. CRC Press, Boca Raton, FL.
20. Lachin, J. M. (1981) Introduction to sample size determination and power analysis for clinical trials. *Control Clin. Trials* 2, 93–113.
21. Zar, J. H. (1996) Biostatistical Analysis, Prentice Hall, Upper Saddle River, NJ.
22. Dawson-Saunders, B. and Trapp, R. G. (1994) *Basic and Clinical Biostatistics*. Appleton & Lange, Norwalk, CT.
23. Meinart, C. L. (1986) *Clinical Trials: Design, Conduct, and Analysis.* Oxford University Press, New York.
24. Friedman, L. M., Furberg, C., and Demets, D. L. (1996) *Fundamentals of Clinical Trials*, 3rd ed. Mosby, St. Louis, MO.
25. Graybill, F. A. (1976) *Theory and Application of the Linear Model*. Duxbury Press, North Scituate, MA.

26. Current Index to Statistics, Editors: Michael Wichura, Klaus Hinkelmann, http://www.statindex.org/
27. nQuery, Statistical Solutions, Janet D. Elashoff, Ph.D., Stonehill Corporate Center, Suite 104, 999 Broadway, Saugus, MA 01906.
28. PASS, Number Cruncher Statistical Systems, Jerry L. Hintze, PhD., 329 North 1000 East, Kaysville, UT 84037.
29. SamplePower, SPSS, Inc., 233 South Wacker Drive, 11th Floor, Chicago, IL 60606.
30. Power and Precision, Michael Borenstein, Director, Biostat, 14 North Dean Street, Englewood, NJ 07631.
31. EAST, Cytel Software Corp., 675 Massachusetts Ave., Cambridge, MA 02139.
32. Egret SIZ, Cytel Software Corp., 675 Massachusetts Ave., Cambridge, MA 02139.
33. Epi Info, The Division of Surveillance and Epidemiology, Epidemiology Program Office, Centers for Disease Control and Prevention (CDC), Atlanta, GA 30333.
34. O'Brien, R. G. and Muller, K. E. (1993) Unified power analysis for t-tests through multivariate hypotheses, in *Applied Analysis of Variance in the Behavioral Sciences* (Edwards, L. K., ed.). Marcel Dekker, New York, pp. 297–344.
35. Odeh, R. E. and Fox, M. (1991) *Sample Size Choice*, 2nd ed. Marcel Dekker, New York.
36. Schoenfeld, D. A. and Richter, J. R. (1982) Nomograms for calculating the number of patients needed for a clinical trial with survival as an endpoint. *Biometrics* **38,** 163–170.
37. Chow, S.-C. and Liu, J. (2000) *Design and Analysis Bioavailability and Bioequivalence Studies*, 2nd ed. Marcel Dekker, New York, p. 139.
38. Goodwin, S. N. and Berlin, J. A. (1994) The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Ann. Intern. Med.* **121,** 202.
39. Winer, B. J. and Michels, K. M. (1991) *Statistical Principles in Experimental Design*, 3rd ed. McGraw-Hill. New York, pp. 405–416.
40. Sham, P. (1998) *Statistics in Human Genetics*. Arnold, London, p. 21.
41. Lin, L. I.-K. (1989) A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **45,** 255–268.
42. Lin, L. I.-K. (1992) Assay validation using the concordance correlation coefficient. *Biometrics* **48,** 599–604.