Journal of the American Association for Laboratory Animal Science
Copyright 2011
by the American Association for Laboratory Animal Science

Vol 50, No 4
July 2011
Pages 445–453

Ethics and Animal Numbers: Informal Analyses, Uncertain Sample Sizes, Inefficient Replications, and Type I Errors

Douglas A Fitts

To obtain approval for the use vertebrate animals in research, an investigator must assure an ethics committee that the proposed number of animals is the minimum necessary to achieve a scientific goal. How does an investigator make that assurance? A power analysis is most accurate when the outcome is known before the study, which it rarely is. A 'pilot study' is appropriate only when the number of animals used is a tiny fraction of the numbers that will be invested in the main study because the data for the pilot animals cannot legitimately be used again in the main study without increasing the rate of type I errors (false discovery). Traditional significance testing requires the investigator to determine the final sample size before any data are collected and then to delay analysis of any of the data until all of the data are final. An investigator often learns at that point either that the sample size was larger than necessary or too small to achieve significance. Subjects cannot be added at this point in the study without increasing type I errors. In addition, journal reviewers may require more replications in quantitative studies than are truly necessary. Sequential stopping rules used with traditional significance tests allow incremental accumulation of data on a biomedical research problem so that significance, replicability, and use of a minimal number of animals can be assured without increasing type I errors.

Abbreviations: SSR, sequential stopping rule.

Animal ethics committees, such as Institutional Animal Care and Use Committees in the United States, must assure that the numbers of animals proposed for use in scientific experiments are justified and reasonable (United States regulations are reviewed in the *IACUC Guidebook*[26]). The goal of this process is to assure that investigators using animals in experimental research have enough subjects to accomplish experimental aims without wasting numerous animals. Among the suggestions[26] for ways to reduce the number of animals are: (1) rational selection of group size (pilot study, power analysis); (2) careful experimental design; (3) maximizing use of each animal; (4) minimizing loss of animals; and (5) statistical analysis (maximum information from minimum number of animals).

The investigator must justify, and the IACUC must approve, all animal numbers proposed for use in a new protocol or full renewal of a protocol. The IACUC must also approve the numbers of animals requested when protocols are amended to add new experiments to the protocol. According to the Office of Laboratory Animal Welfare,[25] the addition of animals to a protocol can itself constitute a significant change to the protocol. In other words, adding animals to a protocol can constitute a significant change even if all procedures that will be used are already included in the protocol.

When an IACUC requests a power analysis as a justification for sample sizes, responses from investigators may include anything from an excellent analysis to bewilderment. Graduates of medical or veterinary schools often do not have strong backgrounds in statistics and, for this reason, the justification of animal numbers causes distress on the part of both investigators and IACUC members. In my experience, a common response to an IACUC's request for a power analysis goes something like, "A power analysis is not possible for our studies because the experiments have not yet been done, and we do not know what the means and standard deviations will be. We will use the fewest number of animals necessary to produce significance. We will conduct the experiment 3 times so that it will be acceptable for publication by a major journal." When pressed about how they will know what the fewest number of animals is, investigators often reply that they will conduct a small study with a few subjects and use the resulting data to conduct a formal power analysis. The investigators then propose to request additional subjects from the committee for the full study by using that pilot analysis.

This response raises several interdependent questions that are the topic of this article: (1) What is an a priori power analysis, and when is it appropriate? (2) How can the investigator and IACUC assure that no more animals will be used than the number necessary to produce significance? and (3) How many times is it necessary to replicate a successful experiment to assure that results are repeatable without wasting animals or repeatedly requesting additional animals from the IACUC?

Requesting animal numbers in stages (for example, a pilot study followed by a main study) consumes time and effort for all concerned and can increase the rate of type I errors if the pilot animals actually are included in the main study.[13] The current article discusses the problems encountered in trying to achieve an optimal sample size and minimize animal usage. A companion article[11] reviews a method, the variable-criteria sequential stopping rule (SSR), that can be used with many ordinary experiments in the biomedical sciences to solve some of these problems.

Received: 20 Jul 2010. Revision requested: 27 Oct 2010. Accepted: 26 Dec 2010.
Office of Animal Welfare, University of Washington, Seattle, Washington.
Email: dfitts@u.washington.edu

Power and Type I or Type II Errors

Concepts. This section reviews the concepts of power and Type I or Type II errors. Readers who are familiar with these concepts can skip to the next section, *Selection and Usage*. Numerous excellent texts, reviews, and bibliographies are available to assist with design and analysis of animal-related data, including a bibliography in the *Guide for the Care and Use of Laboratory Animals*[18] and an entire issue of *ILAR Journal* devoted to the topics of sample size determination,[4] experimental design,[8,14,17,19] and statistical analysis.[8,27] 'Statistical analysis' of a planned experiment can mean a number of different things, but the meaning familiar to many biomedical and biobehavioral researchers is the null hypothesis significance test. Bayesian and other types of statistical procedures also are available for analyzing the same types of data, but biomedical researchers are less familiar with these methods even if they may sometimes be more appropriate analyses.[6] The present review concerns the null hypothesis significance test because of its prevalence in the field.

In a significance test, the investigator states as the 'null' hypothesis that there is no effect of a treatment or no relationship between the variables in the population. A statistic then is calculated from the data, and the statistic is assigned a '*P* value' based on the known probability distribution of the statistic when the null hypothesis is true. The *P* value is the probability of obtaining a statistic as extreme or more extreme when the null hypothesis is true. If the *P* is less than a sufficiently small value, called alpha (for example, 0.05), that was determined in advance, the investigator can 'reject the null hypothesis' and conclude that the treatment really does have an effect or that there actually is a relationship between the variables. Alpha is the probability that the null hypothesis will be rejected when the null hypothesis is actually true. A rejection of the null hypothesis when it is true is therefore an error of inference, known as a type I error or a 'false discovery.' An alpha of 0.05 is a statement that the investigator is willing to accept a type I error 5% of the time if the null hypothesis is actually true.

The *P* value gives information on the likelihood that the investigator could find a significant result in an exact replication,[16] but it does not provide any information about how large or important a difference or relationship is. With sufficient sample size, even tiny and unimportant effects can be discovered as "highly significant"[24] and "likely to be replicated." Therefore, a null hypothesis significance test should be used only when the investigator is satisfied to learn whether there is some difference or relationship between the groups or conditions, and if so, in what direction.[12]

Power in a null hypothesis significance test is the probability that the null hypothesis will be rejected correctly given that there is a true difference or relationship in the population. This probability can be calculated directly if the population parameters are known (for example, means, standard deviations, correlations). Because these parameters are not known in advance, the a priori power analysis requested by the IACUC will always be based on some kind of estimation of these parameters from previous testing and educated guessing. Power is related directly to sample size, so larger sample sizes will increase power. Even with high power in an experiment, there is always the possibility that the investigator will fail to correctly reject the null hypothesis. A type II error occurs when one fails to discover a significant effect when the null hypothesis is false, and the probability of this type II error is denoted as β, which is the complement of power (that is, power = 1 – β).

The following example illustrates the relationship between power and sample size. In a dependent-samples *t* test based on difference scores from a single set of subjects, the power of the test depends on the mean and standard deviation of the difference scores, the sample size, and the alpha chosen for the test. The *t* is given by the following formulas, in which M_D is the sample mean of the difference scores, μ_D is the mean of the population according to the null hypothesis, s_D is the sample standard deviation of the difference scores, se_D is the standard error of the difference scores, and *N* is the number of difference scores (that is, pairs of scores).

$$t = (M_D - \mu_D)/se_D;\ se_D = (s_D/\sqrt{N})$$

The μ_D is included so that investigators can test whether the mean difference is significantly different from any hypothesized population value. If that hypothesized population value is 0 (that is, $\mu_D = 0$), then the *t* is a simple ratio expressing the mean of the differences in standard error units. The standard error of the differences is the standard deviation of the differences divided by the square root of the sample size. A larger mean difference increases the numerator, and a larger sample size decreases the denominator. Both of these increase the value of *t* and therefore increase power (a large *t* is required for significance). A larger standard deviation increases the denominator and therefore decreases the size of *t* and the power of the test.

Alpha affects power by setting the value of the critical *t* required for significance. With everything else being equal, reducing the alpha from 0.05 to 0.01 makes rejecting the false null hypothesis more difficult (a larger obtained *t* is required for the 0.01 level). A decreased probability of rejection of a false null hypothesis is the same thing as a reduction in power. To compensate for the loss of power by selecting the more conservative alpha (for example, 0.01 instead of 0.05), the investigator must increase sample size (use more animals in the experiment).

Selection and usage. If the goal of the IACUC is to conserve research animals, is it ethically defensible to allow an investigator to choose a smaller alpha so that more animals are required for significance? The answer is a qualified 'yes.' Using a 'loose' alpha such as 0.10 may give more power to the test and require fewer subjects, but doing so increases the probability that a type I error will be published. The publication of an error could encourage other investigators to repeat the experiment to elaborate on the false findings, leading to further waste of animals. The level of alpha is arbitrary to a degree, and some research situations have greater risks associated with the publication of a false discovery. On the other hand, if the investigator is going to replicate the experiment 3 times to be certain that an effect exists, the use of a more conservative alpha for each of the 3 replications will probably require an excessive use of animals. The customary choice of 0.05 for alpha is based on balancing the likelihood of type I errors with that of type II errors. A decrease in one type of error increases the rate of the other type of error. A typical a priori type II error for much basic biomedical research is approximately 0.20, indicating the power is approximately 0.80, but a priori power anywhere from 50% to 95% is fairly common.

Some investigators wish to set both type I and type II errors to be very low simultaneously (for example, 0.01 or 0.05). Doing so seems logical at face value to avoid publication of false discoveries yet not miss the opportunity to detect a significant effect. The problem is that the only way to reduce both errors to, for example, 0.01 is to increase the sample size greatly. This problem is illustrated in Figure 1, where the power of an independent groups *t* test with a 2-tailed alpha of 0.05 is plotted as a function of the standardized size of the effect (that is, *d*, the

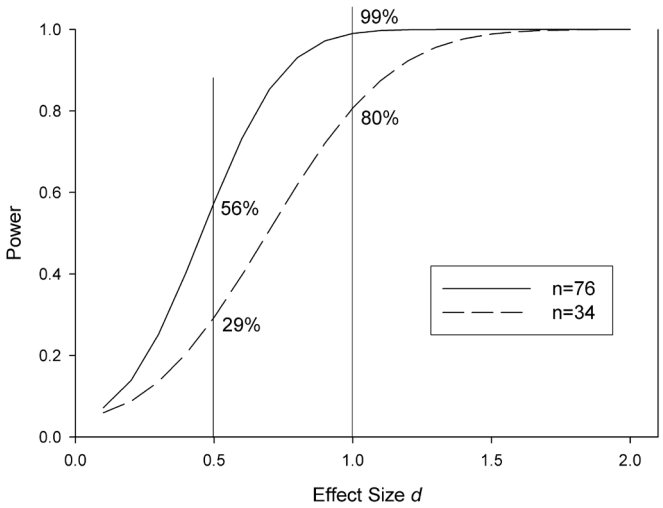

Figure 1. Power as a function of effect size *d* (difference between means divided by standard deviation) for different total sample sizes in a 2-tailed *t* test with 2 independent groups and $\alpha = 0.05$. If an effect of 1 SD is the smallest interesting effect, designing the experiment to detect this effect with high power, such as 99%, can waste animals, because the test continues to detect trivial effects, such as 0.5, with a high frequency (56%). Setting a power of 80% for the smallest interesting effect causes the detection of trivial effects to decline more steeply. Both tests have high power for detecting relatively larger effects, such as 1.5 SD.

difference between the means divided by the within-groups standard deviation[2]).

Suppose for the sake of this example that the difference the investigator hopes to find is quite large, greater than 1.5 SD, and that the smallest difference between the means that would be interesting is about 1 SD. To set a 99% chance of detecting the smallest interesting difference, 76 total subjects would be needed in the 2 groups combined, based on a sample-size calculator.[7] Note that power remains high (Figure 1), well below an effect size of 1.0, and the probability of detecting an effect of only 0.5 SD is still 56%, indicating that so many animals have been used that this trivial difference will still be detected more than half the time. If instead a criterion of 80% power is used to detect the smallest interesting difference, 34 total subjects would be used, and the power falls off rapidly at effect sizes smaller than 1 SD. Both strategies yield excellent power in the range where the investigator actually hopes to find a difference, that is, 1.5 SD and greater. As noted previously, the *P* value does not discriminate between meaningful and trivial differences.[24] Animal experiments that are designed to be so powerful that they can detect even trivial differences are wasteful in terms of animal use.

What Is an A Priori Power Analysis and When Is It Appropriate?

Power analysis. When they are asked by the IACUC to give a power analysis to justify sample sizes, investigators often explain that a power analysis is not appropriate because they will not know the means and standard deviations of the groups until they have conducted the test. Means and standard deviations are required to calculate power or sample size, and few investigators do experiments for which the results are known beforehand. Actual means and standard deviations are required for a posthoc power analysis in which the goal is to determine the amount of power in a test that has already been conducted. However, an IACUC needs an a priori power analysis, which is always assumed to use estimates for the means and standard deviations rather than the actual means and standard deviations. An investigator can always estimate the means and standard deviations of a null hypothesis test to some extent, so an a priori analysis is always possible if a significance test is planned. The accuracy of the analysis depends on the accuracy of the estimates.

An a priori power (or sample-size) analysis in the context of an IACUC protocol is a formal means of communicating information across disciplinary boundaries about the anticipated effect sizes and necessary sample sizes in a way that can be understood by all. IACUC members have varied backgrounds, and none of them may be expert in the particular area of a given protocol. The investigator writing the protocol does have experience and expertise and is familiar with the dependent measurements that are proposed. The investigator knows better than most IACUC members how variable these measurements tend to be and how big an effect must be in order to be considered important in the field. An a priori power analysis can be used as a formal method to communicate to IACUC members about the size that an effect must be in order to be considered important. For example, a social psychologist might be very excited about a correlation coefficient that would disappoint a physiologist. When effect-size information is presented in this standardized way, it is easy to translate knowledge of effect sizes and variability into an estimate of sample sizes in a way that is understood by all.

The IACUC does not expect investigators to be able to predict perfectly the outcome of an experiment. What the IACUC wants is for investigators to use the best available evidence based on personal experience and the literature to give a detailed explanation of how the investigators determined their sample sizes. If an experiment of this sort has been published previously, or if pilot data are available, the standard deviations from those data can be used to estimate sample size for the next experiment. Even if no similar experiment has been done before, the standard deviation of the dependent or outcome variable often is known (such as the standard deviation of normal body temperature for a certain line of rats), and this information can be used in an a priori power analysis by assuming equal standard deviations in all groups. If a treatment is suspected to increase the variance as well as the mean value for a group, this expectation can be built into the sample size analysis to increase power to offset the increased variability. The effect size should be based on the minimal effect that would be considered important instead of the effect size that has been observed in the past. This way, the experiment will be powered to detect any meaningful effect. As demonstrated in Figure 1, the power value should be selected so that the probability of detecting less interesting effects is low.

The actual amount of power invested in an experiment is of interest to the IACUC, and the IACUC should require investigators to alter the sample size when the experiment is obviously either underpowered or overpowered. As noted earlier, there is no accepted standard for the amount of power that must be selected in all cases, and the appropriate power may depend on a variety of scientific issues. In general, the power value should be high rather than low. The issues involved in determining the power in an experiment should be based on science and logic rather than cost or expediency.

A large amount of power is good, but it is possible to have too much. A power curve has a steep slope in the low-power end and a shallow slope at the high-power end. In the low-power range, the addition of a few subjects can add a lot of power to the experiment, but the value of each additional subject diminishes rapidly at the high power end of the distribution. For example,

Figure 2 illustrates power as a function of total sample size for a 2-tailed *t* test with an alpha of 0.05 and an effect size equal to 1 SD. To increase the power of the test from 60% to 70% requires 5 additional subjects (from 22 to 27 total subjects), so each additional subject increases power by 2 percentage points. In the same test, an increase of 22 subjects (from 54 to 76 total subjects) is required to increase the power from 95% to 99%, so each additional subject increases power by only 0.18% points. This effect is well past the point of diminishing returns. Therefore, animals should be added when they provide a large gain of power but not when the gain is small. The value of 0.8 is the approximate point in a power curve where the gain per subject begins to become shallow, so that each additional subject adds less and less power. An investigator should be able to justify the level of power in an experiment, and requests for exceptional power should require exceptional justification.

When previous data from a nearly identical experiment are available, a power analysis will give a strong estimate of the sample size that will be required. Presenting this fact to the IACUC by saying something like, "Because Smith and Jones found significance with 20 subjects per group, we will use 20 subjects per group in our experiment" does not give the IACUC enough information to judge whether the investigator has selected the best sample size, because Smith and Jones may have observed a *P* of 0.0001. In that case, the original experiment was probably overpowered, and the new study should use fewer subjects. Instead, the investigator should report the standard deviations found by Smith and Jones and conduct a sample size analysis with known levels of alpha and power for detecting the smallest meaningful effect. This analysis can be conducted in a few seconds with a computer program.[7]

When little is known about the standard deviation or size of effect before the experiment is conducted, investigators may feel uncomfortable reporting a power analysis because doing so seems like guessing, and they do not want to mislead the IACUC. Investigators may feel more comfortable simply stating, "We will use 6 subjects per group because we usually find significance in these sorts of experiments with 6 subjects." One reason that the IACUC prefers a power analysis in such cases is that the power analysis makes certain facts explicit to the committee that a simple guess of 6 subjects does not. This guess will not mislead the IACUC if the investigator simply reports a low level of confidence. A power analysis provides more information to the IACUC than does the simple guess, because the analysis explicitly indicates the level of risk (types I and II errors), the desired power, and the desired size of the effect in standardized units. For example, if the investigator specifies a 2-tailed, 2-sample *t* test with alpha of 0.05 with a power of 0.80 and sample sizes of 6 per group, one can work backward in a sample size calculator[7] to conclude that the investigator is interested in a difference between the 2 means of greater than 2 SD. If the investigator is really interested in effects as small as 1 SD, the IACUC should question the design with 6 subjects per group as being underpowered.

So, when is an a priori power analysis appropriate? If the investigator plans to use null hypothesis significance testing to analyze the experiment, the power of the statistic will always be estimable to some degree, and a power analysis to determine sample size will always be possible. Both the investigator and the ethics committee are aware that the reported effect size is an estimate and may be inaccurate. For this reason, the investigator should plan the experiment for the smallest effect that would be considered important to include sufficient sample size to detect any important effect. If the investigator does not plan to use null hypothesis significance testing to analyze the experiment, it is inappropriate for the IACUC to request a power analysis. Some other method will be required to justify animal numbers.

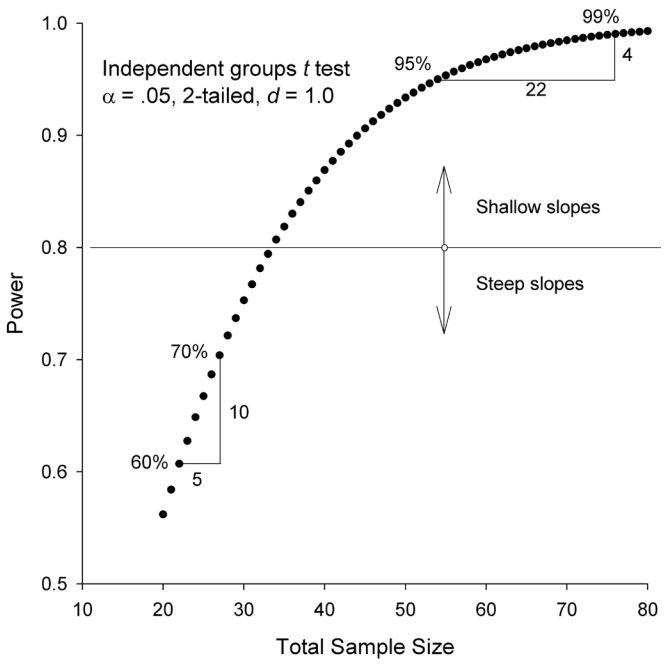

Figure 2. Power as a function of total sample size. A power of 0.8 is the approximate point where the addition of each individual subject begins to add less power to the test. An increase of power from approximately 60% to 70% requires only 5 subjects, whereas an increase from 95% to 99% requires 22 subjects. *d*, difference between means divided by standard deviation.

Pilot study. A power analysis may be the best way to communicate information about sample size, but it is not necessarily always an excellent way to determine sample size. As the amount of solid prior information in a power analysis declines, so does confidence in the power analysis as a good indicator of the appropriate sample size. If an investigator plans sufficient sample size for the smallest meaningful effect given the existing assumptions, and if either the effect size is much larger or the standard deviations of the samples are much smaller than anticipated, a significant result with a very small *P* value (much less than 0.05) will be achieved. In this case, a smaller sample size would have sufficed. Conversely, if the size of the standard deviation is larger than anticipated, a *P* value of 0.06 may be obtained instead of less than 0.05. In this case, we can conclude nothing from the null hypothesis significance test, and animals have been wasted. This inefficiency is built into the null hypothesis test and is elaborated following.

When the foundation for a power analysis is particularly weak, and the investigator suspects that a large number of animals will be required to conduct an experiment, the investigator may request animals for a pilot study instead of a full study. Pilot studies are recommended by regulatory bodies such as the Office of Laboratory Animal Welfare (*IACUC Guidebook*[26]). The data from the pilot animals will then be used to conduct a power analysis with a stronger foundation, and if the results are promising, another request for a larger number of animals will be made to the IACUC. Alternatively, the IACUC may request a pilot study when many animals are likely to be involved and the best sample size is uncertain.

If the only point of a pilot study is to determine an optimal sample size, its use can actually waste animals because: (1) adding animals to the pilot study and testing again at the 0.05

level will increase the type I error rate above 0.05,[1,9,13,28] thus increasing the chance that animals will have been used unnecessarily; and (2) omitting the pilot study from the published data can needlessly duplicate the use of animals in testing. The waste is not great if the number of pilot animals is only a small fraction of the total that will finally be used, but the waste becomes more significant if the 'pilot' study involves a third to a half of all animals used. Reapplication to the IACUC for additional animals would not be necessary if more dynamic methods of determining sample size were available. Sequential stopping rules (SSR) provide such a method,[1,9,10,13,28] but they are largely unknown by the general population of researchers at this time.

Other excellent reasons may support conducting a pilot study, such as exploring doses or technical issues, and these may not waste animals as long as knowledge is gained.

The Fixed-Stopping Rule and Its Abuse

Null hypothesis significance tests originally were intended to be conducted using a fixed-stopping rule. For example, suppose an investigator designs an experiment with 6 animals in each of 2 groups as determined by power analysis using a formula or computer program. (Many such programs are available for a fee or for free. For example, G*Power[7] has extended features such as easy calculation of the power values in Figure 2.) The investigator selects a null hypothesis that the means in the 2 populations are equal and selects an alpha of 0.05 for a 2-tailed t test. The number of degrees of freedom for this test is 10 (6 + 6 – 2); therefore $t(10) = 2.06$, $P = 0.066$. Because the obtained P value is not less than the designated alpha of 0.05, the investigator cannot reject the null hypothesis. Because the true effect of the treatment in the population is unknown, there are 2 possibilities after the t test: the investigator made a correct decision that there is no effect of the treatment, or the investigator made a type II error and narrowly missed detecting a true significant effect.

The experiment ends at that point. The fixed-stopping rule is based on a set of probabilities with the critical assumption that the researcher will never conduct more than one test. However, since the inception of the null hypothesis test, researchers have been tempted to examine the results of the first test and then a second or third test with successively more subjects based on the outcome of the first or second test. The assumption is that perhaps, with a few more subjects, it will emerge that the means are actually much closer or farther apart than they now appear. An addition of subjects should make things clearer. A tenure review or a grant application may rely on a definitive outcome, and there are no funds or time to reproduce the entire experiment from the beginning. If animal subjects have been used, it seems a huge waste of life to stop at this point. Many statisticians have denounced the null hypothesis test for this and other reasons.[12,15,20-22,24] Such results should never be reported simply as 'not significant' or '$P > 0.05$' without explanation. Claiming a 'marginally significant effect' or a 'trend toward' a significant effect is not technically accurate because 'trend analysis' has a specialized meaning, and a trend is either significant or not significant—just like other null hypothesis experiments. These phrases should never be used without citing the means and standard deviations and the actual obtained P value for the test.[5] A future meta-analyst may find reason to get excited about 2 P values of approximately 0.066 in 2 different papers, but the meta-analyst will never observe this phenomenon unless both obtained P values are published.

The procedure of testing with successively larger sample sizes until one finds a P value less than 0.05 is excellent for detecting effects if they really do exist, but is also excellent for detecting effects if they do not exist. The successive procedure has high power and is efficient in the use of subjects. However, if the null hypothesis is true and there is actually no effect of the treatment in the population, the rate of type I errors increases rapidly. If a significant effect emerges with this procedure, the actual rate of type I errors in the experiment may be unknown, so interpreting how meaningful the result may be is impossible.

Effects of Sequential Sampling

Figure 3 illustrates how type I errors accumulate when an investigator uses sequential sampling with a criterion for significance of 0.05 at each test (Figure 3, left). In this example, a computer was used to conduct 10,000 independent-groups t tests where data were sampled randomly ($n = 10$ per group) from 2 populations with identical means. Of these 10,000 t tests at the 5% level of significance, 5% were significant.

To simulate sequential sampling, the computation used a larger probability of 0.36, called the 'upper criterion,' as the dividing line between those experiments for which sample size would be increased and those experiments for which the experiment would stop. In other words, if the P value from an experiment exceeded 0.36, the outcome would not be considered significant. The choice of this value was arbitrary. Usually investigators do not have a set probability in mind, but make this decision informally by inspecting the data.

The area of each bar is proportional to the number of tests conducted at that sample size, and the colored portions of the bars are proportional to the number of tests that were significant, uncertain, and not significant, respectively. Of the 10,000 tests, 31% fell in the 'uncertain' range between P values of 0.05 and 0.36 when the sample size was 10 (Figure 3, leftmost bar). The fixed-stopping rule mandates stopping here. However, to each of these 3100 t tests in the uncertain range, a single additional subject was added and the analysis conducted again with $n = 11$. Note that the proportions are no longer 5% significant and 31% uncertain. If a result with the first 10 subjects was in the uncertain range, the addition of one more subject is not likely to change things much. That is, the second test is not independent from the first test. Nevertheless, a few more of the tests with $n = 11$ were now significant. Another subject then was added to tests that remained in the uncertain region for $n = 12$ and $n = 13$, with an additional increment of significant results at each step. For the entire procedure, 8.46% of the 10,000 t experiments were significant at a P level of less than 0.05. Many of these experiments were significant with $n = 10$, but some were not significant until $n = 13$. The actual rate of type I errors is 8.46% even though a nominal alpha of 5% may be reported in the statistics section of the manuscript. The number of errors increased from approximately 500 to 846 of 10,000 experiments by increasing the sample size from 10 to only 13. Sequential sampling like this always increases the Type I error rate.

In summary, with the customary fixed-stopping rule, a sample size must be determined in advance, and all of the data for all subjects must be collected before the statistical test is conducted once, at the end of the experiment, for better or worse. To add additional sample size at this stage is incorrect, because doing so will greatly increase the probability of a Type I error.[1,9,13] Increased type I errors confuse the literature and waste animals. Investigators who favor pilot studies followed up by 'beefed up' sample sizes when the results of the pilot are promising should be careful about controlling type I errors. One way to do this is

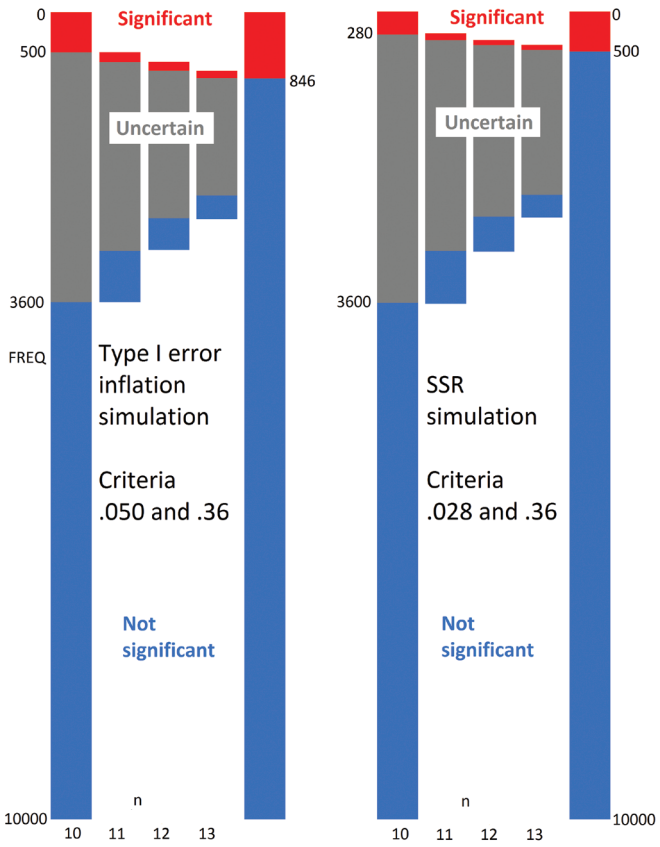

Figure 3. Frequency of errors during sequential testing when the null hypothesis is true. (Left) The areas of the bars are proportional to the number of 10,000 simulated experiments that were significant (less than 0.05), not significant (greater than 0.36), or uncertain (between 0.05 and 0.36) after a *t* test with the null hypothesis true. The leftmost bar includes all 10,000 experiments conducted with $n = 10$. According to the fixed stopping rule, experiments should always be stopped after this first test when the proportion of type I errors equals alpha (0.05). Instead, one subject then was added to all experiments in the uncertain region, and the test was redone after $n = 11$, $n = 12$, and $n = 13$. The fifth bar shows the final decision on all 10,000 experiments. The addition of subjects to experiments in the uncertain region increased the actual rate of Type I errors by 69% from 0.05 to 0.0846 because each successive test included new errors. (Right) An SSR approach using criteria of 0.028 and 0.36 instead of 0.05 and 0.36. The SSR assumes that new subjects will be added to uncertain experiments. Sequential testing of the uncertain region at the 0.028-level produces an error rate of 0.05 for all experiments. The use of a criterion less than 0.05 compensates for the inflation of alpha and allows one to use sequential testing with an overall $\alpha = 0.05$. The individual criteria are specific to the desired sample sizes and can be determined from a published table.

to use a SSR[1,9,10,13,28] as introduced at the end of this article and summarized in the companion article.[11]

How Can an Investigator Assure that No More Animals Will Be Used Than the Number Necessary to Produce Significance?

The phrase, "...than the number necessary to produce significance," is not appropriate to all planned experiments. 'Significance' implies a null hypothesis significance test, and not all experiments should be analyzed in that manner. The question implies that the experimenter will be satisfied to know that a difference or relationship exists and would not be interested in how large the difference or relationship is. The broader question should be worded, "...than the number necessary to achieve the scientific objectives." Investigators whose objective is "to create a confidence interval of a certain size" instead of "to produce significance" need to make the rationale for this clear to the IACUC, especially if the IACUC has the expectation that everyone should be using a power analysis with a significance test to justify animal usage. IACUC members should be aware that some investigators need to use more animals than a power analysis would suggest so that the confidence interval for the obtained effect size will be of a predetermined width to achieve their scientific goals. The investigator needs to explain clearly why the increased accuracy is necessary as a part of the rationale for animal numbers in the IACUC protocol. Investigators with a genuine need for accuracy will have no problem supplying a coherent justification. Simple methods are available to estimate the sample size required to produce a confidence interval of a certain width,[23] and these methods will help to assure that the experiment uses the minimal number of animals to achieve this scientific goal.

Assuming that the investigator has determined that a null hypothesis test is appropriate for the experiment, a number of methods of varying validity can be used to determine sample size. Those already discussed include: 1) always using the same sample size for the same type of study; 2) using the same number of animals used by other investigators in a published paper on the same topic; 3) conducting a power analysis; 4) conducting a pilot study; and 5) testing sequentially with increasingly greater sample sizes until the result becomes significant. This last method is the only method that can assure that no more animals than necessary will be used. However, this method increases type I errors and therefore potentially wastes animals if the criterion used for each test is the same as alpha (for example,

0.05). The fixed-stopping rule without sequential sampling makes it difficult for an investigator to assure the IACUC that a minimal number of subjects will be used to produce significance unless the actual effect size observed in the data is very close to the effect size that was used in the power analysis. This process is inefficient because any fixed sample size that is determined before the experiment is conducted could be very wrong for a particular sample, and the investigator may have used either too few or too many subjects to prove a point.

Many statisticians favor the publication of parameter estimates and confidence intervals instead of null hypothesis significance tests for this very reason. From a predictive standpoint, a *P* value of 0.051 is just as valuable as a *P* value of 0.049. Regression equations from 2 studies with these 2 *P* values will have virtually identical success at predicting future scores from a criterion, for example. The probability that each experiment came from populations with equal means is virtually identical. With the fixed-stopping rule, however, one study is published and has an influence on the field ($P = 0.049$) and one is not ($P = 0.051$). Alternative procedures are being proposed. One such procedure focuses on the probability of replication[3,20,21] instead of an arbitrary decision as to whether the experiment is significant or not. These new methods are not yet widely known or understood by many biomedical researchers doing animal studies.

The fact remains that null hypothesis significance tests are used widely in the biomedical literature. Reviewers continue to question the meaning of a confidence interval that includes the null value. The Instructions for Authors of this very journal state: "If the *P* value is not statistically significant, there is no difference." Furthermore, investigators like the null hypothesis test as a decision-making tool for drawing a definite conclusion from a study (within stated error limits) concerning the existence and direction of an effect.[12] However, its use with the fixed stopping rule makes it difficult to establish a minimal necessary sample size.

How Many Times Must a Successful Experiment Be Replicated to Assure that Results Are Repeatable without Wasting Animals?

Probability of replication. Replication is a critical part of the scientific method, but when ethical considerations attend the use of every subject, how many replications are actually necessary and how many are a waste of animals must be considered carefully. If the goal is to demonstrate that an effect is likely to be repeated significantly in an exact replication, an investigator can adjust the alpha of an experiment to a much more conservative level, for example 0.005 instead of 0.05. When the *P* from a significance test is less than 0.005, the probability is greater than 80% that an exact replication will be significant at the 0.05 level.[16] When an experiment has achieved a *P* value of less than 0.005, the investigator should consider carefully whether further replication is necessary.

However, using 0.005 as one's level of significance in traditionally analyzed experiments with the fixed-stopping rule can lead to an enormous waste of subjects if the hypothesized effect is actually tiny or nonexistent. Suppose an investigator has planned an experiment using an independent-groups *t* test and, after a power analysis, learns that the sample-size recommendation for the fixed-stopping rule is 20 per group at the 0.005-level. In this scenario, a decision about the null hypothesis will not be made until all 40 animals in the 2 groups have been tested. Thus, many animals are used to detect absence of an effect.

Independent replications. Replication is not relevant only for statistics; it is also important with regard to factors such as which experimenter is conducting the trial, the stock of drug being used in a drug study, the batch of animals obtained from the breeder, the time of day, and the weather conditions when the experiment is conducted. No matter how rare a *P* value is in a single experiment, it still may be necessary to conduct an independent replication in some types of experiments to demonstrate that the effect survives changes in poorly controlled variables. An independent replication will help to assure that another investigator will be able to replicate the effect.

A *P* value is not only an indicator of significance but also an indicator of the probability that an exact replication will be significant. As noted in the previous section, an obtained *P* value of less than 0.005 in an experiment means that there is an 80% chance that an exact replication will be significant at the 0.05-level. The obtained *P* in an experiment is the best estimate of the mean and median of all *P* values in the population of identical experiments. [16] If the obtained *P* of 0.05 is in fact exactly the median of the population of *P* values for all identical experiments, it indicates that half of identical experiments (replications) will be significant ($P < 0.05$) and half will not be significant ($P > 0.05$). Therefore, if the obtained *P* is exactly the same as alpha, there is only a 50% chance of replicating the experiment with a significant result with the same alpha. This situation is problematic for investigators who anticipate that their experiments might be repeated by others. A decision to replicate a significant finding should be influenced by the obtained *P* value on the first test. As indicated earlier, if *P* is less than 0.005, the result is already highly likely to be replicated, and an actual replication is probably not necessary and may be a waste of animals. If, instead, *P* is approximately equal to 0.05, a replication is probably a good idea to assure that the probability of replicating the result is better than 50:50.

When independent replications are conducted, all replications should be reported in the scientific publication. The best way to do so is to average across all of the replications instead of reporting a single test that is 'representative' of several tests, because the average of all observations is likely most 'representative' and avoids bias. Apart from a completely random selection, any choice has the potential to introduce bias. If a representative test is presented, the method for selecting the presented test should be explained.

Sometimes, independent replications that are conducted identically do not have identical results. This situation may or may not present an opportunity for learning. For example, if a first experiment is significant and a second is not, there are several possible explanations: (1) the first experiment may have been a type I error; (2) the second experiment may have been a type II error; or (3) subtle and uncontrolled but critical differences existed between the procedures. A discrepancy between replications may be a good reason for adding sample size, but because this decision is made after evaluating the data, the only way to accomplish it legitimately without increasing type I errors is to use a stopping rule such as the variable-criteria SSR (see following).

'Three times' rule. In my experience, some investigators report to the IACUC that they must conduct each experiment at least 3 times to have confidence in the data and to be able to publish in a top journal. The investigators justify their group sizes with a power analysis using 80% to 95% power and then multiply the total number of animals by 3 for their request.

If finding type I errors (false discovery) is a concern, the correct way to compensate is to reduce alpha from 0.05 to some-

thing more conservative, such as 0.01 or 0.02. Some investigators who say that they are concerned about producing false-positive results try to combat this possibility by increasing power from 0.8 to 0.95 without changing alpha (thereby increasing sample size). Because the investigators keep testing for significance at the 0.05-level, they will always have a 5% chance of producing a false-positive result, no matter what the power. Reducing alpha also will require an increase in sample size, but the resulting *P* obviously must be tested at that more conservative criterion to reduce type I errors.

I have systematically examined the Instructions for Authors in most biomedical journals with high impact factors, and no journal that I examined required a certain number of replications. All journals require a valid statistical argument, and some state that it is desirable to replicate the results. Generally the journal's reviewers, not the journals themselves, insist on and perpetuate the 'three times' rule. Is it a good rule? This question is important because, even as important as replications are to science, another imperative is to limit animal use to the number necessary to convincingly demonstrate a scientific principle.

The 'three times' rule is probably a good practice when presenting semiquantitative data such as histologic findings. Demonstrating 3 times that an antibody binds specifically to a ligand is fairly convincing. However, when the data are quantitative and are associated with probability values for type I and type II errors for each test, the 'three times' rule can be excessive.

Some experiments may require additional animals to identify the dose or refine the procedure. These are legitimate uses of animals, but they are not replications. These uses should be requested from the IACUC as animals necessary to work out technical problems, not as animals required for replications. If the technical problems require use of 3 times the number of animals implied by a power analysis to complete one experiment appropriately, the IACUC should be aware of and confront this problem. The Veterinary Services unit at the institution, which carries some institutional memory, may be able to provide solutions that can prevent the use of animals in redundant pilot studies.

Sequential-Stopping Rules (SSR)

In Figure 3 (left), type I errors in 10,000 simulated experiments grew from 500 to 846 as the sample size per group increased from 10 to 13 when every *t* test was conducted at the 0.05-level. Therefore, the actual overall Type I error rate inflated from 0.05 to 0.0846 because of the sequential sampling. In Figure 3 (right), the errors increase from 280 to 500 in the same 4 tests (that is, n = 10, 11, 12, 13) if the criterion *P* for significance in each test is 0.028 instead of 0.05. Therefore, the overall type I error rate for these tests increased from 0.028 to 0.05. If an investigator were to use this approach to test from sample sizes of 10 to 13 by increments of 1, the overall type I error rate would still inflate, but it would inflate to the desired 0.05. The investigator would have to remember to test at the 0.028-level instead of the 0.05-level. The variable-criteria SSR uses computer simulations such as these to determine the criterion *P* values for many different starting and stopping sample sizes that are relevant to small-sample research problems, and these were collected into a table.[9] All an investigator needs to do to use the variable-criteria SSR is to look up the appropriate stopping criteria in the table and test with those criteria instead of with 0.05; the overall type I error rate will then be controlled at 0.05. The original articles[9,10] provide details on how the variable-criteria SSR were derived. The companion article[11] gives more user-friendly explanation with examples.

Sequential sampling is an excellent technique for detecting significant effects if they actually exist. The problem is that incorrect sequential sampling detects too many significant effects when the null hypothesis is true. A correct use of SSR takes advantage of the benefits of sequential sampling yet controls the tendency to inflate type I errors. SSR are efficient with sample size, because the testing begins with a relatively small number of subjects and stops when significance is achieved. On average, this process will use fewer subjects than would have been used based on a power analysis if the original power analysis was correct in its estimation of the effect size. With large sample sizes and large effect sizes, the SSR uses as many as 30% fewer subjects than does the fixed-stopping rule.[9]

SSR are particularly effective when the size of the anticipated effect is unknown to the point that a usual power analysis will not provide a good estimate of the anticipated sample size with high confidence. The SSR can begin with a relatively low sample size and continue testing until the experiment is stopped by one of the stopping rules. This efficiently and seamlessly converts a pilot study into a main study without increasing type I errors. In effect, every SSR study begins with a pilot study and proceeds to a full study if results are promising. All of the animals required for the full study can be requested at the outset, thus reducing paperwork.

If desired, the SSR can be used with the 0.005-level of significance to demonstrate repeatability[16] without the drawback of the fixed stopping rule (see above). When used with the SSR, the experiment may be stopped early if the null hypothesis is true without using all animals allocated to the experiment. A companion paper[11] provides examples.

Conclusions

The numbers of animals approved for scientific experiments must be justified and reasonable. The *IACUC Guidebook*[26] lists "Rational selection of group size", including "Pilot studies to estimate variability and evaluate procedures and effects" and "Power analysis" among the means to this end and concludes that "Appropriate use of statistical software can generate maximum information from minimal numbers of animals." A pilot study or a power analysis used with the fixed stopping rule can actually use unnecessary numbers of animals under some circumstances. Power analysis applies to null hypothesis significance tests, and significance tests are not always the most appropriate statistical analysis. Replication is an important part of science, but mechanically conducting every quantitative experiment 3 times because of an unwritten rule can waste animals and resources. Null hypothesis experiments with extremely small *P* values are far less likely to require replication than are experiments with *P* values near the selected alpha (for example, 0.05). Sequential sampling and testing holds promise for use with significance tests to minimize animal use in experiments because it is powerful, flexible, and far more efficient with animal subjects than is the fixed-stopping rule. However, an informal and casual use of sequential testing at the 0.05-level should not be done because it greatly inflates the rate of type I errors far above 0.05. Instead, an SSR such as the variable-criteria SSR can be used to exploit the assets of sequential sampling without inflating type I errors. The method helps an investigator assure that no more than the minimal number of animals will be used when significance tests are appropriate. SSR can require up to 30% fewer animals than the fixed stopping rule with the same amount of statistical power, and it can be used to ensure that experiments are repeatable and significant without unnecessary animal use.

References

1. **Botella J, Ximenez C, Revuelta J, Suero M.** 2006. Optimization of sample size in controlled experiments: the CLAST rule. Behav Res Methods **38:**65–76.
2. **Cohen J.** 1988. Statistical power analysis for the behavioral sciences, 2nd ed. Hillsdale (NJ): Erlbaum.
3. **Cumming G.** 2005. Understanding the average probability of replication: comment on Killeen (2005). Psychol Sci **16:**1002–1004.
4. **Dell RB, Holleran S, Ramakrishnan R.** 2002. Sample size determination. ILAR J **43:**207–213.
5. **Desbiens NA.** 2003. A novel use for the word 'trend' in the clinical trial literature. Am J Med Sci **326:**61–65.
6. **Erceg-Hurn DM, Mirosevich VM.** 2008. Modern robust statistical methods. An easy way to maximize the accuracy and power of your research. Am Psychol **63:**591–601.
7. **Faul F, Erdfelder E, Lang A-G, Buchner A.** 2007. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods **39:**175–191.
8. **Festing MFW, Altman DG.** 2002. Guidelines for the design and statistical analysis of experiments using laboratory animals. ILAR J **43:**244–258.
9. **Fitts DA.** 2010. Improved stopping rules for the design of efficient small-sample experiments in biomedical and biobehavioral research. Behav Res Methods **42:**3–22.
10. **Fitts DA.** 2010. The variable-criteria sequential stopping rule: generality to unequal sample sizes, unequal variances, or to large ANOVAs. Behav Res Methods **42:**918–929.
11. **Fitts DA.** 2011. Minimizing animal numbers: the variable-criteria sequential stopping rule. Comp Med **61:**206–218.
12. **Frick RW.** 1996. The appropriate use of null hypothesis testing. Psychol Methods **1:**379–390.
13. **Frick RW.** 1998. A better stopping rule for conventional statistical tests. Behav Res Methods **30:**690–697.
14. **Gaines Das RE.** 2002. The role of ancillary variables in the design, analysis, and interpretation of animal experiments. ILAR J **43:**214–222.
15. **Goodman SN.** 1999. Toward evidence-based medical statistics. 1: the *P* value fallacy. Ann Intern Med **130:**995–1004.
16. **Greenwald AG, Gonzalez R, Harris RJ, Guthrie D.** 1996. Effect sizes and *P* values: what should be reported and what should be replicated? Psychophysiology **33:**175–183.
17. **Howard BR.** 2002. The control of variability. ILAR J **43:**194–201.
18. **Institute for Laboratory Animal Research.** 2010. Guide for the care and use of laboratory animals, 8th ed. Washington (DC): National Academies Press.
19. **Johnson PD, Besselsen DG.** 2002. Practical aspects to experimental design in animal research. ILAR J **43:**202–206.
20. **Killeen PR.** 2005. An alternative to null-hypothesis significance tests. Psychol Sci **16:**345–353.
21. **Killeen PR.** 2006. Beyond statistical inference: a decision theory for science. Psychon Bull Rev **13:**549–562.
22. **Loftus GR.** 1996. Psychology will be a much better science when we change the way we analyze data. Curr Dir Psychol Sci **5:**161–171.
23. **Maxwell SE, Kelley K, Rausch JR.** 2008. Sample size planning for statistical power and accuracy in parameter estimation. Annu Rev Psychol **59:**537–563.
24. **Meehl PE.** 1967. Theory-testing in psychology and physics: a methodological paradox. Philos Sci **34:**103–115.
25. **Office of Laboratory Animal Welfare.** [Internet]. PHS policy on humane care and use of laboratory animals. Frequently asked questions. Question D9: what is considered a significant change to a project that would require IACUC review? [Cited 13 July 2010]. Available at: http://grants.nih.gov/grants/olaw/faqs.htm#d9.
26. **Office of Laboratory Animal Welfare, Applied Research Ethics National Association.** 2002. Institutional animal care and use committee guidebook, 2nd ed. Bethesda (MD): Office of Laboratory Animal Welfare.
27. **Shaw R, Festing MF, Peers I, Furlong L.** 2002. The use of factorial designs to optimize animal experiments and reduce animal use. ILAR J **43:**223–232.
28. **Ximenez C, Revuelta J.** 2007. Extending the CLAST sequential rule to one-way ANOVA under group sampling. Behav Res Methods **39:**86–100.