

# How many pigs? Statistical power considerations in swine nutrition experiments<sup>1</sup>

D. K. Aaron<sup>2</sup> and V. W. Hays

University of Kentucky, Lexington 40546-0215

**ABSTRACT:** Replication refers to the assignment of more than one experimental unit to the same treatment. Each replication of a treatment is an independent observation; thus, each replication involves a different experimental unit. In swine nutrition research, the experimental unit may be an individual animal, as in sow reproduction experiments, or a group of animals, as in growing-finishing pig experiments. In either case, calculation of the number of replicates needed to give an accurate and reliable outcome is an important step in deriving an experimental protocol. Although investigators often seem to choose replication arbitrarily on the basis of cost or availability of animals, housing considerations, convenience, or tradition, the question of how many pigs (i.e., how much replication is necessary) is a statistical one that has a statistical answer. A power analysis, performed while designing an experiment, will provide an investigator with an estimate of the number of replicates needed for an experiment of known

power and sensitivity. This a priori, or prospective, power analysis ensures that an investigator does not waste time and resources carrying out an experiment that has little chance of finding a significant effect, if one exists. It also makes sure resources are not wasted by including more experimental units than are necessary to detect an effect. A retrospective, or a posteriori, power analysis may also be conducted. If no significant effects are found in an experiment, an investigator can assess the observed power of the experiment, or may determine the size of treatment effect that could have been detected using the standard deviation and number of replicates in the experiment. The latter may be useful in explaining results. However, the former may be misleading because a high *P*-value will invariably result in a low observed power, and little new information will be gained from the post hoc power analysis. In most cases, the time for making power calculations is before, not after, an experiment is conducted.

Key Words: Replication, Statistical Power, Swine Nutrition

©2004 American Society of Animal Science. All rights reserved. J. Anim. Sci. 2004. 82(E. Suppl.):E245–E254

## Introduction

One of the questions most often posed to consulting statisticians is, “how many replications do I need?” If an investigator’s resources are limited by practical or economical constraints, the question may become, “will I be able to show a significant effect with a low number of replicates?” In either case, the question is really about statistical power, with the investigator’s main interest being to establish validity of the alternate, or research, hypothesis. Unfortunately, most researchers have had little exposure to the concept of statistical power, and many statistical texts provide only a cursory

discussion of the topic. Consequently, as noted by Kraemer and Thiemann (1987), investigators are generally unable to answer such questions themselves. Furthermore, they may not anticipate the information a consulting statistician needs to provide complete and accurate answers. The primary objective of this paper is to provide researchers in swine nutrition with the means to determine a valuable piece of information: how many pigs (or in many cases, pens of pigs) are needed for an experiment of known power and sensitivity. This a priori, or prospective, power analysis, conducted as part of a preexperiment protocol, will ensure that an investigator does not waste time and resources carrying out an experiment that has little chance of finding a significant effect, if one truly exists. It also makes sure resources are not wasted by including more replicates than are necessary to detect an effect. This is the most important type of power analysis. In many cases, however, the issue of power does not arise until an experiment has been conducted and no significant effects found. In interpreting such results, an investiga-

<sup>1</sup>This article was presented at the 2003 ADSA-ASAS-AMPA meeting as part of the Contemporary Issues symposium “Designing Animal Experiments for Power.”

<sup>2</sup>Correspondence: 208 W. P. Garrigus Bldg. (phone: 859-257-7553; fax: 859-257-5318; e-mail: daaron@uky.edu).

Received July 10, 2003.

Accepted September 24, 2003.

tor may wish to conduct a retrospective, or *a posteriori*, power analysis. Thus, the secondary objective of the present paper is to discuss potential uses and limitations of this second type of power analysis.

### Some Basics

An understanding of some basic statistical concepts and the essence of classical hypothesis testing is a necessary precursor to a discussion of statistical power. Consider the experimental unit; it is the smallest division of experimental material to which a treatment (or treatment combination in factorial experiments) can be assigned in a single act of randomization. In other words, it is the smallest entity receiving a treatment, provided two such entities could receive different treatments. In swine nutrition research, the experimental unit may be an individual animal, as in sow reproduction experiments, or a group (pen) of animals, as in growing-finishing experiments. Correct definition of the experimental unit is important because it is variation among experimental units treated alike that provides the unbiased estimate of error for evaluating treatment effects.

Replication refers to the assignment of more than one experimental unit to the same treatment. Each replication is an independent observation; thus, each replication involves a different experimental unit. Correct definition of the experimental unit determines the entity to be replicated. Consider the following example. An experiment is conducted to compare effects of four different diets on the performance of growing-finishing pigs. Four pens of the same size are available and each will house eight pigs of the desired age and weight. The investigator randomly assigns eight pigs to each pen and then randomly assigns diets to pens. The investigator believes “pig” is the experimental unit and that there are eight replications. However, because diets were assigned to pens, and all pigs in the same pen receive the same diet, “pen” constitutes the experimental unit. As a result, the experiment has no replication, and further assumptions are needed before valid conclusions can be drawn.

The basic function of replication is to provide an estimate of experimental error. In general, increasing the number of replications results in a more accurate estimate of experimental error and more precise estimates of treatment effects. For a particular experiment, the number of replications needed depends on the magnitude of the difference or effect to be detected at a specified significance level, the power, and the variability of the experimental material.

Power analyses, as discussed in this article, revolve around the investigator’s interest in establishing validity of the alternate, or research, hypothesis. Therefore, a review of classical hypothesis testing is useful. In classical hypothesis testing, a decision is made to either accept or reject the null hypothesis. Rejection of the null hypothesis points to validity of the alternate, or

**Table 1.** Possible outcomes of the classical hypothesis test

Decision	Null hypothesis	
	True	False
Accept null hypothesis	No error Probability = $1 - \alpha$	Type II error Probability = $\beta$
Reject null hypothesis	Type I error Probability = $\alpha$ (significance level)	No error Probability = $1 - \beta$ (power)

research, hypothesis. There are two possibilities for the underlying reality (Cousens and Marshall, 1987): either the null hypothesis is true or it is false. If the null hypothesis is accepted when it is true or, conversely, rejected when it is false, no error will have been made. This leaves two possibilities for error: a Type I error is made when a true null hypothesis is rejected; a Type II error is made when a false null hypothesis is accepted. Possible outcomes of the hypothesis test are often summarized in a contingency table (Table 1).

Classical hypothesis testing, as taught in introductory statistics courses, can be summarized in the following stepwise process.

**Step 1: *Formulate hypotheses.*** The null hypothesis is one of no difference, no effect, no change. It is the hypothesis to be tested and is the opposite of the alternate, or research, hypothesis, which represents what the investigator truly believes and seeks to verify.

**Step 2: *Decide on Type I error rate,  $\alpha$ .*** This rate is referred to as the *significance level* of the test. Because classical hypothesis testing is intended to reduce the risk of false claims,  $\alpha$  is set as low as possible. Typical yet arbitrary values are 0.01, 0.05, and 0.10.

**Step 3: *Collect and summarize data; calculate the appropriate test statistic ( $t$ ,  $F$ , etc.).***

**Step 4: *Determine the critical value of the test statistic given the value of  $\alpha$  and size of the experiment, identify the rejection region, and state the decision rule.*** If the test statistic falls within the rejection region, the null hypothesis will be rejected in favor of the alternate.

**Step 5: *Make decision and draw conclusions.***

**Step 6: *Compute  $P$ -value and interpret.*** The  $P$ -value is the observed significance level of the test. It is the probability (assuming the null hypothesis is true) of observing a test statistic as large or larger than that observed from the experimental, or sample, data. It describes the strength of evidence against the null hypothesis; the smaller the  $P$ -value, the stronger the evidence against the null hypothesis.

As noted previously, classical hypothesis testing is intended to reduce the risk of false claims. Thus,  $\alpha$  is set to as low a value as possible to guard against Type I errors. Conversely, establishing a value for  $\beta$ , the probability of a Type II error, is rare, and nonsignificance is often used to indicate a true lack of effect. As a result, Type II errors often occur (Table 1). As noted by Cousens and Marshall (1987), investigators must realize that a nonsignificant difference indicates that

if there is a real difference it is smaller than could be detected as significant in that particular experiment. It does not mean that treatments do not differ. Classical hypothesis testing is inappropriate if the intent is to prove that some treatment has no effect. In such cases, equivalence testing should be used (Dixon, 1998).

“Power” in this context refers to the ability of the experiment to detect real differences, if they exist, at the desired significance level. In other words, it is the probability ( $1 - \beta$ ; Table 1) of rejecting the null hypothesis when it is false. Given two experiments, the one that can detect the smaller treatment effect has the greater power. Size of the experiment, as measured by the number of replications, is highly dependent on the desired power. In turn, power depends on the magnitude of the difference to be detected, the significance level, and the size of experimental error. As  $\alpha$  is reduced, the probability of rejecting false, as well as true, null hypotheses is lowered. A reduced Type I error rate is accompanied by reduced power. The chance of detecting smaller differences is increased as power is increased, and this occurs when experimental error is decreased.

### The Use of Power Analyses in Swine Nutrition Research

Power calculations can be made during either the planning or the analysis stage of an experiment. In either stage, essential information includes 1) significance level, 2) size of the difference or effect to be detected, 3) power to detect the effect, 4) variation in response, and 5) number of replications or sample size. These five pieces of information are not independent; any four of them automatically determines the fifth. This forms the basis for the power calculations that will be discussed here. The *a priori* approach, which occurs during the planning stage of an experiment, is the most important and will be addressed first. During this discussion, it will be assumed that the treatment and design structures of the experiment are known and that the only remaining question is the number of replications. Second, the *a posteriori* approach, which occurs during the analysis stage, will be discussed as a means of interpreting nonsignificant results. In both approaches, discussion will be limited to completely randomized designs, such as might be used in sow reproduction experiments, and randomized complete block designs, such as might be used in growing-finishing pig experiments. In the former, the individual sow will be considered the experimental unit; in the latter, a pen of pigs will be the experimental unit.

#### Planning

The *a priori* approach (prospective power analysis) provides an estimate of the number of replications necessary for an experiment of known power and sensitivity, where “sensitivity” refers to the minimal treatment

response that will be detectable (Berndtson, 1991). It helps ensure that an investigator does not waste time and resources on an experiment that has little chance of finding a significant effect, if one truly exists, and ensures that resources are not wasted by including more replications than necessary to detect an effect. This is especially important with livestock, such as swine, for which the cost per replicate is high, both in terms of capital required to purchase and house the animals and the labor needed to care for them and collect data (Morris, 1999). Ethical considerations may also come into play (Erb, 1990; Van Wilgenburg et al., 2003). For example, in experiments involving surgical modifications, either using so few animals that the smallest important effect cannot be observed or using more animals than are needed may be considered an improper use of the animals and a waste of research funds. In reality, using more replications than are needed is rarely a problem in studies involving livestock.

An *a priori* power analysis is a critical part of a good experimental design; however, Morris (1999) observed that it is frequently the most neglected part of a preexperimental protocol. In reality, investigators seem to choose the number of replicates arbitrarily on the basis of cost or availability of animals, housing considerations, convenience, or tradition. Nevertheless, the answer to the question of how many pigs is a statistical one. Reasons given by investigators for *not* conducting a prospective power analysis include “I’ve always used this many,” “I don’t know what variance to use,” “I don’t know how big a difference to expect,” “These are all the pigs (pens) available,” and “I don’t know how.” The overriding reason, however, is that they know that power calculations will likely yield much larger numbers than are desirable, feasible, or even possible, given the resources available.

These excuses point to six hurdles that must be overcome if an *a priori* power analysis is to be conducted. First, the number of treatments must be determined. The choice of treatments that are to be compared is one of the most difficult and important decisions made during the planning stage. Ambiguous results can generally be avoided if the choice of treatments is based on the specific objectives of the experiment. Expense, time, and level of complexity are also important. Second, hypotheses must be specified. As with the choice of treatments, this is a function of the experiment’s objectives. Third, the level of significance must be chosen. Typically, but arbitrarily, levels of 0.01, 0.05, and 0.10 are used, but whatever level of significance, the choice should be based on the relative consequences of rejecting the null hypothesis when it is true (Type I error) and accepting the null hypothesis when it is false (Type II error). The other three hurdles include deciding on the smallest difference that is worth detecting in a particular experiment, selecting the appropriate power, and estimating the variability in response. These are more difficult to overcome and, unlike the first three,

are not usually addressed in introductory statistics courses.

In deciding on the minimal treatment response to be detected, an investigator must consider what might be biologically or economically meaningful. This is often difficult to ascertain and depends on the specific situation; however, the minimal treatment effect must be large enough to make the experiment worthwhile biologically and/or economically. As the size of the difference to be detected decreases, the number of replications required for an experiment of known power will increase. For example, if a difference in a rate of gain of 40 g/d can be detected with 4 replicates, an experiment of approximately 16 replicates will be needed to detect half of this difference, 20 g/d, because the standard errors are in the ratio 2:1 (assuming the same variance). However, if this difference is halved again, to 10 g/d, approximately 64 replicates will be necessary to detect this difference. An investigator will have to determine whether the smaller difference is worth the cost of increased replication.

In swine nutrition experiments, guidelines for determining expected differences may be obtained from previous work. For example, in sow reproduction studies, estimated average litter size is 12 to 14 pigs at birth and 10 to 12 pigs at weaning (personal communication, G. L. Cromwell, University of Kentucky). Expected differences are typically in the range of 5 to 10% of the mean, which would be 0.6 to 1.2 pigs per litter for an average litter size of 12 pigs. In growing-finishing pig experiments, normal ADG is approximately 0.80 kg/d (personal communication, G. L. Cromwell). For practical diets, with no deficiencies, an expected difference of 5 to 10% would be 0.04 to 0.08 kg/d. A difference of 10% is generally assumed to be meaningful. If one of the treatment groups consists of a negative control, in which pigs are deficient in some nutrient, the expected difference may be as large as 50%. In these examples, the minimum difference to be detected has been expressed as a percentage of the mean response.

In determining the appropriate power, the idea is to have a reasonable chance of detecting the stated minimum difference. A target power of 80% is common and can be used as a minimal value. Some statisticians argue for higher powers, such as 85, 90, or even 95%. As power increases, however, the required number of replications increases. Therefore, it is rare with animal experiments to set power at values larger than 80%.

The last, and most difficult, hurdle to overcome involves the question of how much variability. An estimate of the amount of variability in the response variable may be obtained by making an educated guess, by conducting a pilot study, or by reviewing previous experiments conducted by the investigator or reported in the scientific literature. The reliability of estimates obtained from previous experiments depends on the following: the similarity of the populations; the number of observations upon which estimates are based; the experimental protocol (instrumentation and how,

**Table 2.** Estimated variability in sow reproduction experiments

Trait	Mean CV, % <sup>a</sup>	Range in CV, % <sup>a</sup>
Total pigs born	28.0 <sup>b</sup>	26.0–30.0 <sup>b</sup>
No. of live pigs at birth	29.0 <sup>b</sup>	28.0–33.0 <sup>b</sup>
	28.8 <sup>c</sup>	12.2–58.1 <sup>c</sup>
Birth wt	18.0 <sup>c</sup>	4.2–67.2 <sup>c</sup>
No. of live pigs at 21 d	36.0 <sup>b</sup>	29.0–39.0 <sup>b</sup>
No. of pigs weaned	36.9 <sup>c</sup>	20.9–66.4 <sup>c</sup>

<sup>a</sup>Coefficient of variation = (pooled standard deviation/mean) × 100. Means and standard deviations based on sow as experimental unit.

<sup>b</sup>Based on five estimates involving 7,925 litters (our unpublished results).

<sup>c</sup>Based on 24 experiments reported in the scientific literature from 1963 to 1978; 9,445 litters for number of live pigs at birth; 9,467 for birth weight; 8,644 litters for number of pigs weaned.

when, and who administers treatments and collects data, feeding management, type of experimental unit, time at which measurements are taken, etc.); the design structure (blocking); sources of variation (such animal attributes as gender, age, and breed); and whether or not covariates are included in the analysis. Treatments do not have to be the same or even similar because the error mean square provides an estimate of the variation among replicates *within* treatments. Furthermore, the analysis of variance requires the assumption of a homogeneous variance; that is, the variance among replicates within treatments is assumed to be the same for all treatments (Steel and Torrie, 1980). Finally, previous experiments do not have to follow the same design structure as the planned experiment, but the investigator must be careful that the correct variance is being estimated (Lenth, 2001a).

The absence of a previous estimate of variability is never a valid reason for not calculating the number of replicates required. Guidelines for estimating the coefficient of variation of future experimental material are presented by Morris (1999). Among mammals, the coefficient for growth rate is approximately 12%; for reproductive traits, 20 to 40%; for characteristics such as milk yield, 20 to 25%; and for linear measures, such as bone length or height, 6%. Also, as noted by Morris (1999) the coefficient of variation of a trait is inversely related to its heritability. Thus, for traits that are highly heritable (e.g., carcass traits), the coefficient of variation would be lower than for traits that are lowly heritable (e.g., reproductive traits). These guidelines provide investigators with a general idea of the amount of variability for different types of traits across species.

More specifically, estimates of variability for response traits measured on a particular species can be obtained by averaging values calculated from previous experiments. Based on data collected at the University of Kentucky (our unpublished results), as well as a review of experiments reported in the scientific literature, estimates of variability were obtained for some of the response variables typically measured in swine nutrition experiments. For the most part, these esti-

**Table 3.** Estimated variability in growing-finishing pig experiments

Trait	Mean CV, % <sup>a</sup>	Range in CV, % <sup>a</sup>
ADG	4.8 <sup>b</sup> 8.3 <sup>c</sup>	2.0–10.9 <sup>b</sup> 2.6–20.3 <sup>c</sup>
ADFI	7.3 <sup>c</sup>	2.1–15.1 <sup>c</sup>
Feed efficiency	4.1 <sup>b</sup> 7.3 <sup>c</sup>	2.0–7.4 <sup>b</sup> 2.1–18.9 <sup>c</sup>

<sup>a</sup>Coefficient of variation = (pooled standard deviation/mean) × 100. Means and standard deviations based on pen as experimental unit.

<sup>b</sup>Summary of 32 experiments (3,425 group-fed pigs, 4 to 8 pigs/group) conducted at University of Kentucky (our unpublished results).

<sup>c</sup>Summary of 33 experiments (3,680 group-fed pigs) reported in scientific literature from 2000 to 2003.

mates are compatible with guidelines presented by Morris (1999). For sow reproduction experiments, a completely randomized design is assumed because it is common to assign treatments completely at random to individually fed sows. The average coefficients of variation for total number of pigs born; number of live pigs at birth; birth weight; number of live pigs at 21 d; and number of pigs weaned are 28, 29, 18, 36, and 37%; respectively (Table 2). These values are based on 29 estimates involving in excess of 16,500 litters. For growing-finishing pig experiments, a randomized complete block design is assumed because experimental units are commonly blocked on factors such as initial age, weight, and gender. Pen is considered the experimental unit. The average coefficients of variation for rate of gain and feed efficiency range from 5 to 8 and 4 to 7%, respectively (Table 3). For very young pigs, these values may range from 7 to 13% (our unpublished results).

Once the above difficulties have been surpassed and the required information obtained, an a priori power analysis can be conducted using formulas (Cochran and Cox, 1957; Steel and Torrie, 1980; Cohen, 1988), reference tables (Cochran and Cox, 1957; Gill, 1978; Steel and Torrie, 1980; Kraemer and Thiemann, 1987; Berndtson, 1991; Aaron and Hays, 2001), or computer software. Tables are typically constructed assuming particular experimental designs and assume two-tailed tests of significance. For example, Cochran and Cox (1957) assume four treatments in a randomized complete block design, whereas Berndtson (1991) assumes two treatments in a completely randomized design. Tables 4 and 5, adapted from Aaron and Hays (2001), assume a completely randomized design (two treatments) and a randomized complete block design (four treatments), respectively. To use tabular values, one must have determined the expected difference as a percentage of mean response, the significance level, and the coefficient of variation. Replication requirements can be obtained directly, or by interpolation, from such tables for any coefficient of variation and experimental sensitivity combination. Also, although sows or pens of pigs are used as experimental units in Tables 4 and 5, respectively, replication frequently involves other types of experimental unit (such as with in vitro studies) for which tabular values are equally appropriate. As noted by Berndtson (1991), however, a distinction must be made between experimental units (i.e., replicates) and sampling units (i.e., measurements within replicates).

Numerous computer software programs exist for calculating the number of replications for an experiment of known power and sensitivity. One program that is readily available on the internet is G\*Power (<http://>

**Table 4.** Estimated number of replications needed in sow reproduction experiments<sup>a,b</sup>

Significance level, % <sup>c</sup>	Average CV <sup>d</sup>	Expected difference in litter size, % of mean <sup>e</sup>				
		4	6	8	10	12
1	40	2,336	1,039	584	374	260
	35	1,788	795	448	287	199
	30	1,314	584	329	211	146
	25	913	406	229	146	102
	20	584	260	146	94	67
5	40	1,570	698	393	252	175
	35	1,202	533	301	193	134
	30	883	393	221	142	99
	25	614	273	154	99	69
	20	393	175	99	63	45
10	40	1,237	550	310	198	138
	35	947	421	237	152	106
	30	696	310	174	112	78
	25	483	215	121	78	55
	20	310	138	78	50	36

<sup>a</sup>Reproduced from Aaron and Hays (2001).

<sup>b</sup>Assumes a completely randomized design with two treatments, two-tailed test of significance, and power of 80%.

<sup>c</sup>Probability of rejecting null hypothesis when it is true.

<sup>d</sup>Coefficient of variation = (pooled standard deviation/mean) × 100.

<sup>e</sup>A 10% difference in litter size represents approximately 1 and 0.8 pig at birth and weaning, respectively.

**Table 5.** Estimated number of replications (blocks) needed in growing-finishing pig experiments<sup>a,b</sup>

Significance level, % <sup>c</sup>	Average CV <sup>d</sup>	Expected difference in rate of gain or feed efficiency, % of mean				
		5.0	7.5	10.0	12.5	15.0
1	15.0	211	97	54	35	25
	12.5	146	67	38	25	18
	10.0	97	43	25	17	12
	7.5	54	25	15	10	7
	5.0	25	12	7	5	4
	2.5	7	4	3	3	3
5	15.0	141	64	36	23	17
	12.5	99	44	26	17	12
	10.0	64	29	17	11	8
	7.5	36	17	10	7	5
	5.0	17	8	5	4	3
	2.5	5	3	3	2	2
10	15.0	111	51	29	19	13
	12.5	73	35	20	13	10
	10.0	51	23	13	9	7
	7.5	29	13	8	6	4
	5.0	13	7	4	3	3
	2.5	4	3	2	2	2

<sup>a</sup>Reproduced from Aaron and Hays (2001).

<sup>b</sup>Assumes a randomized complete block design with four treatments, two-tailed test of significance, and power of 80%.

<sup>c</sup>Probability of rejecting null hypothesis when it is true.

<sup>d</sup>Coefficient of variation = (pooled standard deviation/mean) × 100. For growing-finishing pigs (group-fed, four to eight pigs/group), values range from 2.5 to 10%. For very young pigs, values range from 7.5 to 15%.

[www.psych.uni-duesseldorf.de/aap/projects/gpower/](http://www.psych.uni-duesseldorf.de/aap/projects/gpower/)). This program is free to download, well-documented, easy to use, and will do both a priori and post hoc power calculations. The input values required include an estimate of the standardized effect size (Cohen, 1988), chosen level of significance, desired power, and number of treatments. Also, available on the internet is a number of replications applet (<http://duke.usask.ca/~rbaker/NumReps.html>). This applet evaluates an equation of Cochran and Cox (1957) and assumes a randomized complete block design. However, results should be very close to those required for a completely randomized design. The following items are required input: type of test (one-tailed or two-tailed), significance level of the proposed test, coefficient of variation, expected difference expressed as a percentage of mean response, number of treatments, and required power. The applet is easy to use and has the advantage that the minimum detectable difference is expressed as a percentage of mean response as opposed to being expressed as a standardized effect size. More will be said about this later.

To illustrate a priori power calculations, consider a proposed sow reproduction experiment. Sows will be individually fed two diets. The response variable is litter size as measured by number of pigs born alive. A coefficient of variation of 29% (Table 2) is assumed, and the investigator would like to be able to detect a difference of at least 10% with 80% power and a significance level of 5%. Using Table 4, the estimated number of sows per treatment is found to be approximately 142. This yields an experiment of size 284 (142 sows for

each of two treatments). Thus, for a population with a coefficient of variation of 29%, 142 sows will be needed per treatment to provide an 80% chance that a 10% treatment response will be detected at the 5% significance level. If the power were increased to 90%, 190 sows per treatment would be required. If it were increased to 95%, 234 sows per treatment would be required. As the power is increased from 80 to 90% and from 90 to 95%, the number of replications required increases by 34 and 65%, respectively. This illustrates how rapidly the number of replications increases as power increases.

Using G\*Power is only slightly more complex. It requires an estimate of the standardized effect size (Cohen, 1988). The standardized effect size ( $f$ ) is calculated as the ratio of the variance of the treatment population means to the population error variance (Myers and Well, 2003) and is an adjunct to the analysis of variance. For a completely randomized design, the effect size can be calculated from a previous experiment as follows:

$$f = \sqrt{\frac{(t-1)(MS_T - MS_E)}{tr(MS_E)}} \quad [1]$$

where  $t$  = number of treatments,  $r$  = number of replications, and  $MS_T$  and  $MS_E$  refer to the treatment and error mean squares, respectively. Expressed more simply,

$$f = \sqrt{\frac{(t-1)(F_T-1)}{tr}} \quad [2]$$

where  $F_T$  is the ratio of  $MS_T$  to  $MS_E$ . Cohen (1988) has suggested that  $f = 0.10, 0.25,$  and  $0.40$  correspond to small, medium, and large effect sizes, respectively. This convention is somewhat controversial. Proponents claim two advantages: 1) pilot or previous data are not required to estimate the variance and 2) the standards for small, medium, and large provide realistic norms (at least in the social sciences). Opponents claim that a standardized effect size, such as  $f$ , has no relevance as a criterion for determining sample size (Elashoff, 2000). And a standardized effect size is harder to relate to the actual units of measure for a particular response variable. For example, given the proposed sow reproduction experiment, what does a small effect size mean as it relates to a difference in number of live pigs born per litter? a medium effect size? a large effect size?

To find the required number of replications for the sow reproduction experiment using G\*Power, the following steps are taken

1. Under "Tests," select "F Tests (ANOVA)"
2. Under "Analysis," select "A priori"
3. Fill in these values:
  - $f = ?$
  - Alpha = 0.05
  - Power = .80
  - Groups = 2
4. Click on "Calculate"

The result will be the total sample size. This will not be exact because G\*Power requires the difference to be expressed as the standardized effect size, and the criteria for the proposed sow reproduction experiment specified the minimum detectable difference as 10% of the mean. Thus, a trial-and-error approach is taken. If a large effect size ( $f = 0.40$ ) is used, the total sample size,  $n$ , is 52, or  $r = 26$  sows per treatment. If a medium effect size ( $f = 0.25$ ) is used,  $n = 128$  and  $r = 64$  sows per treatment. If a small effect size ( $f = 0.10$ ) is used,  $n = 788$  and  $r = 394$  sows per treatment. As discussed previously, the smaller the size of the difference to be detected, expressed here as the standardized effect size, the more replications that will be required. Using Table 4, the number of replications for the proposed experiment was approximated to be 142 sows per treatment. Comparing this value with the number approximated for a small effect size (394) and a medium effect size (64), it could be concluded that a 10% difference corresponds to an effect size somewhere between small and medium.

As a second example, consider a growing-finishing pig experiment being designed to compare the ADG of pigs fed four diets in a randomized complete block design. Pigs will be group-fed (4 to 8 pigs per pen), with pens being blocked according to weight and gender (a

chunk-type blocking factor as described by Ott [1975]). The investigator would like to have an 80% chance of detecting a minimum difference of 10% at a 5% significance level. A coefficient of 8% (Table 3) is assumed. Because each block is a replication, the question actually is, how many blocks are required? Assuming sphericity, power calculations are conducted as with a completely randomized design. Using Table 4, it is determined that approximately 11 replications (blocks), or a total of 44 pens, will be required to provide an 80% chance that a 10% treatment response would be detected at the 5% level of significance.

The number of replications applet will yield the same estimate. Using the applet, the following values are input:

Choose 1- or 2-tailed test:	2
Enter proposed significance level:	0.05
Enter expected coefficient of variation:	8.0
Enter expected difference (as % of mean):	10.0
Enter number of treatments:	4
Enter desired power:	0.80
Click on "Calculate Number of Replications."	
The result is 11.	

A direct answer cannot be obtained for randomized complete block experiments using G\*Power. Instead, a trial-and-error approach (select "Other  $F$  tests") must be employed.

The number of replications required is often quite large, and in reality, resources are likely to be limited. If the number of replicates calculated is prohibitively large, an investigator has several options. First, ask whether the difference is too small to be of biological or economical importance. If the minimum detectable difference is smaller than necessary, the planned study may be more powerful than necessary. Second, can the Type I error rate be increased from 0.01 to 0.05? Can it be increased from 0.05 to 0.10? In other words, how serious would the consequences be if the null hypothesis is rejected when it is actually true? Third, is the power set too high? Typically, 80% is used as a minimal value. Finally, is the variability of the response variable overestimated? Could it be reduced?

If it is determined that all input values are appropriate, there are three possibilities. First, find a satisfactory way to reduce the amount of variability. This might be accomplished by choosing a more precise measuring device, using more homogeneous experimental material, choosing a design structure that will provide more error control (i.e., blocking), or by using one or more covariates to account for extraneous variability. Second, amend treatments so that a larger difference can be expected. Third, go back to the drawing board and narrow the scope of the experiment or argue for a bigger budget.

In many cases, the investigator may ask, what good is a power analysis if I have only a set number of pigs available? When the maximum sample size is set by

spatial, physical, or economical constraints, a power analysis may be useful to determine whether sufficient power exists for specified values of the level of significance, sensitivity required, and anticipated variation. Ultimately, the investigator may have to determine whether the experiment is worth pursuing. The goal should be to establish balance in what is statistically desirable and what is practically possible.

*Other Considerations.* In conducting an a priori power analysis, there are other important considerations. The number of treatments must be decided on. Additional treatments increase total sample size, but they also increase error degrees of freedom and, as a result, may decrease the size of the error variance. This, in turn will increase the power of the experiment. However, according to Berndtson (1991), the effect on the number of replicates is minimal.

The type of means comparison procedure may also affect power calculations. Gill (1989) estimates that replication requirements may be reduced by as much as 20 to 30% if orthogonal contrasts are used instead of pairwise comparisons. Ultimately, this decision should be made based on the nature of the treatments (Aaron and Hays, 2001). One-tailed vs. two-tailed tests could be considered. Using a one-tailed test will reduce replication requirements within limits, but, given that the outcome of the experiment is unknown in advance, a two-tailed test is often more appropriate.

The choice of experimental unit is another factor that affects power in an experiment. As discussed earlier, the experimental unit can be an individual or a group. It is the smallest unit receiving a single treatment provided two such units could receive different treatments. In some experiments, pen is correctly recognized as the experimental unit for response variables like ADFI or gain:feed, which are measured on the pen of pigs, but not for traits such as longissimus dorsi area or plasma urea nitrogen, which are measured on the individual pig. Pen is both the experimental unit and the sampling unit for ADFI and G:F, but for traits such as loin muscle area or plasma urea nitrogen, pen is the experimental unit and pig the sampling unit. In either case, replication involves increasing the number of pigs per treatment. Increasing the number of pigs per pen (i.e., within-replicate sampling) will increase power up to a point, but, if a given number of pigs can be fed individually, power can be increased more by using the individual pig as the experimental unit and having more replications than by using the same number of pigs in fewer experimental units. Treating pigs individually maximizes the use of animals, but maximizing the efficiency of total resources may necessitate the grouping of pigs. In addition, pigs treated individually may not respond to treatment in the same manner as pigs treated in groups. Specifically, variability in response may be greater among individually fed pigs than among group-fed pigs.

Special considerations are also required for more-complex designs, factorial treatment structures, experi-

ments involving repeated measures, use of covariates, and experiments involving multiple response variables. The last is more often than not the case with animal experiments. Multiple response variables will exhibit different degrees of variation, and thus will require different numbers of replications. To further complicate things, different minimum detectable differences may be of interest. The simplest approach is to take the response variable expected to express the greatest degree of variability and having the smallest minimum detectable difference and calculate the number of replications required for known power and sensitivity for that variable. For example, in a sow reproduction experiment with two dietary treatments, both number of live pigs at birth and litter birth weight may be of interest. The number of live pigs at birth is expected to have a coefficient of variation of approximately 29%, whereas the corresponding value for litter birth weight is only 18%. For number of live pigs at birth, it was herein determined that approximately 142 sows per treatment would be required to provide an 80% chance of detecting a 10% treatment response at a 5% significance level. For litter birth weight, however, only 52 sows per treatment would have been required to detect a 10% treatment response. With 142 sows per treatment, the power for detecting a 10% difference in litter birth weight would be over 95%. Whereas power is overestimated for the trait with the smaller coefficient of variation, a minimum power is ensured for the trait with the larger coefficient of variation.

### *Interpreting Results*

A second kind of power analysis may also be useful. If no significant effects are found in an experiment, the investigator can assess post hoc the observed (actual) power of the experiment or determine the size of treatment effect that could have been detected using the standard deviation and number of replicates in the experiment. If used appropriately, this a posteriori, or retrospective, power analysis can be very useful in interpreting results.

Interpretation of results is not difficult when all treatments being compared are significantly different from each other. However, when nonsignificant results are observed, care must be exercised in interpreting the results. A common mistake is to say the treatment effect was nonsignificant and, therefore, does not exist. A nonsignificant statistical result does not mean a difference does not exist.

To calculate the observed power of an experiment, the existing analysis is used to estimate both the effect size and population variance. Consider the following hypothetical sow reproduction experiment. Two dietary treatments (0 and 200 ppb Cr) were randomly allotted to individually fed sows such that there were 10 sows per treatment. The response variable was litter size as measured by number of pigs born alive. Upon completion of the experiment, an *F*-ratio of 1.92 with a corres-

**Table 6.** Analysis of variance for a hypothetical experiment evaluating the effect of Cr supplementation on number of pigs born alive

Source of variation	df	Sum of squares	Mean squares	F-ratio	P-value
Treatment	1	3.2	3.20	1.92	0.1828
Error	18	30.0	1.67		
Total	19	33.2			

ponding  $P$ -value of 0.1818 was obtained (Table 6). Clearly, this was not significant at a 5% level of significance. The observed difference between means was 0.8 pigs per litter (10.8 vs. 11.6). Using Eq. [2], the estimated effect size is calculated to be 0.17. Using the Cohen (1988) guidelines, this is judged to be of small size.

Observed power can be calculated by G\*Power using the following steps:

1. Under “Tests,” select “F Tests (ANOVA)”
2. Under “Analysis,” select “Post hoc”
3. Fill in these values from the completed experiment:  
 $f = 0.17$   
 $\text{Alpha} = 0.05$   
 $\text{Total sample size} = 20$   
 $\text{Groups} = 2$
4. Click on “Calculate”

The results are Critical  $F = 4.4139$  and Power = 0.1112. Thus, the observed power is approximately 11%. This might be interpreted (Myers and Well, 2003) as follows: given the sample size used in the present experiment, and assuming the mean square error is a reasonable estimate of the population variance, there is a probability of about 0.11 that the null hypothesis will be rejected if treatment effects in the population are of the order of magnitude estimated from the present data. In other words, if the treatment population means are about as different as the sample means suggest, there is still a 0.89 ( $1 - 0.11$ ) probability of making a Type II error.

Calculation of the observed power is controversial. Myers and Well (2003) present post hoc power calculations as a useful supplement to hypothesis testing in cases where significant results are not detected. If the observed power to detect an effect of the size found in the experiment is high, it can be safely concluded that the treatment has no effect. If the observed power is low, results will not be sufficient to say there is no effect. Although this sounds reasonable, there is one major problem. As the  $P$ -value increases, the observed power decreases (Thomas, 1997; Hoenig and Heisey, 2001; Lenth, 2001a; Quinn and Keough, 2002). In fact, Hoenig and Heisey (2001) demonstrated that the observed power has a 1:1 relationship with the  $P$ -value. Therefore, when the  $P$ -value is high, the observed power will always be low, and vice versa. It can be shown that

when the observed  $P$ -value is equal to the preselected value of  $\alpha$ , the observed power is approximately 0.50. A good analogy was made by Lenth (2001b): “If my car made it to the top of a hill, then it is powerful enough to climb that hill; if it didn’t make it to the top of that hill, then it obviously isn’t powerful enough.” Post hoc power calculations are like that; they are not necessarily wrong—they just do not contribute new information and may be somewhat misleading.

A more appropriate use of an a posteriori power analysis is for calculation of the minimum detectable difference for a given power. This can be done using results of the last sow reproduction experiment and interpolating from the tabular data presented in the first three tables of Berndtson (1991). Assuming the coefficient of variation in the population was approximately 12%, ( $[\sqrt{1.67}]/11.2 \times 100$ ), at least a 17% treatment response

would have been needed for 80% certainty of statistical significance at the 5% level with 10 sows per treatment. Corresponding differences of 20 and 22% would have been needed for experiments of 90 and 95% power, respectively. From these calculations, it is conceivable that rather large treatment differences (two or more pigs per litter) might go undetected in an experiment of this size with this amount of variability. Thus, the nonsignificant result may be due to the true absence of a treatment response or may be a reflection of inadequate power and sensitivity. Results of the experiment should be regarded as inconclusive and the nonsignificant result interpreted cautiously. Rather than claiming no difference between the two dietary treatments, it should simply be concluded that, given the size of the experiment, there is insufficient evidence for declaring a treatment effect.

## Conclusion

The time for power calculations is before an experiment is conducted. An a priori power analysis will determine the necessary resources, in terms of number of replications, and will provide a means of determining whether the experiment, as planned, is feasible. However, investigators must remember that all calculations are based on estimated values. Accordingly, the calculated number of replications should also be considered an estimate. Furthermore, calculations tell an investigator how many replications are needed at the end of the experiment. If death losses or other experimental problems are anticipated, more replicates will be needed to begin the experiment. Additionally, the values of  $\alpha$  and  $\beta$  are arbitrary. These should be based on the relative consequences of Type I and Type II errors; more often than not, they are simply conventional values. The minimum detectable difference may also be arbitrary. Ideally, it is the smallest difference that would be biologically or economically meaningful. In practice, this value may be hard to define. Although these admonishments may seem to imply that the num-

ber of replicates generated by the power analysis is little more than a guess, the bottom line is that it is the best guess possible. And, the more care that is exercised in obtaining the necessary input values—whether to be used with reference tables or computer software—the better that guess will be.

Finally, the discussion herein has assumed that the design and treatment structures appropriate for a given experiment are known and that the only remaining question is the number of replications. As part of the preexperimental protocol, careful consideration should be given to the treatments to be included and the experimental units to which the treatments will be applied. These experimental units are the replicates for conducting the hypothesis test. Also, the design structure should be selected such that adequate error control is provided, given the nature of the experimental material. Investigators should recognize the wisdom of Cousins and Marshall (1987): “Good experimentation depends as much on an understanding of statistics as it does on an unraveling of the biology.”

### Implications

Careful planning and organization before an experiment is begun can maximize the amount of information gained. Investigators should focus on using an a priori, or prospective, power analysis to determine the appropriate number of replications, during the planning stage of an experiment. When nonsignificant results are found during the analysis stage, an a posteriori, or retrospective, power analysis can be useful for determining the size of the difference that could have been detected given the size of the experiment and the observed variability. Calculating observed power after an experiment has been conducted and nonsignificant results obtained contribute no new information, and may, in fact, be misleading.

### Literature Cited

Aaron, D. K., and V. W. Hays. 2001. Statistical techniques for the design and analysis of swine nutrition experiments. Pages 881–

- 901 in *Swine Nutrition*. 2nd ed. A. J. Lewis and L. L. Southern, ed. CRC Press, Boca Raton.
- Berndtson, W. E. 1991. A simple, rapid and reliable method for selecting or assessing the number of replicates for animal experiments. *J. Anim. Sci.* 69:67–76.
- Cochran, W. G., and G. M. Cox. 1957. *Experimental Designs*. 2nd ed. John Wiley and Sons, New York.
- Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ.
- Cousens, R., and C. Marshall. 1987. Dangers in testing statistical hypotheses. *Ann. Appl. Biol.* 111:469–476.
- Dixon, P. M. 1998. Assessing effect and no effect with equivalence tests. Pages 275–301 in *Risk Assessment: Logic and Measurement*. M. C. Newman and C. L. Strojjan, ed. Ann Arbor Press, Chelsea, MI.
- Elashoff, J. 2000. nQuery Advisor Release 4.0. Statistical Solutions. Software for MS-DOS Systems. Cork, Ireland.
- Erb, H. N. 1990. A statistical approach for calculating the minimum number of animals needed in research. *ILAR News* 32(1):11–16.
- Gill, J. L. 1978. *Design and Analysis of Experiments in the Animal and Medical Sciences*, Vol. 3. Iowa State University Press, Ames.
- Gill, J. L. 1989. Statistical aspects of design and analysis of experiments with animals in pens. *J. Anim. Breed. Genet.* 106:321–332.
- Hoenig, J. M., and D. M. Heisey. 2001. The abuse of power: The pervasive fallacy of power calculations in data analysis. *The American Statistician* 55:19–24.
- Kraemer, H. C., and S. Thiemann. 1987. *How Many Subjects?* Sage Publications, Newbury Park, CA.
- Lenth, R. V. 2001a. Some practical guidelines for effective sample size calculation. *The American Statistician* 55:187–193.
- Lenth, R. V. 2001b. Two Sample-Size Practices That I Don't Recommend. Available: <http://www.stat.uiowa.edu/~rlenth/Power/>. Accessed Mar. 24, 2003.
- Myers, J. L., and A. D. Well. 2003. *Research Design and Statistical Analysis*. 2nd ed. Lawrence Erlbaum Associates, Inc., Mahwah, NJ.
- Morris, T. R. 1999. *Experimental Design and Analysis in Animal Sciences*. CABI Publishing, New York.
- Ott, E. R. 1975. *Process Quality Control*. McGraw-Hill, New York.
- Quinn, G., and M. Keough. 2002. *Experimental Design and Data Analysis for Biologists*. Cambridge University Press, Cambridge.
- Steel, R. G. D., and J. H. Torrie. 1980. *Principles and Procedures of Statistics: A Biometrical Approach*. 2nd ed. McGraw-Hill Book Co., New York.
- Thomas, L. 1997. Retrospective power analysis. *Conserv. Biol.* 11:276–280.
- Van Wilgenburg, H., P. G. van Schaick Zillesen, and I. Krulichova. 2003. Sample Power and ExpDesign: Tools for improving design of animal experiments. *Lab. Anim.* 32(3):39–43.