

# Power of test considerations for beef cattle experiments: A review<sup>1,2</sup>

C. R. Richardson<sup>\*3</sup>, G. A. Nunnery<sup>\*</sup>, D. B. Wester<sup>†</sup>, N. A. Cole<sup>‡</sup>, and M. L. Galyean<sup>\*</sup>

<sup>\*</sup>Departments of Animal and Food Sciences and <sup>†</sup>Range, Wildlife, and Fisheries Management, Texas Tech University, Lubbock 79409 and <sup>‡</sup>ARS,USDA, Bushland, TX 79012

**ABSTRACT:** The use of power tests in the planning and design of beef cattle experiments provides critical information on sample sizes necessary to detect treatment differences at a predetermined significance ( $\alpha$ ) level. Retrospective power analysis provides additional information about previous experiments that may be helpful in designing subsequent investigations. However, in retrospective power analyses, power is inversely related to observed significance level. Benefits of prospective and retrospective power analyses in beef cattle experiments are similar to those for other species; however, because of differences in the methods and conditions involved, considerations for the use of power test procedures are specific for beef cattle research. Retrospective power analyses were conducted on 78 published experiments and on two unpublished experiments. Experiments were compiled into categories that represented group (or pen) feeding, individual feeding,

and metabolism studies. Estimated power in pen feeding experiments using randomized block designs (RBD,  $n = 30$ ) was less than 0.80 for ADG and feed efficiency (FE), but not different from 0.80 for completely random designs (CRD,  $n = 4$ ). Furthermore, estimated power was less for ADG than for FE in both design types. For individual feeding experiments using RBD ( $n = 4$ ), power was not different from 0.80 for either ADG or FE; however, for CRD ( $n = 18$ ), power was less than 0.80 for both ADG and FE. Power was similar for ADG and FE for both RBD and CRD in individual feeding experiments. In metabolism experiments, estimated power for nitrogen retention was less than 0.80 for Latin square designs ( $n = 20$ ) but not for CRD ( $n = 4$ ). Comparisons of power between experimental design types were likely influenced by the number of experiments involved. These results indicate that retrospective power in beef cattle experiments is affected by design type, and response variable measured.

Key Words: Beef Cattle, Experimental Design, Statistics

©2004 American Society of Animal Science. All rights reserved. J. Anim. Sci. 2004. 82(E. Suppl.):E214–E222

## Introduction

Power of the test considerations used in beef cattle research studies are distinct from those for other livestock species because of major differences in the methods and conditions involved. Sample size, number of true replications, and experimental design have been discussed in beef cattle research for many years (Henderson, 1969; Norton, 1969; Gill, 1980). Several considerations have been used in designing experiments to decrease experimental error and the probability of com-

mitting a Type II error (Meyer et al., 1960; Lofgreen et al., 1962). A primary goal of beef experimentation is to reject the null hypothesis of no difference between treatments when it is false; however, many beef experiments seem to be conducted without apparent calculations of prospective power of the test considerations, which decreases the value of the planning phase of experimentation and may result in a need for additional studies. In addition, estimating power retrospectively from beef research provides information for comparing relative power for different types of experiments (pen feeding, individual feeding, and metabolism) and offers the opportunity for appraisal of different types of beef cattle research. Power of the test in beef experiments is affected by experimental design, sample size, response variables being measured, and size of true differences between treatments. Other factors that likely affect power include pen type, seasonal and regional effects, and diet composition.

Although several publications are available on the subject of power tests for use by researchers (Steidl et al., 1997; Gerard et al., 1998; Novak and Haslberger,

<sup>1</sup>This article was presented at the 2003 ASAS-ADSA-AMPA meeting as part of the Contemporary Issues symposium "Designing Animal Experiments for Power." Approved for publication by the College of Agric. Sci. and Nat. Resources, Texas Tech Univ., Pub. No. T-5-446.

<sup>2</sup>Appreciation is expressed to M. S. Brown, West Texas A&M Univ., Canyon, for assistance in compiling the pen-feeding database.

<sup>3</sup>Correspondence: Box 42141 (phone: 806-742-2516; fax: 806-742-4003; E-mail: reed.richardson@ttu.edu).

Received July 10, 2003.

Accepted October 8, 2003.

2000; Kuiper et al., 2002), none specifically addresses beef cattle experiments and the associated factors that make them unique. Shortcomings of power calculations as a means to evaluate research also have been reported (Thomas, 1997; Hoenig and Heisey, 2001). Our objectives were to analyze data from publications of beef cattle research to 1) estimate retrospective power for selected response variables and 2) compare retrospective power between types of experiments (pen, individual, metabolism) and types of experimental design (completely random designs, randomized block designs, and Latin square designs).

## Methods

Data were compiled on subject areas of ionophores (pen feeding), general nutrition (individual feeding), and nitrogen utilization (metabolism), and were used to estimate the power of each experiment involved in these three types of beef cattle research methods. Beef cattle articles published in the *Journal of Animal Science* from 1974 through May 2003 were screened for inclusion in the database for calculation of retrospective power. Only articles that reported individual treatment means were included. These totaled 78 articles. In addition, two manuscripts in review (our unpublished data) were included, yielding a total of 80 experiments.

Both prospective and retrospective power analyses are explored in this paper. Prospective power analysis is reviewed through a Monte Carlo exercise. This is explained more thoroughly in the Results and Discussion section. The procedures used to retrospectively analyze power in experiments reviewed are explained below.

### Retrospective Power Analysis

To estimate power retrospectively, the following information was used: 1) an observed  $F$ -statistic (and its degrees of freedom) associated with a test of the null hypothesis of no treatment effect in the original data set, 2) a stated  $\alpha$  level, and 3) the noncentrality parameter associated with the  $F$ -statistic. The observed  $F$ -statistic can be calculated using treatment means, samples sizes, and the standard error of the mean. This calculation assumes that variances of treatment means are homogeneous, and in the case of randomized block designs, both block and treatment effects are fixed.

There are several estimators of the noncentrality parameter. We used the estimator described by Johnson et al. (1995):

$$\hat{\lambda} = (v_1(v_2 - 2)F / v_2) - v_1 \quad [1]$$

where  $v_1$  and  $v_2$  are numerator and denominator degrees of freedom of the calculated  $F$ -statistic. Power is given by (Graybill, 1976):

$$\text{Power} = \int_{F_{\alpha;v_1,v_2}}^{\infty} F(w;v_1,v_2;\lambda)dw \quad [2]$$

where  $F_{\alpha;v_1,v_2}$  = the upper  $\alpha$  probability point of the central  $F$ -distribution with  $v_1$  and  $v_2$  degrees of freedom;  $F(w;v_1, v_2; \lambda)$  is the probability density function of the noncentral  $F$ -distribution; and  $w$  is the  $F$ -statistic; when  $\hat{\lambda}$  is used in Eq. 2, then power is estimated. It should be noted that  $\hat{\lambda}$  (Eq. 1) is an unbiased estimator of the noncentrality parameter. However, negative estimates are possible, in which case power cannot be estimated; in these cases,  $\hat{\lambda}$  was set to zero (this biases the estimator). Other common estimators of power, namely  $\hat{\lambda}_1 = v_1 F$  (used by the software "GPOWER") and  $\hat{\lambda}_2 = \frac{(t-1)(TRTMS - MSE)}{MSE}$ , where  $t$  = number of treatments, TRTMS = the treatment mean square, and MSE = the experimental error mean square (e.g., Winer et al. 1991; Kirk, 1995), are also biased estimators of the true noncentrality parameter.

For each of the studies in the database, power was estimated using Eq. 2 with the noncentrality parameter estimated by Eq. 1. This estimate of power was then used for further analyses. In particular, we wished to provide descriptive information about estimated power in beef cattle research as it is related to type of experimental design, kind of experiment (e.g., pen-fed animals, individually fed animals, and metabolism trials), and response variable (e.g., ADG, feed efficiency [FE], and nitrogen retention). We are not aware of any theoretical studies of the distribution of power as a random variable. Thus, in our statistical tests that used power as the dependent variable, we employed nonparametric tests. For example, many texts suggest that power of 0.80 or greater is desirable. In our analysis, we tested the hypothesis that power equals 0.80 with a one-sample, Wilcoxon signed-ranks test. Similarly, when we compared power of ADG to power of FE, we used a Wilcoxon paired-samples test, and when we compared power in randomized block designs (RBD) to power in completely random designs (CRD) for a particular response variable (e.g., ADG), we used a Kruskal-Wallis test.

## Results and Discussion

*Prospective Power Analysis—A Monte Carlo Exercise.* Prospective evaluation of power in the planning and design phase of beef cattle experiments provides researchers with critical information on the appropriate design type and sample sizes necessary to detect a treatment effect at a predetermined significance ( $\alpha$ ) level. For example, suppose a researcher is studying ADG by 300-kg crossbred steers. Even when steers of a similar breeding and weight class are fed the same control diet, ADG will not be the same for each animal because of variability resulting from experimental error (i.e., variation among experimental units treated the same).

Now, suppose that it is known that the variance in ADG among animals treated alike is  $\sigma^2 = 0.01$ . If the researcher is interested in studying ADG by animals fed three different diets, he or she randomly selects animals from the target population (300 kg crossbred steers) and randomly assigns them to diets, with  $r$  animals per diet. The experimental design is therefore a CRD, with  $t = 3$  treatments and  $r$  replications per treatment. Under the assumptions that experimental errors are normally and independently distributed with a common variance in each treatment, an ANOVA and its accompanying  $F$ -test is used to test the null hypothesis that there is no difference in mean ADG among the three treatments. The alternative hypothesis is that the three treatment means are not equal.

Two distinct situations are possible. First, suppose that there is no true difference in ADG among the three diets (i.e., the null hypothesis is true). Because of variability in ADG, even among animals treated alike, it is possible that an experiment might yield data that would lead the investigator to reject the null hypothesis that there is no difference in ADG among the three diets—this would be a Type I error, and its probability of commission is controlled by selecting a desired  $\alpha$  level, typically,  $\alpha = 0.05$ .

A second kind of error is possible. Suppose that there actually is a difference in ADG among treatments. Again, because of experimental error, it is possible that an experiment might yield data that would lead the investigator not to reject the null hypothesis—this would be a Type II error, and its probability is denoted  $\beta$ .

Clearly, it is undesirable to reject the null hypothesis when it is in fact true. Likewise, it is undesirable to conclude that there is no difference among treatment means when in fact a true difference exists. This leads to the concept of statistical power: the power of a test is the probability of rejecting the null hypothesis when it is in fact false. Power is denoted by  $1 - \beta$ . It is clear that the power of a test depends, in part, on the exact nature of the inequality among treatment means.

Suppose that in addition to knowing that  $\sigma^2 = 0.01$ , the investigator knows that the true treatment means of ADG are  $\mu_1 = 0$ ,  $\mu_2 = 0.1$ , and  $\mu_3 = 0.2$ , with an overall population mean of  $\mu = 0.1$ . In this example, it is also true that the differences between each treatment mean and the grand mean are  $\tau_1 = \mu_1 - \mu = -0.1$ ;  $\tau_2 = \mu_2 - \mu = 0.0$ ; and  $\tau_3 = \mu_3 - \mu = 0.1$ . Suppose that logistic considerations and available resources are such that it is possible to have  $r = 6$  replications per treatment. Even though it might be known that treatment means differ, it is nevertheless possible that, because of experimental error, an experiment might yield data that would lead to the conclusion that there is no difference among treatment means (i.e., a Type II error). With this possibility in mind, the investigator might ask, "Given that I know that the true population means are different, and further given that ADG varies even among animals on the same diet, what is probability

of rejecting the null hypothesis with six replications per treatment?"

When the population parameters are known, power can be calculated. In this example, with  $\sigma^2 = 0.01$ , and  $\sum_{i=1}^t \tau_i^2 = 0.02$ , the noncentrality parameter is (Graybill, 1976):

$$\lambda = \frac{E[SS(H_0)]}{2\sigma^2} - \frac{df(H_0)}{2} \quad [3]$$

where  $E[SS(H_0)]$  is the expected value of the sum of squares associated with the null hypothesis,  $df(H_0)$  are the degrees of freedom associated with the null hypothesis, and  $\sigma^2$  is the experimental error. For a CRD,

$$E[SS(H_0)] = (t - 1)\sigma^2 + r \sum_{i=1}^t \tau_i^2$$

when  $\sum_{i=1}^t \tau_i = 0$ . In this example,  $\lambda = 6$ . A final parameter,

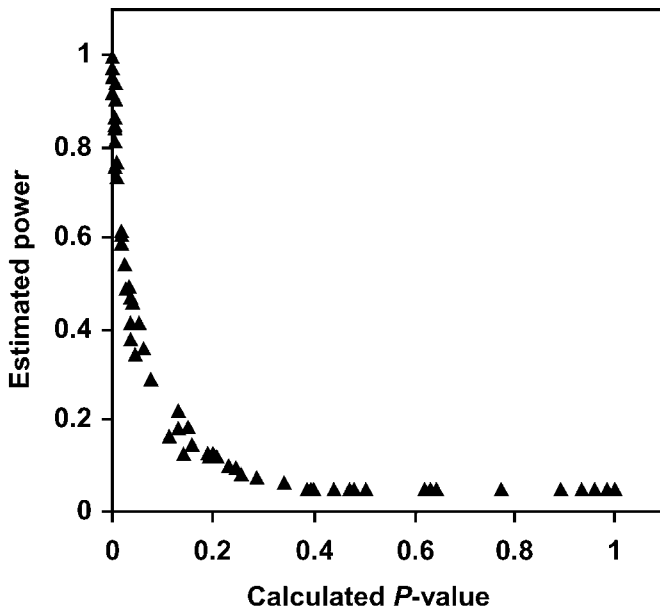
usually denoted  $\phi$ , is calculated by  $\phi = \sqrt{2\lambda/n_1}$ , where  $n_1 = df(H_0) + 1$ . In this example,  $\phi = 2.0$ . With  $\phi$ , power can be determined by consulting Table T-11 in Graybill (1976). In this example, power is 0.8049; thus, the probability of rejecting the null hypothesis is 80.49%.

The foregoing example can be empirically supported using Monte Carlo simulation methods. The IML procedure of SAS (SAS Inst., Inc., Cary, NC) was used to write a program that would generate experimental errors from a normal distribution with a variance of 0.01. Six experimental errors were assigned randomly to each of three treatments; then, a value of 0.1 was added to each experimental error in Treatment 2, and a value of 0.2 was added to each experimental error in Treatment 3. With randomly generated data, an ANOVA was completed with an  $F$ -statistic. This process was repeated 100,000 times in the Monte Carlo simulation.

Out of these 100,000 simulated experiments, the null hypothesis was rejected 80,497 times. Thus, the empirical power of the test statistic was 80.497%. The theoretical power, following Graybill (1976; see above) was 80.49%. We conclude that when population parameters are known, and assumptions underlying the  $F$ -test are satisfied, power can be determined.

A total of 100,000 experiments were simulated in the Monte Carlo exercise, and the results showed clearly that for these 100,000 experiments, there was an 80.497% chance that the null hypothesis was rejected. Clearly, this concept deals with the population of experiments.

For each of these experiments, power also can be retrospectively estimated. It is instructive to consider the possibilities that emerge from post hoc power estimation. For example, in one of the simulated experiments, the calculated  $F$ -statistic was  $F = 8.0699$ . This result can be used to estimate the noncentrality parameter using Eq. 1, from which power can be estimated



**Figure 1.** Relationship of  $P$ -value ( $\alpha$ ) to power in beef cattle pen-feeding experiments (includes both ADG and feed efficiency values,  $n = 34$ ).

using Eq. 2. For this particular simulated experiment, power is retrospectively estimated to be 80.49023%, a value very close to the true power in this example. However, another of the experiments in the Monte Carlo simulation yielded a calculated  $F$ -statistic of  $F = 14.731091$ , with an estimated power of 98.057%. And still another of the simulated experiments yielded a calculated  $F$ -statistic of  $F = 1.2001352$ , with an estimated power of 5.499%.

These results clearly demonstrate that estimated power is a random variable. Thus, any given set of experimental data can be used to estimate power, and the resulting estimate may or may not be close to true power. This is a consequence of the inherent variability in the experimental errors associated with the observed dependent variable.

*A Disclaimer on Interpretation of Retrospective Power.* Retrospective power is related to observed significance level (Figure 1). It is important to appreciate that when the observed significance level is large (i.e., nonsignificant), then retrospectively estimated power will be low; conversely, when the observed significance level is small (i.e., significant), then retrospectively estimated power will be high. Hoenig and Heisey (2001) provided a theoretical explanation of this relationship, and Figure 1 provides empirical support. The importance of this relationship between observed significance level and retrospectively estimated power cannot be overstated. For example, some practitioners advocate estimating power following a nonsignificant test result, with the following in mind: if power is relatively high, then this tends to support the null hypothesis (i.e., because power is relatively high, then one would probably have found a difference if it existed, but because the

null was not rejected, then there probably was not a difference). In a similar manner, if observed power is relatively low and the null hypothesis is not rejected, then a common interpretation is that perhaps there may actually have been a difference, but the low power of the test made it unlikely to reject the null hypothesis. Hoenig and Heisey (2001) termed this “thought experiment” the “power approach paradox.” Given the relationship between observed significance level and retrospective power estimation (Figure 1), it is clear that “computing the observed power after observing the  $P$ -value should cause nothing to change our interpretation of the  $P$ -value” (Hoenig and Heisey, 2001).

*Sample Size Estimation.* One concept that is related to power deals with estimation of sample size to detect a specified difference among treatment means. Continuing the example above, suppose the investigator knows that the variability among experimental units treated alike is  $\sigma^2 = 0.01$ . Further, suppose that the investigator would like to conduct an experiment (using a CRD) studying differences in ADG among three treatments. The investigator wishes to know how many experimental units are needed so that a difference between the largest and the smallest means of 0.2 kg is statistically significant at the 5% significance level with a power of 80%. Thus, four pieces of information are needed: 1) experimental error; 2) the maximum difference between treatments that is desired to detect; 3) a significance level; and 4) power. With this information, the tables provided by Bowman and Kastenbaum (1975) can be used to determine the necessary sample size. In particular, following the notation of Bowman and Kastenbaum (1975), let  $\mu_{\max} = \xi_{\max} = 0.20$  and  $\mu_{\min} = \xi_{\min} = 0.00$ . Then let  $\tau = (\xi_{\max} - \xi_{\min})/\sigma = 2$  for the present example. With  $k = 3$  treatments,  $\alpha = 0.05$ , and  $\beta = 0.20$ , Table 1 in Bowman and Kastenbaum (1975, page 142) can be consulted. From this table, the number of experimental units per treatment is equal to  $r = 5.947 \approx 6$ . The Monte Carlo simulation described above empirically confirms the accuracy of the Bowman and Kastenbaum (1975) tables.

The tables in Bowman and Kastenbaum (1975) do not include all experimental conditions. For example, for single-factor CRD, these tables include situations for  $t = 2$  through 11, and then for  $t = 13, 15, 20, 25, 30, 40, 50,$  and  $60$ ; other numbers of treatments are not included. Desu and Raghavarao (1990) provided the following approximation for sample size. Let

$$m^* = 2\{\sqrt{\chi_{1-\alpha; (t-1)}^2 - (t-2)} + z_{\beta}\}^2(\sigma/\Delta)^2$$

where  $\chi_{1-\alpha; (t-1)}^2$  is the  $100(1 - \alpha)$  percentile point of the  $\chi^2$  distribution with  $(t - 1)$  df;  $z_{\beta}$  is the  $\beta$  percentile point of the standard normal distribution;  $\sigma$  is the square root of the experimental error variance, and  $\Delta$  is the difference between the largest and smallest of  $t$  treatment means. The sample size needed per treatment is  $r = [m^*] + 1$ , where  $[\cdot]$  is the greatest integer function.

**Table 1.** Description of beef cattle experiments included in the pen-feeding database<sup>a</sup>

Experiment No. <sup>b</sup>	Design <sup>c</sup>	No. of animals	No. of treatments	No. of blocks or replications <sup>d</sup>	Power	
					ADG	FE <sup>e</sup>
1	RBD	800	4	20	0.752	1.000
2	RBD	192	4	6	0.078	0.919
3	RBD	280	4	7	0.538	1.000
4	RBD	800	4	20	0.411	1.000
5	RBD	400	4	10	0.360	1.000
6	RBD	320	4	10	0.063	0.360
7	RBD	320	4	10	0.222	0.606
8	CRD	105	5	3	0.185	0.590
9	CRD	126	6	3	0.050	0.098
10	RBD	80	4	4	0.050	0.147
11	RBD	120	4	5	0.998	0.124
12	RBD	114	6	3	0.850	0.087
13	RBD	120	6	2	0.050	0.125
14	RBD	216	6	6	0.492	0.471
15	RBD	120	4	5	0.102	0.119
16	RBD	96	4	4	0.121	0.901
17	RBD	96	4	4	0.050	0.050
18	RBD	128	4	4	0.050	0.179
19	RBD	96	4	3	0.347	0.05
20	RBD	120	5	3	0.768	0.815
21	RBD	140	5	4	0.290	0.488
22	RBD	200	5	4	0.050	0.050
23	RBD	72	4	3	0.938	0.378
24	RBD	72	4	3	0.168	0.592
25	CRD	96	4	4	0.050	0.131
26	CRD	64	4	4	0.050	0.734
27	RBD	96	4	4	0.413	0.975
28	RBD	144	4	6	0.050	0.841
29	RBD	100	4	5	0.050	0.050
30	RBD	96	4	4	0.050	0.050
31	RBD	192	8	6	0.456	0.814
32	RBD	80	4	4	0.178	0.734
33	RBD	96	4	6	0.952	0.863
34	RBD	200	5	5	0.183	0.617

<sup>a</sup>Compiled from selected articles on ionophores in the *Journal of Animal Science*.

<sup>b</sup>Indicates the order of compilation in the database. A complete list is available from the authors.

<sup>c</sup>RBD = randomized block design; CRD = completely random design.

<sup>d</sup>Represents true replications of experimental units.

<sup>e</sup>FE = feed efficiency.

In this example,  $m^* = 4.73$ , so the approximate sample size is five replications per treatment.

Another method for determining the number of replicates required in a proposed animal experiment is detailed by Berndtson (1991). Similar to the previous methods, four pieces of information must be known in order to determine sample size. Three of these (difference between treatments, significance level, and power) are the same in each of these methods. However, in the Berndtson (1991) approach, a coefficient of variation involving the control treatment is used in tables he provided to estimate sample sizes. In addition to requiring a coefficient of variation that involves only the control treatment, his tables provide estimated sample sizes only for two-treatment experiments and only for situations involving a 5% significance level.

Because of inherent differences between research facilities, and thus in the experimental error associated with beef cattle experiments, it is to be expected that the estimated number of replications needed to achieve

a desired power will vary between facilities. Continuing with the above example, suppose that a beef nutritionist wants to study ADG as affected by three treatments in a CRD; he or she wishes to know how many replications are needed to detect a difference among treatments if the largest difference between treatment means is 0.2 kg/d, with  $\alpha = 0.05$  and power = 0.80. She or he estimates experimental error at their research facility to be  $\sigma^2 = 0.0036$ . Using the tables of Bowman and Kastenbaum (1975),  $\tau = 3.33$  and thus  $r = 3$  replications would be used in this experiment. Now, suppose that this nutritionist is collaborating with a colleague at a different facility, where the estimate of experimental error is  $\sigma^2 = 0.02$ . Using the tables of Bowman and Kastenbaum (1973),  $\tau = 1.414$  and  $r = 11$  replications would be used in the experiment at the second facility. Thus, even when the same experiment is repeated at two different facilities, estimated sample sizes will be affected by the inherent variability in the response variable associated with each facility. Finally, it should also

**Table 2.** Estimated power in beef cattle pen-feeding experiments on animal performance variables<sup>a</sup>

Experimental design <sup>b</sup>	Response variable <sup>c</sup>	No. of experiments	Estimated power			Test of Ho: power = 0.80
			Mean	Minimum	Maximum	$P >  S $ <sup>d</sup>
RBD	ADG	30	0.3360	0.0500	0.9982	0.001
	FE	30	0.5135	0.0500	1.0000	0.001
CRD	ADG	4	0.0837	0.0500	0.1849	0.125
	FE	4	0.3881	0.0977	0.7337	0.125

<sup>a</sup>Data were pooled within design type for the response variables studied. RBD = randomized block design; CRD = completely random design; FE = feed efficiency.

<sup>b</sup>For CRD – Ho: ADG = FE →  $P < 0.001$  (n = 4; Wilcoxon paired test); for RBD – Ho: ADG = FE →  $P < 0.004$  (n = 30; Wilcoxon paired test).

<sup>c</sup>For ADG – Ho: CRD = RBD →  $P < 0.073$  (n = 4, 30; Kruskal-Wallis test); for FE – Ho: CRD = RBD →  $P < 0.520$  (n = 4, 30; Kruskal-Wallis test).

<sup>d</sup> $P$ -value for a Wilcoxon signed ranks test.

be appreciated that if true experimental error at the first facility was in fact  $\sigma^2 = 0.01$ , but a value of 0.0036 was used to estimate sample sizes, then an experiment with  $r = 3$  replications would in reality have a power of 0.3857 rather than the nominal value of 0.80. Likewise, if the (overestimate) of 0.02 was used when the actual experimental error was 0.01, then an experiment with  $r = 11$  replications would in reality have a power exceeding 0.99. For these reasons, it is imperative that researchers collect and retain historical data associated with their facilities, and that every effort is made to

provide an accurate estimate of experimental error to be used in designing future experiments.

*Pen-Feeding Experiments.* A description of experiments included in the pen feeding database and retrospective power values for ADG and FE are given in Table 1. Thirty pen-feeding evaluations of animal performance in our database used RBD. Only four studies included used a CRD.

A comparison of the influence of design type on response variables and power data is presented in Table 2. In RBD, the estimated power associated with ADG

**Table 3.** Description of beef cattle experiments included in the individual feeding database<sup>a</sup>

Experiment No. <sup>b</sup>	Design <sup>c</sup>	No. of animals	No. of treatments	No. of blocks or replications <sup>d</sup>	Power	
					ADG	FE <sup>e</sup>
1	RBD	60	6	5	0.050	0.050
2	RBD	60	6	5	0.412	0.175
3	CRD	60	3	20	0.110	0.050
4	CRD	20	4	5	0.050	0.050
5	CRD	20	4	5	0.351	0.050
6	CRD	33	3	11	0.198	0.050
7	CRD	27	3	9	0.050	0.050
8	CRD	44	4	11	0.130	0.126
9	CRD	45	5	9	0.935	0.080
10	RBD	60	5	4	0.050	0.050
11	CRD	56	7	8	0.784	0.989
12	CRD	36	3	12	0.946	0.050
13	CRD	36	3	12	0.998	0.050
14	CRD	36	3	12	0.568	0.909
15	CRD	72	6	12	0.999	0.999
16	RBD	16	4	4	0.050	0.143
17	CRD	54	6	9	0.930	0.395
18	CRD	18	3	6	0.050	0.050
19	CRD	120	10	12	0.050	0.050
20	CRD	16	4	4	0.050	0.198
21	CRD	60	5	12	0.055	0.318
22	CRD	24	2	12	0.050	0.050

<sup>a</sup>Compiled from published (n = 20) and unpublished (n = 2) sources.

<sup>b</sup>Indicates the order of compilation in the data base. A complete listing is available from the authors.

<sup>c</sup>RBD = randomized block design; CRD = completely random design.

<sup>d</sup>Represents true replications of experimental units.

<sup>e</sup>FE = feed efficiency.

**Table 4.** Estimated power in beef cattle individual-feeding experiments on animal performance variables<sup>a</sup>

Experimental design <sup>b</sup>	Response variable <sup>c</sup>	No. of experiments	Estimated power			Test of Ho: power = 0.80
			Mean	Minimum	Maximum	$P >  S $ <sup>d</sup>
CRD	ADG	18	0.4059	0.0500	0.9999	0.003
	FE	18	0.2508	0.0500	0.9999	0.001
RBD	ADG	4	0.1404	0.0500	0.4115	0.125
	FE	4	0.1045	0.0500	0.1749	0.125

<sup>a</sup>Data were pooled within design type for the response variables studied. RBD = randomized block design; CRD = completely random design; FE = feed efficiency.

<sup>b</sup>For CRD – Ho: ADG = FE →  $P < 0.260$  (n = 18; Wilcoxon paired test); for RBD – Ho: ADG = FE →  $P < 0.999$  (n = 4; Wilcoxon paired test).

<sup>c</sup>For ADG – Ho: CRD = RBD →  $P < 0.158$  (n = 18, 4; Kruskal-Wallis test); for FE – Ho: CRD = RBD →  $P < 0.853$  (n = 18, 4; Kruskal-Wallis test).

<sup>d</sup>P-value for a Wilcoxon signed ranks test.

was less ( $P < 0.001$ ) than 0.80. Similarly, estimated power associated with FE was less ( $P < 0.001$ ) than 0.80. In CRD, estimated power did not differ ( $P > 0.125$ ) from 0.80 for either ADG or FE; however, sample sizes associated with these tests limit conclusions.

In RBD, estimated power associated with tests of ADG was less ( $P < 0.004$ ) than the power associated with tests of FE. Similarly, in CRD, estimated power was less ( $P < 0.001$ ) for ADG than for FE.

In studies of FE, estimated power did not differ ( $P > 0.520$ ) for CRD and RBD. Nonetheless, there was an

indication ( $P < 0.073$ ) that RBD were more powerful than CRD when ADG was measured.

*Individual Feeding Experiments.* A description of experiments included in the individual feeding database and retrospective power values for ADG and FE are given in Table 3. Eighteen studies of animal performance that were based on individual animals used a CRD, whereas only four studies used a RBD. A comparison of the influence of design type on response variables and power data is presented in Table 4. In RBD, estimated power did not differ ( $P > 0.125$ ) from 0.80 for

**Table 5.** Description of beef cattle experiments included in the metabolism database<sup>a</sup>

Experiment No. <sup>b</sup>	Design <sup>c</sup>	No. of animals	No. of treatments	Power
				Nitrogen retention
1	CRD	8	2	0.050
2	LS	6	3	0.050
3	LS	8	4	0.945
4	CRD	20	4	0.148
5	LS	8	8	0.801
6	LS	4	4	0.858
7	LS	12	4	0.227
8	LS	4	4	0.143
9	CRD	12	3	0.924
10	LS	5	5	0.741
11	LS	6	3	0.050
12	CRD	16	4	0.816
13	LS	6	6	0.449
14	LS	4	4	0.050
15	LS	5	5	0.463
16	LS	6	6	0.244
17	LS	5	5	0.201
18	LS	4	4	0.972
19	LS	5	5	0.999
20	LS	5	5	0.994
21	LS	6	3	0.050
22	LS	6	3	0.050
23	LS	5	5	0.050
24	LS	4	4	0.050

<sup>a</sup>Compiled from selected articles on nitrogen utilization in the *Journal of Animal Science*.

<sup>b</sup>Indicates the order of compilation in the database. A complete listing is available from the authors.

<sup>c</sup>LS = Latin square; CRD = completely random design.

**Table 6.** Estimated power in beef cattle metabolism experiments on nitrogen retention<sup>a</sup>

Experimental design <sup>bc</sup>	No. of experiments	Estimated power			Test of Ho: power = 0.80
		Mean	Minimum	Maximum	$P >  S $ <sup>d</sup>
Latin square	20	0.4193	0.0500	0.9994	0.002
CRD	4	0.4844	0.0500	0.9238	0.625

<sup>a</sup>Data were pooled within design type for the response variables studied.

<sup>b</sup>LS = Latin square; CRD = completely random design.

<sup>c</sup>For N\_RET - Ho: LS = CRD  $\rightarrow P < 0.844$  (n = 20, 4; Kruskal-Wallis test).

<sup>d</sup>P-value for a Wilcoxon signed ranks test.

ADG or FE. However, in CRD, estimated power was less than 0.80 when ADG was measured ( $P < 0.003$ ), as well as when FE was measured ( $P < 0.001$ ).

In RBD, estimated power did not differ ( $P > 0.999$ ) for tests of ADG and tests of FE. Likewise, in CRD, there was no difference ( $P > 0.260$ ) in the power associated with ADG and FE.

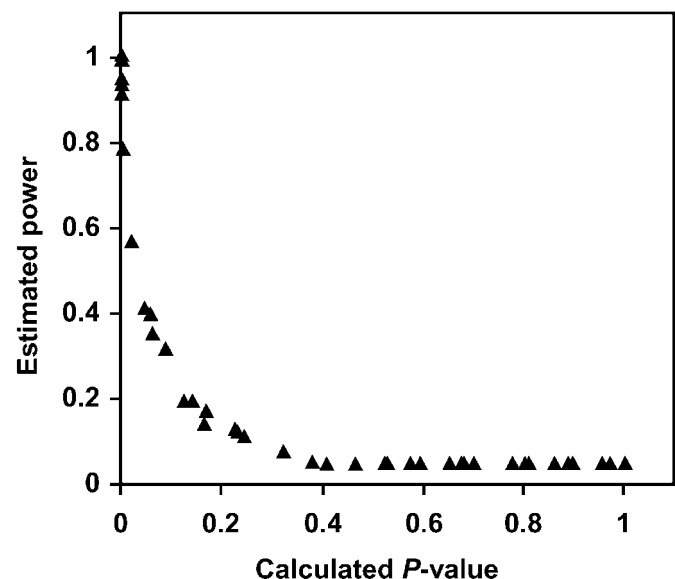
In evaluation of FE, power did not differ ( $P > 0.853$ ) in CRD and RBD. Similarly, in studies of ADG, power did not differ ( $P > 0.158$ ) in CRD and RBD.

The relationship of estimated  $P$ -value to retrospective power in individual feeding experiments for ADG and FE is graphically presented in Figure 2.

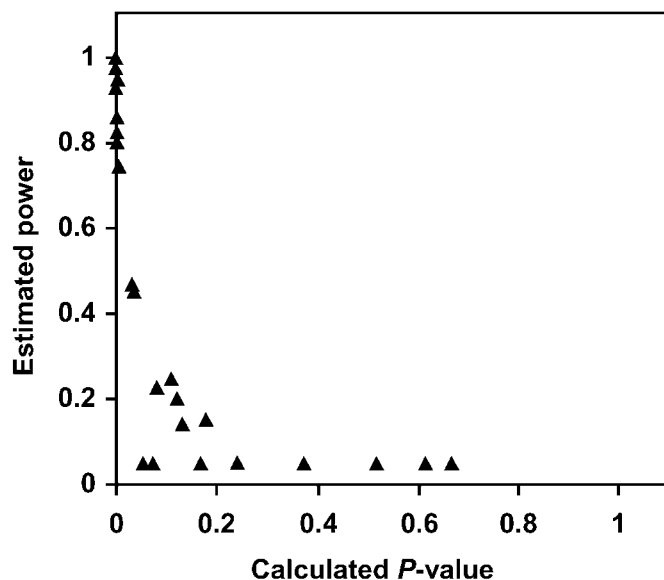
At this point, it is important to note a few items of importance with respect to pen-based and individual animal-based studies. Each of these types of studies is important to beef cattle research, and one is not necessarily better than the other. The primary factor influencing which type of feeding regimen to use is facility design; one can only conduct the type of research allowed by the facility. With this stated, there are factors that need to be taken into account when choosing between pen and individual animal feeding, given that the option exists. Pen-based studies are likely the most appropriate choice when overall animal performance is of interest, with real-world application being the ultimate goal. Because a group of animals is fed, values for DMI should more accurately reflect practical conditions and consequently, so should values for FE. However, even in pen settings, there may be differences resulting from the size of pens and the number of animals in the pen that affect applicability of the results. In contrast, individual animal feeding trials are typically conducted when more basic questions are of interest, and animal-specific measurements, which would not be readily obtained in a group setting, need to be measured. Although the feed intake by individually fed animals may be measured very precisely, it may not accurately reflect the intake by groups of animals. This is because individually fed animals often lack the driving forces of social interaction and competition that likely affect feed intake in groups.

It is also important to address the manner in which pen data are analyzed, especially the analysis of ADG and carcass data. In research settings, it is often tempting to analyze ADG and carcass measurements on an

individual-animal basis because individual BW of each animal is known. When conducting pen-based studies, however, all variables should be analyzed with pen as the experimental unit. This is because treatments are applied to the pen of animals, not to the individual animal. Consider a completely randomized design with pen as the experimental unit; suppose that there are five animals per pen. In this setting, there are two sources of unexplained variation: 1) variation among animals within a pen (called "sampling error") and 2) variation among pens within a treatment ("experimental error"). Perhaps the most important assumption for the ANOVA is independence of experimental errors. When pens are randomly assigned to treatments, it is reasonable to assume independence of experimental errors; the correct error term for testing treatment effects is the experimental error mean square. It is also reasonable to suggest that social interactions among animals within a pen might result in correlated sampling errors. However, the nonindependence of sampling errors within a pen does not bias the  $F$ -test on the treatment effects when the experimental error mean



**Figure 2.** Relationship of  $P$ -value ( $\alpha$ ) to power in beef cattle individual feeding experiments (includes both ADG and feed efficiency values, n = 22).



**Figure 3.** Relationship of  $P$ -value ( $\alpha$ ) to power in beef cattle metabolism experiments for nitrogen retention ( $n = 24$ ).

square is used to test treatment effects (although it can affect the power of this test). When individual animals are used as the experimental unit instead of pen, this leads to an analysis wherein the errors comprising the error (residual) mean square (i.e., the denominator of the  $F$ -test) are not independent. This violation of independence has serious consequences on the performance of the  $F$ -statistic in detecting differences among treatment means. Therefore, it is imperative that the  $F$ -test on the treatment effects use the appropriate error mean square.

*Metabolism Experiments.* Our database included 20 studies of nitrogen retention that used a Latin square and four studies that used a CRD (Table 5). A comparison of design type on nitrogen retention and power data for metabolism experiments are presented in Table 6. For Latin squares, estimated power was less ( $P < 0.002$ ) than 0.80. However, for CRD, estimated power did not differ ( $P > 0.625$ ) from 0.80. There was no difference ( $P > 0.844$ ) in estimated power between Latin squares and CRD.

The relationship of  $P$ -value to retrospective power in metabolism experiments for nitrogen retention is graphically presented in Figure 3.

### Implications

Prospective and retrospective power analyses are fundamentally different, and researchers should be aware of this to avoid misinterpretation of comparative power values. Differences in retrospective power in beef cattle experiments are affected by experimental design, the number of experiments in each design type, and the response variable under investigation. These data

show a higher estimate of retrospective power for feed efficiency than for average daily gain in pen feeding experiments. Low estimated retrospective power for a given experiment, or for a group of experiments, may indicate inappropriate experimental design or inaccurate experimental technique, leading to high estimates of experimental error. Nonetheless, a properly designed and conducted experiment can produce a low retrospective estimate of power. Our results show that even in a population of experiments for which true power was relatively high, individual experiments could produce low estimates of power because estimated power is a random variable. Finally, estimated power may be low because true treatment effects are in fact small.

### Literature Cited

- 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.
- Bowman, K. O., and M. A. Kastenbaum. 1975. Sample size requirement: Single and double classification experiments. Pages 111–232 in *Selected Tables in Mathematical Statistics*. Vol. 3. Institute of Mathematical Statistics, ed. Am. Math. Soc., Providence, RI.
- Desu, M. M., and D. Raghavarao. 1990. *Sample Size Methodology*. Academic Press, Boston, MA.
- Gerard, P. D., D. R. Smith, and G. Weerakkody. 1998. Limits of retrospective power analysis. *J. Wildl. Manag.* 62:801–807.
- Gill, J. L. 1980. Evaluation of statistical designs and analysis of experiments. *J. Dairy Sci.* 64:1494–1519.
- Graybill, F. A. 1976. Pages 520–521 in *Theory and Application of the Linear Model*. Duxbury Press, North Scituate, MA.
- Henderson, C. R. 1969. Design and analysis of animal science experiments. In *Techniques and Procedures in Animal Swine Research*. 2nd ed. Am. Soc. Anim. Sci., Savoy, IL.
- Hoening, 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.
- Johnson, N. L., S. Kotz, and N. Balakrishnan. 1995. Page 495 in *Continuous Univariate Distributions*. Vol. 2. John Wiley and Sons, NY.
- Kirk, R. E. 1995. *Experimental Design, Procedures for the Behavioral Sciences*. 3rd ed. Brooks/Cole Publ., Pacific Grove, CA.
- Kuiper, H. A., P. J. M. Noteborn, E. J. Kok, and G. A. Kleter. 2002. Safety aspects of novel foods. *Food Res. Int.* 35:267–271.
- Kuiper, H. A., G. A. Kleter, P. J. M. Noteborn, and E. J. Kok. 2002. Substantial equivalence—An appropriate paradigm for the safety assessment of genetically modified foods. *Toxicology* 181/182:427–431.
- Lofgreen, G. P., J. L. Hull, and K. K. Otagaki. 1962. Estimation of empty body weight of beef cattle. *J. Anim. Sci.* 21:20–24.
- Meyer, J. H., G. P. Lofgreen, and W. N. Garrett. 1960. A proposed method for removing sources of errors in beef cattle experiments. *J. Anim. Sci.* 19:1123–1131.
- Norton, H. W. 1969. Assessment of the efficacy and safety of drugs used in animal feeds. In *The Use of Drugs in Animal Feeds*. Proc. of Symp. Nat. Acad. Sci. Pub. No. 1679, Washington, DC.
- Novak, W. K., and A. G. Haslberger. 2002. Substantial equivalence of antinutrients and inherent plant toxins in genetically modified novel foods. *Food and Chem. Toxicology*. 38:473–483.
- Steidl, R. J. 1997. Statistical power analysis in wildlife research. *J. Wildl. Manag.* 61:270–279.
- Thomas, L. 1997. Retrospective power analysis. *Conserv. Biol.* 11:276–280.
- Winer, B. J., D. R. Brown, and K. M. Michels. 1991. *Statistical Principles in Experimental Design*. 3rd ed. McGraw-Hill, NY.