The Professional Animal Scientist 32 (2016):19–30; http://dx.doi.org/10.15232/pas.2015-01424

SYMPOSIUM

Invited Review: Designing a grazing experiment that can reliably detect meaningful differences[1]

R. R. Reuter*[2,3] PAS, and C. A. Moffet†[2,4]
*Department of Animal Science, Oklahoma State University, Stillwater 74078; and
†Samuel Roberts Noble Foundation Inc., Ardmore, OK 73401

ABSTRACT

Grazing trials are critical to advance the understanding and management of complex systems involving land, plants, and animals. Experiments provide data that are directly useful to managers and that are required for parameterizing and ground-truthing models. However, grazing trials are expensive to conduct. Because of their resource-consuming nature, many facilities and studies are often lacking in adequate replication. In light of this, grazing researchers should focus on designing experiments that have adequate statistical power to detect the differences among treatments that they suspect actually exist or that are meaningful to managers or other researchers. Power in grazing trials could potentially be improved by increased tolerance of type I or II errors; increasing the df in the denominator of F-tests; and reducing experimental, sampling, and measurement errors. These could be accomplished in multiple ways through experimental design choices, selection of variables to measure, and how carefully measurements are made. Suggestions for these approaches and decision aid tools to evaluate their effectiveness are provided. A final approach to increasing power of grazing experiments is to plan for and use meta-analysis to its full potential. Ultimately, grazing researchers must continually seek and apply new techniques to achieve adequate power for the specific resources they have and the specific questions they seek to answer.

Key words: analysis, error, method, power, statistics

INTRODUCTION

Grazing trials are a vital component to understanding and improving management of grazing systems. Many traits of grazing systems are difficult to model with our current understanding of animal behavior, physiological response to grazing, plant and soil responses under grazing, and so on. However, grazing trials are also expensive to conduct. They typically require larger land areas than small-plot, agronomic trials or pen-based, feedyard trials. Many times, the key variable of interest may be one best measured on a landscape scale, such as native range species composition, necessitating large areas for research. These large areas many times require inputs and management in relation to their size (fertilizers, fences, and so on). In addition, the animals needed to implement the desired treatments are often expensive to buy and maintain. Because of these factors, infrastructure to do experiments is often limited, further increasing the "opportunity cost" of doing a grazing trial; one experiment typically physically displaces another potential experiment (Riewe et al., 1989).

[1] This review article is based on a presentation by the authors in the symposium "Design, Analysis, and Execution of Quality Grazing Research" at the meeting of the American Society of Animal Science Southern Section, February 1, 2015, Atlanta, Georgia, in conjunction with the SERA-41 Beef Cattle Forage Utilization Group and the South Central Chapter of the American Registry of Professional Animal Scientists. Appreciation is expressed to the symposium sponsors: The Samuel Roberts Noble Foundation, Elanco Animal Health—Stocker District, Vigortone, and Barenbrug USA. Publication of this manuscript was sponsored by Elanco Animal Health—Stocker District.
[2] Authors share senior authorship.
[3] Corresponding author: ryan.reuter@okstate.edu
[4] Current address: USDA-ARS, Southern Plains Range Research Station, Woodward, OK, 73801.

Fewer students are being instructed in the methods available to conduct grazing experiments. Ironically, the more grazing research that is done, the harder it becomes to find differences. This is because the large effects are easiest to find and therefore found first. Grazing research in the future will continue to chase ever smaller effects (Giesbrecht, 1989). Because grazing trials are both expensive and important, the most should be made of the inputs, time, and effort put into them. One way to ensure the most is made of grazing trials is to design the proposed studies to have adequate power to detect differences among treatments that the experimenters would consider meaningful or likely. That is to say, the resources might best be used for another experiment if the proposed study is unlikely to have sufficient power to detect differences that the experimenters would want to be able to detect. It is important for the experimenter to think about the treatment differences they seek to estimate. If for example, the entire population of herds getting a treatment, let us call it treatment A, were accurately measured for the response of interest and the entire population of herds getting another treatments, let us call it treatment B, were accurately measured for the same response, it is very unlikely that the difference in the mean response would be 0. But the difference may be too small to be biologically or economically important. However, if the true difference between treatments is biologically or economically important, it is preferable to be able to detect that difference. Considering the potential value of livestock gain, it is not unusual for treatment differences to be large enough to be meaningful. There is a significant penalty for making a type II error when the difference is meaningful.

POWER: A REVIEW

Prospective statistical power is the probability of a test or study to reject a false null hypothesis, i.e., to avoid a type II error. A study that has sufficient power (often researchers will consider 80% sufficient) provides evidence to detect a difference that is real most the time. That is, it will detect the difference 80% of the time when power is 80%. Power is influenced by several factors, mainly, the true magnitude of the differences among treatments, the type I error rate used in the test (α), variability in the measured response, and sample size. (Often, the point of doing an experiment is to determine the magnitude of the differences among treatments, and we do not know this value a priori. In this case, the investigator should choose a magnitude that would be considered a meaningful difference, either from a biological or economic perspective.) These 4 factors can be manipulated by researchers to achieve a desired level of power. Additional discussion of power, including equations for calculating it, can be found in Steele et al. (1997).

To illustrate the types of errors and what power is, Monte Carlo experiments were conducted. The Monte Carlo experiments were conducted by first generating a synthetic population of steers where the true ADG for each steer given the 3 treatments was known. In the base case, the true steer ADG for the control treatment was sampled from a random normal distribution (mean = 0.65 kg/d, SD = 0.15 kg/d). The true steer ADG for the plus 10% and plus 30% treatments were equal to the control ADG plus 0.065 and 0.195 kg/d, respectively (i.e., the effects are additive and the plus 10% and plus 30% treatments always add 0.065 and 0.195 kg/d to the control ADG). For each simulated 90-d experiment in the base case, 162 steers were sampled and treatments were assigned to steers completely at random, so that there were 54 steers in each treatment. True initial BW for the sampled steers were between 165 and 291 kg. Measured initial BW was the true initial BW plus measurement error (assumed 0 in the base case). Measured final BW was the true initial weight plus the true gain (true ADG times 90 d) plus measurement error (again, assumed 0 in the base case). Finally, measured ADG was the measured gain (measured final weight minus measured initial weight) divided by 90 d. An ANOVA was performed to determine whether ADG differed among treatments. Furthermore, the probability of a difference for all pairwise contrasts was also determined. Each simulated experiment was repeated 1,000 times, and the distribution of *P*-values for the *F*-test and all pairwise contrasts were evaluated to characterize power.

P-values from 4,000 contrasts made in the Monte Carlo experiments are shown in Figure 1. This includes 3,000 contrasts using the base case with control versus 10% as "10%," 10% versus 30% as "20%," and control versus 30% as "30%" and 1,000 contrasts from another scenario where the true effect of the plus 10% treatment had been replaced with 0 (i.e., no true effect). As expected, in the case of null or no difference among treatments, *P*-values are uniformly distributed in the range of 0 to 1, and 5% of the experiments produced *P*-values of 0.05 or less (i.e., below the dashed line). These 5% of experiments would be interpreted as having a significant effect of treatment, when in fact there was no effect. This is a type I error. It can be seen from this figure that α controls the type I error rate in the case of the null hypothesis being true.

Power becomes an issue when the null hypothesis is false, i.e., there is a true difference among treatments. This can be seen in the case of the 10, 20, and 30% treatments. The distribution of *P*-values still ranges from 0 to 1 but is now skewed toward 0. In the 10% difference case, 26% of the *P*-values are less than or equal to α, indicating that this experimental design detected the difference 26% of the time (i.e., power). In the 30% treatment, there is a similar scenario, except that power is now 96% because there is a bigger difference that is more likely to be detected. Now, 96% of the *P*-values are below or equal to α. Stated another way, the type II error rate was controlled to 4%, and out of 25 experiments conducted, only 1