J. Range Manage.
55: 526-529 November 2002

Viewpoint: Improving range science through the appropriate use of statistics

WILLIAM R. GOULD AND ROBERT L. STEINER

Authors are Associate Professors in the University Statistics Center, Department of Economics, New Mexico State University, Las Cruces, N.M. 88003-8001.

Abstract

We examined a stratified random sample of articles published over 3 decades of the *Journal of Range Management* to study the applications and changes in statistical methodology employed by range scientists. Our objectives were to characterize the philosophical nature of statistics use in range science and to identify strengths and weaknesses inherent in these approaches. In each article, we examined the research design efficacy and whether the statistical analysis was adeptly used to convey the relevant information. The majority of articles we examined were conducted appropriately. In general, we found more emphasis has been placed on statistical testing than effect size estimation in the last decade. On an average, 82 tests or means comparisons (s.e. = 20) were presented in each article during the 1990's. Articles that reported an effect size via a sample mean frequently did not report an associated standard error. Research designs lacked adequate descriptions in several cases, making it difficult to determine if the appropriate analysis was performed. Improper identification of the experimental or sampling unit and/or the interdependence of observations occurred in all decades. We recommend increased inferential use of confidence intervals and suggest that the practical significance (as opposed to statistical significance) of results be considered more often. Improvements in the 'science' of range science can be made by greater understanding and communication of statistical concepts through consultation with statisticians.

Key Words: effect size, estimation, p-value, repeated measures, Type I error

Resumen

Examinamos una muestra aleatoria estratificada de artículos publicados durante tres décadas en el *Journal of Range Management* para estudiar las aplicaciones y los cambios en la metodología estadística empleada por los científicos de manejo de pastizales. Nuestros objetivos fueron caracterizar la naturaleza filosófica del uso de la estadística en la ciencia de los pastizales e identificar las fortalezas y debilidades inherentes a estas estrategias. En cada articulo, examinamos la eficacia del diseño de la investigación y si el análisis estadístico se uso hábilmente para conducir a información relevante. La mayoría de los artículos que examinamos se condujeron apropiadamente. En general, encontramos que en la última década se ha puesto mas énfasis en las pruebas estadísticas que en el efecto del tamaño de la estimación. En promedio 82 pruebas o comparaciones de medias (s.e. = 20) se presentaron en cada artículo durante la década de 1990. Los artículos que reportaron un efecto de tamaño vía media de la muestra frecuentemente no reportaron un error estándar asociado. En varios casos los diseños de la investigación carecieron de descripciones adecuadas dificultando el determinar si se condujo un análisis estadístico apropiado. La identificación inadecuada de la unidad experimental o de muestreo o la interdependencia de las observaciones ocurrió en todas las décadas. Recomendamos el aumento del uso inferencial de los intervalos de confianza y sugerimos que la significancia práctica (contrario a la significancia estadística) de los resultados debe ser considerada mas a menudo. Se pueden hacer mejoras en la "ciencia" de manejo de pastizales mediante un mayor entendimiento y comunicación de los conceptos estadísticos a través de la consulta con los estadísticos

This paper is motivated by our consulting experience with agricultural, biological, and environmental scientists in the southwest. Combined, we have over 15 years of consulting experience and have worked with dozens of faculty and hundreds of students involved in natural resources. We have found that many students and faculty have an aversion to statistics or a misunderstanding of the role statistics plays in the research process. For example, it is not uncommon for statistical help to be initially solicited after a data set has been collected, leading us to making recommendations that can limit the inferential power of their work (e.g., when randomization was not invoked). As a result, we have been viewed as unrealistic in our desire for scientific rigor and as barriers to publishing research outcomes. This specific study focuses on the use and abuse of statistics in the *Journal of Range Management* over the past 3 decades. Our purpose is not to implicate specific individuals, entire departments or the field of range science. Indeed, our observations apply more generally to many professional fields outside range science, but reviews of statistics use have been made by others in other disciplines (Harlow et al. 1997, Cherry 1998, Anderson et al. 2000).

Anderson et al. (2000) documented the overuse of hypothesis testing in *Ecology* and the *Journal of Wildlife Management*. They concluded that the vast majority of statistical hypothesis tests are conducted on null hypotheses that are clearly false. Nester (1996) suggested several reasons for the indiscriminate use of hypothesis tests. (1) They appear to be objective and exact; (2) they are readily executed with statistical software packages; (3) we are taught

Authors wish to thank Dr. R. Pieper for lending us his journals and his encouragement.

Manuscript accepted 27 Jan. 02.

to use them and everyone seems to use them; (4) some journal editors and supervisors demand them. Anderson et al. (2000) suggest that too much weight is given to statistical tests and that there is not enough emphasis on effect sizes (estimates of magnitudes of effects), directionality of differences, and biological importance. In other words, identifying statistical significance via hypothesis tests, (i.e., reporting of a p-value by itself) provides little information in considering real scientific hypotheses. Furthermore, when statistical hypothesis tests are conducted, the importance of evaluating the assumptions underlying those tests cannot be overstated. Application of statistical methodologies to nonrandom data from observational studies must be clearly described and considered with caution (Cherry 1998).

In evaluating the articles, we asked ourselves 1) Was the research sufficiently described so that it was repeatable? 2) Were randomization, replication and controls or comparisons properly used in experiments? 3) Given a clear description of the study design and treatment arrangement, was an adequate (as opposed to optimal) analysis performed? 4) Were results reported with sufficient detail (e.g., measures of precision, effect size, test statistics, degrees of freedom, etc.). Our evaluation of the range science literature differs from reviews made by others (Cherry 1998, Anderson et al. 2000) in that we examined the experimental and/or sampling designs implemented, methods of analysis, reporting of results and interpretation in both a quantitative and qualitative fashion. We did note when the aforementioned problems in reporting results occurred, but have also attempted to assess all of the statistical machinery underlying range science studies.

The articles we read covered a variety of topics, including habitat use by animals, effects of fire and herbicides on vegetation and soils, food preference studies, nutrient analysis, drought and grazing effects on grasses, evaluation of technology (e.g., pedometers), resource conflicts (e.g., perceived economic damage by ungulates), etc. While we agree with Guthrey et al. (2001) that states the research hypothesis should be given more weight than statistical hypotheses, we have not judged the value or scope of the research itself because we are not range scientists. We excluded technical notes, book reviews, viewpoints, management notes, presidential addresses, comment papers and rebuttals, and invited synthesis papers from the collection we evaluated. As our intent is not to embarrass specific authors or institutions, we refer to specific articles by year only. Exact citations of the examples presented are available upon request.

Methods

We selected a stratified multistage random sample of 54 journal articles from the *Journal of Range Management*. Decades of the 1970s, 1980s, and 1990s served as strata, from which 3 years were randomly selected. We stratified by decade to ensure samples were selected from each decade so that we might identify trends in statistical usage over these time frames. From each of the selected years, 2 issues were randomly selected, from which 3 articles were randomly selected for examination. Simple random samples were selected at each stage using a random number table. Each article was read and evaluated by 1 of us. Quantitatively, we tallied the number of statistical tests or means comparisons, the rate at which appropriate standard errors were reported with means, and the frequency with which P-values were reported without an accompanying test statistic and degrees of freedom.

Evidence of statistical testing was usually indicated in the results section of papers in 1 of 2 manners. Either a declarative statement was made and accompanied with a p-value, or tables of means were presented with superscripts indicating statistical differences. In some articles, the actual number of means comparisons was unclear because the multiple comparison procedure used was unspecified. In such cases, we only recorded the number of means to be compared.

The rate at which standard errors were not reported with means was a frequency measure whereby if an article had at least 1 such occurrence, it was flagged. Only articles that presented at least 1 mean were included in our frequency measure. In a few cases, a single standard error was reported for a collection of means under a complex design structure that would have different variance components. We flagged these instances as failures to report appropriate standard errors.

We use the term 'naked p-value' to indicate values that are reported by themselves without a corresponding test statistic or associated degrees of freedom (e.g., $P = 0.028$). Our definition differs from that used by Anderson et al. (2001b) in which they consider a p-value naked if it lacks an effect size, its direction, and a measure of its precision. Reporting the test statistic and degrees of freedom allows one to evaluate if the test was performed appropriately (e.g., no pseudoreplication occurred). In addition, we noted if only a range was given for the p-value ($P > 0.05$ or $P < 0.10$). We do not wish to perpetuate the misinterpretation of p-values as representing the strength of evidence for the alternative hypothesis or the probability that the null hypothesis is false. However, we believe that knowledge of its exact value more accurately describes the degree of consistency of the data with the null hypothesis (Ellison 1996).

Qualitatively, we determined if the design used was clearly stated and sufficiently detailed so that the study could be replicated. We questioned if the randomization was executed appropriately and replication recognized at the correct level. When statistical tests were used, we determined if they were described adequately and whether or not the practical significance of their result was considered in addition to statistical significance.

Results

The majority of articles we examined were commendable on many measures. Controls or comparisons were used in most experimental studies from all decades. Often, the locations of sampling units were randomized within plots. In several cases, a statistician had been either acknowledged or included as a co-author. However, we found there is room for improvement regarding the statistical components of range science research studies. For example, only occasionally did authors mention that their data met the assumptions underlying the analyses.

The number of means comparisons and/or statistical tests has increased over the past 3 decades, averaging 51 tests per articles in the 1970s (s.e. = 20), 60 tests in the 1980s (s.e. = 15) and 82 tests in the 1990s (s.e. = 20). These results are much higher than those recently reported for other journals. For instance, Anderson et al. (2000) reported that over the period from 1978 to 1997, the average number of p-values per Ecology article ranged from 10 (s.e. = 3) to 44 (s.e. = 8). While some individual articles exceeded 200 p-values, in general more statistical tests are being reported in the *Journal of Range Management*. Statistical testing in the *Journal of Range Management* also exceeds that reported in the *Journal of Wildlife Management*, where the average ranged from 31 (s.e. = 6) to 56 (s.e.= 16)

during 1994-1998 (Anderson et al. 2000).

It appears there is a belief that statistical testing is necessary for a study to be scientifically valid (Cherry 1998). Several notable individuals, including statisticians (Yates 1951, Cox 1977) have recognized the overuse of statistical testing in the literature over many years. In contrast, Johnson (1999) notes the lack of use of ordinary confidence intervals, despite being more informative than p-values. We found many articles with large tables of means that were compared within rows and columns. One article from 1992 contained a whopping 328 statistical test results. The potential for Type I error in such cases is extremely large, leading to spurious effects described by Anderson et al. (2001a).

When p-values were reported within an article, they were naked in most cases, although the frequency of such practice appears to be declining (1970s: 80%, 1980s: 69%, 1990s: 58%). In some cases, the effect size or direction was not reported, leaving one to wonder how large the difference or treatment effect was. For example, in a 1978 article, the following results were reported, "Lotebushes used by quail averaged 3.8 m^3 and were significantly ($P < 0.05$) larger than plants randomly chosen". There is no indication of how large randomly chosen plants were and how much variability there was in these sample means.

Sample means were commonly reported in articles (even when no statistical tests were performed), but they were rarely accompanied with a measure of their variability, i.e., a standard error. Standard errors were reported along with means in 6.8% of the articles from the 1970s, 11.6% of the articles from the 1980s and 12.6% of the articles from the 1990s. When standard errors were reported, often there was no mention of evaluating homogeneity of variances, creating a potential for inaccurate values. On many occasions, large tables of means were presented without accompanying standard errors. To be fair, several studies reported the sample mean along with the sample standard deviation, s. Such practice is reasonable if describing the sample is the intention; no inferential process is being initiated. The sample standard deviation is a descriptive statistic whereas the standard error is an inferential statistic (Anderson et al. 2001b). However, when the following combination is reported $\overline{x} \pm s$ there is no meaningful interpretation for this interval as an interval estimator for the true population mean. The standard error or the standard error multiplied by a t-statistic (for a given confidence level) should be used for such constructions.

Qualitative Observations

We noted a variety of misinterpretations of statistics, a listing of which is beyond the scope of our study. We have categorized the most frequent types of mistakes in 3 areas. We noted that several papers lacked an adequate description regarding the treatment application and analysis methodology. For example, in a 1978 article the following statement appears, "Seven treatments and one control were used to evaluate the effect of fire on quail habitat." There is no indication of the design or the treatment structure. Were the treatments randomly assigned? Replication is never mentioned, although it appears there is none. The article then states, "Thirty plants were selected in each of the seven treatment areas plus the control. Fifteen of the lotebushes selected were used by quail and 15 were randomly chosen." One might ask if only lotebushes were sampled from and whether or not the 15 lotebushes used by quail were randomly selected from all lotebushes used by quail. Two of the most basic principles of experimental design appear to have been ignored or at least not adequately described. We refer the reader to Wester (1992) for an excellent discussion about design principles and their use in range science. The first sentence of the last paragraph of the methods states, "Both parametric and nonparametric tests were used to evaluate the data." With the exception of mentioning the use of Spearman's Rho Test for correlating home range size with covey size and woody plant density, no other information is given regarding what testing procedures were used in the study.

Another common mistake made in the papers we examined was the failure to recognize the correlation of observations observed on the same experimental unit over time. Repeated measures designs are often used unknowingly and are not analyzed accordingly, despite a SRM presentation and proceedings paper by Engemen et al. (1986). For example, in a 1989 article, 10 bulls were randomly sampled from 2 cattle herds (one sedentary and the other migratory). Fecal samples were collected biweekly, pooled, and analyzed for fecal nitrogen and fecal acid detergent fiber (ADF). The bulls were weighed on a monthly basis. Average percent weight change and fecal measurements were then correlated without regard to the lack of independence between monthly measurements. Gurevitch and Chester (1986) emphasize that ignoring the correlative structure among observations from the same individual can lead to faulty test results. Furthermore, only under certain conditions (Huynh and Feldt 1970, Milliken and Johnson 1992) can repeated measures data be analyzed via univariate split-plot analysis.

Finally, pseudoreplication issues plagued several papers. Pseudoreplication is a pervasive problem in many scientific areas and has been repeatedly warned against in the ecological literature (Hurlbert 1984, Heffner et al. 1996) and range science literature (Brown and Waller 1986, Wester 1992). Walker and Richardson (1986) clarified the differences between pseudoreplication and true replication in grazing system studies, the key to which lies in identification of the experimental unit. We repeat their plea for proper reporting of results when replication was not achieved due to logistical difficulties. As an example, we refer to the paper on cattle live weight changes described before. Ten bulls were randomly selected from a migratory herd and a sedentary herd. The 10 bulls represent replicates with respect to their specific herds, but do not represent replicates with respect to the 'treatment' of herd type (migratory or sedentary). Furthermore, by pooling fecal samples from the 10 bulls within a herd type, they no longer are useful as replications for their respective herds.

Recommendations

Use of confidence intervals as interval estimators, rather than relying on single point estimators and tests between them, is more informative because it inherently gives the effect size and a measure of its precision. Displaying such values in figures is particularly appealing because of the ease with which one can compare the various responses at different treatment levels. Confidence intervals can still be used to test statistical hypotheses, but they have the added advantages mentioned earlier. The current editor of the *Journal of Wildlife Management* has instructed future authors to present measures of central tendency and dispersion in lieu of excessive use of p-values (Brennan 2001). When reporting the results from a statistical significance test, include the actual p-value (not a range), along with a test statistic and its degrees of freedom. Clearly distinguish between an observational study and an experiment when describing the research,

so that p-values under the former can be viewed with a greater degree of skepticism. Additionally, go beyond statistical significance and elaborate on the practical significance of the results. Brennan (2001) suggests more research is needed to determine what effect size has on a meaningful impact on a system. Finally, seek statistical advice at the beginning of a study. The most important time for statistical input is during the planning stages of a study rather than after a data set has been collected.

Literature Cited

Anderson, D.R., K.P. Burnham, and W.L. Thompson. 2000. Null hypothesis testing: problems, prevalence and an alternative. J. Wildl. Manage. 64:912–923.

Anderson, D.R., K.P. Burnham, W.R. Gould, and S. Cherry. 2001a. Concerns about finding effects that are actually spurious. Wildl. Soc. Bull. 29:311–316.

Anderson, D.R., W.A. Link, D.H. Johnson, and K.P. Burnham. 2001b. Suggestions for presenting the results of data analyses. J. Wildl. Manage. 65:373–378.

Brennan, L.A. 2001. New methods for data analysis and presentation of results. J. Wildl. Manage. 65:172.

Brown, M.A. and S.S. Waller. 1986. The impact of experimental design on the application of grazing research results- an exposition. J. Range Manage. 39:197–200.

Cherry, S. 1998. Statistical tests in publications of the Wildlife Society. Wildl. Soc. Bull. 26:947–953.

Cox, D.R. 1977. The role of significance tests. Scand. J. Stat. 4:49-70

Ellison, A.M. 1996. An introduction to Bayesian inference for ecological research and environmental decision-making. Ecol. Appl. 6:1036–1046.

Engemen, R.M., D.E. Palmquist, and L.L. McDonald. 1986. The use of repeated measurement designs in field studies. pp. 59–66 *In* Statistical Analyses and Modeling of Grazing Systems Symposium Proceedings. February 11, 1986, Kissimmee, Fla. Soc. for Range Manage., Denver, Colo.

Gurevitch, J. and S.T. Chester, Jr. 1986. Analysis of repeated measures experiments. Ecol. 67:251–255.

Guthery, F.S., J.J. Lusk, and M.J. Peterson. 2001. The fall of the null hypothesis: liabilities and opportunities. J. Wildl. Manage. 65: 379–384.

Harlow, L.L., S.A. Mulaik, and J.H. Steiger (eds.). 1997. What if there were no significance tests? Lawrence Erlbaum Assoc., Mahwah, N.J.

Heffner, R.A., M.J. Butler, and C.K. Reilly. 1996. Pseudoreplication revisited. Ecol. 77: 2558–2562.

Hurlbert, S.H. 1984. Pseudoreplication and the design of ecological field experiments. Ecol. Monogr. 54:187–211.

Huynh, H. and L.S. Feldt. 1970. Conditions under which mean square ratios in repeated measures designs have exact F-distributions. J. Amer. Stat. Assoc. 65:1582–1589.

Johnson, D.H. 1999. The insignificance of statistical significance testing. J. Wildl. Manage. 63:763–772.

Milliken, G.A. and D.E. Johnson. 1992. Analysis of messy data: designed experiments. Chapman and Hall, New York, N.Y.

Nester, M.R. 1996. An applied statistician's creed. Appl. Stat. 45:401–410.

Walker, J.W. and E.W. Richardson. 1986. Replication in grazing studies- why bother? pp. 51–58 *In* Statistical Analyses and Modeling of Grazing Systems Symposium Proceedings. February 11, 1986, Kissimmee, Fla. Soci. for Range Manage., Denver, Colo.

Wester, D.B. 1992. Viewpoint: replication, randomization and statistics in range science. J. Range Manage. 45:285–290.

Yates, F. 1951. The influence of statistical methods for research workers on the development of the science of statistics. J. Amer. Stat. Assoc. 46:19–34.