Type I and Type II Errors in the Real World

Authors: Berger, Joel, and Kock, Michael D.

Source: Journal of Wildlife Diseases, 25(3) : 451-454

Published By: Wildlife Disease Association

URL: https://doi.org/10.7589/0090-3558-25.3.451

Journal of Wildlife Diseases, 25(3), 1989, pp. 451–454

RESPONSE TO LETTER TO THE EDITOR . . .

Type I and Type II Errors in the Real World

Biostatistics are used in two principle fashions, to test hypotheses and to estimate population parameters, both of which are used to gain reliable knowledge. Individuals involved in a nascent science like the management of wild animals and plants have an obvious need to use such knowledge, but they frequently do not have the statistical backgrounds necessary to discriminate among possible biases in the presentation of data. Hence, any errors about the reliability of results may be serious. Halverson and Teare (1989) suggest that our (Berger and Kock, 1988) statement ". . . carfentanil had no long term negative effect on the survival of bull bison" is without basis, and therefore "there is a danger that the unfounded conclusions will misdirect the decisions of wildlife managers working in the field." While we agree with Halverson and Teare (1989) that care must be taken in all analyses and that Type II errors can be very serious, we believe that there are good reasons why studies, even those with small samples, should focus on hypothesis testing. Our response considers three issues: acquisition of reliable knowledge, Type I and Type II errors, and the utility of our data for wildlife management.

Gaining knowledge

Reliable knowledge may be defined as those sets of ideas that are consistent with facts (Romesburg, 1981). Obviously any error involving the statistical analysis of data will not lead to reliable knowledge. But, how is reliable knowledge obtained? Many texts deal with the scientific method and argue that experiments must be designed to examine hypotheses about supposed causal links of factors. These include solving methodological and logistical problems, and minimizing statistical errors, but if one is not to be misled more has to be done than to report accurately the statistical parameters of a study. In the strictest sense, one needs to test hypotheses because science cannot be advanced when a hypothesis is not falsifiable (Popper, 1968); thus, while Halverson and Teare argue that estimation of parameters is useful (and we agree), it would not be prudent to adopt parameter estimation as an alternative to hypothesis testing. They should go hand in hand.

Statistical tests and Type I and Type II errors

This raises the question of whether studies employing small samples are useful because their statistical power is limited (Cohen, 1977). Halverson and Teare (1989) correctly point out that we accepted a null hypothesis rather than failing to reject it. And, they bring forth a valid point that our sample (21 treated and 30 control bison) may be too small to permit much power when testing for a Type II error. Biologists often use samples similar in size to ours; for example, queries about sex ratio adjustment in animals by Berman (1988) and Silk (1983). Often there is reliance on alpha levels of ≤ 0.05 since a limit is then set for the probability of falsely rejecting the null hypothesis (a Type I error as defined by Sokal and Rohlf, 1981). Quite obviously, reliable knowledge is best gathered by minimizing the chances of both types of error (I or II), but it is also important to know which type of error is most serious.

When true null hypotheses (e.g., in our case, that carfentanil has no effect on overwinter survival in male bison; Berger and Kock, 1988) are tested, Type II errors may be more common than Type I, and hence one tends to err on the side of ignorance by reporting the lack of detection of an experimental treatment. Connor and Sim-

berloff (1986) point out that to avoid the championing of favored ideas the prudent approach is to err on the side of ignorance, which is one of the premises of the American judiciary system—that one is innocent until proven guilty. On the other hand, when testing products such as pesticides or drugs, Type II errors can be costly because reporting no effect when there is one may entail major expenses (Toft and Shea, 1983). This is exactly the case presented by Halverson and Teare (1989) concerning our data.

Nevertheless, we contend, as have Connor and Simberloff (1986), that the more costly error depends on the type of inquiry. Had our initial investigation been designed to detect the absolute minimum difference in overwinter survival given a 0.076 difference between treatment and control groups, we would have had to sample 684 animals, replicate our procedures at numerous sites, or work with different subsamples at Badlands over many years. Instead, we chose to report "There was no statistical difference between the two samples" (Berger and Kock, 1988) and draw explicit attention to other possible factors which may have affected our results, such as density or climate. Perhaps, it would have been more appropriate to suggest that in the absence of true replication and knowledge about the home ranges or condition of bulls, we failed to detect any statistical difference between the two samples but, because our sample sizes were limited, it was not possible to say with certainty whether we may accept or reject the null hypothesis. Quite simply, we have been unable to report statistically that an effect of carfentanil on overwinter survival exists.

The claim that no basis exists for rejecting our null hypothesis (no treatment effect) is bolstered by data provided by the National Park Service (R. Klukas, unpubl. data) at Wind Cave National Park (WCNP), South Dakota, an area where free-roaming adult male bison were immobilized with carfentanil during November 1986 (2 mo after we immobilized bison in Badlands National Park). Of 25 immobilized bulls, 24 survived the winter. The exact number of bulls living in the park was unknown, but about 150 were estimated. Hence, 125 were not immobilized. None died over the winter (Klukas, pers. comm.); to be conservative and avoid overinflating the estimate, we assume that 100 existed and serve here as controls. A comparison between the control and treated sample reveals overwinter survival to be 100 and 96%, respectively, samples which are not statistically different (G_{adj} = 1.73; NS; see below for further statistical details). Thus, using data from two geographically separated bison populations we have been unable to detect an effect of carfentanil on overwinter survival.

We could have gained statistical power by pooling data from the two bison populations (which, if done, results in a significant effect of carfentanil; G_{adj} = 4.02; $P < 0.05$) but this would be incorrect for two primary reasons: (1) the samples are not true replicates because environmental conditions, management practices, and the period of winter weather experienced by bison differed between sites; and (2) pooling the data results in the loss of information (Sokal and Rohlf, 1981).

Interestingly, Halverson and Teare (1989) criticized our use of a Chi Square Test with an adjustment for small samples, although support exists for this test when cell frequencies are <10 (Bruning and Kintz, 1977). Instead they suggest the Fisher Exact Test. Its use is somewhat controversial (Berkson, 1978) and it is simply the wrong test for our data because it is based on the hypergeometric distribution which assumes that both column and row totals are fixed (Sokal and Rohlf, 1981). In our analysis it is true that column totals are fixed (the number of individuals in each category, treated and untreated), but the number of animals which lived or died is not. Sokal and Rohlf (1981, p. 735), in a later edition than that cited by Halverson and Teare (1989), recommend the *G*-test

of independence when marginal totals are fixed for one criterion.

Utility of our data for wildlife management

Almost 60 yr ago Fisher and Wishart (1930) admonished, "No one would now dream of testing the response to a treatment by comparing two plots, one treated and the other untreated." Yet, the present state of wildlife management is in most cases reduced to contrasts between populations where treatments are not replicated though samples may be, a process designated "pseudoreplication" (Hurlbert, 1984). This is unfortunate, but when dealing with large, wild mammals replication is usually not an option. We advise wildlife managers to accept our original data for what they are, a comparison between 21 (treatment) and 30 (control) animals, which is why sample sizes are reported in the first place. We have now provided additional information on overwinter survival from another area, again based on samples which suggest a trivial and non-significant effect (one animal died of 25 tested). It seems clear, at least to us, that in most scientific writing it is accepted convention that errors of both types (I and II) may occur, but attempting to minimize the probabilities of these errors does not necessitate reporting the probabilities with each analysis. Differences exist between scientific and statistical hypotheses, and wildlife managers need some idea on which to base decisions. Halverson and Teare (1989) offer an illustrative example in which Type II errors may be important, and suggest that confidence limits may be a useful alternative to hypothesis testing, a suggestion not without basis in some aspects of testing for disease in natural populations (Wehausen, 1987). But, in their zeal to improve our statistical presentation, they overlooked an assumption important to their recommended test and how reliable knowledge may be obtained.

In the real world, it is not easy to employ true replication in studies of large mammals. We have tried to do the next best thing, report the results of two similar studies, neither of which has been able to detect effects of a treatment. We believe that it is important for wildlife managers to know this, as well as the probability of accepting a false null hypothesis. Halverson and Teare (1989) bring forth a useful point about sample sizes and Type II errors; without their concern we would not have uncovered the Wind Cave data, which parallel our original findings.

We thank Drs. Wendy Green, Stephen Jenkins, Aron Rothstein, and John Wehausen for their comments on this note.

LITERATURE CITED

BERGER, J., AND M. D. KOCK. 1988. Overwinter survival of carfentanil-immobilized male bison. Journal of Wildlife Diseases 24: 555–556.

BERKSON, J. 1978. In dispraise of the exact test. Journal of Statistical Planning and Inference 2: 27–42.

BERMAN, C. M. 1988. Maternal condition and offspring sex ratio in a group of free-ranging rhesus monkeys: An eleven year study. American Naturalist 131: 307–328.

BRUNING, J. L., AND B. L. KINTZ. 1977. Computational handbook of statistics, 2nd ed. Scott, Foresman, and Company, Glenview, Illinois, 308 pp.

COHEN, J. 1977. Statistical power analysis for the behavioral sciences. Academic Press, New York, New York, 474 pp.

CONNOR, E. F., AND D. SIMBERLOFF. 1986. Competition, scientific method, and null models in ecology. American Scientist 74: 155–162.

FISHER, R. A., AND J. WISHART. 1930. The arrangement of field experiments and the statistical reduction of the results. Imperial Bureau of Soil Science (London), Technical Communication 10: 1–23.

HALVERSON, T. G., AND J. A. TEARE. 1989. Carfentanil and overwinter survival in bison: The alternative hypothesis. Journal of Wildlife Diseases 25: 448–450.

HURLBERT, S. H. 1984. Pseudoreplication and the design of ecological field experiments. Ecological Monographs 54: 187–211.

POPPER, K. R. 1968. The logic of scientific discovery. Harper and Row, New York, New York, 480 pp.

ROMESBURG, H. C. 1981. Wildlife science: Gaining reliable knowledge. The Journal of Wildlife Management 45: 293–313.

SILK, J. B. 1983. Local resource competition and

facultative adjustment of sex ratios in relation to competitive abilities. American Naturalist 121: 56–66.

SOKAL, R. R., AND F. J. ROHLF. 1981. Biometry, 2nd ed. W. H. Freeman and Company, New York, New York, 859 pp.

TOFT, C. A., AND P. J. SHEA. 1983. Detecting community-wide patterns: Estimating power strengthens statistical inference. American Naturalist 122: 618–625.

WEHAUSEN, J. D. 1987. Some probabilities associated with sampling for diseases in bighorn sheep. Desert Bighorn Council 1987 Transactions 31: 8–10.

Joel Berger[1,2] **and Michael D. Kock,**[3] [1] Departments of Range, Wildlife, and Forestry, and Biology, University of Nevada, Reno, Nevada 89512, USA; [2] Conservation and Research Center, Front Royal, Virginia 22630, USA; [3] Faculty of Veterinary Science; P.O. Box MP 167; University of Zimbabwe, Mt. Pleasant, Harare, Zimbabwe.