INFECTION AND IMMUNITY, Dec. 2003, p. 6689–6692 Vol. 71, No. 12
0019-9567/03/$08.00+0 DOI: 10.1128/IAI.71.12.6689–6692.2003

GUEST COMMENTARY

Review of the Use of Statistics in *Infection and Immunity*

Cara H. Olsen*

Uniformed Services University of the Health Sciences, Bethesda, Maryland

The use of statistics in scientific and medical journals has been subjected to considerable review in recent years. Many journals have published systematic reviews of statistical methods (4–8, 12). These reviews indicate room for improvement. Typically, at least half of the published scientific articles that use statistical methods contain statistical errors. Common errors include failing to document the statistical methods used or using an improper method to test a statistical hypothesis.

This study analyzes articles in *Infection and Immunity* for appropriateness of statistical analysis and reporting. I reviewed all 141 articles from two issues, January 2002 (volume 70, no. 1) and July 2002 (volume 70, no. 7); listed the statistical analyses used; and identified errors in analysis and reporting. Errors were defined with respect to the current instructions to authors (2), accepted statistical practice (1, 11), and professional judgment.

The statistical errors identified in *Infection and Immunity* are comparable to those found in similar journals: 54% of the articles reviewed contained errors of analysis (20%), reporting (22%), or both (12%). The most common analysis errors are failure to adjust or account for multiple comparisons (27 studies), reporting a conclusion based on observation without conducting a statistical test (20 studies), and use of statistical tests that assume a normal distribution on data that follow a skewed distribution (at least 11 studies). The most common reporting errors are unlabeled or inappropriate error bars or measures of variability (15 studies) and failure to describe the statistical tests performed (12 studies). These errors are discussed more fully below, with examples and suggestions for improvement.

ERRORS IN ANALYSIS

Errors in analysis may lead to misinterpretation of the data and faulty conclusions. The most common such error was failure to adjust or account for multiple comparisons. When more than two experimental groups are compared with each other or with a common control, it is usually necessary to adjust the significance level or *P* value to account for multiple comparisons. Two common scenarios are comparing multiple treatments to a control and comparing two treatments at several different time points. While 24 studies used an appropriate adjustment for multiple comparisons, 26 studies, nearly 1 in 5, reported multiple comparisons without adjusting *P* values. Typically, the authors calculated a separate *t* (or similar) test for each comparison, using a significance level of 0.05 for each comparison. The overall error rate in this case equals $1 - (1 - 0.05)^k$, where *k* is the number of comparisons (9). Therefore, if three such independent comparisons were made for a single experiment (e.g., three treatments to a common control or treatment versus control for three time points), the probability of finding a falsely significant result for any of the comparisons performed increases from 0.05 to 0.14. If more treatment groups or time points are involved, the probability of false significance increases even more.

Decisions regarding whether and how to adjust for multiple comparisons depend on several factors, including whether the comparisons were planned before the study began, how many comparisons are being made, and the study design. A simple adjustment, the Bonferroni adjustment, involves multiplying the observed *P* values by the number of comparisons made. This adjustment is somewhat conservative, especially if many comparisons are made or the comparisons are not independent (which is typically the case). The Bonferroni adjustment may result in a reduced ability to detect significant differences. For example, if a treatment is compared to a control at eight time points, the unadjusted *P* value for each comparison must be less than 0.00625 to achieve statistical significance at the 5% level under the Bonferroni adjustment. If this is a concern, other adjustments, such as Tukey's or Student-Newman-Keuls, can be performed by using standard statistical software.

If authors state that there is a significant difference among groups, they must also conduct and report a corresponding significance test. Even when the observed difference between groups appears to be quite large, there is still a possibility that the difference is due to chance. Statistical analysis is the only way to quantify the likelihood that the observed difference is not due to chance alone. Twenty studies reported significant differences in Results that were not supported by statistical tests or *P* values in the text or tables. Authors must describe which test they used, report the effect size (the appropriate measure of the magnitude of the difference, usually the difference or ratio between groups), and give a measure of significance, usually a *P* value, or a confidence interval for the difference.

Statistical significance of differences among groups cannot be determined by examining whether the error bars around the group means overlap. While only one study made this claim explicitly, authors of many of the 20 studies with unsubstanti-

* Mailing address: Uniformed Services University of the Health Sciences, Bethesda, MD 20814-4799. Phone: (301) 295-9468. Fax: (301) 295-1933. E-mail: colsen@usuhs.mil.

ated differences based their claims on visual examination of bar charts. Standard deviation bars contain no information about the precision of the mean. Standard error bars correspond roughly to a 67% confidence interval and should not be interpreted as indicating significant differences. Even error bars that represent 95% confidence intervals should not be used to determine statistically significant differences among groups, since there can be considerable overlap of confidence intervals even when there is a statistically significant difference among groups (10).

Many variables in biomedical research are not normally distributed, and many of them are positively skewed. Examples of variables that are usually skewed include cell counts, CFU counts, titers, and percentages. These variables should not be analyzed by using standard parametric tests such as *t* tests and analysis of variance (ANOVA) unless an examination of the data shows that they follow a normal distribution. Two alternative approaches are possible. If taking the log of the data, or using another transformation, results in a normal distribution, parametric tests may be performed on the transformed data. Otherwise, nonparametric tests should be used. Eight studies reported parametric tests on non-normal variables, and several other studies did not provide sufficient information to determine whether a different approach should have been used.

When variables are log transformed and analysis is performed on the transformed variables, the antilog of the result is often calculated to obtain the geometric mean. When the geometric mean is reported, it is not appropriate to report the antilog of the standard error of the mean of the logged data as a measure of variability. Instead, confidence limits (e.g., endpoints of the 95 or 99% confidence interval) should be calculated on the log scale. The antilog of the confidence limits may be presented as confidence limits on the original scale (1).

Data analyzed by nonparametric statistics (e.g., Mann-Whitney U test, Wilcoxon signed-rank test, Kruskal-Wallis test, etc.) should be reported in tables or depicted in figures as the median along with an appropriate range (minimum and maximum values, upper and lower quartiles or quintiles, etc.). Means and standard deviations (or standard errors) are not appropriate for reporting data analyzed by nonparametric statistics; if the distribution of the data is sufficiently non-normal to require nonparametric analysis, the mean and standard deviation will not provide a useful description of the location and range of the data. Of 17 studies that reported one of the nonparametric tests listed above, only 3 reported the median and upper and lower quartiles in figures or tables. One of these three studies reported quartiles for tabular data but not for the medians presented in bar charts. Eight studies reported the mean, four reported the mean of the logged data (the log of the geometric mean), and two reported the geometric mean. While such summaries are not recommended in this case, under certain circumstances, for example, when antibody titers are being summarized, the geometric mean is a good estimate of the median and is therefore acceptable (3).

The importance of using transformed data or nonparametric statistical methods for data that are not normally distributed can be illustrated by a simple numerical example. Suppose a researcher obtains the following data in an experiment: control group, 1, 2, 4, 8, and 16; treatment group, 0.25, 1, 4, 16, and 64. The treatment and control groups have the same median (4.0) and geometric mean (4.0), but the mean of the treatment group (17.1) is almost three times as high as the mean of the control group (6.2). While the difference in means is not statistically significant on the basis of a two-sided *t* test for independent samples assuming unequal variances ($t = 0.88$, $P = 0.43$), the absolute difference between the means might lead the researcher to the mistaken conclusion that the treatment has a clinically meaningful effect on the average outcome and should be studied further.

Two other types of errors, while not as common as those discussed above, were important enough to warrant mention. When multiple samples are taken from the same experimental source (e.g., multiple serum samples drawn from the same animal or patient, etc.), the use of statistics that assume independence of the samples is inappropriate. One advantage of using multiple samples from the same source is that when each source is compared to itself, variability among subjects has less influence on the results. This advantage is only realized when the appropriate statistical test is used. The paired *t* test, repeated-measures ANOVA, Wilcoxon signed-rank test, and McNemar's test are common significance tests that can be used on this type of data and correspond, respectively, to the two-sample *t* test, ANOVA, Mann-Whitney U test, and chi-square test for independent samples. Many studies failed to provide enough information to determine whether a test for independent or dependent samples was appropriate, and two studies used tests for independent samples when it was clear that each subject provided multiple cell samples. Authors should clearly indicate when multiple samples are derived from the same source and should use an appropriate statistical test.

When several independent experiments are performed, the authors should summarize each experiment or consider combining the data rather than present a single representative experiment. If it is appropriate to combine the data, authors should use a statistical method, such as randomized block ANOVA, that accounts for variability among experiments. Several studies reported results from one of several independent experiments. This is problematic because there is no indication of how the authors chose which experiment to present. They could have chosen the results that best supported their hypothesis and discarded the rest.

ERRORS IN REPORTING

Errors in reporting do not necessarily indicate that the statistical analysis was incomplete or inappropriate, simply that not enough information was given to determine whether this was the case. The most common such error was failure to describe the variability in the sample. Measures of variability such as standard deviation should accompany means, medians, and other data summaries. The standard deviation of the data is appropriate in most circumstances because it describes the variation among observations in the sample. When data are analyzed by using nonparametric or distribution-free methods, such as the Mann-Whitney U test, it is generally more appropriate to report the median values along with the minimum and maximum values, 25th and 75th percentiles, or other central range.

Thirty-five (40%) of the 88 studies that reported means along with a measure of variability reported the standard error

of the mean instead of the standard deviation. The standard error describes the precision with which the sample mean estimates the true population mean but does not provide direct information about the variability in the sample. Because the interpretation of the standard error is different from that of the standard deviation, it is critical to indicate which summary is reported.

Twelve studies failed to identify which measure of dispersion was reported, and three failed to report any measure of variability at all. For example, many studies presented bar charts in which each bar represents a group of subjects receiving the same treatment and the height of each bar represents the mean response of the subjects in each group. While most such charts included error bars, it is impossible to interpret the error bars unless the authors indicate whether they represent standard deviations, standard errors, or some other measure of variability. In the text, the average response for a group is often reported as the mean ± a measure of variation. Authors and reviewers should be sure that both of these terms are clearly defined, for example, by stating in Materials and Methods that the "results are expressed as means ± standard deviations".

All significance tests should be defined in Materials and Methods and in the text, tables, or charts in which the results appear. Defining the tests allows the reader to evaluate whether the appropriate test was chosen and to interpret the results correctly.

Some tests, such as Student's *t* test and the Wilcoxon test, have several variants, and authors should indicate which variant was chosen. When Student's *t* test is used, the authors should indicate whether the version for paired samples or independent samples was used, whether the test was two sided or one sided, what significance level was chosen (e.g., "*P* values less than 0.05 were considered statistically significant"), and, in the case of the independent-sample test, whether the variances in the two groups being compared were assumed to be equal. For most other significance tests, it is sufficient to report the name of the test, the significance level, and whether it is one sided or two sided.

Most studies failed to report these details, and 12 studies failed to report even the name of each reported significance test. Six studies described more than one significance test in Materials and Methods but did not indicate when each was used, and six studies did not even mention which tests were used. Often, *P* values indicating significant differences between groups were reported in tables or the text with no accompanying explanation. Of the 44 studies that reported *t* tests, only 19 indicated whether they used the form of the test appropriate for matched pairs (6 studies) or independent samples (13 studies). One study reported use of the Wilcoxon test to compare two groups but did not indicate whether Wilcoxon's rank sum test (for independent samples) or Wilcoxon's signed-rank test (for paired samples) was performed. One study indicated that ANOVA would be used when the data were normally distributed and the Mann-Whitney U test would be used for other comparisons. While this is an appropriate analysis strategy, the study did not indicate which test was used for each reported comparison.

FURTHER ISSUES

It is also important for researchers to decide on a statistical analysis strategy when planning the study. It is tempting to try several different approaches and report the one that has the greatest statistical significance; this approach leads to increased reporting of falsely significant results. Statistical tests should be chosen before the data are analyzed, and the choice should be based on the study design and distribution of the data, not the results.

Researchers often conclude that if the difference between two treatment groups in their study is not statistically significant, there is no difference between the treatments. In fact, it is often the case that the study did not have sufficient power to detect a difference between the groups. Power is the probability that a study can detect a statistically significant difference between treatment groups if the treatments differ, and it depends on the sample size, the magnitude of the treatment effect, and the variability among subjects. Studies reported in *Infection and Immunity* tend to have small sample sizes, especially when animals are used as subjects. Many studies use only three or four animals per treatment group; samples this small do not have sufficient power for meaningful statistical comparisons among groups unless the variability among subjects is small and the effect of the treatment is large. For example, when a two-sided *t* test for independent samples is used to compare two treatment groups with a significance level of 5%, a sample size of three animals per group will have 80% power to detect a difference equal to 3.1 standard deviations between the groups. A sample size of four animals per group will have 80% power to detect a difference equal to 2.4 standard deviations. Investigators are usually interested in detecting differences much smaller in magnitude, such as a difference equal to one-half of a standard deviation. A sample size of three will have 7% power to detect a one-half standard deviation difference, while a sample of size four will have a power of only 9%. When several groups are being compared, statistical power may be greater if the hypothesis can be simplified; e.g., conducting a test for trend across all groups instead of comparing each group separately to a control. Sample sizes should be calculated in advance by considering the variation among subjects and the smallest difference that is clinically important; choosing a sample size on the basis of cost or convenience alone may result in wasted resources and meaningless statistics.

In summary, while most of the statistics reported in *Infection and Immunity* are fairly straightforward comparisons of treatment groups, even these simple comparisons are often analyzed or reported incorrectly. Authors and reviewers should be aware of common pitfalls. Recognition and understanding of these errors should help researchers choose appropriate statistical methods and use them correctly, thereby improving the quality of the published research.

REFERENCES

1. **Altman, D. G., S. M. Gore, M. J. Gardner, and S. J. Pocock.** 1983. Statistical guidelines for contributors to medical journals. BMJ **286:**1489–1493.
2. **American Society for Microbiology.** 2002. Instructions to authors. Infect. Immun. **70:**iii–xvi.
3. **Armitage, P., and G. Berry.** 1994. Statistical methods in medical research, 3rd ed. Blackwell, Oxford, England.

4. **Avram, M. J., C. A. Shanks, M. H. Dykes, A. K. Ronai, and W. M. Stiers.** 1985. Statistical methods in anesthesia articles: an evaluation of two American journals during two six-month periods. Anesth. Analg. **64:**607–611.
5. **Cruess, D. F.** 1989. Review of use of statistics in the American Journal of Tropical Medicine and Hygiene for January-December 1988. Am. J. Trop. Med. Hyg. **41:**619–626.
6. **Emerson, J. D., and G. A. Colditz.** 1983. Use of statistical analysis in the New England Journal of Medicine. N. Engl. J. Med. **309:**709–713.
7. **Felson, D. T., L. A. Cupples, and R. F. Meenan.** 1984. Misuse of statistical methods in Arthritis and Rheumatism: 1982 versus 1967–68. Arthritis Rheum. **27:**1018–1022.
8. **MacArthur, R. D., and G. G. Jackson.** 1984. An evaluation of the use of statistical methodology in the Journal of Infectious Diseases. J. Infect. Dis. **149:**349–354.
9. **Ott, R. L., and M. T. Longnecker.** 2001. An introduction to statistical methods and data analysis, 5th ed. Brooks/Cole, Pacific Grove, Calif.
10. **Schenker, N., and J. Gentleman.** 2001. On judging the significance of differences by examining the overlap between confidence intervals. Am. Statistician **55:**182–186.
11. **van Belle, G.** 2002. Statistical rules of thumb. John Wiley & Sons, Inc., New York, N.Y.
12. **White, S.** 1979. Statistical errors in papers in the British Journal of Psychiatry. Br. J. Psychiatry **135:**336–342.

The views expressed in this Commentary do not necessarily reflect the views of the journal or of ASM. Editor: D. L Burns.