Common Statistical Errors Even YOU Can Find*

Part 1: Errors in Descriptive Statistics and in Interpreting Probability Values

Tom Lang, MA
Tom Lang Communications

"Critical reviewers of the biomedical literature have consistently found that about half the articles that used statistical methods did so incorrectly."[1]

Statistical probability was first discussed in the medical literature in the 1930s.[2] Since then, researchers in several fields of medicine have found high rates of statistical errors in large numbers of scientific articles, even in the best journals.[3-6] The problem of poor statistical reporting is, in fact, long-standing, widespread, potentially serious, and almost unknown, despite the fact that most errors concern basic statistical concepts and can be easily avoided by following a few guidelines.[7]

The problem of poor statistical reporting has received more attention with the rise of the evidence-based medicine movement. Evidence-based medicine depends on the quality of published research; that is, evidence-based medicine is literature-based medicine. As a result, several groups have proposed reporting guidelines for different types of trials,[8-10] and a comprehensive set of guidelines for reporting statistics in medicine has been compiled from an extensive review of the literature.[11]

In a series of articles, I will describe several of the more common statistical errors found in the biomedical literature, errors that can be identified even by those who know little about statistics. These guidelines are but the tip of the iceberg; readers who want to know more about the iceberg should consult more detailed texts,[11] as well as other references cited in this series.

The field of statistics can be divided into two broad areas: **descriptive statistics**, which is concerned with how to describe samples of data collected in a research study, and **inferential statistics**, which is concerned with how to estimate (or infer) from the sample the characteristics of the population from which the sample was selected. In this article, I describe errors made in defining variables, in summarizing the data collected about these variables, and in interpreting probability (*P*) values.

Errors in Descriptive Statistics

Error #1: Not Defining Each Variable in Measurable Terms

Science is measurement. Researchers need to tell us what they measured—the variables—and how they measured them, by providing the **operational definition** of each variable. For example, one operational (measurable) definition of hypertension is a systolic blood pressure of 140 mm Hg or higher, and an operational definition of obesity is a body mass index above 27.3 for women and above 27.8 for men.

Variables relating to concepts or behaviors may be more difficult to measure. Depression defined as a score of more than 50 on the Zung Depression Inventory is operationally defined, but how well the Inventory actually measures depression can be debated. In one major U.S. survey, a "current smoker" is anyone who smoked one cigarette in the 30 days before the survey. Although this definition is not an obvious one, it is nevertheless an "operational" one, and we at least know who "current smokers" are in the survey, even if we disagree with the definition.

Error #2: Not Providing the Level of Measurement of Each Variable

Level of measurement refers to how much information is collected about the variable. For practical purposes, there are three levels of measurement: nominal, ordinal, and continuous. At the lowest level are **nominal data**, which consist of two or more nominal, or named, categories that have no inherent order. Blood type defined as type A, B, AB, or O is measured at the nominal level of measurement.

Ordinal data consist of categories that *do* have an inherent order and can be sensibly ranked. A person may

*This series is based on 10 articles first translated and published in Japanese by Yamada Medical Information, Inc. (YMI, Inc.), of Tokyo, Japan. Copyright for the Japanese articles is held by YMI, Inc. The *AMWA Journal* gratefully acknowledges the role of YMI in making these articles available to English-speaking audiences.

be described as short, medium, or tall. We may not know the exact height of the patients studied, but we do know that a person in the tall category is taller than one in the medium category, who, in turn, is taller than one in the short category.

Continuous data consist of values along a continuous measurement scale, such as height measured in centimeters or as blood pressure measured in millimeters of mercury. Continuous data are the highest level of measurement because they tell how far each point value is from any other value on the same scale.

Researchers need to specify the level of measurement for each variable. For example, they may wish to characterize a patient's blood pressure as a nominal variable (either elevated or not elevated), as an ordinal variable (hypotensive, normotensive, or hypertensive), or as a continuous variable (the systolic pressure in millimeters of mercury). The levels of measurement of response and explanatory variables are important because they determine the type of statistical test that can be used to analyze relationships. Different combinations of levels of measurement require different statistical tests.

Error #3: Dividing Continuous Data into Ordinal Categories Without Explaining Why or How the Categories Were Created

To simplify statistical analyses, continuous data, such as height measured in centimeters, are often separated into two or more ordinal categories, such as short, medium, and tall. Reducing the level of measurement in this way also reduces the precision of the measurements, however, as well as reducing the variability in the data. Authors should explain why they chose to lose this precision. In addition, they should explain how the boundaries of the ordinal categories were determined, to avoid the appearance of bias. In some cases, the boundaries (or cut points) that define the categories can be chosen to favor certain results.

Error #4: Using the Mean and Standard Deviation to Describe Continuous Data That Are Not Normally Distributed

Unlike nominal and ordinal data, which are easily summarized as the number or percent of observations in each category, continuous data can be graphed to form distributions. Distributions are usually described with a value summarizing the bulk of the data—the mean, median, or mode—and a range of values that represent the variation of the data around the summary value—the range, the interpercentile range, or the standard deviation (SD).

Normal distributions are appropriately described with any of the above descriptive statistics, although the mean and the SD are used most commonly. In fact, the mean and the SD should be used *only* to describe approximately normal distributions. By definition, about 67% of the values of a normal distribution are within ±1 SD of the mean, and about 95% are within ± 2 SDs. **Non-normal** or **skewed distributions,** however, are *not* appropriately described with the mean and the SD. The **median** value (the value that divides observations into an upper and a lower half) and the **interquartile range** (the range of values that include the middle 50% of the observations) are more appropriate for describing non-normally distributed data.

Most biologic data are not normally distributed, so the median and interquartile range should be more common than the mean and the SD. A useful rule of thumb is that if the SD is greater than half of the mean (and negative values are not possible), the data are not normally distributed.

Error #5: Using the Standard Error of the Mean (SEM) As a Descriptive Statistic

Unlike the mean and the SD, which are *descriptive statistics* for a *sample* of (normally distributed) data, the **standard error of the mean (SEM)** is a *measure of precision* for an estimated characteristic of a *population.* (One SEM on either side of the estimate is essentially a 67% confidence interval [see later]. However, the SEM is often reported instead of the SD. The SEM is always smaller than the SD, and so its use makes measurements look more precise than they are. In addition, the preferred measure of precision in the life sciences is the 95% confidence interval. Thus, measurements (when normally distributed) should be described with the mean and SD, not SEM, and an estimate should be accompanied by the 95% confidence interval, not the SEM.

Errors in Interpreting Probability (*P*) Values

"We think of tests of significance more as methods of reporting than for making decisions because much more must go into making medical policy than the results of a significance test."[12]

Probability (*P*) values can be thought of as the amount of evidence in favor of chance as the explanation for the difference between groups. When the probability is small, usually less than five times in 100, chance is rejected as the cause, and the difference is attributed to the intervention under study; that is, *P* values indicate mathematical probability, not biologic importance. Probability values are compared to the alpha level that

defines the threshold of statistical significance. Alpha is often set at 0.05. A *P* value below alpha is "statistically significant"; a *P* value above alpha is "not significant at the 0.05 level." This all-or-none interpretation of a *P* value and the fact that any alpha level is arbitrary are other causes of misinterpretation.

A *P* value can help to decide whether, say, two groups are significantly different. The *lack* of statistical significance, however, does not necessarily mean that the groups are similar. Concluding that groups are equivalent because they do not differ significantly is another common misinterpretation.

Error #6: Reporting Only *P* Values for Results

The problems described have led journals to recommend reporting the 95% confidence interval for the difference between groups (that is, for the "estimate") instead of, or in addition to, the *P* value for the difference.[13] The following examples show the usefulness of confidence intervals.[11]

- *The effect of the drug on lowering diastolic blood pressure was statistically significant* ($P<0.05$). Here, the *P* value could be 0.049; statistically significant (at the 0.05 level) but so close to 0.05 that it should be interpreted similarly to a *P* value of, say, 0.051, which is *not* statistically significant. In addition, we do not know by how much the drug lowered diastolic pressure; that is, we cannot judge the clinical importance of the reduction.
- *The mean diastolic blood pressure of the treatment group dropped from 110 to 92 mm Hg* ($P=0.02$). This presentation is the most typical. The values before and after the test are given, but not the difference. The mean drop—the 18-mm Hg difference—is statistically significant, but it is also an *estimate* of the drug's effectiveness, and without a 95% confidence interval, the precision (and therefore the usefulness) of the estimate cannot be determined.
- *The drug lowered diastolic blood pressure by a mean of 18 mm Hg, from 110 to 92 mm Hg* (95% *CI = 2 to 34 mm Hg*; $P=0.02$). The confidence interval indicates that if the drug were to be tested on 100 samples similar to the one reported, the average drop in blood pressure would range between 2 and 34 mm Hg in 95 of the 100 samples. A drop of only 2 mm Hg is not clinically important, but a drop of 34 mm Hg is. So, although the *mean* drop in blood pressure in this particular study was statistically significant, the expected difference in blood pressures may not always be clinically important; that is, the study results are actually *inconclusive*. For conclusive results, more patients probably need to be studied to narrow the confidence interval until *all* or *none* of the values are clinically important.

Error #7: Not Confirming That the Assumptions of Statistical Tests Were Met

Statistical tests may not give accurate results if their assumptions are violated.[14] For this reason, both the name of the test and a statement that its assumptions were met by the data should be included when reporting statistical analyses. The most common errors are

- Using parametric tests (which require data to be normally distributed) when the data are skewed. In particular, when comparing two groups, the Student *t* test is often used when the Wilcoxon rank sum test (or another nonparametric test that does not assume normally distributed data) is more appropriate.
- Using tests for independent samples on paired samples, which require tests for paired data. Again, the Student *t* test is often used when a paired *t* test is required.
- Using linear regression analysis without establishing that the relationship between variables is, in fact, linear. (The assumption of linearity may be tested by what is called an analysis of "residuals."[11])

Error #8: Interpreting Nonstatistically Significant Results As "Negative" When They Are, in Fact, Inconclusive

A researcher who finds no statistically significant difference between experimental groups must decide whether the lack of difference means that the groups were, in fact, similar (the intervention made no difference), or that too few data were collected to detect a meaningful difference. This decision is usually made with a **power calculation**, which determines how many subjects must be studied to have a given chance of detecting a given difference, if such a difference is there to be detected.

Unfortunately, many studies reporting nonstatistically significant findings are underpowered and, therefore, do not provide conclusive answers.[15] The researchers found no difference, but neither can they rule out the existence of a difference. Absence of proof is not proof of absence.

In adequately powered studies, statistically insignificant results are truly negative: the groups being compared are, in fact, similar because no difference was found, but a difference *could* have been found had it existed in the data. Adequate power is especially important in equivalence trials (or noninferiority trials), which are conducted to establish that one drug is as good as another.

Error #9: Not Reporting Whether or How Adjustments Were Made for Multiple Hypothesis Tests

Many studies report several *P* values, which increases the risk of making a **type I error**: concluding that the difference found is the result of an intervention when chance is a more likely explanation. For example, to compare each of six groups to all the others, 15 pair-wise statistical tests—15 *P* values—are needed. Without adjusting for these multiple tests, the chance of making a type I error rises from 5 times in 100 (the typical alpha level of 0.05) to 55 times in 100 (an alpha of 0.55).

The multiple testing problem may be encountered when

- **Establishing group equivalence** by testing each of several baseline characteristics for differences between groups (hoping to find none)
- Performing **multiple pair-wise comparisons**, which occurs when three or more groups of data are compared two at a time in separate analyses
- Testing **multiple endpoints** that are influenced by the same set of explanatory variables
- Performing **secondary analyses** of relationships observed during the study but not identified in the original study design
- Performing **subgroup analyses** not planned in the original study
- Performing **interim analyses of accumulating data** (one or more endpoints measured at several different times)
- **Comparing groups at multiple time points** with a series of individual group comparisons (repeated-measures procedures)

Adjusting for multiple comparisons is sometimes optional. However, readers need to know whether or not adjustments were made and, if so, what adjustments were involved.[16] The Bonferroni correction is a common adjustment, for example.

Multiple testing is often needed to explore new relationships among data; however, exploratory analyses should be reported as exploratory. Data dredging—performing *undisclosed* analyses to compute many *P* values to find *something* that is statistically significant (and, therefore, worth reporting)—is poor science.

Error #10: Confusing Statistical Significance with Biologic Importance

As described here, many researchers interpret a statistically significant *P* value as indicating a biologically important result.[17] In fact, *P* values have no biologic interpretation. The nature and size of the difference must be judged to determine biologic importance. Perhaps the best way to remember this most common of statistical errors, as well as to close this article, is with a quote from statistician John Yancy: "It has been said that a fellow with one leg frozen in ice and the other leg in boiling water is comfortable—on average."[18]

References

1. Glantz SA. Biostatistics: how to detect, correct and prevent errors in the medical literature. *Circulation.* 1980;61:1-7.
2. Mainland D. Chance and the blood count. *Can Med Assoc J.* 1934; June:656-658.
3. Schor S, Karten I. Statistical evaluation of medical journal manuscripts. *JAMA.* 1966;195:1145.
4. White SJ. Statistical errors in papers in the *Br J Psychiatry.* 1979;135:336-342.
5. Hemminki E. Quality of reports of clinical trials submitted by the drug industry to the Finnish and Swedish control authorities. *Eur J Clin Pharmacol.* 1981;19:157-165.
6. Gore SM, Jones G, Thompson SG. The *Lancet*'s statistical review process: areas for improvement by authors. *Lancet.* 1992;340:100-102.
7. George SL. Statistics in medical journals: a survey of current policies and proposals for editors. *Med Pediatric Oncol.* 1985;13:109-112.
8. Altman DG, Schulz KF, Moher D, et al, for the CONSORT Group. The CONSORT statement: revised recommendations for improving the quality of parallel-group randomized trials. *Ann Intern Med.* 2001;134:657-62; *Lancet.* 2001;357:1191-1194; *JAMA.* 2001;285:1987-1991.
9. Stroup D, Berlin J, Morton S, et al. Meta-analysis of observational studies in epidemiology [MOOSE]: a proposal for reporting. *JAMA.* 2000;283:2008-2012.
10. Moher D, Cook DJ, Eastwood S, et al, for the QUORUM Group. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUORUM statement. *Lancet.* 1999;354:1896-1900.

11. Lang T, Secic M. *How to Report Statistics in Medicine: Annotated Guidelines for Authors, Editors, and Reviewers.* Philadelphia: American College of Physicians; 1997.

12. Mosteller F, Gilbert JP, McPeek B. Reporting standards and research strategies for controlled trials. *Control Clin Trials.* 1980;1:37-58.

13. Bailar JC, Mosteller F. Guidelines for statistical reporting in articles for medical journals. *Ann Intern Med.* 1988;108:266-273.

14. DerSimonian R, Charette LJ, McPeek B, Mosteller F. Reporting on methods in clinical trails. *N Engl J Med.* 1982;306:1332-1337.

15. Gotzsche PC. Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal antiinflammatory drugs in rheumatoid arthritis. *Control Clin Trials.* 1989;10:31-56.

16. Pocock SJ, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials: a survey of three medical journals. *N Engl J Med.* 1987;317:426-432.

17. Ellenbaas RM, Ellenbaas JK, Cuddy PG. Evaluating the medical literature, part II: statistical analysis. *Ann Emerg Med.* 1983;12:610-620.

18. Yancy JM. Ten rules for reading clinical research reports. *Am J Surg.* 1990;159:553-559.