COMMON STATISTICAL ERRORS EVEN YOU CAN FIND*
PART 2: ERRORS IN MULTIVARIATE ANALYSES AND IN INTERPRETING DIFFERENCES BETWEEN GROUPS

Tom Lang, MA
Tom Lang Communications

This article is the second in a series in which I describe several of the more common statistical errors in the biomedical literature. The first article in the series focused on 10 errors in descriptive statistics and in interpreting probability, or *P* values.[1] Here, I provide an overview of multivariate analyses (regression analysis and analysis of variance, or ANOVA) and describe nine errors in interpreting differences between groups.

An Overview of Multivariate Analyses

The most common forms of multivariate analyses in medicine are regression analysis and ANOVA. The two methods are similar. Both are used in studies involving two or more explanatory variables. In general, ANOVA is used to assess *categorical* explanatory variables, whereas regression analysis is used to assess *continuous* explanatory variables. When a study includes both categorical and continuous and explanatory variables, the analysis may be called either multiple regression analysis or analysis of covariance (ANCOVA). The results of multivariate procedures are referred to as models (equations), because they seek to describe the mathematical relationships among the variables so that one value can be predicted from the others.

The most common types of multiple regression analysis are the following:

- **Linear regression**, in which two or more explanatory variables are used to predict the value of a continuous response variable
- **Logistic regression**, in which two or more explanatory variables are used to predict the value of a binomial response variable (alive or dead, healed or not healed)
- **Cox proportional hazards regression**, in which two or more explanatory variables are used to predict the time to an event (such as the time from surgery to death)

The most common ANOVA procedures are one-way ANOVA, two-way ANOVA, multi-way ANOVA, ANCOVA, and repeated-measures ANOVA.[2] Unfortunately, these procedures take more space to explain.

- **One-way ANOVA** assesses the effect of *a single categorical explanatory variable* (sometimes called a factor) on a single continuous response variable. The factor (category) also has three or more alternatives (or levels or values; for example, the category of blood type has four alternatives: A, B, AB, or O). When there are only two alternatives (two groups), this analysis reduces to the Student *t* test.

  Example: Women with osteoporosis have been randomly assigned to one of three groups: a standard treatment, a new treatment, or a placebo. The response variable is the change in bone mineral density, a continuous variable. The explanatory variable is the form of treatment, which distinguishes each group. The results can be analyzed with one-way ANOVA.
- **Two-way ANOVA** assesses the effect of *two categorical explanatory variables* (again, sometimes called factors) on a single continuous response variable.

  Example: Suppose age was included in the previous example as a second explanatory variable. Age is coded as one of four ordinal categories: 30 to 40 years old, 41 to 50 years old, 51 to 60 years old, and 61 years old or more. With two categorical variables, treatment (or group) and age, the data can be analyzed with two-way ANOVA.
- **Multi-way ANOVA** assesses the effect of *three or more categorical explanatory variables* (still called factors) on a single continuous response variable.

  Example: To the previous example, the addition of more categorical explanatory variables, such as diet

*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.

(vegetarian or nonvegetarian) and alcohol consumption (less than 2 ounces of alcohol per day, 2 to 5 ounces per day, or 6 ounces or more per day), would move the analysis from two-way to four-way ANOVA, or simply, multi-way ANOVA.

- **ANCOVA** assesses the effect of one or more categorical explanatory variables *while controlling for the effects of some other (possibly continuous) explanatory variables* (now called covariates) on a single continuous response variable.

  Example: To the previous example, we now may wish to control for the severity of disease. Women with more severe osteoporosis may have different bone mineral densities than women with less severe disease. If we are to study the relationship between treatment and age on bone mineral density, we must control for disease severity. We thus add another (categorical) explanatory variable, disease severity (mild, moderate, and severe). The analysis is now called analysis of covariance (ANCOVA).

- **Repeated-measures ANOVA** is used to assess several paired, or repeated, measurements of the same subjects under different conditions (such as blood pressure measurements taken while the patient is supine, sitting, and standing) or at different points over time (such as muscle strength measured 1, 5, 10, and 20 days after surgery).

  Example: Again, building on the previous example, suppose we have measurements of bone mineral density for all patients at the onset of symptoms and at 6 and 12 months after the onset of symptoms. Time can now be added to the ANOVA model as an explanatory variable. Here, time is a repeated measure; although each woman belongs to a single treatment group and to a single age category, each has bone density measurements at three points in time (0, 6, and 12 months).

Error #11. Not Confirming That the Data Met the Assumptions of ANOVA

ANOVA assumes that the response variable is approximately normally distributed within each level of the explanatory variable and that the variability of these distributions is approximately the same. Because most biologic data are not normally distributed,[3-9] the data may need to be mathematically transformed into distributions that are more normally distributed. Alternatively, a nonparametric form of ANOVA can be used. For example, skewed data should probably be analyzed with the Wilcoxon rank-sum test, rather than with one-way ANOVA, and by the Kruskal-Wallis test, rather than with multi-way ANOVA. (The assumptions of regression analyses are mentioned in Error #9 in the first article of this series.)

Error #12. Not Identifying the Procedure Used to Adjust for Multiple Comparisons in ANOVA

ANOVA is a group comparison that determines whether a statistically significant difference occurs somewhere among the groups studied. If a significant difference occurs, ANOVA is followed by a multiple comparison procedure that compares combinations of groups to determine which groups differ statistically. Common multiple comparison procedures include Tukey's procedure, Student-Neuman-Keuls procedure, Scheffe's method, and Fisher's least-significant method; there are many others.

Error #13. Not Testing the Explanatory Variables for Interaction or Colinearity

Two explanatory variables are said to interact if the effect of one of the response variables depends on the level of the other. For example, alcohol and barbiturates can interact to cause death, even if the amounts of each—by themselves—are not lethal. Interaction implies that the factors should be considered together, not separately. Thus, an analysis of the causes of death from drug overdose would have one factor for blood alcohol level, one for blood barbiturate level, and an interaction term that represents the fact that the effect of alcohol on death depends in part on barbiturate level.

Two variables are said to be colinear if they are highly associated and therefore provide the same information in the model. Systolic and diastolic blood pressure, for example, may contribute such similar information to the model that only one need be used. Testing for interaction and colinearity is usually necessary only in large studies with several explanatory variables.

Error #14. Not Indicating the Goodness-of-Fit of the Model to the Data

Goodness-of-fit indicates how well the model expresses the relationships observed in the data. Examining the residuals (the differences between the observed values and those estimated by the model) helps to determine the fit of the model. The results of the analysis of residuals need not be reported; a statement that the residuals were examined and that the model did (or did not) appropriately fit the data will suffice.

In multiple regression analysis (not ANOVA), the value of R^2 should be reported. This value indicates how much of the variation in the response variable is explained by the factors included in the model. Thus, the higher, the better.

Error #15. Not Reporting Whether and How the Model Was Validated

Multivariate models can be validated or tested against a similar set of data to show that they explain what they seek to explain. One method of validation, used with large samples, is to develop the model on, say, 70% of the data and to compare it with another model based on the remaining 30%. Another method involves removing the data from one subject at a time and recalculating the model. The coefficients and the predictive validity of all the models (there may be hundreds) can then be compared. Such methods are called jackknife procedures. A third method involves developing a new model on a new set of comparable data to determine whether the results are similar.

Errors in Interpreting Differences Between Groups

The majority of biomedical research studies are interested in *differences*, either in one or more groups over time or between two or more groups at the same time. Differences are of interest, for example, when they indicate that one intervention might be more effective than another. Differences can be presented in several forms, however, some of which can be misleading. Here, I describe some of the more common forms, how they can be misinterpreted, and what additional information is needed to prevent these misinterpretations.

Error #16. Not Reporting Confidence Intervals with Estimates

When interpreting any difference, whether it is statistically significant or not, the *direction* and *magnitude* of the difference should be evaluated for its clinical importance. However, because a study is based on a *sample* of the population of interest, rather than on a *census* of the population, its results are actually *estimates* of the differences expected if the study were to be repeated on the entire population. Thus, another factor that should be considered when evaluating differences is the *precision of the estimate.*

In clinical research, the most common measure of precision for an estimate is the 95% confidence interval. In the following example,[2] evaluating only the estimated size of the difference can be misleading. For this reason, journals now 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.[10]

"*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 pretest and posttest values 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).*" In essence, 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 fall between 2 and 34 mm Hg in 95 of the 100 samples. *(See Letter to the Editor and response on page 135.)* 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, these 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 its values are clinically important.

Error #17. Reporting Only Relative Differences and Not Absolute Ones

The absolute difference between groups is simply the *mathematical difference between their values*, whereas the relative difference is the *absolute difference expressed as a percentage.* By themselves, relative differences can mislead because they can make differences appear to be larger or smaller than they really are.[11] For example, a 50% survival rate could mean that two of four patients survived or that 2,000 of 4,000 survived. The absolute difference in survival is two in the smaller study and 2,000 in the larger one. Thus, although both studies show the same relative difference, the absolute difference of the first study is probably too small to justify meaningful conclusions.

In a scientific article, the numerators and denominators should be apparent for all percentages so that the absolute differences can be determined.[2] This need is especially important when the numbers are less than 100, because the percentages are larger than the actual numbers they represent. "A third of the rats lived, 33% died, and the last one got away." Here, 33% is one of three rats. In the following, more serious example,[12] readers given the absolute difference usually judge the drug to be far less effective than do readers given the relative difference. "In the Helsinki study of hypercholesterolemic men, after 5 years, 84 of 2030 patients on placebo (4.1%) had heart attacks, whereas only 56 of 2051 men treated

with gemfibrozil (2.7%) had heart attacks ($P < 0.02$)" Here, the **absolute difference** (and therefore, the "absolute risk reduction" in heart attack) was **1.4%**; that is, the difference between the frequency of heart attacks in the two groups was 1.4% (4.1% - 2.7% = 1.4%). However, the **relative difference** (and therefore, the "relative risk reduction" in heart attack) was **34%**; that is, 1.4 is 34% of the 4.1% of men in the control group who had heart attacks (1.4%/4.1% = 34%).

Error #18. Not Differentiating Between Unit of Observation and the Number of Patients Improved

The unit of observation or the unit of analysis is what is being studied. In clinical research, the unit of observation is usually the patient. However, sometimes the unit is something other than the patient. The problem comes when, say, differences are reported for the unit of observation but not for the number of patients in whom differences occurred. For example, if a drug markedly improves mean glomerular filtration rate in patients with renal disease, it may also be helpful to know how many patients actually improved.

This issue can be illustrated with a simple example (Figure 1), in which the results can be reported as a mean *decrease* from time 1 to time 2 or as an *increase* in two of three (66%!) patients. Both results are technically correct, but reporting only one can be misleading because the mean change is the result of an unusual response in a single patient.

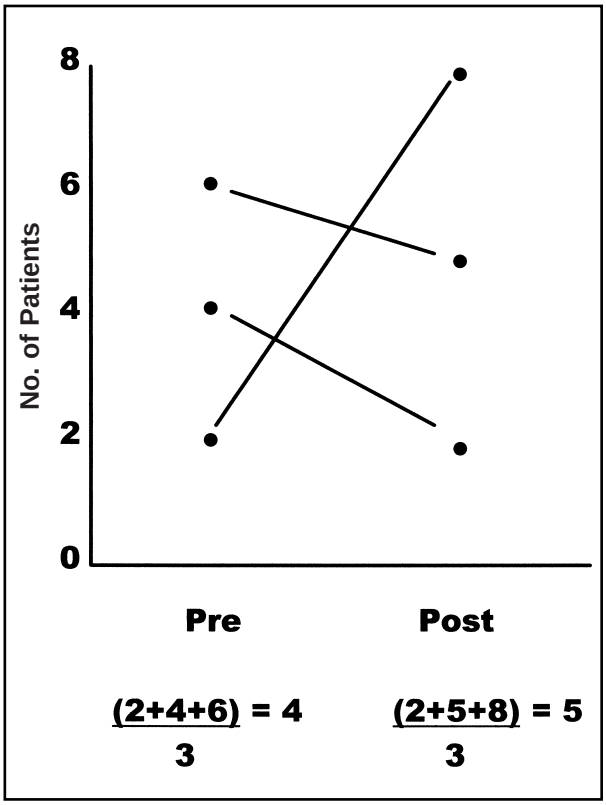

Figure 1

Error #19. Confusing Post-hoc Analyses with Planned Analyses

Post-hoc analyses are analyses performed after investigators have reviewed the study data; that is, post-hoc analyses are *exploratory analyses* suggested by the data and are not planned in advance of data collection. Exploratory analyses are necessary to make the most of the data collection effort. The problem comes when these analyses are presented as planned, primary analyses, rather than as exploratory analyses. Differences detected by post-hoc analyses should be evaluated more critically than differences detected by the planned analyses.

The number of exploratory analyses can sometimes be large. As mentioned in Error #9,[1] generating multiple *P* values greatly increases the chance of finding a significant *P* value *somewhere* in the data. Exploratory analyses are thus sometimes called data dredging or "fishing expeditions" when the real search is for *any* significant *P* value rather than meaningful differences in the data. "Hypothesis-generating studies (sometimes referred to as 'fishing expeditions') should be identified as such. If the fishing expedition catches a boot, the fishermen should throw it back, not claim that they were fishing for boots."[13]

References

1. Common Statistical errors even you can find. Part 1: errors in descriptive statistics and in interpreting probability values. *AMWA J.* 2003;18:67-71.
2. Lang T, Secic M. *How to Report Statistics in Medicine: Annotated Guidelines for Authors, Editors, and Reviewers.* Philadelphia, Pa: American College of Physicians, 1997.
3. Haines SJ. Six statistical suggestions for surgeons. *Neurosurgery.* 1981;9:414-418.
4. Griner PF, Mayewski RJ, Mushlin AI, Greenland P. Selection and interpretation of diagnostic tests and procedures. *Ann Intern Med.* 1981;94(4 part 2):557-592.

5. Evans M, Pollock AV. Trials on trial. A review of trials of antibiotic prophylaxis. *Arch Surg.* 1984;119:109-113.

6. Feinstein AR. X and iprP: an improved summary for scientific communication [editorial]. *J Chron Dis.* 1987;40:283-288.

7. Hall JC, Hill D, Watts JM. Misuse of statistical methods in the Australasian surgical literature. *Aust NZ J Surg.* 1982;52:541-543.

8. Hall JC. The other side of statistical significance: a review of type II errors in the Australian medical literature. *Aust NZ J Med.* 1982;12:7-9.

9. Wulff HR, Andersen B, Brandenhoff P, Guttler F. What do doctors know about statistics? *Stat Med.* 1987;6:3-10.

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

11. Guyatt GH, Sackett DL, Cook DJ. Users' guide to the medical literature. II. How to use an article about therapy or prevention. B. What were the results and will they help me in caring for my patients? *JAMA.* 1994;271:59-63.

12. Brett AS. Treating hypercholesterolemia: how should practicing physicians interpret the published data for patients? *N Engl J Med.* 1989:321;676-680.

13. Mills JL. Data torturing [letter]. *N Engl J Med.* 1993;329:1196-1199.