7 Statistical Issues that Researchers Shouldn't Worry (So Much) About

By Karen Grace-Martin
Founder & President

THE ANALYSIS
FACTOR

About the Author

Karen Grace-Martin is the founder and president of The *Analysis Factor*. She is a professional statistical consultant with Masters degrees in both applied statistics and social psychology. Her career started in psychology research, where her frustration in applying statistics to her data led her to take more and more statistics classes. She soon realized that her favorite part of research was data analysis, and she's never looked back. Her background in experimental research and working with real data has been invaluable in understanding the challenges that researchers face in using statistics and has spurred her passion for deciphering statistics for academic researchers.

Karen was a professional statistical consultant at Cornell University for seven years before founding The Analysis Factor. Karen has worked with clients from undergraduate honors students on their first research project to tenured Ivy League professors, as well as non-profits and businesses. Her ability to understand what researchers need and to explain technical information at the level of the researcher's understanding has been one of her strongest assets as a consultant. She treats all clients with respect, and derives genuine satisfaction from the relief she hears in their voices when they realize that someone can help them.

Before consulting, Karen taught statistics courses for economics, psychology, and sociology majors at the University of California, Santa Barbara and Santa Barbara City College. Karen has also developed and presented many statistics workshops, most recently on missing data, logistic regression, and interpreting regression parameters.

Karen has co-written an introductory statistics textbook with sociologist Stephen Sweet: *Data Analysis with SPSS*. It focuses on statistical concepts and data analysis practices, without the endless calculations that often obscure them.

Introduction

I have devoted my entire professional life to excellence in research and data analysis in academic settings. My passion is in helping academic researchers learn, apply, and practice statistics.

In my work with thousands of researchers over the past 9 years, I have seen these issues come up again and again. While some of them are legitimate concerns in very limited situations, most are not.

I'm sure some come from the emphasis statistics professors have made on watching out for those legitimate concerns. Others come from misinterpretations or misremembering on the part of students. This ebook will help you separate the legitimate statistical concerns from the ones that you shouldn't worry (so much) about!

Unique Vantage Point

In my experience, academic researchers are usually not statistics experts. Nor should they be—they are experts in their own field.

One advantage of my position as a statistical consultant is I have the vantage point of seeing the different analysis conventions in different fields. Researchers never have this perspective.

Another is I see the same statistical issues come up for many researchers. As a researcher, each statistical issue is new the first time you experience it. Researchers just don't have the same access to statistical problems that an experienced consultant does.

My goal with this ebook is to share my vantage point to clear up some common statistical misconceptions I see many researchers have.

Acknowledgments

I would like to thank Rebecca Smyth and John Mordigal for their very helpful comments on an earlier draft.

Issue #1: Making the analysis complicated

T-tests and correlations are absolutely fabulous when they are really what you need, despite what your committee says. Your analysis is not there to impress or confuse. It is there to test your hypothesis. Simpler is always better. Some hypotheses can be tested different ways—always use the simplest one that gives you the information you need.

Issue #2: .05 as a cutoff

> "Statistical analysis is a tool to be used in helping you find the answer...it's not the answer itself"
>
> - John Mordigal

I realize I'm walking on sacred ground here, but the p-value is a probability—an indication of how likely you're falsely claiming an effect. .04 and .06 are really, really close. They tell you pretty much the same thing about the likelihood of a Type I error.

Yes, you set .05 as an upper limit of the risk you're willing to take that you're wrong about your claim of an effect. But it's not reasonable to make assertions of effects with a p-value of .03 without looking at the size of the effect. Would the size of the effect make any difference in the real world?

Every effect is significant in giant surveys with tens of thousands of participants. Remember that statistical significance does not indicate scientific importance.

Issue #3: Retrospective Power Analysis

There really isn't any point in calculating power after the analysis has been found insignificant. I know you want to claim that there really is an effect, but you didn't have enough power.

At best, it doesn't add any new information. That's clear from the p-value. At worst, it's significance fishing. You might get away with it, but it isn't good practice.

For more information, see: Lenth, R. V. (2000), "Two Sample-Size Practices that I don't Recommend,'' presented at the *Joint Statistical Meetings*, available at : http://www.cs.uiowa.edu/~rlenth/Power/2badHabits.pdf

Issue #4: Unequal Sample Sizes

In your statistics class, your professor made a big deal about unequal sample sizes in one-way Analysis of Variance (ANOVA) because you had to calculate everything by hand. Sums of squares require a different formula if sample sizes are unequal, but SPSS (and other statistical software) will automatically use the right formula.

The only issue in one-way ANOVA is that *very* unequal sample sizes can increase the risk of violating the homogeneity of variance assumption—the assumption that the population variance of every group is equal. ANOVA *is* considered robust to moderate departures from this assumption, but the departure needs to stay smaller when the sample sizes are very different.

Real issues with unequal sample sizes *do* occur in *factorial* ANOVA, if the sample sizes are confounded in the two (or more) factors. For example, in a two-way ANOVA, let's say that your two independent variables (factors) are age (young vs. old) and marital status (married vs. not). If there are twice as many young people as old and the young group has a much larger percentage of singles than the older group, the effect of marital status cannot be distinguished from the effect of age.

Power, the ability to reject an incorrect null hypothesis, is based on the smallest sample size, so while it doesn't hurt power to have more observations in the larger group, it doesn't help either.

Issue #5: Multicollinearity

Multicollinearity occurs when two or more predictor variables in a regression model are redundant. It is a real problem, and it *can* do terrible things to your results. However, the dangers of multicollinearity seem to have been so drummed into students' minds that it has created a panic.

Unless you have a very controlled, designed experiment, your study will have *some* degree of multicollinearity. This is common, normal, and expected.

It is not a problem, but it does affect the interpretation of model parameter estimates (compared to a model where all predictors are independent). Researchers need to keep the associations among predictors in mind as they interpret model parameter estimates (regression coefficients or mean differences). Each regression coefficient represents the effect of its predictor on the response, *above and beyond* the effect of all other predictors in the model.

Multicolllinearity becomes a problem if two or more of your variables are measuring *exactly* the same thing. This is *quite rare*. High correlations among predictor variables may suggest severe multicollinearity, but it is NOT a reliable indicator that it exists.

The real problem with severe multicollearity is that redundant information in predictors hugely inflates the variance of parameter estimates (regression coefficients, group means, etc.). This means that standard errors become enormous and t values end up at 0, making everything insignificant. The best way to check for severe multicollinearity is using condition indices. It is easily done with the 'collinearity diagnostics' option in SPSS regression analysis.

A very nice article that explains in more detail how to diagnose multicollinearity and use condition indices is at http://cscu.cornell.edu/news/statnews/stnews65.pdf.

Issue #6: The Distribution of Independent Variables

There are NO assumptions in any linear model about the distribution of the *independent* variables. Independent variables can be skewed, continuous, count, bimodal, categorical, ordinal—anything.

Yes, you only get meaningful parameter estimates from nominal (unordered categories) or numerical (continuous or discrete) independent variables. So ordinal predictors need to be either reduced to unordered categories or assumed to be numerical.

But no, the model makes no assumptions about their distribution.

It is useful, however, to understand the distribution of independent variables to find influential outliers or concentrated values. It is always good practice to run univarite descriptives and graphs about any variable you plan to use in a regression model.

Issue #7: The Distribution of Dependent Variables

> "To set up interval estimates and make tests, however, we need to make an assumption about the form of the distribution of the ε_i. The standard assumption is that the error terms ε_i are normally distributed."
>
> - Neter, Kutner, Nachtsheim, & Wasserman, 1996, p. 29

The assumptions of normality and homogeneity of variance for linear models concerns the *residuals* (the error terms), NOT the dependent variable. The normality of the residuals "*implies* that the Y_i are independent normal random variables" (Neter, Kutner, Nachtsheim, & Wasserman, 1996), but it is not always true because the Y_i (the dependent variable) are affected by the X_i (the independent variables).

The distribution of the dependent variable can tell you what the distribution of the residuals is *not*—you just can't get normal residuals from a binary dependent variable. But it cannot always tell what the distribution of the residuals *is*.

For example, even if the residuals are normally distributed, a binary categorical independent variable with a big effect will result in a dependent variable made up of two side-by-side normal distributions (one for each level of the independent variable). This would make the dependent variable a continuous, bimodal distribution. A look at only the distribution of the dependent variable would lead a researcher to believe that a linear model is inappropriate, when, in fact, it is.

Neter, Kutner, Nachtsheim, & Wasserman's *Applied Linear Regression Models*, and its expansion, *Applied Linear Statistical Models,* is an excellent resource about linear models, including assumptions. Any edition of the book is a good resource, and the order of the authors often changes from one edition to another. Non-current editions are reasonably priced, and make an excellent addition to statistical resource libraries.