Overcoming the Fear of Statistics: Survival Skills for Researchers

Karen B. Williams PhD, RDH

Introduction

One of the most common complaints I hear from clinician-researchers is that statistics are difficult to understand and apply. Misstatements such as "differences were highly significant, with p=0.008" or "our study proved X causes Y" reinforce common misperceptions associated with statistics. These statements illustrate 2 common fallacies. The first is that smaller p values can be interpreted as a larger effect, and, that a small p value is evidence of "truth." In order to understand why these assumptions are fallacies, it is important to know what the p value represents.

The accepted convention for separating potential explanations (X causes Y) from chance happenings is testing the null hypothesis. One can think of testing the null hypothesis as a "ritualized exercise of devil's advocacy."[1] The null hypothesis is an artificial argument – that any difference between treatment groups is due to chance, assuming that the treatment has no effect. Researchers hope that this likelihood is small. The p value derived from statistical testing is an estimate – the probability that, assuming the intervention is not effective, that treatment groups are different simply due to chance variation. If a small p value (conventionally <0.05) is obtained, then the researcher rejects the assumption of difference due to chance and accepts the alternative - that differences are likely due to the treatment.

Groups can differ simply due to chance. Two common sources that contribute to this are sampling error and measurement error. Sampling error occurs when groups are inherently different by chance. Random assignment can reduce this error, but does not ensure group equivalence with respect to all factors that might influence the outcome. Measurement error can exist depending on how, when, where and by whom outcomes are measured. Either source of error can introduce doubt as to whether change in the outcome (Y) is solely attributable to the intervention (X). Thus, it is not possible to prove causality. We can, however, estimate the probability (p) that observed differences between groups are based on "chance" using the null hypothesis.

Getting significant differences (p<0.05) is influenced by three factors: magnitude of effect, sample size, and variation in the data. Because sample size influences p value, a small p cannot be simply equated with large effect size. Results from a study with 1,000 subjects will always have a much smaller p value than one with 100 subjects, given the same magnitude of difference between groups. Power of a statistical test - the likelihood of rejecting the null hypothesis when there is a real difference - is influenced by the number of observations/sample size.

Effect size is about actual differences. It can be determined from raw data (e.g., difference between group means) or standardized (raw effect size divided by the standard deviation). It is helpful for researchers to think about raw effect size as the minimally important difference (MID), that is clinically meaningful. The standardized effect size, which takes into account the variance, can be interpreted as a measure of "importance". Thus, it gives an objective estimate of the strength of association between the outcome and intervention/treatment. Common effect size measures include r^2, eta square, odds ratio and Cohen's d.

Statistical Decision Making

So, why do clinicians often equate a statistically significant p value with truth about causality? Humans innately have a need for certainty. When individuals feel uncertain or there are multiple cues that need to be considered simultaneously, individuals often rely on one-dimensional rule-based decision making.[2] Such is the case with statistical analysis and interpretation.[3] Several researchers have criticized this "fantasy" of statistical testing as proving effectiveness, and have called for logical interpretation

of data along with use of the p value, effect size estimate and replication of findings.[4,5]

CONSORT Guidelines and Improved CONSORT guidelines now encourage researchers to provide information about the MID when publishing. They also suggest that MID be defined in advance and used as the effect size for designing clinical trials.[6] Despite changes in publication standards and improved statistical techniques available, clinicians and researchers still tend to fear statistics and make rash judgments about the meaningfulness of statistics.

Therefore, the remainder of this paper will discuss issues that may help demystify statistical testing and provide clinician-researchers with realistic strategies for improving the quality of one's own research efforts.

The Logic of Establishing Causality

Establishing the causality between an intervention and outcome requires that 5 tenets be met. First, there must be a logical or biologically plausible relationship between the cause and the outcome. Second, exposure to the cause must precede development of the outcome. Third, there has to be evidence of strength of association. Fourth, and critically relevant to both proper design and statistical testing, is that there has to be a lack of competing explanations for the results. Last, evidence must be replicated. A single study does not provide sufficient evidence to support causality.

Study design is critical to making causal statements. Having a comparison group (or better yet, a control group if possible) is necessary to tease apart whether any observed changes are attributable to the treatment/intervention. While the statistical test (and associated p value) can give us an estimate of chance differences, it alone is insufficient. One must consider why treatment versus comparison groups might (or might not) differ. Some common reasons include:

- Individuals in the respective groups looked similar but differed in subtle ways that were undetectable but important.
- Changes observed over time could be natural occurrences (e.g. aphthous ulcers and healing)
- Measurement was flawed or unequally implemented
- Study length was insufficient to capture impact over time
- Not all subjects were available for all observation periods or differentially dropped from the study (missing data).
- There were too few subjects to capture a difference if it existed or there were so many subjects that even a trivial difference would be statistically significant.

Statistical Tests as Part of a Logical Argument

One of the most compelling books in print today is Statistics as Principled Argument.[1] Abelson argues for use of applied logic and good judgment along with hypothesis testing to make good decisions about study results. Psychologists have shown that people are highly susceptible to confirmation bias. Confirmation bias results when people selectively focus on information that reinforces preexisting ideas, thus resulting in overestimating the influence of systematic factors (like an imposed treatment) and underestimating influence of alternative explanations, including chance. This may cause individuals to conclude that an intervention is effective, especially if there is a p value from a statistical test of <0.05, without thoughtful consideration of other factors.

Since very few clinical researchers have the depth of understanding that underlies the field of methods and biostatistics, they are likely unaware of how a conceptual model, study design and measurement can be used to their maximal benefit to answer meaningful research questions. Actively seeking out a consultation with a biostatistician with experience in the broad field of health-related research is one of the most effective ways to overcome a fear of statistics.

Getting a Statistical Consult

Obtaining a statistical consult during the design phase of a study is one of the best ways to maximize efficiency in the research process. Many institutions have statistical consultation services or individuals who can provide these consults. Find someone at your institution who is knowledgeable with whom you can discuss your project.

Once identified, prepare for the consultation in advance. Be prepared for the questions that the statistician might ask about previous research. In the literature, be attentive to how results may have changed over time. An inter-

esting observation about study results is that effects often decrease over time. Lehrer suggests that "truth wears off" over time because our illusions about the meaningfulness of various research questions declines over time.[7] Being able to articulate this trend will be important for study design and power analysis. Getting the right estimate for sample size initially improves the likelihood of getting meaningful results.

In advance, draft an abstract that summarizes your proposed project using the PICO format.[8]

(P) Population: Who is the population being studied?
(I) Intervention: What is the intervention or exposure variable?
(C) Comparison or Control Group: What is the most appropriate comparison or control group?
(O) Primary Outcome Measure: What outcomes are feasible to measure?

A good consultation will usually result in modifying some aspects of your original research plan. So, be prepared to capture recommendations either in writing or audio recording. Clarify issues that are confusing at that time. A good consultant will help identify potential confounding variables that should be controlled either by design or statistically. Make sure you leave with an understanding of how design, measurement and statistical analysis fit together. Once you have drafted your proposal, get confirmation from the consultant that you have "gotten it right."

Make sure you discuss how to set up your data for analysis. The statistical analysis plan, design of the study, capture of confounders, number and type of outcome measures, and statistical software will dictate the appropriate format. Unless you are completely comfortable with statistical software and the analysis plan, do not do this on your own. There is nothing more frustrating than to have all of your data entered, only to find it is not in an analyzable format.

Conclusion

Most importantly, leave your apprehension at the door and look at the consultation as a unique opportunity to engage in creative planning. Statistics are wonderful tools, but only if used correctly. Statistical analysis programs manage the computational aspects but do not overcome bad design and incorrect analyses. Approach the research process just as you would plan a trip to a foreign country and you can avert the fear of statistics and pain of failure.

References

1. Abelson RP. Statistics as Principled Argument. New York, Taylor and Francis Group, Psychology Press. 1995.

2. Erick J, Boomer J, Smith J, Ashby F. Information-integration category learning and the human uncertainty response. *Mem Cogn*. 39: 536-554, 2011.

3. West CP, Ficalora RD. Clinician attitudes towards biostatistics. *Mayo Clinic Proceedings*. 82(8); 939-943, 2007.

4. Carver RP. A case against statistical significance testing. *Harvard Educ Review*. 48(3): 378-399, 1978.

5. Carver RP. A case against statistical significance testing, revisited. *J Experimental Educ*. 61(4): 287-292, 1993.

6. Zwarenstein M, Treweek S, Gagnier J, Altmam DG, Tunis S, Haynes B, Oxman AD, Moher D. Improving the reporting of pragmatic trials: an extension of the CONSORT statement. *BMJ* 337: 1-8, 2008.

7. Lehrer J. The truth wears off: Is there something wrong with the scientific method? New York Times. Available at: http://www.newyorker.com/reporting/2010/12/13/101213fa_fact_lehrer Accessed August 2011.

8. Duke University. Asking a Well-Built Clinical Question. Introduction to Evidence-based Practice. Available at: http://guides.mclibrary.duke.edu/content.php?pid=431451&sid=3529524 Accessed August 20, 2014.