International Journal of Aquatic Research and Education

Volume 5 | Number 2 Article 2

5-1-2011

# Statistical Inference and Evidence-Based Science

Stephen J. Langendorfer
*Bowling Green State University*, slangen@bgsu.edu



## Recommended Citation

Langendorfer, Stephen J. (2011) "Statistical Inference and Evidence-Based Science," *International Journal of Aquatic Research and Education*: Vol. 5: No. 2, Article 2.
DOI: https://doi.org/10.25035/ijare.05.02.02
Available at: https://scholarworks.bgsu.edu/ijare/vol5/iss2/2

*International Journal of Aquatic Research and Education,* 2011, 5, 143-146


# Statistical Inference and Evidence-Based Science

I presume that many readers may have heard some variation of the quote, attributed to British Prime Minister Benjamin Disreali and popularized by American humorist, Mark Twain (a.k.a., Samuel Clemens), when referring to confusion generated by the use and misuse of quantitative figures. "There are three kinds of mistruth: lies, damned lies, and statistics" (Twain, 1906).

One of the "rites of passage" associated with obtaining a graduate degree is being required to complete multiple statistics classes. I try to sympathize with my current students when I reflect on how little I could recall after finishing my first course in tests and measurements as an undergraduate. During my Masters program at Purdue University, I gained a completely undeserved reputation for being a "statistics whiz," bestowed upon me by my fellow student and oft co-conspirator, Larry Bruya (who fulfills my personal definition of a "true friend" wherein a "friend" is said to be one who will bail you out of jail, while a "true friend" is one who sits in the jail cell with you and proclaims, "Golly, that was fun!"). Larry and I took the same first-level statistics class together at Purdue and in the evenings while studying, he would quiz me about what each day's topic meant. I was too dumb to realize that Larry wasn't asking rhetorical questions to challenge me, but that he really didn't know the answers. I figured I didn't want to appear stupid, so I started concocting answers and in the process figured out how to actively learn statistics! Thanks, Larry. There is a sequel to this story decades later. Whenever I make some kind of pronouncement in his presence, Larry has learned to inquire, "Do you really know the answer or are you just making that up?!" Such an inquiry never fails to result in gales of laughter while Larry explains to whoever is gathered our personal story about what he affectionately calls "making up crap about statistics."

An important realization to come from any discussion about statistics, with or without any notion of lying or even just "making up crap," is that comprehending statistics can legitimately be quite confusing, even to those with some basic knowledge. They can be utterly mystifying to those without a degree of quantitative literacy in probability, laws of chance, and elementary statistical procedures. Worse, when statistics have been misused (say it ain't so!) simply to support one's preconceived opinion, all trust in them can go right out the window so that the validity of all statistics becomes suspect.

## Statistics as Tools

There are all sorts of numbers that pass for statistics, rightly or wrongly: Individual scores, percentages, percentiles, standard scores, means, medians, modes, ranges, standard deviations, variances, T scores, t tests, analysis of variance, correlation, multivariate analysis of variance, analysis of covariance, and multiple regression, *ad infinitum*. An important first realization is that any statistic is merely a tool like a hammer, screwdriver, shovel, or rake. Like any tool, a statistic can be both used and misused. Applying the appropriate statistic to the right research question is critical in scientific study. That necessitates fundamental knowledge about many kinds of statistics. As the old saying goes, "If your only tool is a hammer, everything begins to look like a nail!" I cannot possibly and do not intend to overview everything that is important about statistics in these several editorial pages. I am addressing several notions that do apply to the scientific rigor associated with research papers submitted to and published by this particular scholarly journal. For more sophisticated understanding of statistics, I encourage readers to seek out various print and online statistical sources including the *Publication Manual of the American Psychological Association* (6th ed., 2010). Chapter 5, "Displaying Results" (pp. 125-167) presents an important summary of how to analyze and present the statistical results of many different kinds of study.

## Descriptive Vs. Inferential Statistics

It may be helpful to consider that all statistics (remember that does not mean all numbers) can be categorized into one of two groups: *descriptive statistics* and *inferential statistics*. I realize that there are also parametric and nonparametric statistics, subcategories of inferential statistics, and I am certain there are other ways to categorize different statistics. For my discussion purposes, I think these two groupings will suffice.

***Descriptive Statistics.*** For most individuals the most straightforward way to summarize or capture the essence of a group of numbers is by using *descriptive statistics*. Examples of descriptive statistics include measures of central tendency (i.e., mean, median, mode) and measures of variability (i.e., range, standard deviation, variance, standard errors) along with simple percentages, percentiles, standard scores, and correlations. The purposes of descriptive statistics are to summarize for a reader the score characteristics of a group or sample of numbers as well as to provide some insight into the scores achieved by individuals within that group. True to their name, they simply describe and summarize the group of numbers: what the distribution of numbers looks like, where its middle score lies, how spread out the scores are, and whether scores are related or associated with other scores.

Importantly, individual descriptive statistics do not, by themselves, give one enough information to generalize those numbers or scores to other groups beyond the immediate sample. They also can be misleading when used in isolation or inappropriately. For example, knowing the mean of a group of scores tells the reader only where the arithmetic average of those scores falls. If the sample of scores is skewed (i.e., asymmetric with outlying scores higher or lower than the mean), then the mean does not give a very appropriate picture of the middle of the group of numbers.

Importantly, the actual way that the numbers are being used, the so-called *measurement scale* associated with the four qualities of the real number system, makes a huge difference in how they should be summarized. I come across misunderstandings and misuses of this constantly. For example, when someone administers a Likert scale survey question, the resulting scores typically represent "ordinal scale" numbers, which only possess two of the four characteristics of a real number. Ordinarily, one should not calculate means and standard deviations or parametric statistics for rank ordered Likert scale data, especially because these data also may be skewed. Instead, for ordinal scale numbers, the median or mode and range of scores along with nonparametric inferential statistics should be employed.

It is customary and certainly appropriate to present measures of central tendency and variability in research studies, although both, not just one, should be used together to describe the characteristics of the group of scores obtained. At the same time, if descriptive statistics alone are calculated and presented, the results from the sample in a study *may not* (please note my emphasis!) be used to generalize the results to other situations or other samples. When one relies solely on descriptive statistics, conclusions may represent only W.A.S.G. ("wild ass scientific guesses"; Pia, personal communication), which may very well lead to the same estimates one would make without calculating any statistics whatsoever!

***Inferential Statistics.*** Generalizing results requires the application of *inferential statistics* such as *t*-tests, ANOVA, or multiple regression. As straightforward as descriptive statistics can be, inferential statistics can be complicated and complex, both to calculate and to understand. They draw upon the well established, but oft poorly understood, laws of probability and chance. If appropriately calculated and interpreted, they should allow someone to understand how well the results of one study may apply to the results of other studies and situations.

The cornerstone of inferential statistics lies in a solid understanding of probability that can answer this question: To what degree are the results under consideration trustworthy (i.e., valid or accurate and reliable or consistent and repeatable) and not just a chance occurrence? Fortunately, astute statisticians (e.g., Sir Ronald Fisher) have described score distributions that allow us to make appropriate judgments about the degree to which a result may be simply luck (i.e., a chance occurrence) or whether it is trustworthy and generalizable.

Modern inferential statistics can be a literal "Tower of Babel" because of their sophistication and complexity. Using powerful modern computing technology, new ways of calculating statistics are ever evolving. We can now calculate on our desktop or laptop computers in minutes what would have required me nights and nights of work on a mainframe computer when I was in graduate school. Regardless of the sophistication of the statistical software, there are some universal principles we can consider about how we know if we should generalize results.

***Errors in Statistical Conclusions.*** First and foremost, we must ask the Type I error question: To what degree do I have confidence that I am correctly choosing to accept a certain outcome and that outcome is in fact not a chance occurrence? I like to think of this first principle in terms akin to how the judicial system in the U.S. establishes criminal innocence or guilt: The accused person is judged to be innocent until proven guilty. Similarly, the inferential system assumes that a result is a chance occurrence (i.e., innocent) unless we can prove that results are highly unlikely (i.e., guilty). Most readers will recognize the alpha ($\alpha$) level of 0.05 as the traditional degree of chance that researchers require before assuming nonchance findings.

The second related inferential question is called the Type II error question: To what degree have I tested a sufficiently large sample and am using sensitive enough measures that I can identify a meaningful, nonchance difference? A recent requirement in scholarly journals that follow the American Psychological Association protocols is that measures of statistical power be tested and identified in a paper along with the alpha level. While we have not always required statistical power to be calculated in the past, that is one requirement that will be enforced in the future to allow us to remain in compliance with APA.

One other final minor controversy has arisen related to how we should detect and describe the existence of significant differences. My own traditional statistical training basically requires the investigator to compare a so-called ρ, or probability level, with the traditional $\alpha < 0.05$. More contemporary approaches are advocating the use of confidence intervals, or estimates of the range in which significantly different values may fall. In point of fact, both *p* values and confidence intervals provide exactly the same statistical inference information in slightly different ways. They both require an understanding of how to avoid both Type I and Type II errors.

In our ongoing mission to provide the very best and strongest aquatic literature, the *International Journal of Aquatic Research and Education* strives to publish the strongest scholarly research possible. This means that while we have and will occasionally publish descriptive work as appropriate, more and more we will be expecting authors to employ contemporary and appropriate inferential statistical procedures so that our readers can be assured that published findings in fact reflect the most rigorous scientific evidence and results.

## Welcome Kevin and Bob

As I mentioned last issue, the *International Journal of Aquatic Research and Education* would be welcoming new members to our Editorial Board as we increase our international diversity, enhance the expertise related to recent popular topics being published in the Journal, and not overburden members who have already served a term. Alert readers already may have noticed the changes to the Editorial Board membership indicated on the back cover of this issue. I hope Editorial Board members, readers, subscribers, authors, and reviewers of *IJARE* will join me in welcoming Dr. Kevin Moran, principal lecturer in education at the University of Auckland, New Zealand, and Dr. Robert Keig Stallman, emeritus professor from the Norwegian School of Sports Science, Oslo, Norway, as the newest members of the Editorial Board. Both of these new Board members are contributors to *IJARE* as frequent authors and reviewers. They also possess expertise in very timely areas of drowning prevention and learning to swim. They both enjoy a well-deserved international reputation as very thoughtful and reflective scientists as well as individuals with a well developed sense of scientific curiosity. Welcome aboard, Kevin and Bob!

*Steve Langendorfer, Editor*
*International Journal*
*of Aquatic Research and Education*

## References

American Psychological Association. (2010). *Publication manual* (6th Ed.). Washington, DC: Author

Twain, M. (1906). Chapters from my autobiography. *North American Review*. Boston, MA.