JSM Environmental Science & Ecology

Editorial

Some Statistical and Research Design Considerations towards Sound Scientific Inquires

Bahram Momen*

ENST Department, University of Maryland, USA

***Corresponding author**

Bahram Momen, ENST Department, University of Maryland, College Park, 20742, USA, Email: bmomen@umd.edu

Submitted: 09 August 2013

Accepted: 30 August 2013

Published: 02 September 2013

STATISTICAL INFERENCE

Scientific endeavors generally involve examination of samples' characteristics to draw conclusions at broader scales, and this requires statistical analysis. If the parameters (characteristics) of the populations to be compared are known, there is no need for statistical analysis to make an inference. The population parameters of interest may simply be compared to gauge their differences with certainty (ignoring Heisenberg's uncertainty principle). Although this seems to be a simple concept, searching the web reveals ample questions regarding the appropriate statistical procedures to compare population parameters.

FUNDAMENTAL ASSUMPTION IN MAKING ANY INFERENCE

Statistical analysis requires specific assumptions without which valid inference may not be made. Some of these assumptions should be strictly examined, followed, and/or enforced, but some may be treated as 'ideal conditions'. An assumption that should be met strictly before making any inference to a larger scale is 'independence' of observational or experimental units within study samples for a desired scope of inference. In observational (mensurative) or experimental (manipulative) studies, sample size or replication, respectively, should indicate the number of independent units. Although there should be no excuse for missing such an important concept especially after Hurlbert's (Ecological Monographs, 54:2,187-211, 1984) and many other follow-up articles, violation of such assumption abounds in presentations at scientific meetings and to a lesser degree in published articles.

REPLICATION, REPLICATION, REPLICATION

Units within a sample are just numbers for a mathematical statistician. However, they may represent watersheds, communities, hospitals, or units of an imaginary populations (those in mind when applying certain treatments), etc., that are hard, expensive, or sometimes impossible to replicate. In these cases, such limitations should be acknowledged, and qualitative, rather than statistical, comparison of the data should be performed. Detection of lack of independence is almost impossible when analyzing the data without proper information regarding the setup of the research and the way data were collected. Fortunately, the researcher has full control over 'independence' by replicating the units that represent groups (treatments) independently for a given scope of inference.

Replication enables calculation of the magnitude of 'noise' against which the magnitude of a 'signal' is compared. Without replication the denominator of any statistical test becomes zero prohibiting further calculations. However, minimal replication allowing statistical calculation to proceed does not suffice. Depending on the availability of resources, replication should be maximized to increase the precision (reproducibility) of estimates and the power of the test to detect a real effect. Although it may not seem intuitive, increased replication does not affect the Type I error on average.

IDEAL CONDITIONS

While independence should be strictly enforced for a desired scope of inference, other ideal conditions are less crucial either due to some mathematical concepts or due to the availability of analytical procedures. The effect of *Central Limit Theorem* (CLT: distribution of means of samples drawn from a non-normal distributions approaches normality with increasing sample size) on alleviation of the lack of normality (making ANOVA robust with regard to moderate departures from normality) seems to be unappreciated by researchers; perhaps because normality of 'what distribution' is rarely explained. In comparative studies, in which sample means are compared, the focus should be on the characteristics (e.g., normality) of the sampling mean distribution (or mean differences) rather than the distributions of the parent populations. This may not be mentioned clearly in applied statistical text books, and hence, researchers spend much effort on the examination of the normality of the collected data or more correctly on 'residuals' (treatment-mean adjusted data), using appropriate routines in statistical packages.

TEST OF IDEAL CONDITIONS AND DATA TRANSFORMATION

Available tests of normality cannot directly test normality of the sampling mean distribution as there is usually one mean for each group (treatment) unless simulated data are used. In addition, all statistical tests within the Frequentist statistics framework are intended to reject the null hypothesis (i.e., they cannot be used to prove the null). In the test of normality, it is hoped that the normality would not be rejected; but not rejecting normality does not mean that normality was proved. Finally, rejection of normality is affected greatly by the sample size. Through simulations it can be seen that normality may be rejected due to large sample size no matter how trivial (of no practical importance) the departure from normality is, while a severe departure from normality may not be detected due to small sample size. Undue emphasis on normality and ignoring the CLT may result in 'data transformation', which changes the nature of the responses, and magnitudes of variances and P-values, and hence, may lead to incorrect conclusions. Moreover, after data transformation, results should be reported and interpreted consistently based on the transformed units. However, this is cumbersome in addition to the fact that in many cases the transformed units may not make biological sense. Search of literature reveals abundant use of unnecessary data transformations and inappropriate reporting of the results.

Homogeneity of variances (HOV) is another ANOVA ideal condition with which researchers may be too obsessed albeit ANOVA is robust in terms of departure from HOV. Moreover, major statistical packages have routines to perform appropriate calculations based on homogeneous or heterogeneous variances. Perhaps a more detrimental issue of concern should be the correlation of variances with the group (treatment) means (as revealed by residual plots), regardless of homogeneous or heterogeneous variances.

CAUSATION OR CORRELATION

Causation versus correlation is another important consideration that has been dissuaded and addressed abundantly, but seems to be neglected frequently. It should be noted that causality can only be established through controlled experiments in which the researcher controls all variables being held constant or varied. Of course, the results of observational studies revealing correlations are of great value and can be used to suggest causation by an informed researcher in the field. Researchers may be aware of the issue but still use vague or inappropriate terminology (e.g., effect, in response to, a function of, related to, associated with, correlated with, etc. used interchangeably) to convey the results.

RESEARCH SETTING AND STATISTICAL ANALYSIS

Statistical analysis and experimental design concepts are linked closely, but differ sufficiently to warrant due attention. Since many educational programs offer only statistical analysis courses or offer statistical analysis courses prior to the experimental design courses, many students with limited time and course requirements may only take one statistical analysis course hoping to do justice to their research and publications. However, it is the design of the experiment that governs the validity of the research and its results. Experimental design or observational approach not only involves some preliminary steps such as setting objectives and scientific hypothesis (those that are falsifiable), realistic treatments levels and structure, use of covariates, etc., but also guides appropriate statistical analysis. The simplest case is the choice of two-independent or paired t-test that would depend on the experimental approach and the way the data were collected. Therefore, knowledge of research methods and experimental design is vital towards conducting a sound research and successful publication.

P-VALUE

I remember having a hard time linking the concept of the probability of an event to occur, such as predicting only 5% chance of rain during an important soccer game, with the fact that it actually occurred (100%). Subsequently, it was the conditional probability concept that became challenging. But perhaps it was not just me; rather Capt. Yossarian (a character in Catch 22 by J. Heller) might have also been surprised to see 99% of his comrades to be sick where he was recovering! Yet more challenging became the concept, use, and level of the P-value usually reported in scientific literature to declare statistical significance (probability of the Type I error). But again, it seems that it is not just me as an *Editorial* article in *Nature Medicine* (11: 1, 2005) acknowledged evidence that the authors of 31% of articles published by *Nature Medicine* in 2000, misunderstood the meaning of the *P* value.

There are more issues related to the use of the P-value beyond misunderstanding its meaning. The P-value routinely calculated by statistical packages and reported in scientific articles is usually intended to show the probability of the Type I error. This P value does not indicate the probability of existence of an effect or lack of it; rather it is a conditional probability (i.e., the probability of the observed results or more extreme ones **pending the null were true** --lack of an effect). The null is usually rejected if the P- value is less than, or equal to, a stated significance level (e.g., 0.01, 0.05, or 0.10) depending on the researchers' liking or the publication venue. While it has been argued that there is nothing magic about these levels, their importance is emphasized to eliminate subjectivity. Of course, these levels were selected subjectively, when no computer program was available, to obviate printing of thousands of pages of tables for critical values at a given range of P. Faced with choosing such levels and assuming that positive results have a greater chance of being published (albeit incorrectly), misuse of statistical analysis may occur even unknowingly to achieve a desired P-value to declare a 'significant' effect. However, current statistical packages report the exact P-value obviating the use of related tables. Therefore, it may just be prudent to allow the scientist, who is conducting the research and is familiar with both the field of study and the research limitations such as sample size, decide the 'existence' of the effect s/he observes and just report the P-value (whatever it may be) for the reader to make her/his own decision. This would perhaps satisfy the suggestion by Higgs (American Scientist, 101:1-9, 2013) to abandon the term 'significance' in scientific literature. Considering the above and allowing the publication of the negative findings could result in a substantial decrease in the misuse and abuse of statistics as well as in clearer reporting of the research protocol and findings.

TYPE I AND II ERRORS

Traditionally, researchers have focused on protecting against and reporting the results, based on the type I error. However, in many recent fields, where there is a risk involved if a real effect is not detected, the Type II error should be emphasized. The relationship between the Type I and II errors is one-way. Protecting against and decreasing the Type I error (by decreasing the significance level) increases the Type II error. And consequently, all the strategies to protect against the Type I error (including use of conservative multiple mean comparison tests) would result in increased Type II errors (and thus decreased power) to detect a real effect. A simple example to illustrate which type of error should be emphasized through the use of conservative or sensitive tests is the choice of an 'alarm system' desirable for a cheap car in a wealthy neighborhood or in an airport. A powerful (sensitive, liberal) alarm system, that may result in frequent 'false positive' (Type I error) in a safe neighborhood may not be needed for a cheap car that can be replaced with not much harm. However, we all favor a very sensitive alarm system at an airport to scream due to any penny in our pocket (increased false positive) hoping to increase the probability of detecting a forbidden item when there is one, and hence, decreasing false negative or Type II error (i.e., increasing power).

FURTHER CONSIDERATIONS

Discussion of using linear, linearized, fixed, random, mixed, and non-liner systems, as well as the choice of Frequentist or Bayesian statistics warrants further and much more detailed attention. If this note is read with so many questions still remaining, it serves a purpose. I frequently hear from fellow faculty members and research scientists as to why their students, who have taken a course in statistical analysis, are unable to be statistically independent. I hope one of these days I can convince them that the field of statistics and experimental design is so broad that no one or several course(s) can make anyone statistically independent. To this, I might add the complexities involved with learning advanced statistical packages and their appropriate routines.

Cite this article

Momen B (2013) Some Statistical and Research Design Considerations towards Sound Scientific Inquires. JSM Environ Sci Ecol 1(1): 1003.