Austin Journal of Environmental Toxicology

Open Access

Austin Publishing Group

Mini Review

Are Environmental Scientists using Statistics Correctly? A Review of Common Mistakes

Rispoli FJ[1]* and Green T[2]
[1]Department of Mathematics, Dowling College, USA
[2]Environment Protection Division, Brookhaven National Laboratory, Upton

***Corresponding author:** Rispoli FJ, Department of Mathematics, 150 Idle Hour Blvd, Oakdale, NY 11769, USA

Received: September 12, 2014; **Accepted:** February 12, 2015; **Published:** February 13, 2015

Abstract

The importance of statistics in environmental toxicology is obvious because much of what is learned about the environment is based on numerical data. Therefore, the appropriate use of data analysis and statistical methods is vital in environmental research. However, a growing body of literature points to persistent statistical errors, flaws, and deficiencies in published scientific work. In this paper we discuss frequently occurring errors noted in scientific literature with the hope of avoiding or at least reducing these mistakes in the future.

Keywords: Statistics in environmental research; Study design; Statistical methods; Model building

Introduction

Statistics and data analysis has long been regarded as a powerful tool in science and is used often in the study of environmental toxicology. The importance of statistics is obvious because much of what is learned about the environment is based on numerical data. Therefore, the appropriate use of data analysis and statistics is vital. However, a growing body of literature points to persistent statistical errors, flaws, and deficiencies in published scientific work [1,2]. For example, in March of 2014 *Scientific American* published an article citing a study in *Nature Neuroscience* that shows that more than half of 314 articles on neuroscience in elite journals during an 18-month period failed to take adequate measures to ensure that statistically significant study results were not erroneous [3]. Hence, at least some of the results in journals like *Nature*, *Science*, *Nature Neuroscience* and *Cell* were likely to be false positives, even after going through the strict peer review process.

In environmental applications the use of incorrect statistical methods may make individuals and organizations vulnerable to being sued for large amounts of money [4]. Often, after an environmental disaster such as a large oil spill or a natural disaster an environmental impact must be calculated based on historical data. It is important to point out that there usually is not a single correct way to gather and analyze this type of data. At best there may be several alternative approaches that are all about equally good. At worst the alternatives involve different assumptions and lead to different conclusions.

In this paper we present a review of common statistical errors, flaws and deficiencies concerning different stages of environmental research. The items presented are intended to help researchers to focus on what is important statistically and present it properly in their research papers. The paper addresses the stages of the research process from start to finish with respect to statistics by considering: study design, statistical methods, model building, documentation and presentation, and interpretation. In each of these sections a list of common flaws is given.

Study Design

The most important phase of any research is the planning and design phase. Proper and complete study design provides the foundation for sound research. At the top level studies can be described as either: observational, experimental involving treatments and controls, and meta-analysis which involve a review of many past studies. There is a vast amount of excellent material targeted at the design of experiments (e.g. [5] & [6]), however most of it is intended for statisticians. But this is not a valid reason to ignore design principles. Errors in this stage can have a negative impact on the validity and reliability of the research results.

When designing a statistical study the primary outcome measure must be reliable. In an ideal world the primary measure should tested for both repeatability and reliability using an ANOVA Gage R & R study [7]. The statistical as well as scientific hypotheses should be pre-specified and explicitly mentioned. In addition, serious consideration must go into determining the sample size. A small sample size may not have the "power" to detect small differences, even if they are statistically significant. The study design must consider expected differences among treatment groups, and what sample size is sufficient to detect such differences. This is extremely important in environmental science and toxicology when it is often very hard to make measurements that are precise and accurate.

- Study design flaws that often arise in research papers are as follows:*Study aims and primary outcome measures are not clearly defined or reliable*
- *The papers fails to report important parameters of the sample such as bias*
- *No a priori sample size calculation or power calculation*
- *Failure to use randomization when identifying experimental and control groups*
- *Use of an inappropriate control group*
- *Inappropriate testing for equality of baseline characteristics*

Statistical Methods

When applying statistical tests it must be clear to the researchers that tests are design for a very special purpose. Each test has a set

Citation: Rispoli FJ and Green T. Are Environmental Scientists using Statistics Correctly? A Review of Common Mistakes. Austin J Environ Toxicol. 2015;1(1): 1003.

of assumptions that must be met for the test to be meaningful. For example, there is an important difference between a pair-wise t-test and a two sample t-test which is often misunderstood. Moreover, each test involves the probability of making a Type I and II error (false positive and false negative conclusions) which are often overlooked. For a good reference see [8]. Another frequently occurring error is failure to test to see if the distribution in question is normal, which is a common assumption for many statistical tests involving a parameter such as the mean. In many areas such as human characteristics and manufacturing data, the normal distribution occurs often. However, in the environment this is not true. The lack of a normal distribution may indicate that it is necessary to use a non-parametric test.

- Statistical methodology flaws that often arise in research papers are as follows:*Use of wrong statistical tests:*

Unpaired tests for paired data or vice versa

Use of an inappropriate test for the hypothesis under investigation

Incompatibility of statistical test with type of data examined

Inappropriate use of parametric methods

- *Typical errors with tests of the mean:*

Failure to prove test assumptions

Improper multiple pair-wise comparisons of more than two groups

Failure to use multivariate techniques to adjust for confounding factors

Model Building

Mathematical models are often constructed using regression analysis to make predictions or assess the impact of various inputs. More than 4,000 hits were obtained with the keywords “multiple regression analysis” in the Science Citation Index within the areas of Environmental Sciences. In many environmental studies the reliability of measurement data is often difficult to control. For example, obtaining toxicity levels such as in the study [9] indicated a wide variation. Statisticians have studied the reliability of regression models and examine the “reliability matrix” [10] to help assess the model. However, rarely is a reliability matrix mentioned in an environmental paper. The conventional way to evaluate regression models is to consider regression model parameters such as R^2 and p-values associated with model coefficients. Using benchmarks for these parameters various decisions are made concerning the accuracy of a model. But for a scientist using a model to make predictions and inferences, these model parameters often do not provide a sufficient assessment. One problem is that R^2 values can be made artificially large by including an excessive number of terms, and p-values only indicate if a term is statistically significant and do not assess the accuracy of parameter estimation.

Let us digress for a moment and consider the origins of regression models which are traced back to Sir Francis Galton (1822-1911) who was trying to predict offspring heights based on data from parents. Galton [11] was interested in predicting heights that showed offspring of tall parents were, on average, not as tall as their parents, and similarly, offspring of short parents were, on average, not as short as their parents. Moreover, the generational average height remained the same. This process where the offspring are viewed as tending toward a population average is now referred to as regression to the mean, however, it was originally called “regression to mediocrity.” It was noted in [12] that Galton’s work compelled Karl Pearson and Alice Lee to study the height regression model. Pearson and Lee were bothered that the model seemed inconsistent with the notion that both parents were equally responsible for height. The coefficients of the mother’s height in the regression equations were invariably higher than the father’s coefficients, and they hypothesized that this is due to the fact that women were shorter than men. But admitting the mother’s measurements are more important than the father’s when predicting the height of an offspring may mean one of two things: The mother is more biologically important than the father, or the mother’s height is more accurately measured than the father’s. The authors in [12] argue that the latter is more plausible if one acknowledges a human behavior that is probably as old as marriage itself-marital infidelity.

Could this be an error embedded in the study design?

Obtaining a model with a reasonable number of terms is somewhat of an art. One rule of thumb is that the sample size should be at least five times the number of predictive terms. Unless there is extreme confidence in the measurement system, we believe that a high degree term (degree 3 or more) should not be included. Indeed, this is consistent with the “sparsity-of-effects-principle” which states that a system is usually dominated by main effects and low order interactions. The sparsity-of-effects-principle has been explained in depth in [13]. Once a potential model is constructed it should be validated as much as possible. Data points excluded from the sample may be use to confirm the model. Another method may be to simulate a small random error in the data points, say 5%, recalculate the regression model, and compare the result to the original model. If there is a significant difference, then the original model is sensitive to small changes in the input which must be considered when making inferences. The perturbation is intended to represent the measurement and systematic errors introduced when performing experiments with limited measurement resolution [14].

- The flaws that often arise in regression analysis are as follows*Key model assumptions that are often not confirmed*

The independent and dependent variables contain measurement errors

Residuals of the model are not independent over time

Residuals are not normally distributed

Residuals do not have mean zero, and do not exhibit constant variation

- *Model does not have a good number of terms*

R^2 is too low, more predictive terms are needed

R^2 is artificially high, too many terms are being used

- *Model is not robust*

Small changes in the input data can lead to large changes in the output

The model has not been validated

The number of input data points is too small

Documentation and Presentation

- All statistical methods applied in a study should be described clearly, accurately and with enough detail to enable a knowledgeable reader with access to the study data to recalculate results. A subsection of any paper should be devoted to issues arising in statistical analysis. Commonly used methods do not need to be described in detail, but unusual techniques should be completely described or referenced. In the world of environmental data that is more often skewed, than it is normal, giving medians, quartiles and ranges is often more meaningful than a standard deviation. It is not acceptable to just give means without any measure of variability. When statistical tests are used the resulting p-value or its equivalent should be reported as an exact value rather than merely stated that p < 0.05. A summary is as follows.*Inadequate graphical or numerical description of the basic data*

Providing a mean with no indication of variability of the data

Giving a standard error instead of standard deviation to describe data

Use of mean or standard deviation to describe non-normal skewed data

- *Inappropriate and poor reporting of results*

Results given only as p-values, no confidence intervals given

"p<0.05" or other arbitrary thresholds instead of reporting exact p-values Numerical information given to an unrealistic level of precision

Interpretation

Even when a statistical study has been well designed and implemented, its result may be misrepresented either by misleading graphics or by concluding statements. Researchers occasionally misinterpret the results of their own studies due to bias. Or perhaps jump to a conclusion that is either are not supported, or insufficiently supported by the study data and statistical results. When studies do not exhibit statistical significance, it is crucial to be careful in drawing conclusions. A lack of statistical significance does not automatically imply no difference. It may be that the sample size is too small.

- A summary of frequent interpretation errors is as follows:*Wrong interpretation of results*

"Non significant" Interpreted as "No effect" or "No difference"

Drawing conclusions not supported by the study data

Significance claimed without data analysis or statistical test mentioned

- *Poor interpretation of results*

Disregard for Type II error when reporting non-significant results

Failure to discuss sources of potential bias and confounding factors

In summary we note that too much is at stake in terms of what we learn about the environment to ignore the issues cited here. There are many environmental papers with an excellent use of statistics. However, there are also too many flawed papers in the scientific literature whose analysis can be improved. We hope that this paper will at least get these studies going in a better direction. Perhaps encouraging researchers to have the paper in question reviewed by a statistician will be a reasonable next step.

References

1. Buhl-Mortenson L. Type-II Statistical Errors in Environmental Science and the Precautionary Principle. Marine Pollution Bulletin. 1996; 32: 528-531.
2. Stix G. Statistical Flaw Punctuates Brain Research in Elite Journals. Scientific American. 2014.
3. Nieuwenhuis S, Forstman BU, Wagenmakers EJ. Erroneous analyses of interactions in neuroscience: a problem of significance. Nature Neuroscience. 2011; 14: 1105-1107.
4. Manly BFJ. Statistics for Environmental Science and Management. Chapman and Hall/CRC. 2001.
5. Montgomery DC, Design and Analysis of Experiments. 7th edn. New York, Wiley. 2009.
6. Canavos GC, Koutrouvelis IA. An Introduction to the Design & Analysis of Experiments. Pearson Prentice Hall. 2008.
7. Burdick RK, Borror CM, Montgomery DC. Design and Analysis of Gauge R and R Studies. 2005.
8. Douglas C. Montgomery. Making Decisions with Confidence Intervals in Random and Mixed ANOVA Models. American Statistical Association and the Society for Industrial and Applied Mathematics. 2005.
9. Ott RL, Longnecker MT. An Introduction to Statistical Methods and Data Analysis. 6th edn. Cengage Learning. 2010.
10. Rispoli F, Angelov A, Badia D, Kumar A, Seal S, Shah V. Understanding the toxicity of aggregated zero valent copper nanoparticles against *Escherichia coli*. J Hazardous Mat. 2010; 185: 212-216.
11. Gleser LJ. The importance of assessing measurement reliability in multivariate regression. Journal of the American Statistical Association. 1992; 87: 696–707.
12. Galton F. Regression towards mediocrity in hereditary stature. Journal of the Anthropological Institute. 1886.
13. Marcello Pagano, Sarah Anoke. Mommy's Baby, Daddy's Maybe: A Closer Look at Regression to the Mean. Chance Magazine. 27.
14. Wu CF, Hamada M. Experiments: Planning, analysis, and parameter design optimization. 2nd edn. Wiley. 2000.

Austin J Environ Toxicol - Volume 1 Issue 1 - 2015
ISSN: 2472-372X | www.austinpublishinggroup.com
Rispoli et al. © All rights are reserved

Citation: Rispoli FJ and Green T. Are Environmental Scientists using Statistics Correctly? A Review of Common Mistakes. Austin J Environ Toxicol. 2015;1(1): 1003.