

SIMPLE OVERSIGHTS THAT LEAD TO ERRORS IN DATA EVALUATION

By James R. Chastain, Jr., PhD, PE, MPH



It has been said that the vast majority of mistakes in statistics (and life) result from a failure to plan. Prior to undertaking a quantitative study of a problem, it is appropriate to develop a clear plan of how the data will be evaluated to support the decision making process. Although this would seem to be common sense, it is a step that is frequently omitted. When the data collection effort is expensive, the planning effort becomes increasingly important. The advent of the internet has provided convenient access to large stores of data, which introduces the issue of effectively judging the “goodness” of the data. So even if the acquisition of the data has little cost associated with it, the accuracy or relevance of the data can affect the outcome in unpredicted ways if not adequately vetted. This article briefly presents a few aspects of statistical planning that might be useful in evaluating or interpreting a data set.

At the outset one must define the point or objective of the study. Again, this may seem like common sense, possibly even self-evident, but it is interesting how a little probing will produce the need to clarify our thinking. A few common questions to ask along these lines would include (but not be limited to) the following.

Are there spatial or temporal variations in Variable Y?

This is probably one of the most common outcomes sought from a data collection effort. Routine statistical analysis methods such as analysis of variance (ANOVA) or regression methods are fairly well suited for describing these relationships. In many cases, this analysis should be viewed as an exploratory effort because underlying mechanisms or drivers may not be evident. However, without this step, it may not be possible to identify these initial relationships.

What is the effect of Factor X on Variable Y?

This is another question that is frequently the result of a statistical analysis. To be formally correct, this should be the result of an experimental (manipulative) design. In that way, confounding variables can be controlled or minimized. If performed in this way, the resulting p-value can be used to meaningfully test for significance of Factor X. Many times, however, observational data is used to test this relationship. While this approach does not necessarily invalidate the results, the resulting inferences are usually weaker because confounding variables are not managed.

Are the measurements of Variable Y consistent with the Hypothesis under consideration?

This question is appropriate when the study is seeking to confirm an assertion (hypothesis) or mathematical model. Data from either an experimental or observational study can be used for this purpose. The challenge under this category is to plainly state the hypothesis. Unfortunately, in environmental studies, it may be difficult to state simple, falsifiable predictions. Therefore, care must be taken when interpreting and making confident assertions about the conclusions in that event.

Using the Measurements of Variable Y, what is the Best Estimate of Parameter θ in Model Z?

This question is somewhat infrequently used, but is a powerful tool in confirming and refining mathematical models of a condition. Parameter estimation is required to develop functional predictive models. While there are a number of ways to do this, a careful statistical approach is probably the best. As with the other categories, by keeping in mind the overall objective of the study, care can be taken to understand and account for accuracy and variance in the underlying data. Otherwise, the predictive power of the model will suffer.

There are certainly other questions or data objectives that can be envisioned in a study. However, these are a few of the more common ones and provide some insight into the types of probing that should be carried out prior to beginning the study effort. Once the data is collected, it must be assessed and interpreted. The following section provides a few considerations to keep in mind when developing the report.

Observations and Cautions When Forming Statistical Conclusions

Understandably, the use of various statistical methods is related to the analyst’s familiarity and depth of knowledge of those methods. A number of studies in reputable, peer-reviewed journals have illustrated that these errors can occur even in otherwise advanced treatment of complex issues. The point here is that the integrity of a well-researched study can be compromised by a fairly fundamental mishandling or misinterpretation of the data. A few of the more common problem areas are summarized in the following.

Mean

Computation of the mean (average) is such a fundamental statistical measure that most people give no thought to computing it as a representation of the central tendency of the data set. The mean does in fact have some amazing mathematical features and deserves a prominent place as a statistical parameter. The Central Limit Theorem assures us that, regardless of the underlying distribution of the data, the true value of the mean will lie within two standard deviations (actually 1.96) of the computed mean 95% of the time in repeated tests.

Unfortunately, while well known and commonly reported, the mean is commonly misunderstood. For example, the common relationships about the mean do not apply to individual values...they apply to the true values of the mean. This may seem to be a subtle issue, but it can make a significant difference in drawing conclusions from the data. For example, if the underlying data distribution is non-symmetrical, expecting some level of data to lie within a specified number of standard deviations will provide erroneous results.

Possibly a better way to present the central tendency of the data set is to compute both the median and the mean. Since the median is a non-parametric parameter, by comparing the two, one can quickly get an indication whether the data is symmetrically distributed. The more the two are separated in value, the more non-symmetrical the distribution. This provides a rough check on whether common statistical relations will be appropriate or not.

Standard Error

The standard error is an important statistical parameter. Many technical journals require that the “mean ± standard error” be included with all data sets. Quantitatively, the standard error is defined as:

$$SE = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{\sqrt{n(n-1)}}$$

It is related to the standard deviation and in fact can be considered the standard deviation of the mean as opposed to the standard deviation of the data set.

Standard error is quite helpful if the data takes a normal (Gaussian) distribution. In this case, the data is symmetrical and as statisticians say “is well behaved”. Unfortunately, in many cases the data will have non-Gaussian or truncated data sets. In these cases, the standard error can’t legitimately be used to draw the common inferences from the data set.

As noted above, the standard error is computed as the square of the difference between mean and individual values. Thus, if there are several exceptionally large (or small) values in the data set, it will have a noticeable impact on the interpretation. This is especially true if the values are outliers or the data set is small (less than 8 data points). In fact, if the data set is less than 6 data points, the standard error is virtually meaningless.

If the statistical parameter of interest is the mean (as opposed to other statistical measures), the good news about the standard error is that as the data set becomes larger, some of the discontinuities begin to resolve. The Central Limit Theorem will smooth some of the data unevenness, and the information about the mean will become more accurate. However, as mentioned above, this use of “± SE” will infer a symmetry to the data set that may not exist.

The key message here is that the sample size needs to be large enough to smooth the data for the standard error to be meaningful. Most texts recommend at least 30 data points.

Confidence Interval

One technique that has been proven to help more accurately portray various parameters (in addition to the mean) is the use of confidence intervals. Most statistical software allows the convenient computation of these values within some stipulated probability (normally 95%). As a general rule, there should be at least 8 data points available in order to stabilize the computational values. Confidence intervals can be used to evaluate the precision of the estimates and the significance of hypothesis tests. It is important to realize, however, that the center of the confidence interval is no more likely to represent the true value of the parameter than any other point within the interval. As with all statistical tests, the reliability of the results are only as good as the validity of the input data (i.e., GIGO).

Recommendations

The key message here is that care must be used when interpreting data and the associated statistics. Computer software packages make it quite easy to compute statistics, regressions and inferences. However, each of those techniques have specific, implicit assumptions and when violated can produce inaccurate inferences. The following are a few rules of thumb to help minimize those problems.

1. Define the purpose of the study. Recognize whether the data is observational or generated from an experiment. Observational data is prone to biases and confounding, so it should be used more to develop hypotheses rather than to define or prove cause/effect relationships.
2. Always plot the data and look for underlying patterns and distributions.
3. In small data sets (less than 6 data points), summarize the data by using minimum, maximum, mean and median. Compare the mean and median for evidence of symmetry. The closer the mean and median, the more likely the distribution is symmetrical. Standard error statistics are ineffective in this range.
4. Larger data sets (8 or more data points) are generally large enough to begin stabilizing statistics, especially if the data set is normally distributed. Even with the larger data sets, non-symmetrical data would probably be better served by minimizing the use of statistics that are based on normal distributions. Techniques such as box-plots, median/mean relationships, and more advanced procedures such as bootstrapping will more accurately characterize the data.
5. With larger data sets, confidence intervals should be computed and reported.

Dr. Jim Chastain is the CEO and President of Chastain-Skillman, Inc. He has a Bachelor of Science in Civil Engineering (honors) and Master of Engineering from the University of Florida and a Master of Public Health and Ph.D. in Public Health from the University of South Florida. He can be reached at (863) 646-1402 or jrchastain@chastainskillman.com.

© 2011 Chastain-Skillman, Inc. This article is taken from the 3rd quarter 2011 issue of Consultant's Update, a publication of Chastain-Skillman, Inc.