Abstract - Statistical analysis of all sorts of data are routinely presented at technical conferences to indicate a
causal effect of positive change based on some previously implemented strategic decision. They are nearly always
wrong. This paper explores some of the problems with statistical analyses and the issue of separating the “signal”, or
the true cause and effect data, from the “noise”, or background data that may be highly correlated, but are totally
unconnected to the cause and effect claimed.

Keywords: Signal, noise, statistical analysis

THE PROBLEM

Any three or more data points can be analyzed by statistical methods. In most cases the intent of those analyses is to
indicate that the results or findings of the analysis are not likely to have occurred solely by chance and that there is
some causal relationship between one set of events and another set of outcomes. This may, of course, be correct. A
wall could be streaked with a paint brush that had just been dipped into fresh paint. This experiment could be
conducted multiple times, hundreds of times, even, and the result would always be the same: there would be paint
left behind on the wall. These observations could then be analyzed statistically, using any favorite statistical model
the analyst wishes to use, and the results will always indicate that the probability of the wall becoming smeared with
paint in the absence of the action is essentially zero and that the probability of it happening as a result of the stated
action is essentially 100% certain. There is very little noise in those data and the results can be relied upon with
great confidence.

We can also send out a survey to several hundred alumni of our individual university and ask them how well their
education assisted them in evolving in their chosen career. We could do a lot of research and independent
verification of the questions on the survey instrument to minimize bias and ambiguity in the questions, and we could
offer multiple incentives to get people to complete the survey within a reasonable specified time frame. The data
gathered from this type of survey, however, are fundamentally flawed and so distorted by noise that they are
essentially worthless. How can that be?

The second case study just cited is flawed because the cohort of people to whom the survey was sent was self-
selected to be part of the Alumni Association. That, at the outset, introduces a favorable bias on the part of the
surveyed alumni. Those who respond may be responding favorably to the survey for three reasons unrelated to the
underlying objective of the survey: they want to please their institute by providing good, encouraging survey results
and/or they feel so good about their educational institution, in general, that they want to say nice things about it for
the use of institute advertising, and/or they believe that they should say something nice because they are being given
a prize or incentive to complete the survey. All three feelings can bias the outcome data significantly.

1 Wentworth Institute of Technology, 550 Huntington Avenue, Boston, MA, 02115, hopcroftf@wit.edu
It may be argued, of course, that there are likely also people who respond because of some bad feelings recently encountered or developed during their educational years and that those people so bias their answers negatively that it serves to offset the positive bias of the other respondents. While there may well be some of that happening, there is no way to know whether there is a true offset and even if there was, the noise surrounding those data points is so loudly biasing the data that the results may not be relied upon.

**EXAMPLES**

Recent advertisements for a particular brand of beer have been running during various football games of national interest. One such ad shows, for example, a group of friends sitting in the stands when a particularly difficult field goal attempt is about to be made by their team. They all turn the labels of their beer bottles out towards the field to “create a force field that will guide the football through the uprights”. They have apparently done this many times before and find that more often than not, when they all do this, the football does, indeed, go through the uprights for a field goal. The implication is that if the number of times this was done and the ball went through the uprights, versus the number of times this was not done and the football failed to go through the uprights, were analyzed there would be a very strong correlation between the event of turning out the labels and success of the field goal attempt. However, the event cannot be related to the outcome in any meaningful way and to suggest a cause and effect relationship, regardless of the statistical significance of the correlations suggested by various statistical analyses of the data, is ludicrous at best and disingenuous at worst. This is clearly noise masquerading as signal.

**STATISTICAL ANALYSES AS ASSESSMENT TOOLS**

The use of statistics to document the validity of data used to assess programs and program components has become a mainstay of the profession in light of the ABET 2000 criteria and subsequent variations thereof. Adequate assessment is, in fact, a key to successful accreditation and reaccreditation and there are precious few good ways to generate or analyze suitable assessment data. Most such data are generated through surveys of current students, recent graduates and alumni, along with surveys of employers of graduates and cooperative education or internship employers where those elements are strong components of the overall program. In the general case, a survey instrument is developed and designed to be comprehensive, but short, all inclusive and sufficient to separate the various respondents into suitable groups or categories for analysis. This is an impossible task at the outset, of course, but that minor inconvenience seldom deters us from charging forward with a survey anyway.

A case in point concerns a recent paper presented to ASEE national for presentation at its June meeting in Atlanta, GA. This particular paper was focused on assessing a new program designed around project-based service learning projects [1]. The underlying question was whether the projects themselves added value to the engineering program and whether specific design skills and a broader set of skills required for graduates to thrive within a global environment were being successfully imparted to the students. The assessment data were generated through a survey carefully crafted and internally tested to eliminate bias and ambiguity in the questions, then distributed electronically to over 1500 alumni with valid email addresses. Just over 1000 alumni who did not have a registered email address were sent a postcard with the link on it. Of a total population of just under 2,600, a total of about 600 people responded.

That is actually a fairly large response pool and it is common to assume that a response pool of that size, a 23% response rate, should give some fairly reliable data and that the statistical analyses should be reliable indicators for the future. Notice, however, that the pool of respondents is self-selected, probably already feels good about the institute (or, conversely, has really bad feelings about the Institute) and that the data are going to have a bias, probably a positive bias, right from the start. Any analyses, then, are going to be unreliable, at best.

A second case study involves another paper submitted to ASEE for presentation in Atlanta. This paper looks one year later at the impact of Leadership and Service Learning Modules in the first year of study on the successes of the
students in subsequent years and in their interest in pursuing various leadership or service components in their educational careers [2]. The Leadership module was very well designed and incorporated a host of excellent service components working at a local children’s museum to develop learning modules for the children. It also included lectures on leadership and other elements of interest and compatibility. The leadership module was optional in that students are required to take two or three of several modules, one of which is the Leadership and Service Learning module.

The study focused on whether those students who chose to take this module were more likely to engage in leadership activities outside the confines of that class and whether those students were also more likely to undertake service activities outside of that class than students who did not take that specific module. Not too surprisingly, the statistical analysis of the data suggests that the answer is yes in both cases.

Think about why that might be. This module is voluntary. Consequently, those who take it are already pre-disposed to leadership roles and may well be looking for either an easy module to take, or a way to improve those skills they know they want to hone. Those who choose not to take that module are predisposed to being followers and are not likely to have exercised significant leadership or service activity in other areas after the class was over. In addition, only 87 students responded to this study survey. Of those the response rate was about 24% for students who took the Leadership module and about 18% for students who did not take that module. (Note the consistency with the response rate for the previously cited study). Since the majority (54 of 87) of the students who responded took the module being tested, there is already a positive bias towards that module and those students were self-selected, which tends to indicate a positive attitude about the module before they start the survey. Consequently, what can be said about the results is, at best, that those who take that module tend to be those who are going to be leaders anyway and that this module may reinforce those skills, but there is no indication of a cause and effect that taking the module creates this desire to be a leader or that this module improves the leadership skills of those who take it. There is a lot of noise here, but very little in the way of true signal.

A much more obvious example occurred several months ago when the author decided that he was tired of carrying a heavy laptop and a heavy briefcase back and forth from school to home. So one day he put both of those onto a wheeled luggage carrier and proceeded to pull the carrier along the sidewalk behind him on the walk to the local train. On that particular day, the train conductor announced part way into the city, that the train was not going to be able to stop at either of the two stops favored by the author and would, instead have to go all the way to the main station downtown – a 2 ½ mile walk back to campus. On no other days has this happened. Consequently, the correlation between this faculty member using the luggage carrier and the train not stopping at either station is extraordinarily high. All the statistical tests that could be done would indicate that since it never happened without the luggage carrier and it always happened when the carrier was used, there must be a clear cause and effect relationship between the two.

Bunk. That is clearly very loud noise masquerading as signal and there is equally as clearly absolutely no causal relationship between the two events, regardless of the statistical inferences that may surround them.

So how can this dilemma be resolved? If we are to do assessments of programs and parts of programs we are going to need to generate data, most commonly with surveys of students, alumni and internship employers and we are going to have to do some statistical analyses of those data when we are done. If the data are so unreliable, what good does it do to spend the time and energy to conduct the surveys and perform the statistical analyses?

The answer is a two-fold response. First, we need to recognize that the data do not always tell us what we want to hear with anything like the veracity with which we would like to hear it. We need to be careful about making sure we are asking the data the right question. Second, we need to retrain to be able to add something to the statistical data to account for the variations and self-inflicted biases that come with the data. Very few data are, in fact, pure; but statistics inherently assume that they are. We need to recognize the uncertainty in predicting future outcomes.
based on historical data and on data generated from self-selecting (and potentially self-serving) respondents – regardless of the statistical significance of those data.

An interesting discussion on this topic comes from a book by Bob E. Hayes, called Measuring Customer Satisfaction [4]. In this book Mr. Hayes discusses a variety of topics related to the design of survey instruments and the analyses of the resulting data. He points out that the best we can do when we want information regarding how our customers feel about a product or service is to ask them some questions about it and see what the results are. Chapter 3 of that text, titled “Reliability and Validity”, presents “measurement issues that demonstrate the importance of careful thought when designing questionnaires to measure perceptions and attitudes.”

Mr. Hayes also distinguishes between “customer satisfaction” and “perceptions of quality”, which he defines as two very different things, but both of which may be used to summarize a set of observable actions related to the product or service. For example, if we observe that our customers (students, in most cases) are generally smiling, laughing and saying positive things when we talk to them, we might reasonable infer from those observations that the students are happy with the services we are providing to them. Similarly, we can infer attitudes and perceptions about the quality of the services provided by observing the responses to carefully crafted questionnaires that our students are willing to complete for us.

It is noted, however, that we can never really know the true underlying level of our student satisfaction. The best we can do is to design and develop appropriate survey instruments that will generate data from which we may infer student satisfaction. Therefore, it is critically important when developing such instruments to consider the measurement issues to ensure that the data derived from those instruments provide accurate information regarding the attitudes and perceptions being tested. This is true when we make physical measurements in the field or in the laboratory as engineers, and it is equally as important when we measure student perceptions as educators.

Reliability is defined as the extent to which measurements are free from random error variance [4]. The data obtained from a questionnaire are comprised of observed scores, but those scores are themselves comprised of a true value and an error value. We use statistical analyses to try to discern the most likely value of the error component so that we can then assign a reliability judgment to the rest of the data. If we have designed the perfect survey instrument, the error score, in the assessment of the overall responses, will be zero, although the error inherent in the responses to any specific question may be significant. The reason for that anomaly is that the perfect survey instrument would have sufficient verification questions imbedded into it to eliminate the error components through correlative analysis of the answers to specific questions. A perfect survey questionnaire has likely never been developed, however, so assuming that the one being analyzed fits that definition is going to allow a lot of noise to masquerade as signal and lead to very bad decisions about the meaning of those data.

Nate Silver has written a definitive treatise on this topic titled “The Signal and the Noise: Why So Many Predictions Fail – But Some Don’t” [10]. Mr. Silver uses a large array of different circumstances in which data are developed and analyzed for a variety of different purposes to illustrate the problems associated with separating the noise from the signal. It is reported that well before the election in 2008 Mr. Silver correctly predicted the presidential winner in 49 of the 50 states and accurately predicted the winner of all 35 senate races the same year. He also created a method for predicting the outcomes of baseball games that was so successful that the system was purchased by “Baseball Prospectus” in 2003 [5]. In his book he looks at successful and unsuccessful attempts over the years to forecast everything from stock prices, weather, and anything else folks want to predict. There was a synopsis of this book published by Burton Malkiel in the online version of the Wall Street Journal on September 24, 2012 [8].

When we use surveys to generate data about the perceptions of our students, our graduates, or anyone else, we are trying to predict, based on the historical satisfaction of past users of whatever we are measuring, how well future users of that information or system will view it. We are, in essence, trying to predict the future. Indeed, the data from past users of the system or service is of no practical use at all unless we are trying to predict the future with it.
If the data indicate current success, we stay with the current system; if not, we try to adjust and change to improve the satisfaction scores in the future.

So why is Mr. Silver so much more successful than other people in forecasting the future? Well in truth, he is not that much better in most areas. He happens to have selected a few very specific areas in which he has spent considerable energy researching and in those areas he is very good. In other areas he is no better than most. He then points out that he does not trust any predictive method or program completely because there are way too many random factors involved in every prediction to be able to create a program or system that will be 100% right 100% of the time – even the systems that he has developed for his own, highly successful, use. He always tempers the system outcomes with some common sense thinking and analysis of what could be introducing random error, how much those errors may influence the outcome, and what the most reasonable outcome scenario is likely to be based on that intuitive modification of the model output data.

Two very specific issues raised by Mr. Silver are the following. The first is what he has called “the out-of-sample problem” [5]. That means that the event the surveyor is concerned about did not exist in the data set. The example he uses is that in 2007 the Federal Reserve Bank predicted a very low probability of a recession in the foreseeable future. But one started a month later. The data from which the prediction had been based were from 1986 to 2006 during which there had been only two very minor recessions. The “big one” had not occurred during the data generating period and therefore it could not have been predicted from those data.

The second problem cited is called “overfitting of the data”. There are patterns which develop in the data, but the patterns do not have any predictive power. They are merely descriptive, and so often, as in the case of the train breakdown versus the use of a luggage carrier cited earlier, they are descriptive of the noise and not the true signal. He also cites the example here of earthquake predictors wherein people try to apply complex equations to eight different variables to a very noisy data set and then try to predict outcomes based on the analyses of those data. It does not work. Similarly, the government regularly publishes data on 45,000 different economic statistics. It is not hard to find correlation between dozens of those indicators and any economic event being analyzed. That does not indicate predictive qualities among the statistics, however; only apparent, and random, correlation over the time frame analyzed.

Weather forecasting is perhaps the most commonly recognized predictive use of statistical data. Every weather forecaster on the television or radio has been analyzing myriads of data from dozens of sources, including several different and differing computer models to come up some notion of what the weather is going to be like in a specific area of the country from several hours in advance to several months in advance. The interesting thing about weather forecasting is that weather forecasters are among the very few people who routinely clothe their forecasts in a cloak of uncertainty. They will say “there is a 40% chance of rain tomorrow” or that “there is a 60% chance of a thunderstorm in your area tonight”, for example. In doing so they are overtly recognizing the inherent uncertainty in their own forecasts due to a variety of random things that can happen in the world to change the path or direction, the magnitude or the timing of any weather system. Sometimes, even a mere hours before a storm is expected to hit a specific location, the models are still telling different stories and the forecast is still very much in doubt.

A corollary to that is the art of predicting global warming and the effects that global warming may have in and on various parts of the world. A classic treatise on this topic was published in 2010 by John Christy, et al, regarding what the observational datasets say about modeled tropospheric temperature trends since 1979. [6] Note that this is a very short, 30 year period, in observational weather data. The analysis of the data in that paper suggests, parentheses added, that “The majority of AR4 (a set of computer based models developed by the Intergovernmental Panel on Climate Change in 2007) simulations tend to portray significantly greater warming in the troposphere relative to the surface than is found in observations.”
A large array of bloggers and others have taken issues with the Christy, et al study and with a variety of other studies on the same topic for failing to properly incorporate the probability of random error in the output data. See for example, comments by Roger Pielke on challenges made to that paper [9], a blog string initiated by Greg Laden [7], and comments by Anne Stark at phys.org [11].

In the end, we are all storytellers when we use statistics to relate historical data to future predictions. We need to be especially cognizant in the world of academics, wherein the public expects truth to be found, that we tell the truth, including the potential risk of random errors that may be present, in any work we produce. When reading the predictions from outcomes, we are admonished by a journalist identifying himself only as “Dr. Denny” [3] that we need to ask certain questions when we analyze the output from any storyteller:

What are their qualifications?
To what audience are they appealing?
What motive have they not revealed?
What is the purpose of the storytelling?
Do their arguments make sense?
What do they offer to suggest that they are credible?
Am I encouraged to think for myself?

These questions are equally as appropriate for researchers as they are for journalists.

REFERENCES

Prof. Francis J. Hopcroft

Prof. Hopcroft has been teaching Civil and Environmental Engineering at Wentworth Institute of Technology in Boston since January of 1994. He joined Wentworth after 26 years in consulting and regulatory agency service. He is a Registered Professional Engineer in all six New England States and a Licensed Site Professional in Massachusetts. He holds two U.S. Patents and has authored parts of three text books and numerous papers on a variety of topics.