This article is part of the IEEE Reliability Society 2008 Annual Technology Report.

The Science and Pitfalls of achieving Verifiable Data

Samuel Keene
Email: S.keene@ieee.org

Background – a life long search

This author has long been a student of reliability successes and, more often; of reliability failures [1]. This article seeks to identify some of the underlying causality in experimental errors and report the "lessons learned", as well as best practices found for assuring data validity. Some of the "lessons learned" came about in testing and qualifying parts for military, space and commercial programs. These part qualification efforts are data collection intensive with lots of opportunities for error. These data errors have come from:

- Drifting in measurement apparatus corrupting the reported data. One case had a resistor heating up in the measurement system that skewed the measured data. Plotting the Measured data from 50 identical devices showed a definite data pattern vs. the random pattern that should have been observed. This data pattern was not observed or acted upon by the technician collecting the data. It was only observed in hindsight.

- Measurement equipment out of calibration. The lab technician was measuring the gate to cathode voltage in the range of hundreds of volts across a triac motor driver. The differential scope probe was subsequently found to read the same level of spikes when the probes were on the same triac terminal. The hundreds of volts were due to a timing offset in the probes.

- Measured data was not representative. Devices tested during development testing were not representative of the production product. The design point determined during the initial development was based on early samples and did establish a robust solution using production samples.

- Data can be confounded where a trend of one variable' effect is mistakenly attributed to another co-varying variable. For example, the cocky rooster crowing every morning believes he brings up the sun each day.

- There can be variable interaction effects that are not properly isolated or recognized. When Firestone tires (condition 1) were underinflated (condition 2) to increase vehicle stability on Ford Explorers (condition 3) rollovers increased. It took all three conditions to create the problem. This is an interaction.

- Lack of control samples to contrast experimental effects, demonstrate measurement repeatability, or maintain traceability of test results.

- Like the Heisenberg uncertainty principle says, "just measuring something affects its value". So it is with experimentation. We want to minimize the effect of the experimenter on the experiment outcome. This measurement effect can be mitigated by using the double blind experimental practice. This is often used in medical studies to determine the efficacy of a new drug, for instance. Then the subjects undergoing the testing don't know if they are getting the new drug or the placebo. Also the Doctors running the test do not who is getting the real drug vs. the placebo. This takes out the effect of Doctor or the patient expectation on the experiment result.

- There are also data transcription and recording errors. This even happens today with all the automated data logging capability that we have. An example of this is the "hottest October on record", which actually turned out to be September data inadvertently repeated. This will be discussed below.

- One can also see errors in the routine polling processes to assess customer preferences. To be done correctly, the polling must be scientifically designed including drawing a statistically significant, representative, and sample from the relevant population that we are trying to assess. The polling sample needs to be randomly drawn to preclude potential bias in the poling results. This is sometimes difficult to achieve and needs a plan and have the plan execution monitored. Poling questions often force decisions into too few selections, and don't allow for the cases of: don't know, don't care, or feel equal between the choices.

"life is the art of drawing sufficient conclusions from insufficient premises"
Samuel Butler 1835-1902

Some recent data problems:
Michael Mann, along with his co-workers, published an estimation of global temperatures from 1000 to 1980 [2]. They arrived at this estimate by combining the results of 112 previous proxy studies. By "proxy studies" I mean tree-ring and isotope and ice core studies that are intended to provide an indirect measurement of temperature in the time before thermometers existed. Mann's results appeared to show a spike in recent temperatures that was unprecedented in the last one thousand years.

Mann's assessment of the data was criticized on several fronts. The first was historical fact: his chart didn't appear to show the well-known medieval warm period, or the so-called little ice age that began around 1400.

Two Canadian researchers, McIntyre and McKitrick, obtained Mann's data and repeated his study. They found numerous grave and astonishing errors in Mann's work, which they detailed in 2003 [3]. For example, two statistical series in Mann's study shared the same data. The data had apparently been inadvertently copied from one series to another. In addition, nineteen other series had had gaps in the data, which Mann's team had then filled in - a fact that had not been disclosed. In addition, all 28 tree ring studies had calculation errors - and so on and so forth. Such that in the end, the Canadians' corrected graph looked quite different: The corrected graph suggests that the global temperature today is very far from the warmest it has been in the last thousand years.

Mann has countered these claims "http://info-pollution.com/mandm.htm**,"** so the debate is on-going.

Another facet of this global warming data controversy was the Goddard Space Information Systems reported (incorrectly) that October 2008 was the hottest October on record. http://hotair.com/archives/2008/11/16/hottest-october-on-record-was-really-a-september/. An excerpt from that report: "A GISS spokesman lamely explained that the reason for the error in the Russian figures (that they had used) … were obtained from another body, and that GISS did not have resources to exercise proper quality control over the data it was supplied with."

"The great danger here is that public policy and law can be launched from a faulty premise. If language be not in accordance with the truth of things, affairs cannot be carried on to success."

Confucius

Is Snopes the final answer?

"Whoever undertakes to set himself up as judge in the field of truth and knowledge is shipwrecked by the laughter of the gods."
Albert Einstein

"Who dares to say that he alone has found the truth?"
Henry Wadsworth Longfellow

For the past few years www.snopes.com <http://www.snopes.com>; has portioned itself, or others have labeled it, as the 'tell all final word' on any comment, claim and email. Wikipedia reports is run by a husband and wife team. No big office of investigators and researchers, no team of lawyers. It's just a mom-and-pop operation that began as a hobby.

David and Barbara Mikkelson in the San Fernando Valley of California started the website about 13 years ago - and they have no formal background or experience in investigative research. It is doubtful that snopes is run without bias and they have been proven wrong (http://patterico.com/2007/05/29/snopes-wrong-again-on-flight-327/). So Snopes is a good starting place but it should not be totally relied upon or considered the final arbitrator. Use it only to lead you to their references where you can link to and read the sources for yourself. Plus, you can always google a subject and do the research yourself.

"If you add to the truth, you subtract from it."
The Talmud

Data discipline:

As scientists, engineers and decision makers we need to routinely question the:

1. measurement requirements
2. Experiment design
3. Measurement process
4. Measurement calibration
5. Gage R&R (see below)
6. Data collected
7. Data legacy depository capability for traceability

Six Sigma uses a programmed Gage Repeatability and Reproducibility (GR%&R) process to validate the measurement capability. See http://en.wikipedia.org/wiki/ANOVA_Gage_R&R. Data measurements are replicated in randomized order by the first measurer and then replicated a second time by an independent measurer. There are six sigma guidelines on what constitutes an adequate data measurement capability. "what gets measured (data gage), gets improved".

We possibly need Six Sigma Gage R&R experimental verification, DOE, best practices, independent and open data and analysis reviews of critical government funded research. A lot of policy ($) rides on the premises formed by this experimentation. It might make sense to triplicate critical environmental research experiments, across diverse teams, with cross reviews. The operating cost: millions. The potential policy cost savings: billions.

"The truth may be puzzling. It may take some work to grapple with. It may be counterintuitive. It may contradict deeply held prejudices. It may not be consonant

with what we desperately want to be true. But our preferences do not determine what's true." Carl Sagan, 'Cosmos'

References

[1] Keene, Samuel, J., "Validating Measurement Data." Evaluation Engineering, November – December 1969.

[2] Mann, M.E., R.S. Bradley, and M.K. Hughes, Global-scale temperature patterns and climate forcing over the past six centuries, Nature, 392, 779-787, 1998.

[3] McIntyre and McKitrick "Corrections to the Mann et. al. (1998) Proxy Data Base and Northern Hemispheric Average Temperature Series", Energy & Environment, Volume 14, Number 6, 1 November 2003 , pp. 751-771(21)