AUTHOR:

Robert Gerst

Partner, Converge Consulting Group Inc.

rgerst@converge-group.com

BIOGRAPHY:

Robert Gerst is Partner in Charge of Operational Excellence and Research Methods at Converge Consulting Group Inc. a Calgary based consulting practice. He is author of the Performance Improvement Toolkit and a member of the American Statistical Association and the American Association for the Advancement of Science.

PAPER/POSTER:

Paper

SESSION:

Significance--How and Why is it Determined?

TITLE:

Was There an Impact?

ABSTRACT:

Impact assessment identifies the likely impact of actions on the "physical-chemical, biological, visual, cultural and socio-economic components" of the environment". This requires a comparison of before and after conditions. Sophisticated designs may also encompass comparisons with targets and control groups, but there is no escaping the basic requirement of making comparisons in any and all impact assessments.

The analytical challenge is deciding whether the differences observed in making these comparisons are sufficient to conclude the impact was important. Where impact assessments rely on quantitative methods, the dividing line between important and not important is often defined through statistical significance and hypothesis testing. Doing so is *lying with statistics*.

Faulty conclusions, destructive decisions and misguided policy are the inevitable consequences—but what are the solutions? Replacing statistical significance with material significance and using control charts to make quantitative comparisons.

DESCRIPTION:

Impact assessment practitioners use statistical techniques like significance and hypothesis testing to help reach conclusions concerning policy or program impact. But statistical significance can't do that. A move to material significance is required.

Was There an Impact?

Impact assessment identifies the likely impact of actions on the "physical-chemical, biological, visual, cultural and socio-economic components" of the environment.[1] It guides decision making by examining, "the difference between what would happen with a proposed action and what would happen without it".[2]

Sometimes it pays to admit the obvious. Impact assessment requires a comparison of conditions before and after. Sophisticated designs may also encompass comparisons with targets and control groups, but there is no escaping the basic requirement of making before and after comparisons in any and all impact assessments.

The Analytic Challenge

The analytical challenge is deciding whether the before and after characteristics differ sufficiently to conclude the impact was important, that is, of material significance[3]. This is a challenge because any two characteristics or measurements will differ from one another. Survey something on Monday and the results will be different when resurveyed on Tuesday. That doesn't mean the difference is important. There will always be differences between: before and after measures, treatment to control group results, or treatment group results to targets. Because of this, measured differences are themselves, no evidence of impact.

A dividing line, defining whether a difference is large enough to represent something important, is required. Otherwise, everything action can be said to have a 'significant' impact. Impact assessment practitioners must decide where to set this line. Where assessments rely on quantitative methods, the dividing line is often defined through statistical significance and hypothesis testing. Doing so amounts to *lying with statistics*.

This is because impact assessment makes inferences about the cause and effect system or process. It, therefore, belongs to the *analytic* class of scientific studies. These are contrasted with *enumerative* studies that describe populations (as opposed to processes).[4] Data from an enumerative study may be used in an analytic study, but the methods used to analyze the data will be different. Statistical significance and hypothesis testing are useful tools in enumerative studies. They are useless in analytic studies, yielding inaccurate and misleading conclusions.

The questions decision-makers, policy analysts, and the public want answered, and typically believe they are buying through impact assessment, are analytic. Specifically, the likelihood some action (H)

---

[1] What is Impact Assessment?, International Association of Impact Assessment. http://www.iaia.org/publicdocuments/special-publications/What%20is%20IA_web.pdf

[2] ibid.

[3] Materially significance means practical importance—equivalent to biological, ecological, social or economic significance.

[4] For more on the distinction between analytic and enumerative studies see, *On Probability As a Basis for Action*, Edwards Deming, The American Statistician, Vol. 29, No.4, 1975, pp 146-152

had, or will have, an impact (O). This is expressed formally as P(H|O)—the probability of the hypothesis given the observations.

Statistical significance and hypothesis testing answer a different question—P(O|H). This is the probability of an observation assuming a hypothesis. Its common expression is the phrase, "What are the chances of this happening?". For example, what are the chances that Fort Chipewyan has the cancer rates it does (O) assuming cancer incidence is distributed randomly (H)?

Cancer Rates in Fort Chipewyan

This is not a hypothetical question. The impact of oil sands development on the people and environment of northern Alberta is a flash point for the global environmental movement. Considerable impact assessment work is being done in the area. Billions of investment dollars are at stake. Of special interest recently, concern among residents of high cancer rates.

To date, the highest profile study on this issue is *Cancer Incidence in Fort Chipewyan, Alberta 1995-2006,* a joint project of Alberta Health Services, Nunee Health Board Society, Alberta Health and Wellness and Health Canada,[5]. The report clearly establishes its purpose and defines its' dividing line stating:

"The purpose of the investigation is to determine if there is an elevated rate of cholangiocarcinoma in Fort Chipewyan and whether there is an elevated rate of cancers overall in Fort Chipewyan."

"For the conclusions of this investigation, an increase was considered statistically significant if there was less than a 5% chance of observing the same number or a higher number of cancers in that community."

In other words, the purpose of this study is answering the enumerative question of P(O|H). Impact is defined by statistical significance, specifically, when the probability of obtaining the result due to chance are less than 5%. This limit was not reached, leading the report to conclude:

"The observed cases of cholangiocarcinoma and colon cancer during the period of investigation (1995-2006) are within the expected range of cancer occurrence."

In other words, the cancer rates were not statistically significant.

But so what? This is an accurate answer to the wrong question. The study conclusions provide the probability of obtaining a result (*observed cases*) assuming a hypothesis (*expected range of cancer occurrence*). In other words, the study determined P(O|H). But what people want to know is whether living in Fort Chipewyan (H) leads to higher cancer rates (O) or P(H|O).

A Statistical Confidence Game

When the Fort Chipewyan study was launched, the provincial government assured residents their concerns would be investigated. By the time it was published, however, the study noted:

"The study was not designed to determine whether living in Fort Chipewyan elevated cancer risk. "

---

[5] Cancer Incidence in Fort Chipewyan, Alberta 1995-2006. Alberta Cancer Board, Division of Population Health and Information Surveillance, Feb 2009 http://www.albertahealthservices.ca/rls/ne-rls-2009-02-06-fort-chipewyan-study.pdf

In the context of the massive political, social, environmental, economic and health issues involved, few noticed that the study failed to address the concerns of residents. What was investigated was of concern to no one--the probability that observed cancer rates were consistent with the various assumptions concerning an underlying statistical distribution.

This is a statistical version of the shell game. A statistical con, whereby policy makers and the public think they are buying an answer to one question, but receive an answer to a subtly different and far less important one. In science, this con is called the fallacy of the transposed conditional. The fallacy occurs whenever;

- enumerative analysis techniques used to describe populations are used to answer analytic questions concerning processes or systems, or
- statistical significance and hypothesis testing is used to determine findings of practical importance or scientific/material significance,

that is, whenever P(O|H) is used to represent P(H|O).

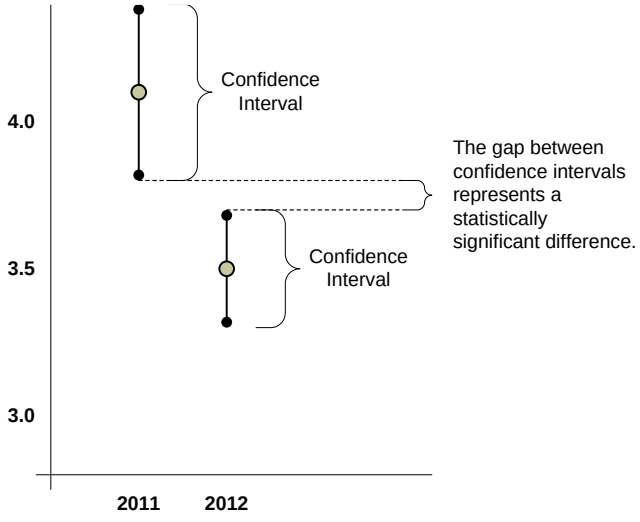
Statistical significance and hypothesis testing tell us nothing about our observations in the real world. Rather, they inform us about the quality or resolution of the measurement system used to make the observations. Thus, when using statistical significance as a stand-in for material significance, we are conducting science by pun, and the joke is on those believing the results. In this case, concluding there is no difference in cancer rates simply because the measurement system wasn't good enough to measure it. That's like concluding there are no dangers on the road when driving at night with the headlights off.

Measuring Water Quality

To better understand the role of statistical testing, let's take an example of evaluating the water quality of a lake. Specifically, estimating the amount of a pollutant. This is an enumerative study because we are describing (estimating a parameter) of a fixed population (the lake). A sample is taken and we obtain a result such as 3.5 ppm +/- 0.2 ppm 19 times out of 20. Perhaps last year's estimate is 4.1 ppm +/- 0.3 ppm 19 times out of twenty. What can we conclude from this?

We can conclude our measurement system had sufficient resolution to detect differences in pollutant levels between this year and last. Differences we already knew where there (to some decimal place). Again, the conclusion concerns the measurement system, not pollutant levels.[6]

Graphical representation of statistically significant differences in water quality.



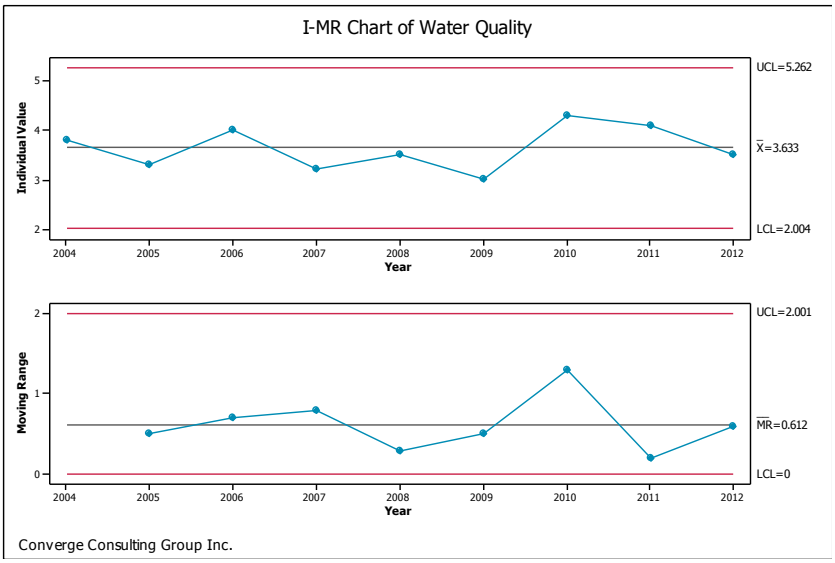
[6] This is true regardless of specificity or sensitivity (power) of the statistical test. Both get the analysis backwards, calculating P(O|H) and making conclusions concerning the measurement system rather than the cause and effect system.

But are these differences trending up? We don't know. Down? Still don't know. Does the difference between this year and last represent something of practical importance? Of biological or ecological significance? Does the difference represent some materially significant shift? No idea. Nothing in the results of the enumerative study, nothing in statistical significance or hypothesis testing, can tells us.

An analytic study is required. What does that look like? Like this. You may recognize it. It's a control chart, in common use in statistical process control applications in industry.

A common misconception is that control charts are limited to industrial, particularly manufacturing, applications. They are, rather, data analysis tools, designed specifically for analytic studies--analyzing data to make conclusions and inferences about a process rather than a population. In other words, control charts help answer P(H|O) They identify whether something had an impact of material significance. Without going into the details of control chart preparation, let's examine how to interpret the chart.

I-mR Control Chart analysis of water quality



Focus on the upper chart as it contains the data of interest, specifically, pollutant levels. The vertical (Y) axis represents pollutant levels and the horizontal (X) axis represents time. The first interpretative step is taking advantage of is the Inter-ocular Trauma Test (ITT). If there is something significant (materially) going on, it should hit you right between the eyes. Despite the statistically significant difference in water quality between 2011 and 2012, these two years don't look unusual in the context of data from previous years presented in the control chart. In other words, while there's a statistically significant difference in pollution levels between the two years, there's no materially significant difference between 2011 and 2012.

We can get a little more sophisticated in our analysis. Lines on the chart corresponding to UCL=5.262 and LCL=2.004 represent the upper and lower control limits. Any data points going above the upper control limit, or below the lower control limit, would be a signal of a special cause of variation within the system--something of material significance. But that doesn't occur here. In fact, water quality in this system has been stable for the past 8 years. There are other tests that can be used to identify material significance in the control chart. Some of these are presented in Control Chart Patterns.

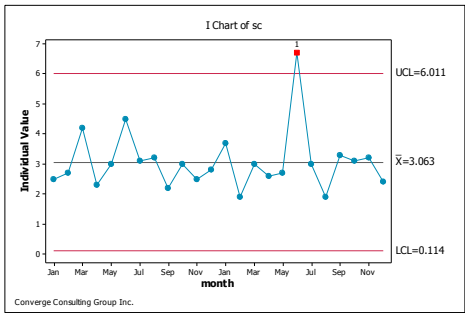
A quick lesson from the control chart is that the before 'baseline' is just that, a line, not a point. The data line in the control chart *is* the baseline and the normal operating parameters are the three lines

representing the mean and upper and lower control limits.[7] So if you haven't got a control chart, you don't have a baseline, and without a baseline, you don't have a 'before' to compare to the 'after'.

Control Chart Patterns

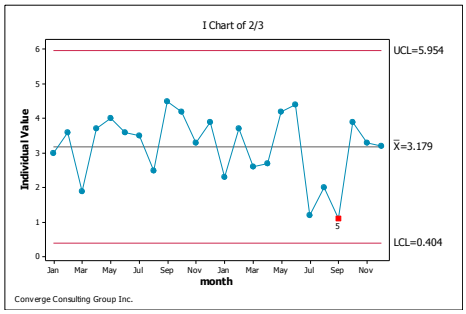
Individuals Control Chart

Type of Pattern/Impact



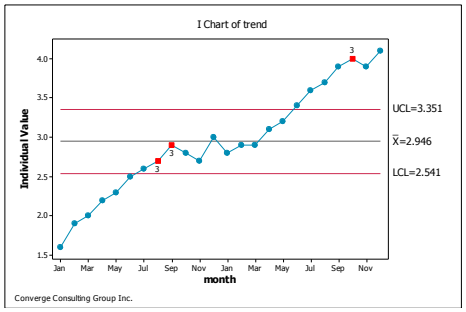
Point Beyond Control Limit

Any data point going above the Upper Control Limit or below the Lower Control Limit is a signal of a likely special cause—something of material significance is 'impacting' the system. In this case, the root cause came and went, impacting the system in June.



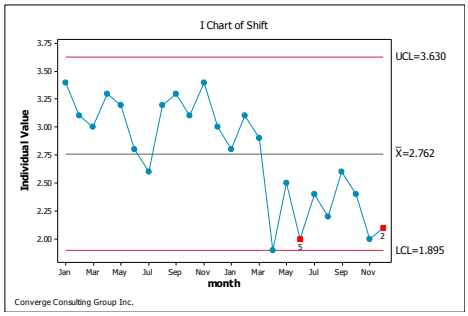
Two of Three Rule

Two of three data points lie close to a control limit, specifically, between 2 and 3 sigma. Also a signal of a special cause of variation. Something has significantly impacted the system.



Trend Test

Seven or more consecutively increasing or decreasing data points. Doesn't necessarily confirm a trend, but rather confirms the lack of one. If you don't have seven consecutively increasing or decreasing data points, you don't have a trend.



Shift Test

Eight or more consecutive data points to one side of the average line or the other. Indicates a special cause has shifted system behavior from one performance level to another. This patter would be displayed if some program was successful at materially reducing say, recidivism rates, pollutant levels or accidents.

[7] Normal here does not mean a normal distribution but rather the collective impact of numerous causes and conditions that come together in different ways and in different years to produce variation in pollutant levels.

Conclusions

When attempting to assess impact quantitatively, statistical significance is not just the wrong tool for the job, it's;

" a virus infecting; (i) academic research in psychology, biology, ecology, education, health, economics, medicine, (ii) industry research, including market and customer research, process improvement, operational & organizational analysis, employee satisfaction research and (iii) government research, including public reporting, policy and program evaluation."[8]

The virus is spreading and is "*Why Most Research Findings Are False*", including those using quantitative methods in impact assessment.[9]

The problem with statistical significance is that it answers the enumerative problem of P(O|H). But people and policy makers typically want answers to the analytic question of P(H|O). In using one for the other we are engaged in a logical fallacy of the transposed conditional. Using statistical significance and hypothesis tests to represent material significance, importance, is *lying with statistics*.

The cure is the control chart. In any instance where statistical significance and hypothesis tests are used, control charts (and related techniques) can and should be used instead. Not because of a preference for using one statistical tool for another, but because control charts are the appropriate tool for analytic studies that make inferences about the cause and effect system.

---

[8] Robert Gerst, *Significance, statistical and otherwise*, in publication. Significance. The Royal Statistical Society and The American Statistical Association.

[9] John P. A. Ioannidis PLOS Medicine. Why Most Research Findings Are False http://www.plosmedicine.org/article/info:doi/10.1371/journal.pmed.0020124