
**Correlation vs. Causation - and What Are the Implications for Our Project?** By Michael Reames and Gabriel Kemeny

In problem solving, accurately establishing and validating root causes are vital to improving processes. A common issue with teams of subject matter experts is they are biased, and are easily drawn to explanations about likely cause-and-effect relationships. What sets these teams apart from unenlightened approaches to root-cause analysis is that they seek statistically significant relationships to suspected causes and unsatisfactory effects. The unenlightened approach can be described as the intuitive or "seat of the pants" approach; one way of visualizing this is the statement "Don't bother me with the data – I know what's going on, and how this problem should be resolved."

It is very useful for process improvement teams to know the difference, then, between correlation and causation. In so doing, the team realizes that correlation is not enough, and that an effective problem-solver seeks to understand causation.

For purposes of this article, an adequate dictionary definition of correlation is:
"a relation existing between two or more things that tend to vary, be associated, or occur together in a way not expected on the basis of chance alone." These "things" can be described mathematically: hence, correlation refers to how closely two sets of information or data are related.

Causation goes a step further by defining "a relationship in which one action or event is the direct consequence of another."

Meeting only one of these standards is not sufficient for validation. A typical approach is using one's experience with a process to suggest a cause-and-effect relationship. Many leaders have achieved success by being decisive, often without good data to back them up. Short of information (data) demonstrating a significant statistical correlation, this knowledge is simply intuitive – "seat of the pants" – expertise. Solutions based on intuition and not a demonstrated correlation may improve the process, but only to the extent that the intuitive guess is correct (i.e., has underlying data supporting it, but which was never collected). In other words, sometimes our intuition is right; but not always.

On the other hand, a strong mathematical relationship (correlation) between two variables does not by itself confirm that one causes the other. For an example, consider the fallacy that storks bring babies. Indeed, in nineteenth-century northern Europe villages and towns, there was a remarkably strong and continuing correlation between the local stork population and the birth of babies. This led to the enduring myth that storks bring babies. How else to explain the correlation?

**Figure 1: Storks bring babies**

Although the fallacy of this myth is obvious, it is impossible to refute the correlation. Perhaps there is an underlying cause that creates both effects, i.e., that increases the stork population and baby births.

A bit of further knowledge (data gathering) is enlightening:

- Northern Europe experiences cold, harsh winters
- Storks prefer protected, relatively warm nesting sites
- Dwellings almost always included fireplaces with chimneys atop the roofs
- The roofs near the chimneys provided protected, warm nesting sites
- As families grew (babies), more dwellings were built (hence, more fireplaces and chimneys)
- Storks were attracted to the abundance of new nesting sites

Hence, an increase in homes provided for the increases in the human population, while at the same time drawing storks into the same areas to take advantage of good nesting sites.

This simple example demonstrates the risk of automatically assuming that a certain indicator has an impact on another indicator. Many times variables and indicators have mutual relationships that are easily proven mathematically (correlation). Even so, that does not necessarily mean that one thing has had an effect on the other.

Many statistical tools assist in establishing a statistical correlation. Scatter diagrams and regression are techniques that visualize the relationship between pairs of variables

(Figures 2 and 3 are examples of scatter diagrams). The strength of the relationship is the correlation coefficient "r" which ranges from +1 to -1:

| | |
|---|---|
| Perfect positive relationship | **+1** |
| No linear relationship | **0** |
| Perfect negative relationship | **-1** |

Hypothesis testing helps us to handle uncertainty objectively. These include t-test, ANOVA (Analysis of Variance), Proportions test, Chi-Square analysis, and Logistic Regression. Each test is based on the characteristics of the gathered data. Applying the appropriate test allows us to confirm or disprove assumptions and to control our risk of making wrong decisions. They help teams to make fact-based decisions about process improvements, rather than intuitive guesses.

In the example of storks and babies, a verified statistical correlation between two variables X and Y is not by itself conclusive. Rather, it may lead to a number of possible alternative conclusions:

1. X affects Y (i.e., the cause creates the effect)
2. Y affects X (i.e., the effect creates the cause)
3. X interacts with Y (each affects the other)
4. Other (or unknown) variables may affect both X and Y
5. A combination of some or all of the above
6. A pure coincidence (highly unlikely)

Thus, although it is exciting for a process improvement team to discover a significant statistical correlation, the team needs to investigate further for possible causation, even if the "statistically correlated" relationship has a large effect.

Proven past process or scientific knowledge (by subject matter experts) may help the team eliminate some (or all) of these alternative conclusions. Too frequently, however, there are cases where there is no expert knowledge available regarding the factors that may affect a particular response. If such knowledge is unavailable, the team can use the scientific method to acquire new knowledge and to discover causation (or to refute it) through experimentation.

The technique known as Design of Experiments DOE) allows the experimenter to manipulate controllable factors (independent variables) at different levels to see their effect on some response (dependent variable). By manipulating inputs to see how the output changes, he/she begins to understand and model the dependent variable (Y) as a function of the independent variable (X).

In summary, just because two things occur together does not prove that one caused the other, even if it seems to make sense. Our intuition often leads us astray when it comes

to distinguishing between causality and correlation. Validation of a root cause is achieved only when two standards are met:

1. There is a statistically significant relationship between the suspected root cause and effect (i.e., a correlation); and
2. Knowledge of the process assures that a causal relationship is feasible and likely to exist.

ProcessGPS has compiled many examples of interesting correlations. All have been drawn from a general Bing/Google search. Consider:

| Correlation | Possible Explanation |
|---|---|
| The cycle time for a step in a time-sensitive process is highly correlated with overall cycle time to the customer. | The particular process step requires 80% of total cycle time in the process. Subject matter experts conclude that this single process step drives overall process cycle time. **CONCLUSION: X AFFECTS Y** |
| The number of literature club meetings attended by a student is correlated with her English grades. We assume that frequency of attendance in club meetings (X) affects the grades (Y). | Many students are encouraged to attend club meetings because of their academic success. Because they get good grades, they get more invitations. **CONCLUSION: Y AFFECTS X** |
| The number of Nobel prizes won by a country (adjusted for population) correlates well with per capita chocolate consumption (the New England journal of Medicine). | Thus far, no adequate explanation (see Figure 2). **PRELIMINARY CONCLUSION: X AND Y MAY AFFECT EACH OTHER; OR AN UNKNOWN VARIABLE MAY AFFECT BOTH X AND Y** |
| The more fire engines sent to a fire, the more damage is done. | Increasing numbers of firefighters respond to larger fires; more water and fire retardants are used; hence, more damage. **CONCLUSION: ANOTHER VARIABLE AFFECTS BOTH X AND Y** |
| As vaccinations increase in the U.S., so have instances of diagnosed autism. | This correlation has been widely dismissed by public health experts. The rise in autism rates is likely due to increased awareness and more expert medical diagnosis, or any of other possible factors that have changed over the past 50 years. **CONCLUSION: OTHER VARIABLES (PERHAPS SEVERAL) AFFECT BOTH X AND Y** |
| Fresh lemons imported to the USA from Mexico and total US Highway Fatality Rate. | No adequate explanation. See Figure 3. Sources: U.S. HGTSA, DPT HS 810 78- U.S. Dept. of Agriculture. **PRELIMINARY CONCLUSION: AN UNKNOWN VARIABLE MAY AFFECT BOTH X AND Y** |
| In early elementary school years, astrological sign is correlated with IQ, but this correlation weakens with age and disappears by adulthood. | **PRELIMINARY CONCLUSION: A PURE COINCIDENCE?** |

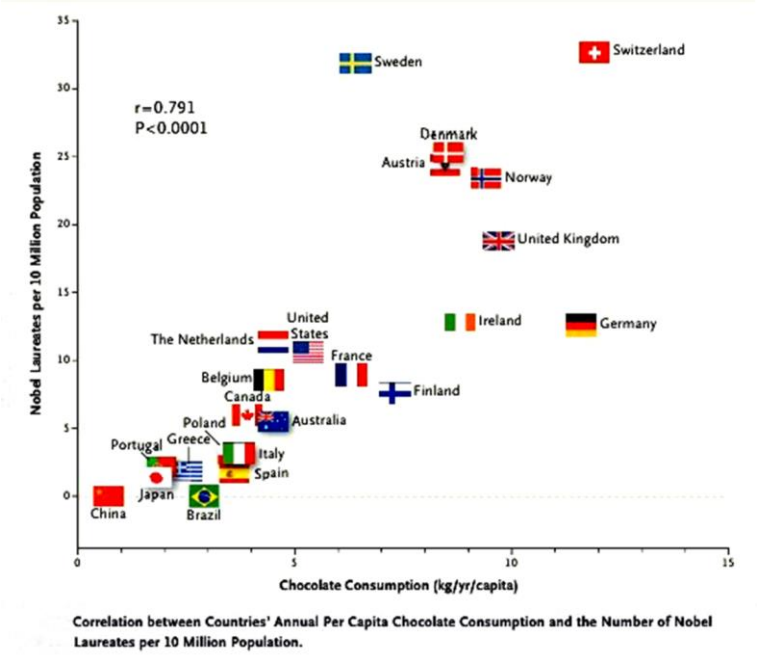


**Figure 2: Annual Per Capital Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population**

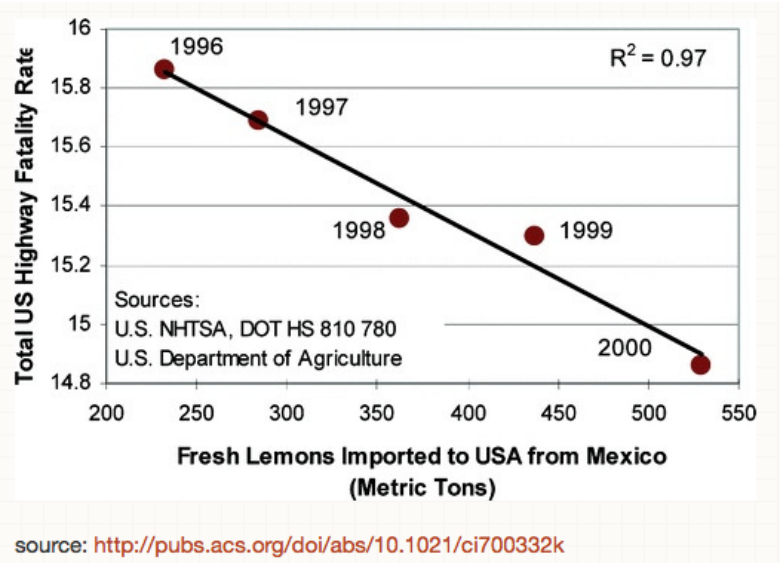


**Figure 3: Lemon Importation from Mexico (Metric Tons) vs. U.S. Highway Fatality Rate**

ProcessGPS
*Guaranteed Project Success*

**Gabe Kemeny and Michael Reames**
**Partners and Co-Founders, ProcessGPS, LLC**

gabe@ProcessGPS.com (801) GPS-0606
michael@ProcessGPS.com (484) 2020-GPS

Website: www.ProcessGPS.com

*Your guaranteed success defines ProcessGPS' success*