Guiding Beacon

Using statistical engineering principles for problem solving

QUALITY PROFESSIONALS are often faced with solving major organizational problems such as: "Customers are complaining about the quality of our product and returning it"; "Our major process is producing an unacceptable amount of defective product"; or "The regulatory agency has identified a major environmental problem associated with one of our operations."

How should quality professionals approach such problems, which are clearly not textbook with one correct answer? Where should they begin the problem-solving effort? What should be considered? How can the projects be set up for success? The fundamentals of statistical engineering can provide valuable guidance for these types of complex problems.

Statistical engineering has been defined as: "The study of how to best utilize statistical concepts, methods and tools, and integrate them with information technology and other relevant sciences to generate improved results."[1] Applications of this discipline produce improved results because statistical engineering is grounded on sound underlying principles that address the critical elements of effective problem solving. In short, the key elements of the principles of statistical engineering are (see Figure 1):

• Proper understanding of the problem context.
• A well-defined strategy for problem solution.
• Evaluation of the pedigree of the associated data and information.
• Integration of sound subject matter knowledge with data analysis.
• Sequential approaches involving the testing of existing hypotheses and development of new hypotheses.

Understanding problem context

Problem context is everything we know about the problem, including its history, what has been tried before, the technology involved and political considerations at play. Too often, data analysis begins with the data. This can be seen in various data analysis competitions, such as those on kaggle.com. Keep in mind, however, that the data are not the problem; the problem is the problem. That is, we should view data as a "how," and the original problem trying to be solved as the "what."

Once we are clear on the problem we are trying to solve and its context, we can determine the type and amount of data needed to solve it. Conversely, if we already have data, clarification of the problem helps determine how the data can be best used to solve the problem. "Data have no meaning in themselves; they are meaningful only in relation to a conceptual model of the phenomenon studied,"[2] wrote George Box, Bill Hunter and Stu Hunter.

In addition, the best technical or business solution is not always the best statistical solution. For example, we may determine from initial analysis of existing data that they are not appropriate or sufficient for solving the problem at hand; additional, better quality data are needed. Performing sophisticated or detailed analysis of the current data would simply waste time at this point.

In other cases, a simple analysis is all that is needed because the answer is obvious from basic graphs. The bottom line is that the context of the problem, not statistical metrics, determines the best business solution and the level of sophistication needed.
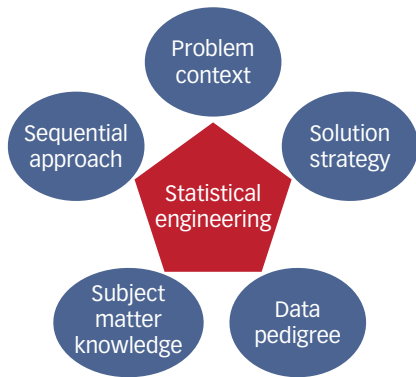
Well-defined strategy

Some practitioners have a favorite tool, whether it is multiple regression, time-series analysis or a nonparametric method.

Underlying principles of statistical engineering / FIGURE 1

Unfortunately, favorite tools can be more of a hindrance than a help to problem solving. Practitioners can fall into the trap captured eloquently by the saying: "If all you have is a hammer, every problem looks like a nail." In other words, practitioners may use favorite tools even when they are not the best approach, or when they are not appropriate at all.

Similarly, the approach of trying various modeling approaches and picking the one that maximizes a quantitative metric, such as R^2 or root mean square error (RMSE), can provide some insight, but rarely produces an actionable model. Most problems are too complex to be adequately reduced to maximizing one quantitative metric. Rather, an overall strategy or plan of attack is needed for complex problems. A strategy is a plan of action designed to achieve a major goal. In other words, it is a high-level game plan to win. For example, some problems have a known solution, and we just need to figure out how to deploy the solution. Others have no known solution, and root causes must be identified and evaluated.[3] The nature of the problem should guide our approach, which requires strategic thinking rather than tools-based thinking.

Unfortunately, we have found that the word "strategy" is rarely used in technical textbooks. Conversely, good problem-solvers develop a plan of attack to match the particular problem, and of course, need to remain flexible to revise their strategy when the unexpected turns up, as is often the case.

For example, authors Box, Hunter, and Hunter provide an overall approach to using design of experiments and model building in a sequential manner to empirically optimize processes.[4] This approach, often referred to as response surface methodology, is not simply a collection of designs and models, but an overall strategy for attacking such problems. Quality and statistical professions need more of these types of strategies.

Data pedigree

Many statistics textbooks discuss the importance of sample size on statistical analyses and provide formulas for determining appropriate sample size. Unfortunately, very little is typically said about data quality. The quote from Box, Hunter and Hunter noted earlier is a welcome but rare exception.

The assumption seems to be that "all data are created equal." If only that were true. Practitioners who have had to collect their own data know how challenging it can be to collect good data. Missing values and variables, poor measurement processes and collinearity between the independent variables are just a few problems typically encountered. No amount of data or sophisticated data mining algorithms will salvage a bad data set.

The key point is that rather than jumping into analyses, practitioners should always carefully consider data quality first: where it came from, how it was collected, who collected it, over what time frame, the measurement system used and the associated science and engineering. This type of information is called the data pedigree because it describes the background and history of the data, much like a show dog's pedigree documents its credentials. Data should always be considered guilty until proven innocent.

In many cases, the data are sufficient to answer some questions, but not to solve the overall problem. More data, collected differently, are often required, based on analysis of the original data. Further, the sophistication of any models developed should be based on the needs of the problem and the data pedigree, and should never be more complex than can be adequately supported by the current data.

Subject matter knowledge

When analyzing data, it can be tempting to put data into the computer, push buttons on sophisticated software packages and believe the resulting output. In addition to the potential problems noted earlier, this approach also ignores the key principle that data should always be interpreted in light of our existing subject matter or "domain" knowledge. This is everything we know about the underlying science or theory of the process of interest. Without a good understanding of the process that produced the data, we are susceptible to making egregious errors in analysis.

George Cobb and Stephen Gelbach illustrated this point with data on heart disease among pipe, cigarette and cigar smokers.[5] A straightforward analysis of the data indicates that cigar and pipe smokers have higher rates of heart disease than cigarette smokers, and that this difference is statistically significant. Does this seem surprising?

A closer look with additional data reveals that pipe smokers are much older than cigarette smokers on average. After the age difference is taken into account, the analysis clearly shows that cigarette smoking is a more dangerous contributor to heart disease than is pipe smoking.

Similarly, while the first author (Hoerl)

Practitioners who have had to collect their own data **know how challenging** it can be **to collect good data.**

was working for Scott Paper Co., his analysis of paper towel data indicated that paper towel strength was positively correlated with absorbency—how much water the towel could hold. When he presented this analysis to Scott engineers, he was laughed off the stage.

It turned out that the fundamental science of papermaking shows that strength and absorbency of towels are inherently negatively correlated. With tail between his legs, Hoerl reevaluated the data pedigree. In this case, the data happened to come from two different papermaking processes. After the data were stratified by papermaking technology, the expected negative correlation was apparent.[6]

In both cases, naive analysis of the data presented produced not only incorrect answers, but conclusions that were the exact opposite of the correct conclusions. Good subject matter knowledge not only helps to avoid such blunders, but also can guide us in identifying the most appropriate data to solve the problem, and then help us interpret it correctly.

Of course, sometimes data will challenge our subject matter knowledge and require us to rethink hypotheses. This is how learning occurs and how data analysis can spark creativity. New theory still must build on previous theory, however. Most old sayings point us in the right direction. However, the saying, "The data speak for themselves" is dead wrong. Data never speak for themselves, but require sound subject matter knowledge to be properly analyzed and interpreted.

Sequential approaches

You might get the impression from statistics textbooks that proper analysis of one data set satisfactorily addresses most problems. Unfortunately, resolving tough problems is usually more difficult than that and requires a sequential approach using multiple data sets and statistical methods. As noted previously, an overall

strategy is required, and this strategy typically incorporates multiple phases and data sets.

For example, the Framingham Heart Study (FHS)[7] was initiated in 1948 and continues to collect data and analyze it today. The FHS has produced major breakthroughs in identifying risk factors for heart disease, such as smoking, high blood pressure, high blood cholesterol and obesity.

By using a sequential mindset, practitioners get to use hindsight to their advantage. That is, after reviewing the initial data, they can quickly see which hypotheses have been validated and what surprises there might be. The surprises lead to reevaluation of existing subject matter knowledge, as noted previously, and to new theories.

New sets of data, specifically identified to evaluate the new theories, then can be collected. The cycle continues with new knowledge gained at each step in the process. Through a sequential approach, data collection and analysis fit into the overall scientific discovery process. Statistical methods test existing hypotheses and also help generate new hypotheses—that is, statistical methods properly applied in a sequential approach spark innovation and creativity.

Guidance for problem solving

Solving large, complex and unstructured problems is difficult and goes well beyond solving typical textbook problems. Guidance is needed so practitioners don't reinvent the wheel with each new problem. The fundamental principles of statistical engineering can guide our efforts and significantly increase our learning curves and chances of success. QP

REFERENCES AND NOTES
1. Roger W. Hoerl and Ronald D. Snee, "Statistical Thinking and Methods in Quality Improvement: A Look Toward the Future," *Quality Engineering*, Vol. 22, No. 3, 2010, pp. 119-129.
2. George E.P. Box, William G. Hunter and J. Stuart Hunter, *Statistics for Experimenters,* John Wiley & Sons, 1978, p. 291.
3. Ronald D. Snee and Roger W. Hoerl, "One Size Does Not Fit All," *Quality Progress*, May 2013, pp. 48-50.
4. George E.P. Box, J. Stuart Hunter and William G. Hunter, *Statistics for Experimenters,* second edition, John Wiley & Sons, 2005.
5. George W. Cobb and Stephen Gelbach, "Statistics in the Courtroom: United States v. Kristen Gilbert," *Statistics: A Guide to the Unknown,* Thomson/Brooks Cole, 2006.
6. For this analysis on paper towel strength, read Roger W. Hoerl and Ronald D. Snee's *Statistical Thinking: Improving Business Performance,* second edition, John Wiley and Sons, 2012, p. 171.
7. For more information about Framingham Heart Study, visit www.framinghamheartstudy.org.

ROGER W. HOERL is a Brate-Peschel assistant professor of statistics at Union College in Schenectady, NY. He has a doctorate in applied statistics from the University of Delaware in Newark. Hoerl is an ASQ fellow, a recipient of the ASQ's Shewhart Medal and Brumbaugh Award and an academician in the International Academy for Quality.

RONALD D. SNEE is president of Snee Associates LLC in Newark, DE. He has a doctorate in applied and mathematical statistics from Rutgers University in New Brunswick, NJ. Snee has received ASQ's Shewhart and Grant Medals. He is an ASQ fellow and an academician in the International Academy for Quality.