

Make Data Matter

A five-pronged approach to analyze process data

IS DATA ANALYSIS an art or a science? Arguments exist for both sides, and many people simply come down in the middle. In my mind, I believe it's both.

Regardless of which view you take, the discussion misses a critical element—the need for an explicitly articulated strategy for data analysis. In fact, the various attitudes toward the nature of data analysis often imply unreflective strategies.

Partisans of data analysis as an art simply might look at the data, manipulate it based on their intuition and experience, and proceed confidently to extract what they believe is useful information. The more scientific folks, with perhaps too much faith in numbers, go straight to statistical software and do some indisputable number crunching.

Those who stand on middle ground—possibly the great majority of practitioners—do a little of both: rely on their insight to manipulate the data, run the numbers, do some further manipulation and rerun the numbers until they achieve what they believe is a satisfactory result. All of those approaches are likely to produce questionable results in terms of what the analysis addresses and the significance of the results.

Five activities

Practitioners can avoid the pitfalls of these unreflective or ad hoc approaches by adopting a clearly articulated, proven strategy for analyzing process data and systematically following that strategy.¹ Such a strategy entails five essential activities:

1. Understanding the context of the analysis.
2. Examining the pedigree of the data.
3. Graphically representing the process.
4. Graphically representing the data.
5. Statistically analyzing the data.

Note that these are iterative, as opposed to sequential, activities. Depending on the circumstances, the order of some of these activities may shift.

For example, in the mutually dependent iterations of this approach, the graphical representation of the process may precede the examination of the pedigree. In any case, most of these activities look forward and backward. The examination of the data's pedigree—where it came from and how it was collected—may drive the analyst back to a fuller exploration of the context of the process to fill out that

“The key to success lies in intimately **knowing the data from the context of the process**, graphically representing it and formulating a model.”

important results that are ready to be fully and persuasively reported.

This approach offers at least three distinct advantages over less structured approaches. First, it is repeatable—it can be used

in any situation that calls for the greater understanding of a process. Second, like sound processes themselves, it's robust—flexible enough to encom-

pass the wide variation of particulars to be found in different situations. Third, and most importantly, it's more likely to produce useful results.

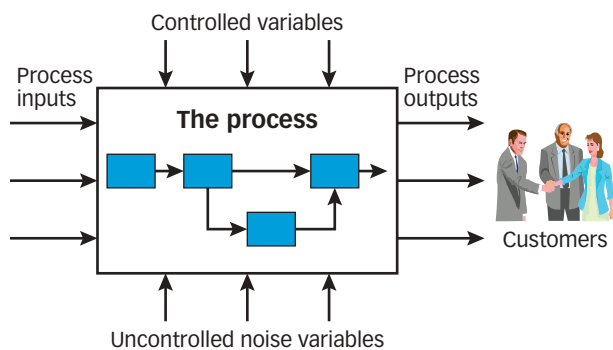
Understanding the context

It's difficult to know precisely how to proceed until you ask the most basic of questions: What is the purpose of the analysis? Are you trying to confirm a hypothesis? For example, a manufacturer that uses raw materials from two different vendors suspects that differences in quality are causing defects in the finished product. Data analysis can confirm or disconfirm the hypothesis and, in this example, identify the offending vendor. Such contexts call for what is sometimes referred to as confirmatory data analysis.

Alternatively, let's say you're trying to solve specific problems, the causes of which you do not understand. For example, a chemical process is producing unacceptable variations in purity from batch to batch. Or a business process, like a bank loan approval process, is taking far too long to complete. Or, perhaps a distributor's percentage of on-time deliveries is fluctuating widely. These contexts call for exploratory data analysis, which must first have a hypothesis to test.

In confirmatory and exploratory analyses of a process, the goal is the same: find the inputs and the controlled and uncon-

A process and its variables / FIGURE 1



pedigree. But the pedigree of the data also points to how the process should be graphically represented. That, in turn, could retrospectively suggest the need for additional types of data and prospectively affect the graphical representation. By engaging iteratively in these activities, you can arrive at

trolled variables that have a major impact on the output of the process.²

Examining the pedigree

Data analysis begins with a data table, which is either provided to or constructed by the analyst. In either case, you should always question the data because data can be, among many other things:

- **Incorrect:** Some of the information is wrong—for example, when someone monitoring a process records the data incorrectly or a measurement device is faulty.
- **Irrelevant:** Some of it is the wrong information—for example, when data on the wrong variables are captured.
- **Incomplete:** Crucial information is missing—for example, when data on an important variable are missing.
- **Misleading/biased:** Data points you in the wrong direction for analysis—for example, when an important variable has been examined only over a short time, thus making it appear to be a constant.

An understanding of the context of the process can guard against these errors, but the context alone is insufficient. Given these and the many other shortcomings that can undermine the value of the data, it is absolutely critical to understand the pedigree of the data—where it came from and how it was collected.

For example, consider a batch manufacturing process in which a sample is taken every shift and carried to an analytical lab where it is tested for purity, and the results are recorded. Thus, the data trail is: Production process ► sampling process ► testing process ► data-logging process.

To understand the resulting data, it is necessary to understand this data trail and the production process parameters. That is the pedigree of the data.

Incomplete understanding of the data's pedigree can lead you down wrong analytical trails. Suppose, for example, a pharmaceutical company is experiencing differences in yield from batch to batch

of a product because of the properties of the raw materials supplied by a vendor. Although the properties for each batch of raw materials are within specifications, the yield nevertheless varies unacceptably. The analyst has been given a data table that includes the properties of the raw materials for each batch of product under consideration. But if the analyst does not know that some raw material batches were analyzed by the vendor's quality assurance lab and some by the manufacturer, then there is a strong possibility the analysis will come up empty. By taking the time to understand the pedigree of the data fully, the analyst can save much frustration and fruitless work.

Graphing the process

A graphical representation of the process shows how the process works from end to end. Such representations fall into two broad categories: flow charts and schematics. A flow chart maps the sequence and flow of the process and often includes icons, such as pictures of a truck to represent a transportation step or smokestacks to indicate a factory.

A schematic representation is designed to exhibit the inputs and the controlled and uncontrolled variables that go into a process to produce its outputs. Both types of representation reinforce one another by suggesting what types of data are needed, where they can be found and how they can be analyzed.

Figure 1 is an elementary schematic representation of a process (such as pharmaceutical, chemical or loan approval). As the analyst knows, the context is unacceptable variations in yield from batch to batch of the finished product. Therefore, "yield" is the key output.

To get an accurate picture of the process again, however, analysts should not simply rely on the context. To find out how the process really works, they should also observe the process first-hand and question the people who operate it. This investigation might also lead the analyst to further

SOME GUIDING PRINCIPLES

- The process provides the context for the problem being studied and the data being analyzed.
- Know the pedigree of the data—the who, what, when, where, why and how of its collection.
- Analysis is defined by how the data were generated.
- Understand the measurement system as well as the process.
- Be aware of human intervention in a process. Humans are often a large source of variation.

refine the pedigree of the data—the who, when and why of its measurement and collection.

With yield as the key output of a manufacturing process, the analyst can now graphically represent the process and fill in the blanks with the sources of possible variation that led to the unacceptable variations in yield. For the inputs, sources of variation might be energy, raw materials and different lots of the same raw materials. Controlled variables that go into the process might include things like temperature, speed of flow and mixing time.

In essence, controlled variables are the things that can be adjusted with a knob or a dial. Uncontrolled variables that go into this process may include human intervention and differences in work teams, production lots, days of the week, machines or even heads on the same machine. In the output of the process, variation may result from the measurement system itself.

A good rule to follow when you have, for example, two production lines doing the same thing or two pieces of equipment performing the same task, is to assume they vary until proven otherwise. That's especially true for the human factor. Experience shows that in creating the initial data table and in the graphical representation of the process, the human element is a frequently overlooked source of variation.

In the aforementioned pharmaceutical

ART OR SCIENCE?

What side do you fall on? Is data analysis an art or a science? Post your thoughts at www.qualityprogress.com or e-mail your comments to editor@asq.org.

The analyst must not only **do data analysis that matters**, but also **make it matter.**

manufacturing process, the analyst may overlook that the process includes three shifts with four different work teams on the shifts.

As a result of the observation and investigation that goes into constructing the graphical representation of the process, however, the analyst makes sure the data table records which team produced which batches on which days and that the data are stratified in the analysis. The failure to take that human element into account results in a highly misleading data table and might obscure the ultimate solution to the problem.

Graphing the data

The graphical representation of the process—and the understanding of the possible sources of variation it helps generate—suggests ways in which the analyst can graphically represent the data. Because data are almost always sequential, a run chart is often needed. In our example, the x-axis would register time and the y-axis would register yield. A scatter plot also may be used, with process variables registered on the x-axis and process outputs registered on the y-axis. Other familiar graphical techniques include box plots, histograms, dot plots and Pareto charts.

In using any of these techniques, the goal is to make sure you are exploring the relationships of potentially important variables and preparing an appropriate graphical representation for purposes of statistical analysis. Plotting the data in different ways can lead to insights and surprises about the sources of variation.

Statistically analyzing the data

The statistical analysis of the data, usually with the aid of statistical software, establishes what factors are statistically significant. For example, are the differences in yield produced by different work

teams statistically significant? What about variations in temperature or flow? What about the measurement system itself?

The key to success lies in intimately knowing the data from the context of the process, graphically representing it and formulating a model that includes the comparisons, relationships, tests and fits you are going to study.

Once you have created the graphics and done the statistical calculations, the results should be checked against the model. Does it account for all of the variation? In short, do the results make sense? If so, you can confidently report your results.³

Beyond analysis to action

The final point about reporting the results offers a reminder that analysis goes beyond the exploratory or confirmatory. The analyst also must be able to display and communicate results to decision makers. The most elegant analysis possible is wasted if it fails to communicate and the organization therefore fails to act. Early in my career, I was asked to analyze whether a chemical company's new product had adversely affected animals in safety studies. Personnel in the company's lab insisted the data from the experiments showed adverse effects, and the company should therefore cease development of the product. Analysts on the company's business side had concluded the data showed no adverse effects. My analysis reached the same conclusion, and in a showdown meeting between the business and the lab personnel, I presented my findings. At the conclusion of my presentation, replete with analytical representations of the statistical significance of the data, the lab director remained unconvinced. So I handed him one final graph: a dot plot that, for some reason, I had not included in my presentation.

He looked at the graph and began to think aloud while everyone in the meeting sat silently. He continued to look and talk and look and talk. At last, he said emphatically, "Maybe there isn't a difference."

In the absence of that persuasive graphical representation and model of the data, the company might have ceased production of what turned out to be a valuable and harmless product. The bottom line is that the analyst must not only do data analysis that matters, but also make it matter. **QP**

© Ronald D. Snee, 2008.

REFERENCES AND NOTE

1. Roger W. Hoerl and Ronald D. Snee, *Statistical Thinking—Improving Business Performance*, Duxbury Press, 2002. Chapter 6 of this book describes a systematic approach for doing regression analysis.
2. Ronald D. Snee, "Develop Useful Models—Finding the Critical Few Xs Allows You to Better Control and Optimize a Process," *Quality Progress*, December 2002, pp. 94-96.
3. Hoerl and Snee, *Statistical Thinking—Improving Business Performance*, see reference 1.



RONALD D. SNEE is principal of performance excellence and lean Six Sigma initiative leader at Tunnell Consulting in King of Prussia, PA. He has a doctorate in applied and mathematical statistics from Rutgers University in New Brunswick, NJ. Snee has received the ASQ Shewhart and Grant medals and is an ASQ fellow.



Founded in 1962 and serving many of the world's leading life sciences firms, Tunnell Consulting integrates strategic, technical, process, and organizational skills to design and implement sustainable solutions that exactly meet client needs. With deep industry knowledge, extensive scientific credentials, and superior measurable results, we consistently boost the operating performance of each unique client we serve.

Headquarters
900 East Eighth Avenue, Suite 106 • King of Prussia, PA 19406
ph 610.337.0820 • www.tunnellconsulting.com

King of Prussia, PA
San Diego, CA

Washington, DC
San Juan, PR