IGNORE THE NUMBERS…

JUST THE FACTS

Too much data—and misguided analysis—often can confuse decision makers.

There are five basic rules for properly analyzing data that must be followed to solve problems, bring value and avoid misinterpretation.

Using these five rules will guarantee successful informed decisions that rely on predictive and actionable information.

EMBRACE THE DATA.

Five basics for eliminating poor data analysis **by Gregory C. McLaughlin**

David Malan via Getty Images

Businesses or organizations cannot function without data. The mantra today is to collect enormous amounts of data, which is referred to as big data. Big data have led to big analytics. Executives are spending plenty on analytics. By one estimation, that spending is at $187 billion—and growing.[1] The problem is that big does not always equal better.

For many, big analytics doesn't seem to answer imprecise questions or offer new insights.[2] The root cause of the problem is an improper recognition of the data's inherent behavior related to its predictability and consistency, which applies to large or small data sets and is rooted in failed data analysis.

These failures result in massive losses—for example, IBM estimates data analysis failures cost up to $3 trillion in the United States alone.[3] Poor data analysis is happening at epidemic proportions and is a result of too much data to examine and review, failure to adhere to assumptions or requirements, lack of knowledge regarding risks and inconsistencies, and poor interpretation that brings little value and frequent error.

Appropriate analysis and interpretation are required to solve new problems and gain new answers—without making the same old statistical mistakes.[4] Using the five rules for data analysis presented in this article provides a solution to this epidemic. A familiar example (the Standard & Poor's, or S&P, stock index) will help in understanding these concepts.

Using these rules will guarantee a successful informed decision that relies on predictive and actionable information.

First rule

Numbers describe what it is, not what happened nor what will happen. Numbers alone cannot solve problems or enable the right or best decisions.

Businesses and decision makers may rely too heavily on numbers. If sales drop 10% from the previous week, for example, would you react, investigate or monitor? What is the best advice if the number changes? The best choice is to investigate. Why? You must identify why the value changed because the value alone won't provide an answer. For example, the S&P index ended the day on May 1, 2019, at 2,676.19. What value does this number provide? Except for its descriptive value, a number provides no information on what to expect the next trading day or any day.

Use data—rather than numbers—to make decisions. Data transform numbers into information by explaining their meaning, placing it within a specified time frame, comparing it to a known value or standard, and establishing its importance and priority. Let the facts, experience and instinct direct your decision rather than just emotion, opinion or feelings.

Data should reveal something about the event, issue, problem, threat or opportunity. Be aware of inaccurate, missing, misused and inappropriate data, as well as issues with data entry.[5] Enron used inaccurate financial data to hide its money (solvency) problems[6]—its collapse and loss of shareholder value were well-publicized.

TABLE 1

S&P index for April 2019 (end of the trading day)

2,867.19	2,867.24	2,873.40	2,879.39	2,892.74	2,895.77
2,878.20	2,888.21	2,888.32	2,907.41	2,905.58	2,907.06
2,900.45	2,905.03	2,907.97	2,933.68	2,927.25	2,926.17
2,939.88	2,943.03	2,945.83			

S&P = Standard and Poor's

Second rule

Data has excellent descriptive properties and should be visualized (for example, charts, graphs, plots or tables).[7]

Analyzing data includes developing a descriptive summarization and visual patterns and behaviors of the data. The results of the analysis permit testing and confirming the data with statistical techniques to develop a long-term predictive profile.

Most analysts and decision makers use a small amount of data to make a decision. They can examine the numbers and come to some conclusions. Visualizing the data, however, is not that difficult and often provides more detail and information. Consider the entire end of day S&P index values for April 2019 (see Table 1). To visualize that data, begin with a simple plot (index value Y versus day X).

The index increases from the first day, but it is not successive. There are up and down days. Plots or graphs reveal how data change over time—an excellent method to visualize consistency. Repeatable data remain the same, inconsistent data change or vary frequently and without a pattern. You can see "spread" in the data with this plot. When graphing and interpreting the data, often it is helpful to examine how the data varies (changes or spreads) against a known or calculated value.

Figure 1 (p. 42) data lack a point or standard from which to make a comparison, judge the rate of increase for the index or forecast the future. One such common point estimate to use as a standard is the mean (or average). These statistics often (but not always) describe the center of the data (the median is another such standard that measures centrality). A simple summary descriptor for the S&P index plot is the average for the month, which is 2,903 (dotted red line). See Figure 2, p. 42.

The final element of a simple data analysis is the determination of shape or form. The shape is related to the long-term behavior of the data. By grouping and arranging the data into intervals, it is possible to observe its form. Histograms, with no fewer than seven equal-width groups, provide an excellent tool for visualization (see Figure 3, p. 43). Although data may be sampled randomly or sequentially, the shape that emerges generally exhibits a specific pattern. This pattern is called a distribution and has typically predictive (probabilistic) capabilities.

To best visualize the shape, however, you should get a larger sample size than 50 observations. Fewer data points could distort the natural behavior of the data. Consistent data exhibits a pattern, and it can be repeated and predicted (forecasted) with reasonable certainty. Knowing the probability distribution and its unique behavior (how the data are spread, shaped and centered) provides valuable information.

To verify a particular probability distribution, use statistical tests to confirm how well the data match a known set of criteria that describes the distribution. Be aware that the size of the sample (small versus large) may yield a different visual pattern.

Histograms and box plots are useful tools to visualize this concept (see Figure 4, p. 44). Also, fundamental statistical analysis may be helpful, as well. Figure 4 tests whether the data follow a normal distribution (Anderson-Darling test of normality)—a typical symmetrical pattern (also called the bell-shaped curve) that, given the author's 40-plus years of experience, exists about 60% of the time.

Do not concern yourself with only the statistics provided. First focus on the visual. For Figure 4, the shape is defined (tested) to be normally distributed (bell shaped). These data now have the potential predictive capability if the variation remains consistent.

Third rule

Patterns, trends or cycles indicate inconsistency in the variation that is unpredictable and changing.

Plotting the data offers a unique approach to understanding their performance. The S&P index data for April 2019 (Figure 5, p. 45) provides additional information. The plot clearly shows two distinct trends from the beginning of the month to its end. Notice that the individual values vary from day to day, with small and significant differences. These differences measure short-term consistency and are due to various factors (financial, business information, news items, and analysts' warnings or encouragement).

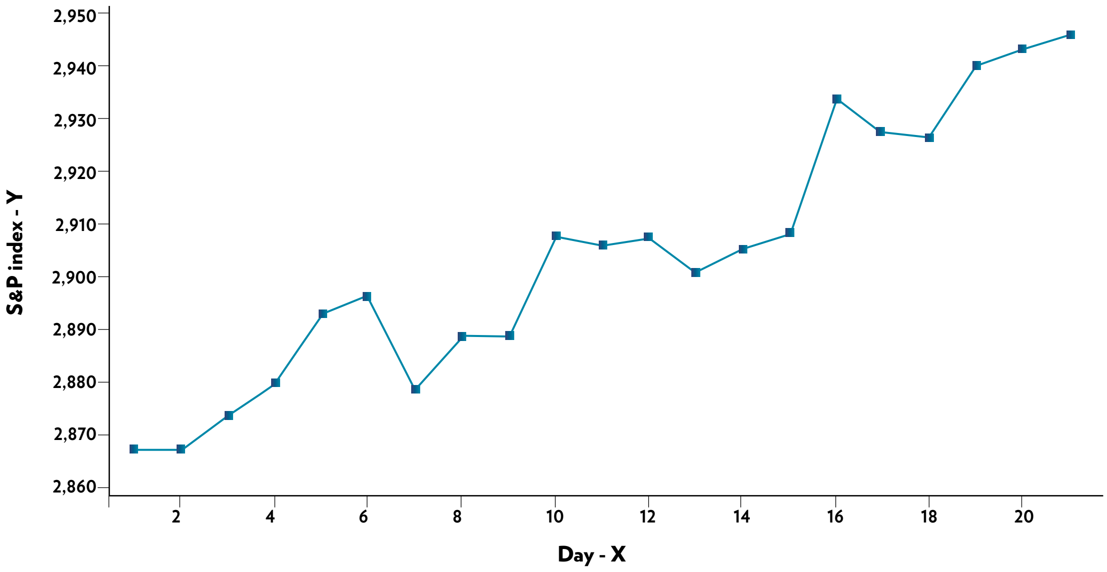
These short-term changes depend on many inputs, which include feelings and emotions. These unknown influences are what make the market inherently unstable (unpredictable—unless the analysis identifies what is driving the market, then short-term results can be very profitable). The trend in Figure 5 is unsustainable because of this inherent inconsistency. The causes and impact of this variance are much less evident during days seven through 13.

Differences also exist between longer-term measures (such as the mean or average) and individual values. In Figure 1,

iStock.com/kjeruff

FIGURE 1

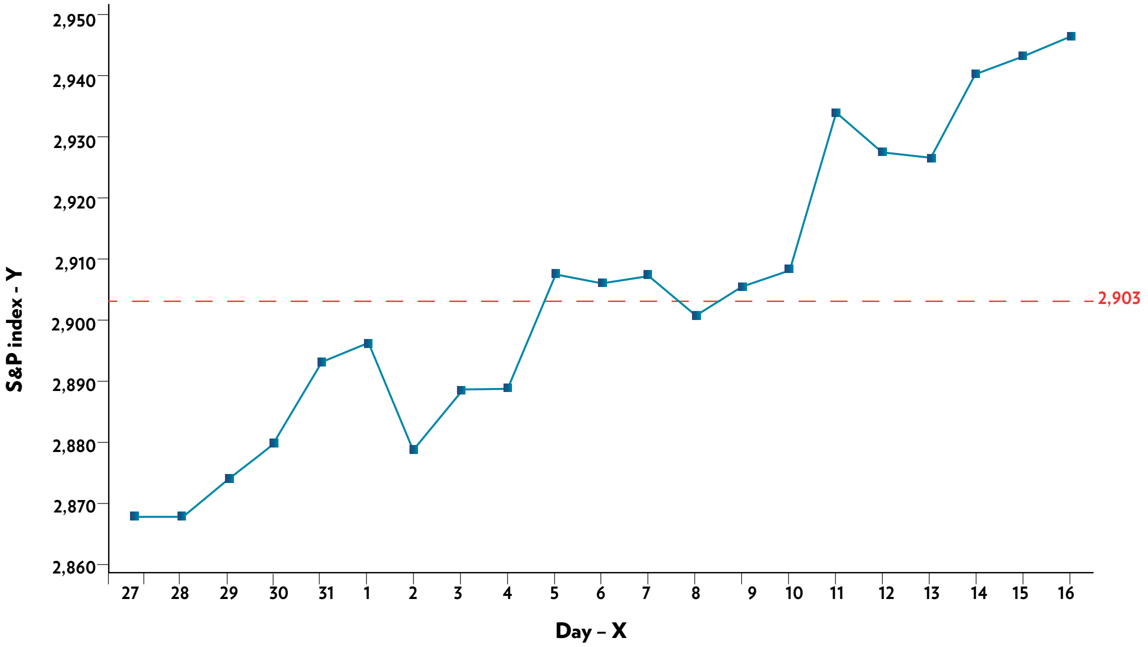
S&P index for April 2019 (end of the trading day)



S&P = Standard and Poor's

FIGURE 2

S&P index with monthly mean



S&P = Standard and Poor's

the mean (average) is only a point or single estimate. How well does the average predict the stock index price? Not well because only a few values are near enough to the center to be predictive.

From Figure 5, what does the trend suggest about future performance? The answer is the pattern will be reversed sometime in the future. The result is that a dilemma occurs in trying to determine whether to buy, sell, wait or search for more information. The answer is to search for various influences that move the market (what causes the prices to change), and either act (buy or sell) or wait. If critical information is missing or incomplete, the decision becomes more of a gamble.

Few understand the inherent error they commit when trying to predict in the presence of instability. Predictability—a desired outcome of statistics—is directly linked to stability. Without stability, a prediction is irrelevant. Control charts (Figure 6, p. 46) help visualize this concept.

Trends or cycles (troughs and peaks) and patterns (recurring data, seasonal or time dependency) are the results of one or more causes or influences that can and do change. Therefore, the future may be unpredictable if these influences remain hidden, fluctuating or unknown.

Moving on, Figure 7 (p. 47) is the next eight days (now, May 2019). The mean (average) is nearly identical, but the pattern is different, and it is changing (falling index).

Why is the index falling? Uncontrolled or unknown influences are affecting the index. Watch any of the business TV shows and analysts will try to diagnose the change, predict the future or stress one factor that seems to drive the market. The fact that stocks are affected by economic conditions, political pressures, emotions and complex buy-sell programs are some of the reasons for the inconsistency.

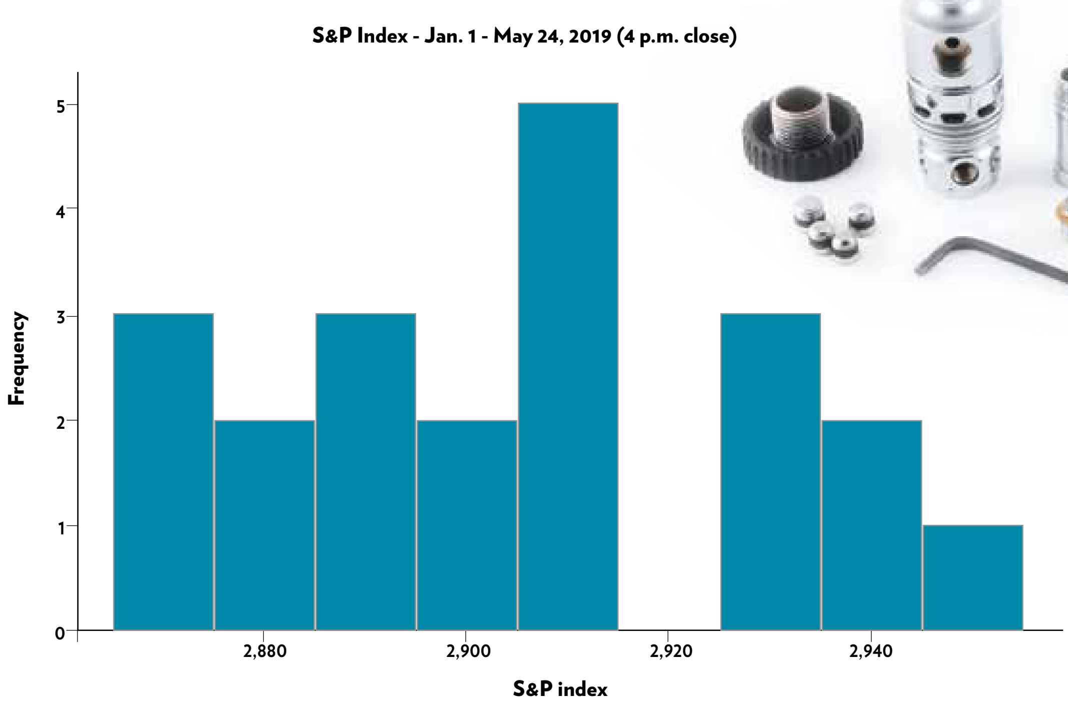
The next eight days confirm this noticeable change (see Online Figure 1, which can be found on this article's webpage at qualityprogress.com).

The May average is now significantly lower. When the inconsistency is present, the statistics will change but have no predictive component (it will not indicate future performance).

Therefore, to determine long-term behavior, a larger sample was extracted. The spread and shape of the S&P index for Jan. 1 to May 20, 2019 (Online Figure 2) is different than for the smaller sample example. The effect of time now influences the result as the shape changes from bell

FIGURE 3

Histogram of S&P index—April 2019

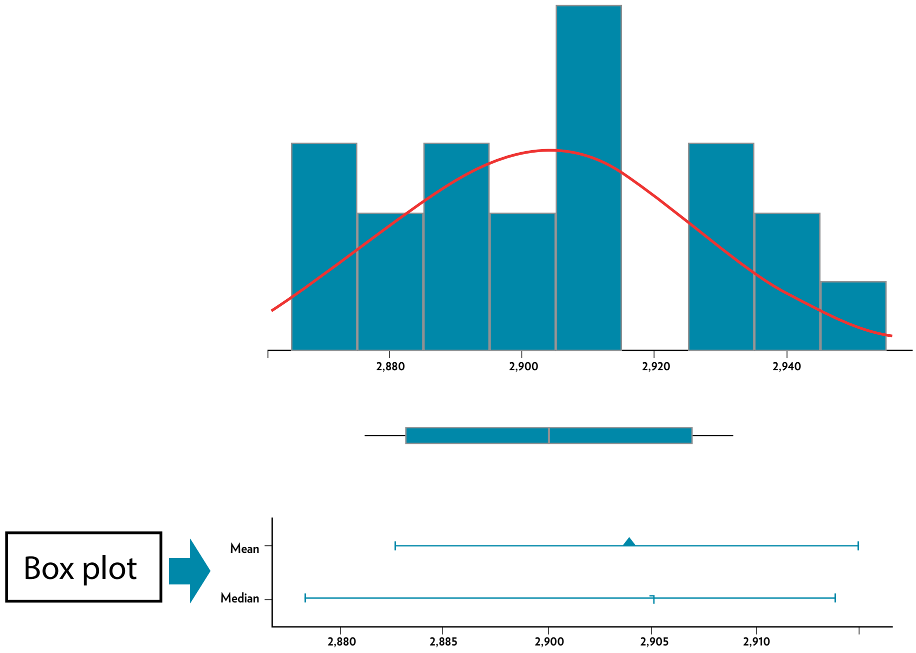


S&P = Standard and Poor's

iStock.com/kjerulff

FIGURE 4

Summary report for S&P index—April 2019



Anderson-Darling Normality Test	
A-Squared	0.35
P-Value	0.434
Mean	2903.8
StDev	24.6
Variance	606.5
Skewness	0.243328
Kurtosis	-0.957623
N	21
Minimum	2867.2
1st Quartile	2883.8
Median	2905.0
3rd Quartile	2926.7
Maximum	2945.8
95% Confidence Interval for Mean	
2892.6	2915.0
95% Confidence Interval for Median	
2888.3	2913.9
95% Confidence Interval for StDev	
18.8	35.6

S&P = Standard and Poor's

shaped (normally distributed) to a more skewed shape (data tests to be non-normal, which are not shown).

These data contain an upward trend, which is inherently unstable. That trend ended in May 2019. Changeable information or data are unpredictable. If you try to predict the S&P with a traditional distribution shape, such as the normal distribution, forecasts will be consistently incorrect.

Fourth rule

Knowing the probability distribution is required to forecast long-term performance.

If the variation is consistent and the distribution verified, the long-term prediction is possible. You cannot forecast data that are changing uncontrollably (this includes periods of stability). A similar fate awaits those who do not understand the predictive nature of a probability distribution.

LEARN MORE

Read more about data collection and data analysis tools at ASQ's Learn About Quality Library. An extensive collection of data collection tools and templates are there to download for you to use immediately. Visit asq.org/quality-resources/data-collection-analysis-tools for more details.

Many assume that data are always normally distributed. That assumption is correct, especially if the measurement is lengths, widths, depths, volumes, weights, flows and scores, to name a few. However, if the measure is something like a survey response, tracking errors, perceptions, importance, value, repair rates or prices, other distributions are better for overall prediction and accuracy.

Online Figure 3 visualizes four distinctive probability distributions, each of which is predictive for a specific measure. Online Table 1 lists typical variables that are non-normal. The list is far from exhaustive, indicating the diversity present in everyday data.

Distributions are either discrete (such as the Poisson) or continuous (such as the familiar normal). To visualize a known distributional shape requires a large number of data points (more than 500 data values). Yet, most decision makers use fewer than 50 samples to make critical decisions. Small samples increase the risk of making an error, and that error increases if the behavior of the data is unknown.

If behavior (that is, shape) can be estimated or predicted, this becomes a method to increase the informational and predictive power of the data. If, however, the data shape is unknown, forecasts and predictions are limited. Validate all distributional forms and try not to approximate a distribution if the behavior is unfamiliar. Using an approximated probability distribution—such as the normal—may simplify the analysis but provide inaccurate results, if unsubstantiated.

Fifth rule

Statistical tests, techniques and experiments are likely to validate and confirm data.

Most decision makers use a small sample from their work to make critical decisions. If the variation is consistent, and the shape is known and validated, the statistics will have excellent predictive properties. The less you know about the distributional properties, the larger (and better) the sample required to verify the results. Assume a shape is dangerous if that shape is untested.

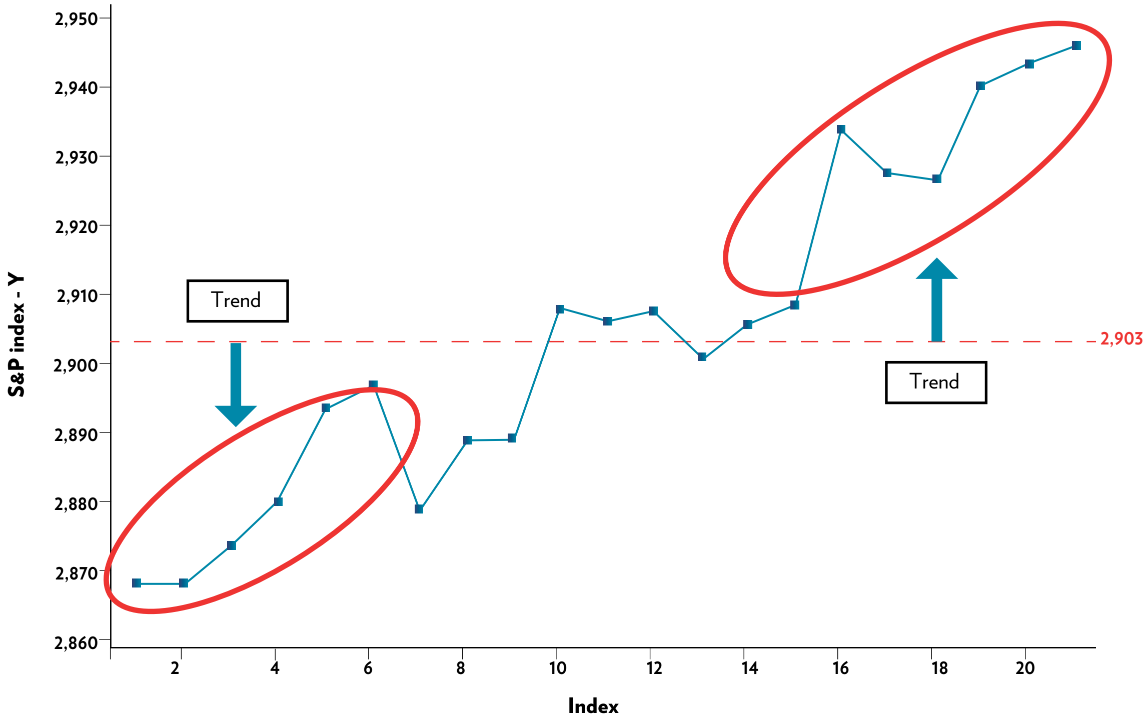
The consequences of doing nothing

Organizations must change the way they analyze and interpret data. Poor analytical techniques yield suspicious results. The normal distribution has become the de facto shape to describe most data. Shape or distributional form and variation rarely are discussed or dismissed as being too complicated. Both are critical to understanding long-term patterns of data. Statistics, such as the mean, are accepted as long-term predictors. Yet, some distributions have a calculated average that is neither a measure of centrality nor predictive. The result leads to an epidemic of poorly analyzed and misinterpreted results, such as:

- Analysts, business publications and TV shows want to analyze every number as if it uniquely describes the

FIGURE 5

S&P index—at the closing bell, April 2019 (trends)



S&P = Standard and Poor's

measure, value or trend of the stock, index or statistic even if those values change daily.

- Medical personnel often want to react to a number, such as a blood pressure reading or test results, without first examining the patient's history or placing the data into its proper context. What impact does this have on the patient?
- The easy application of statistical software has created so-called "experts" who lack a basic understanding of probability and statistics.
- Relying on the mean or average is handy and predictive when the data come from a symmetric distribution, but may have little meaning for non-normal distributions or data with changing variation. The statistical theory does not substantiate the claim that all data eventually would become normally distributed.

Flawed analysis, for the most part, stems not from fraud or formal misconduct, but more normal misbehavior: miscalculation, poor study design or self-serving data analysis.[8]

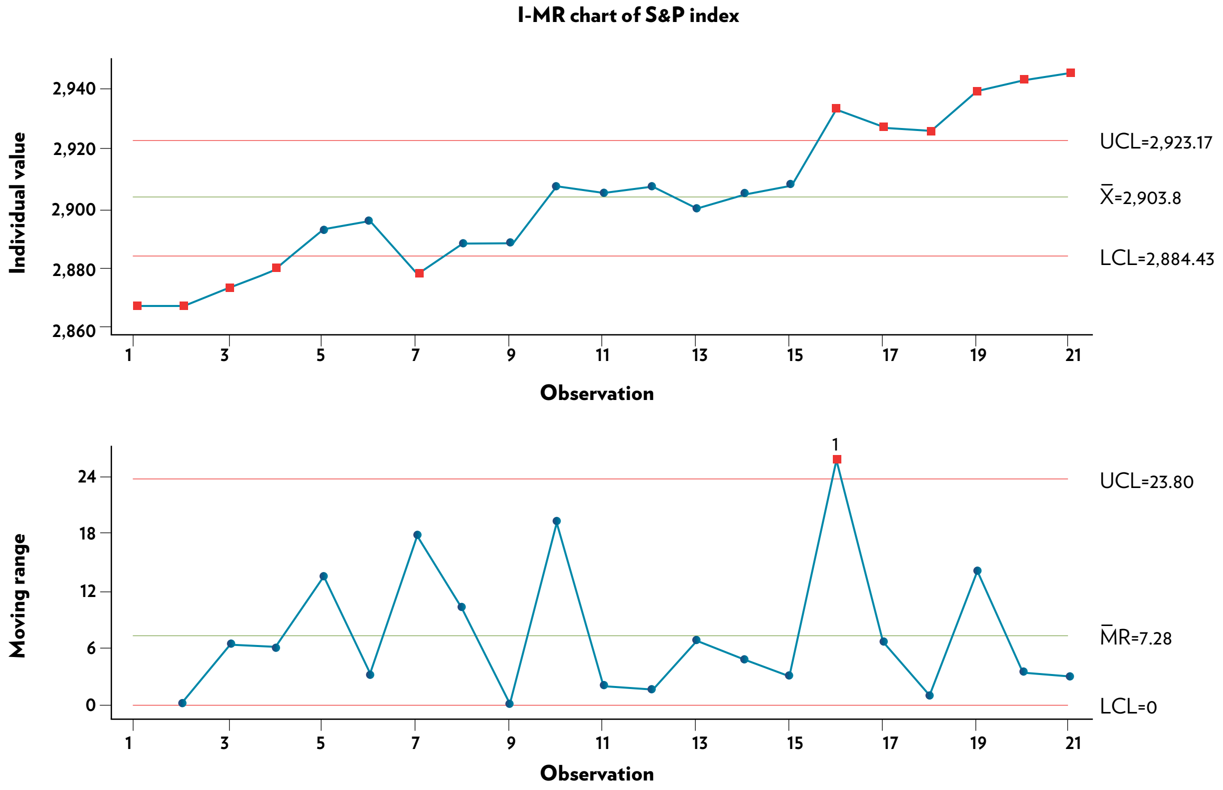
Recommendations

Insist on data analysis and interpretation. Ask "why?" concerning any pattern or trend that changes, underperforms or surprises. Do not be impressed with statistical output that is too complex to decipher or cannot be fully explained by more straightforward data analysis. Initiate a culture of data fluency that supports informed decision making.

Data analysis includes interpretation, which places the results within a context that management can use easily.

FIGURE 6

Confirmation of instability and unpredictable patterns—control chart



I-MR = individual-moving range
S&P = Standard and Poor's

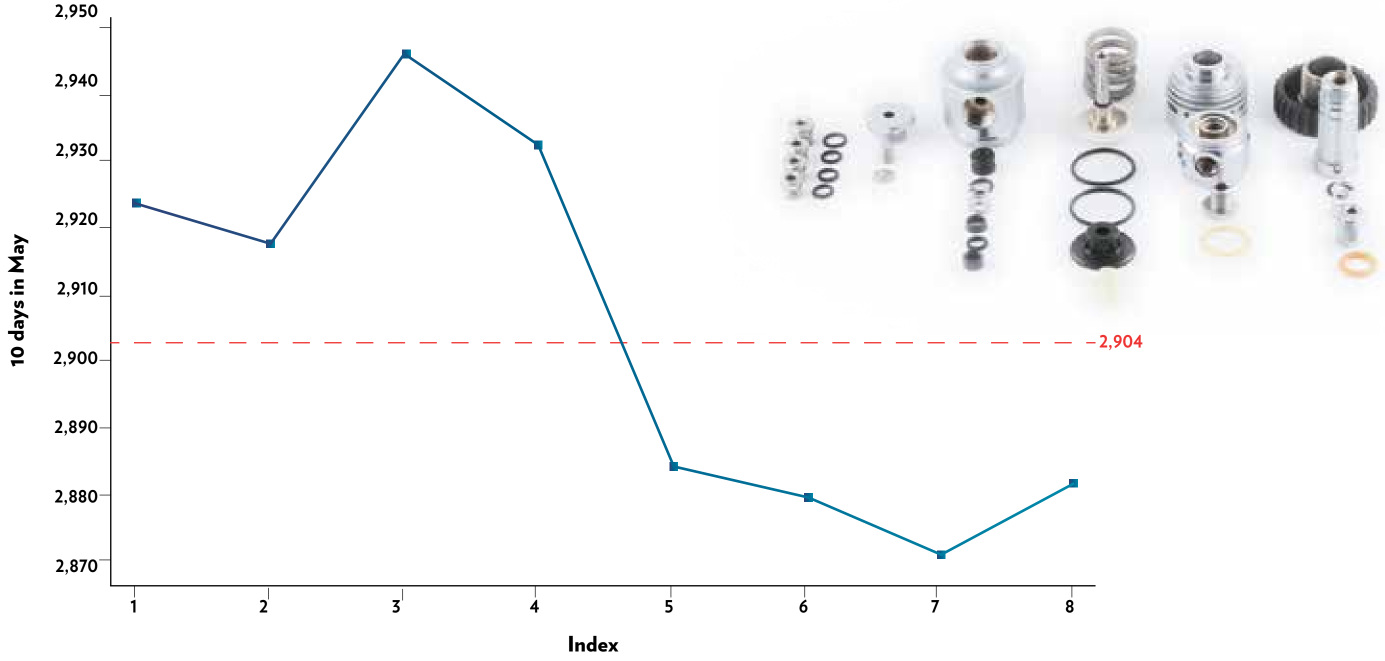
LCL = lower control limit
UCL = upper control limit

MR = moving range
X = mean

iStock.com/kjerulff

FIGURE 7

First 8 days of May 2019—S&P index at the closing bell



S&P = Standard and Poor's

Ask for the charts and graphs, and require that these be fully interpreted and validated. Using these rules will guarantee a successful informed decision that relies on predictive and actionable information. **QP**

REFERENCES AND NOTE

1. William Markow, Soumya Braganza and Bledi Taska, "The Quant Crunch," Burning Glass Technologies Inc., 2019, www.ibm.com/downloads/cas/3RL3VXGA.
2. Andy Capaloff, "Big Data + Big Analysis = Big Bust," Curatti, 2014, https://curatti.com/big-data-bad-analysis-big-bust.
3. Jonathan Low, "How Bad Data Costs the U.S. $3 Trillion Per Year," The Lowdown blog, 2016.
4. Tim Hartford, "Big Data: Are We Making a Mistake?" *Financial Times*, Dec. 1, 2014.
5. Mahak Vasudev, "What Is Bad Data—and Its Side Effects?" Business 2 Community, Feb. 21, 2015, https://tinyurl.com/what-is-bad-data.
6. Matthew Zajechowski, "The Lessons We Can Learn From Bad Data Mistakes Made Throughout History," SmartDataCollective, May 25, 2017, https://tinyurl.com/learn-from-bad-data.
7. All figures were created originally using Minitab 17 software.
8. John P.A. Ioannidis, "Why Most Published Research Results Are False?" PLOS Medicine, Aug. 30, 2005.

WATCH MORE

ASQ**TV** has a host of videos related to data analysis, including one that offers advice on gathering and analyzing data in organizations, tips on using Likert scales, and a case study on leveraging data to help the bottom line. Visit videos.asq.org/likert-scales-and-data-analysis to access the video, which features an interview with Chris McMillian, senior corporate performance analyst for the City of Fayetteville in North Carolina.

Gregory C. McLaughlin is chief analyst at McLaughlin Partners LLC in Fort Lauderdale, FL, and a contributor to the Malcolm Baldrige National Quality Award. He holds a doctorate in business administration from Nova Southeastern University in Fort Lauderdale and a master's degree in statistics from Florida State University in Tallahassee. He is a senior member of ASQ and an ASQ-certified Six Sigma Black Belt, quality engineer and quality systems lead auditor. McLaughlin is the author five books about innovation, including *Dubai: The Epicenter of Modern Innovation* (Productivity Press, 2017).