General Considerations for Working with Data

Of Elephants, Measurements, and Statistical Tools

By Katie Daisey

Q What are the general considerations before collecting data?

A There's an Indian parable about several blind men who encounter an elephant. It goes:

> *A group of blind men heard about a strange animal called an elephant. Out of curiosity, they said, "We must inspect and know it by touch, of which we are capable." So, they sought it out, and when they found it, they groped about it. The first man, whose hand landed on the trunk, said, "This being is like a thick snake." Another man, whose hand was upon its leg, said, "The elephant is a pillar like a tree trunk." The blind man who placed his hand upon its side said the elephant "is like a wall." Another, who felt its tail, described it as a rope. The last felt its tusk, stating the elephant is that which is hard, smooth, and like a spear.*

This parable tells many truths for the area of statistics and the pitfalls that can occur when transforming data into knowledge.

SAMPLING

Like the blind men, we can only use our tools to examine what is directly in front of us. In the parable, we had five blind men who, upon each taking a single measurement, found data that seemed to be unrelated. This is the argument for having an adequate sample size. It's possible that on continued examination of the creature, they would begin to build a clearer picture of an elephant.

The main issue in this case is: How many samples are enough? If we know something about what types of animals exist, we might be able to guess how many blind men we need before we have a complete picture of any creature they might examine. Exchange "distribution" for "type of animal," and the application to general sampling theory becomes clear. This has been treated quite a few times, including in previous Data Points columns, so I leave the mathematical treatment to those articles. But a better understanding of what we are examining leads to a better understanding of what comprises an adequate sample size.

Also clear is the impact of poor sample planning. Imagine only the first blind man making several measurements, all in the same general location of the elephant's trunk. He would be convinced, and have strong statistical data to support, the premise that the creature he examined was very similar to a snake. Any attempt to apply this knowledge to another area of the creature, or to the elephant as a whole, would fail immediately. This appears quite a bit when measuring changing systems. There may be limited times and locations to physically access the system, maybe with only a single gage located near the end of a process. Attempts to represent the beginning of the process using only samples collected at the end of the process would lead one astray.

When deciding whether a sample is representative or not, care must be taken, as this will also determine how representative any learnings from that sample will be.

ERROR (AND UNCERTAINTY)

The concepts of systematic and random error are well studied in most quantitative fields, but I wish to expand upon the nature of these errors. The most common systematic error is an offset, where the true value is consistently off from the measured value, M, giving $(M + E)$ across the entire range of possible measurements, but there also exists proportional error where the measured value is affected by the value of the actual value (E^*M). Systematic errors can also arise from drift in a system (experimental or measurement).

The majority of the time, random error will be IID (Independently and identically distributed) and Gaussian, which are statistical terms to describe a measurement that is normally distributed and not reliant on another measurement. But these are not required properties of random error. Random error can change depending upon the magnitude of the value being measured, typically becoming more broadly distributed at larger values. Random error can also have a non-normal distribution.

While it is important to understand the types of error present in our measurements so we can determine the appropriate statistical tests to apply, it is also important to understand the practical impact of such errors. Knowing that our measurement system has a

proportional systematic error with a more broadly distributed random error at higher values, we might be less willing to trust a single measurement at high values than at low values.

CORRELATIONS

In grade-school science and mathematics, we often speak about independent variables (those that are manipulated by an experimenter) and dependent variables (those that are affected by those changes). These terms can deceive us into believing that independent variables affect the dependent variable independently from each other, and even worse, independently from all of the other variables we did not measure.

Consider the "independent" observations of the mysterious creature. The confusion clearly arises because of the unmeasured but correlated variable of location on the elephant. Without acknowledging this missing, confounding variable, the data collected seems nonsensical. Correlated dependent variables should be checked for routinely, as issues arise when applying statistical models designed for uncorrelated data and drawing incorrect conclusions. As always, correlation does not imply causation.

LIMITATIONS

What color is the elephant?
Our blind men currently lack the tools necessary to answer this question. But if a pink elephant were to exist in the herd, it might cause immediate problems for the herd's survival. It is important to remember that conclusions are drawn on the strength of what is measured. Often, what is analyzed is what is easy to measure, whether or not that actually addresses the information of interest.

That is often a direct consequence of "operationalization" of our measurements. Sometimes, the information of importance is easily measurable (quantitative or qualitative), but it can be much more difficult. For instance, the texture of our elephant as felt by hands is difficult to directly measure. A standard scale from smooth to rough could be developed and the observers rigorously trained. Perhaps texture is a multi-faceted parameter, and it is necessary to include the hardness and temperature of the surface (along with numerous additional variables). The translation from an individual understanding to transferable data requires special attention.

TRUTH

The statistical moral of this parable is to be very careful and precise when designing and undertaking studies and experiments. Without a solid understanding of the assumptions and nature of the data collected, a naïve application of statistical tools can look quite similar to a group of blind men examining an elephant. ■

Katie Daisey is a scientist (chemometrician) for Arkema Inc. A member of ASTM International since 2018, she is recording secretary for the committee on quality and statistics (E11).

Dean V. Neubauer, the Data Points column coordinator, is engineering fellow and chief statistician, Corning Inc. A member at large on the executive subcommittee of the committee on quality and statistics (E11), he is an ASTM International fellow and a past chair of the E11 committee.

CORRECTION

In the March/April Data Points, "The Weibull Model — Building on Reliability" (online at www.astm.org/weibull-model), Stephen Luko and Dean Neubauer answer the question, "What is the Weibull distribution and how is it used in data analysis?" Of the article's example about an aerospace device, Luko and Neubauer write, "This is not a safety-related issue, and the manufacturer has agreed to a warranty time of 1,500 hours. The table values show an estimated reliability at $t = 1500$ cycles of about 99.5%, assuring the manufacturer and the customer of this value." "Hours" should read "cycles" in this sentence, but the math and numerals are otherwise correct (although inconsistent as to comma/no comma) as published.