

SUMMARIZING YOUR DATA...OR WHAT DOES A MEAN MEAN?

By James R. Chastain, Jr., PhD, PE, MPH

One of the consequences of our digital age is that data not only proliferates, it can absolutely overwhelm. For raw data to be useful it must be converted into useful information. Much of this data is numerical in nature and one of the ways it is interpreted is by the use of statistics or statistical inference. The mere mention of the word “statistics” causes many people to fall into a catatonic state, which is unfortunate. Certainly, statistical practice can be somewhat arcane and mathematically precise, but the fundamentals are quite accessible and are a part of most professionals’ academic background. Two common queries that are typically made are to (1) summarize the data and (2) detect meaningful trends. This article will discuss a few important points to consider when summarizing data sets.

After thinking about it for a moment, the reason that difficulty exists in interpreting data is that the data changes. If every piece of data was the same, then interpretation wouldn’t be a problem. However, the fact is that variability does exist...and variability in the numbers can lead to variability in interpretation which, in turn, can result in erroneous analysis or decisions. So how can summary statistics help with this task?

PRELIMINARY ISSUES

Before analyzing the data, the first thing that needs to be done is to characterize it. Normally this will only take a couple of minutes, but it helps to highlight the types of computations that are appropriate for the data. Thus, at the outset, the following should be identified: (1) the data type, (2) the data extent, and (3) the data distribution.

The data type refers to the nature of the data. Is it discrete (whole numbers) or continuous (includes fractional values)? When considering more sophisticated analyses, the data type can be an important distinction. Although we won’t go into it in any detail in this article, it is helpful to also classify the data according to whether it is nominal, ordinal, interval or ratio. Many times this is a function of whether the data is coming from a specific “measurement” or whether it is the output of something like an opinion survey. For the purposes of this article, it is assumed that interval or ratio data (the typical numerical “measured” data) is being used.

The data extent refers to whether the data measures the whole source or just part of it. In other words, does the data describe the whole population or is it a sample of the population? This is an important distinction. For example, if the height of everyone in a room is measured and recorded...is that a population or a sample? To answer the question, another key question must be asked, ‘how is the data going to be used?’. If one is just interested in understanding the height variation of people in that room, then it is a population measurement. Why? It is because everyone (in the population) has been measured and we have complete information about the group. On the other hand, if the analyst is using the people in the room to estimate the height variation for people in the state of Florida, then the group is a sample of the population and not the population itself. Thus, there is additional uncertainty in the statistics that must be accounted for. Misunderstanding this distinction is a common error and can lead to improper data summaries.

The data distribution simply refers to the shape of the data as it is plotted. Most everyone has used histograms to help visualize data. It is found, by experience, that different data sources can assume common shapes which can be approximated mathematically. These mathematical models can then be used to more efficiently summarize the data. But if the analysis is performed using the wrong data distribution, it is easy to see how errors in interpretation can be made.

CENTRAL TENDENCY

As a first step it is necessary to see where the “center of gravity” of the data lies. To characterize the data by a single number, this center of gravity value will represent the expected value of the system.

It seems almost instinctive, when faced with a ream of numbers, to compute the average. The average, or “**arithmetic mean**”, is one of the first statistical parameters taught in school. It is merely the sum of all the data points in a data set, divided by the number of points in that data set. In most spreadsheets or calculators it can be automatically computed. Thus, it is commonly used because it’s an easy value to obtain and the value is easy to interpret and explain.

VARIANCE

Unfortunately, the mean as a single point estimate may not adequately express the nature of the data. The next task, then, is to develop an estimate of the variation of the data from the mean (center of gravity). Ideally, what we would like to do is get an average of the variation so that we can represent the data variation by a single number too. Unfortunately, we can’t do this directly because a simple average of the deviation from the mean will always equal zero. This is because values above the mean always equal the values below the mean (that’s why it’s the average!). So statisticians developed a clever way to get around the problem. Rather than average the difference between the mean and each value, they averaged the square of the difference between the mean and each data value. This will not equal zero (unless there is no variance) because the square of any deviation (above or below the mean) will be positive so nothing cancels out. This value is called the **variance**.

The only problem with the variance is that it’s a squared term and is difficult to interpret when compared to the mean. However, if the square root of the variance is taken, the resulting value has the same dimensions as the mean. This value, the **standard deviation**, can be added to or subtracted from the mean to give a sense of the data’s deviation from the mean.

With these simple operations, we have an estimate of the data’s center of gravity (mean) and an idea of the variation of the data from that center of gravity (standard deviation). Up to this point there’s probably not much that’s been presented that you didn’t know already. However, many times, there are subtle aspects of basic statistics that need to be understood so that misinterpretations don’t bias the use of the information.

THE AMAZING CENTRAL LIMIT THEOREM

One of the truly amazing relationships to come from mathematical statistics is the Central Limit Theorem. In essence it states that, for a sample from any distribution, the true mean lies within ± 1.96 times the standard deviation of the data 95% of the time given a large set of samples. What makes this amazing is that it applies to any data distribution...not just the normal (Bell Curve) distribution. So it doesn’t matter what the underlying distribution is (ex. normal, uniform, triangular, weibull, non-uniform, etc.); the mean will be located by the Central Limit Theorem with a stated probability. Of course the more elements in the data set, generally the smaller (tighter) the standard deviation becomes, so the estimate becomes more reliable.

HOW IS THIS MISUSED?

The most common reason for mistakes being made with this theorem is that it is assumed that 95% of the values lie between ± 1.96 times the standard deviation from the mean. This only applies to data sets that encompass the entire population (all possible values). Most data that is generated in practical settings are samples of the population. The central limit theorem says that, for repeated samples, the actual mean lies within ± 1.96 standard deviations of the computed mean (with 95% probability). Note the central limit theorem attempts to locate the mean (only) and not all of the data values.

Thus, when summarizing data, be clear about the nature of the data set and the assertions that you make. In other words, is it the mean that is being discussed or is it the data distribution? If this isn’t clearly

understood, significant errors can be made in the data analysis.

IS THE MEAN MEANINGFUL?

As valuable as the mean is, there are situations where it can mislead. The mean is most useful with data distributions that are symmetrical, that is the data is a mirror image on either side of the mean. However, many times the data that is encountered in practice is skewed or, in other words, there are more data points on one side of the mean than the other. It is common to find this situation when dealing with environmental or economic data. For example, consider an environmental system that is in compliance most of the time but has a one-time pollutant spill. The average of all those values might indicate that the system was routinely out of compliance because the average was higher than the permit value. This would be untrue though. The system only had one non-compliance event that skewed the average.

In cases where the distribution is highly skewed, many times it is better to use the **median** instead of the mean. The median represents the data point where 50% of the data points are above it and 50% are below it. Thus, the median is based on a sorted or ranked set of the data. As an example, if there are 99 data points in the ranked data set, then 49 points would be above the median and 49 points would be below it.

Using the same process, the variance can be described by recording the value at any desired percentile (ex. 10th percentile, 20th percentile, etc.). Thus, any percentile (also called quantile) can be located and recorded. When these values are plotted, they can provide a good sense of the data distribution.

The key point here is that the median limits the effect of extreme values on the magnitude of the measure of central tendency. Also, note that since the median is based on the ranking process, it is not dependent on the underlying distribution of the data (i.e. it is non-parametric or distribution-free).

SO WHAT DOES ONE DO?

Given a long list of numbers, what initial steps should one take to summarize the data? The following steps might be helpful.

1. Understand what the data represents.
 - a. Does it encompass the whole population of the set of interest or is it actually just a sample of the population?
 - b. Is the data independent (i.e., selected randomly)? This is critical if it is a sample.
2. Compute basic statistics.
 - a. Compute Mean and Standard Deviation.
 - i. If the data is population-based, compute the confidence limits for the data set.
 - ii. If the data is a sample, compute the confidence limits for the mean.
 - b. Compute the Median and desired Quantiles.
 - c. Compare Mean and Median values.
 - i. Look at magnitude of deviation. The greater the deviation, the larger the data is skewed or non-symmetrical.
 - ii. Identify extreme values that drive the mean. What happens to the mean if those value(s) are eliminated? Does the existence of those values make sense and should additional study be done to understand them?
3. Plot the data.
 - a. Plot a dot plot, histogram and/or probability plot of the data to form a concept of the data distribution. Look for patterns or clusters.
 - b. Can the data be classified into a common distribution?
 - c. Plot the Median and other percentile values. How does that compare to the parametric distribution?
 - d. Identify extreme values. Can they be explained?

By using this simple approach, better insight into the data can be gained. Many times it will also identify questions that can be used to initiate other tests or queries to further clarify underlying causes or events.

And, by understanding the limitations on the interpretation of the mean, it can help one avoid making faulty judgments based on incorrectly summarized data.

Dr. Jim Chastain is the President of Chastain-Skillman, Inc. He has a BSCE (honors) and Master of Engineering from the University of Florida. He also has a Masters of Public Health and Ph.D. from the University of South Florida. He is a registered Professional Engineer with over 30 years of experience and is a Diplomate of the American Academy of Environmental Engineers. He can be reached at (863) 646-1402 or jrchastain@chastainskillman.com.

© 2005 Chastain-Skillman, Inc. This article is taken from the 4th quarter 2005 issue of *Consultant's Update*, a publication of Chastain-Skillman, Inc.