More Good Reasons to Look at the Data

By Tanya Hoskin, a statistician in the Mayo Clinic Department of Health Sciences Research who provides consultations through the Mayo Clinic CTSA BERD Resource.

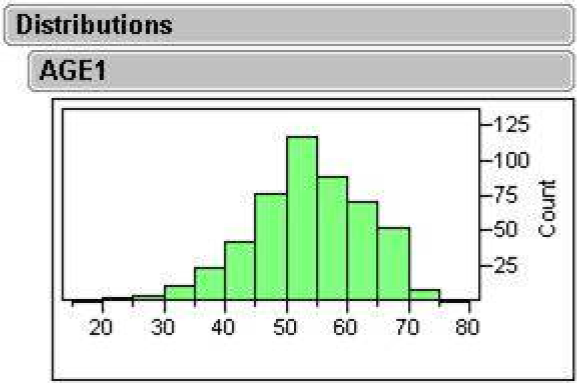
The article Always Look at the Data suggested that you look at your data to find problems, such as invalid codes or impossible values. Another article, Data Types, explained how the type of data helps determine what statistical procedures are appropriate. This statistical tip will be a combination of those two topics. Here, we will think about how the overall look of our data might also help us determine which statistical procedures to use. In other words, we are going to think about aspects of the data distribution. We will focus on quantitative, continuous data.

What is a distribution?

You might hear the word “distribution” all the time in discussions about data, but how often do you think about what it means? Basically, we have a clinical question we want to answer. In order to answer that question, we collect data from a sample of individuals from the population. It is not very helpful to look at each individual’s value separately, so we need a way to look at the values for the whole sample at once. The distribution of a variable shows us a summary of all the values in a single picture. In other words, the distribution shows us how the values are distributed or where they fall in the range of possible values.

The histogram in Figure 1 shows the distribution of the age variable (AGE1) for a sample of 500 individuals. The vertical bars represent the number of patients out of the total sample who have ages in each interval. Here the intervals span 5 years. The tallest bar shows us that approximately 120 of the individuals had an age between 50 and 55 years. This bar is approximately in the middle of the range of values. Without calculating the average age of our patients, we might guess that it is between 50 and 55 just by looking at this picture. We also know that most patients were between 30 and 70 years old. There were only a few patients younger than 30 or older than 70 based on the very short bars for those intervals. Clearly, we gain a lot of information quickly by looking at the distribution, and this information is certainly more helpful than a list containing each individual’s age.

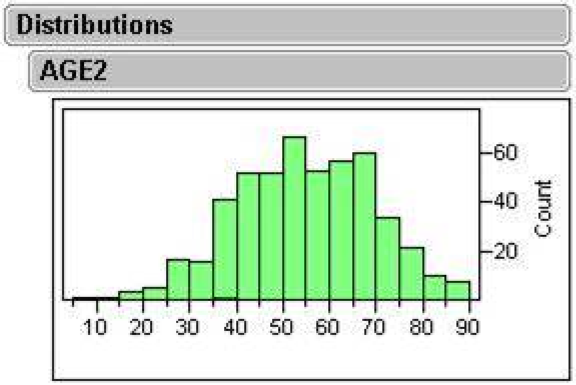
Figure 1



Center, spread and shape

The center, spread and shape of a data distribution are the three key pieces of information we can assess by looking at a histogram. “Center” simply refers to the middle of the distribution, or an estimate of what a typical value would be for these individuals. Based on Figure 1, we said that the center of the distribution seems to be somewhere between 50 and 55. “Spread” is simply how “spread out” or variable the data is. In other words, were a wide range of values observed or do patients generally have values near the center of the distribution? Figure 2 contains the distribution of ages (AGE2) for a different sample of 500 patients. The distribution in Figure 2 has more spread than the distribution in Figure 1. Can you see why? Well, the center of the distribution is about the same, somewhere between 50 and 55, but we observed a larger range of values. The youngest patient in Figure 2 is between 5 and 10 years of age, while the oldest is between 85 and 90. Also, we see more patients in intervals that are further from the center (e.g., 25 to 30 years). These observations tell us that there is more variability or spread in age among the patients from Figure 2 compared to Figure 1.

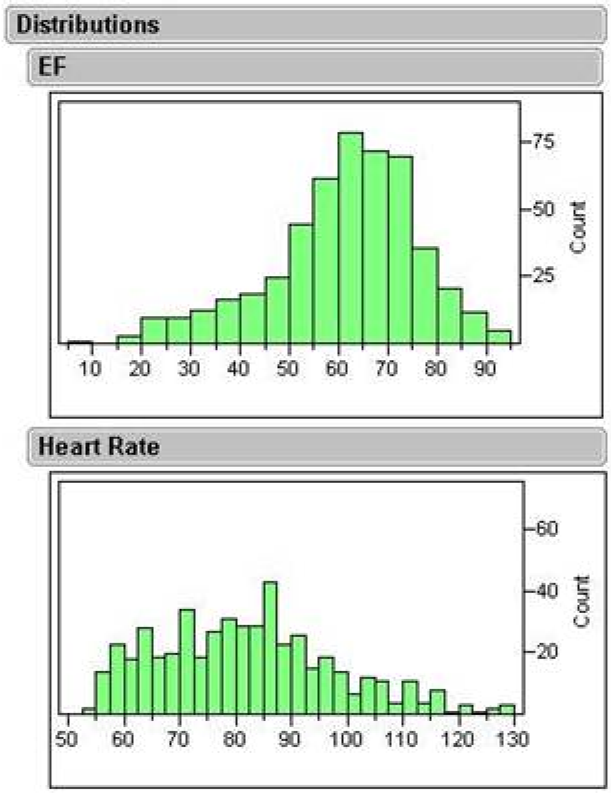
Figure 2



If you were to describe the shape of the histograms in Figures 1 and 2, you might say that both are shaped like a mound. The general appearance of the histogram is one thing to note about the shape of the distribution; in many cases it will be a mound, but occasionally you might see a shape that looks like two mounds (called a bimodal distribution) or some other non-mound shape. A second very important property to assess with regard to the shape of a distribution is its symmetry. In other words, could you cut the histogram in half and have one side that looked roughly like the other side? If a distribution is not symmetric, we might describe it as either right or left skewed.

Figure 3 shows two examples of distributions with skewed shapes. The distribution of the variable ejection fraction (EF) could be described as left skewed; the distribution of the variable heart rate could be described as right skewed. The “right” and “left” parts of the description refer to the side of the histogram that trails out farther than the other side.

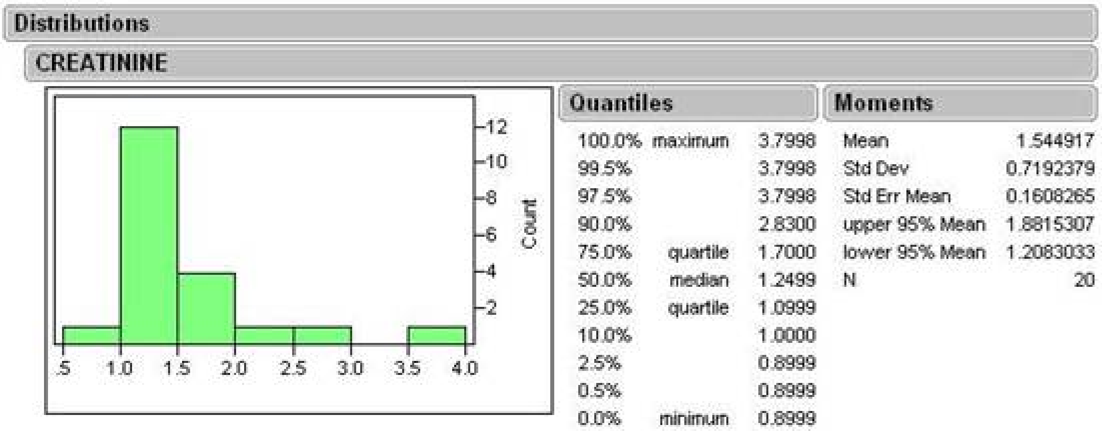
Figure 3



Shape helps determine appropriate statistical procedures: a simple example

Look at the distribution of creatinine in Figure 4. It is based on a sample of 20 individuals and clearly has a right skew. The sample mean and median are 1.5 and 1.2, respectively. In this example, these two measures of center are quite different. Which one should you report? Which value is more representative of a "typical value" for these individuals?

Figure 4



Quantiles

100.0%	maximum	3.7998
99.5%		3.7998
97.5%		3.7998
90.0%		2.8300
75.0%	quartile	1.7000
50.0%	median	1.2499
25.0%	quartile	1.0999
10.0%		1.0000
2.5%		0.8999
0.5%		0.8999
0.0%	minimum	0.8999

Moments

Mean	1.544917
Std Dev	0.7192379
Std Err Mean	0.1608265
upper 95% Mean	1.8815307
lower 95% Mean	1.2083033
N	20

For highly skewed distributions or those with unusually large or small values (i.e., outliers), the median typically is a more appropriate summary statistic to describe the center of the distribution. Why? Since the mean is calculated by adding all of the individual values and dividing by the sample size, it is strongly influenced by extreme observations in the tails of the distribution, especially for small sample sizes. The median, on the other hand, is simply the value that has half of the data points falling below it and half above, so it is not affected by the magnitude of extreme observations. In a perfectly symmetric distribution, the mean and median are equal. In a skewed distribution, the mean is pulled in the direction of the skew. Thus, the mean is larger than the median if the distribution is right skewed and smaller than the median if the distribution is left skewed.

When one reports the median rather than the mean, the range (minimum and maximum) or interquartile range (25th and 75th percentiles) is the appropriate measure of spread. For the distribution of creatinine in Figure 4, we might report a median creatinine of 1.2 with a range from 0.9 to 3.8.

The normal distribution

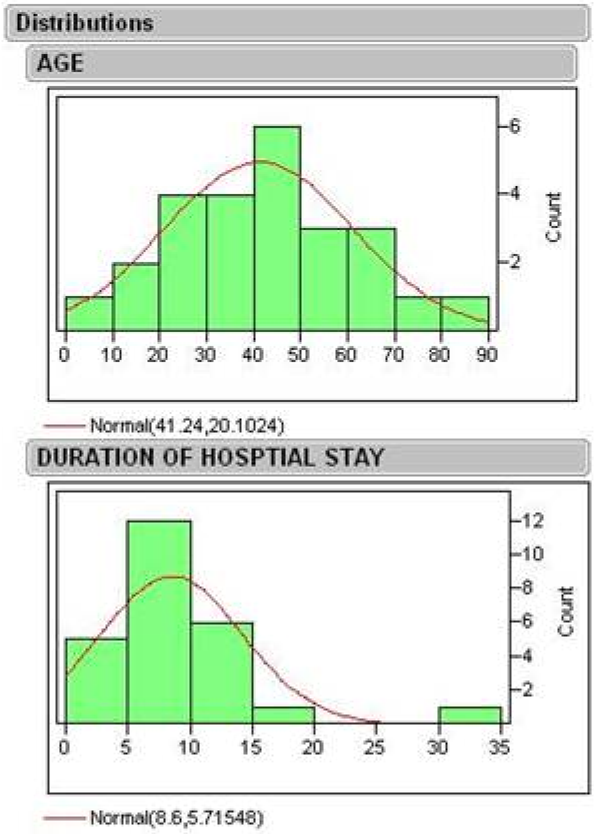
In order to talk about more advanced examples of how the shape of the data distribution affects our choice of statistical procedures, we must discuss the normal probability distribution. There are assumptions underlying many common statistical tests and procedures, and the most common assumptions are those related to the normal probability distribution. Specifically, many statistical procedures assume your data is normally distributed.

The normal probability distribution is a theoretical, mathematically defined distribution that explicitly defines how likely certain ranges of values are to occur. It is perfectly symmetric and shaped like a bell. A normal probability distribution curve overlays the two histograms in Figure 5. You can see that the histogram more closely resembles the shape of the normal curve for the variable age than for the variable duration of hospital stay. We might say that the age distribution is approximately normal but that duration of hospital stay is not normal and is right skewed.

The normal distribution shows up in a surprising number of natural phenomena. For example, human hippocampal volumes are approximately normal. It also shows up in many of our derived statistical calculations. By exploiting the desirable properties of the normal distribution, statisticians have developed many of the statistical procedures that are so helpful for answering scientific questions.

If the normal distribution is at the foundation of so many of our statistical procedures, you might wonder how we deal with the fact that the histograms are rarely, if ever, perfectly normal. We will discuss this topic in the next quarterly statistical tip, but the short answer is that “approximately normal” is often good enough, and we have tools to help when distributions are clearly not normal.

Figure 5



More information

The Mayo Clinic CTSA provides a biostatistical consulting service through its BERD Resource. More information can be found on the BERD home page.