# Useful Histograms

## No data have meaning apart from their context

Donald J. Wheeler

Context involves both the background for the data and how the data behave. This behavior of the data is most easily seen by using two complementary graphs—the running record and the histogram. Here I address the secrets of creating useful histograms.

The first graph mentioned above—the running record or time series—is incorporated into the *XmR* chart. The virtue of this graph is that nothing can hide—each value has to sink or swim on its own. By showing how the values vary over time the time series graph draws our eyes in direction our minds want to go. In this way the running record is like a movie of the data.

In contrast, the histogram of the data is like a group photo of the whole family. Here we see how each value compares with all the related values. But a family photo is not very good if some people are not looking at the camera and others are hidden by those in front. So how do we sharpen up our family photo?

Our example will consist of the call length times for 108 calls answered between 10:00 am and 11:00 am on February 12 at a customer service center. Since these calls were handled by 22 operators working in parallel with each other there is no unique time-order sequence for these data, so we immediately turn to the histogram.

| Operator | Call Lengths (in minutes) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 10.0 | 10.7 | 11.4 | 10.7 | 9.8 | | | | |
| 2 | 9.1 | 7.4 | 6.6 | 5.8 | 2.8 | 16.9 | | | |
| 3 | 13.2 | 12.3 | 10.8 | 10.1 | | | | | |
| 4 | 10.9 | 12.2 | 10.6 | 12.4 | | | | | |
| 5 | 14.2 | 22.0 | 13.9 | | | | | | |
| 6 | 11.7 | 10.4 | 9.7 | 8.0 | 6.8 | 6.1 | | | |
| 7 | 5.4 | 3.1 | 16.8 | 3.2 | 3.1 | 5.8 | 6.3 | 6.0 | |
| 8 | 6.9 | 7.6 | 9.0 | 13.4 | 15.3 | | | | |
| 9 | 6.1 | 10.3 | 10.4 | 7.0 | 11.2 | | | | |
| 10 | 7.5 | 11.5 | 12.0 | 8.1 | 12.7 | | | | |
| 11 | 13.1 | 9.3 | 15.0 | 6.1 | | | | | |
| 12 | 10.0 | 19.7 | 6.6 | | | | | | |
| 13 | 11.0 | 1.1 | 2.0 | 12.5 | 9.1 | 5.7 | 15.3 | | |
| 14 | 10.0 | 6.2 | 6.6 | 7.3 | 1.4 | 2.1 | 3.2 | 5.4 | 5.7 |
| 15 | 15.8 | 18.3 | 6.1 | | | | | | |
| 16 | 24.7 | 11.3 | 7.6 | 3.2 | | | | | |
| 17 | 5.6 | 9.3 | 10.3 | 7.0 | 8.0 | | | | |
| 18 | 9.2 | 6.1 | 6.7 | 8.5 | 10.5 | | | | |
| 19 | 17.8 | 7.7 | 9.7 | | | | | | |
| 20 | 16.0 | 9.2 | 3.2 | 7.4 | 9.3 | | | | |
| 21 | 11.3 | 12.8 | 11.4 | | | | | | |
| 22 | 10.2 | 9.6 | 8.7 | 1.9 | 6.4 | 7.2 | | | |

**Figure 1: Call Length Times**

When we put the data into the computer we get a histogram something like Figure 2. By grouping the observations into intervals this histogram shows how the data "pile up" within each interval. This allows us to see that these data have a slightly lop-sided mound shape. Since call lengths cannot be negative this shape is what we might have expected for these data.

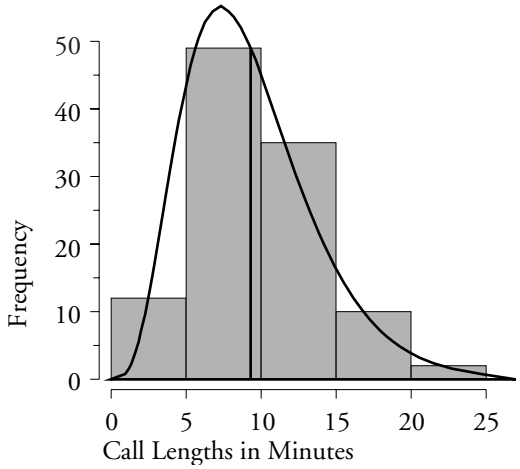


**Figure 2: Histogram for Call Lengths**

The art of creating a *useful* histogram requires a balance between the amount of "piling up" of the data and the choice of the interval size for the horizontal axis. To make this choice John Tukey recommended starting your analysis by creating a "Stem and Leaf" plot.

## STEM AND LEAF PLOTS

A stem-and-leaf plot is a way to organize your data. It is a cross between a table of values and a histogram. The idea is to split the observations into left-hand and right-hand portions. The left-hand portions will form the stem for our plot while the right-hand portions will be the leaves on the stem.

From Figure 1 we see that these data are recorded to one decimal place and from Figure 2 we see that the call lengths can range up to 24.9 minutes. So that gives us two choices for splitting these numbers. We can use the tens position for our split, or we can use the decimal places for our splits. If we used the tens position of each number to define our left-hand portions then our stem would have only three values: 0 to represent the values from 0.0 to 9.9, 1 to represent the values from 10.0 to 19.9, and 2 to represent the values from 20.0 to 29.9.

Since three stem values is too few for a good stem and leaf plot we consider using the values to the left of the decimal to define the left-hand portion of each number. Now we would need 25 values ranging from 0 to 24 on our stem.

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24

**Figure 3: Stem Values Defined by Using the Decimal Point for the Split**

Since we write numbers from left to right it is generally easier to create stem-and-leaf plots in a vertical format. Thus we commonly rotate the stem 90 degrees as shown in panel (a) of Figure 4. Next we read the data and place the right-hand portion of each value as a "leaf" on the stem.

The first value is 10.0, so we place a "0" leaf beside the 10 position on the stem to mark this value. Moving down the first column of Figure 1 the second value is 9.1, so a "1" leaf is placed beside the 9 position on the stem to mark this value. These first two values are shown in Panel (a) of Figure 4.

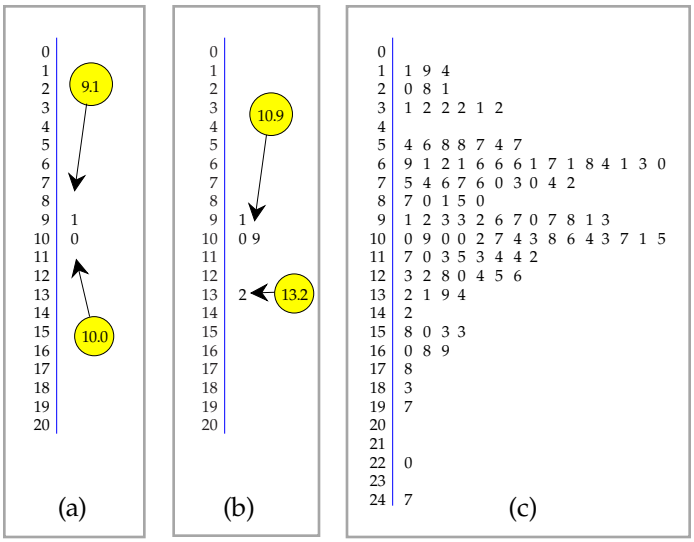


**Figure 4: Creating a Stem and Leaf Plot**

The next two values in the first column of Figure 1 are 13.2 and 10.9. The placement of these two leaves is shown in Panel (b) of Figure 4. (Practice: The first values for operators 5 and 6 in Figure 1 are 14.2 and 11.7. For these two values place the appropriate leaves in Panel (b) of Figure 4, and check your work in Panel (c).) Continuing in this manner the remainder of the 108 values are added to get the plot shown in Panel (c) of Figure 4.

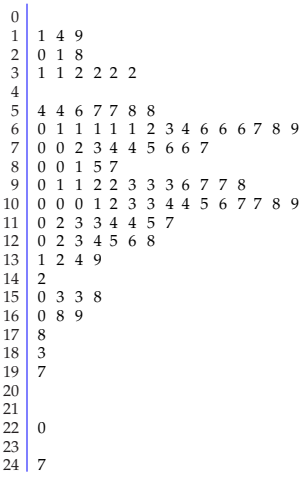


**Figure 5: A Stem-and-Leaf Plot for the Call Length Data**

Finally, by numerically sorting the values on each row of Panel (c) of Figure 4, we can completely arrange the data in order of magnitude as shown in Figure 5. This stem-and-leaf plot is an elegant combination of table and graph. It simultaneously preserves each value, arranges

the values in order, and reveals the shape of the histogram.

With a glance at Figure 5 we learn that the gap around 4 minutes is actually over two minutes long (from 3.2 minutes to 5.4 minutes). The interpretation of this gap would be that any call that lasts more than three minutes is complex enough for it to take over five minutes to complete.

For the calls that last more than 5 minutes there are two mounds, one centered around 6 minutes and another centered around 10 minutes, with a distinct valley at 8 minutes. Thus, these data have at least three distinct clusters, and all of this is revealed by the simplest of graphs.

## A USEFUL HISTOGRAM

As a result of creating the stem-and-leaf plot we now know that the detail within these data is made visible with one-minute intervals. Using larger intervals, such as the five minute intervals of Figure 2, will obscure the interesting detail that the call lengths have multiple modalities. Rounding the call lengths to the nearest minute we get the histogram in Figure 6 where the modalities at 3, 6, and 10 minutes are apparent.

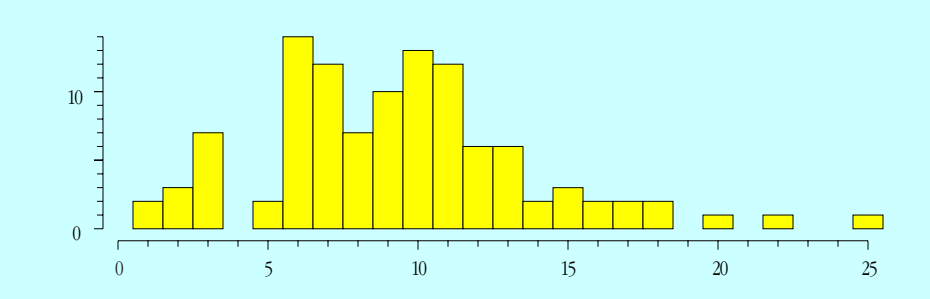


**Figure 6: A Useful Histogram for the Call Length Data**

The histogram is the "tissue of the data." When data have a time-order sequence, the histogram complements the time series and *XmR* chart. When data do not have a known time order, the histogram still shows the all the data in relationship to the other values. When we use intervals that are small enough to capture the interesting details within the data we have a useful representation of the data. It is like a family photo where everyone can be seen and all are smiling and looking at the camera.

So the major trick to getting useful histograms is to use intervals that are small enough to show the interesting details. Just what constitutes an interesting detail will depend upon the context for the data and the judgment of the analyst. Since neither of these aspects of the data will be known to your software, you may well have to manipulate the interval size in your software to get a useful histogram of your data.

Once you have a histogram with appropriately-sized intervals you may add specification limits and even natural process limits as needed to tell the story contained within your data. However, most programs will also decorate your histogram by superimposing some probability model.

DECORATED HISTOGRAMS

Three examples of a decorated histogram are shown in Figure 7. In each case the probability model superimposed on the histogram is an example of non-data ink. These curves are pure decoration. They have nothing to do with the data or the process that produced your data. They are nothing but a figment of someone's imagination.

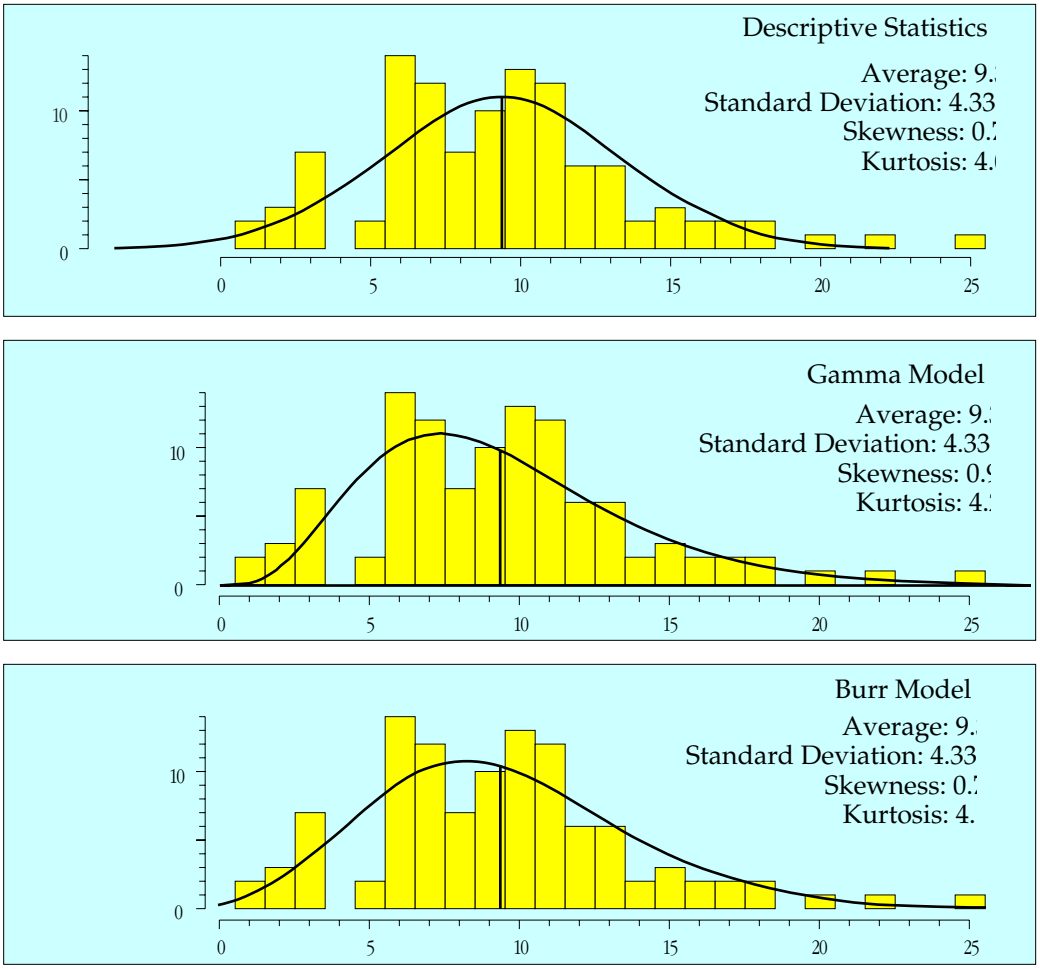


**Figure 7: Three Decorated Histograms for the Call Length Data**

Probability models do not generate your data. Never have, never will. At best, a probability model is merely a mathematical approximation of reality. At the worst, a probability model can be totally misleading. Here we have three different unimodal distributions thrown over a trimodal histogram. The histogram with its three humps is the reality. Each of these probability models ignores this reality and compulsively substitutes a bit of wishful thinking.

The second problem with these superimposed curves has to do with the way we see things—our eyes naturally tend to follow the smooth curves. As we do this our attention is drawn away from the gaps at 4 minutes and 8 minutes. In contrast, in Figure 6 our eyes tend to follow the tops of the bars, which makes the gaps easier to recognize. Thus, superimposed curves actually distract, interrupt, and can even distort the message of the histogram.

Finally, ignoring for the moment the fact that all three models shown are wrong for these data, how would you ever figure out which model provides the best fit for these data? You can always come up with many different curves to fit any one histogram. When you have a few hundred observed values, or less, you simply do not have enough information to make an informed choice between the different mathematical approximations. This means that the particular curve which your computer program may display on a histogram is more of a reflection of the assumptions of the computer programmer than a revelation of the characteristics of the data.

Edward R. Tufte coined a technical term for these superimposed curves—he called them "chartjunk." And that is exactly what they are—non-data ink. Since non-data ink obscures, distracts, and complicates your graph, and since the whole objective of using a graph is to communicate the message in the data, all non-data ink is chartjunk and should be rigorously avoided. So if you use a software package to create histograms, turn off the chartjunk!

## ITS NOT ABOUT THE MODEL

"But don't we need to know how the data are distributed?"

Yes you do, and that is exactly what the histogram does. It shows you how your data are distributed. It is your reality. Nothing more is required. When a histogram is created using appropriate-sized intervals it can reveal details that do not show up on the running record or *XmR* chart.

However, if by asking how the data are distributed you are thinking about branding and tagging your data with a probability model, then you are mistaken. When we start fitting models to our data we essentially impose our ideas upon the data, rather than listening to those data. Probability models are simply mathematical abstractions used to work out theoretical relationships. Such models are useful in *developing* techniques for the analysis of data, *but they have no direct role to play in the analysis of your data*.

This means, among other things, that you do not need to test your data for normality prior to using a Student's t-test, or a process behavior chart, or any other of a long list of robust statistical techniques. Neither do you have to fit a probability model to your data to compute the "correct" capability ratio. (For more on this topic see my column for June 2011.) If you have been taught otherwise, then you have been taught by a novice, and you need to apply for a refund of your tuition.

Moreover, any use of a probability model is equivalent to assuming predictability for the process generating your data. This is because the first step in defining a probability model is to assume that the random variables are independent and identically distributed. As we have seen in recent articles this is equivalent to assuming that the physical process is being operated predictably. Since a process behavior chart is the *only* operational definition of predictable operation, assuming a probability model for a histogram is like getting the cart in front of the horse.

Finally, histograms can take on many shapes that probability models do not match simply because histograms can come from unpredictable processes as well as predictable ones. Useful histograms show the warts and pimples in your data, unlike probability models which may well hide the interesting details.

## UNDERSTANDING THE PROCESS

By creating a histogram of the call lengths from Figure 1 we have gained some insight into the system. But is this one hour typical of other hours? Can we extrapolate from this one hour to characterize the process as a whole? To answer these questions we could plot the average call lengths for each hour on an *XmR* chart.

Say that the average call lengths for the previous 12 hours are: 9.29, 9.55, 9.84, 8.84, 8.95, 9.21, 10.30, 9.39, 8.40, 8.29, 8.50, and 7.68 minutes. These 12 values are the basis for the limits shown on the *XmR* chart in Figure 8, while the data from Figure 1 is shown as the last point.

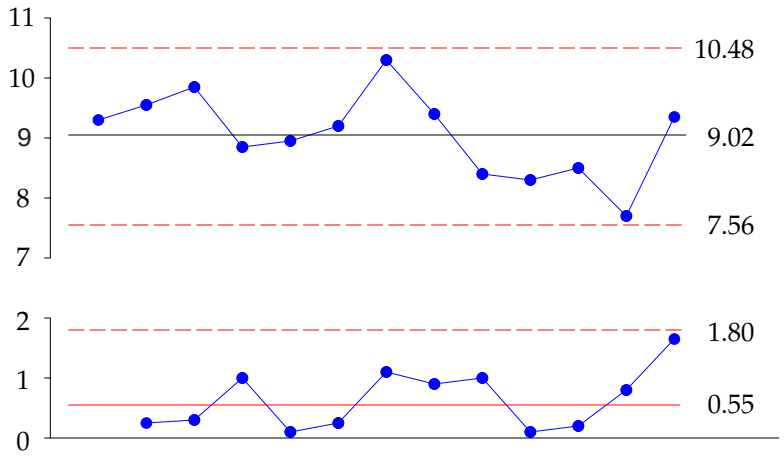


**Figure 8: XmR Chart for Hourly Average Call Lengths**

With an average call length of 9.02 minutes it is unreasonable to expect operators to average more than about 6.7 calls per hour. Therefore, 22 operators will be able to handle an average of up to 146 calls per hour. Higher volumes will require more operators. There is no evidence that this last value represents any change in the operation of the system.

## SUMMARY

A useful histogram will have intervals that are:

1. easy for the reader to interpret,
2. narrow enough to reveal the interesting details, and
3. wide enough to show how the data "pile up."

Thus, the art of creating a useful histogram involves finding the right balance between the simplification of the overall picture and the loss of detail that is necessarily part of that simplification. This balance point will differ with different data sets. It will depend upon the structure present in the data, the amount of data, and the purpose of the histogram. Constructing a stem-and-leaf plot is one of the best ways to get this feeling for the data, and so the stem-and-leaf plot can be considered to be a preliminary step in preparing a useful histogram.

If the data consist of fewer than 20 or 30 values, a simple dot-diagram may suffice. When the data consist of hundreds of values, some grouping will generally improve the picture. One common guideline is that there should be about 20 categories, or fewer, in the region where the

bulk of the data occur. In the past, elaborate formulas were given as guidelines for establishing the intervals for the horizontal axis of the histogram, but today it is usually simpler to try different sized intervals and let a computer do the work.

Finally, avoid the creation of misleading histograms by getting rid of the superimposed probability models supplied by your unknowing software.