Observations®
THE TECHNICAL JOURNAL
for SAS Software Users
www.sas.com/obs

Perusing Process Data with JMP® Histograms

Karen Copeland

Karen Copeland is an industrial statistician who works as an independent statistical consultant. Her areas of expertise include the design and analysis of experiments, general data discovery, and the transfer of statistical knowledge to practical knowledge for her clients. She has a Ph.D. in mathematics sciences from Clemson University. Karen has been using JMP® software for four years.

Abstract

As process data becomes more readily available, techniques for making sense of large amounts of data are needed. This article looks at a simple technique of sifting through process data with JMP histograms. The interactive nature of JMP software facilitates this technique and is discussed. Examples are given from a team session conducted at a chemical plant using JMP® 3.2.2 run on Windows 95.

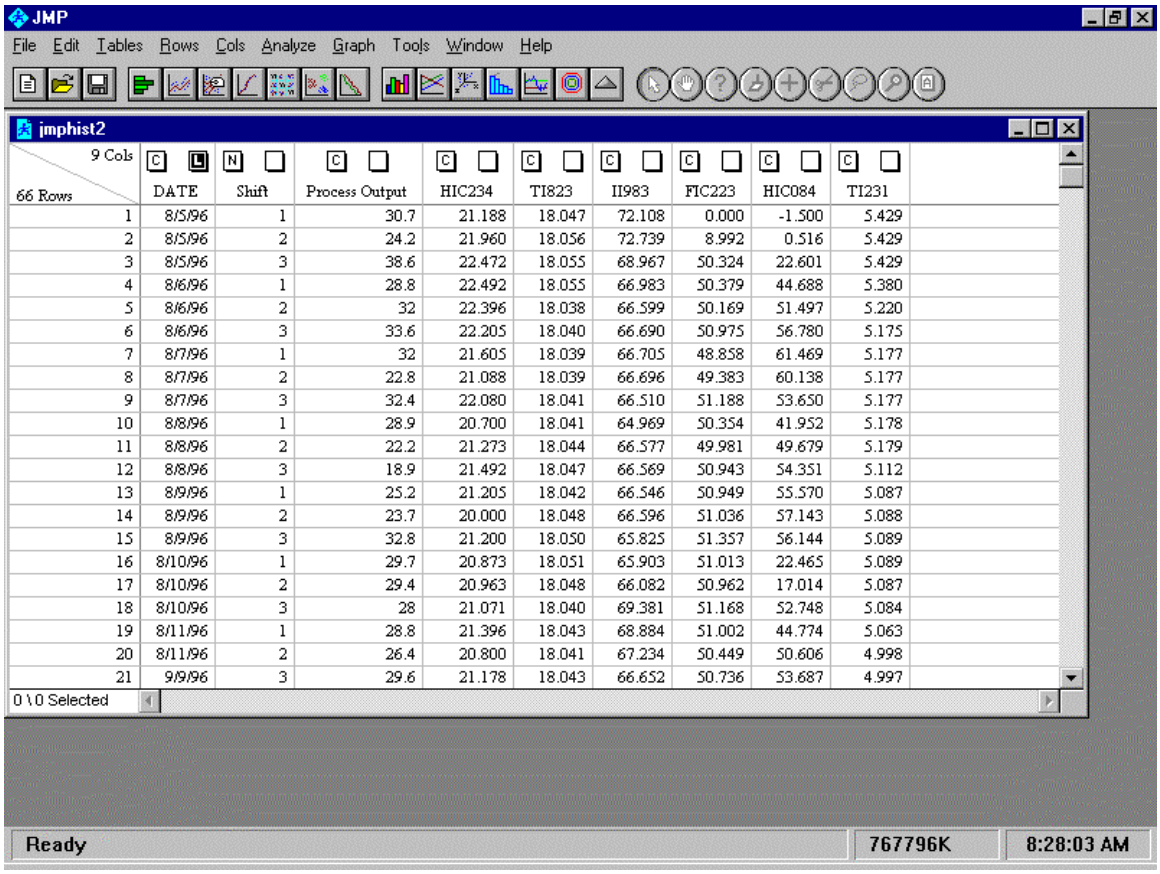
Contents

Introduction

One surprisingly simple way to begin sifting through process data is the use of JMP histograms and box plots. For each process variable, a histogram/box plot is constructed and examined using the interactive features of JMP software. This analysis is an ideal tool for a team problem-solving session. In a relatively short amount of time, a statistical/software expert and process experts can look for suspicious process variation, consider simple relationships among variables, and rule out process variables with little variation.

Histogram Basics

To begin, one must have their process data in a JMP data table. The data should be arranged in columns, with one column for each process variable, and additionally, at least one column for an identifier such as the date and/or time that corresponds to when the data was collected. The table shown in Display 1 was constructed from a real data table for illustrative purposes only. The values are averages of readings taken every five minutes over an eight hour shift of a chemical production process. The column names, which appear cryptic, are typical of what I have encountered with production data. In this data set, most of the values are readings from probes and/or control devices located throughout the process. For example, TI823 stands for temperature indicator #823, which is a temperature probe numbered 823. The process engineers will either know where and what this probe is measuring or they will have a diagram from which that information can be obtained. It is beneficial to have the process engineers, supply a “key” to the variable names. This would be a document that lists the variable names along with descriptive information. This information could be included in the JMP table by utilizing the **Notes** feature found in the **Column Info** dialog box.



Display 1 Example data table

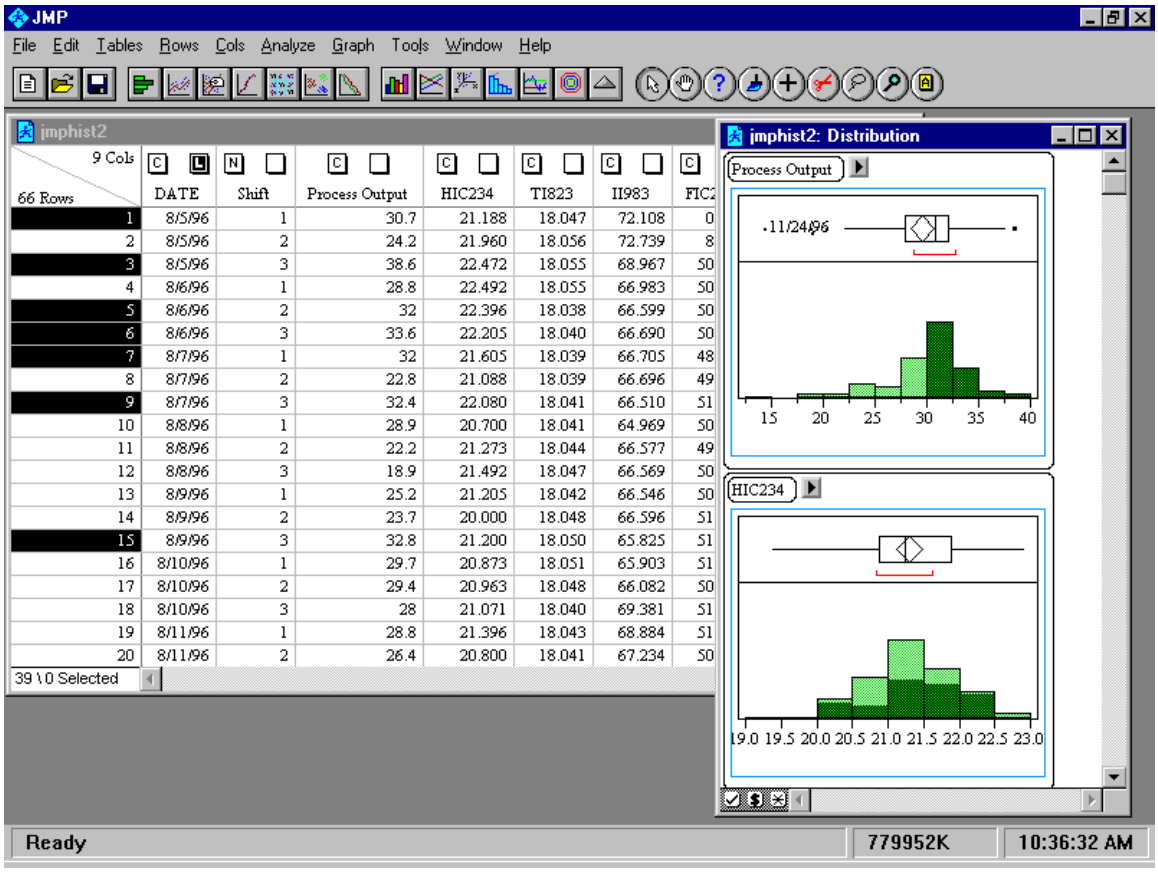
A histogram is a graphical representation of the distribution of a variable. To construct a histogram with JMP software, select **Distribution of Y** from the **Analyze** menu, or click on the tool bar button corresponding to the Distribution of Y (PC only). This action will result in a dialog box. Simply highlight the variable(s) of interest from the column list on the left and click on the **Add** button to add them to the list on the right side. When completed with the variable selection process, click on **OK**. The default output for continuous variables includes graphical output and text output. It is the graphical output that is of most interest here. To hide the text output from the screen, select **Text Report** from the display options pop-up menu (i.e., the check mark menu found in the lower left corner of the output window). Also, depending on your preference, you can change the orientation of the histogram from vertical to horizontal by selecting the **Horizontal Layout** option from the display options pop-up menu. For the **Distribution of Y** platform, all of the default settings of items found in the display options pop-up menu can be changed. To do so, select **Preferences** from the main **File** menu. Then from the **Analyze** pop-up menu in the **Preference** window, select **Distribution of Y**. Set the desired items to fit your preferences and select **Save**.

Interactive Histograms

There are two JMP software features that facilitate learning from histograms. First, there is the ability to identify extreme data values, which are indicated in box plots as single points plotted beyond the "whiskers". Clicking on a point will identify the corresponding data table row number for that point. In addition, the corresponding row in the data table will now be highlighted. This feature is especially effective with the proper choice of a label column (such as the date). To select a column to use as the data label, click on the right square found in the data table at the top of the desired column, and select **Label** from the role assignment pop-up menu. That column will now have an L in the right column heading square, indicating

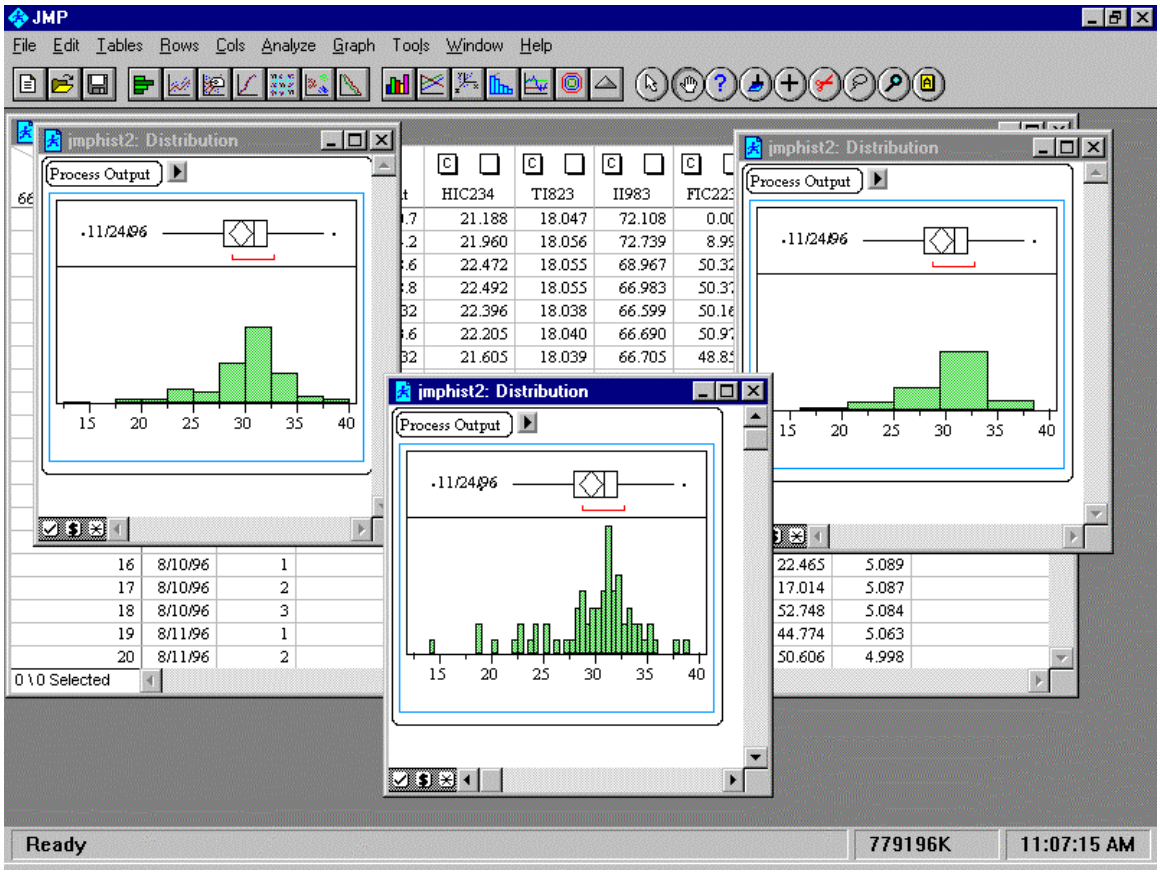
that it is the label column. This is illustrated in Display 1 by the DATE column. Now when a point is selected on the box plot, the entry from the label column, rather than the row number, will be shown.

The second feature of interest is the interactive shading of multiple histograms. On an individual histogram, one can highlight a bar or multiple bars by clicking on the bar (hold the shift key while clicking for multiple bars). The shading of the histogram also highlights the corresponding rows in the data table, as well as all other histograms in the same or in different, yet open, analysis windows. This allows one to highlight a range of values on one histogram, such as high production rates, and then look at other histograms to see the distribution of the corresponding values. These features are illustrated in Display 2.



Display 2 Sample JMP output illustrating the interactive histogram shading

To prepare for an interactive histogram session, a sufficient amount of process data is needed. The question of how much is sufficient will depend on the scope of the problems under consideration, and perhaps more importantly, how much is available. Histograms are sensitive to changes in bin width and midpoints for small amounts of data. One can easily change the bin width and/or midpoints in JMP histograms by using the hand tool. Select the hand tool from the **Tools** pull-down menu or from the tool bar (PC only). While holding down the first mouse button, drag the hand up to decrease or down to increase the bin width. Dragging left and right will shift the midpoints of the bins. The larger the data set, the less sensitive the histogram shape will be to these changes. Display 3 shows three histograms, all of the same data, where each has different bin widths and/or midpoints.

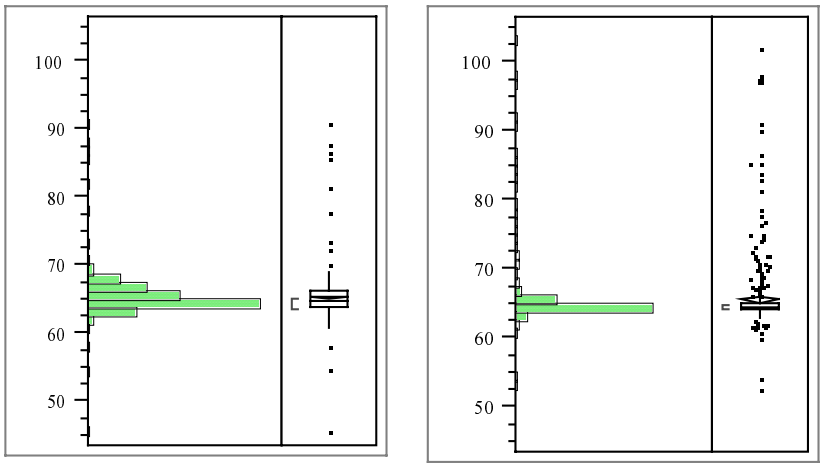


Display 3 Illustration of changing histogram bin widths and midpoints

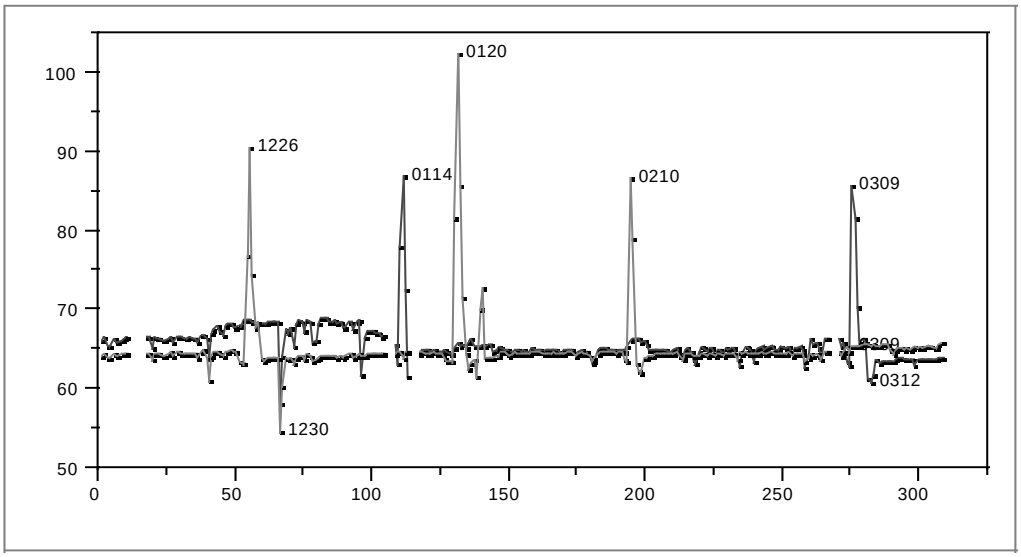
Examples

What might a team expect to learn from a **Distribution of Y** analysis session? Given the right group of participants, one can expect to come away from such a session with a number of process improvement ideas, proof of process improvement successes, and most likely questions about the process that need further research. For example, a team consisting of myself (the statistician), the plant SPC/Quality coordinator, the process manager, a process engineer, and an operations engineer met after collecting over three months of data. The data consisted of eight-hour shift averages (calculated from values collected every 5 minutes) from a chemical production process. In two hours, we looked at over 100 process variables with two goals in mind. First, the team wanted to identify any variables relating directly to periods with high production output, and second, the team wanted to identify variables that we could exclude from further analysis (e.g., variables with little or no variation, such as a set point that is never changed).

Many discoveries exceeding our goals were made in those two hours. For example, in Output 1, the distribution of two temperature gauges, one on each of two similar tanks, show nearly identical average values but different variation. An overlay plot (graph menu) further showed that the temperature with more variation started out higher than the other, made a step change that aligned the two temperatures, and then made another step change that put it below the other temperature. See Output 2. This was an action item for the engineers to investigate.

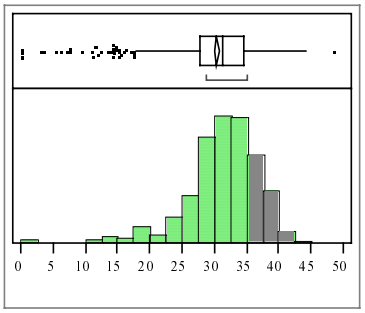


Output 1 Histograms of two temperature probes, one on each of two similar tanks, indicating a difference in temperature variation

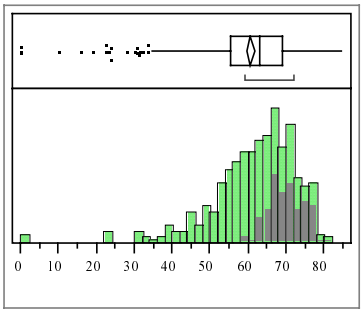


Output 2 Overlay plot of two temperature probes on each of two similar tanks

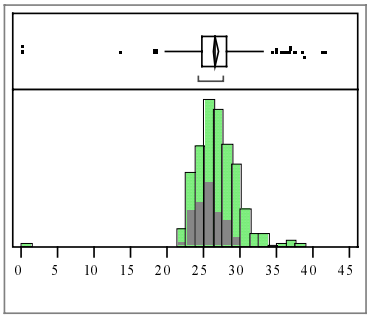
Output 3 illustrates the use of the histogram shading to look at relationships between variables. The histogram in Output 3a is shaded for high production rates (recall the goal of looking for process variables that were clearly linked to high output). The histogram in Output 3b illustrates a process variable that tends to be high when production rates are high as the shading is concentrated to the high side of the histogram. The histogram in Output 3c was a surprise to the engineers because they held the belief that high production rates were achieved when this process variable was maxed out. However, the histograms illustrates high production rates are actually achieved when this process variable is running at a moderate level. This finding led to a change in operating procedure to maintain this process variable at the ideal level identified in the histogram.



Output 3a: Production rate, shaded for desired level



Output 3b: Process variable that correlates to high production rates



Output 3c: Process variable that correlates to production rates contrary to previous beliefs

Conclusion

Given a large amount of process data, interactive histograms provide a simple first look at the data, and by looking at the data, one can begin to learn from the data. This technique can be used to investigate a particular problem, such as why a quality characteristic was high in a given time period, or as a starting place for further work, such as process modeling or monitoring schemes.

Questions and comments should be directed to:

Karen A. Copeland, Ph.D.
230 Shandwick Place
Alpharetta, GA 30004
Phone: (678)297-7738
E-mail: copeland@inetnow.net