Much of statistical research consists of measuring or estimating the properties of a given variable or variables. In many cases, what we really want to know is not, say, what the mean and standard deviation of a particular variable are but how that variable is distributed, and estimators such as the mean and standard deviation are only of value to the extent they help us judge this. In almost all such cases, the best technique is simply to look at the data, that is, to use a histogram. PROC CHART and PROC GCHART provides us with the ability to plot histograms quickly and easily. Unfortunately, neither PROC CHART nor PROC GCHART’S default values are the best, that is, they do not allow us to use all of the information in our data. Fortunately, by doing some data step and macro programming we can get all or almost all of the information the data possesses.

Histograms are extremely powerful techniques for representing and analyzing distributions. A histogram is a graph of vertical bars where the areas are proportional to the frequencies represented. A distribution is a description of a variable that relates the probability of the variable taking a given value or range of values to each value or range of values the variable might take. A variable is simply a value that can vary as opposed to a constant which cannot. A probability is a number between zero and one inclusive where zero indicates that an event cannot occur and one indicates that the event must occur.

If the variable is discrete, that is, if the variable can only take on certain specific values, the procedure is particularly easy. For each value one simply counts the number of times that value occurred and plots (see Exhibit 1).

When the data is continuous, the procedure is slightly more complicated. The statistician must set up bins which are non-overlapping ranges of values. For each bin one counts the number of times a value falls within that bin and plots. The height of each bar on a histogram represents the density of the distribution for the bar's width. If the bars are equally wide, the heights of the bars represent density per bin, that is, per unit of the underlying variable. Histograms should generally be adjusted for sample size. When this is done the area or heights, as the case may be, represent proportions or probabilities (see Exhibit 2).

If equal bin widths are used there would seem to be only two subtleties: how wide to make the bins and where to locate the edges. PROC CHART and PROC GCHART allow the investigator to make these decisions explicitly, or to default and let SAS® make them. Unfortunately, while SAS has reasonably good default values, the values are virtually never optimal. Worse, it’s hardly ever obvious what optimal values are.

Edge location is generally not a problem. In most real world situations at least one bin edge is important. For example, it may be important to know what proportion of the population is larger than zero. Once one edge location has been decided the other edge locations follow naturally.

Neither PROC CHART nor PROC GCHART allow you to specify and present the bin edges. Rather, they demand that the data be analyzed and presented by the midpoints of the bins. While in many cases, this is an inconvenience, it rarely presents a serious problem. In my experience, in almost all cases PROC CHART and PROC GCHART midpoint selection default values are eminently reasonable. When they are not reasonable both the bin edges or midpoints and the number of bins must be changed.

Selecting the appropriate bin width is a subtler problem. If too many bins are used, the histogram will be too rough. It will vary widely and randomly from the distribution it is representing. If too few bins are used, the histogram will be too smooth. The histogram will now systematically overestimate and underestimate the distributions it is representing (see Exhibit 3).

Unfortunately, the optimal bin width depends on the distribution the histogram is representing. This is never known, of course. Indeed, we generally use histograms precisely because we don’t know how the variable of interest is distributed.

Most beginning statistical tests recommend selecting an arbitrary number of bins between 6 and 15. Slightly more advanced or comprehensive tests generally involve Sturge’s rule.

Here the number of bins (B) is:

\[ B = 1 + 3.3 \log_{10} N \]

Where:

\[ N = \text{the sample size} \]

\[ \log_{10} \text{is the common or base 10 logarithm.} \]

If, for example, \( N = 144 \) then:

\[ B = 1 + 3.3 \times 1.144 \]

\[ = 1 + 3.3 \times 2.1584 \]

\[ = 1 + 7.1227 \]

\[ = 8.1227 \]

Or, rounding down, 8.0. To calculate bin widths, divide the range of the sample by the number of bins. This is precisely the approach SAS takes in calculating a default value for.
Recent research by Terrell and Scott\(^1\) has shown that, given that the distribution does not exist outside of some finite interval, at least \((2n)\exp(1/3)\) bins are required for an optimal histogram. In most of the most interesting cases this will not be true. A normal distribution, for example, has no highest or lowest value. On the other hand, in all cases our histograms will be charts of samples, rather than charts of the distributions we are really interested in, and samples always have a highest and lowest point. Dividing the range \((R)\) of the sample by the minimum number of bins gives the maximum bin size, of course. For example, if \(R = 1000\) and \(N = 144\), then:

\[
\frac{1000}{(2\times144)^{1/3}} = 1000/(288)^{1/3} = 100/6.6039 = 151.4267
\]

Or, rounding down, 150.

If the data is normally distributed, and much real world data is normally or approximately normally distributed, the optimal bin width \((w)\), according to David Scott\(^2\), is:

\[
w = 3.49Sn^{1/3}
\]

where:

\(S\) = the sample standard deviation

For example, if \(S = 14.524\%\) and \(N = 144\), then:

\[
w = 3.49 \times 14.524 \times 144^{-1/3} = 3.49 \times 14.524 \times 0.1908 = 9.671
\]

Scott's technique will not torture non-normal data into the shape of a normal curve. Therefore, if, after the data is plotted, a reasonable approximation of a normal curve appears, Scott's technique was appropriate. However, when the histogram does not resemble a rough approximation of a normal curve, Scott's technique is inappropriate and the data might be replotted using a different bin width.

Unfortunately, when the data is not normally distributed, the mathematics are, at best, quite complicated and for many distributions there is no known optimal solution.

If the statistician only wants to chart one variable, whose sample size he knows, the best approach is certainly to use the appropriate formula above and calculate the value by hand and set PROC CHART'S or PROC GCHART'S levels equals option to the recommended number of bins. More important than using the correct bin width or midpoint location is the production of comparable charts over BY variables, of course. Unfortunately, PROC CHART'S levels equals option does not insure that the charts produced will be comparable. If the levels equals options is used all of the charts will have the same number of bins but they will not be the same bins, that is, they will not have the same widths or locations.

However, if the user is going to calculate a number of variables, with or without BY variables, it is better to explicitly calculate the bin midpoints and set them in a midpoints equals option. These values can be calculated in a DATA STEP or a PROC MEANS and DATA STEP and then transferred to the PROC CHART or PROC GCHART by using the SAS MACRO capability.

Scott's and Sturge's techniques, though valuable, give the impression of more accuracy than they possess. Histograms are graphic techniques and the final choice of bin width and edge location must be made on graphic criteria. At least three criteria are important. First, the histogram should not imply greater accuracy than exists in the data. If the data can only be measured to within, say, a hundred units, the bin width must not be smaller than a hundred units. Second, the edge locations chosen should minimize the number of cases falling on or near the bin boundaries. Third, the bin range should have some psychological, economical or other value. If, for example, Scott's technique suggests a bin width of 108, round up to 110.00 or, better, down to 100.00.

Unfortunately, whatever techniques are used to choose bin widths and select bin edges, the histograms themselves will invariably give the impression of an accuracy they do not possess. A histogram seems to be a graph of a distribution of interest, yet, in all of the most interesting cases, it is not (see Exhibit 4). Rather, it is a graph of a sample or estimate of the distribution and like all estimates it should minimize the number of cases falling on or near the bin boundaries. If the data is normally distributed, and much real world data is normally or approximately normally distributed, the optimal bin width \((w)\), according to David Scott\(^2\), is:

\[
w = 3.49Sn^{1/3}
\]

where:

\(S\) = the sample standard deviation

For example, if \(S = 14.524\%\) and \(N = 144\), then:

\[
w = 3.49 \times 14.524 \times 144^{-1/3} = 3.49 \times 14.524 \times 0.1908 = 9.671
\]

Scott's technique will not torture non-normal data into the shape of a normal curve. Therefore, if, after the data is plotted, a reasonable approximation of a normal curve appears, Scott's technique was appropriate. However, when the histogram does not resemble a rough approximation of a normal curve, Scott's technique is inappropriate and the data might be replotted using a different bin width.

Unfortunately, when the data is not normally distributed, the mathematics are, at best, quite complicated and for many distributions there is no known optimal solution.

If the statistician only wants to chart one variable, whose sample size he knows, the best approach is certainly to use the appropriate formula above and calculate the value by hand and set PROC CHART'S or PROC GCHART'S levels equals option to the recommended number of bins. More important than using the correct bin width or midpoint location is the production of comparable charts over BY variables, of course. Unfortunately, PROC CHART'S levels equals option does not insure that the charts produced will be comparable. If the levels equals options is used all of the charts will have the same number of bins but they will not be the same bins, that is, they will not have the same widths or locations.

However, if the user is going to calculate a number of variables, with or without BY variables, it is better to explicitly calculate the bin midpoints and set them in a midpoints equals option. These values can be calculated in a DATA STEP or a PROC MEANS and DATA STEP and then transferred to the PROC CHART or PROC GCHART by using the SAS MACRO capability.

Scott's and Sturge's techniques, though valuable, give the impression of more accuracy than they possess. Histograms are graphic techniques and the final choice of bin width and edge location must be made on graphic criteria. At least three criteria are important. First, the histogram should not imply greater accuracy than exists in the data. If the data can only be measured to within, say, a hundred units, the bin width must not be smaller than a hundred units. Second, the edge locations chosen should minimize the number of cases falling on or near the bin boundaries. Third, the bin range should have some psychological, economical or other value. If, for example, Scott's technique suggests a bin width of 108, round up to 110.00 or, better, down to 100.00.

Unfortunately, whatever techniques are used to choose bin widths and select bin edges, the histograms themselves will invariably give the impression of an accuracy they do not possess. A histogram seems to be a graph of a distribution of interest, yet, in all of the most interesting cases, it is not (see Exhibit 4). Rather, it is a graph of a sample or estimate of the distribution and like all estimates it will contain two types of errors: statistical and nonstatistical. Statistical error is a function of the sample size and the technique used to make the estimate. The larger the sample is and the larger the sample is as a proportion of the population of interest, the smaller, usually, the statistical error. Statistical technique can be used to minimize statistical error independent of the sample size. More correctly, highly sophisticated techniques use more of the available information than less sophisticated techniques do. For example, the techniques discussed here estimate the height of a bin by counting the values within that bin. More advanced techniques will estimate the height for a given bin by comparing the count for that bin with counts of the data in the other bins. Bins close to each other will affect each other heavily while bins far from each other will only affect each other a bit. For example, the value of bin 6 will be estimated primarily by the count in bin 6 but also by the counts in bins 5 and 7. Bin 6 will also be affected, at least marginally, by the values in bins 1 and 10. At the same time, the count in bin 6 will affect the estimates of the values of bins 5, 7 and so on (see Exhibit 5). In other words, a complicated weighted averaging process is used.

Currently, there is no way to implement these techniques using either PROC CHART or PROC GCHART. If the statistician wants to use them he must program them himself. Unfortunately,
Most of the available techniques vary from the complicated to the extremely complicated. However, at least one quick but not too dirty approach is available. This technique uses the actual count for a given bin as a preliminary estimate for that bin, then it sees how far that estimate deviates from the value one might 'expect' for that bin based on a moving average of the adjoining bins' estimates. The values for each bin are then adjusted toward what might be expected, the adjustment being based on how much the bins as a group deviate from their expected values. Technically, the adjustments are made using James-Stein's estimators. Strictly speaking the data is not adjusted. Rather, a pair of summary variables is created and these variables are used to create the chart. One variable is simply a listing of the midpoints of the bins of the variable of interest. It is this variable that is used as the charted variable in the PROC CHART or PROC GCHART. A second variable contains the counts or adjusted counts for each bin. This variable is used in the Proc FREQ equal option. The midpoints equal options must also be specified as noted above.

Nonstatistical error is a function of the sampling procedure. A histogram is a graph of a distribution but it is not necessarily a graph of the one desired. Good researchers worry a lot about what sample they should be taking and what sample they really took.

REFERENCES


SAS is a registered trademark of SAS Institute Inc., Cary, NC, U.S.A.