Chapter 143

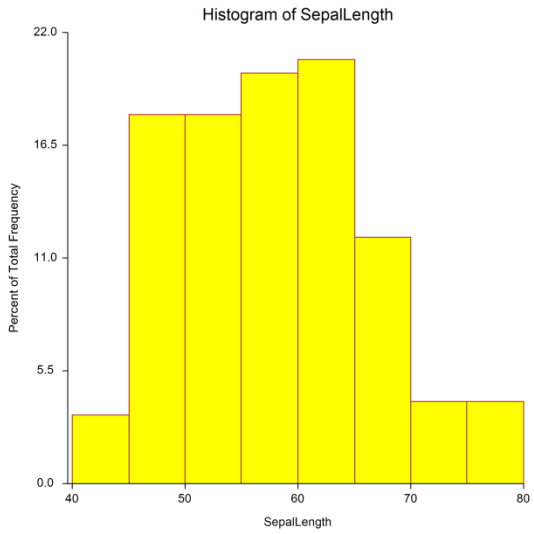
Histograms

Introduction

The word *histogram* comes from the Greek *histos*, meaning pole or mast, and *gram*, which means chart or graph. Hence, the direct definition of "histogram" is "pole chart." Perhaps this word was chosen because a histogram looks like several poles standing side-by-side.

A histogram is used to display the distribution of data values along the real number line. It competes with the probability plot as a method of assessing normality. A histogram is created by dividing up the range of the data into a small number of intervals or bins. The number of observations falling in each interval is counted. This gives a frequency distribution.

A *histogram* is a graph of the frequency distribution in which the vertical axis represents the count (frequency) and the horizontal axis represents the possible range of the data values.



Density Trace

The histogram is widely used and needs little explanation. However, it does have its drawbacks. First, the number and width of the intervals are a subjective decision, yet they have a high impact on the appearance of the histogram. Slightly different boundary values can sometimes give dramatically different looking histograms, especially when the number of values used to create the histogram is small.

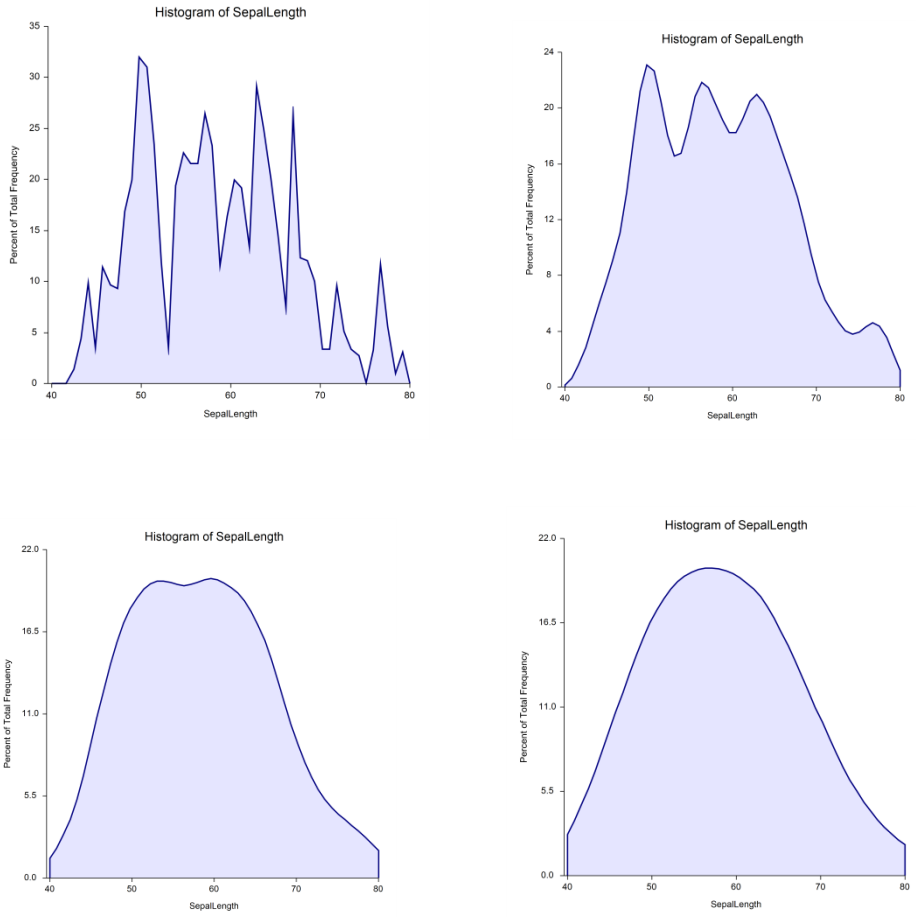
Another problem with the histogram is that the rectangles make it appear that the data are spread uniformly throughout the interval, which is rarely the case. Also, the "skyscraper" look of the histogram doesn't resemble the rather smooth nature of the data's true distribution.

Histograms

These issues with the histogram have brought many new innovations. One of the more popular display techniques for showing the distribution of data is the density trace.

Density refers to the relative frequency (concentration) of data points along the data range. Mathematically, the density at a value x is defined as the fraction of data values per unit of measurement that lie in an interval centered at x. Once you choose a suitable interval width, you can calculate the density at any (and every) x value. If you calculate the density at, say, 50 values and connect them, you'll have a density trace.

In **NCSS**, the interval width is specified as a percentage. As you increase the percentage, you increase the amount of data included in each density calculation. This increases the smoothness of the chart. The following four density traces were made of the same data at increasing percentage smoothness.



As the interval width is increased, data points further and further from the center value are included. In order to decrease the weight of points that are far removed from the center value, we use a weighting scheme that weights points proportionally to their distance from the center value. The weight function used is half the cosine function with its peak at the center value. It decreases symmetrically to zero, after which a weight of zero is applied. Hence, points have a smaller and smaller impact on the density trace as they are further and further from the center.

Another way to think of the density trace is to imagine that you construct 1000 histograms of the same data using slightly different boundary positions and take the average rectangle height at each of 50 values along the data range. This would give you a smoothed histogram that has many of the same properties of the density trace. Hence, the density trace should be thought of as a smoothed histogram in which interval width and number of bins do not come into play.

Data Structure

A histogram is constructed from the values of a single column. A second variable may be used to divide the first variable into groups (e.g., age group or gender). In this procedure, a separate histogram is produced for each group.

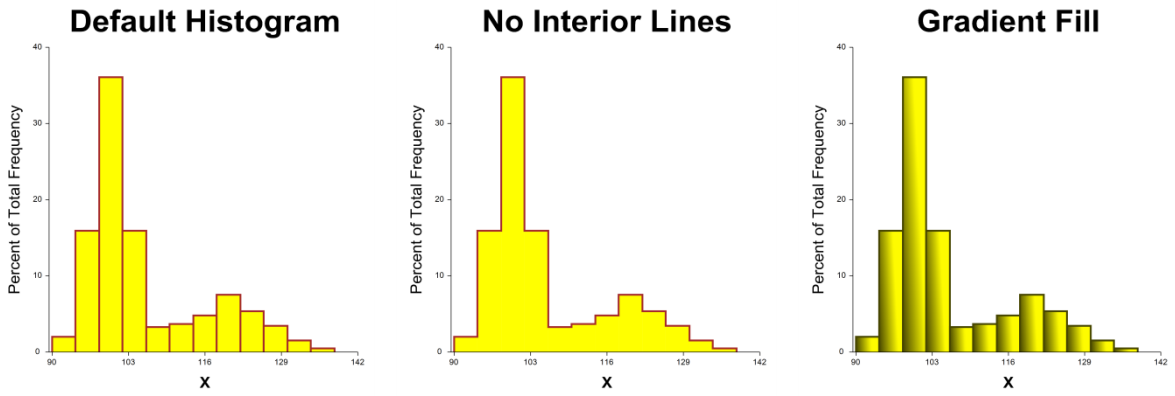
Histogram Window Options

This section describes the specific options available on the Histogram window, which is displayed when the Histogram button is clicked. Common options, such as axes, labels, legends, and titles are documented in the Graphics Components chapter.

Histogram Tab

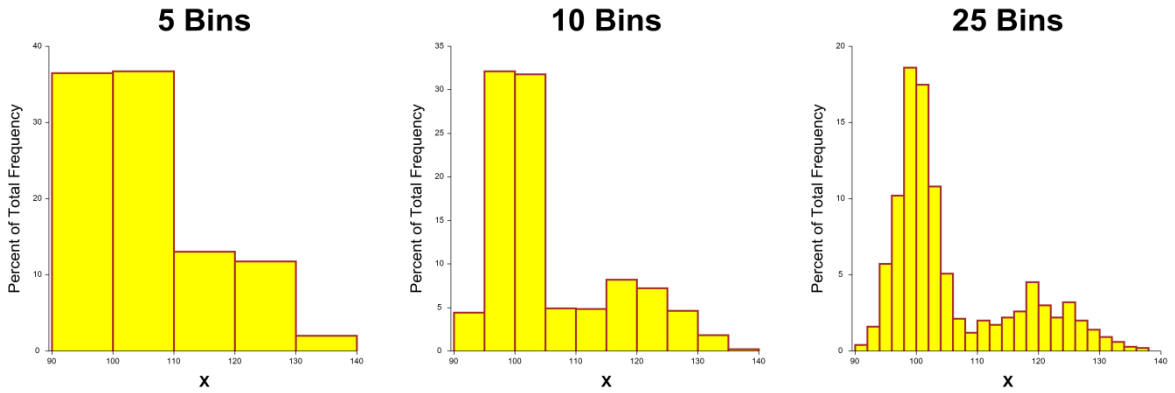
Format Section

You can modify the color of the histogram and its outline using the options in this section. The third example uses a brown to yellow gradient fill.



Bins Section

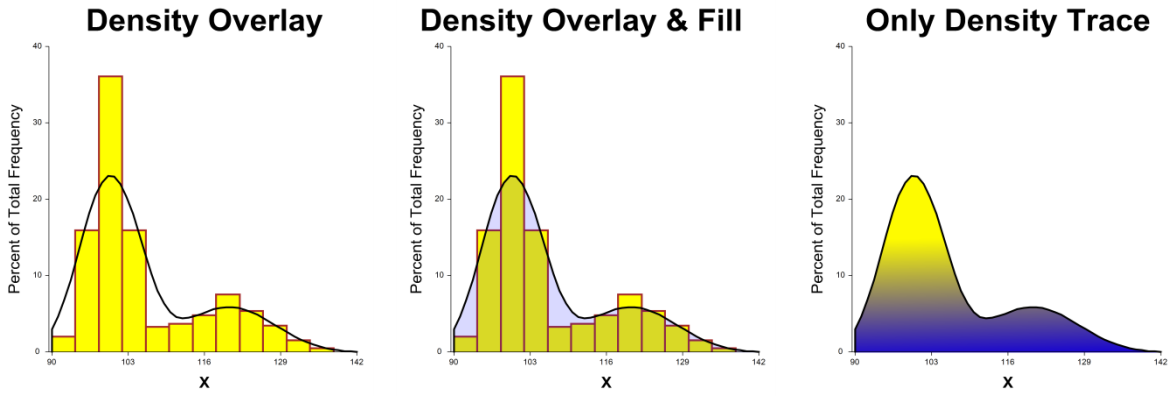
You can specify the number of bins (bars) of the histogram in several ways.



Overlays Tab

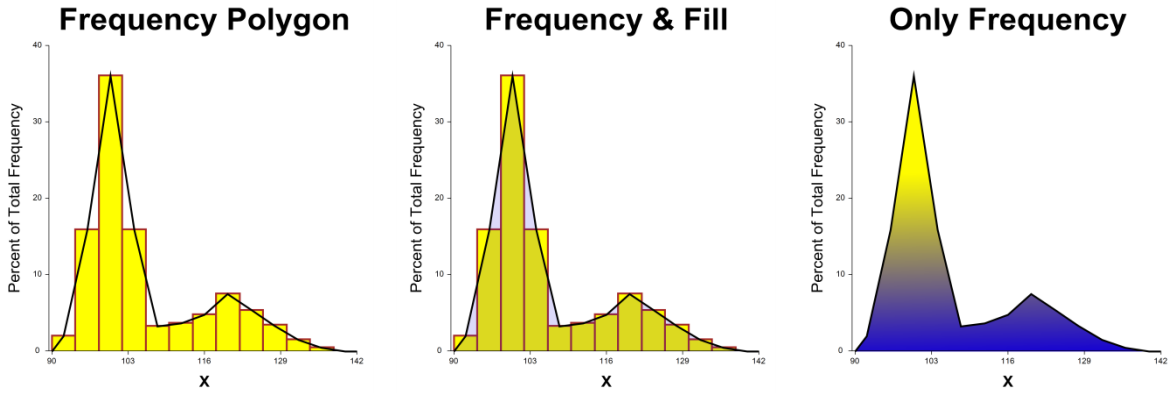
Density Section

You can add a density trace line over the histogram. This serves as a smoothed histogram. Note that the third example uses a blue to yellow gradient fill.



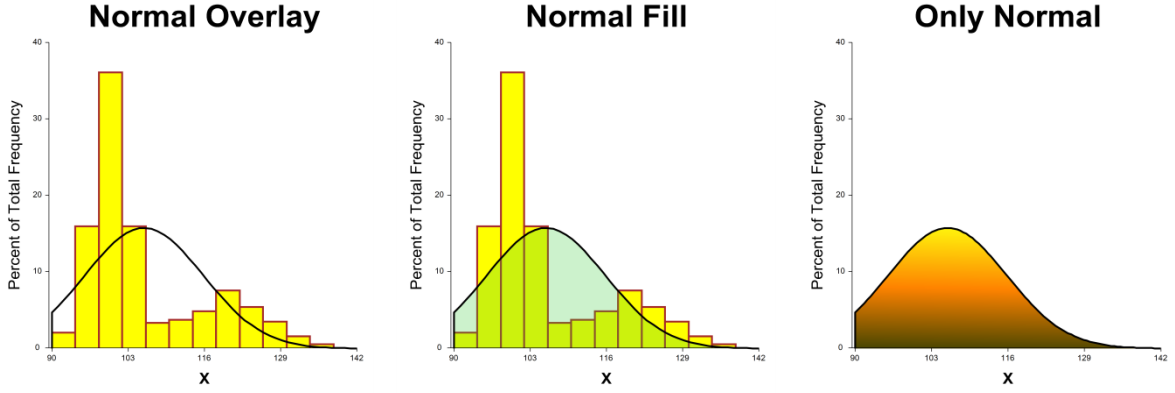
Frequency Section

You can add the frequency polygon line over the histogram. This line connects the top midpoints of each bar.



Normal Density Section

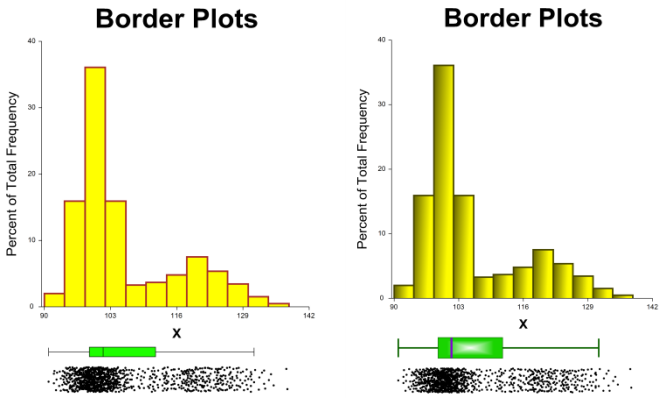
You can add a normal density line over the histogram. This line is based on the data's mean and standard deviation. Note the impact of the brown to orange to yellow gradient in the third example.



Border Plots Tab

X Axis Section

You can add a box plot and a dot plot underneath the histogram to give a very clear picture of the density of the data.



Titles, Legend, X Axis, Y Axis, Grid Lines, and Background Tabs

Details on setting the options in these tabs are given in the Graphics Components chapter.

Example 1 – Creating a Histogram

This section presents an example of how to generate a histogram. The data used are from the Fisher dataset. We will create a histogram of *SepalLength*.

Setup

To run this example, complete the following steps:

1 Open the Fisher example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Fisher** and click **OK**.

2 Specify the Histograms procedure options

- Find and open the **Histograms** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 1** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables Tab	
Data Variable(s)	**SepalLength**
Histogram Format *(Click the Button)*	
Overlays Tab	
Outline (Density)	**Checked**
Border Plots Tab	
Box Plot	**Checked**
Dot Plot	**Checked**
Report Options (*in the Toolbar*)	
Variable Labels	**Column Names**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Histogram Output

Histogram Section