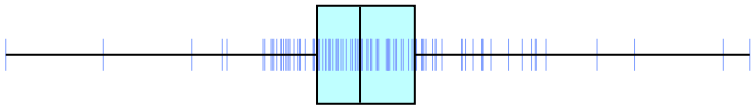
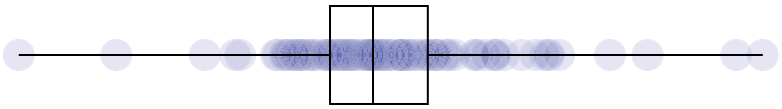
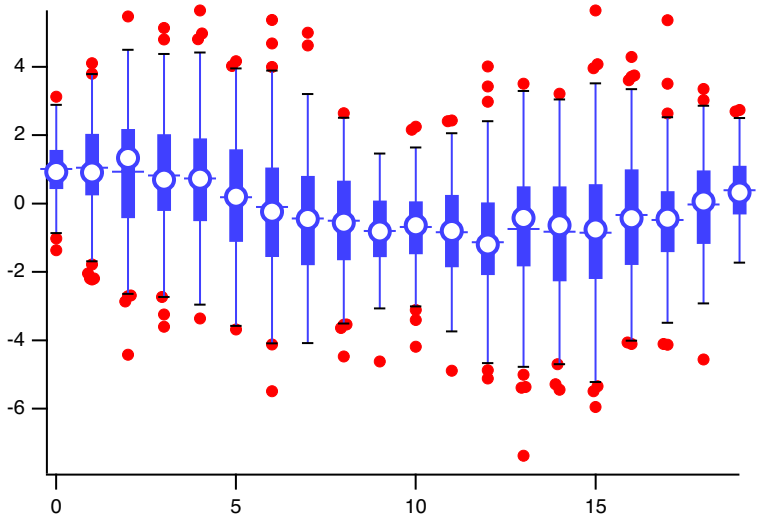


Large markers with transparency can also give a sense of data density. This plot uses the solid circle marker (number 19) with a blue color with alpha set to 0.1. The fill color for the box is turned off, as the background will show through the transparent markers and affect the color.



Sometimes the number of datasets and data points can be overwhelming. Here is an alternative look that makes a more compact display:



The whisker method is set to Mean ± f*SD, with the factor set to 2, so the whiskers show a span of 2 standard deviations about the mean. The outlier method is set to Whiskers, so only data points beyond the ends of the whiskers are shown. That limits the shear number of data points on the plot.

There are many datasets, so the boxes are quite narrow and filled to make a solid box. Instead of a hard-to-see median line, the median is shown as a white-filled circle marker. To prevent the box outlines from showing on top of the white marker fill, the Draw Median Marker On Top checkbox is turned on in the Markers tab of the Modify Box Plot dialog.

Finally, the mean is shown as a horizontal bar marker. The marker size is set large enough that it shows outside the median circle marker.

Box Plot Reference

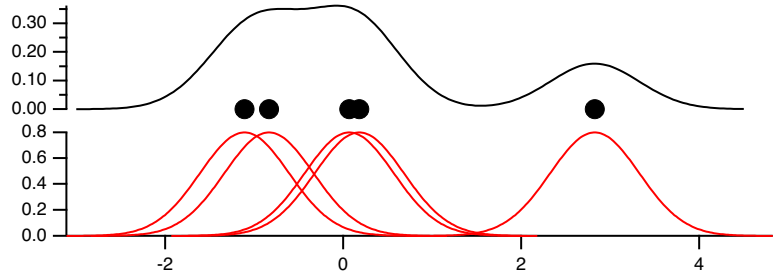
Tukey, John W., Exploratory Data Analysis, Addison-Wesley Publishing Company, 1977.

Violin Plots

A violin plot, also called a bean plot, is a way to summarize the distribution of data. A violin plot shows the distribution of a dataset using a kernel density estimate (KDE). The KDE creates a smooth estimate of the

underlying data distribution by summing some kernel function, one function per data point. The summed curve is then normalized to an area of 1.0 so that it is an estimate of the probability distribution function for the dataset.

Suppose you have five points drawn from a Gaussian distribution, such as the points represented by the black dots here:

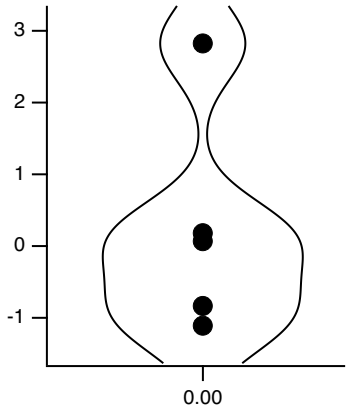


Generate a Gaussian curve for each point (the red curves), then sum the curves and normalize for an area of one (the black curve). In this plot we have arbitrarily selected a standard deviation for the red curves of 0.5; this is referred to as the bandwidth when computing a KDE curve or violin plot.

Now we have a smooth curve that gives a possible representation of the underlying distribution from which the data points were drawn. Quite possibly the kernel bandwidth we used was too small, and unjustified by the small number of points. The choice of kernel function, Gaussian in this example, and the width of the kernel are somewhat arbitrary. The Gaussian kernel is in some sense "smooth" and reflects our bias that most data follow a Gaussian distribution. Others are possible.

You can compute a KDE for your datasets yourself using the **StatsKDE** operation, but to get from there to a violin plot is quite tedious. Igor does a lot of this work for you when you create a violin plot trace by executing the steps described under **Creating Box Plots and Violin Plots** on page II-330.

In a violin plot, the curve is in general plotted vertically and reflected across the midline to give a plot that looks somewhat like a violin or a green bean pod with lumps for each seed. Here is Igor's violin plot of the five points shown above, modified to show the raw data points and to use black coloring:



In this plot, the width of the curves is not meaningful except in a relative way. The centerline indicates zero estimated probability density. Because it is the only dataset included in the trace, and it is not a category plot, it is positioned at zero on the X axis.

For the following illustrations, we made two representative fake datasets: