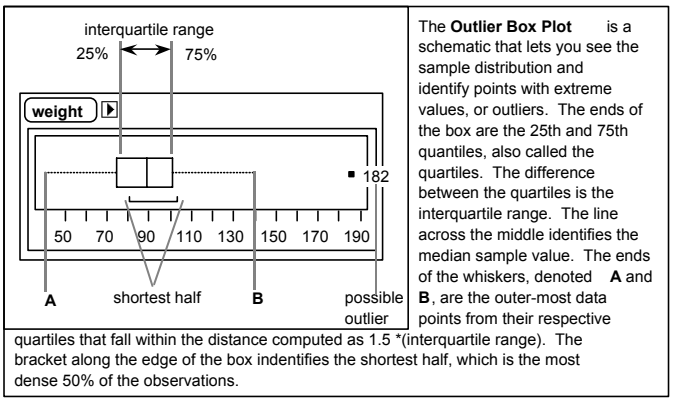
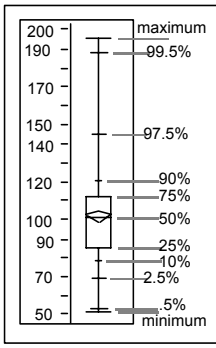
Boxplots (also called box-and-whisker plots)

Advantages: -compact, concise, simple to draw.

Disadvantages: -obscure many finer features of distribution
-emphasize tails of distribution (which are most uncertain/unstable)
-there are different conventions for what boxplot symbols mean.
Here are two common conventions:



The **Outlier Box Plot** is a schematic that lets you see the sample distribution and identify points with extreme values, or outliers. The ends of the box are the 25th and 75th quantiles, also called the quartiles. The difference between the quartiles is the interquartile range. The line across the middle identifies the median sample value. The ends of the whiskers, denoted **A** and **B**, are the outer-most data points from their respective quartiles that fall within the distance computed as 1.5 *(interquartile range). The bracket along the edge of the box indentifies the shortest half, which is the most dense 50% of the observations.



The quantile box plot shows selected quantiles on the response axis. The box shows the median as a line across the middle and the quartiles (25th and 75th percentiles) as its ends. The means diamond identifies the mean of the sample and the 95% confidence interval about the mean.

Boxplots sometimes include notches that describe the expected range of variability of the median. The notches are defined by the median, plus or minus its standard error:

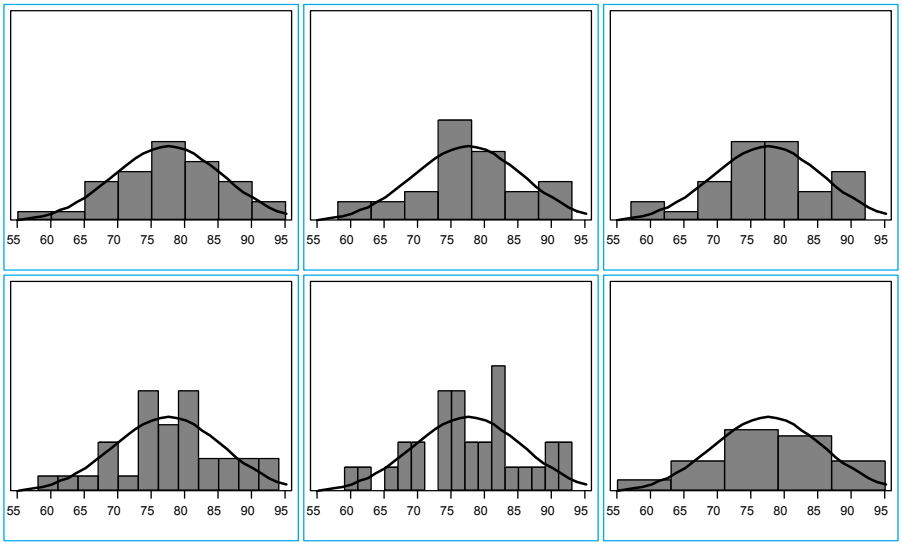
$$notch\ edges = median \pm 1.57 \frac{interquartile\ range}{\sqrt{n}}$$

Where the interquartile range is the difference between the 75th and 25th quantiles, and n is the number of observations. If the notches between two plots do not overlap, then (under certain restrictive statistical assumptions) the medians can be judged to be different with 95% confidence.

Histograms

Advantages: -widely used, familiar, needs no explanation
-simple to draw/plot

Disadvantages: -contain no information on distribution of variables within bins
-sensitive to number/width/placement of bins.
For example, all of these histograms represent *exactly the same data*:



Hazards: -if bin width changes within a histogram, the results can be wildly misleading.

Density traces (approximate probability density functions)

Advantages:
-familiar and easily explained
-represent the density of the data in intuitively obvious form
-avoid histograms' sensitivity to bin width and bin placement

Disadvantages:
-smoothness depends on arbitrary choice of window width

How to (#1): For a distribution of observations x_i, i=1..n, define the local data density at any point x as:

$$\textit{local density at } x = \frac{\textit{number of } x_i \textit{ such that } x - h/2 < x_i < x + h/2}{h}$$

Then plot this density as a function of x (note that the variable x need not be one of the observations x_i). This yields the average number of observations per unit of measurement, averaged over a window of width h centered around the point x^*. You must choose the window width h that you want to average the density over. Larger values of h give a smoother curve, but (for that reason) will obscure any abrupt changes in data density. Smaller values of h will show more detail, which may be spurious, particularly if n is small.

The data density calculated above is the number of observations per unit of measurement. In some circumstances (such as comparing data sets of different sizes) one wants instead the fraction of observations per unit of measurement. That can be obtained simply by replacing h by hn in the denominator of the expression above.

How to (#2): The window over which the data are averaged above is square; each observation in the window counts equally, whether it is close to the center at x, or near one of the edges at $x \pm h/2$. This leads to roughness in the density trace, as individual data points enter and leave as the window is scanned across the x axis. A smoother trace can be obtained if points near the edge of the window are weighted less. One such weighting scheme is as follows:

First, calculate the distance between each x_i and x, normalized by the window width h:

$$u_i = \frac{x_i - x}{h}$$

Next, weight each observation, depending on how close it falls to the center of the window (you can check to verify that the average weight within the window is 1, as it should be):

$$w_i = \begin{cases} 2(\cos \pi u_i)^2 & \textit{if } |u_i| < 1/2 \\ 0 & \textit{otherwise} \end{cases}$$

Finally, sum these weights and divide by the window width h.

$$\textit{local density at } x = \frac{1}{h} \sum_{i=1}^{n} w_i$$

Then plot density as a function of x. As above[1], if you want the data density in fractions per unit of measurement, divide by hn rather than h when you sum the weights (or, for that matter, if you want percent per unit of measurement, replace h with $hn/100$). Here, as above, you must choose the window width h. Remember, any trace that smooths the data will inevitably broaden the apparent distribution, and obscure sharp features. Since any smoothing is a form of distortion, you must choose an amount of smoothing that renders the distribution intelligible without distorting its relevant features.

Quantile (or percentile) plots (approximate cumulative distribution functions)

Advantages:
-display all the data, and thus portray distributions as precisely and completely as they can be known, given the available observations.
-do not require arbitrary choices of smoothing parameters

[1]The astute will notice that the square window above is equivalent to the procedure outlined here, if the weighting function is replaced by w_i =1 for $|u_i|<1/2$ and w_i=0 otherwise.

Disadvantages: -differences between distributions are often obscured

How to: The p^{th} quantile, $Q(p)$, of a distribution is the value of x such that a fraction p of the observations x_i are less than x, and a fraction 1-p of the x_i's are greater than x. When p is measured in percent rather than in fractions, quantiles are called percentiles (i.e., the 0.75 quantile is the same as the 75^{th} percentile)..

Sort the n observations x_i into ascending order, so that x_1 is the smallest and x_n is the largest. For each i, calculate the fraction of the set of observations that have x values smaller than x_i, as follows: $p_i=(i-0.5)/n$ (For percentiles, multiply p_i by 100)[2].

Plot the ordered values of the observations x_i as a function of the fractions p_i, connecting adjacent data points with straight lines. This yields a quantile plot. Conventions vary; some plot the data values on the horizontal axis and the quantiles or percentiles on the vertical axis, while others plot the quantiles/percentiles on the vertical axis and the data values on the horizontal. The latter is more commonly associated with cumulative distribution functions. The slope of the cumulative distribution function is the probability density function (that is, the rate at which numbers of observations accumulate with increasing x), so the smoothed slope of the quantile plot should approximate the density trace (see above).

For each of the i, the p_i^{th} quantile is simply the i^{th} sorted observation: $Q(p_i)=x_i$. For a fraction (or percentage) that lies between p_i and p_{i+1}, the quantile $Q(p)$ is approximated by linear interpolation:

$$Q(p)=\frac{p_{i+1}-p}{p_{i+1}-p_i}Q(p_i)+\frac{p-p_i}{p_{i+1}-p_i}Q(p_{i+1})$$

where i = integer part of $(np+0.5)$

or

$$Q(p)=x_i+(x_{i+1}-x_i)(np-i+0.5)$$

One-dimensional scatterplots

Advantages: -compact and easily constructed

Disadvantages: -features of distribution cannot easily be extracted

-points often overprinted--can be resolved by *jittering* (randomly displacing plotted points a small distance along y-axis.

Displaying distributions of categorical data

When the observations are categorical (e.g. healthy/ill/dead) rather than continuous, the available options are limited. These options are all familar enough that they need neither explanation or commentary.

-use a table to enumerate the numbers of observations in each category

-use a bar chart to show the numbers visually (since the categories are discrete, the concept of data "density" does not apply, nor does the complaint that histograms "distort" data density)

-use a pie diagram or stacked bar chart to illustrate the proportion of observations in each category. Note that it is difficult to precisely judge proportions from these plots.

References:

Chambers, J. M., W. S. Cleveland, B. Kleiner and P. A. Tukey, *Graphical Methods for Data Analysis*, 395 pp., Wadsworth & Brooks/Cole Publishing Co., 1983 (chapter 2).

Tufte, E. R., *The Visual Display of Quantitative Information*, 197 pp., Graphics Press, 1983.

[2] Various other formulas have been suggested for p_i, for a variety of arcane reasons. For a summary of these, see Cunnane, C., Unbiased plotting positions - a review, *Journal of Hydrology*, **37**, 205-222, 1978. For the record, Cunnane's formula is $p_i=(i-0.4)/(n+0.2)$. Neither Cunnane's nor the other proposed formulas differ materially from the formula used here, for most purposes. One formula that is biased and generally should *not* be used is $p_i=i/(n+1)$.

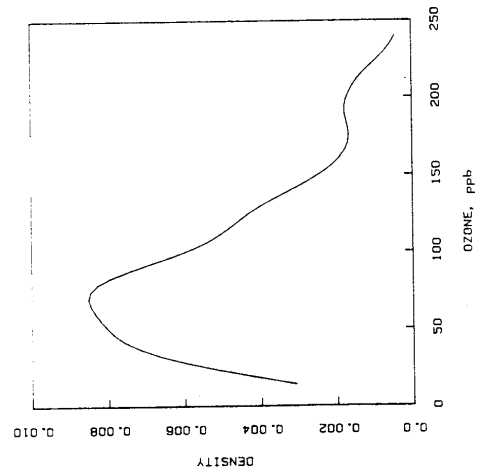


Figure 2.20 Density trace for the ozone data computed with the cosine W function and $h = 75$. A box plot of the data is added below the picture.

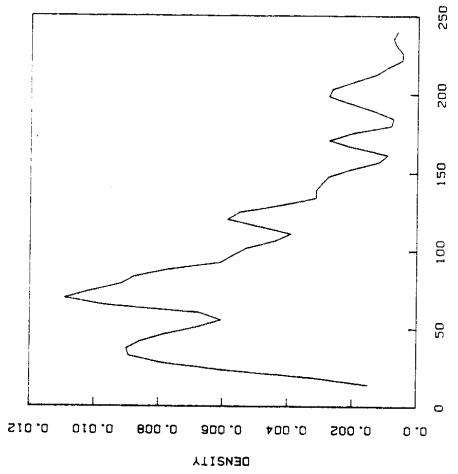


Figure 2.21 Density trace for the ozone data computed with the cosine W function and $h = 25$.

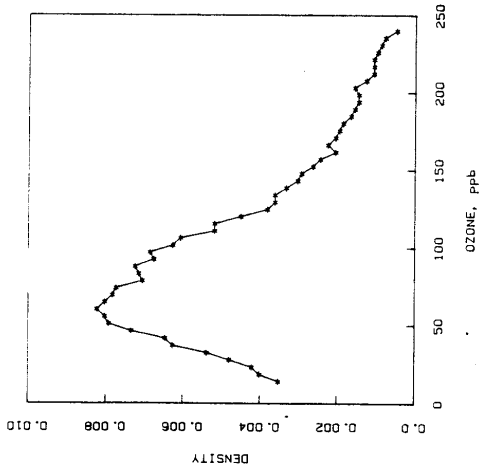


Figure 2.17 Density trace for the ozone data computed with the boxcar W function and $h = 75$.

$$W(u) = \begin{cases} 1 & \text{if } |u| \leq \tfrac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

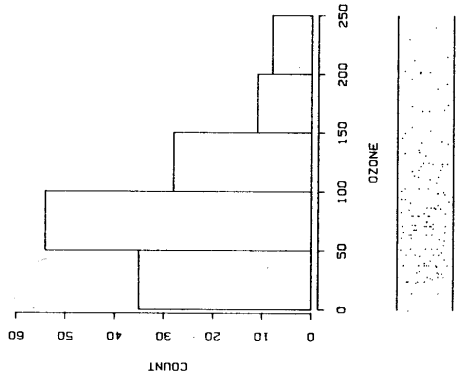


Figure 2.11 Histogram of the ozone data, with a jittered one-dimensional scatter plot.

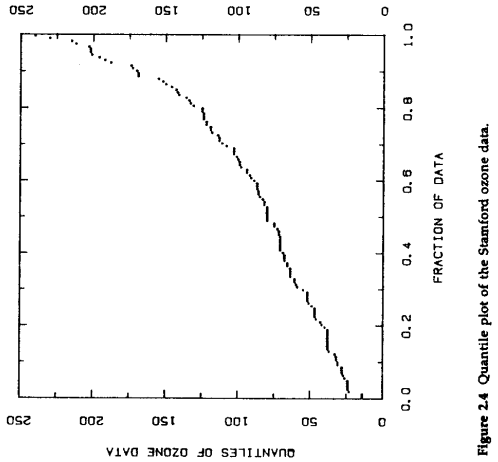


Figure 2.4 Quantile plot of the Stamford ozone data.

The figures above show exactly the same distribution of data values, plotted many different ways. Source: J.M. Chambers, W.S. Cleveland, B. Kleiner, P.A. Tukey, *Graphical Methods for Data Analysis*, Wadsworth & Brooks/Cole, 1983.