GRAPHICAL DISPLAYS FOR UNIVARIATE DATA

William G. Jacoby
ICPSR and Michigan State University

ICPSR Summer Program
July 27-29, 2009

http://polisci.msu.edu/jacoby/icpsr/graphics

I. **A Univariate Graph is a *Model* of the Data**

A. Objective is to highlight salient points without distorting features or imposing undue assumptions

B. Avoid excessive detail of individual data points, focus on overall structure within the data set

II. **Univariate Scatterplot**

A. Each observation plotted as point along axis representing range of variable values

B. Overplotting problems

C. Advantages and disadvantages of unvariate scatterplot

III. **The Histogram**

A. Most commonly-used univariate graph

B. Histogram is a two-dimensional diagram, with one axis representing the range of the variable, and the other axis representing the data density at positions within the range

C. If data are relatively continuous, then observations are grouped into mutually exclusive and exhaustive categories, called "bins"

D. Histograms provide a great deal of information about the distribution and data values

E. Disadvantages of the histogram, most arising from the binning process

IV. **Smoothed Histograms**

A. Shows local densities as a smooth, continuous function of the original data values

B. Comparison to traditional histogram and constructing a smoothed histogram

C. The kernel function

D. The "sliding window"

E. Effect of specific kernel function on shape of smoothed histogram

F. Effect of bandwidth on shape of smoothed histogram

V. **The Quantile Plot**

A. Quantiles are the order statistics for a data distribution

B. Quantile plot is a two-dimensional display— a scatterplot of data values versus their position within the ordered distribution

C. Comparison to histogram, for various distribution shapes

D. Interpretation, and advantages, of a quantile plot

VI. **Box Plots (Box-and-Whisker Diagrams)**

A. Focuses primarily on "important" quartiles

B. Definitions and components of the box plot

C. Box plot shows main features of a data distribution

D. This kind of display has many advantages, with only one real disadvantage

VII. **Dot Plots**

A. Useful display when data values are associated with identifying information, such as a label or an index number.

B. Two-dimensional plot of data labels versus data values

C. Dot plots present a great deal of information and are very easy to interpret

D. Dot plots can be used in a variety of data analysis contexts

E. The Dot plot has several advantages over its "competitors," pie charts and bar charts

F. Dot plots and arbitrary origin of data values

VIII. **Conclusions: How Does One Select the Appropriate Univariate Display?**

A. Subjective, somewhat idiosyncratic factors play a big role

B. Different kinds of displays better for different aspects of visual perception

C. Despite advantages of other methods, the histogram will probably continue to dominate univariate statistical graphics