Data types and graphics

- What graphics are appropriate for
  - Continuous (metric) variables
  - Discrete variables
  - Ordinal variables
  - Nominal variables
  - Binary variables
  - Time series variables
  - Spatial variables
  - …

Univariate continuous — features

- Features to look for

  [errors (text, impossible values, missing values …)]
  - outliers
  - symmetry
  - favoured values
  - modes
  - groupings/clusters
  - gaps
  - valleys
  - …

Univariate continuous — possible displays for data

- Dotplot/Rugplot
- Stem and leaf plot
- Histogram (self-weighted histogram)
- Barchart
- Density estimate
- Distribution function
- Boxplot
- PP and QQ plots
- Shorth plot
- Pareto plot, Lorenz curve …

Dotplot (Rugplot)

- A symbol/glyph (usually a point) is drawn for each case at its value. Can be vertical or horizontal.
- A rugplot has a line at each point.
- Effective for
  - outliers, gaps
  - small datasets with unique values
- Ineffective for
  - symmetry, favoured values, modes, groupings, valleys
  - large datasets (jittering or heaping can be tried for medium-sized datasets)

Stem and leaf plot

- Cases are sorted by value and split into stem and leaf. Stems are listed in the left column and leaves in the right.
- Of historical interest because of Tukey's use in "EDA"
- Of value for timetable displays, e.g. in Japan

Histogram (1)

- Divide range of values into equally sized bins (unequally sized bins are confusing to interpret and offer only a poor man's density estimate) and plot bars with a height equal to bin frequency.
- Relative frequency scaling is also possible.
- Binwidth and anchor point are key parameters. No one histogram can reveal all possible features. Drawing several is often informative.
- The bins are half-open intervals. Whether they are open to the left or right can make a difference to the display.

Histogram (2)

- Effective for
  - symmetry, modes, groups/clusters, valleys
  - large datasets
- Can be useful for
  - favoured values, gaps (if parameters are right)
- Poor for
  - outliers
  - low frequency bins in large datasets
- Selfweighted histograms use sums of case values instead of numbers of cases for bar heights. They are useful for importance measures.

Barchart

- Plot the data as a barchart, as if the values were nominal
- Useful for checking variables with possible errors
- Useful for identifying favoured values
- The x axis no longer has metric meaning (which can be confusing).

Density estimates (1)

- Plot a nonparametric estimate of the density for the population from which the data have come using
  - kernel methods
  - logspline approach
  - clustering methods
  - average shifted histograms

  OR
- Plot a parametric estimate assuming a density form and estimating the parameters from the data

Density estimates (2)

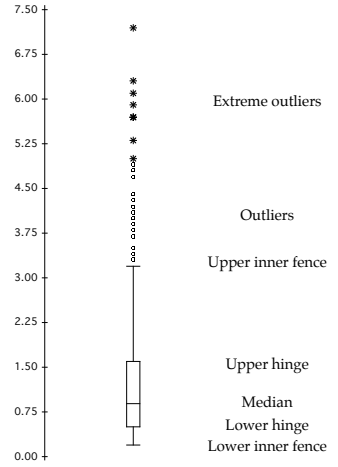
- Depend heavily on bandwidth/window used
- Estimation problems with
  - boundaries (e.g., no values ≤ 0)
- Useful for
  - symmetry, modes, , groupings/clusterings, valleys
  - getting a picture of the likely population distribution
- Not so useful for
  - outliers, gaps, favoured values
  - picturing the raw data

Distribution function

- Empirical distribution functions are step functions.
- Since all distribution functions are monotone non-decreasing, they look fairly similar.
- Useful for showing if there is stochastic dominance when comparing two distributions.

Boxplot

- Every mark in a boxplot represents an actual value
- Useful for
  - outliers, symmetry
  - summarising a distribution in little space
  - comparing distributions
- Not useful for
  - modes, gaps, valleys, grouping/clusters, favoured values



PP Plot

- In a PP plot two distribution functions are plotted against one another.
- For a single sample, the empirical distribution function of the data may be plotted against a theoretical distribution.
- PP plots are for comparing distributions:
  - location
  - scale
  - shape
  - tails
- QQ plots are generally preferred for this purpose.

QQ Plot

- In a QQ plot the quantiles of two distributions are plotted against one another.
- Comparing an empirical distribution with a theoretical one leads to an issue of which theoretical quantiles should be used. For a sample of size n, the quantiles {k/(n+1); k=1,…n} are a reasonable choice. Many other choices are possible.
- QQ plots are good for comparing distributions and especially for assessing goodness of fit to theoretical distributions.

Shorth plot

- The shorth is the shortest interval containing half the distribution.
- α-shorth is the shortest interval containing a proportion α of the distribution.
- The shorth plot draws the α-shorth lengths for various α's as a function of x (i.e. x must be in the interval) and uses a reversed y axis
- Possibly good for detecting modes

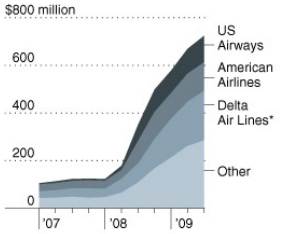
The New York Times

April 7, 2010

Making Money On Bags . . .

Airlines' revenue from fees for checking luggage has soared.

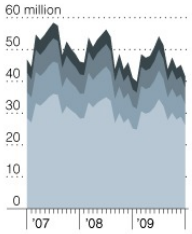
QUARTERLY REVENUE FROM BAGGAGE FEES



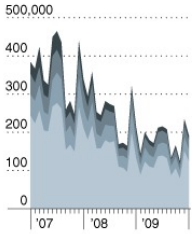
. . . And Losing Fewer of Them

Passenger traffic has fallen during the economic downturn, but not nearly as much as reports of mishandled baggage.

NUMBER OF PASSENGERS, MONTHLY



REPORTS OF MISHANDLED BAGGAGE, MONTHLY



Source: Transportation Department

*Includes Northwest Airlines, which recently merged with Delta.

Figure 2 – Percentage of live births by mother's age and type of registration, England and Wales, 2008