# Visualizing numeric variables – histograms
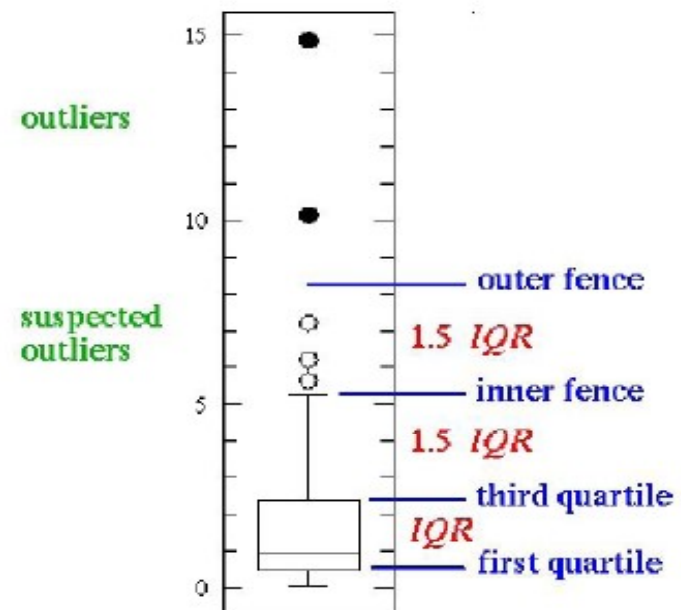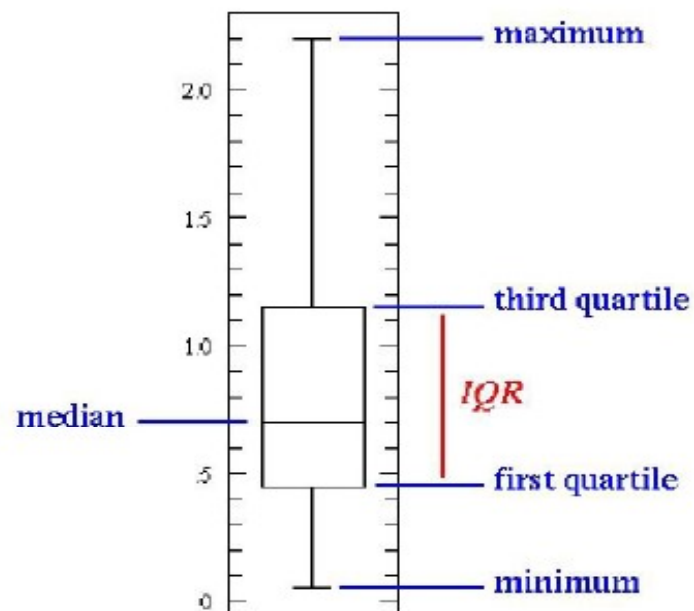
- A **histogram** is another way to graphically depict the spread of a numeric variable.

- hist(attribute-name, main=" ")

- The histogram is composed of a series of bars with heights indicating the count, or f**requency**

Sample code is in **histogram-example.r**

# Box plot



- The first **quartile** ($Q_1$) is defined as the middle number between the smallest number and the median of the data set
- The second **quartile** ($Q_2$) is the median of the data
- The third **quartile** ($Q_3$) is the middle value between the median and the highest value of the data set

- The horizontal lines forming the box in the middle of each figure represent Q1, Q2 (the median), and Q3 when reading the plot from bottom-to-top.

- The median is denoted by the dark line,

# Visualizing numeric variables – boxplots

- Visualizing numeric variables can be helpful for diagnosing many problems with data.

- A common visualization of the five-number summary is a **boxplot**.

- The boxplot displays the center and spread of a numeric variable in a format that allows you to quickly obtain a sense of the range and skew of a variable, or compare it to other variables.

- boxplot(attributename,main="name ")

Sample code is in **boxplot-example.r**

- A histogram is **symmetric** in shape if the right and left sides have essentially the same shape.

- When the right side of the histogram, containing the larger half of the observations in the data, extends a greater distance than the left side, the histogram is referred to as **skewed to the right.**

- The histogram is **skewed to the left** when its left side extends a much larger distance than the right side.

- The measures of central tendency related for a given set of measurements depends on the **skewness** of the data.
- If the distribution is mound-shaped and symmetrical about a single peak, the mode ($Mo$), median ($Md$), mean (m), and Trimmed mean($TM$) will all be the same.
- This is shown using a smooth curve and population quantities.
- If the distribution is skewed, having a long tail in one direction and a single peak, the mean is pulled in the direction of the tail; the median falls between the mode and the mean; and depending on the degree of trimming,
- The trimmed mean usually falls between the median and the mean.
- The following figures illustrate this for distributions skewed to the left and to the right.
- If mean value is greater than median this implies that the distribution of the attribute is right skewed.

$M_o$
$M_d$
$\mu$
TM
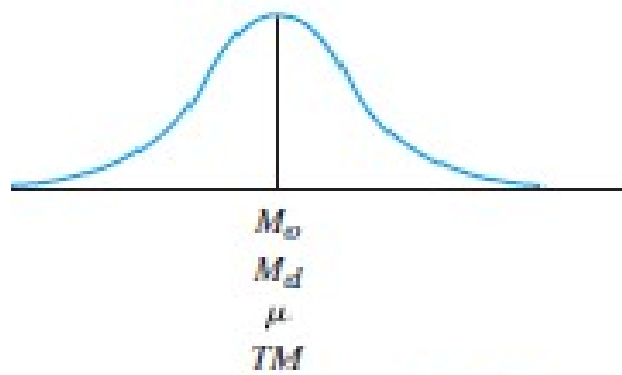
(a) A mound-shaped distribution

$\mu$  $M_d M_o$
  TM
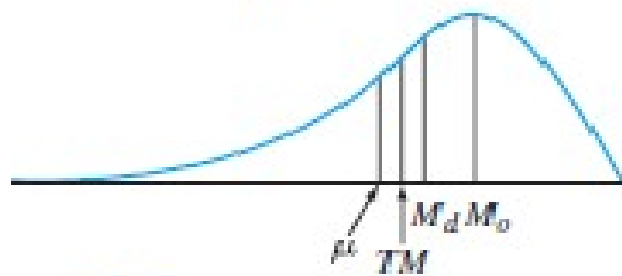
(b) A distribution skewed to the left

$M_o$ $M_d$
  TM  $\mu$
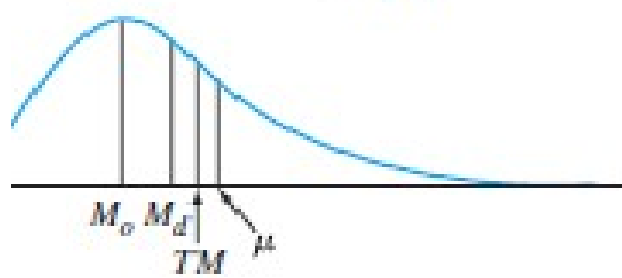
(c) A distribution skewed to the right

# Visualizing Qualitative variables – barplot

Sample code is in **barplot-example.r**

# Visualizing relationships – scatterplots

- A **scatterplot** is a diagram that visualizes a bivariate relationship.

- It is a two-dimensional figure in which dots are drawn on a coordinate plane using the values of one feature to provide the horizontal *x* coordinates, and the values of another feature to provide the vertical *y* coordinates.

- Patterns in the placement of dots reveal underlying associations between the two features.

- We'll use the plot() function for this.

Sample code is in **linechart-example.r**