Statistical Descriptions of data

- Statistical descriptions can be used to identify properties of the data and highlight which data values should be treated as noise or outliers.
- There are three areas of basic statistical descriptions
    1. Measures of central tendency
    2. Dispersion of the data
    3. Graphic displays of basic statistical descriptions

Measuring the Central Tendency:

1. Mean:

The most common and effective numeric measure of the "center" of a set of data is the (arithmetic) mean. Let x1,x2, : : : ,xN be a set of N values or observations, such as for some numeric attribute X, like salary.

The mean of this set of values is

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}.$$

2. Median:

Gives Middle value if odd number of values, or average of the middle two values otherwise.

3. Mode:

Value that occurs most frequently in the data

4. Midrange:

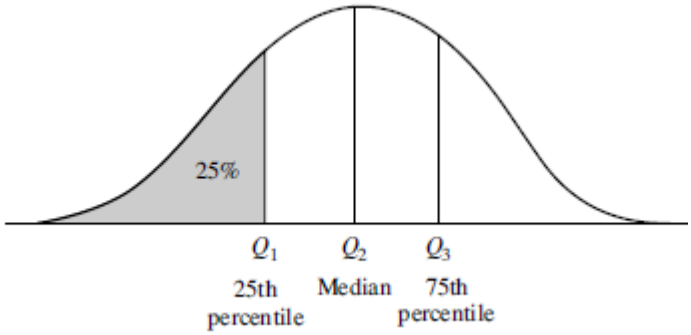
The midrange can also be used to assess the central tendency of a numeric data set. It is the average of the largest and smallest values in the set. This measure is easy to compute using the SQL aggregate functions, max() and min().

Measuring the Dispersion of Data

Range, Quartiles, and Interquartile Range

- Quantiles are points taken at regular intervals of a data distribution, dividing it into essentially equalsize consecutive sets.

- The 2-quantile is the data point dividing the lower and upper halves of the data distribution. It corresponds to the median. The 4-quantiles are the three data points that split the data distribution into four equal parts; each part represents one-fourth of the data distribution. They are more commonly referred to as quartiles.
- The 100-quantiles are more commonly referred to as percentiles; they divide the data distribution into 100 equal-sized consecutive sets. The median, quartiles, and percentiles are the most widely used forms of quantiles.



- The distance between the first and third quartiles is a simple measure of spread that gives the range covered by the middle half of the data. This distance is called the interquartile range (IQR) and is defined as $IQR = Q_3 - Q_1$.

Five-Number Summary, Boxplots

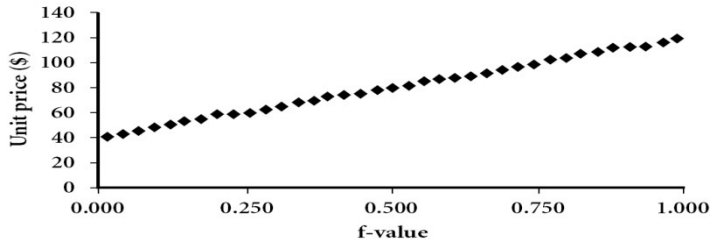
- The five-number summary of a distribution consists of the median (Q2), the quartiles Q1 and Q3, and the smallest and largest individual observations, written in the order of Minimum, Q1, Median, Q3, Maximum.
- Boxplots are a popular way of visualizing a distribution. A boxplot incorporates the five-number summary as follows:
  - ✓ Typically, the ends of the box are at the quartiles so that the box length is the interquartile range.
  - ✓ The median is marked by a line within the box.
  - ✓ Two lines (called whiskers) outside the box extend to the smallest (Minimum) and largest (Maximum) observations.

Graphic Displays of Basic Statistical Descriptions of Data

This include quantile plots, quantile–quantile plots, histograms, and scatter plots. Such graphs are helpful for the visual inspection of data, which is useful for data preprocessing. The first three of these show univariate distributions (i.e., data for one attribute), while scatter plots show bivariate distributions (i.e., involving two attributes).

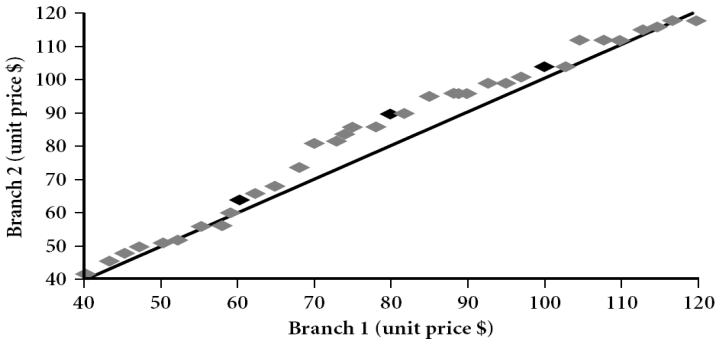
Quantile Plot

- Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)
- Plots quantile information
    - For a data x_i data sorted in increasing order, f_i indicates that approximately 100 f_i% of the data are below or equal to the value x_i



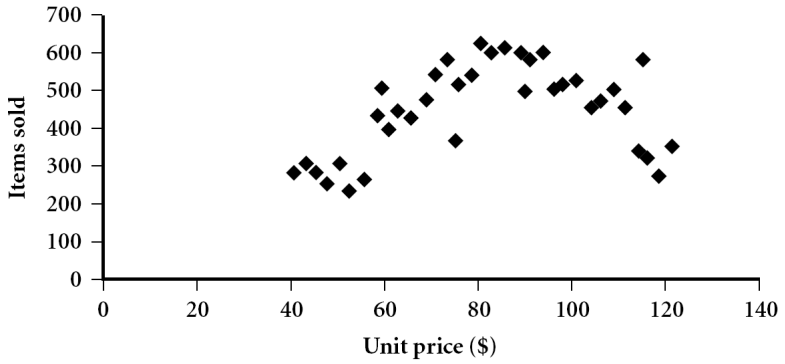
Quantile-Quantile (Q-Q) Plot

- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another.
- View: Is there is a shift in going from one distribution to another?
- Example shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile. Unit prices of items sold at Branch 1 tend to be lower than those at Branch 2.



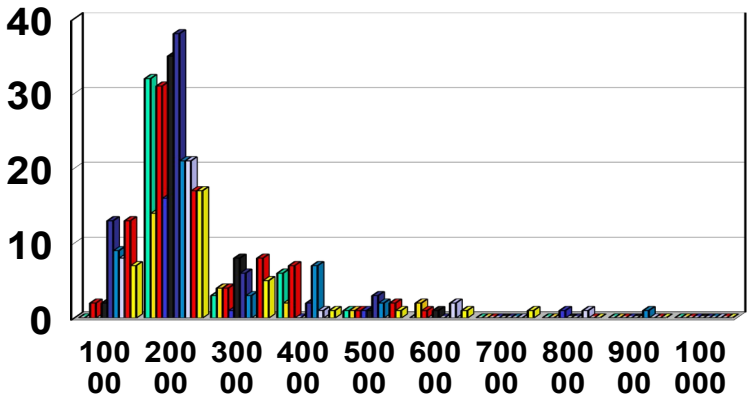
Scatter plot

- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane

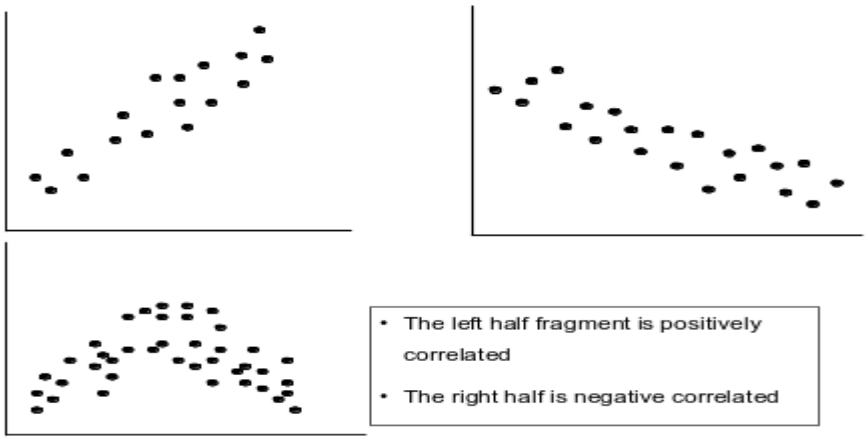


Histogram Analysis

- Histogram: Graph display of tabulated frequencies, shown as bars.
- It shows what proportion of cases fall into each of several categories.
- Differs from a bar chart in that it is the *area* of the bar that denotes the value, not the height as in bar charts, a crucial distinction when the categories are not of uniform width.
- The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent.



Positively and Negatively Correlated Data



- The left half fragment is positively correlated
- The right half is negative correlated

Uncorrelated Data