# BUAN/OPRE 6359: Advanced Statistics for Data Science

## Lecture 1: Descriptive Statistics

In our daily lives, data is generated and collected around us. For example:

- a device we wear that is counting steps,
- your phone apps collecting details of your usage,
- the geographic location app identifying when you arrive to a store,
- the internet search engine matching your buying preferences with different offers

**Data analysis:**

- Being able to discriminate from all the raw data available to us, which data is valuable, and which is not.
- Turn raw data into knowledge.

**Basic definitions:**

- A variable is any characteristics, number, or quantity that can be measured or counted.
  - e.g., the final grade for this course
- The values of the variable are the range of possible values/possible attributes that the variable can assume.
  - e.g., the final grade for this course is a real number between 0 and 100.
- Data are the observed values/characteristics of a variable.
  - e.g., the final grades in this class: {92, 74, 71, 83, 93, 55, 62,....}

**Data** can be classified as either categorical or quantitative.

- Categorical data can be grouped by specific categories.
- Quantitative data use numeric values to indicate how much or how many about a specific characteristic (variable). Quantitative variables can be further classified as discrete or continuous.
  - Discrete variable takes on a countable number of values. For example, the number of students coming to class on time.
  - Continuous variable takes on an uncountable number of numerical values. For example, the height of a person.

**Example:** Determine whether the following data are quantitative or categorical.

(a) The number of Toyotas imported monthly by the United States

(b) Gender

(c) Distance between two cities

(d) Height

(e) The size of a soft drink (small, medium, or large) ordered in a restaurant

(f) Age

(g) Marital status (single, married, widowed, divorce)

(h) Marital status (0, 1, 2, 3)

(i) Your final (numerical) grade for this course

(j) Your final (letter) grade for this course

**Example[1]:** The Cherry Blossom 10 Mile Run is a road race held in Washington, D.C. in April each year. (The name comes from the famous cherry trees that are in bloom in April in Washington.) The results of this race are published. This data frame contains the results from the 2005 race. Data are collected on the following variables.

- state: state of residence for the runner.
- time: official time from starting gun to finish line.
- net: the recorded time (in seconds) from when the runner crossed the starting line to when the runner crossed the finish line. This is generally less than the official time because of the large number of runners in the race: it takes time to reach the starting line after the gun has gone off.
- age: age of runner in years.
- sex: a factor with levels F and M.

[1] https://cran.r-project.org/web/packages/mosaicData/mosaicData.pdf

**Graphical/tabular summary:**

- Frequency distribution: a tabular summary of categorical data showing the count of observations in each of several non-overlapping categories or classes

- Relative frequency distribution: a tabular summary of categorical data showing the proportion (relative count) of observations in each of several non-overlapping categories or classes

  - Example (continued): Create the frequency and the relative frequency tables for the data collected on categorized age ([10,20), [20,30),...) using R.

- Cross tablulation: a table showing the frequency of each combination of the two variables' values

  - Example (continued): Create a cross tabulation for two variables: sex and categorized age using R.

- Bar/Pie chart: a graphical summary of categorical data showing the frequency/relative frequency of several non-overlapping categories or classes

  - Example (continued): Create a bar chart and a pie chart of the data collected on categorized age using R. Create a bar chart of categorized age for the two sexes in one chart.

- Histogram: a graphical summary of quantitative data showing the count of observations in each of a series of intervals

  - Example (continued): Create a histogram of the data collected on net using R.

- Scatter chart: a graphical summary of two quantitative variables

  - Example (continued): Create a scatter plot of the data collected on age and net using R.

- Line chart (time series plot): a graphical summary of time-series data (data that are collected over successive points in time)

  - Example[2]: Create a line chart of the monthly average retail price of gasoline since 1976 in GasPrice.rds data file.

[2] Statistics for Management and Economics ($11^{\text{th}}$) by Gerald Keller

**Numerical measure:** a single number that describes data

- Measure of central location
- Measure of relative standing
- Measure of variability
- Measure of association (linear relationship)

**Measures of central location:**

- Arithmetic Mean (average)

$$\text{sample} : \bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}, \quad \text{population} : \mu = \frac{x_1 + x_2 + ... + x_N}{N}$$

  - sensitive to outliers (extreme values)
- Median: the middle value in the sorted data
  - if there is an even number of values, then median is the mean of the two middle values
- Mode: the value with the highest frequency in the data

**Example (continued):** Calculate the mean, median and mode of the data collected on age using R.

**Measures of relative standing:** provide information about the position of a value relative to the entire dataset.

- Percentile: The $p^{th}$ percentile is the value for which $p$ percent of the values in the dataset are less than and $(100 - p)\%$ are greater than that value.
- Quartile: special names for the $25^{th}$, $50^{th}$, and $75^{th}$ percentiles
  - The first quartile $(Q_1) = 25^{th}$
  - The second quartile $(Q_2) = 50^{th}$
  - The third quartile $(Q_3) = 75^{th}$

**Example (continued):** Calculate the $40^{th}$ percentile and the quartiles for the data collected on age using R.

**Measures of variability:**

- Range: measures the spread of all the data values

$$\text{range} = \text{largest observation} - \text{smallest observation}$$

- Interquartile Range (IQR): measures the spread of the middle 50% of the observations

$$IQR = Q_3 - Q_1$$

- Variance: average squared deviation of data values from the mean

$$\text{sample} : s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}, \quad \text{population} : \sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

- Standard deviation: square root of the variance

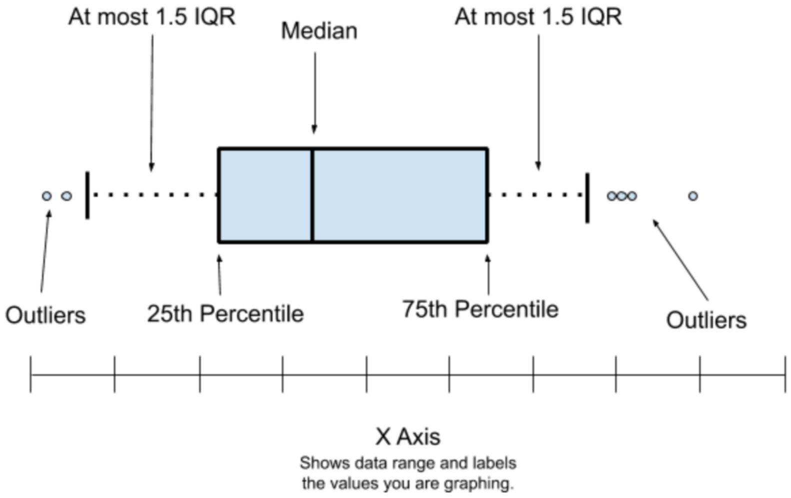
$$\text{sample} : s = \sqrt{s^2}, \quad \text{population} : \sigma = \sqrt{\sigma^2}$$

- Coefficient of variation: relative size of the standard deviation

$$\text{sample} : CV = \frac{s}{\bar{x}}, \quad \text{population} : CV = \frac{\sigma}{\mu}$$

**Example (continued):** Calculate the measures of variability for the data collected on age using R.

**Box plot:** a graphical summary of the distribution of data



Source[3]

[3] https://publiclab.org/notes/mimiss/06-18-2019/creating-a-boxplot-to-identify-outliers-using-codap

**Example (continued):** Create a box plot of the data collected on net for the two sexes using R.

**Measure of association:** provides information as to the strength and direction of a linear relationship between two variables (if one exists).

- Covariance

$$\text{sample} : s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}, \quad \text{population} : \sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}$$

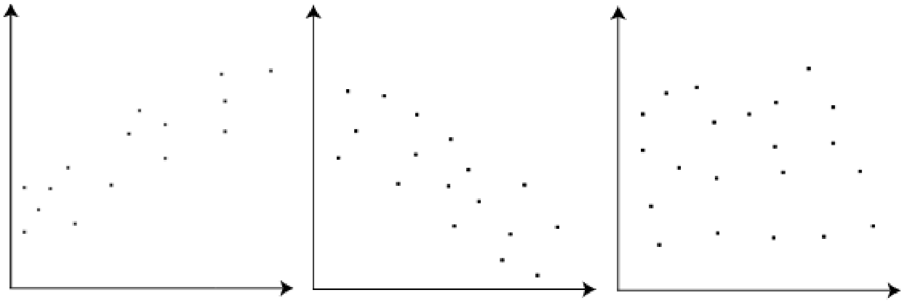
  - If $x$ and $y$ move in the same direction, $s_{xy}$ is large and positive.
  - If $x$ and $y$ move in opposite directions, $s_{xy}$ is large and negative.
  - If $x$ and $y$ move in no linear pattern, the $s_{xy}$ is a small number.

- Correlation coefficient

$$\text{sample} : r = \frac{s_{xy}}{s_x s_y}, \quad \text{population} : \rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

$-1 \leq r, \rho \leq 1$

  - Correlation coefficient close to $-1 \Rightarrow$ strong negative linear relationship
  - Correlation coefficient close to $+1 \Rightarrow$ strong positive linear relationship
  - Correlation coefficient close to $0 \Rightarrow$ weak linear relationship



**Example (continued):** Is there a linear relationship between the variables age and net?