Process Modeling by Statistical Methods (0905331)
01- Introduction

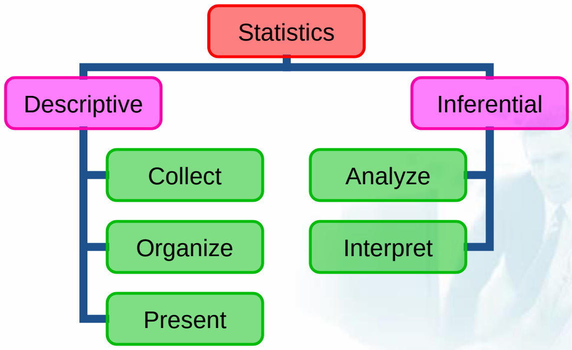
Dr. Ali Khalaf Al-matar
Chemical Engineering Department
University of Jordan
aalmatar@yahoo.com

Outline

- Definition and applications of statistics
- Types of statistical data
- Population and sample
- Probability and its relation to statistics
- Sampling methods
- Basic definitions
  - Measures of central tendency
  - Measures of dispersion
  - Measures of skewness
- Histogram plots

What is Statistics?

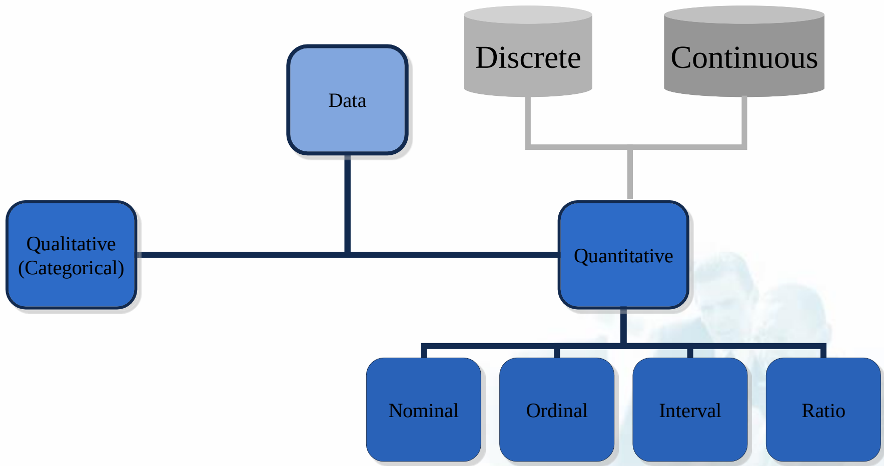
Statistics is the science dealing with data to assist in making more effective decisions in the face of **uncertainty.**



Why Study Statistics?

- Numerical information is everywhere!
- Statistical methods are used to make decisions that affect our lives
- To understand why decisions are made and how such decisions affect you.

Classification of Statistical Data



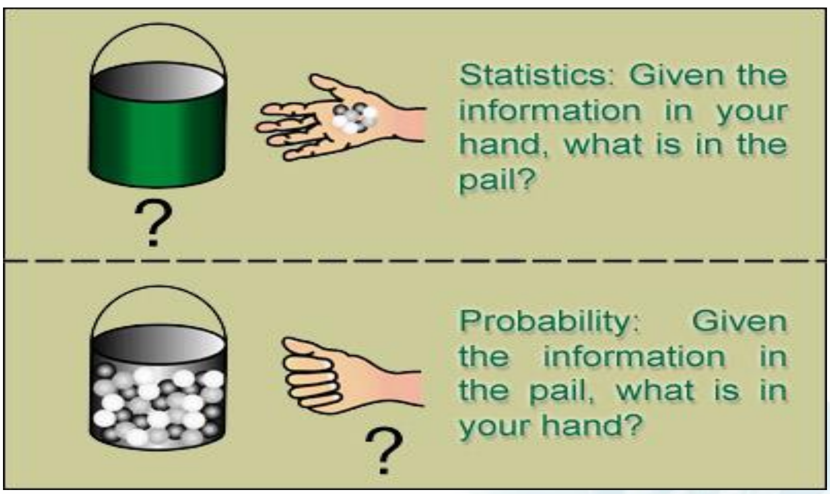
Population & Sample

- **Population** is the collection consisting of all possible outcomes of an experiment, measurement, or observation.
- **Sample** is a subset or a part of the population.

Probability

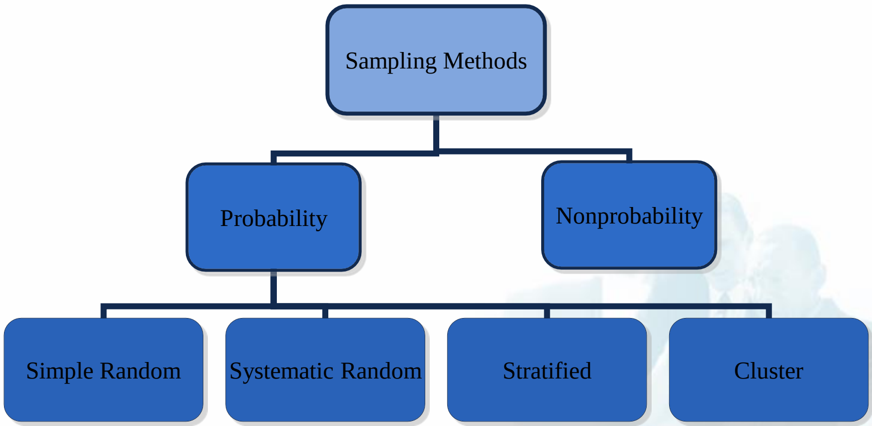
- Probability refers to the study of **randomness** and **uncertainty**.
- Any process or action that generates a **unique** observation out from the set of possible observations is called an **experiment.**
- The use of the term "experiment" does not necessarily mean that it has to be performed using specialized laboratory equipment. A computer experiment, or a coin toss is referred to as an experiment.

Statistics & Probability

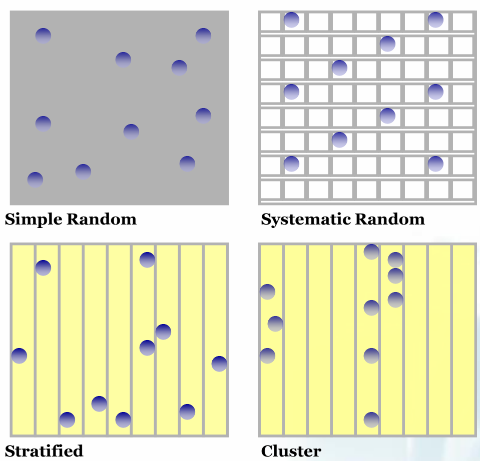


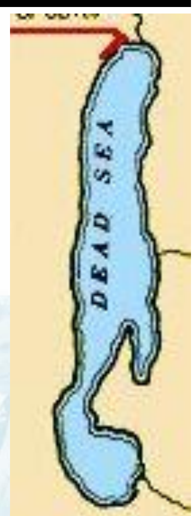
From MIT OCW 15.075 Spring 2003

Sampling Methods



Sampling Methods





Measures of Central Tendency: Mean

Sample mean (Arithmetic) $\bar{X} = \dfrac{\sum_{i=1}^{n} X_i}{n}$

Geometric mean $\bar{X}_g = \left[\prod_{i=1}^{n} X_i\right]^{1/n}$

Harmonic mean $\bar{X}_h = \dfrac{n}{\sum_{i=1}^{n} \dfrac{1}{X_i}}$

Population mean $\mu = \dfrac{\sum_{i=1}^{N} X_i}{N}$

- Other types of means:
  - Log mean temperature in heat transfer (log based averaging).
  - Sauter mean drop size in liquid-liquid dispersions (Volume to surface area ratio)

Central Tendency: Median

- The mean is **very sensitive to outliers or extreme values in the data**.
- Median is defined as a rigorous estimator of central tendency.
  - Sort the data in ascending order
  - The median is defined such that 50% of the values are above, and 50% are below it.

$$\tilde{X} = \begin{cases} x_{(n+1)/2} & ,n \text{ odd} \\ \dfrac{x_{n/2} + x_{n/2+1}}{2}) & ,n \text{ even} \end{cases}$$

Central Tendency: Mode

- Mode: the value of the most frequently encountered observation.
  - Mathematically, it is the most frequent value i.e. the **maximum** of the distribution.
- All the analysis above refer to sample values. To obtain population values just replace the number of observations with the proper population values.

Measures of Dispersion: Range

- Averages provide information about **central tendency** but it does not provide any information about the **spread** of the data
- Range is the difference between the maximum and minimum values of the random variable we are interested in

$$\text{Range} = x_{\text{Max}} - x_{\text{Min}}$$

Dispersion: Variance and Standard Deviation

Sample variance	$s^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}$	Population variance	$\sigma^2 = \frac{\sum_{i=1}^{N}\left(X_i - \mu\right)^2}{N}$
Sample standard deviation	$s = +\sqrt{s^2}$	Population standard deviation	$\sigma = +\sqrt{\sigma^2}$

Measures of Dispersion: Coefficient of variation

- The **Coefficient of Variation** (*CV*) is the ratio of standard deviation to the arithmetic mean. Expressed in percent

$$CV = 100\frac{S}{\overline{X}}\%$$

- A good measure for comparing different values of means and standard deviations (**relative**).

Measures of Symmetry: Skewness

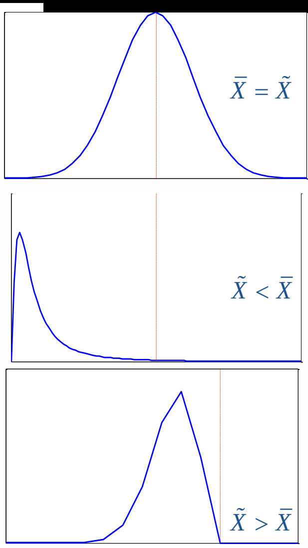
- Measures of dispersion give some values about the spread of a distribution.
- They don't provide any info about the shape of the distribution around the mean or median.
- Coefficient of skewness (*Sk*) is defined to alleviate such lack of info

$$Sk = \frac{3(\text{mean} - \text{median})}{\text{standard deviation}} = \frac{3(\bar{X} - \tilde{X})}{s}$$

Symmetric

Positive (right skew)

Negative (left skew)



Trends and measures of central tendency

Symmetric →
mean = median.
If one mode exists (unimodal) →
mean = median = mode

Skewed to the right
mode < median < mean

Skewed to the left
mode > median > mean

Frequency Distribution

- Generate a **frequency distribution**
  - Divide the range of data into intervals (**class intervals**, **cells** or **bins**).
  - Choose a number of bins. Use the **square root** of the **number of observations** (n).
  - Find the frequency of observations in each bin
- Relative frequency (**normalized**): the observed frequency in each bin divided by the total number of observations.
- Cumulative frequency: the height of each bar is the total number of observations that are less than or equal to the upper limit of the bin.

Bin Selection

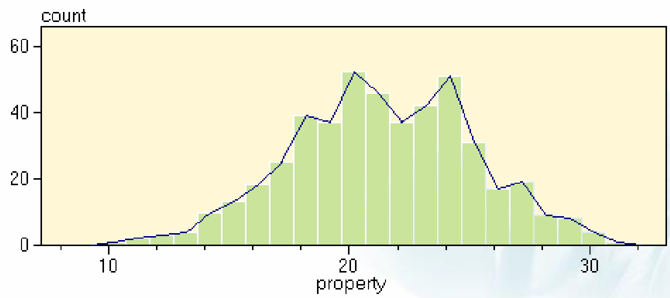
- Use the **square root** of the **number of observations** (n), or
- Freedman-Diaconis rule

$$\text{Bin size} = 2 \cdot \text{IQR}(x) \cdot n^{-1/3}$$

  - x is the data
  - IQR is the interquartile range of the data
  - n is the number of observations in the sample.

Histogram Plots (Frequency Distribution)

- Histograms are plots of frequency versus the property of interest.
  - Good for display of the shape (trend) of the data for relatively large samples ($n \geq 100$).



Frequency polygons

- Basically the same as histograms where the rules valid for histograms are also valid for frequency polygons.
- Smoother alternative to histograms.
- Can be constructed from histograms by joining the midpoints of the histogram bars with lines.
- The areas below the histogram and the frequency polygon are equal.
- In general, one should prefer to use histograms rather than frequency polygons, as the width of the classes cannot be seen well in frequency polygons.

Box Plots

- Graphical display that simultaneously describes several important features of a data set
  - Center
  - Spread
  - Departure from symmetry
  - Identification of outliers.