STA255: Statistical Theory

Chapter 1: What Is Statistics?

Summer 2017

What is statistics?

- Statistics may be defined as the art of science that deals with collecting, analyzing, presenting, and interpreting data in order to help managers make better decisions.
- Data (plural of 'datum') are the results of measurements, can be the basis of graphs, images, or observations of a set of variables.
- Theory of Statistics provides a basis for the whole range of techniques that are used within applications of Statistics.

What is statistics?

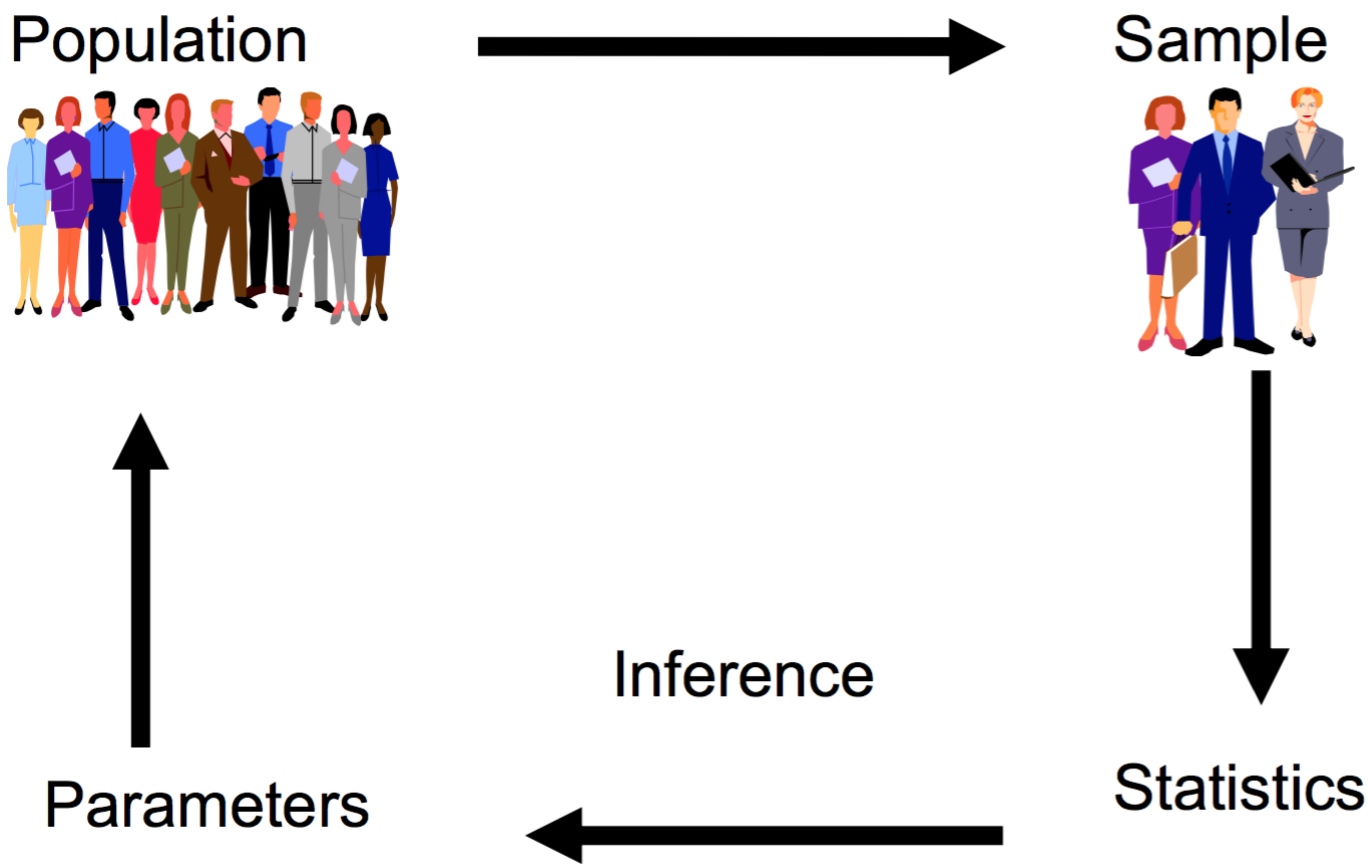
There are two main areas of Statistics:

- Descriptive Statistics:
  - Collecting, organizing, summarizing, presenting data.
- Inferential statistics:
  - Drawing conclusions and/or making decisions concerning a population based on sample data.

Basic Concepts

- A population is the entire collection of objects or outcomes about which information is sought.
- A sample is a subset of a population, containing the objects or outcomes that are actually observed.
- Example: UofT registrar office sample 100 undergraduate students from entire undergraduate students at UofT and ask how much tuition fee they pay for next academic year.
- The population of interest: The entire undergrad. students at UofT. The sample: 100 undergrad students selected.

Statistics and Parameters



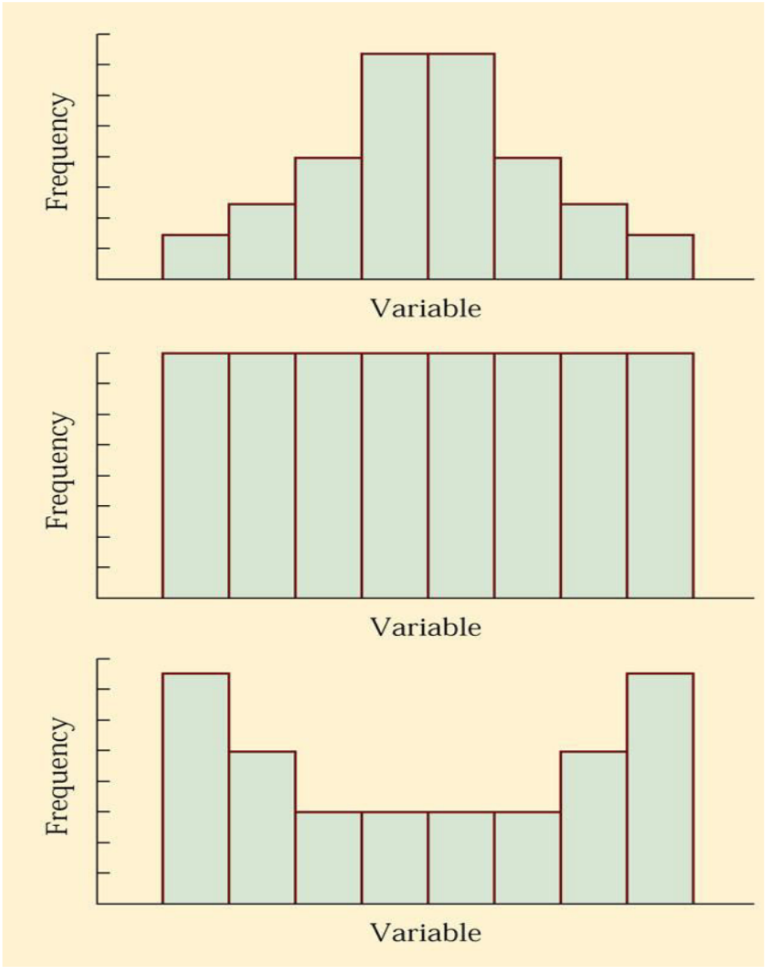
Histogram

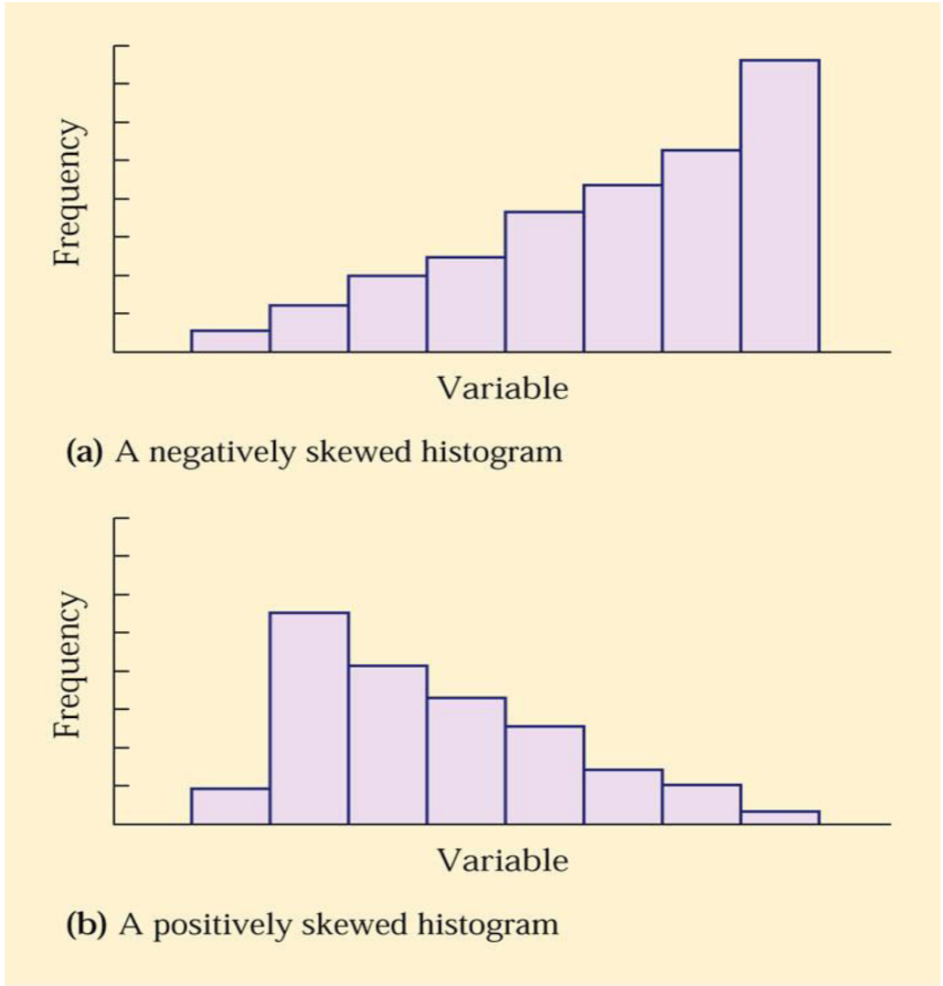
- Graphical display that gives an idea of the shape of the data distribution.
- What to Look For: Central or typical value, extent of spread or variation, general shape, location and number of peaks, presence of gaps and outliers.

Shapes of Histograms

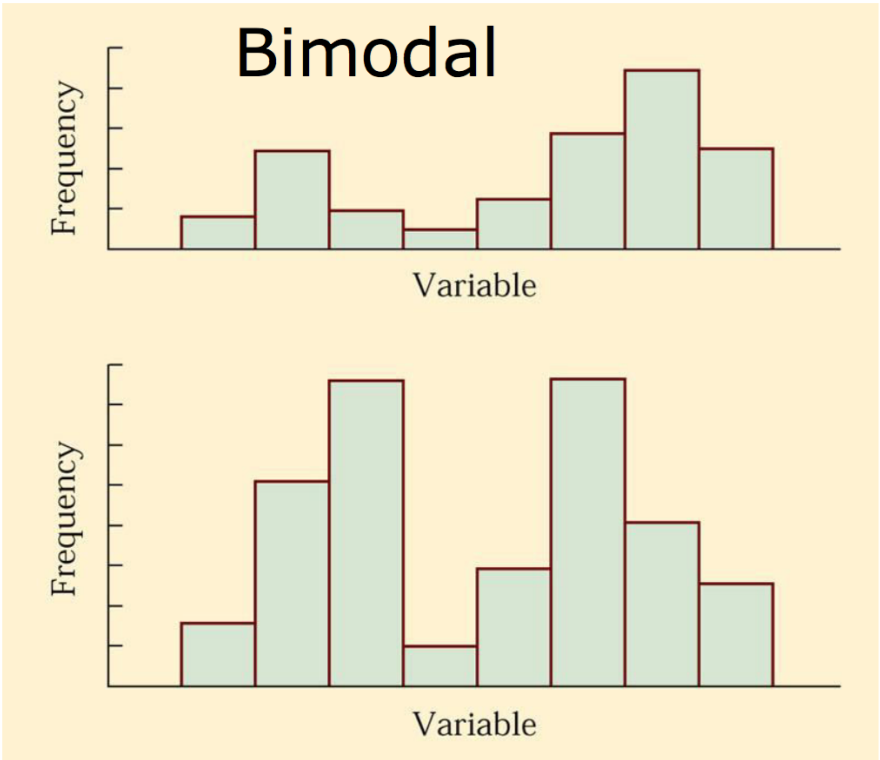
- A histogram is symmetric if its right half is a mirror image of its left half.
- Histograms that are not symmetric are referred to as skewed.
- A histogram with a long right-hand tail is said to be skewed to the right, or positively skewed.
- A histogram with a long left-hand tail is said to be skewed to the left, or negatively skewed.
- A histogram is unimodal if it has only one peak (or mode), and bimodal if it has two clearly distinct modes. In principle, a histogram can have more than two modes.

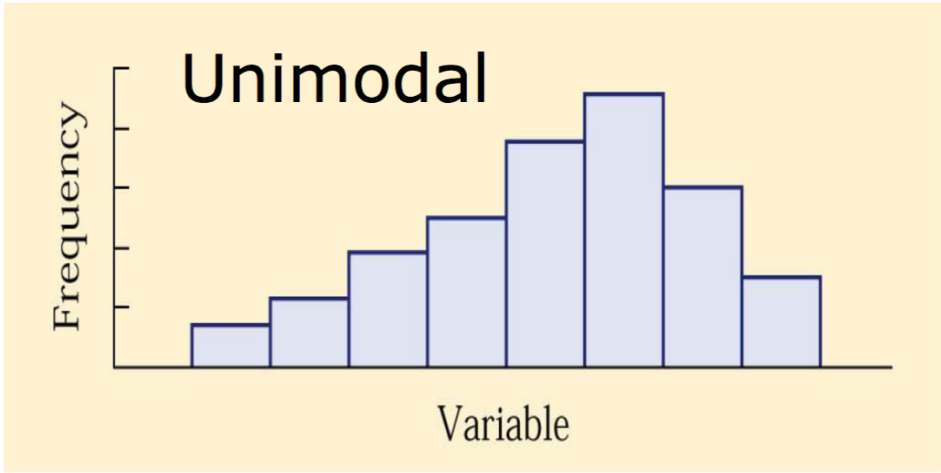
Shapes of Histograms

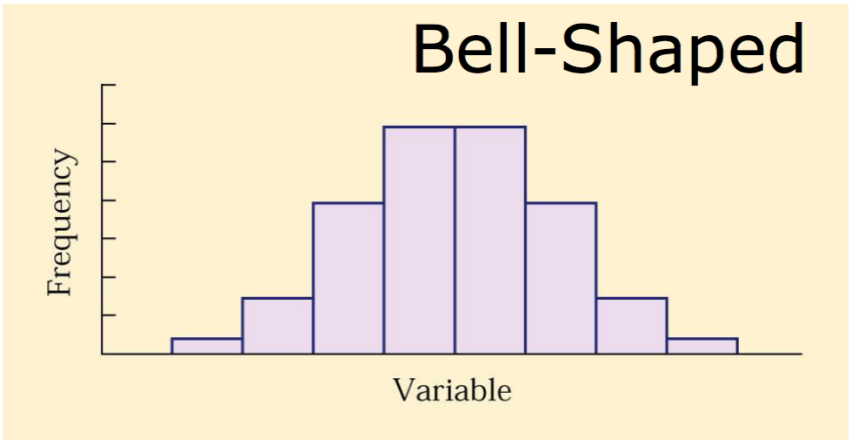




Shapes of Histograms







Measures of Centre: Mean

- The Mean is the average of data values. Let $y_1, \ldots, y_n$ be a sample of n measured responses,
- Sample mean:

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

- Population mean:

$$\mu = \frac{\sum_{i=1}^{N} y_i}{N}$$

where$n$ is the sample size and N is the population size.

- Sometimes a sample may contain a few points that are much larger or smaller than the rest. Such points are called outliers and may affect the mean.

Measures of Spread (dispersion): Variance

- Average of squared deviations of values from the mean.
- Population variance:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(y_i - \mu)^2}{N}$$

- Sample variance:

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

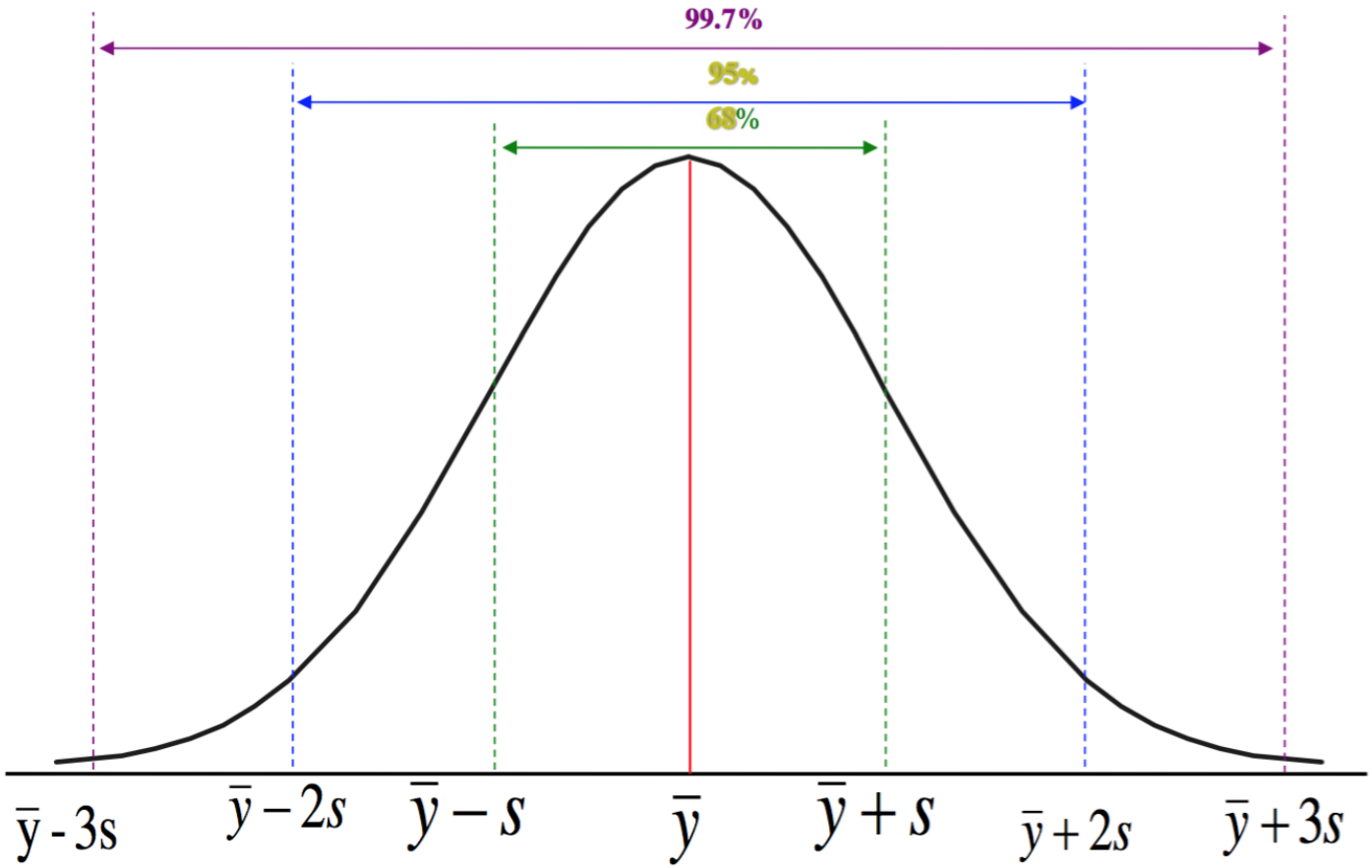
- The sample variance (s^2) is a reasonable estimate of the population variance (σ^2)

Measures of Spread: Standard Deviation

- Most commonly used measure of variation.
- The square root of the variance.
- Shows variation about the mean.
- Has the same units as the original data.
- Sample standard deviation:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

The Empirical Rule



The Empirical Rule

- For a distribution of measurements that is approximately normal (bell shaped), it follows that the interval with end points
  - $\mu \pm \sigma$ contains approximately 68% of the measurements.
  - $\mu \pm 2\sigma$ contains approximately 95% of the measurements.
  - $\mu \pm 3\sigma$ contains almost all of the measurements.