FoxMIS CONNECT | INNOVATE

Fox School of Business
TEMPLE UNIVERSITY®

MIS2502:
Data Analytics
Descriptive Statistics and Hypothesis Testing

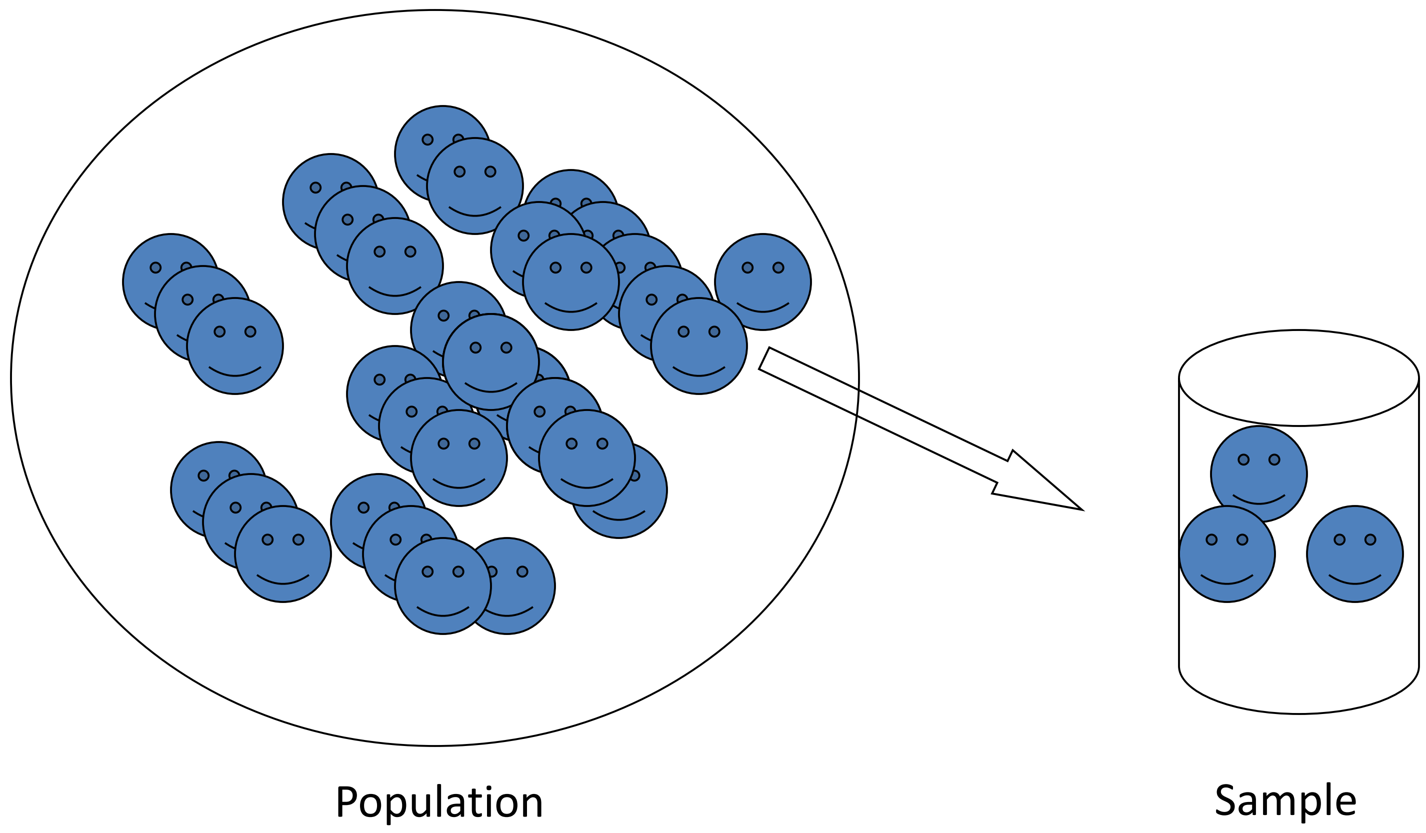
Alvin Zuyin Zheng
zheng@temple.edu
http://community.mis.temple.edu/zuyinzheng/

Descriptive Statistics

Descriptive Statistics:

- Tools for summarizing, organizing, and simplifying data
- What data tells us about the population
  - Measures of Central Tendency
  - Measures of Dispersion
  - Correlation
  - Tables & Graphs

Sample vs. Population



Central tendency

- Mean (average)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Notation:
$n = the\ number\ of\ values$
$x_1, x_2, \ldots, x_n = \text{the values}$

- Median
  - The “middle” of a sorted list of numbers.
- Mode
  - The value that appears most often.

Dispersion

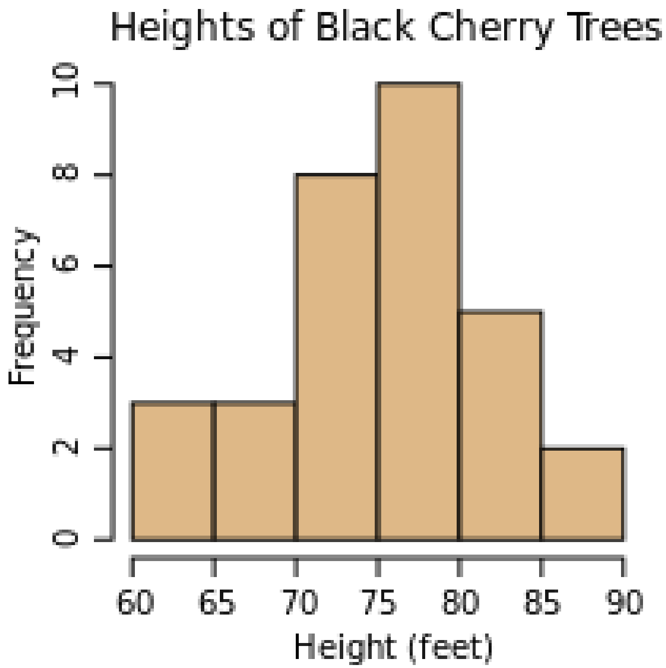
- Range
  - i.e. max - min
- $Variance$

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

- Standard deviation

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

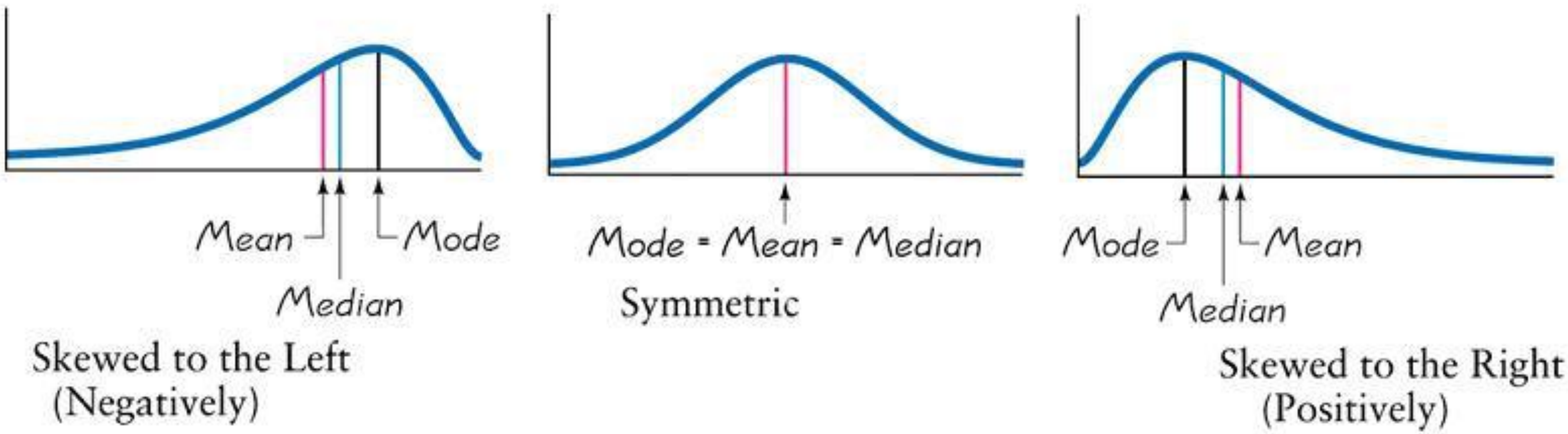
Histogram



A **histogram** is a graphical representation of the distribution of data.

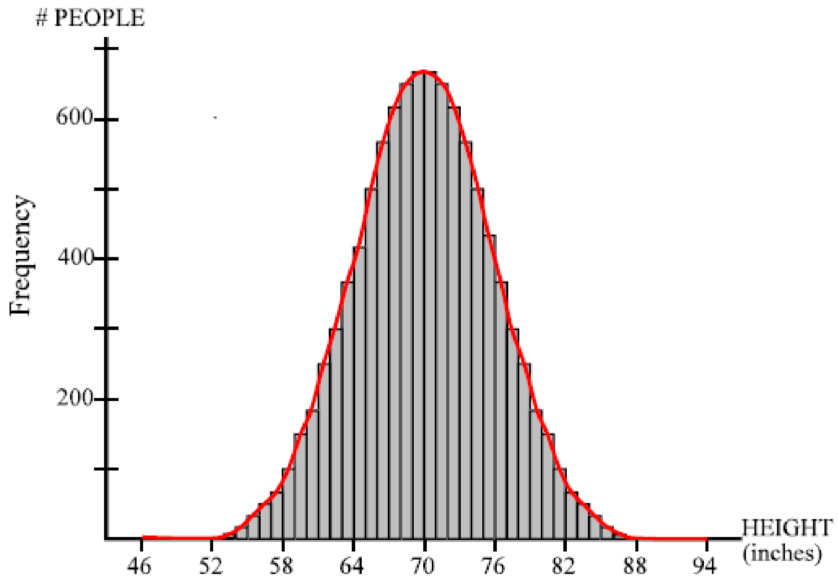
Skewness

Skewness is a measure of the asymmetry of the distribution.



http://rchsbowman.wordpress.com/2010/08/30/statistics-notes-the-shape-of-distributions/

Normal Distribution



- Symmetric
- Bell-shaped

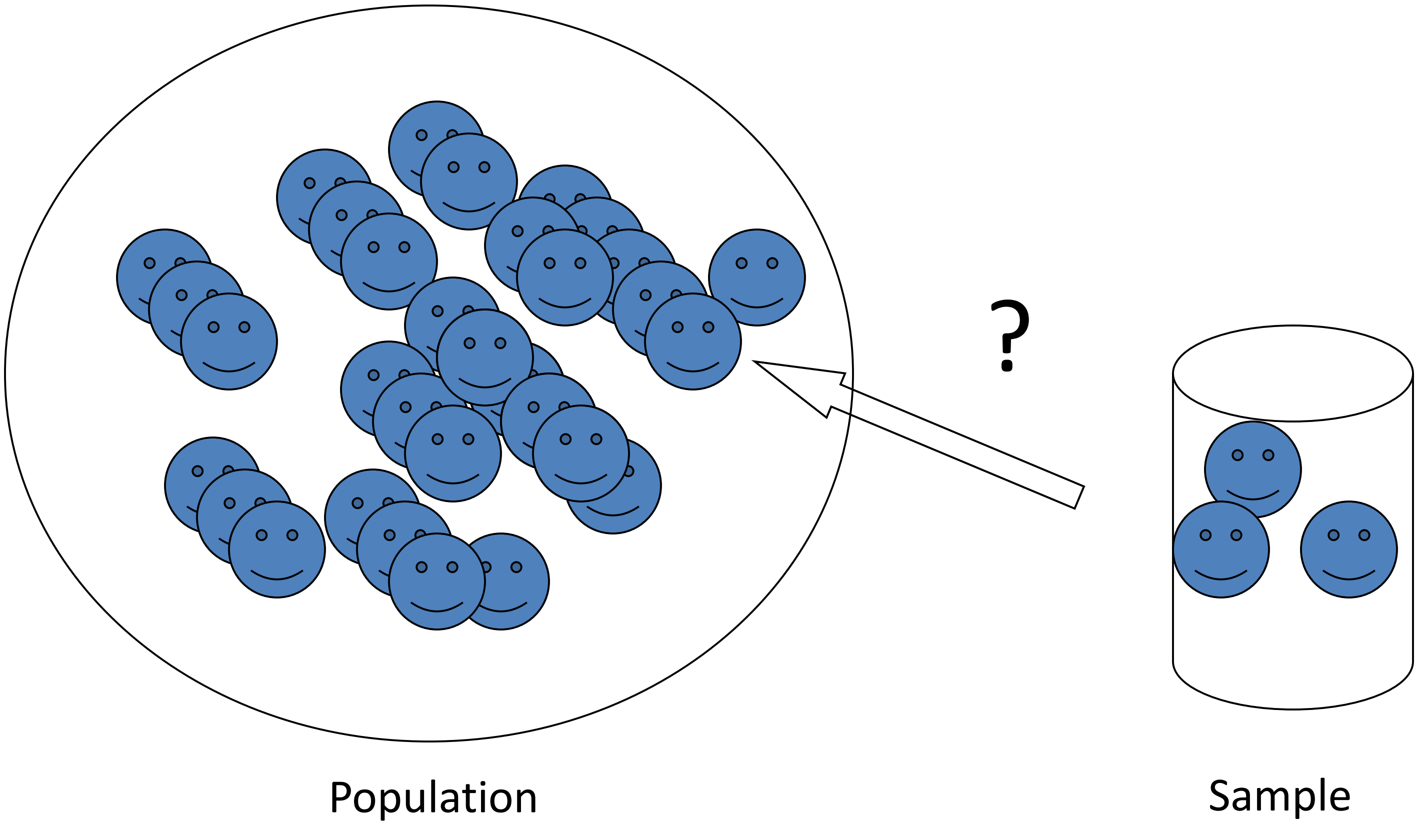
Correlation

- The degree to which two variables have a tendency to vary together
  - Can be positive or negative
  - Range: -1 to +1, 0 means no correlation

Example 1: The more you study, the better you do on the exam (Positive Correlation)

Example 2: The more I talk about correlation, the less you want to be here (Negative Correlation)

Learning from Sample



Hypothesis testing

- **Hypothesis Testing:** A technique for using data to validate/invalidate a claim about a population.

Example 1:
Population mean (Is the average delivery time of 30 mins really true?)

- Null Hypothesis: H_0: $\mu = 30$ (The average delivery time is 30 mins)
- Alternative Hypothesis: H_1: $\mu \neq 30$ (The average delivery time is different from 30 mins)

Example 2:
The difference in two population means (Is it true that the average income is the same in the neighborhood A versus neighborhood B?)

- Null Hypothesis: H_0: $\mu_A = \mu_B$(The average income is the same in the neighborhood A versus neighborhood B)
- Alternative Hypothesis: H_0: $\mu_A \neq \mu_B$(The average income is different in the neighborhood A versus neighborhood B)

Steps in Hypothesis Testing

1. State the Null hypothesis.
2. State the Alternative hypothesis.
3. Calculate the test statistics and p-values.
4. Decision rule for rejection of the Null hypothesis using p-values.

Source: Wissel, et al (2001)

p-values

Hypothesis testing uses **p-values** to weigh the strength of the evidence against the null hypothesis (what the data are telling you about the population).

The p-value is a number between 0 and 1.

- **A small *p*-value (typically ≤ 0.05)** indicates strong evidence against the null hypothesis, so you **reject the null hypothesis**.
- **A large *p*-value (> 0.05)** indicates weak evidence against the null hypothesis, so you **fail to reject the null hypothesis**.

Back to Example 1

Example 1:

Population mean (Is the average delivery time of 30 mins really true?)

- Null Hypothesis: H_0: $\mu = 30$ (The average delivery time is 30 mins)
- Alternative Hypothesis: H_1: $\mu \neq 30$ (The average delivery time is different from 30 mins)

- If ***p*-value ≤ 0.05**, you **reject the null hypothesis.** Therefore, the average delivery time is statistically different from 30 mins.

- If ***p*-value> 0.05,** you **fail to reject the null hypothesis.** Therefore, there is insufficient evidence to conclude that the average delivery time is different from 30 mins.

Back to Example 2

Example 2:
The difference in two population means (Is it true that the average income is the same in the neighborhood A versus neighborhood B?)

- Null Hypothesis: H_0: $\mu_A = \mu_B$(The average income is the same in the neighborhood A versus neighborhood B)
- Alternative Hypothesis: H_0: $\mu_A \neq \mu_B$(The average income is different in the neighborhood A versus neighborhood B)

- If ***p*-value ≤ 0.05**, you **reject the null hypothesis.** Therefore, The average income is significantly different in the neighborhood A versus neighborhood B.

- If ***p*-value> 0.05,** you **fail to reject the null hypothesis.** Therefore, there is insufficient evidence that the average income is different in the neighborhood A versus neighborhood B.