Statistics for Business

Background: Descriptive Statistics

Panagiotis Th. Konstantinou

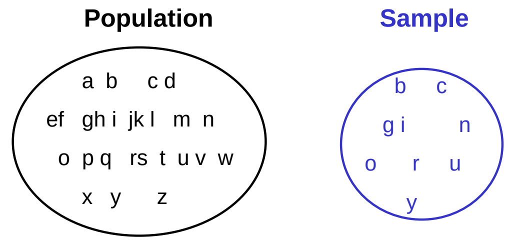
MSc in International Shipping, Finance and Management,

Athens University of Economics and Business

First Draft: July 15, 2015. **This Draft**: August 30, 2023.

Key Concepts

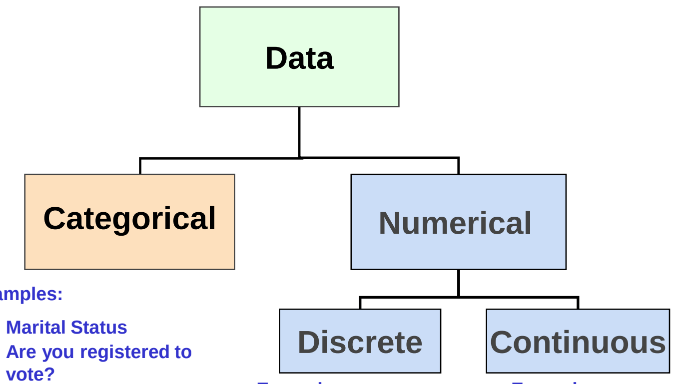
- A **population** is the collection of all items of interest or under investigation (N represents the population size)
- A **sample** is an observed subset of the population (n represents the sample size)
- A **parameter** is a specific characteristic of a population
- A **statistic** is a specific characteristic of a sample



Values calculated using population data are called parameters

Values computed from sample data are called statistics

Data Types



Examples:

- Marital Status
- Are you registered to vote?
- Eye Color

(Defined categories or groups)

Examples:

- Number of Children
- Defects per hour

(Counted items)

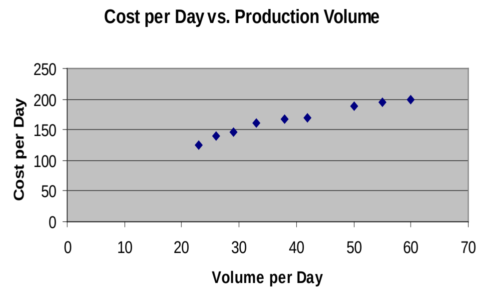
Examples:

- Weight
- Voltage

(Measured characteristics)

Relationships Between Variables

Volume per day	Cost per day
23	125
26	140
29	146
33	160
38	167
42	170
50	188
55	195
60	200



Investment Category	Investor A	Investor B	Investor C	Total
Stocks	46.5	55	27.5	**129**
Bonds	32.0	44	19.0	**95**
CD	15.5	20	13.5	**49**
Savings	16.0	28	7.0	**51**
Total	**110.0**	**147**	**67.0**	**324**

Describing Data Numerically

- Describing Data Numerically
  - Central Tendency
    - Arithmetic Mean
    - Median
    - Mode
  - Variation
    - Range
    - Interquartile Range
    - Variance
    - Standard Deviation
    - Coefficient of Variation

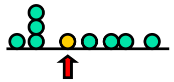
Measures of Central Tendency

- Central Tendency
  - Mean
  - Median
  - Mode

Mean: $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$

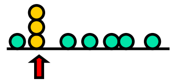
Arithmetic average

Median:



Midpoint of ranked values

Mode:



Most frequently observed value

- Median position $\frac{n+1}{2}$ position in the ordered data
  - If the number of values is odd, the median is the middle number
  - If the number of values is even, the median is the average of the two middle numbers

Measures of Central Tendency

Example

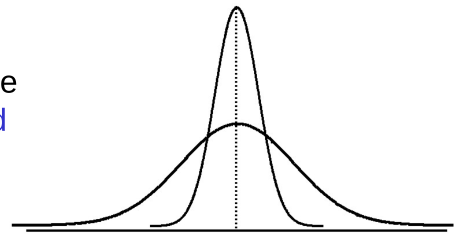
House Prices	
	\$2,000,000
	500,000
	300,000
	100,000
	100,000
Sum	\$3,000,000

- **Mean**: \$3,000,000/5 = \$600,000
- **Median**: middle value of ranked data = \$300,000
- **Mode**: most frequent value = \$100,000

Measures of Variability

- Variation
  - Range
  - Interquartile Range
  - Variance
  - Standard Deviation
  - Coefficient of Variation

- Measures of variation give information on the spread or variability of the data values.



Same center, different variation

Variance

- ***Population Variance***: Average of squared deviations of values from the mean

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

where

  - $\mu =$ population mean
  - $N =$ population size
  - $X_i = i-$th value of the variable X

- ***Sample Variance***: Average (approximately) of squared deviations of values from the sample mean:

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

where

  - $\bar{x} =$ sample mean/average
  - $n =$ sample size
  - $x_i = i-$th value of the variable X

Standard Deviation

- ***Population Standard Deviation***: Most commonly used measure of variation
  - Shows variation about the mean
  - Has the *same units as the original data*

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

- ***Sample Standard Deviation***: Most commonly used measure of variation
  - Shows variation about the *sample* mean
  - Has the *same units as the original data*

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

Standard Deviation

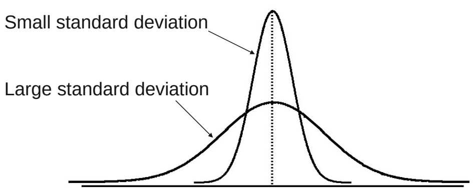
Example: Sample Standard Deviation Computation

- Sample Data (x_i) : 10 12 14 15 17 18 18 24
- $n = 8$ and sample mean $= \bar{x} = 16$
- So the standard deviation is

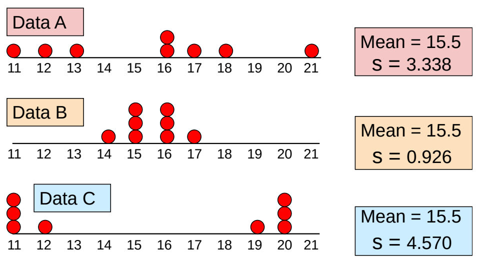
$$\begin{aligned} s &= \sqrt{\frac{(10-\bar{x})^2 + (12-\bar{x})^2 + (14-\bar{x})^2 + \cdots + (24-\bar{x})^2}{n-1}} \\ &= \sqrt{\frac{(10-16)^2 + (12-16)^2 + (14-16)^2 + \cdots + (24-16)^2}{8-1}} \\ &= \sqrt{\frac{126}{7}} = 4.2426 \end{aligned}$$

- This is a measure of the "**average**" scatter around the (sample) mean.

Comparing Standard Deviations

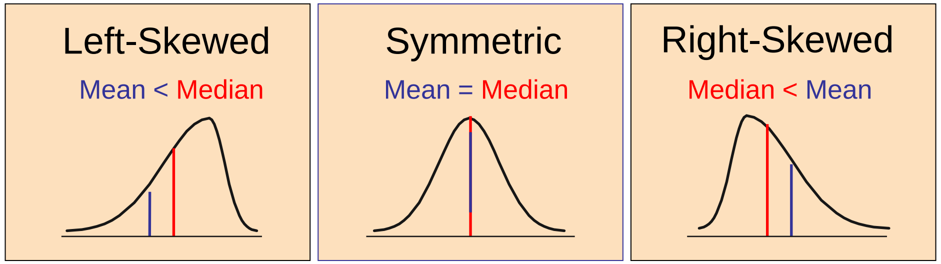


- The smaller the standard deviation, the more concentrated are the values around the mean.



- Same mean, different standard deviations.

Shape of a Distribution



- Describes how data are distributed
- Measures of **shape**:
  - Symmetric or skewed
  - Left = Negative (mass of distr. concentrated on the right of figure); Right = Positive (mass of distr. concentrated on the left of figure).

$$SK = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}$$

Coefficient of Variation

- Measures relative variation and is always in percentage (%)
- Shows variation **relative to mean**
- Can be used to compare two or more sets of data **measured in different units**

$$CV = \left(\frac{s_x}{\bar{x}}\right) \cdot 100\%$$

- **Stock A**:
  - Avg price last year = \$50
  - Standard deviation = \$5

$$CV_A = \left(\frac{\$5}{\$50}\right) \cdot 100\% = 10\%$$

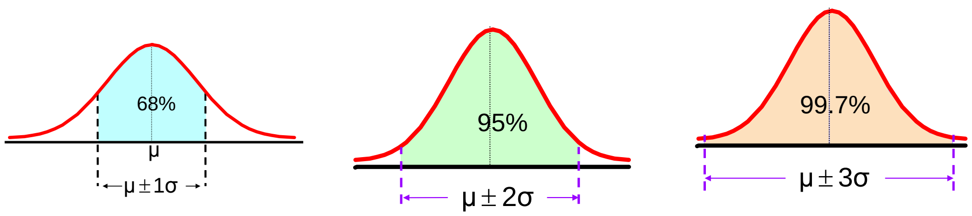
- **Stock B**:
  - Avg. price last year = \$100
  - Standard deviation = \$5

$$CV_B = \left(\frac{\$5}{\$100}\right) \cdot 100\% = 5\%$$

- Both stocks have the same standard deviation, but stock B is less variable relative to its price

The Empirical Rule

If the data distribution is bell-shaped, then the interval:



- $\mu \pm 1\sigma$ contains about 68% of the values in the population or the sample
- $\mu \pm 2\sigma$ contains about 95% of the values in the population or the sample
- $\mu \pm 3\sigma$ contains almost all (about 99.7%) of the values in the population or the sample.

Covariance

- The covariance measures the strength of the linear relationship between **two variables**
- The ***population covariance***:

$$\mathrm{Cov}(X, Y) = \sigma_{XY} = \frac{\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y)}{N}.$$

- The ***sample covariance***:

$$\widehat{\mathrm{Cov}(x, y)} = s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1}.$$

- Only concerned with the strength of the relationship
- No causal effect is implied
  - $\mathrm{Cov}(x, y) > 0$, x and y tend to move in the ***same*** direction
  - $\mathrm{Cov}(x, y) < 0$, x and y tend to move in ***opposite*** directions

Correlation Coefficients

- The correlation coefficient measures the relative strength of the linear relationship between **two variables**
- The ***population correlation coefficient***:

$$\mathrm{Corr}(X, Y) = \rho_{XY} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

- The ***sample correlation coefficient***:

$$\widehat{\mathrm{Corr}(x, y)} = r_{xy} = \frac{\widehat{\mathrm{Cov}(x, y)}}{s_x s_y}.$$

- Unit free and ranges between -1 and 1
  - The closer to -1, the stronger the negative linear relationship
  - The closer to 1, the stronger the positive linear relationship
  - The closer to 0, the weaker any positive linear relationship

Correlation Coefficients

Examples