Economic Risk and Decision Analysis for Oil and Gas Industry
CE81.9008

School of Engineering and Technology
Asian Institute of Technology

January Semester

Presented by
Dr. Thitisak Boonpramote

Department of Mining and Petroleum Engineering, Chulalongkorn University

Review of Descriptive Statistics

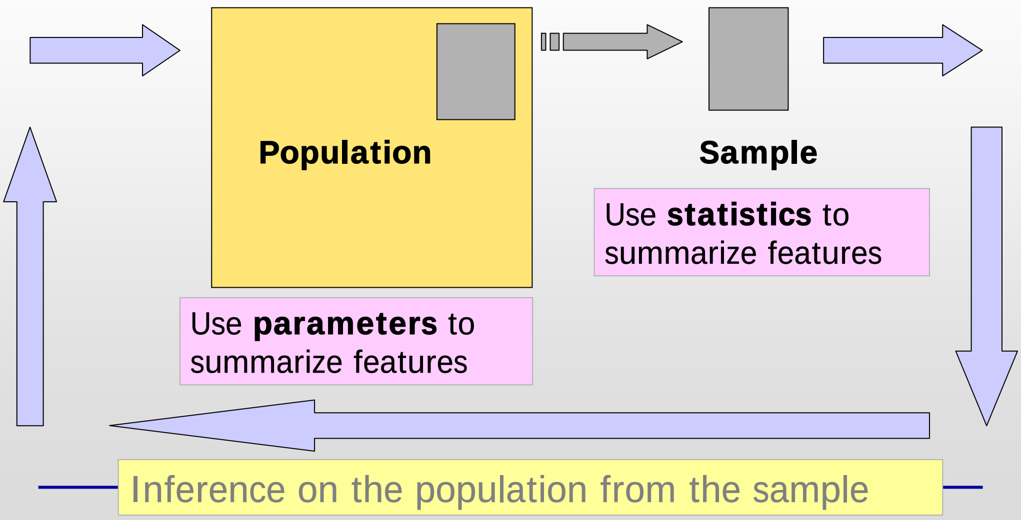
Why you needs to know about Statistics

- To know how to properly present information
- To know how to draw conclusions about populations based on sample information
- To know how to improve processes
- To know how to obtain reliable forecasts

Key Definitions

- A population (universe) is the collection of things under consideration.
  - data set that contains all possible items of interest
- A sample is a portion of the population selected for analysis.
  - data set that contains only a few random or otherwise representative elements of a data set
- A parameter is a summary measure computed to describe a characteristic of **the population.**
  - measures of central tendency and other statistical characteristics that describe a population
- A statistic is a summary measure computed to describe a characteristic of **the sample**
  - corresponding measures and statistical characteristics that describe a sample

Population and Sample

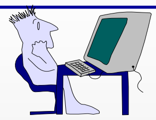


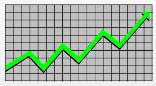
Statistical Methods

- **Descriptive statistics**
  - Collecting and describing data
  - What to describe?
    - What is the “location” or “center” of the data? (“measures of location”)
    - How do the data vary? (“measures of variability”)
- **Inferential statistics**
  - Drawing conclusions and/or making decisions concerning a population based only on sample data

Descriptive Statistics

- Collect data
  - e.g. Survey
- Present data
  - e.g. Tables and graphs
- Characterize data
  - e.g. Sample mean = $\frac{\sum X_i}{n}$



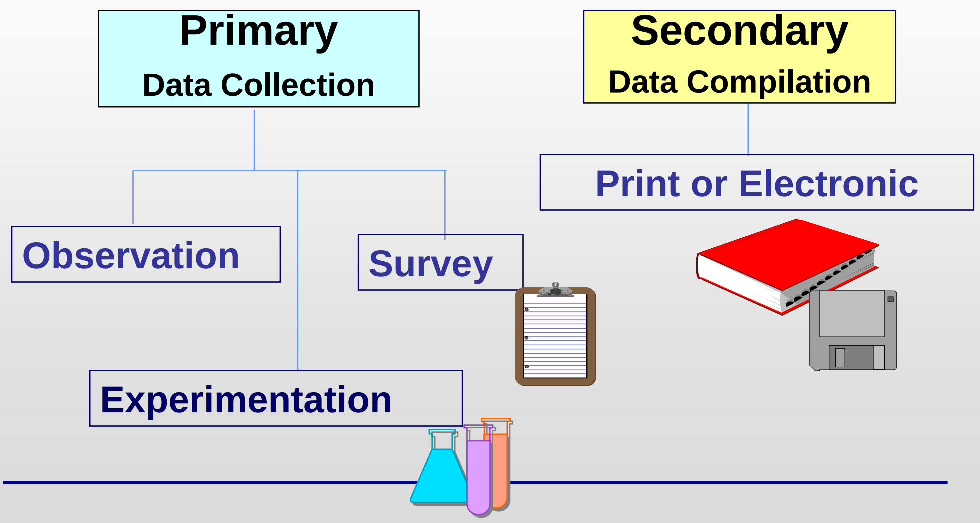


Inferential Statistics

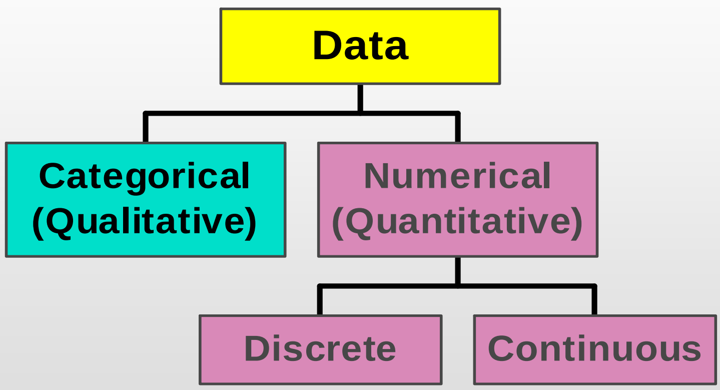
- Estimation
  - e.g.: Estimate the population mean weight using the sample mean weight
- Hypothesis testing
  - e.g.: Test the claim that the population mean weight is 120 pounds

Drawing conclusions and/or making decisions concerning a population based on sample results.

Data Sources



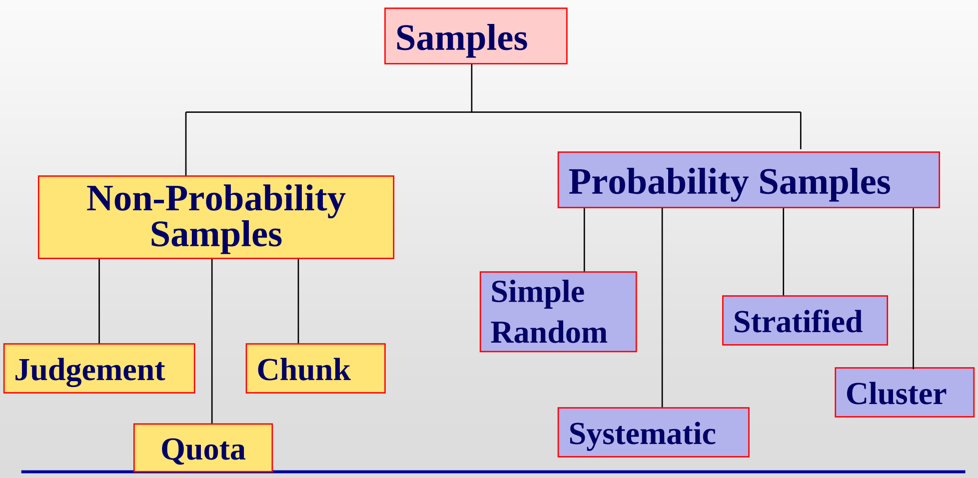
Types of Data



Reasons for Drawing a Sample

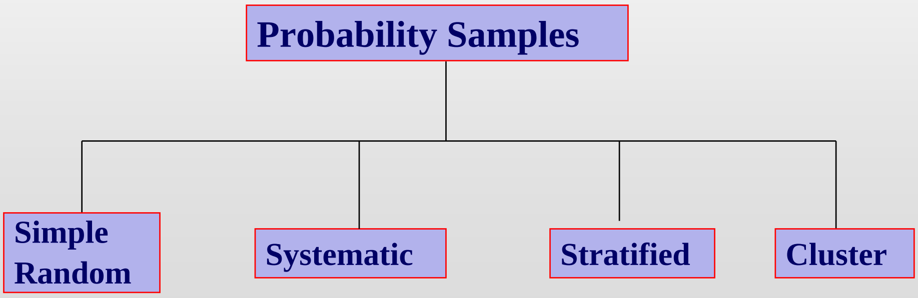
- Less time consuming than a census
- Less costly to administer than a census
- Less cumbersome and more practical to administer than a census of the targeted population

Types of Sampling Methods



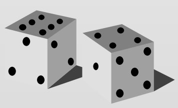
Probability Sampling

- Subjects of the sample are chosen based on known probabilities



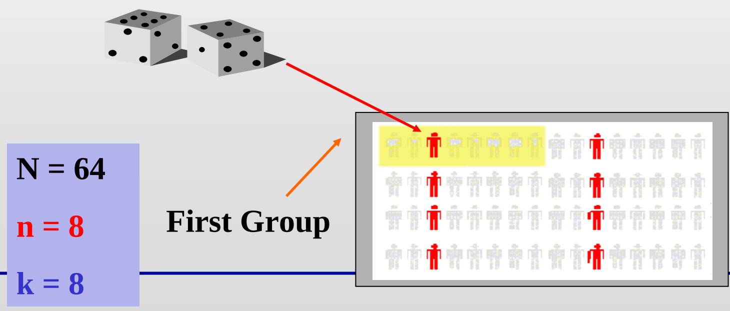
Simple Random Samples

- Every individual or item from the frame has an **equal chance** of being selected
- Selection may be **with replacement** or **without replacement**
- Samples obtained from **table of random numbers** or **computer random number generators**



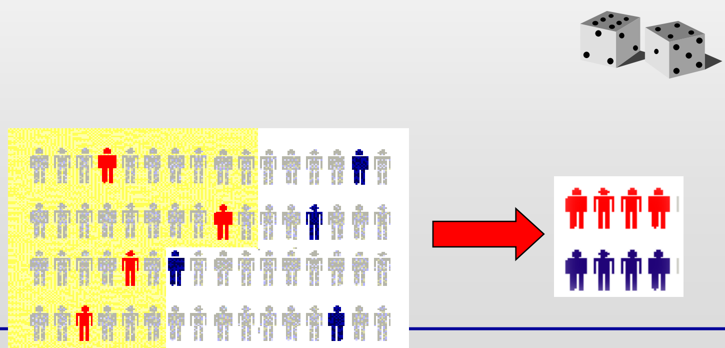
Systematic Samples

- Decide on sample size: n
- Divide frame of N individuals into groups of k individuals: $k=n/N$
- Randomly select one individual from the 1st group
- Select every k-th individual thereafter



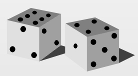
Stratified Samples

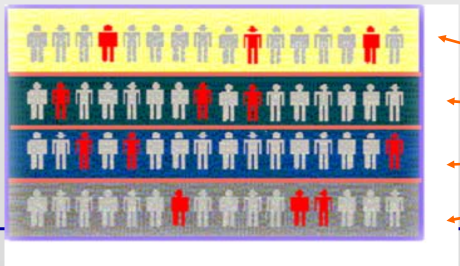
- Population divided into two or more groups according to **some common characteristic**
- Simple random sample selected from each group
- The two or more samples are combined into one



Cluster Samples

- Population divided into several “clusters,” each representative of the population
- Simple random sample selected from each
- The samples are combined into one



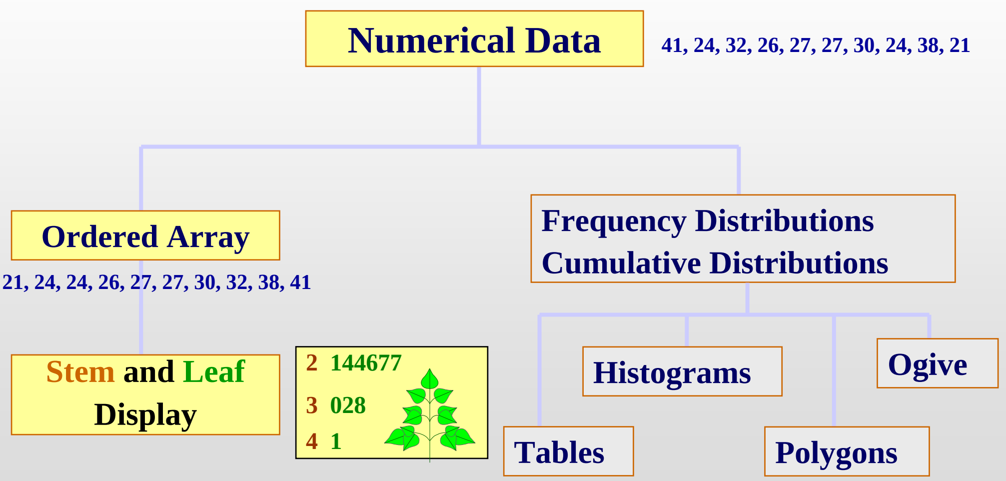


Population divided into 4 clusters.

Advantages and Disadvantages

- **Simple random sample and systematic sample**
  - Simple to use
  - May not be a good representation of the population’s underlying characteristics
- **Stratified sample**
  - Ensures representation of individuals across the entire population
- **Cluster sample**
  - More cost effective
  - Less efficient (need larger sample to acquire the same level of precision)

Organizing Numerical Data



Organizing Numerical Data

(continued)

- Data in *raw* form (as collected):
  24, 26, 24, 21, 27, 27, 30, 41, 32, 38
- Data in *ordered array* from *smallest to largest*:
  21, 24, 24, 26, 27, 27, **30**, 32, 38, 41
- Stem-and-leaf display:

2	144677
3	028
4	1

Tabulating and Graphing Numerical Data

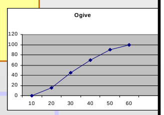
Numerical Data 41, 24, 32, 26, 27, 27, 30, 24, 38, 21

Ordered Array
21, 24, 24, 26, 27, 27, 30, 32, 38, 41

Stem and Leaf Display

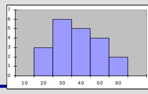
2	144677
3	028
4	1

Frequency Distributions
Cumulative Distributions



Ogive

Histograms



Tables

Polygons

Tabulating Numerical Data: Frequency Distributions

- Sort raw data in ascending order:
  12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58
- Find range: **58 - 12 = 46**
- Select number of classes: **5 (usually between 5 and 15)**
- Compute class interval (width): **10 (46/5 then round up)**
- Determine class boundaries (limits): **10, 20, 30, 40, 50, 60**
- Compute class midpoints: **15, 25, 35, 45, 55**
- Count observations & assign to classes

Frequency Distributions, Relative Frequency Distributions and Percentage Distributions

Data in ordered array:

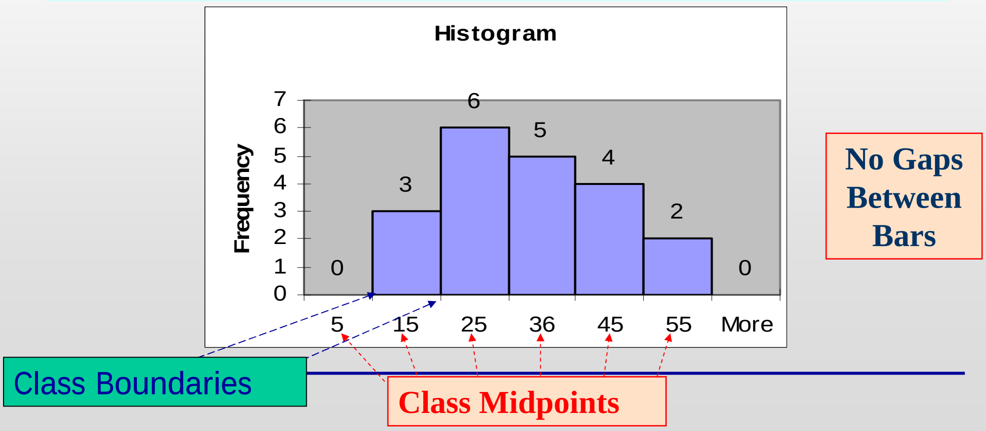
12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

Class	Frequency	Relative Frequency	Percentage
10 but under 20	3	.15	15
20 but under 30	6	.30	30
30 but under 40	5	.25	25
40 but under 50	4	.20	20
50 but under 60	2	.10	10
Total	**20**	**1**	**100**

Graphing Numerical Data: The Histogram

Data in ordered array:

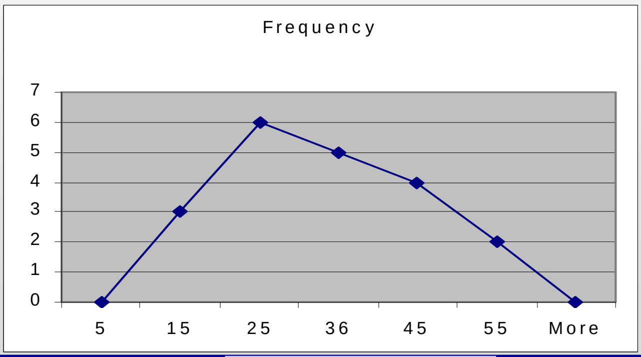
12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58



Graphing Numerical Data: The Frequency Polygon

Data in ordered array:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58



Class Midpoints

Tabulating Numerical Data: Cumulative Frequency

Data in ordered array:

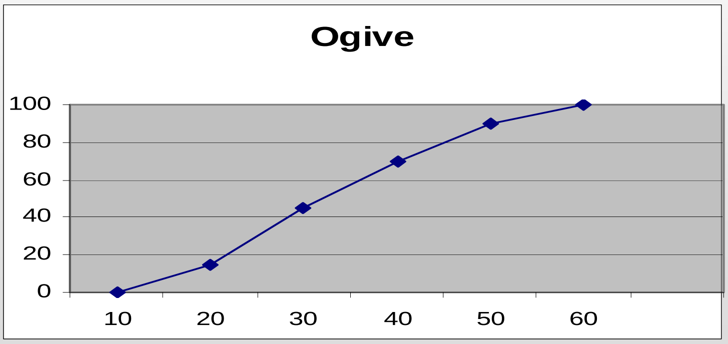
12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

Class	Cumulative Frequency	Cumulative % Frequency
10 but under 20	3	15
20 but under 30	9	45
30 but under 40	14	70
40 but under 50	18	90
50 but under 60	20	100

Graphing Numerical Data: The Ogive (Cumulative % Polygon)

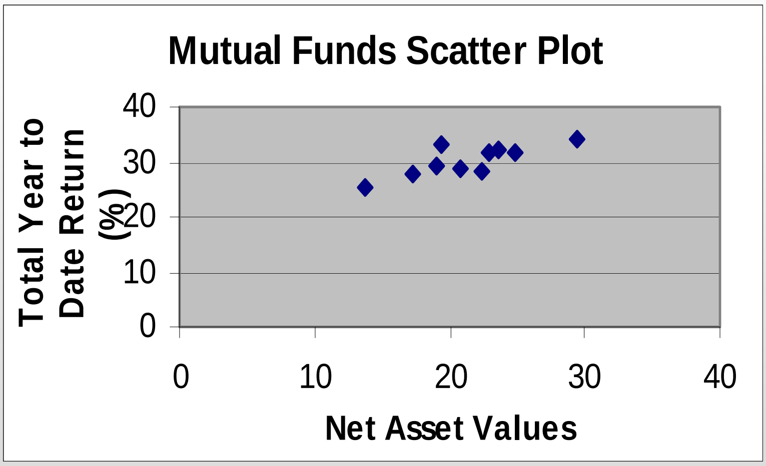
Data in ordered array:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58



Class Boundaries (*Not Midpoints*)

Graphing Bivariate Numerical Data (Scatter Plot)



Tabulating and Graphing Categorical Data: Univariate Data

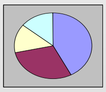
Categorical Data

- **Tabulating Data** — The **Summary Table**
- **Graphing Data**
  - **Pie Charts**
  - **Bar Charts**
  - **Pareto Diagram**

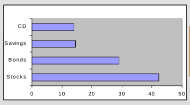
Graphing Categorical Data: Univariate Data

Categorical Data

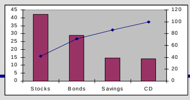
- **Tabulating Data** — The Summary Table
- **Graphing Data**
  - **Pie Charts**



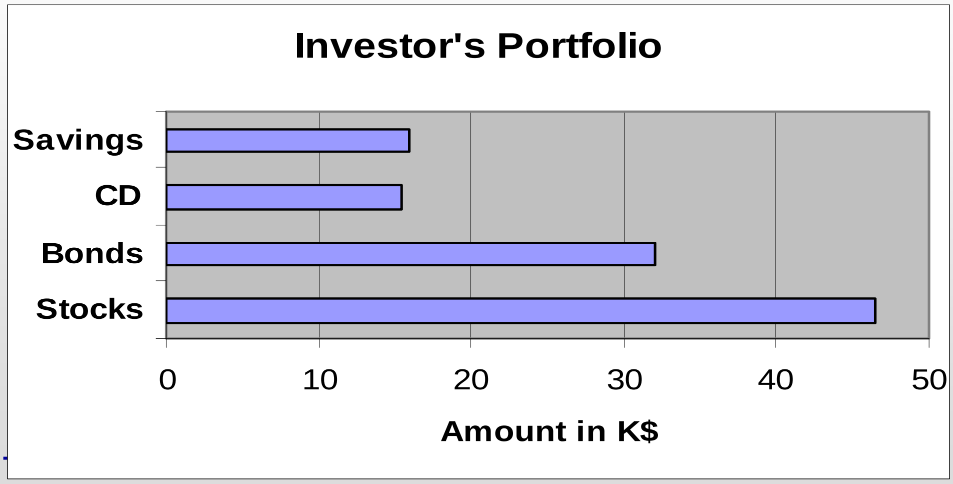
  - **Bar Charts**



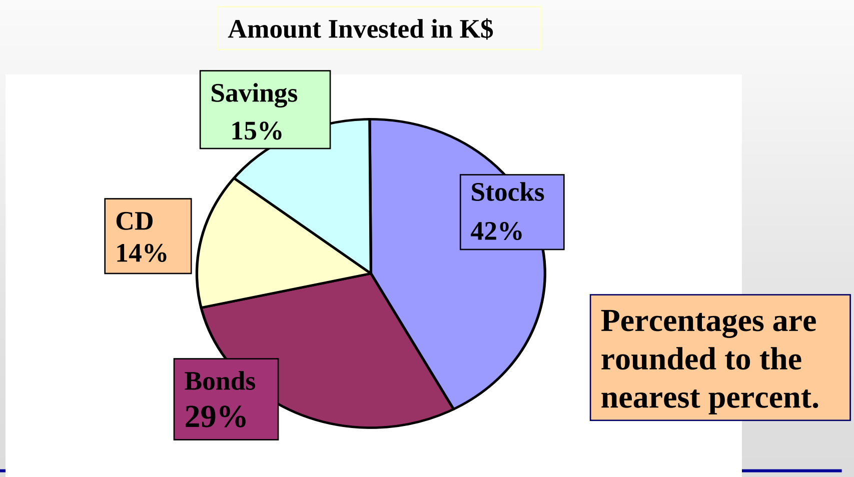
  - **Pareto Diagram**



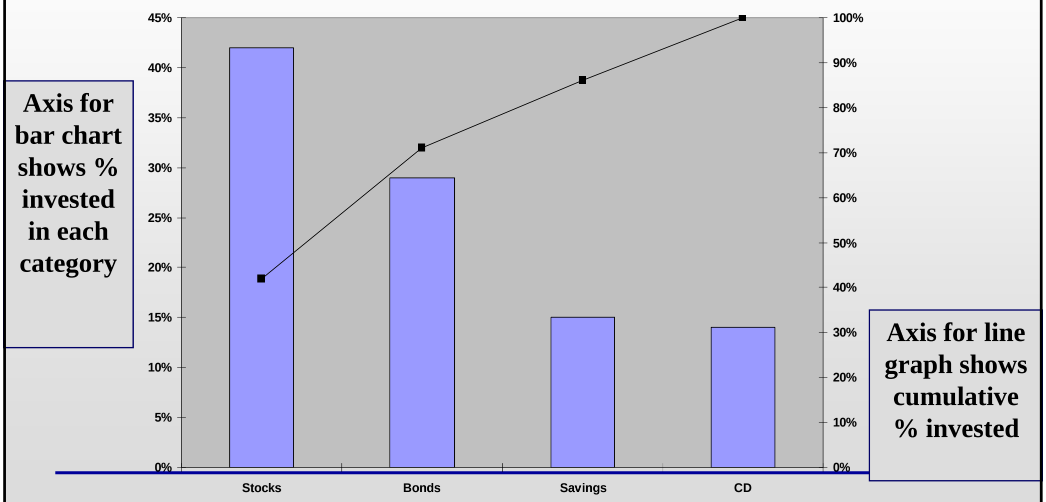
Bar Chart



Pie Chart



Pareto Diagram

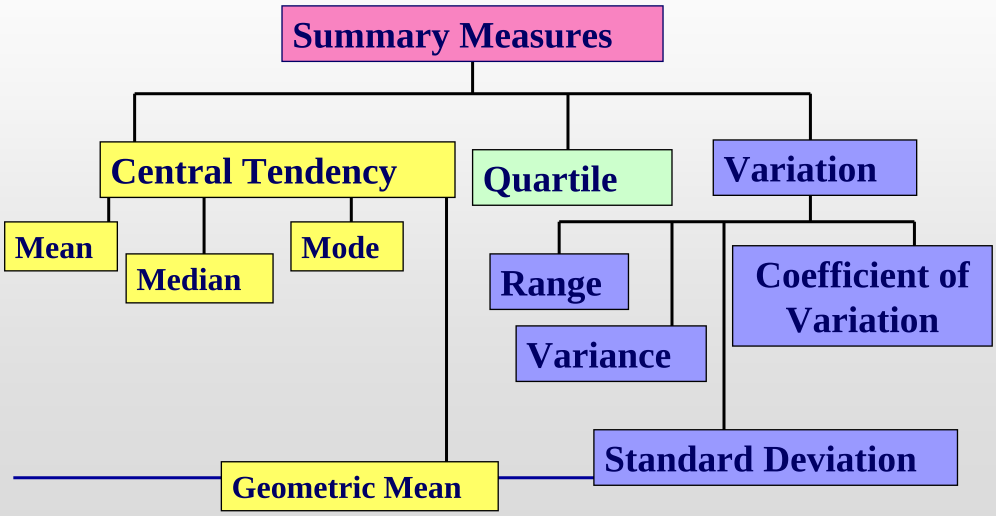


Tabulating and Graphing Bivariate Categorical Data

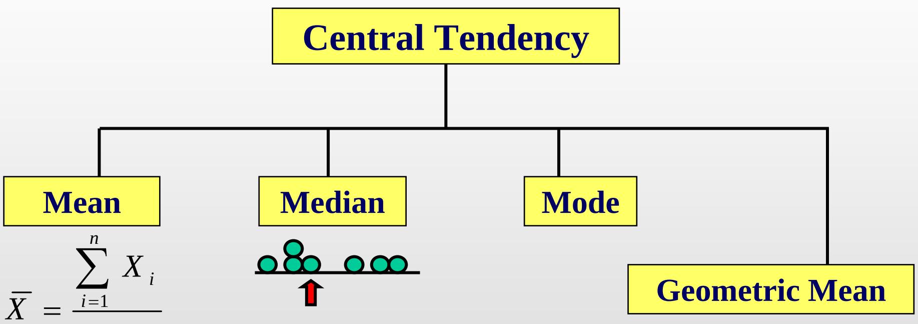
- Contingency tables: investment in thousands of dollars

Investment Category	Investor A	Investor B	Investor C	Total
Stocks	46.5	55	27.5	**129**
Bonds	32	44	19	**95**
CD	15.5	20	13.5	**49**
Savings	16	28	7	**51**
Total	**110**	**147**	**67**	**324**

Summary Measures



Measures of Central Tendency



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

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

$$\bar{X}_G = \left(X_1 \times X_2 \times \cdots \times X_n \right)^{1/n}$$

Mean

- Also called arithmetic mean
  - Symbol μ represents population mean
  - Symbol $\overline{X}$ represents sample mean
  - Commonly called average
- Calculated by adding values of all items in data set and dividing by total number of items in set

$$\overline{X} = \frac{\boxed{\sum_{i=1}^{n} X_i}}{\boxed{n}}$$

Mean of sample

Sum of all data elements in sample

Number of data elements in sample

Mean

(continued)

- Mean of data values
  - Sample mean

Sample Size

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

  - Population mean

Population Size

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{X_1 + X_2 + \cdots + X_N}{N}$$

Properties of Mean

- Sum of deviations from mean is zero

Population $$\sum_{i=1}^{N} \left(X_i - \overline{X} \right) = 0$$

- Sum of squared deviations minimized when deviations are measured from mean

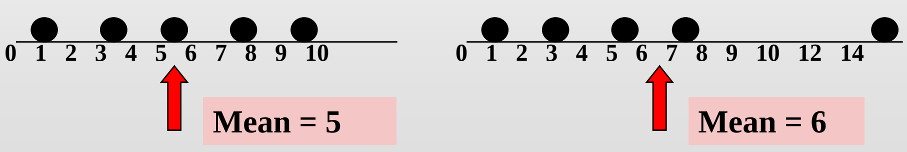
$$\sum_{i=1}^{N} \left(X_i - \overline{X} \right)^2 = \text{minimum}$$

Mean may be influenced by extreme values.

Mean

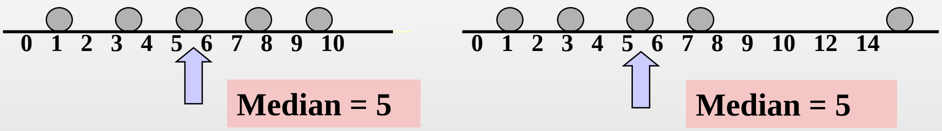
(continued)

Affected by extreme values (outliers)



Median

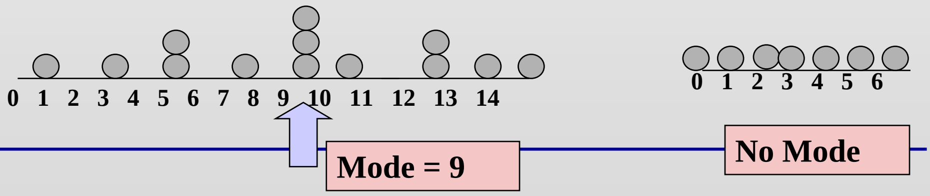
- Robust measure of central tendency
- Not affected by extreme values



- In an ordered array, the median is the “middle” number
  - If n or N is odd, the median is the middle number
  - If n or N is even, the median is the average of the two middle numbers

Mode

- Poor measure of central tendency in most cases–does not take into account values of other data elements
- Value that **occurs most often**
- Not affected by extreme values
- Used for either numerical or categorical data
- There may be no mode
- There may be several modes e.g., bimodal (two modes)



Calculate Footage Drilled

- 20 bits drilled 2,013 ft
- Determine mean, median, mode

Bit number	Ft. Drilled
1	53
2	69
3	72
4	76
5	80
6	89
	90
	95
9	102
10	102

Mode
(most frequent)

Bit number	Ft. Drilled
11	105
12	108
13	109
14	110
15	115
1	
1	
1	
1	
20	139

Median

$$\frac{102+105}{2} = 103.5 \text{ ft}$$

Calculate Footage Drilled

- 20 bits drilled 2,013 ft
- Determine mean, median, mode

$$= \sum_{i=1}^{n} x_i$$

Bit number	Ft. Drilled
1	53
8	95
9	102
10	102

Mean

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{2{,}013}{20} = 100.65 \text{ ft}$$

Bit number	Ft. Drilled
11	105
	108
	109
	110
	115
	116
	123
18	125
19	135
$n =$ 20	139

Geometric Mean (G_m)

- *N*th root of product of individual data elements of data set with *N* elements

$$G_m = \sqrt[n]{X_1 \bullet X_2 \bullet X_3 \bullet \ldots \bullet X_n}$$

  - Calculation simplified using logarithms

$$G_m = anti\ \log\left(\frac{1}{n}\sum_{i=1}^{n}\log X_i\right) = 10^{\left(\frac{1}{n}\sum_{i=1}^{n}\log X_i\right)}$$

  - In terms of natural logarithms

$$G_m = e^{\left(\frac{1}{n}\sum_{i=1}^{n}\ln X_i\right)}$$

Properties of the Geometric Mean

- Biased toward smaller values; appropriate for skewed data sets (asymmetrical distributions)
- Not affected as much as arithmetic mean by extreme values
- Undefined for data sets with negative or zero values

Geometric Mean

- Useful in the measure of **rate of change of a variable over time**

$$\bar{X}_G = \left(X_1 \times X_2 \times \cdots \times X_n\right)^{1/n}$$

- Geometric mean **rate of return**
  - Measures the status of an investment over time

$$\bar{R}_G = \left[\left(1+R_1\right)\times\left(1+R_2\right)\times\cdots\times\left(1+R_n\right)\right]^{1/n} - 1$$

Example

An investment of $100,000 declined to $50,000 at the end of year one and rebounded to $100,000 at end of year two:

$$X_1 = \$100,000 \qquad X_2 = \$50,000 \qquad X_3 = \$100,000$$

Average rate of return:

$$\bar{X} = \frac{(-50\%)+(100\%)}{2} = 25\%$$

Geometric rate of return:

$$\bar{R}_G = \left[\left(1+(-50\%)\right)\times\left(1+(100\%)\right)\right]^{1/2} - 1$$

$$= \left[(0.50)\times(2)\right]^{1/2} - 1 = 1^{1/2} - 1 = 0\%$$

Harmonic Mean, H_m

- Reciprocal of arithmetic mean of reciprocals of data elements in data set

$$H_m = \frac{n}{\sum_{i=1}^{n} \frac{1}{X_i}}$$

Quadratic Mean, Q_m (Root Mean Square)

$$Q_m = \left(\frac{\sum_{i=1}^{n} X_i^2}{n} \right)^{1/2}$$

Weighted Average

- Averages in which data elements are weighted by frequency of occurrence

$$X_w = \frac{\sum_{i=1}^{n} \boxed{w_i} X_i}{\sum_{i=1}^{n} \boxed{w_i}}$$

Weighting factor of element X_i

Weighted Average

- Weighted geometric mean (G_{wm})

$$G_{wm} = anti\ \log\left(\frac{\sum_{i=1}^{n} w_i \log X_i}{\sum_{i=1}^{n} w_i}\right)$$

- Weighted harmonic mean (H_{wm})

$$H_{wm} = \left(\frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} \frac{w_i}{X_i}}\right)$$

Calculate Footage Drilled

- Determine geometric, harmonic, quadratic means of bit record

Bit	Ft (X_i)	Log X_i	1/X_i	X_i^2
1	53	1.7243	0.0189	2,809
2	69	1.8388	0.0145	4,761
3	72	1.8573	0.0139	5,184
4	76	1.8808	0.0132	5,776
5	80	1.9031	0.0125	6,400
6	89	1.9494	0.0112	7,921
…	…	…	…	…
20	139	2.1430	0.0072	19,321
n = 20	2,013	39.8215	0.21057	212,435

Calculate Footage Drilled

- Determine geometric, harmonic, quadratic means of bit record

Bit	Ft (X_i)	Log X_i	1/X_i	X_i^2
n = 20		39.8215	0.21057	212,435

- Geometric mean

$$G_m = 10^{\left(\frac{1}{n}\sum_{i=1}^{n}\log X_i\right)} = 10^{\left(\frac{1}{20}\times 39.8215\right)}$$

$$= 97.97 \text{ ft}$$

Calculate Footage Drilled

- Determine geometric, harmonic, quadratic means of bit record

Bit	Ft (X_i)	Log X_i	1/X_i	X_i^2
n = 20		39.8215	0.21057	212,435

- Harmonic mean

$$H_m = \frac{n}{\sum_{i=1}^{n} \frac{1}{X_i}} = \frac{20}{0.2105} = 95.01 \text{ ft}$$

Calculate Footage Drilled

- Determine geometric, harmonic, quadratic means of bit record

Bit	Ft (X_i)	Log X_i	1/X_i	X_i^2
n = 20		39.8215	0.21057	212,435

- Quadratic mean

$$Q_m = \frac{\sum_{i=1}^{n} X_i^2}{n} = \frac{212,435}{20} = 103.6 \text{ ft}$$

Calculate Weighted-Average Porosity

- 22-ft pay zone
- Calculate weighted-average porosity

Porosity (ϕ), %	Thickness (h), ft	$\phi \times h$
15.2	2	30.4
	3	31.5
	5	66.0
	2	37.2
	6	68.4
	4	50.4
	Σh = 22	$\Sigma \phi h$ = 283.9

$$\phi_w = \frac{\sum_{i=1}^{n} \phi_i h_i}{\sum_{i=1}^{n} h_i} = \frac{283.9}{22} = 12.9$$

Weighted porosity ($\phi \times h$)

Calculate Weighted-Average Cost of Capital

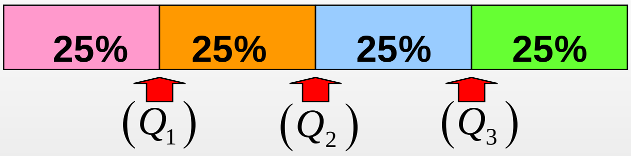
- Company will invest in $500,000 project
  - $150,000 equity at cost of 8%
  - 350,000 long-term debt at 18%
- Calculate weighted-average cost of capital

$$i_w = \frac{150{,}000 \times 0.08 + 350{,}000 \times 0.18}{500{,}000}$$

$$= \frac{12{,}000 + 63{,}000}{500{,}000} \times 100 = 15\%$$

Quartiles

- Split Ordered Data into 4 Quarters

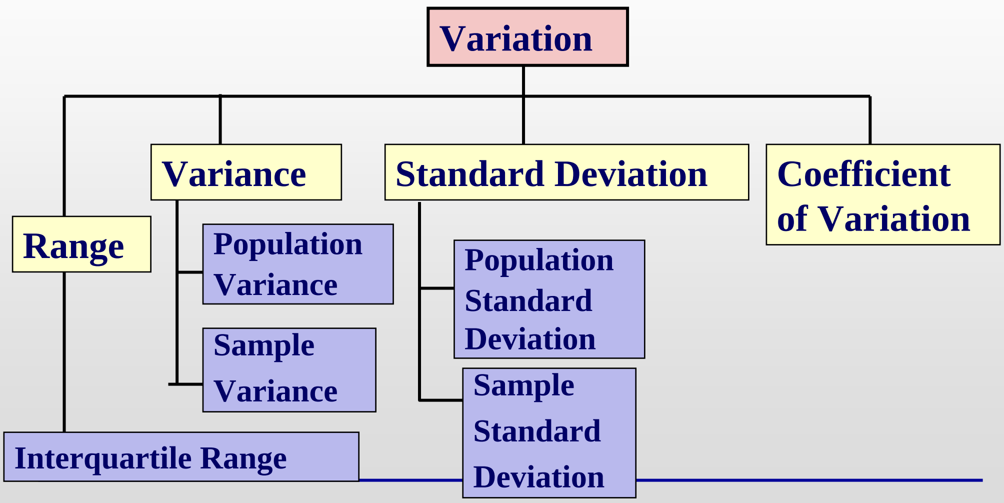


- Position of i-th Quartile $(Q_i) = \frac{i(n+1)}{4}$
- Q_1 and Q_3 Are Measures of Noncentral Location
- Q_2 = Median, A Measure of Central Tendency

Data in Ordered Array: 11 12 13 16 16 17 18 21 22

$$\text{Position of } Q_1 = \frac{1(9+1)}{4} = 2.5 \qquad Q_1 = \frac{(12+13)}{2} = 12.5$$

Measures of Variation



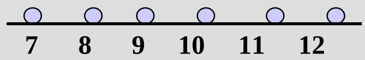
Range

- Measure of variation
- Difference between the largest and the smallest observations:

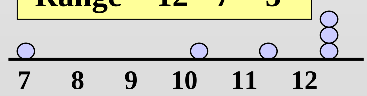
$$\text{Range} = X_{\text{Largest}} - X_{\text{Smallest}}$$

- Ignores the way in which data are distributed
- Not particularly useful measure of dispersion, since it uses only two values from data set

Range = 12 - 7 = 5



Range = 12 - 7 = 5



Interquartile Range

- Measure of variation
- Also known as **midspread**
  - Spread in the middle 50%
- Difference between the first and third quartiles

Data in Ordered Array: 11 12 13 16 16 17 17 18 21

$$\text{Interquartile Range} = Q_3 - Q_1 = 17.5 - 12.5 = 5$$

- Not affected by extreme values

Variance

- Important measure of variation
- Shows variation about the mean
  - Sample variance:

$$S^2 = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2}{n-1}$$

  - Population variance:

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

Standard Deviation

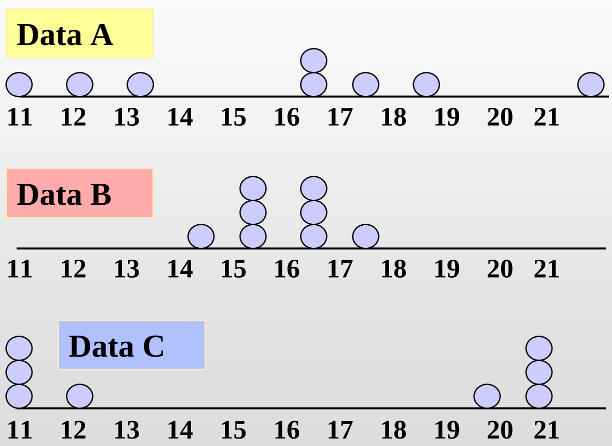
- Most important measure of variation
- Shows variation about the mean
- Has the same units as the original data
  - Sample standard deviation:

$$S = \sqrt{\frac{\sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2}{n-1}}$$

  - Population standard deviation:

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

Comparing Standard Deviations



Mean = 15.5
s = 3.338

Mean = 15.5
s = .9258

Mean = 15.5
s = 4.57

Mean Absolute Deviation, d_m

- Average deviation of data from the mean over all observations

$$d_m = \frac{\boxed{\sum_{i=1}^{n} \left| X_i - \bar{X} \right|}}{n}$$

Absolute

- For symmetric (bell-shaped) distributions

$$s = 1.25 d_m$$

Coefficient of Variation

- Measures relative variation
- Always in percentage (%)
- Shows **variation relative to mean**
- Expresses standard deviation as fraction of percentage of mean
- Is used to compare two or more sets of data measured in different units

$$CV = \left(\frac{S}{\overline{X}} \right) 100\%$$

Comparing Coefficient of Variation

- Stock A:
  - Average price last year = $50
  - Standard deviation = $5
- Stock B:
  - Average price last year = $100
  - Standard deviation = $5

- Coefficient of variation:
  - Stock A: $CV = \left(\frac{S}{\overline{X}} \right) 100\% = \left(\frac{\$5}{\$50} \right) 100\% = 10\%$
  - Stock B: $CV = \left(\frac{S}{\overline{X}} \right) 100\% = \left(\frac{\$5}{\$100} \right) 100\% = 5\%$

Calculate Statistical Values for Drilling

- Determine range, standard deviation, for bit record

Bit	Ft (X)	$X-\bar{X}$	$(X-\bar{X})^2$	X^2	$\lvert X-\bar{X}\rvert$
1	53	(47.65)	2,270.52	2,809	47.65
2	69	(31.65)	1,001.72	4,761	31.65
3	72	(28.65)	820.82	5,184	28.65
4	76	(24.65)	607.62	5,776	24.65
5	80	(20.65)	426.42	6,400	20.65
6	89	(11.65)	135.72	7,921	11.65
...	...	...	...	...	...
20	139	38.35	1,470.72	19,321	38.35
n = 20	2,013	0	9,826.55	212,435	362.4

Calculate Statistical Values for Drilling

- Determine range, standard deviation for bit record

Bit	Ft (X)	$X-\bar{X}$	$(X-\bar{X})^2$	X^2	$\lvert X-\bar{X}\rvert$
n = 20	2,013	0	9,826.55	212,435	362.4

  - Range

$$R = X_{max} - X_{min} = 139 - 53 = 86$$

Calculate Statistical Values for Drilling

- Determine range, standard deviation for bit record

Bit	Ft (X)	$X-\overline{X}$	$(X-\overline{X})^2$	X^2	$\lvert X-\overline{X}\rvert$
n = 20	2,013	0	9,826.55	212,435	362.4

- Standard deviation

$$s=\left[\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{n-1}\right]^{1/2}=\left[\frac{9{,}826.55}{20-1}\right]^{1/2}=22.7417 \quad \text{ft}$$

Calculate Statistical Values for Drilling

- Determine range, standard deviation for bit record

Bit	Ft (			X^2	$\lvert X-\overline{X}\rvert$
n = 20	2,01			12,435	362.4

Arithmetic mean

$\overline{X}$ = 100.65 ft

- Standard deviation (alternate method)

$$s=\left[\frac{\sum_{i=1}^{n}X_i^{\,2}}{n-1}-\frac{n}{n-1}\overline{X}^2\right]^{1/2}=\left[\frac{212{,}435}{20-1}-\frac{20}{20-1}\times 100.65^2\right]^{1/2}$$

$$=\left[11{,}180.79-1.053\times 10{,}13042\right]^{1/2}=22.66 \quad \text{ft}$$

Calculate Statistical Values for Drilling

- Determine variance, mean absolute deviation for bit record

Bit	Ft (X)	$X-\bar{X}$	$(X-\bar{X})^2$	X^2	$\lvert X-\bar{X}\rvert$
n = 20	2,013	0	9,826.55	212,435	362.4

  - Variance (square of standard deviation)

$$s^2 = 22.7417^2 = 517.19$$

Calculate Statistical Values for Drilling

- Determine variance, mean absolute deviation for bit record

Bit	Ft (X)	$X-\bar{X}$	$(X-\bar{X})^2$	X^2	$\lvert X-\bar{X}\rvert$
n = 20	2,013	0	9,826.55	212,435	362.4

  - Mean absolute deviation

$$d_m = \frac{\sum_{i=1}^{n} \left| X_i - \bar{X} \right|}{n} = \frac{362.40}{20} = 18.12$$

Calculate Footage Drilled

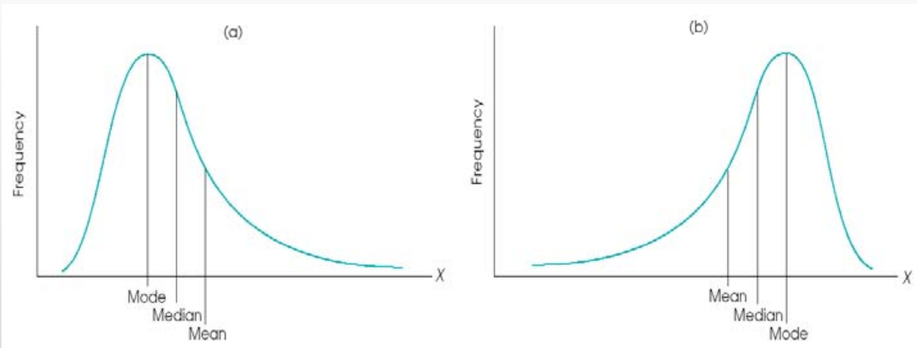
- Determine coefficient of variation for bit record

Bit	Ft	Arithmetic mean	X^2	$\lvert X-\overline{X}\rvert$
n = 20	2,(	$\overline{X}$ = 100.65 ft	212,435	362.4

- Coefficient of variation

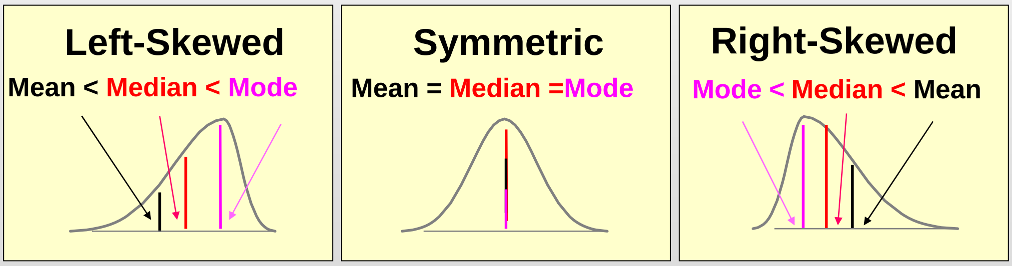
$$v = \frac{s}{\overline{X}} \times 100 = \frac{22.74}{100.65} \times 100 = 22.59\%$$

Central Tendencies and Distribution Shape



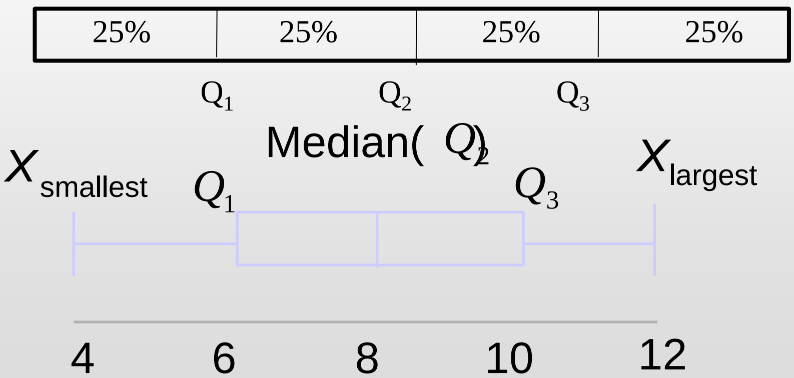
Shape of a Distribution

- Describes how data is distributed
- Measures of shape
  - Symmetric or skewed

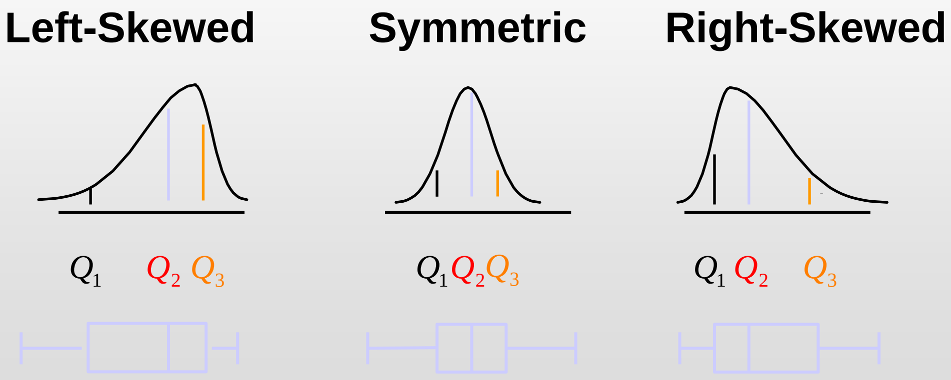


Exploratory Data Analysis

- Box-and-whisker plot
  - Graphical display of data using 5-number summary
  - A plot that shows the center, spread and skewness of data set



Distribution Shape and Box-and-Whisker Plot



Coefficient of Correlation

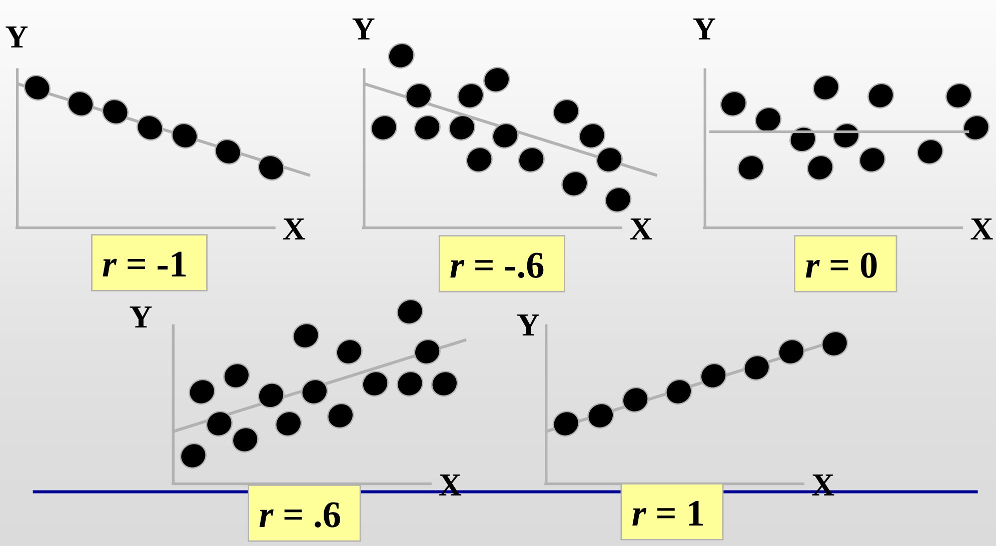
- Measures the strength of the linear relationship between **two quantitative variables**

$$r = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}$$

Features of Correlation Coefficient

- Unit free
- 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

Scatter Plots of Data with Various Correlation Coefficients



Pitfalls in Numerical Descriptive Measures

- Data analysis is objective
  - Should report the summary measures that best meet the assumptions about the data set
- Data interpretation is subjective
  - Should be done in fair, neutral and clear manner

Ethical Considerations

Numerical descriptive measures:

- Should document both good and bad results
- Should be presented in a fair, objective and neutral manner
- Should not use inappropriate summary measures to distort facts