Chapter 2 Describing the Distribution of a Variable

Chapter 2: **Statistical Concepts**

- **Distribution of a single variable:**
  - Basic Concepts (population and sample, data sets, variables, and observations, types of data)
  - Descriptive measures for categorical variables
  - Descriptive measures for numerical variables
  - Outliers and Missing values
- **Finding relationships among variables:**
  - Categorical variables
  - Numerical variables
  - Categorical variables and a Numerical variable
- **Sampling and distributions:**
  - Terminology
  - Estimation
  - Confidence Interval estimation
  - Sampling distributions
  - Confidence interval,
  - Hypothesis testing, Chi-square test for independence

Basic Concepts

- Several important concepts
  - Populations and samples
  - Data sets, Variables and observations
  - Types of data

Populations and Samples

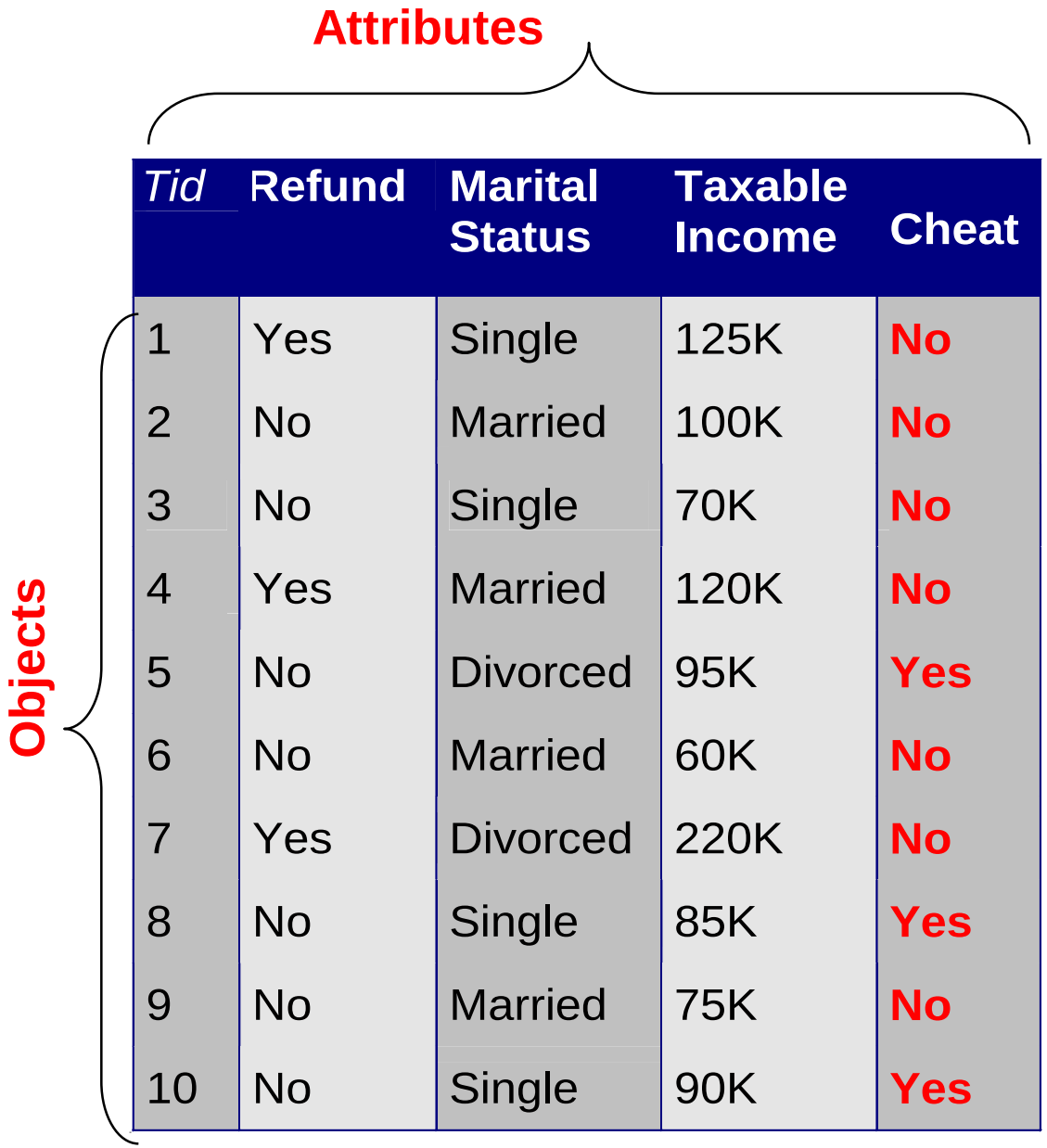
- A **population** includes all of the entities of interest in a study (people, households, machines, etc.).
  - Examples
    - All potential voters in a presidential election
    - All subscribers to cable television
    - All invoices submitted for Medicare reimbursement by nursing homes
- A **sample** is a subset of the population, often randomly chosen and preferably representative of the population as a whole.

Data Sets, Variables, and Observations

- A **data set** is usually a rectangular array of data, with variables in columns and observations in rows.
- A **variable** (or field or attribute) is a characteristic of members of a population, such as height, gender, or salary.
- An **observation** (or case or record) is a list of all variable values for a single member of a population.

What is Data Set?

- Collection of ***data objects*** and their ***attributes***
- An ***attribute*** is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, dimension, or feature
- A collection of attributes describe an ***object***
  - Object is also known as record, point, case, sample, entity, or instance



Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Important Characteristics of Structured Data

- **Dimensionality**
  - It is the number of attributes that the objects in the dataset posses.
  - Data with small number of dimensions tend to be qualitatively different than moderate or high-dimensional data
  - The difficulties lies in the high dimensionality data are sometimes referred to as the "Curse of dimensionality". And due to this dimensionality reduction is required.
- **Sparsity**
  - In database, sparsity and density describe the number of cells in a table that are empty or zero(sparsity) and that contain information(density).
  - For some data sets, such as those with symmetric features, most attibutes of an object have value 0, in many cases fewer than 1% of the entries are non-zero

Important Characteristics of Structured Data

- **Resolution**
  - It is frequently possible to obtain data at difeerent levels of resolution and often the properties of the data are different at different resoultion
  - For instance, the surface of Earth seems very uneven at a resoultion of few meters, but is relatively smooth at a resolution of 10KMs.
  - The patterns in the data also depend on the level of resolution
    - If the resolution is too fine, a pattern may not be visible or may be buried in noise
    - If the resoulution is too coarse, the pattern may disappear.

  For eg. variations in atmosheric pressure on a scale of hours reflect the movement of storms and other weather systems, whereas on a scale of months such phenomenon are not detectable

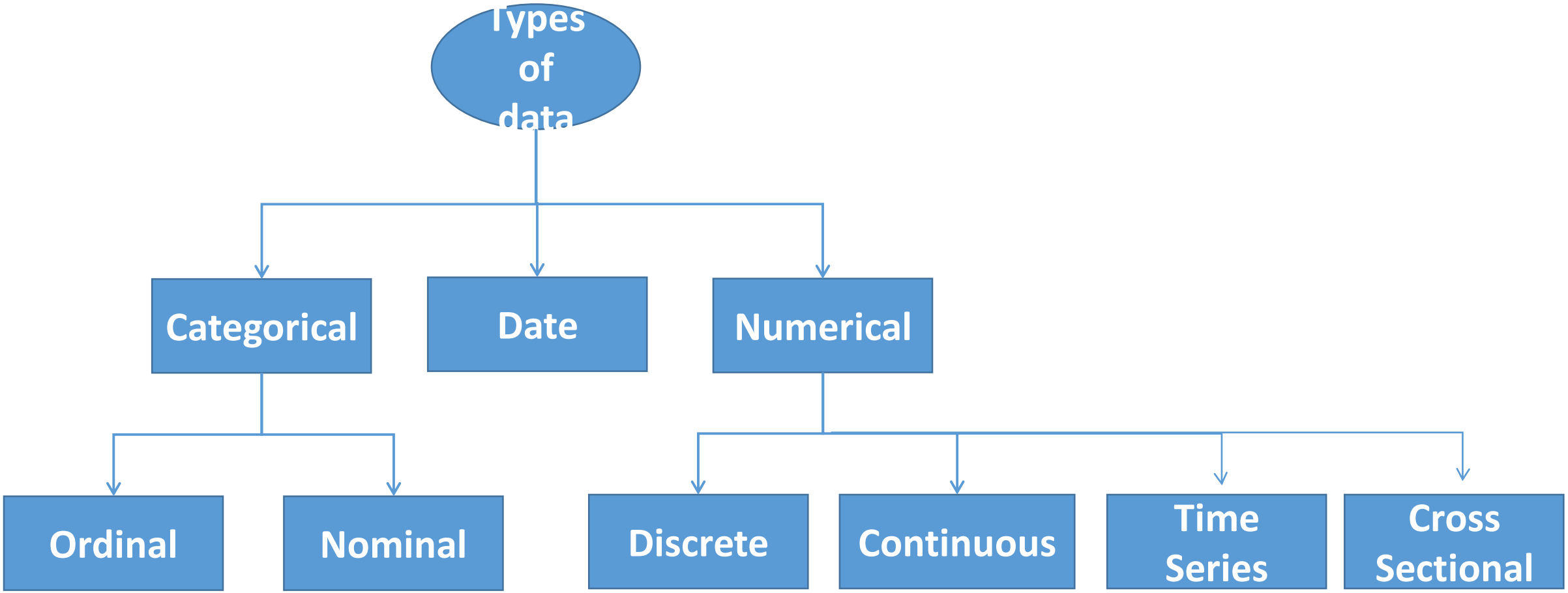
Resolution

Data from an Environmental Survey

- **Objective:** To illustrate variables and observations in a data set.
- **Solution:** Data set includes observations on 30 people who responded to a questionnaire on the president's environmental policies.
- Variables include age, gender, state, children, salary, and opinion.
  - Include a row that lists variable names.
  - Include a column that shows an index of the observation.

	A	B	C	D	E	F	G
1	**Person**	**Age**	**Gender**	**State**	**Children**	**Salary**	**Opinion**
2	1	35	Male	Minnesota	1	$65,400	5
3	2	61	Female	Texas	2	$62,000	1
4	3	35	Male	Ohio	0	$63,200	3
5	4	37	Male	Florida	2	$52,000	5
6	5	32	Female	California	3	$81,400	1
7	6	33	Female	New York	3	$46,300	5
8	7	65	Female	Minnesota	2	$49,600	1
9	8	45	Male	New York	1	$45,900	5
10	9	40	Male	Texas	3	$47,700	4
11	10	32	Female	Texas	1	$59,900	4

Types of data



Types of Data

- A variable is **numeric** if meaningful arithmetic operation can be performed on it.
- Otherwise, the variable is **categorical**.
- There is also a third **data type**, a **date** variable.
  - Excel® stores dates as numbers, but dates are treated differently from typical numbers.
- A categorical variable is **ordinal** if there is a natural ordering of its possible values.
- If there is no natural ordering, it is **nominal**.

Data Types (Categorical)

- Categorical variables can be coded numerically.
- A **dummy variable** is a 0–1 coded variable for a specific category.
  - It is coded as 1 for all observations in that category and 0 for all observations not in that category.
- A **binned** (or **discretized**) **variable** corresponds to a numerical variable that has been categorized into discrete categories.
- These categories are usually called **bins**.

Data Types

	A	B	C	D	E	F	G	H	I	J	K	L
1	**Person**	**Age**	**Gender**	**State**	**Children**	**Salary**	**Opinion**					
2	1	Middle-aged	1	Minnesota	1	$65,400	Strongly agree					
3	2	Elderly	0	Texas	2	$62,000	Strongly disagree					
4	3	Middle-aged	1	Ohio	0	$63,200	Neutral					
5	4	Middle-aged	1	Florida	2	$52,000	Strongly agree					
6	5	Young	0	California	3	$81,400	Strongly disagree					
7	6	Young	0	New York	3	$46,300	Strongly agree					
8	7	Elderly	0	Minnesota	2	$49,600	Strongly disagree					
9	8	Middle-aged	1	New York	1	$45,900	Strongly agree		Age lookup table (range name AgeLookup)			
10	9	Middle-aged	1	Texas	3	$47,700	Agree		0	Young		
11	10	Young	0	Texas	1	$59,900	Agree		35	Middle-aged		
12	11	Middle-aged	1	New York	1	$48,100	Agree		60	Elderly		
13	12	Middle-aged	0	Virginia	0	$58,100	Neutral					
14	13	Middle-aged	0	Illinois	2	$56,000	Strongly disagree		Opinion lookup table (range name OpinionLookup)			
15	14	Middle-aged	0	Virginia	2	$53,400	Strongly disagree		1	Strongly disagree		
16	15	Middle-aged	0	New York	2	$39,000	Disagree		2	Disagree		
17	16	Middle-aged	1	Michigan	1	$61,500	Disagree		3	Neutral		
18	17	Middle-aged	1	Ohio	0	$37,700	Strongly disagree		4	Agree		
19	18	Middle-aged	0	Michigan	2	$36,700	Agree		5	Strongly agree		
28	27	Young	1	Illinois	3	$45,400	Disagree					
29	28	Elderly	1	Michigan	2	$53,900	Strongly disagree					
30	29	Middle-aged	1	California	1	$44,100	Neutral					
31	30	Middle-aged	0	New York	2	$31,000	Agree					

Note the formulas in columns B, C, and G that generate this recoded data. The formulas in columns B and G are based on the lookup tables below.

Data Types (Numeric)

- A numerical variable is **discrete** if it results from a count, such as the number of children.
- A **continuous** variable is the result of an essentially continuous measurement, such as weight or height.
- **Cross-sectional** data are data on a cross-section of a population at a distinct point in time.
- **Time series** data are data collected over time.

Discrete Vs Continuous Variable

- Continuous variables would take forever to count. In fact, we would get to forever and never finish counting them. For example, take an age. We can't count "age". Because it would literally take forever. For example, it could be 37 years, 9 months, 6 days, 5 hours, 4 seconds, 5 milliseconds, 6 nanoseconds, 77 picoseconds…and so on.
- **[A person's age in years. A baby's age in months, Time, Time in wrist watch,** Temperature in Arizona**]**

Discrete Variable	**Continuous Variable**
It is a variable whose value is obtained by counting.	It is a variable whose value is obtained by measuring.
Examples: Number of planets around the Sun Number of students in a class	Examples: Number of stars in the space Height or weight of the students in a particular class
Range of specified numbers is complete.	Range of specified numbers is incomplete, i.e. infinite.
It assumes a distinct or a separate value.	It assumes any value between two values.

Cross-sectional Data Vs Time series

In cross sectional data, there are several variables at the same point in time. Data set with maximum temperature, humidity, wind speed of few cities on a single day is an example of a cross sectional data.

City	Maximum Temperature	Humidity	Wind Speed
City A	29	60%	20mph
City B	27	65%	26mph
City C	30	60%	21mph

Another example is the sales revenue, sales volume, number of customers and expenses of an organization in the past month. Cross sectional data takes the form of Xi. Expanding the data from several months will convert the cross sectional data to time series data.

Cross-sectional vs Time series

Time series data are data collected over a period of time.

	A	B	C
1	**Quarter**	**Revenue**	
2	Q1-2015	$1,026,000	
3	Q2-2015	$1,056,000	
4	Q3-2015	$1,182,000	
5	Q4-2015	$2,861,000	
6	Q1-2016	$1,172,000	
7	Q2-2016	$1,249,000	
8	Q3-2016	$1,346,000	
9	Q4-2016	$3,402,000	
10	Q1-2017	$1,286,000	
11	Q2-2017	$1,317,000	
12	Q3-2017	$1,449,000	
13	Q4-2017	$3,893,000	
14	Q1-2018	$1,462,000	
15	Q2-2018	$1,452,000	
16	Q3-2018	$1,631,000	
17	Q4-2018	$4,200,000	

Categorical Data (Qualitative Attribute Types)

- **Nominal:** Nominal means "relating to names." The values of a nominal attribute are symbols or names of things for example,
  - *Hair_color* = {*auburn*, *black*, *blond*, *brown*, *grey*, *red*, *white*}
  - marital status, occupation, ID numbers, zip codes

  The values do not have any meaningful order about them.

- **Binary:** Nominal attribute with only 2 states (0 and 1), where 0 typically means that the attribute is absent, and 1 means that it is present. Binary attributes are referred to as Boolean if the two states correspond to true and false.
  - ***Symmetric binary:*** both outcomes equally important
    - e.g., gender
  - ***Asymmetric binary:*** outcomes not equally important.
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)

Chapter 2: **Statistical Concepts**

- **Distribution of a single variable:**
  - Basic Concepts (population and sample, data sets, variables, and observations, types of data)
  - Descriptive measures for categorical variables
  - Descriptive measures for numerical variables
  - Outliers and Missing values
- **Finding relationships among variables:**
  - Categorical variables
  - Numerical variables
  - Categorical variables and a Numerical variable
- **Sampling and distributions:**
  - Terminology
  - Estimation
  - Confidence Interval estimation
  - Sampling distributions
  - Confidence interval,
  - Hypothesis testing, Chi-square test for independence

Data Analysis: Introduction...

- In the field of data, there is nothing more important than understanding the data that needs to be analyzed. In order to understand the data, it is important to understand the purpose of the analysis because this will help to save time and dictate how to go about analyzing the data.
- **Exploratory data analysis (EDA)** refers to the critical process of performing initial investigations on data to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. It can be classified as **univariate**, **bivariate**, and **multivariate** analysis.
- So, the goal is to present data in a form that makes sense to people. Tools that are used to do this include:
  - **Graphs:** bar charts, pie charts, histograms, scatter charts, and time series graphs.
  - **Numerical summary measures:** counts, percentages, averages, and measures of variability
  - **Tables of summary measures:** totals, averages, and counts, grouped by categories

Data Analysis: Introduction...

Univariate data and its analysis:

- Data consists of only **one variable** like *Height, age, income.*
- The analysis of univariate data is the simplest form of analysis since the information deals with only one quantity that changes.
- It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it.
- The description of patterns in this type of data can be made by:
  - Central tendency measures (mean, median and mode),
  - Dispersion or spread of data (range, minimum, maximum, quartiles, variance and standard deviation)
  - Frequency distribution tables: Box Plot, histograms, pie charts, frequency polygon and bar charts.

Data Analysis: Introduction...

Bivariate data and its analysis

- This type of data involves two different variables.
- The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship among the two variables: independent and dependent.
- Example of bivariate data can be temperature and ice cream sales in summer season.

Temp (in Celsius)	Ice cream sales
20	2000
25	2500
35	5000

- The relationship is visible from the table that temperature and sales i.e. they are directly proportional to each other
- Thus bivariate data analysis involves comparisons, relationships, causes and explanations.
- These variables are often plotted on X and Y axis on the graph for better understanding

Bivariate data and its analysis...

Types of bivariant data analysis:

- **Numerical and Numerical:** When both variables (independent and dependent) are numerical values. The technique used: Correlation Coefficient analysis
- **Categorical and categorical:** When both variables are categorical.
  Technique used : Chi-Square
- **Categorical and numerical:**
  - When categorical variable is independent (input) and numeric variable is dependent (output):
    - In categorical variable if the **category is two** (e.g. Good & Bad): T-Test
    - In categorical variable if the **category is > two** (e.g. Good, Bad, Worst):- ANOVA
  - When numeric variable is independent (input) and categorical variable is dependent (output): Logistic Regression

	Categorical	Numerical
Categorical	**Chi-Square**	**T-test ANOVA**
Numerical	**Point biserial Correlation & Logistic Regression**	**Correlation**

Data Analysis: Introduction...

Multivariate data and its analysis

- When the data involves three or more variables, it is categorized under multivariate.
- Example of this type of data is suppose an advertiser wants to compare the popularity of four advertisements on a website, then their click rates could be measured for both men and women and relationships between variables can then be examined.
- The ways to perform analysis on this data depends on the goals to be achieved.
- Some of the techniques are **regression analysis**, **path analysis**, **factor analysis** and **multivariate analysis of variance (MANOVA)**.

Data Analysis: Introduction...

Univariate, Bivariate and Multivariate analysis

Univariate	Bivariate	Multivariate
It summarize single variable at a time.	It summarize two variables	It summarize more than 2 variables.
It does not deal with causes and relationships.	It deal with causes and relationships.	It deal with causes and relationships.
It does not contain any dependent variable.	It does contain only one dependent variable.	It is similar to bivariate but it contains more than 2 dependent variables.
The main purpose is to describe.	The main purpose is to explain.	The main purpose is to study the relationship among them.

Descriptive Measures for Categorical Variables

- There are only a few possibilities for describing a categorical variable, all based on counting:
  - Count the number of categories.
  - Give the categories names.
  - Count the number of observations in each category. (The resulting counts can be reported as “raw counts” or as percentages of totals.)
    - Once you have the counts, you can display them graphically, usually in a column chart or a pie chart.

Descriptive Measures for Categorical Variables

Supermarket Data Example

- **Objective**: To summarize categorical variables in a large data set.
- **Solution**: Data set contains transactions made by supermarket customers over a two-year period.
- Children, Units Sold, and Revenue are numerical.
- Purchase Date is a date variable.
- Transaction and Customer ID are used only to identify.
- All of the other variables are categorical.

	A	B	C	D	E	F	G	H	I	J	K	O	P
1	Transaction	Purchase Date	Customer ID	Gender	Marital Status	Homeowner	Children	Annual Income	City	State or Province	Country	Units Sold	Revenue
2	1	12/18/2016	7223	F	S	Y	2	$30K - $50K	Los Angeles	CA	USA	5	$27.38
3	2	12/20/2016	7841	M	M	Y	5	$70K - $90K	Los Angeles	CA	USA	5	$14.90
4	3	12/21/2016	8374	F	M	N	2	$50K - $70K	Bremerton	WA	USA	3	$5.52
5	4	12/21/2016	9619	M	M	Y	3	$30K - $50K	Portland	OR	USA	4	$4.44
6	5	12/22/2016	1900	F	S	Y	3	$130K - $150K	Beverly Hills	CA	USA	4	$14.00
7	6	12/22/2016	6696	F	M	Y	3	$10K - $30K	Beverly Hills	CA	USA	3	$4.37
8	7	12/23/2016	9673	M	S	Y	2	$30K - $50K	Salem	OR	USA	4	$13.78
9	8	12/25/2016	354	F	M	Y	2	$150K +	Yakima	WA	USA	6	$7.34
10	9	12/25/2016	1293	M	M	Y	3	$10K - $30K	Bellingham	WA	USA	1	$2.41
11	10	12/25/2016	7938	M	S	N	1	$50K - $70K	San Diego	CA	USA	2	$8.96

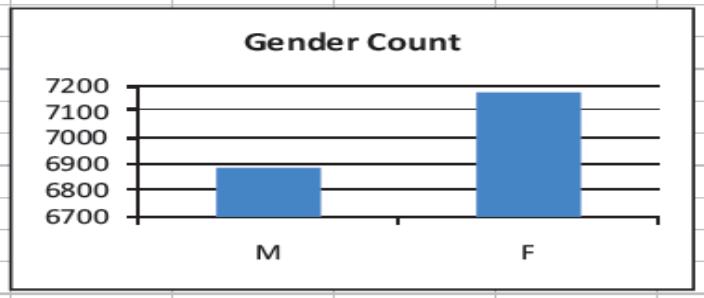
Descriptive Measures for Categorical Variables

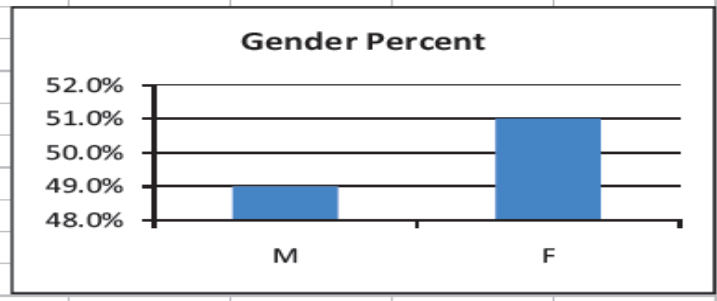
Supermarket Sales Example

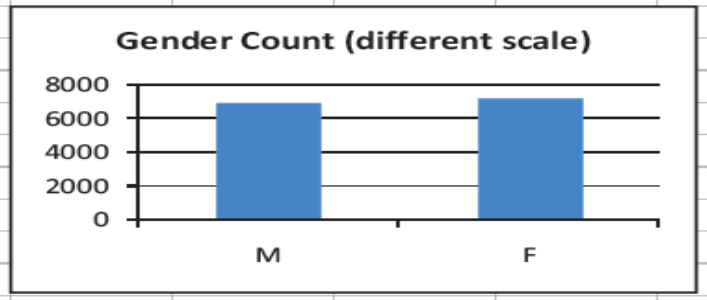
To get the counts in column S, use the Excel® function, COUNTIF.

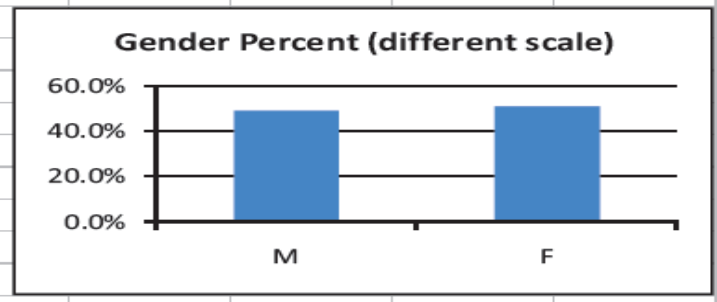
- To get the percentages in column T, divide each count by the total number of observations.
- Keep charts simple so that the information they contain emerges as clearly as possible.

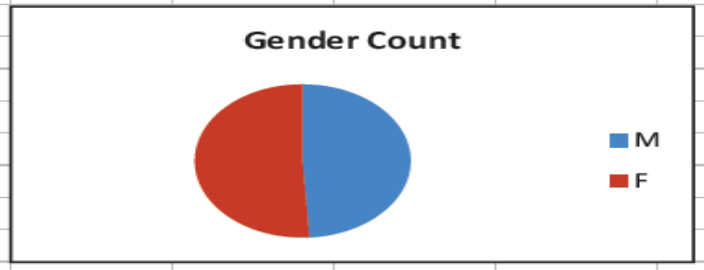
	R	S	T
1	Categorical summaries		
2	Gender	Count	Percent
3	M	6889	49.0%
4	F	7170	51.0%
5			100.0%
6			
7	Marital Status	Count	Percent
8	S	7193	51.2%
9	M	6866	48.8%
10			100.0%
11			
12	Homeowner	Count	Percent
13	Y	8444	60.1%
14	N	5615	39.9%
15			100.0%
16			
17	Annual Income	Count	Percent
18	$10K - $30K	3090	22.0%
19	$30K - $50K	4601	32.7%
20	$50K - $70K	2370	16.9%
21	$70K - $90K	1709	12.2%
22	$90K - $110K	613	4.4%
23	$110K - $130K	643	4.6%
24	$130K - $150K	760	5.4%
25	$150K +	273	1.9%
26			100.0%

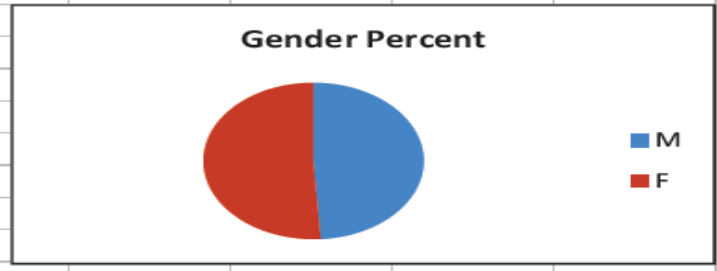












Descriptive Measures for Categorical Variables...

Supermarket Sales Example

Another efficient way to find counts for a categorical variable is to use dummy (0–1) variables.

- Recode each variable so that one category is replaced by 1 and all others by 0.
  - This can be done using a simple IF formula.
- Find the count of that category by summing the 0s and 1s.
- Find the percentage of that category by averaging the 0s and 1s.

	A	B	C	D	E
1	Transaction	Purchase Date	Customer ID	Gender	Gender Dummy for M
2	1	12/18/2016	7223	F	0
3	2	12/20/2016	7841	M	1
4	3	12/21/2016	8374	F	0
5	4	12/21/2016	9619	M	1
6	5	12/22/2016	1900	F	0
7	6	12/22/2016	6696	F	0
14058	14057	12/31/2018	250	M	1
14059	14058	12/31/2018	6153	F	0
14060	14059	12/31/2018	3656	M	1
14061				Count	6889
14062				Percent	49.0%

Descriptive Measures for Numerical Variables

- There are many ways to summarize numerical variables, both with numerical summary measures and with charts.

Let a numerical variable such as Salary, where there is one observation for each person. Our basic goal is to learn how these salaries are distributed across people by asking:

1. What are the most “typical” salaries?
2. How spread out are the salaries?
3. What are the “extreme” salaries on either end?
4. Is a chart of the salaries symmetric about some middle value, or is it skewed in one direction?
5. Does the chart of salaries have any other peculiar features besides possible skewness?

Descriptive Measures for Numeric Variable...

- Measures of Central Tendency [**Mean, Median, Mode**]
- Measures of data Disperssion/Measures of Variability [**Minimum, Maximum, Percentiles, and Quartiles, Five Number Summary**]
- Empirical Rules for Interpreting Standard Deviation
- Measures of Shape

Descriptive Measures for Numeric Variable...

Measuring the Central Tendency

1. **Mean** (algebraic measure) (sample vs. population):

- The most common and effective numeric measure of the "center" of a set of data is the ***(arithmetic) mean***. Let $x_1, x_2, \ldots\ldots, x_N$ be a set of ***N*** values or ***observations***, such as for some numeric attribute *X*, like salary.

The **mean** of this set of values is

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

Note: *N* is population size.

- Sometimes, each value x_i in a set may be associated with a **weight w_i** for $i = 1, \ldots, N$. The weights reflect the significance, importance, or occurrence frequency attached to their respective values. In this case, we can compute

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_N x_N}{w_1 + w_2 + \cdots + w_N}$$

This is called the ***weighted arithmetic mean*** *or the* ***weighted average.***

Descriptive Measures for Numeric Variable...

Measuring the Central Tendency

- A **trimmed mean** (sometimes called a ***truncated mean***) is similar to a mean, but it trims any outliers. Outliers can affect the mean (especially if there are just one or two very large values), so a trimmed mean can often be a better fit for data sets with erratic high or low values or for extremely skewed distributions. Even a small number of extreme values can corrupt the mean.
- For example, the mean salary at a company may be substantially pushed up by that of a few highly paid managers. Similarly, the mean score of a class in an exam could be pulled down quite a bit by a few very low scores.
- Which is the mean obtained after chopping off values at the high and low extremes.
- Example: Find the trimmed 20% mean for the following test scores: 60, 81, 83, 91, 99.
  - Step 1: Trim the top and bottom 20% from the data. That leaves us with the middle three values: 60, 81, 83, 91, 99.
  - Step 2: Find the mean with the remaining values. The mean is (81 + 83 + 91) / 3) = 85.

Descriptive Measures for Numeric Variable...

Measuring the Central Tendency

2. Median

- The median of a set of data is the middlemost number in the set. The median is also the number that is halfway into the set.
- To find the median, the data should first be arranged in order from least to greatest.
- Middle value if odd number of values, or average of the middle two values otherwise

- **What will be the median estimated by interpolation (for *grouped data*)?**

Descriptive Measures for Numeric Variable...

Measuring the Central Tendency

3. Mode

The mode for a set of data is the value that occurs most frequently in the set. Therefore, it can be determined for qualitative and quantitative attributes.

- It is possible for the greatest frequency to correspond to several different values, which results in more than one mode.
- Data sets with one, two, or three modes are respectively called **unimodal**, **bimodal,** and **trimodal**.
- In general, a data set with two or more modes is **multimodal**.
- At the other extreme, if each data value occurs only once, then there is no mode.
- Example:
  - 53, 56, 56, 58, 58, 59, 59, 60, 61, 61, **62, 62, 62**, 64, 65, 65, 67, 68, 68, 70
  - 62 appears three times, more often than the other values, so **Mode = 62**

Descriptive Measures for Numeric Variable...

Mean, Median and Mode from Grouped Frequencies

The Race.....This starts with some raw data (not a grouped frequency yet)..

Alex timed 21 people in the sprint race, to the nearest second:

59, 65, 61, 62, 53, 55, 60, 70, 64, 56, 58, 58, 62, 62, 68, 65, 56, 59, 68, 61, 67

To find the Mean Alex adds up all the numbers, then divides by how many numbers:

Mean = (59+65+61+62+53+55+60+70+64+56+58+58+62+62+68+65+56 +59+68+ 61+ 67) /21

= 61.38095...

Descriptive Measures for Numeric Variable...

Mean, Median and Mode from Grouped Frequencies

- To find the **Median** Alex places the numbers in value order and finds the middle number.
- In this case the median is the 11th number:

53, 55, 56, 56, 58, 58, 59, 59, 60, 61, **61**, 62, 62, 62, 64, 65, 65, 67, 68, 68, 70

- **Median** = 61

- To find the **Mode**, or modal value, Alex places the numbers in value order then counts how many of each number. The Mode is the number which appears most often (there can be more than one mode):

53, 55, 56, 56, 58, 58, 59, 59, 60, 61, 61, **62, 62, 62**, 64, 65, 65, 67, 68, 68, 70

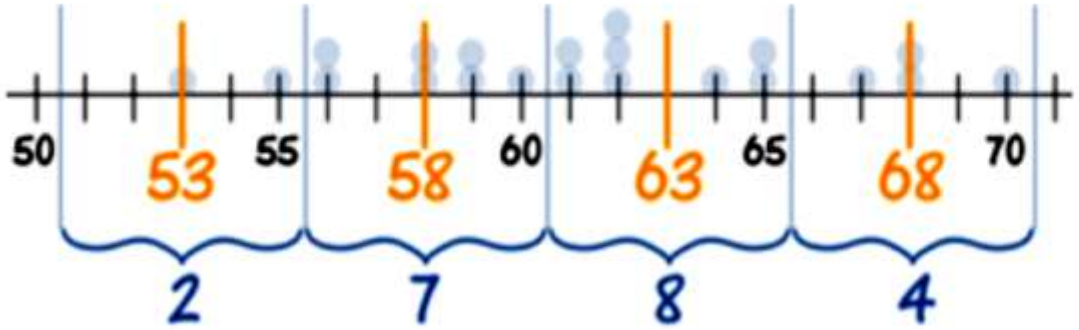
- 62 appears three times, more often than the other values, so **Mode** = 62

Descriptive Measures for Numeric Variable...

Estimating the Mean from Grouped Data

The groups (51-55, 56-60, etc), also called **class intervals**, are of **width** 5

The **midpoints** are in the middle of each class: 53, 58, 63 and 68



Seconds	Frequency
51 - 55	2
56 - 60	7
61 - 65	8
66 - 70	4

We can estimate the Mean by using the ***midpoints.***

Let's now make the table using midpoints:

Midpoint	Frequency
53	2
58	7
63	8
68	4

Descriptive Measures for Numeric Variable...

Estimating the Mean from Grouped Data

- Our thinking is: "2 people took 53 sec, 7 people took 58 sec, 8 people took 63 sec and 4 took 68 sec". In other words we imagine the data looks like this:

53, 53, 58, 58, 58, 58, 58, 58, 58, 63, 63, 63, 63, 63, 63, 63, 63, 68, 68, 68, 68

Midpoint	Frequency
53	2
58	7
63	8
68	4

- Then we add them all up and divide by 21. The quick way to do it is to multiply each midpoint by each frequency:

Midpoint x	Frequency f	Midpoint × Frequency fx
53	2	106
58	7	406
63	8	504
68	4	272
Totals:	**21**	**1288**

And then our estimate of the mean time to complete the race is:

$$\text{Estimated Mean} = \frac{1288}{21} = 61.333\ldots$$

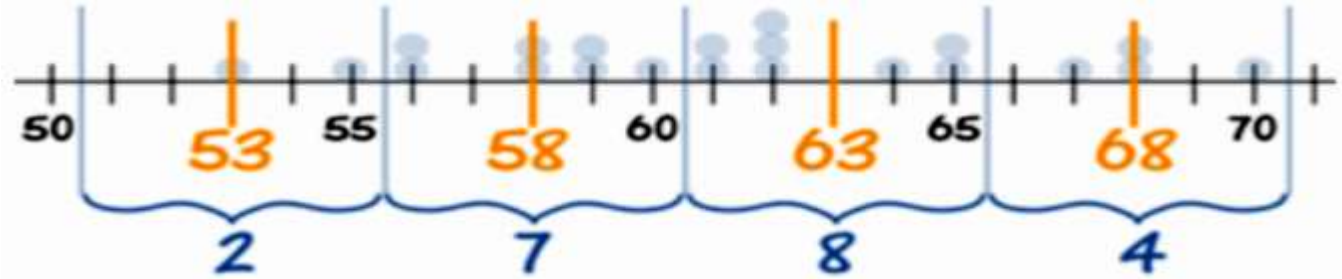
Descriptive Measures for Numeric Variable...

Estimating the Median from Grouped Data

Let's look at our data again:

Seconds	Frequency
51 - 55	2
56 - 60	7
61 - 65	8
66 - 70	4

- The median is the middle value, which in our case is the 11th one, which is in the 61 - 65 group:
- We can say "the **median group** is 61 - 65"
- But if we want an estimated **Median value** we need to look more closely at the 61 - 65 group.



We call it "61 - 65", but it really includes values from 60.5 up to (but not including) 65.5.

Why? Well, the values are in whole seconds, so a real time of 60.5 is measured as 61. Likewise 65.4 is measured as 65.

Descriptive Measures for Numeric Variable...

Estimating the Median from Grouped Data

Seconds	Frequency
51 - 55	2
56 - 60	7
61 - 65	8
66 - 70	4

- At 60.5 we already have 9 runners, and by the next boundary at 65.5 we have 17 runners.
- By drawing a straight line in between we can pick out where the median frequency of n/2 runners is:

And this handy formula does the calculation:

$$\text{Estimated Median} = L + \frac{(n/2) - B}{G} \times w$$

where:

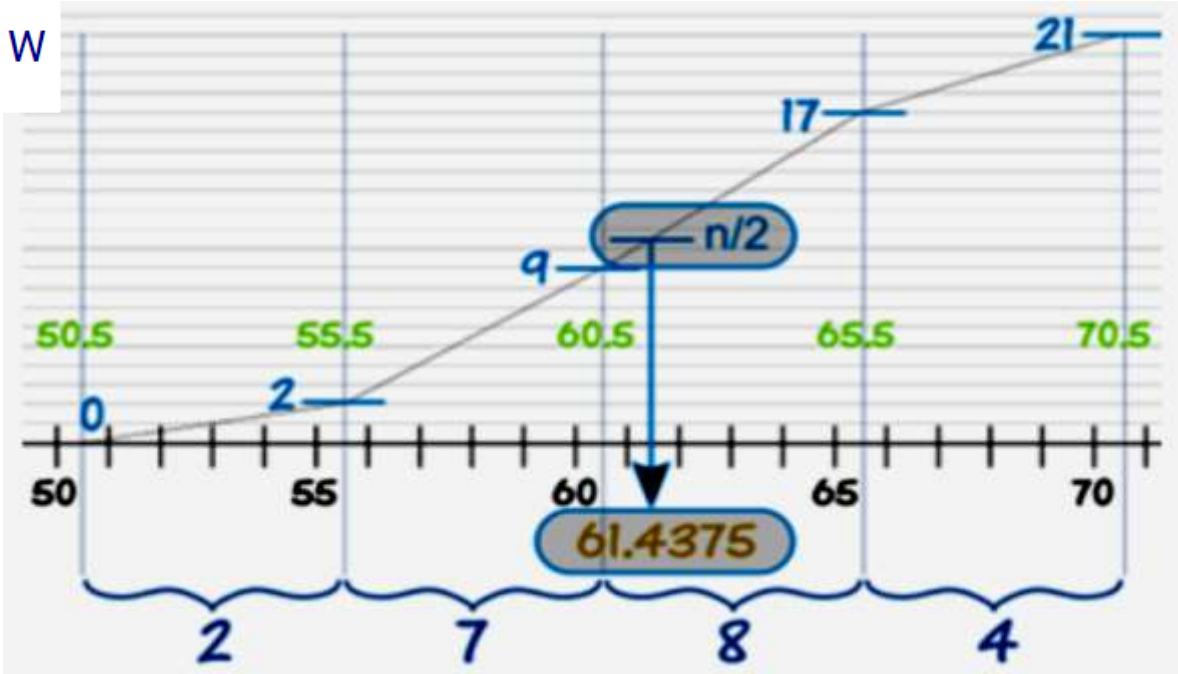
L is the lower class boundary of the group containing the median

n is the total number of values

B is the cumulative frequency of the groups before the median group

G is the frequency of the median group

w is the group width



Descriptive Measures for Numeric Variable...

For our example:

Seconds	Frequency
51 - 55	2
56 - 60	7
61 - 65	8
66 - 70	4

L = 60.5

n = 21

B = 2 + 7 = 9

G = 8

w = 5

$$median = L_1 + (\frac{n/2 - (\sum freq)l}{freq_{median}})width$$

Estimated Median $= 60.5 + \frac{(21/2) - 9}{8} \times 5$

$= 60.5 + 0.9375$

$= \mathbf{61.4375}$

age	*frequency*
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

Descriptive Measures for Numeric Variable...

Estimating the Mode from Grouped Data

- Again, looking at our data:
- We can easily find the **modal group** (***the group with the highest frequency***), which is 61 - 65
- We can say **"the modal group is 61 - 65"**

Seconds	Frequency
51 - 55	2
56 - 60	7
61 - 65	8
66 - 70	4

- But the actual Mode may not even be in that group! Or there may be more than one mode. Without the raw data we don't really know. But, we can estimate the Mode using the following formula:

$$\text{Estimated Mode} = L + \frac{f_m - f_{m-1}}{(f_m - f_{m-1}) + (f_m - f_{m+1})} \times w$$

where:

L is the lower class boundary of the modal group

$\mathbf{f_m}$ is the frequency of the modal group

$\mathbf{f_{m-1}}$ is the frequency of the group before the modal group

w is the group width

$\mathbf{f_{m+1}}$ is the frequency of the group after the modal group

Descriptive Measures for Numeric Variable...

For our example:

Seconds	Frequency
51 - 55	2
56 - 60	7
61 - 65	8
66 - 70	4

$L = 60.5$

$f_{m-1} = 7$

$f_m = 8$

$f_{m+1} = 4$

$w = 5$

$$\text{Estimated Mode} = 60.5 + \frac{8 - 7}{(8 - 7) + (8 - 4)} \times 5$$

$$= 60.5 + (1/5) \times 5$$

$$= \mathbf{61.5}$$

Our final result is:

- Estimated Mean: **61.333...**
- Estimated Median: **61.4375**
- Estimated Mode: **61.5**

(Compare that with the true Mean, Median and Mode of **61.38..., 61 and 62** that we got at the very start.)

Descriptive Measures for Numeric Variable...

Baby Carrots Example

Example: You grew fifty baby carrots using special soil. You dig them up and measure their lengths (to the nearest mm) and group the results:

Length (mm)	Frequency
150 - 154	5
155 - 159	2
160 - 164	6
165 - 169	8
170 - 174	9
175 - 179	11
180 - 184	6
185 - 189	3

Age Example

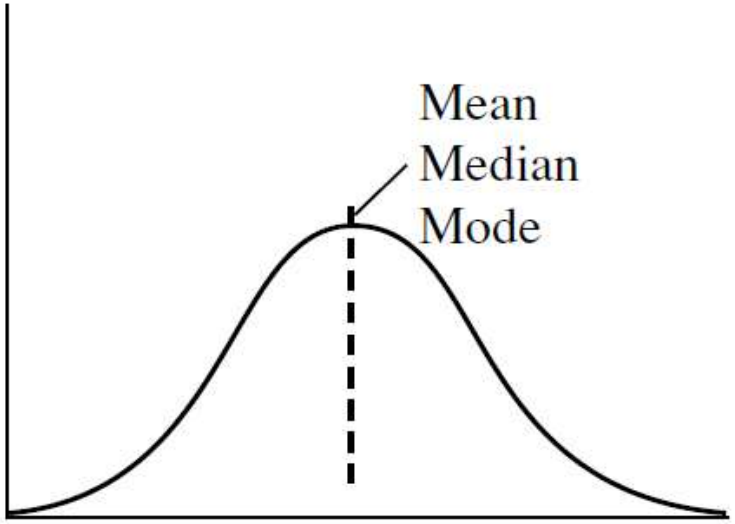
The ages of the 112 people who live on a tropical island are grouped as follows:

Age	Number
0 - 9	20
10 - 19	21
20 - 29	23
30 - 39	16
40 - 49	11
50 - 59	10
60 - 69	7
70 - 79	3
80 - 89	1

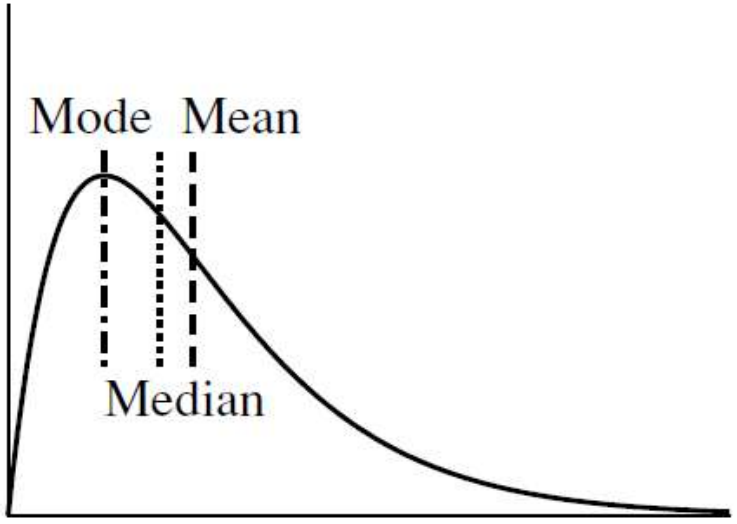
Descriptive Measures for Numeric Variable...

Symmetric vs. Skewed Data

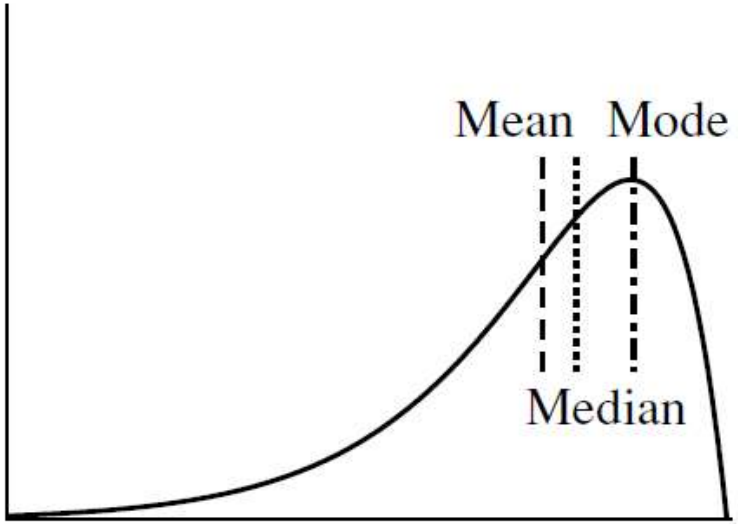
- Median, mean and mode of ***symmetric, positively*** *and* ***negatively skewed*** data
- Data in most real applications are not **symmetric (a)**. They may instead be either ***positively skewed (b)***, where the mode occurs at a value that is smaller than the median or ***negatively skewed (c)***, where the mode occurs at a value greater than the median.



(a) Symmetric data



(b) Positively skewed data



(c) Negatively skewed data

Descriptive Measures for Numeric Variable...

Measures of data Disperssion/Measures of Variability

- In addition to knowing where the center of distribution is, it is often helpful to know the degree to which individual values lies around teh center. This is known as **variability/disperssion**.
- There are a few ways to measure the variability:
  - **Variance**
  - **Standard Deviation**
  - **Quartiles, Interquartile Range, Five Number summary**

Descriptive Measures for Numeric Variable...

Measuring the Dispersion of Data (cont..)

Range

- Let $x_1, x_2, \ldots, x_N$ be a set of observations for some numeric attribute, X.
- The range of the set is the difference between the ***largest (max())*** and ***smallest (min())*** values.

Variance

variance is a measure of dispersion that takes into account the spread of all data points in a data set. iT is a measure of how data points differ from the mean.

The **sample variance** is denoted by s^2, and the **population variance** by σ^2.

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \text{mean})^2}{n-1} \qquad \sigma^2 = \frac{\sum_{i=1}^{n}(X_i - \text{mean})^2}{n}$$

Descriptive Measures for Numeric Variable...

Measuring the Dispersion of Data (cont..)

Variance....

If all observations are close to the mean, their squared deviations from the mean—and the variance—will be relatively small.

If at least a few of the observations are far from the mean, their squared deviations from the mean—and the variance—will be large.

- **Coefficient of variation:** A coefficient of variation is an effective metric for quickly evaluating the relative dispersion of the data points around a sample mean. The coefficient of variation (CV) is the SD divided by the mean.

$$CV = \frac{\sigma}{\mu}$$

σ = population standard deviation

μ = population mean

Descriptive Measures for Numeric Variable...

Measuring the Dispersion of Data (cont..)

Standard Deviation

- The standard deviation (usually abbreviated SD, sd, or just s) of a bunch of numbers tells you how much the individual numbers tend to differ (in either direction) from the mean. It's calculated as follows:

$$SD = sd = s = \sqrt{\frac{\sum_i (d_i)^2}{N-1}}, \text{ where } d_i = X_i - \overline{X}$$

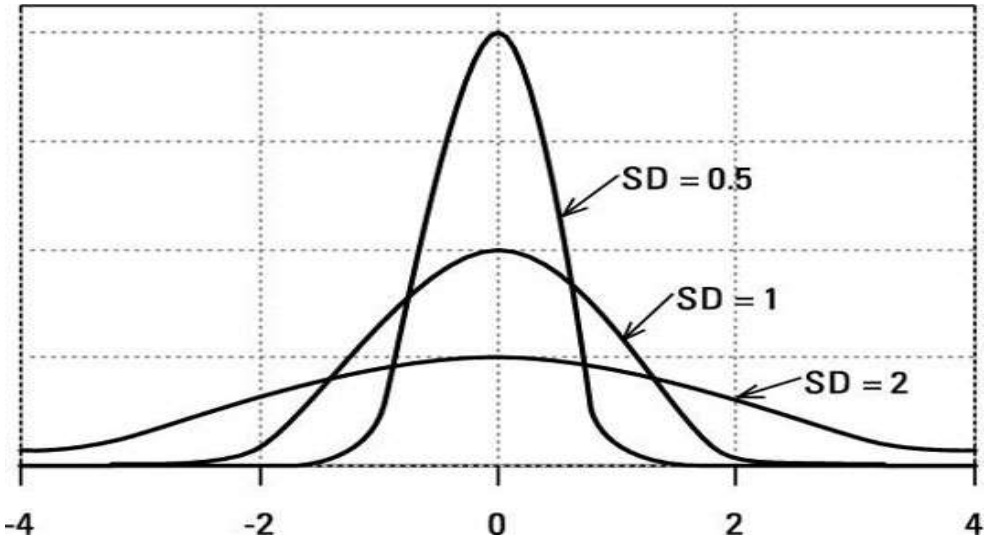
This formula is saying that you calculate the standard deviation of a set of N numbers (X_i) by subtracting the mean from each value to get the deviation (d_i) of each value from the mean, squaring each of these deviations, adding up the $(d_i)^2$ terms, dividing by $N-1$, and then taking the square root.

Descriptive Measures for Numeric Variable...

Measuring the Dispersion of Data (cont..)

Standard Deviation (cont.)

- This is almost identical to the formula for the root-mean-square deviation of the points from the mean, except that it has ***N – 1*** in the denominator instead of ***N***.
- This difference occurs because the sample mean is used as an approximation of the true population mean (which you don't know). If the true mean were available to use, the denominator would be ***N***.
- When talking about population distributions, the SD describes the width of the distribution curve. The figure shows three normal distributions. They all have a mean of zero, but they have different standard deviations and, therefore, different widths. Each distribution curve has a total area of exactly 1.0, so the peak height is smaller when the SD is larger.



For an IQ example (84, 84, 89, 91, 110, 114, and 116) where the mean is 98.3, you calculate the SD as follows:

$$SD = \sqrt{\frac{(84-98.3)^2 + (84-98.3)^2 + \ldots + (116-98.3)^2}{7-1}} = 14.4$$

Standard deviations are very sensitive to extreme values (outliers) in the data. For example, if the highest value in the IQ dataset had been 150 instead of 116, the SD would have gone up from 14.4 to 23.9.

Descriptive Measures for Numeric Variable...

Why n-1 in Standard Deviation?

Bessel's correction

Why divide by n-1 rather than n?

You compute the difference between each value and the mean of those values. You don't know the true mean of the population; all you know is the mean of your sample. Except for the rare cases where the sample mean happens to equal the population mean, the data will be closer to the sample mean than it will be to the true population mean.

So the value you compute will probably be a bit smaller (and can't be larger) than what it would be if you used the true population mean.

To make up for this, divide by n-1 rather than n. This is called **Bessel's correction.**

But why n-1? If you knew the sample mean, and all but one of the values, you could calculate what that last value must be. Statisticians say there are n-1 degrees of freedom.

Descriptive Measures for Numeric Variable...

Measuring the Dispersion of Data (cont..)

Quartiles

It divide an ordered data set into four equal parts.

The values which divide each part are called the first, second, and third ***quartiles***; they are denoted by Q1, Q2, and Q3, respectively.

Q1 is the middle value of the first half of the ordered data set, Q2 is the median value in the set, and Q3 is the middle value in the second half of the ordered data set.

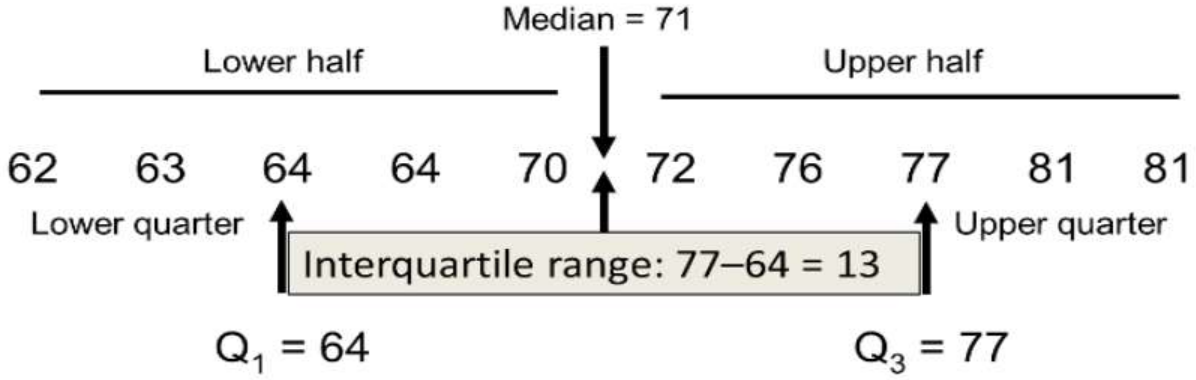
Interquartile range (IQR)

A good measure of the spread of data is the ***interquartile range (IQR)*** or the difference between Q3 and Q1. This gives us the width of the box, as well. A small width means more consistent data values since it indicates less variation in the data or that data values are closer together. So, **IQR = Q3 - Q1**

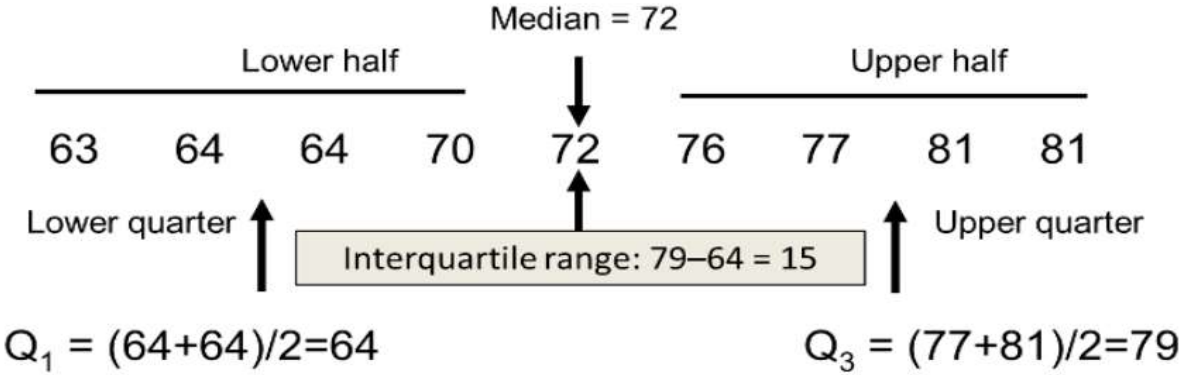
Descriptive Measures for Numeric Variable...

Interquartile Range = Q3-Q1

- **With an Even Sample Size:**
- For the sample (n=10) the median diastolic blood pressure is 71 (50% of the values are above 71, and 50% are below).
- The quartiles can be determined in the same way we determined the median, except we consider each half of the data set separately.



- **With an Odd Sample Size:**
- For the sample (n=10) the median diastolic blood pressure is 72 (50% of the values are above 72, and 50% are below).
- When the sample size is odd, the median and quartiles are determined in the same way.
- Suppose in the previous example, the lowest value (62) were excluded, and the sample size was n=9. The median and quartiles are indicated below.



Descriptive Measures for Numeric Variable...

Outliers and Tukey Fences:

Tukey Fences

- **When there are no outliers** in a sample, the **mean and standard deviation** are used to summarize a typical value and the variability in the sample, respectively.
- **When there are outliers in a sample**, the **median and interquartile range** are used to summarize a typical value and the variability in the sample, respectively.

- Outliers are values **below Q1-1.5(Q3-Q1)** or **above Q3+1.5(Q3-Q1)** or equivalently, values below **Q1-1.5 IQR** or above **Q3+1.5 IQR**.

- In previous example, for the diastolic blood pressures, the lower limit is 64 - 1.5(77-64) = 44.5 and the upper limit is 77 + 1.5(77-64) = 96.5. The diastolic blood pressures range from 62 to 81. Therefore there are no outliers.

Descriptive Measures for Numeric Variable...

Example : The Full Framingham Cohort Data

- The Framingham Heart Study is a long-term, ongoing cardiovascular cohort study on residents of the city of Framingham, Massachusetts. The study began in 1948 with 5,209 adult subjects from Framingham, and is now on its third generation of participants.
- Table 1 displays the means, standard deviations, medians, quartiles and interquartile ranges for each of the continuous variables in the subsample of n=10 participants who attended the seventh examination of the Framingham Offspring Study.

Table 1 - Summary Statistics on n=10

Characteristic	Mean	Standard Deviation	Median	Q1	Q3	IQR
Systolic Blood Pressure	121.2	11.1	122.5	113.0	127.0	14.0
Diastolic Blood Pressure	71.3	7.2	71.0	64.0	77.0	13.0
Total Serum Cholesterol	202.3	37.7	206.5	163.0	227.0	64.0
Weight	176.0	33.0	169.5	151.0	206.0	55.0
Height	67.175	4.205	69.375	63.0	70.0	7.0
Body Mass Index	27.26	3.10	26.60	24.9	29.6	4.7

Descriptive Measures for Numeric Variable...

- Table 2 displays the observed minimum and maximum values along with the limits to determine outliers using the quartile rule for each of the variables in the subsample of n=10 participants.
- Are there outliers in any of the variables? Which statistics are most appropriate to summarize the average or typical value and the dispersion?

Table 2 - Limits for Assessing Outliers in Characteristics Measured in the n=10 Participants

Characteristic	Minimum	Maximum	Lower Limit[1]	Upper Limit[2]
Systolic Blood Pressure	105	141	92	148
Diastolic Blood Pressure	62	81	44.5	96.5
Total Serum Cholesterol	150	275	67	323
Weight	138	235	68.5	288.5
Height	60.75	72.00	52.5	80.5
Body Mass Index	22.8	31.9	17.85	36.65

[1] Determined by $Q_1-1.5(Q_3-Q_1)$

[2] Determined by $Q_3+1.5(Q_3-Q_1)$

Since there are no suspected outliers in the subsample of n=10 participants, the mean and standard deviation are the most appropriate statistics to summarize average values and dispersion, respectively, of each of these characteristics.

Descriptive Measures for Numeric Variable...

For clarity, we have so far used a very small subset of the Framingham Offspring Cohort to illustrate calculations of summary statistics and determination of outliers. For your interest, Table 3 displays the means, standard deviations, medians, quartiles and interquartile ranges for each of the continuous variable displayed in Table 1 in the full sample (n=3,539) of participants who attended the seventh examination of the Framingham Offspring Study.

Table 3-Summary Statistics on Sample of (n=3,539) Participants

Characteristic	Mean $\overline{X}$	Standard Deviation (s)	Median	Q1	Q3	IQR
Systolic Blood Pressure	127.3	19.0	125.0	114.0	138.0	24.0
Diastolic Blood Pressure	74.0	9.9	74.0	67.0	80.0	13.0
Total Serum Cholesterol	200.3	36.8	198.0	175.0	223.0	48.0
Weight	174.4	38.7	170.0	146.0	198.0	52.0
Height	65.957	3.749	65.750	63.000	68.750	5.75
Body Mass Index	28.15	5.32	27.40	24.5	30.8	6.3

Descriptive Measures for Numeric Variable...

- Table 4 displays the observed minimum and maximum values along with the limits to determine outliers using the quartile rule for each of the variables in the full sample (n=3,539).

Table 4 - Limits for Assessing Outliers in Characteristics Presented in Table 3

			Tukey Fences	
Characteristic	**Minimum**	**Maximum**	**Lower Limit[1]**	**Upper Limit[2]**
Systolic Blood Pressure	81.0	216.0	78	174
Diastolic Blood Pressure	41.0	114.0	47.5	99.5
Total Serum Cholesterol	83.0	357.0	103	295
Weight	90.0	375.0	68.0	276.0
Height	55.00	78.75	54.4	77.4
Body Mass Index	15.8	64.0	15.05	40.25

[1] Determined by $Q_1-1.5(Q_3-Q_1)$

[2] Determined by $Q_3+1.5(Q_3-Q_1)$

Are there outliers in any of the variables?

Which statistics are most appropriate to summarize the average or typical values and the dispersion for each variable?

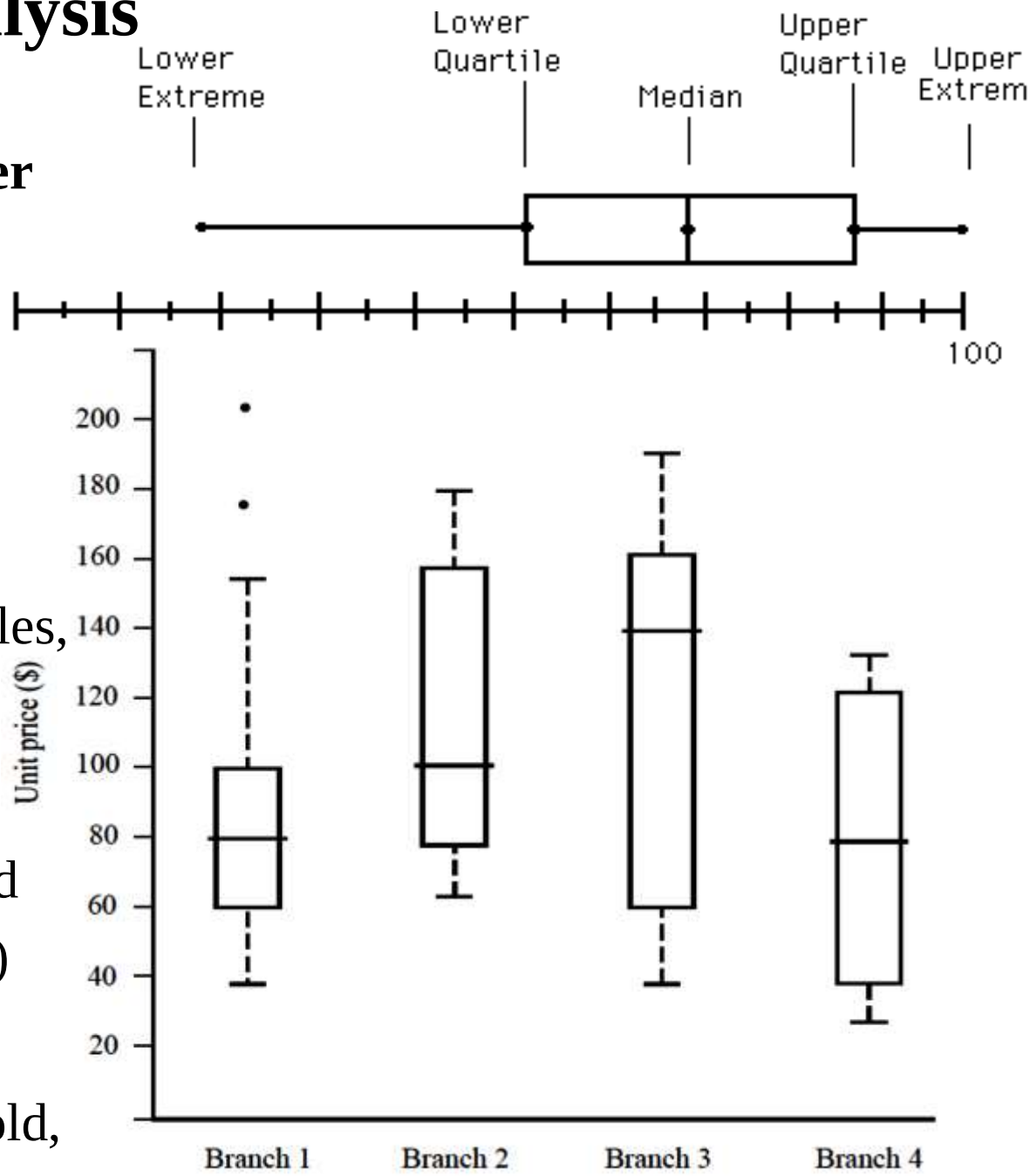
Descriptive Measures for Numeric Variable...

Observations on example.....

- In the full sample, each of the characteristics has outliers on the upper end of the distribution as the maximum values exceed the upper limits in each case. There are also outliers on the low end for diastolic blood pressure and total cholesterol, since the minimums are below the lower limits.
- For some of these characteristics, the difference between the upper limit and the maximum (or the lower limit and the minimum) is small (e.g., height, systolic and diastolic blood pressures), while for others (e.g., total cholesterol, weight and body mass index) the difference is much larger. This method for determining outliers is a popular one but not generally applied as a hard and fast rule. In this application it would be reasonable to present means and standard deviations for height, systolic and diastolic blood pressures and medians and interquartile ranges for total cholesterol, weight and body mass index.

Boxplot Analysis

- **Boxplots are a popular way of visualizing a distribution. A boxplot incorporates the five-number summary as follows:**
- **Five-number summary** of a distribution
  - Minimum, Q1, Median, Q3, Maximum
- **Boxplot** Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - The median is marked by a line within the box
  - Two lines (called **whiskers**) outside the box extend to the smallest (Minimum) and largest (Maximum) observations.
  - **Outliers:** points beyond a specified outlier threshold, plotted individually

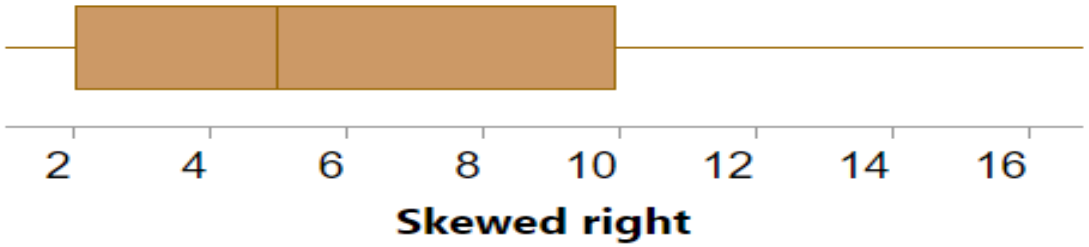


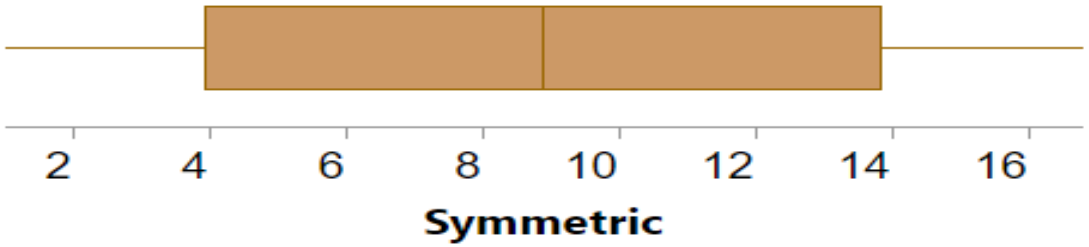
Boxplot for the unit price data for items sold at four branches of AllElectronics during a given time period.

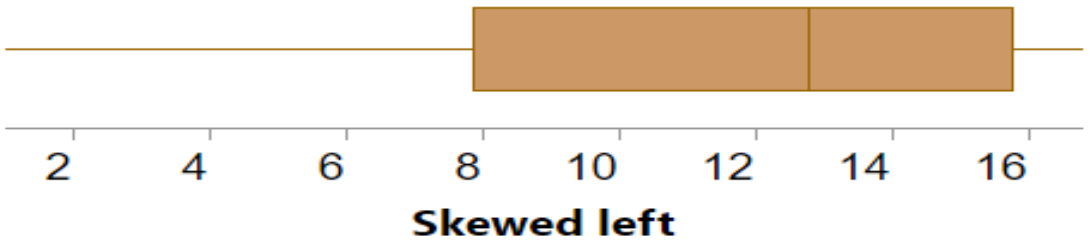
Descriptive Measures for Numeric Variable...

Boxplot Analysis (cont..)

Finally, boxplots often provide information about the shape of a data set. The examples below show some common patterns.







Each of the boxplots illustrates a different **skewness pattern.**

If most of the observations are concentrated on the low end of the scale, the distribution is skewed right; and vice versa.

If a distribution is symmetric, the observations will be evenly split at the median, as shown in the middle figure.

Descriptive Measures for Numeric Variable...

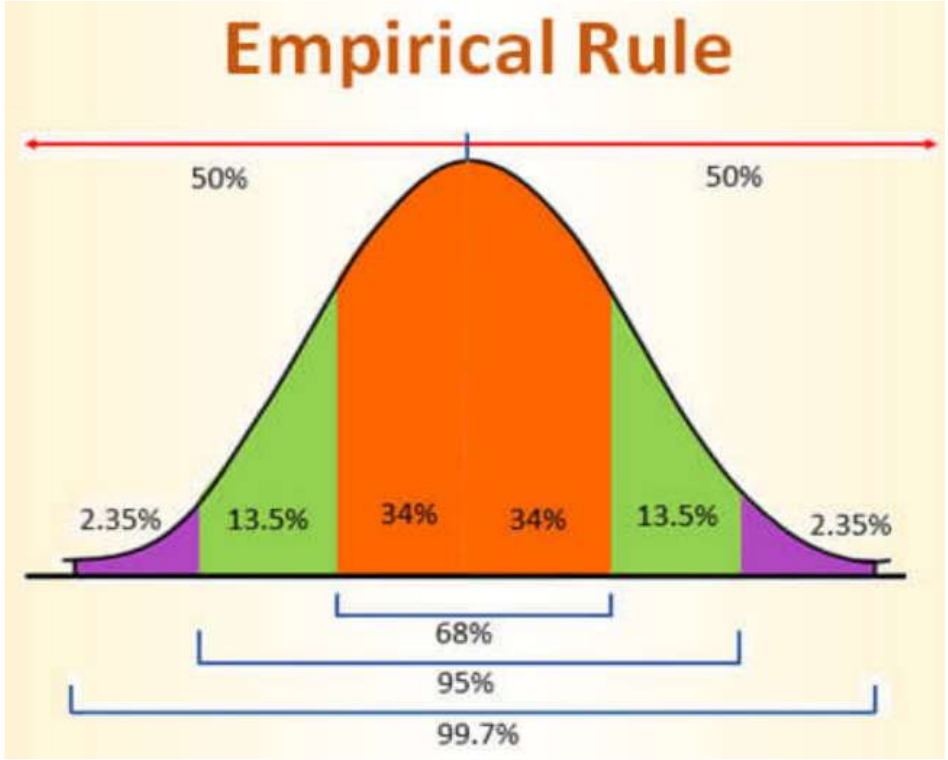
Empirical Rules for Interpreting Standard Deviation

- The interpretation of the standard deviation can be stated as three **empirical rules**.
  - "Empirical" means that they are based on commonly observed data, as opposed to theoretical mathematical arguments.
  - If the values of a variable are approximately *normally* distributed (symmetric and bell-shaped), then the following rules hold:
    - Approximately **68%** of the observations are **within one standard deviation of the mean.**
    - Approximately **95%** of the observations are **within two standard deviations of the mean.**
    - Approximately **99.7%** of the observations are **within three standard deviations of the mean.**

Descriptive Measures for Numeric Variable...

The beauty of the normal curve:

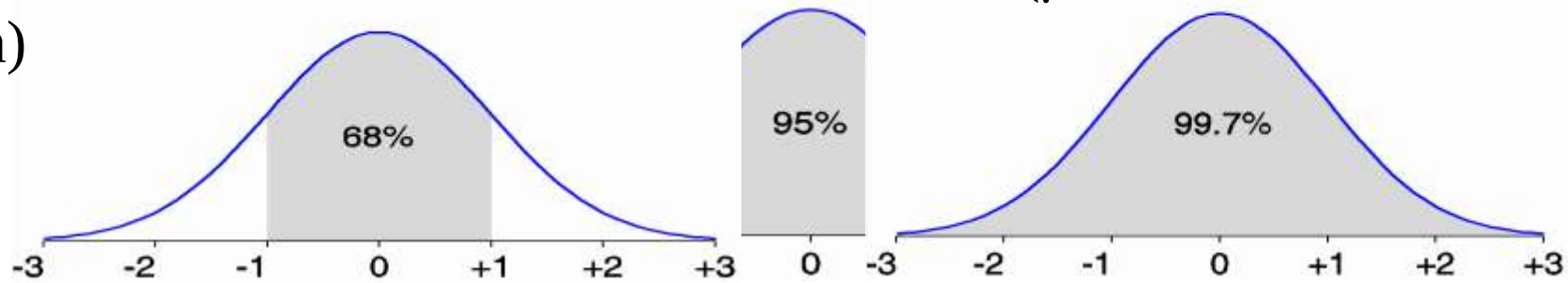
(EmpiricalRules for Interpreting Standard Deviation)



68-95-99.7 Rule

No matter what μ and σ are,

- the area between μ-σ and μ+σ is about 68%; the area between μ-2σ and μ+2σ is about 95%;
- and the area between μ-3σ and μ+3σ is about 99.7%.
- Almost all values fall within 3 standard deviations. (μ: mean, σ: standard deviation)



SKEWNESS
&
KURTOSIS

(Measure of Shape)

Concept of Skewness

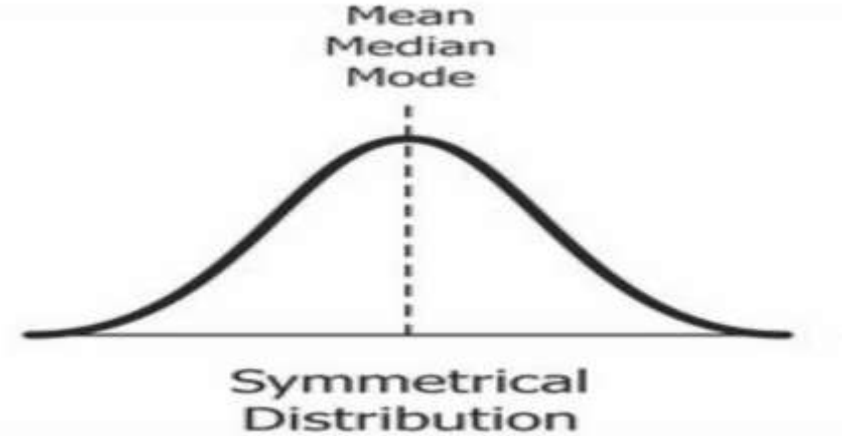
A distribution is said to be skewed-when the mean, median and mode fall at different position in the distribution and the balance (or center of gravity) is shifted to one side or the other i.e. to the left or to the right.

Th concept of skewness helps us to understand the relationship between three measures-

- **Mean.**
- **Median.**
- **Mode.**

Symmetrical Distribution

- A frequency distribution is said to be symmetrical if the frequencies are equally distributed on both the sides of central value.
- A symmetrical distribution may be either bell – shaped or U shaped.
- In symmetrical distribution, the values of mean, median and mode are equal i.e. **Mean=Median=Mode**



Skewed Distribution

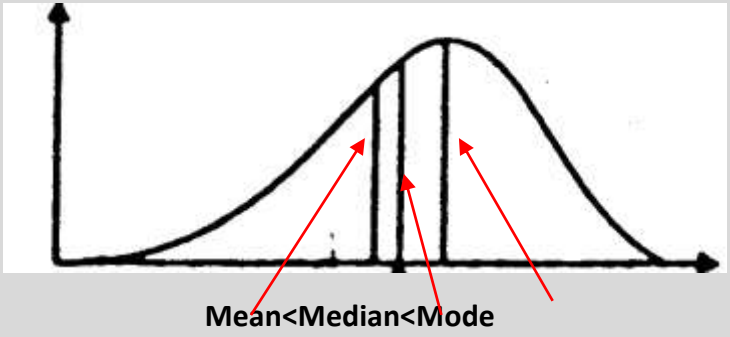
- A frequency distribution is said to be skewed if the frequencies are not equally distributed on both the sides of the central value.

- A skewed distribution may be-

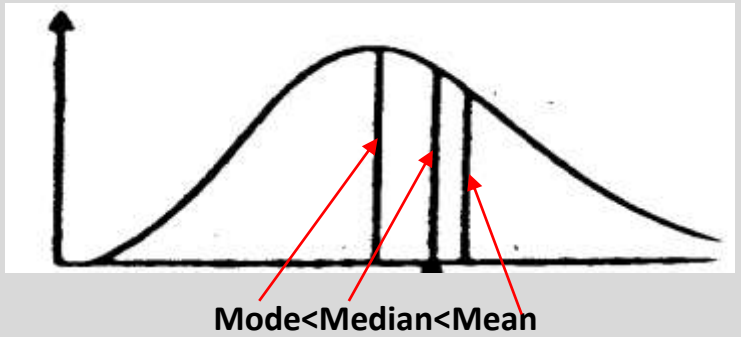
  - **Positively Skewed**
  - **Negatively Skewed**

Skewed Distribution

- **<u>Negatively Skewed</u>**
- In this, the distribution is skewed to **the left (negative)**
- Here, **Mode** exceeds Mean and Median.



- **<u>Positively Skewed</u>**
- In this, the distribution is skewed to **the right (positive)**
- Here, **Mean** exceeds Mode and Median.



Tests of Skewness

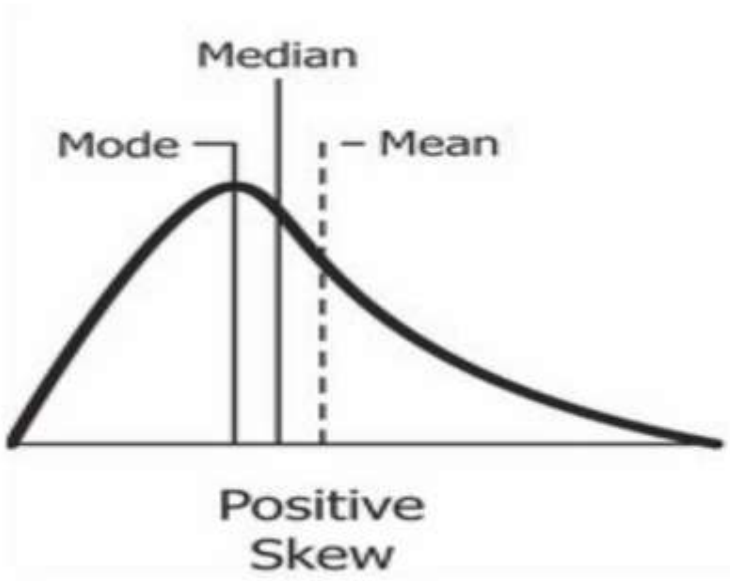
In order to ascertain whether a distribution is skewed or not the following tests may be applied. Skewness is present if:

1. The values of mean, median and mode do not coincide.
2. When the data are plotted on a graph they do not give the normal bell shaped form i.e. when cut along a vertical line through the center the two halves are not equal.
3. The sum of the positive deviations from the median is not equal to the sum of the negative deviations.
4. Quartiles are not equidistant from the median.
5. Frequencies are not equally distributed at points of equal deviation from the mode.

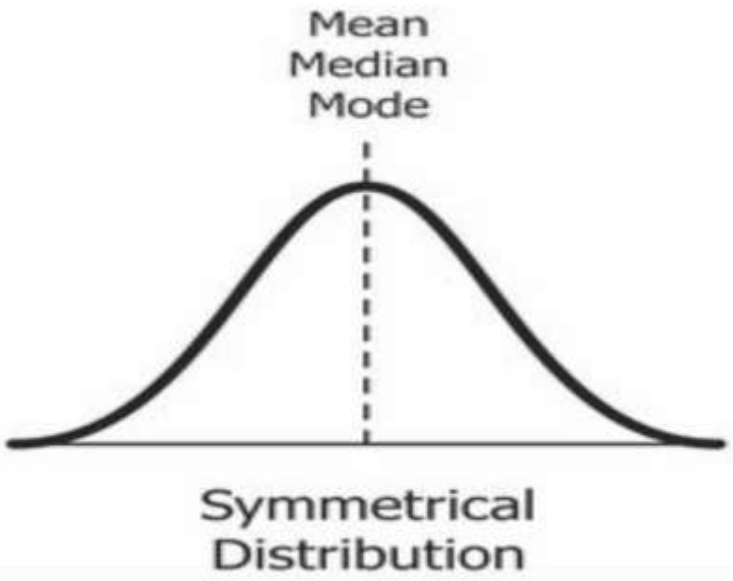
Graphical Measures of Skewness

- Measures of skewness help us to know to what degree and in which direction (positive or negative) the frequency distribution has a departure from symmetry.
- Positive or negative skewness can be detected graphically (as below) depending on whether the right tail or the left tail is longer but, we don't get idea of the magnitude
- Hence some statistical measures are required to find the magnitude of lack of symmetry

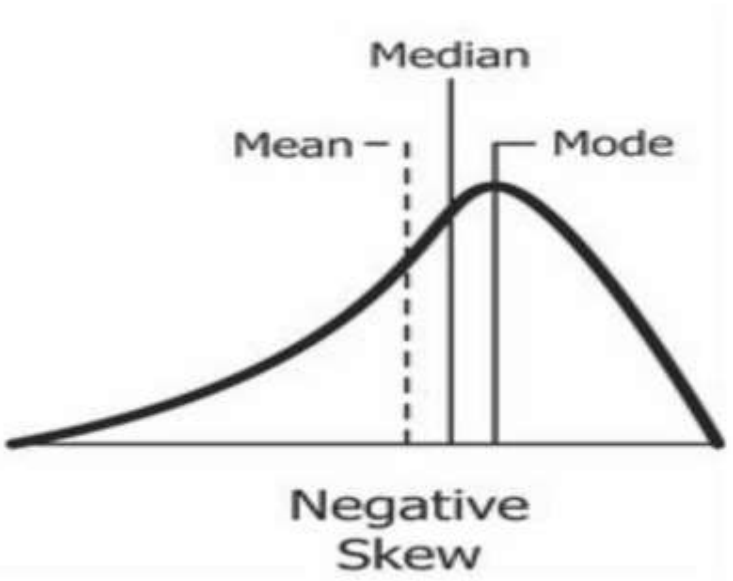
Mean> Median> Mode



Mean=Median=Mode



Mean<Median<Mode



Statistical Measures of Skewness

Absolute Measures of Skewness

- Absolute measures of dispersion include: Range, quartile deviation, mean deviation, standard deviation and variance.
- Absolute measures of dispersion use the original units of data. They are most useful for understanding the dispersion within the sample.

Relative Measures of Skewness

- In order to make comparison of the skewness among two or more distributions/data sets relative measures of skewness are calculated.
- Relative measure are always in terms of ratios or percentage. One of the relative measures of dispersion is the ratio of the standard deviation to the mean.

Following are the absolute measures of skewness:

- Skewness (Sk) = Mean–Median
- Skewness (Sk) = Mean–Mode
- Skewness (Sk) = (Q3-Q2)-(Q2-Q1)

There are four measures of skewness:

- β and γ Coefficient of skewness
- **Karl Pearson's Coefficient of skewness**
- Bowley's Coefficient of skewness
- Kelly's Coefficient of skewness

Karl Pearson's Coefficient of Skewness……01

- This method is most frequently used for measuring skewness. The formula for measuring coefficient of skewness is given by

$$SK_P = \frac{\text{Mean} - \text{Mode}}{\sigma}$$

Where,

$\mathbf{SK_P}$ = Karl Pearson's Coefficient of skewness,

$\boldsymbol{\sigma}$ = standard deviation.

Note: Ccoefficient of skewness lies between -3 to +3.

Karl Pearson's Coefficient of Skewness…..

In case the mode is indeterminate, the coefficient of skewness is:

$$SK_p = \frac{\text{Mean} - (3\ \text{Median} - 2\ \text{Mean})}{\sigma}$$

Now this formula is equal to

$$SK_P = \frac{3(\text{Mean} - \text{Median})}{\sigma}$$

When the distribution is **symmetrical**: The value of coefficient of skewness is **zero**,

When the distribution is **positively skewed:** The value of coefficient of skewness is **positive.**

When the distribution is **negatively skewed: T**he value of coefficient of skewness is **negative**,

Example:

Question: For a distribution Karl Pearson's coefficient of skewness is 0.64, standard deviation is 13 and mean is 59.2 Find mode and median.

Solution: We have given

$S_k = 0.64$, $\sigma = 13$ and Mean = 59.2

Therefore by using formula

$$S_k = \frac{Mean - Mode}{\sigma}$$

$$0.64 = \frac{59.2 - Mode}{13}$$

Mode = 59.20 – 8.32 = 50.88
Mode = 3 Median – 2 Mean
50.88 = 3 Median - 2 (59.2)

$$\text{Median} = \frac{50.88+118.4}{3} = \frac{169.28}{3} = 56.42$$

SKEWNESS EXAMPLE:
https://www.wallstreetmojo.com/skewness-formula/

https://byjus.com/skewness-formula/

Kurtosis

•Kurtosis is another measure of the shape of a frequency curve. It is a Greek word, which means bulginess.

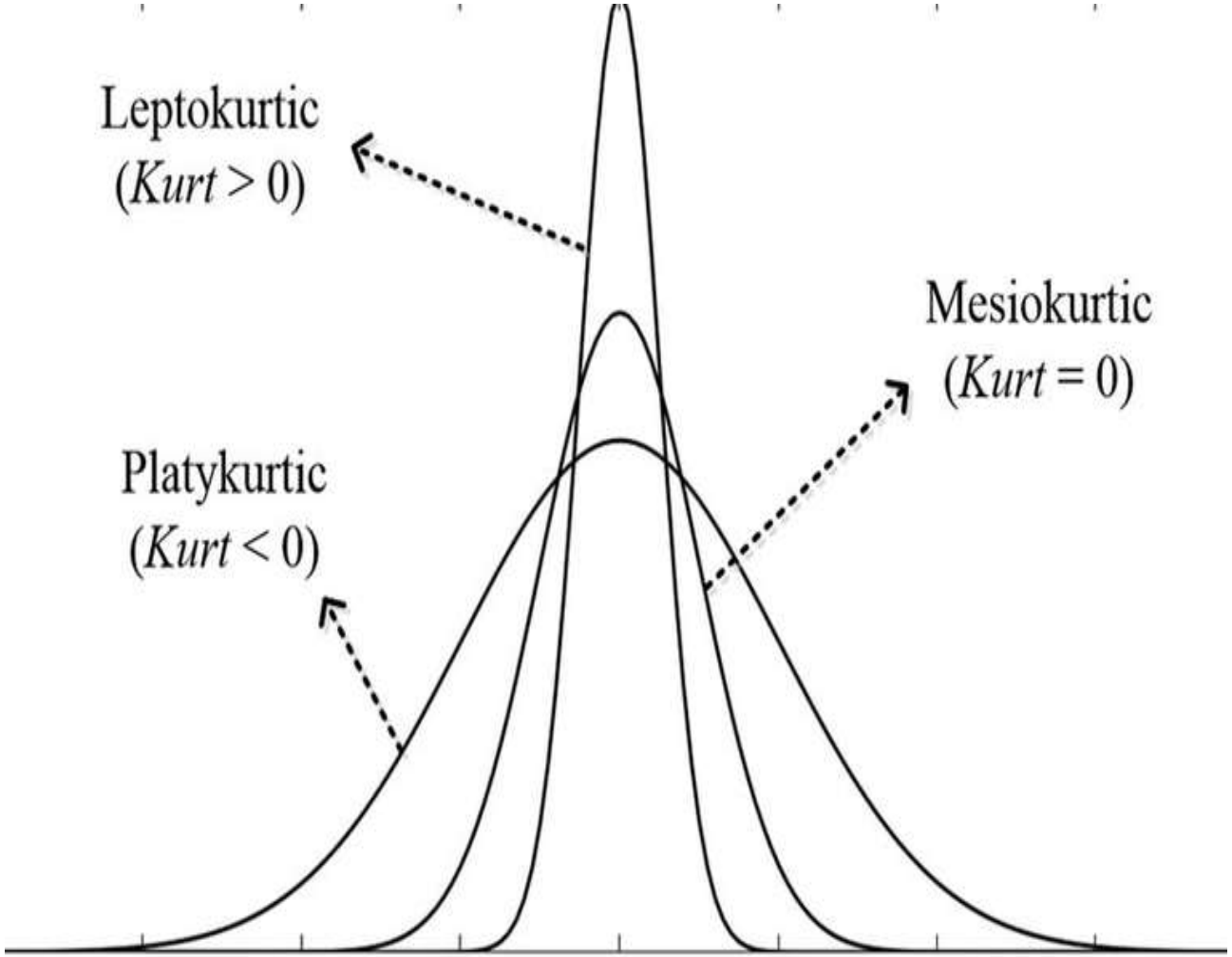
•While skewness signifies the extent of asymmetry, kurtosis measures the degree of peakedness of a frequency distribution.

•Karl Pearson classified curves into three types on the basis of the shape of their peaks. These are:-

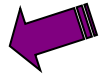
- **Leptokurtic**
- **Mesokurtic**
- **Platykurtic**

Kurtosis

- When the peak of a curve becomes relatively high then that curve is called **Leptokurtic.**
- When the curve is flat-topped, then it is called **Platykurtic**.
- Since normal curve is neither very peaked nor very flat topped, so it is taken as a basis for comparison.
- This normal curve is called **Mesokurtic**.



Chapter 2: **Statistical Concepts**

- **Distribution of a single variable:**
  - Basic Concepts (population and sample, data sets, variables, and observations, types of data)
  - Descriptive measures for categorical variables
  - Descriptive measures for numerical variables
  - Outliers and Missing values 
- **Finding relationships among variables:**
  - Categorical variables
  - Numerical variables
  - Categorical variables and a Numerical variable
- **Sampling and distributions:**
  - Terminology
  - Estimation
  - Confidence Interval estimation
  - Sampling distributions
  - Confidence interval,
  - Hypothesis testing, Chi-square test for independence

Outlier

- An outlier is a value in an entire observation that lies well outside of the norm.
- Some statisticians define an outlier as any value more than three standard deviations from the mean, but this is only a rule of thumb.
- Even if values are not unusual by themselves, there still might be unusual combinations of values.
- When dealing with outliers, it is best to run the analyses two ways: with the outliers and without them.
- Outliers can be considered as extreme values, and for any particular data set, it can be decided how extreme a value needs to be to qualify as an outlier

- For example, let us consider a row of data [10,15,22,330,30,45,60]. In this dataset, we can easily conclude that 330 is way off from the rest of the values in the dataset, thus 330 is an outlier. It was easy to figure out the outlier in such a small dataset, but when the dataset is huge, we need various methods to determine whether a certain value is an outlier or necessary information.

Outlier...

Types of outliers : There are three types of outliers

• **Global Outliers:** The data point or points whose values are far outside everything else in the dataset are global outliers. Suppose we look at a taxi service company's number of rides every day. The rides suddenly dropped to zero due to the pandemic-induced lockdown. This sudden decrease in the number is a global outlier for the taxi company.

• **Collective Outliers:** Some data points collectively as a whole deviates from the dataset. These data points individually may not be a global or contextual outlier, but they behave as outliers when aggregated together. For example, closing all shops in a neighborhood is a collective outlier as individual shops keep on opening and closing, but all shops together never close down; hence, this scenario will be considered a collective outlier.

• **Contextual Outliers:** Contextual outliers are those values of data points that deviate quite a lot from the rest of the data points that are in the same context, however, in a different context, it may not be an outlier at all. For example, a sudden surge in orders for an e-commerce site at night can be a contextual outlier

Outliers can lead to vague or misleading predictions while using machine learning models. Specific models like linear regression, logistic regression, and support vector machines are susceptible to outliers. Outliers decrease the mathematical power of these models, and thus the output of the models becomes unreliable.

Outlier...

What causes Outliers?

- Data Entry Errors:
- Measurement Error:
- Experimental Error:
- Data Processing Error:
- Natural Outlier:

What is the impact of Outliers on a dataset?

Without Outlier	With Outlier
4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7	4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7,300
Mean = 5.45	Mean = 30.00
Median = 5.00	Median = 5.50
Mode = 5.00	Mode = 5.00
Standard Deviation = 1.04	Standard Deviation = 85.03

Outlier...

How to detect outliers

1. Sorting method
2. Using visualizations
3. Statistical outlier detection
4. Using the interquartile range

1. Sorting method:

Trimmed Mean: is calculated by eliminating extreme observations at both ends of the sample (between 10%-25%)

e.g. 1, 6, 12, 14, 20, 24, 36, 100 [Regular Mean: 26.62]

~~1, 6~~, 12, 14, 20, 24, ~~36, 100~~

=> 12, 14, 20, 24 (After removal of 25% from end)[Now, Trimmed Mean: 17.5]

2. Using Visualization:

Most commonly used method to detect outliers is visualization. Various visualization methods, like Box-plot, Histogram, Scatter Plot can be used to detect outliers.

Outlier...

3. Statistical Outlier Detection: [Using Z Score/ Standard Deviation]

When the data follow a normal distribution, then the standard deviation of the data, or the equivalent z-score can be used to detect outliers.

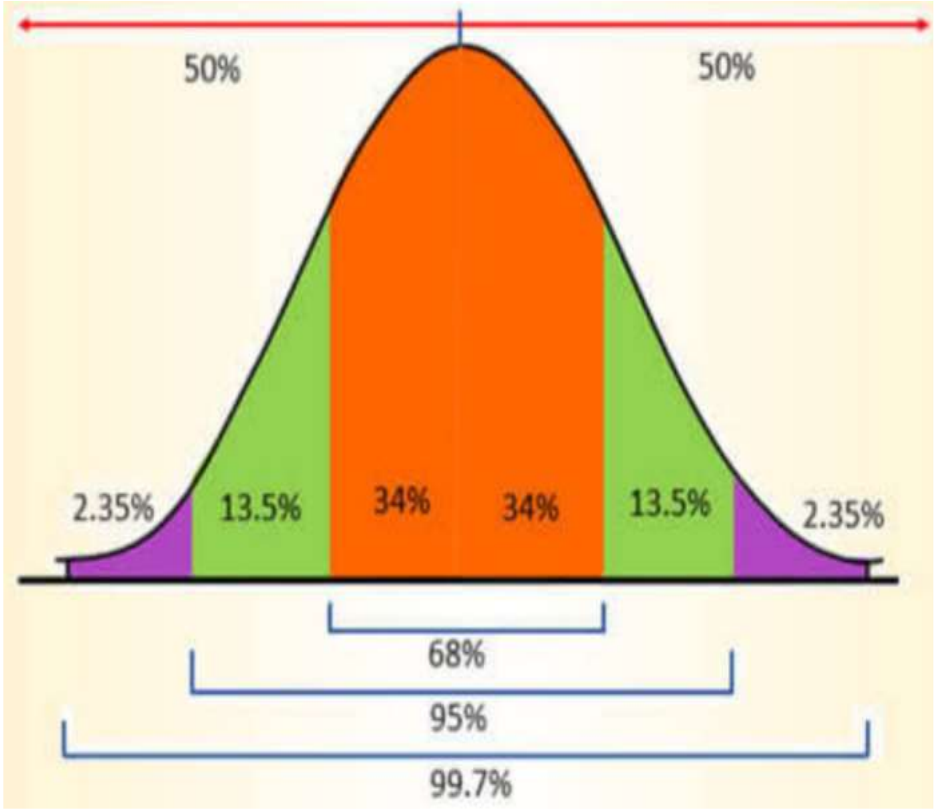
a. Using Standard Deviation:

In statistics, standard deviation measures the spread of data around the mean, and it captures how far away from the mean the data points are.

Let's denote the **standard deviation by σ, and the mean by**

One approach to outlier detection is to set the lower limit to three standard deviations below the mean (μ - 3*σ), and the upper limit to three standard deviations above the mean (μ + 3*σ). Any data point that falls outside this range is detected as an outlier.

As 99.7% of the data typically lies within three standard deviations, the **number of outliers will be close to 0.3%** of the size of the dataset.



Outlier...

b. Using Z-Score:

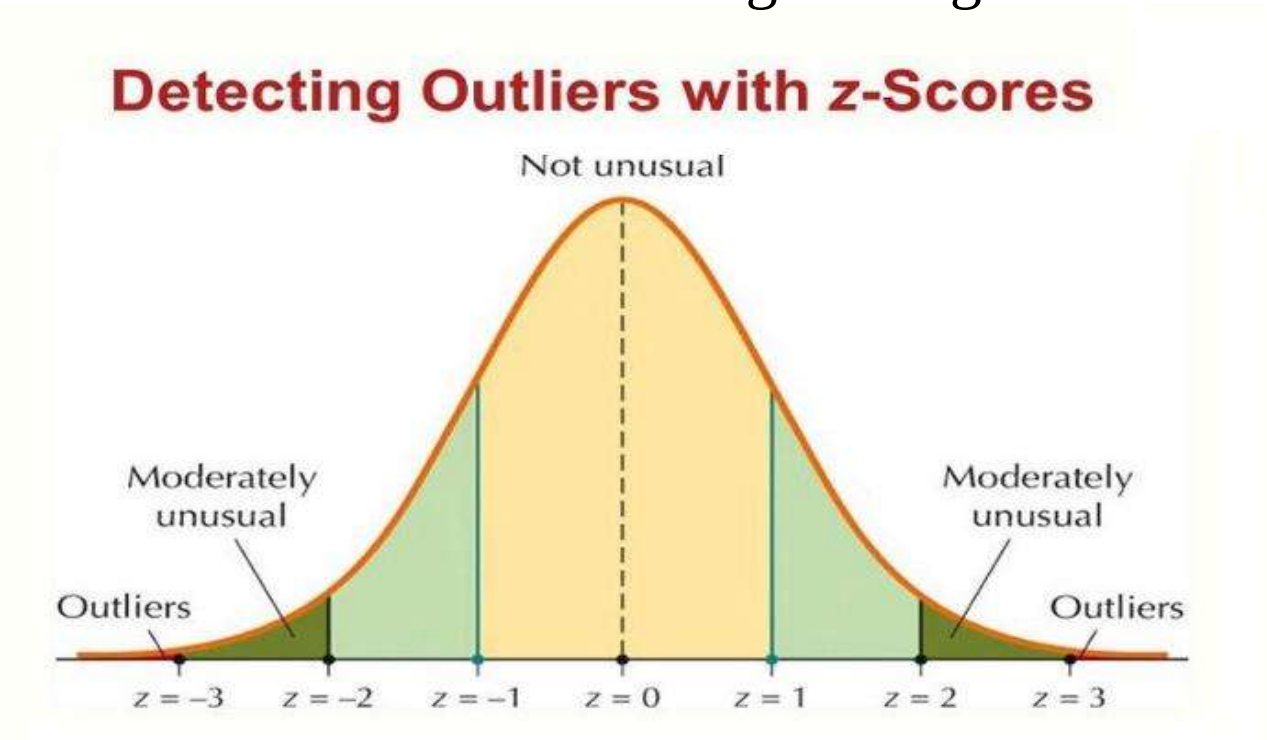
A z score is a standard score that tells you how many standard deviations away from the mean an individual value (x) lies:

- A positive z score means that your x value is greater than the mean.
- A negative z score means that your x value is less than the mean.
- A z score of zero means that your x value is equal to the mean.

As a rule of thumb, values with a z score greater than 3 or less than –3 are often determined to be outliers. The Z-score of each data points can be found. If a value is as a high enough or low enough z score, it can be considered an outlier.

$$Z\ Score = \frac{(\text{Observation} - \text{Mean})}{\text{Standard Deviation}}$$

$$Z\ Score = \frac{X - \mu}{\sigma}$$



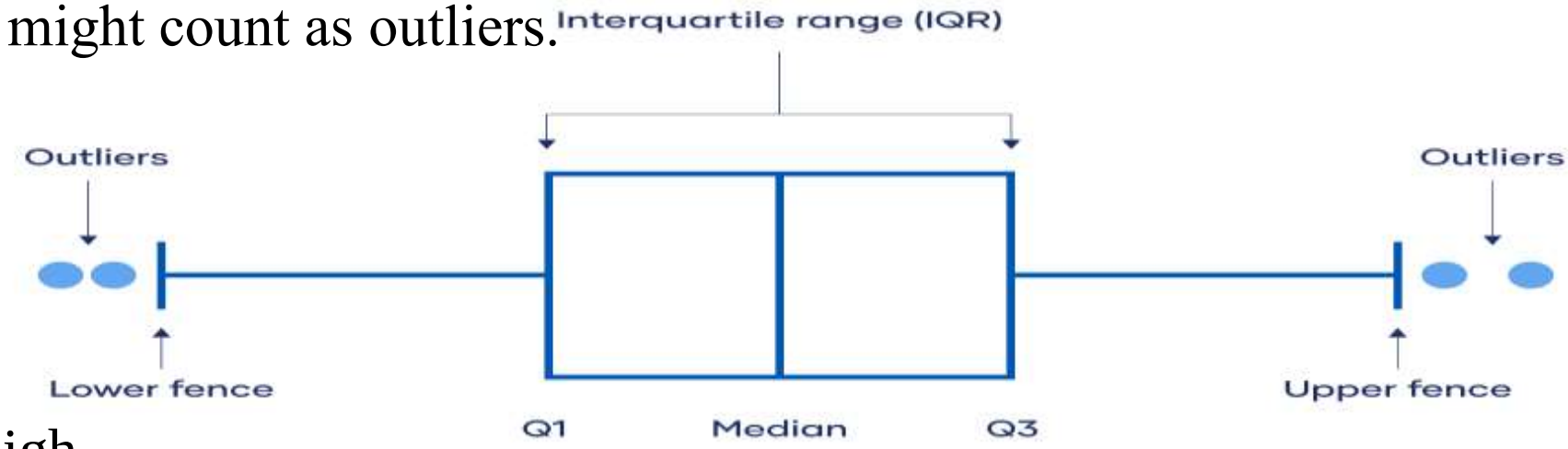
Outlier...

4. Using the interquartile range

The interquartile range (IQR) tells the range of the middle half of the dataset. IQR can be used to create "fences" around the data and the outliers can be defined as the data that fall outside those fences.

This method is helpful if you have a few values on the extreme ends of your dataset, but you aren't sure whether any of them might count as outliers.

Interquartile range method:



1. Sort your data from low to high
2. Identify the first quartile (Q1), the median, and the third quartile (Q3).
3. Calculate your IQR = Q3 – Q1
4. Calculate your upper fence = Q3 + (1.5 * IQR)
5. Calculate your lower fence = Q1 – (1.5 * IQR)
6. Use your fences to highlight any outliers, all values that fall outside your fences.

Your outliers are any values greater than your upper fence or less than your lower fence.

Outlier...

How to Remove Outliers

Deleting observations: We delete outlier values if it is due to data entry error, data processir error or outlier observations are very small in numbers. We can also use trimming at both en to remove outliers.

Transforming and binning values: Transforming variables can also eliminate outlier Natural log of a value reduces the variation caused by extreme values.

Binning is also a form of variable transformation.

Missing Data

- **Data is not available/ Missing data may be due to**
  - ✓E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
  - ✓Information is not collected :(e.g., people decline to give their age and weight)
  - ✓Attributes may not be applicable to all cases (e.g., annual income is not applicable to children)
  - ✓equipment malfunction
  - ✓inconsistent with other recorded data and thus deleted
  - ✓data not entered due to misunderstanding
  - ✓certain data may not be considered important at the time of entry
  - ✓not register history or changes of the data

How to Handle Missing Data?

- **Ignore the tuple**: usually done when class label is missing (assuming the task is classification—not effective in certain cases)
- Fill in the missing value **manually**: tedious + infeasible?

- Use a **global constant** to fill in the missing value: e.g., “unknown”, a new class?! simple but not foolproof.
- Use the **central tendency (mean/median)** to fill in the missing value. Normal->Mean, Skewed->Median
- Use the **most probable value** to fill in the missing value: inference-based such as regression, Bayesian formula, decision tree

Bias the data

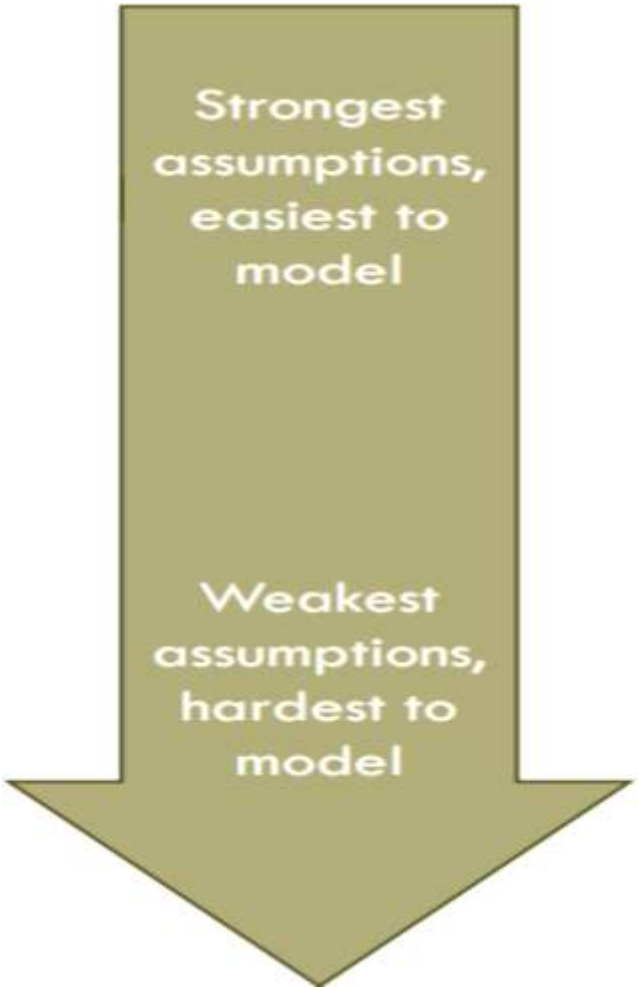
Types of Missing Values

Why missing data is a problem?

Ans: It creates bias in the data. because we don't know that the data is missing randomly/missedout/intensionally.

*Bias data: produce lack of prdictivity & trustworthyness

- **Missing completely at random (MCAR)**
- **Missing at Random (MAR)**
- **Missing Not at Random (MNAR)**



Missing Completely at Random (MCAR) (Types of Missing Values…)

Assumption: If a person has missing data then it is completely unrelated to the other information in the data. The missingness on the variable is completely unsystematic.

- Missingness of a value is independent of attributes
- Fill in values based on the attribute
- Analysis may be unbiased overall

Example when we take a random sample of a population, where each member has the same chance of being included in the sample.

ID	Gender	Age	Income
1	Male	Under 30	Low
2	Female	Under 30	Low
3	Female	30 or more	High
4	Female	30 or more	
5	Female	30 or more	High

When data is missing completely at random, it means that we can undertake analyses using only observations that have complete data (provided we have enough of such observations).

Missing at Random (MAR) Types of Missing Values…

- Missingness is related to other variables
- Fill in values based on other values
- Almost always produces a bias in the analysis

Example of MAR is when we take a sample from a population, where the probability to be included depends on some known property.

A simple predictive model is that income can be predicted based on gender and age. Looking at the table, we note that our missing value is for a Female aged 30 or more, and observations say the other females aged 30 or more have a High income. As a result, we can predict that the missing value should be High.

ID	Gender	Age	Income
1	Male	Under 30	Low
2	Female	Under 30	Low
3	Female	30 or more	High
4	Female	30 or more	
5	Female	30 or more	High

There is a systematic relationship between the inclination of missing values and the observed data. All that is required is a probabilistic relationship

Missing not at Random (MNAR) - Nonignorable

Types of Missing Values…

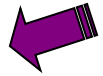
- Missingness is related to unobserved measurements and they are not random
- The missing values are related to the values of that variable itself, even after controlling for other variables.

MNAR means that the probability of being missing varies for reasons that are unknown to us.

Example: when smoking status is not recorded in patients admitted as an emergency with an intention (not random), then it is more likely to have worse outcomes from surgery.

Strategies to handle MNAR are to find more data about the causes for the missingness, or to perform what-if analyses to see how sensitive the results are under various scenarios.

Chapter 2: **Statistical Concepts**

- **Distribution of a single variable:**
  - Basic Concepts (population and sample, data sets, variables, and observations, types of data)
  - Descriptive measures for categorical variables
  - Descriptive measures for numerical variables
  - Outliers and Missing values
- **Finding relationships among variables:** 
  - Categorical variables
  - Categorical variables and a Numerical variable
  - Numerical variables
- **Sampling and distributions:**
  - Terminology
  - Estimation
  - Confidence Interval estimation
  - Sampling distributions
  - Confidence interval,
  - Hypothesis testing, Chi-square test for independence

Finding relationships among variables

This is an important first step in any exploratory data analysis.

- To look closely at variables one at a time, but it is almost never the last step.
- The primary interest is usually in relationships between variables.

For a variable salary, the entire focus was on how salaries were distributed over some range.

It is natural to ask what drives salaries.

- *Does it depend on qualitative factors, such as;*
  - ❖ *Player's team or position?*
- *Does it depend on quantitative factors, such as;*
  - ❖ *Number of hits the player gets or the number of strikeouts?*

Name	Team	Position	Salary
Mike Trout	Los Angeles Angels	Outfielder	$3,40,83,333
Clayton Kershaw	Los Angeles Dodgers	Pitcher	$3,40,00,000
Zack Greinke	Arizona Diamondbacks	Pitcher	$3,19,54,483
Miguel Cabrera	Detroit Tigers	First Baseman	$3,00,00,000
David Price	Boston Red Sox	Pitcher	$3,00,00,000
Jake Arrieta	Philadelphia Phillies	Pitcher	$3,00,00,000
Yoenis Cespedes	New York Mets	Outfielder	$2,90,00,000
Justin Verlander	Houston Astros	Pitcher	$2,80,00,000
Jon Lester	Chicago Cubs	Pitcher	$2,75,00,000
Albert Pujols	Los Angeles Angels	First Baseman	$2,70,00,000
Felix Hernandez	Seattle Mariners	Pitcher	$2,68,57,143

To answer, it is required to examine relationships between various variables and salary.

Types of Relationships among Variables:

a. Categorical vs Categorical
b. Categorical vs Numerical
c. Numerical vs Numerical

1. Relationships Among Categorical Variables

(Categorical vs Categorical)

Consider a data set with two categorical variables:

Smoking and Drinking.

Smoking	Drinking
Non Smoker (NS)	Non Drinker (ND)
Occasional Smoker (OS)	Occasional Drinker (OD)
Heavy Smoker (HS)	Heavy Drinker (HD)

	A	B	C
1	Person	Smoking	Drinking
2	1	NS	OD
3	2	NS	HD
4	3	OS	HD
5	4	HS	ND
6	5	NS	OD
7	6	NS	ND
8	7	NS	OD
9	8	NS	ND
10	9	OS	HD
11	10	HS	HD

Do the data indicate that smoking and drinking habits are related? For example,

- *Do nondrinkers tend to be nonsmokers?*
- *Do heavy smokers tend to be heavy drinkers?*

1. Relationships Among Categorical Variables...

- The most meaningful way to describe a categorical variable is with counts, possibly expressed as percentages of totals, and corresponding charts of the counts.
- The counts of the categories of either variable can be found separately, but more importantly, it is required to find counts of the joint categories of the two variables, such as the count of all nondrinkers who are also nonsmokers. It is customary to display all such counts in a table called a **crosstabs /contingency table.**

Do the data indicate that smoking and drinking habits are related? For example,

- Do nondrinkers tend to be nonsmokers?
- Do heavy smokers tend to be heavy drinkers?

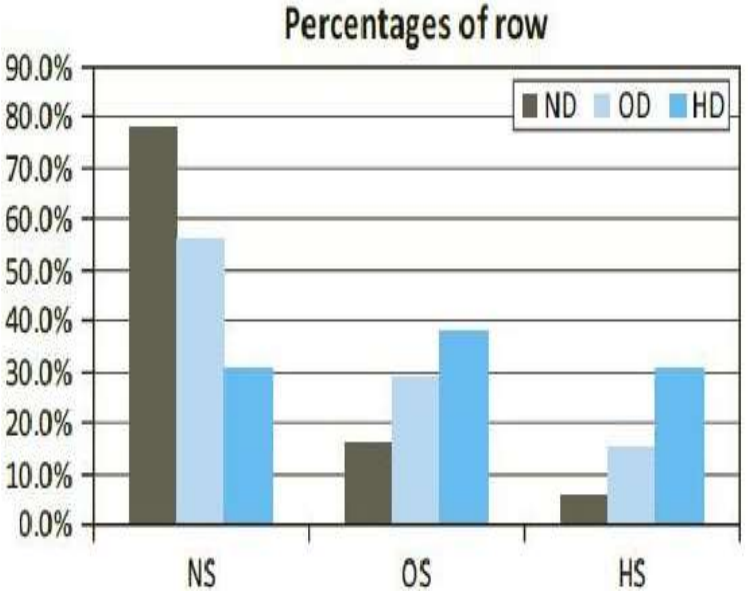
The 1st two arguments are for the condition on smoking; the 2nd two are for the condition on drinking.

- **A contingency table displays how two categorical variables are related in a table with how many individuals fall in each combination of categories.**
- **The categories of one variable define the rows and categories of the other variable define the columns of the table.**

1. Relationships Among Categorical Variables...

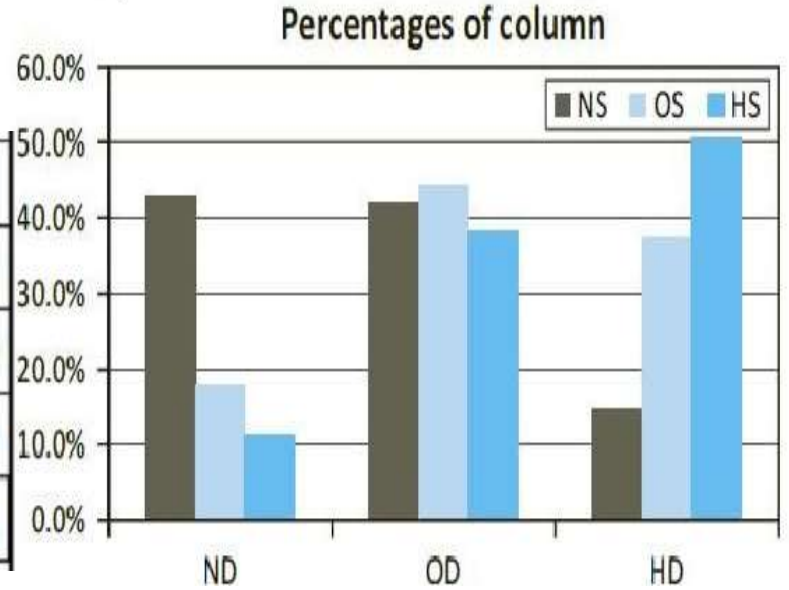
Contigency Table in terms of % of Row

	NS	OS	HS	Total
ND	78.0%	16.0%	6.0%	100.0%
OD	56.0%	29.0%	15.0%	100.0%
HD	31.0%	38.0%	31.0%	100.0%



Contigency Table in terms of % of Col

	NS	OS	HS
ND	43.1%	18.1%	11.3%
OD	42.0%	44.4%	38.1%
HD	14 9%	37 4%	50 6%
Total	100.0%	100.0%	100.0%



Contigency Table

	NS	OS	HS	Total
ND	2118	435	163	2716
OD	2061	1067	552	3680
HD	733	899	733	2365
Total	4912	2401	1448	8761

Total count or percentage across the columns or rows of contigency table is called as marginal (frequency marginal or percentage marginal respectively)

1. Relationships Among Categorical Variables...

Process to find relation/association between categorical variables is

1. **Set up hypothesis:**
   - Null hypothesis: Assumes that there is no association between the two variables.
   - Alternative hypothesis: Assumes that there is an association between the two variables
2. **Collect Data**
3. **Define the significance level and on the basis of significance level proove or reject the Null hypothesis**

Methods to find Relationships Among Categorical Variables

-- Chi-square test

– Cramer’s V

– Bonferroni correction

2. Relationships Among Categorical and Numerical Variables...

It describes a very common situation where the goal is to break down a numerical variable by a categorical variable.

It occurs when comparison of a numerical measure with two or more subpopulations is made.

Examples:

- *The subpopulations are males and females, and the numerical measure is salary.*
- *The subpopulations are different regions of the country, and the numerical measure is the cost of living.*
- *The subpopulations are different days of the week, and the numerical measure is the number of customers going to a particular fast-food chain.*
- *The subpopulations are different machines in a manufacturing plant, and the numerical measure is the number of defective parts produced per day.*
- *The subpopulations are patients who have taken a new drug and those who have taken a placebo, and the numerical measure is the recovery rate from a particular disease.*
- *The subpopulations are undergraduates with various majors (business, English, history, and so on), and the numerical measure is the starting salary after graduating.*

2. Relationships Among Categorical and Numerical Variables...

Stacked

	A	B
1	Gender	Salary
2	Male	81600
3	Female	61600
4	Female	64300
5	Female	71900
6	Male	76300
7	Female	68200
8	Male	60900
9	Female	78600
10	Female	81700
11	Male	60200
12	Female	69200
13	Male	59000
14	Male	68600
15	Male	51900
16	Female	64100
17	Male	67600
18	Female	81100
19	Female	77000
20	Female	58800
21	Female	87800
22	Male	78900

- There are two possible data formats, **stacked** and **unstacked**.
- The data are stacked if there are two long variables, such as gender and salary. The idea is that the male salaries are stacked in with the female salaries.
- One will occasionally see data in unstacked format, when there are two short variables, such as male salary and female salary.
- To understand the relationship, box plot is used for each category.

Unstacked

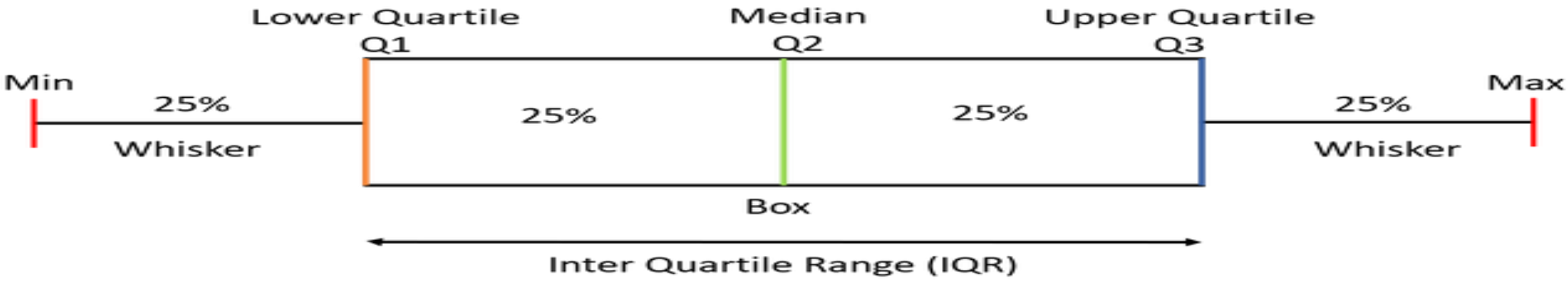
	A	B
1	Female Salary	Male Salary
2	61600	81600
3	64300	76300
4	71900	60900
5	68200	60200
6	78600	59000
7	81700	68600
8	69200	51900
9	64100	67600
10	81100	78900
11	77000	
12	58800	
13	87800	

Narrow, or stacked data is presented with one column containing all the values and another column listing the context of the value.

Wide, or unstacked data is presented with each different data variable in a separate column.

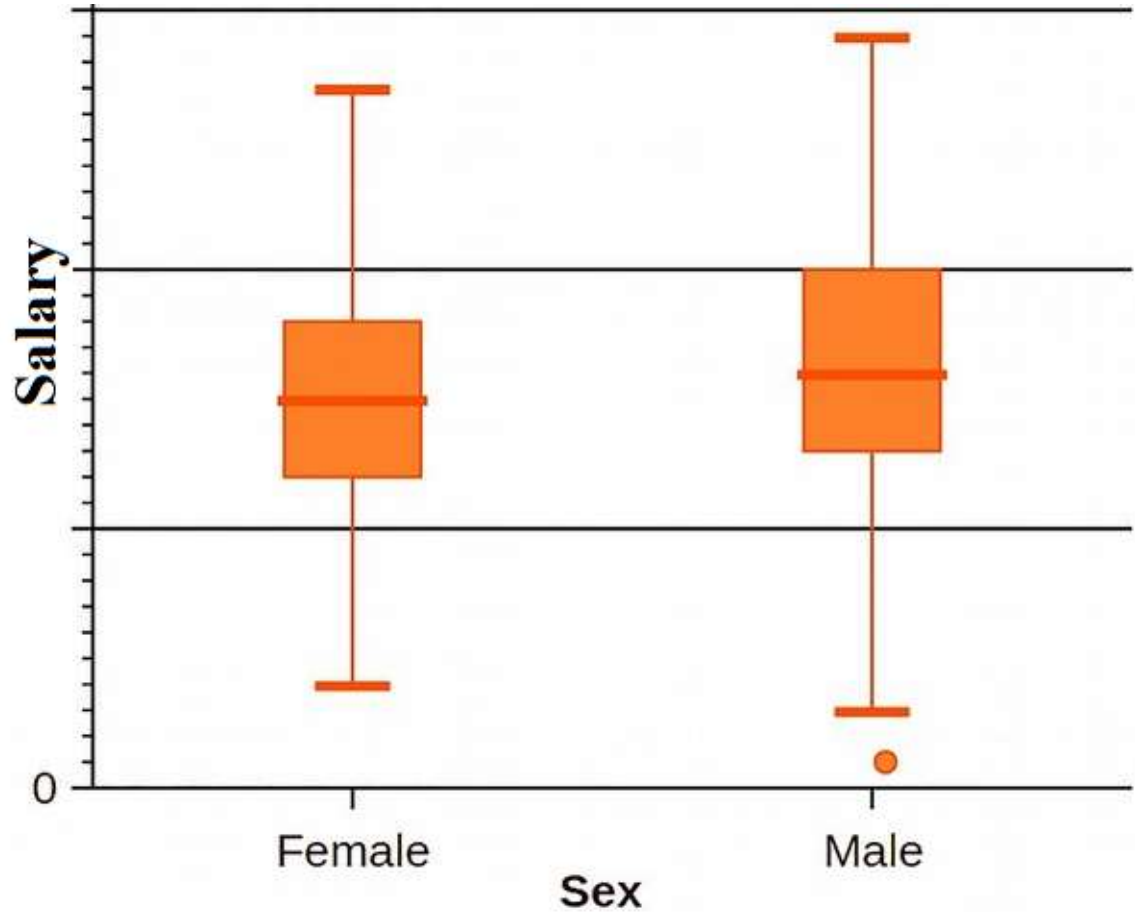
2. Relationships Among Categorical and Numerical Variables...

Box Plot:



A box plot gives a five-number summary of a set of data which is:

- Minimum – It is the minimum value in the dataset excluding the outliers
- First Quartile (Q1) – 25% of the data lies below the First (lower) Quartile.
- Median (Q2) – It is the mid-point of the dataset. Half of the values lie below it and half above.
- Third Quartile (Q3) – 75% of the data lies below the Third (Upper) Quartile.
- Maximum – It is the maximum value in the



2. Relationships Among Categorical and Numerical Variables...

Class Exercise

Find the relationships of salary between male and female of the sample by illustrating with the box plot.

1	Gender	Salary
2	Male	81600
3	Female	61600
4	Female	64300
5	Female	71900
6	Male	76300
7	Female	68200
8	Male	60900
9	Female	78600
10	Female	81700
11	Male	60200
12	Female	69200
13	Male	59000
14	Male	68600
15	Male	51900

2. Relationships Among Categorical and Numerical Variables...

Data of baseball salaries

Name	Team	Position	Salary
Justin Verlander	Detroit Tigers	Pitcher	$2,80,00,000
Zack Greinke	Los Angeles Dodgers	Pitcher	$2,70,00,000
Josh Hamilton	Los Angeles Angels	Outfielder	$2,50,00,000
Cliff Lee	Philadelphia Phillies	Pitcher	$2,50,00,000
Felix Hernandez	Seattle Mariners	Pitcher	$2,48,57,142
Albert Pujols	Los Angeles Angels	First baseman	$2,40,00,000
Robinson Cano	Seattle Mariners	Second baseman	$2,40,00,000
Clayton Kershaw	Los Angeles Dodgers	Pitcher	$2,40,00,000
Cole Hamels	Philadelphia Phillies	Pitcher	$2,35,00,000
Mark Teixeira	New York Yankees	First baseman	$2,31,25,000
Joe Mauer	Minnesota Twins	First baseman	$2,30,00,000
CC Sabathia	New York Yankees	Pitcher	$2,30,00,000
Miguel Cabrera	Detroit Tigers	First baseman	$2,20,00,000
Masahiro Tanaka	New York Yankees	Pitcher	$2,20,00,000

- Do pitchers (or any other positions) earn more than others?
- Does one league pay more than the other, or do any divisions pay more than others?
- How does the Yankee's payroll compare to the others?

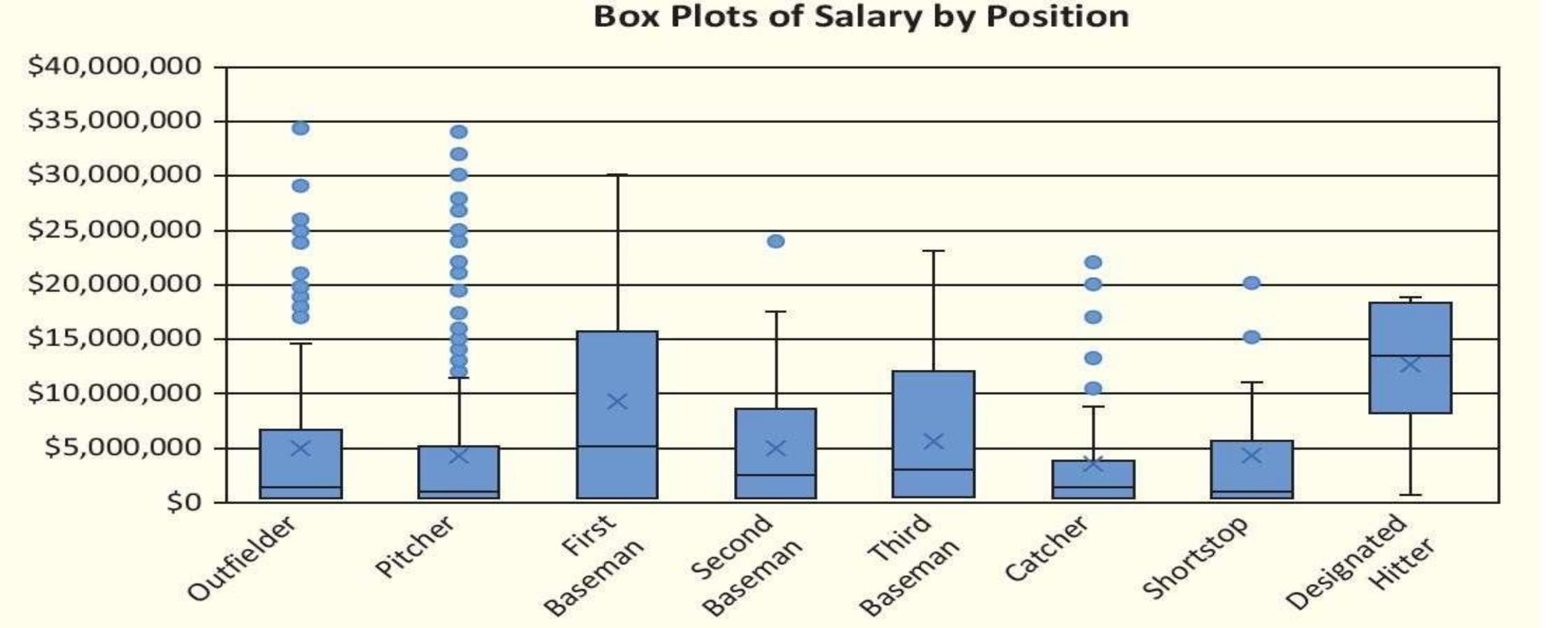
2. Relationships Among Categorical and Numerical Variables...

	Salary (Catcher)	Salary (Center Fielder)	Salary (Designated Hitter)	Salary (First Baseman)	Salary (Left Fielder)	Salary (Pitcher)	Salary (Right Fielder)	Salary (Second Baseman)	Salary (Shortstop)
One Variable Summary	Salary 2015 Data	Salary 2015 Data	Salary 2015 Data	Salary 2015 Data	Salary 2015 Data	Salary 2015 Data	Salary 2015 Data	Salary 2015 Data	Salary 2015 Data
Mean	$2690741.01	$4102195.20	$8364880.86	$8790404.78	$5582728.84	$3755287.58	$5715874.29	$3588563.07	$3693277.48
Std. Dev.	$3752384.07	$4875749.01	$6273012.23	$7961098.01	$6757322.64	$5136810.49	$5664412.46	$4756997.85	$5357972.10
Median	$1000000.00	$1650000.00	$6666666.00	$6500000.00	$2500000.00	$1312500.00	$2666666.00	$1800000.00	$850000.00
Minimum	$507500.00	$507500.00	$512500.00	$511000.00	$507500.00	$507500.00	$507500.00	$507500.00	$507500.00
Count	70	59	7	41	50	443	45	61	50
1st Quartile	$518290.00	$514500.00	$2950000.00	$2000000.00	$522500.00	$518000.00	$550000.00	$510900.00	$513543.00
3rd Quartile	$3100000.00	$6214285.00	$14250000.00	$14000000.00	$6900000.00	$5000000.00	$9500000.00	$5000000.00	$3175000.00

This table lists each of the requested summary measures for each of the **nine positions** in the data set.

If one wants to see salaries broken down **by team** or any other categorical variable, you can easily run this analysis again and choose a different Categorical variable.

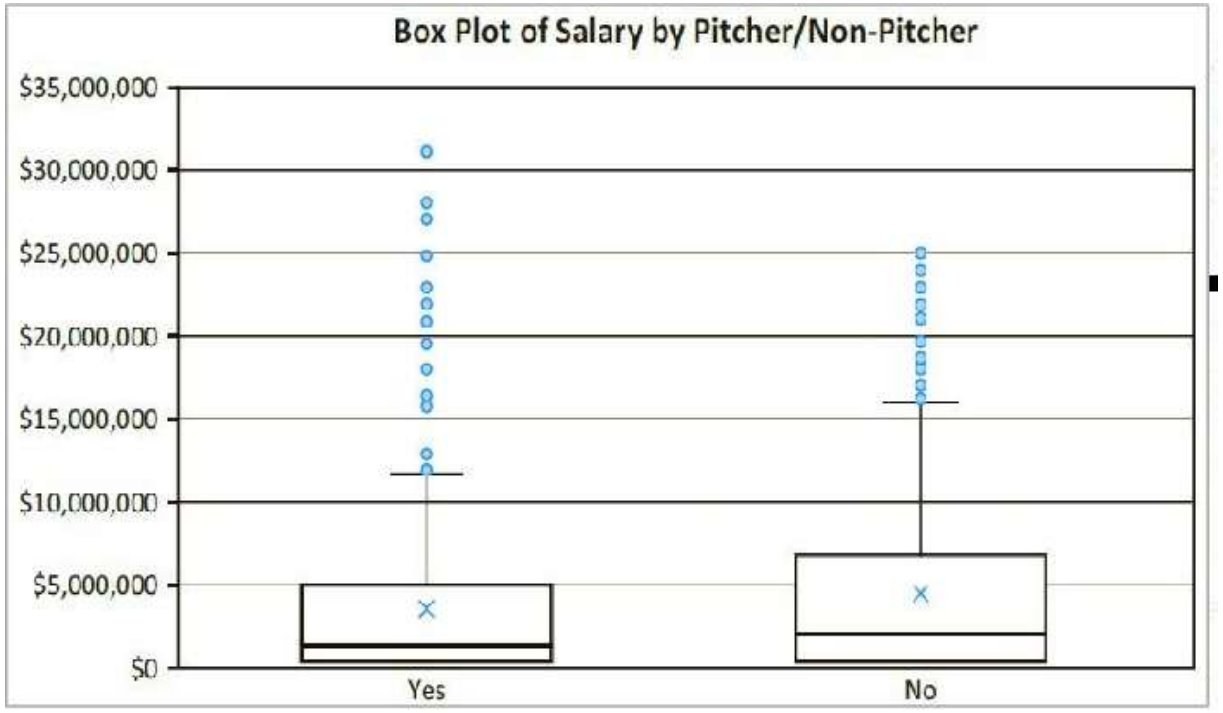
2. Relationships Among Categorical and Numerical Variables...

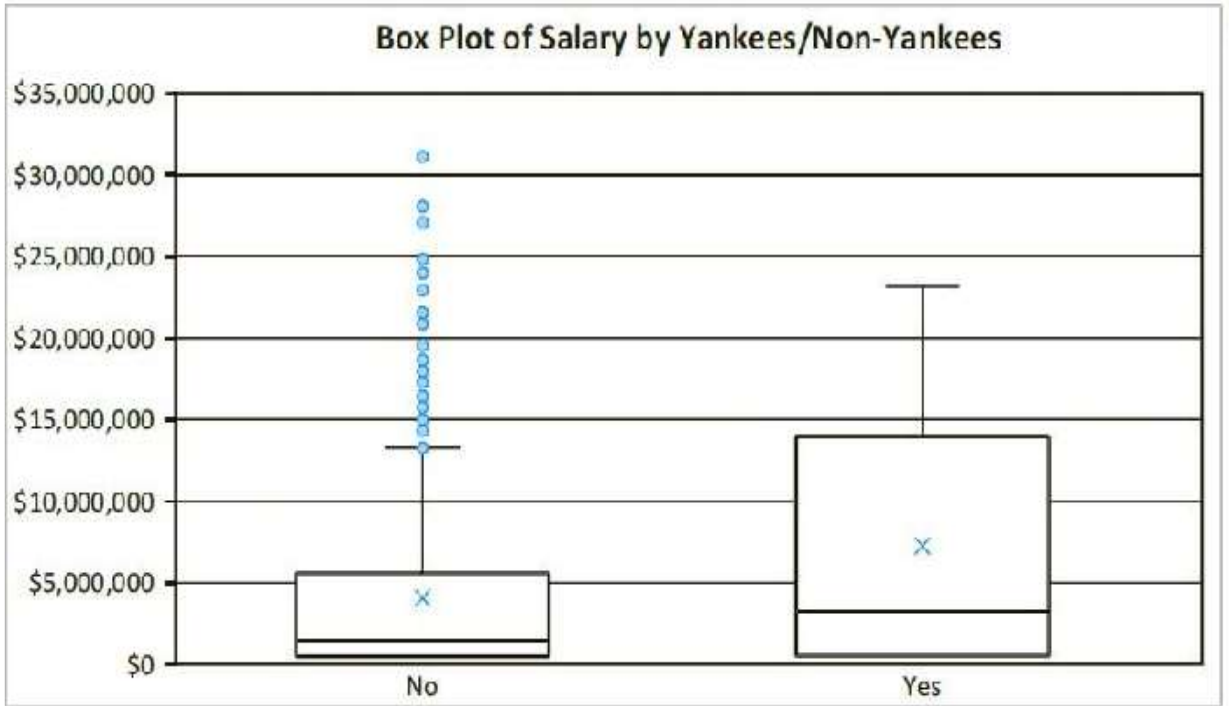


From these box plots, we can conclude the following:

- *Catchers salary is the lowest compared to others in the group.*
- *Pitchers have more outliers than others in the group. Disignated hitter have no outlier and their salaries are more*

2. Relationships Among Categorical and Numerical Variables...





Pitchers make somewhat less than other players, although there are many outliers in each group.

The Yankees payroll is indeed much larger than the payrolls for the rest of the teams. In fact, it is so large that its stars' salaries aren't even considered outliers relative to the rest of the team.

2. Relationships Among Categorical and Numerical Variables...

- Out of all the correlation coefficients the coefficient required to find for categorical vs numerical is the trickiest one.
- Some of the methods to understand a continuous and categorical are significantly correlated —

  - **Point biserial correlation:** It is a special case of the Pearson. It is used when one wants to measure the relationship between a continuous variable and a dichotomous variable, or one that has two values (i.e. male/female, yes/no, true/false)
  - **Logistic regression:** Logistic regression is a supervised learning used to calculate or predict the probability of a binary (yes/no, True/False) categorical event on the basis of one numerical event.

[In both of the above methods it estimate/predict categorical variable from numeric/continuous variable]

  - **Kruskal Wallis H Test(Or parametric forms such as t-test or ANOVA):** This is the family of methods to estimate association between a continuous and discrete variable which rely on estimating/predicting the continuous variable through/from the categorical variable.

3. Relationships among numerical variables

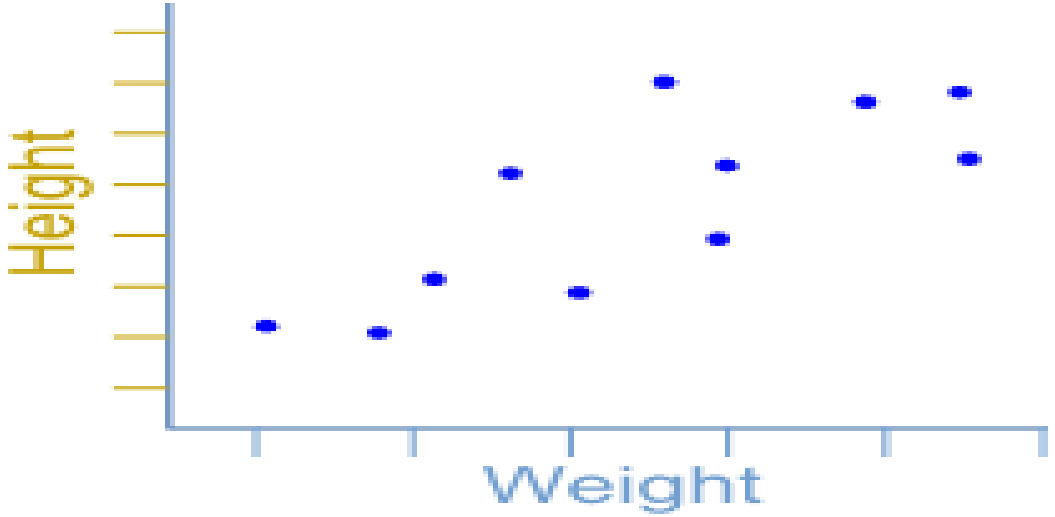
- To study relationships among numerical variables, a new type of chart, called a **scatterplot**, and two new summary measures, **covariance** and **correlation**, are used.
- However, they are appropriate only for truly numerical variables, not for categorical variables that have been coded numerically.

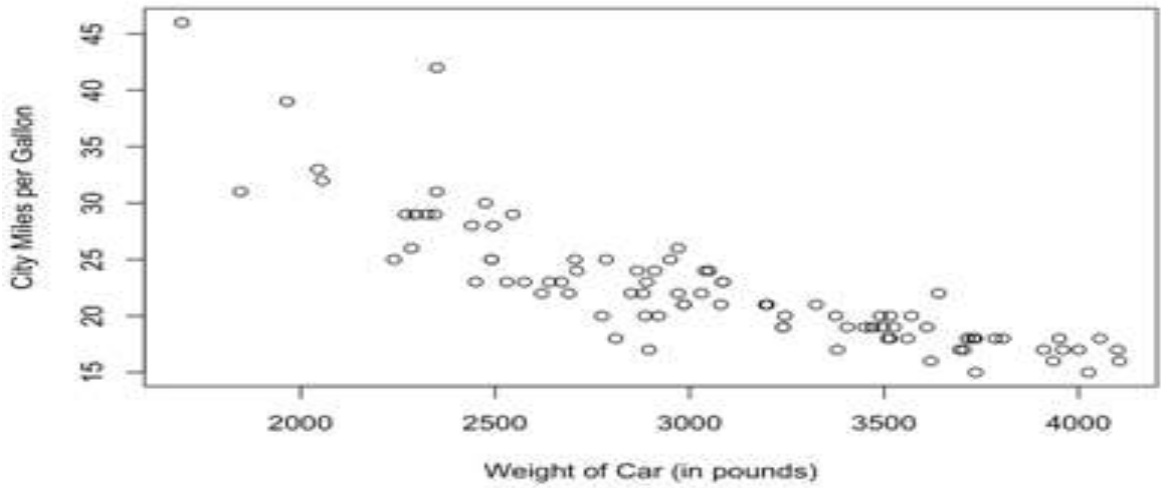
Correlation	Covariance
Indicates the direction and strength	Indicates the direction
It can be between – 1 to + 1	It can be between - ∞ to + ∞
Positive correlation coefficient close to 1 indicates a strong positive correlation and a value close to -1 indicates a strong negative correlation.	Positive covariance indicates an increase in one variable tends to increase the other variable and vice versa
There are multiple variations of a correlation coefficient i.e., the **Pearsons correlation coefficient,** the Spearman's Rho and the Kendall's Tau.	

3. Relationships among numerical variables...

Scatterplots

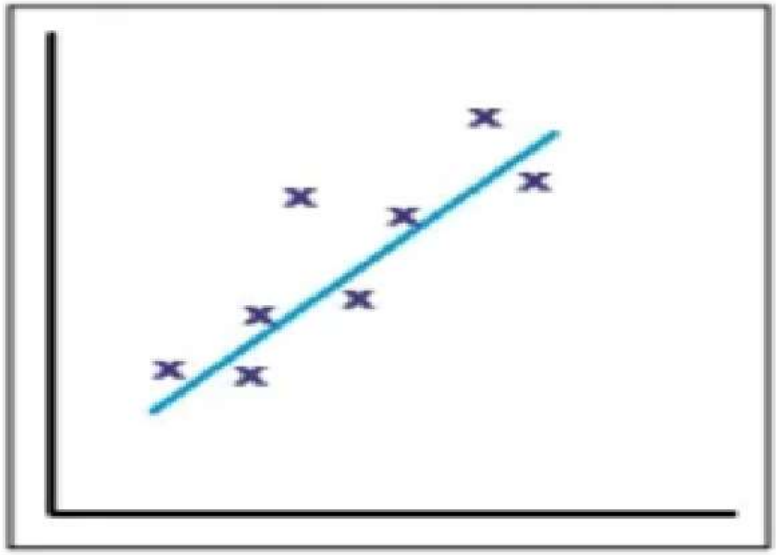
- A scatterplot is a scatter of points, where each point denotes the values of an observation for two selected variables.
- There must have the same number of observations, and the values for any observation should be naturally paired.
- It is a graphical method for detecting relationships between two numerical variables.
- The two variables are often labeled generically as X and Y, so a scatterplot is sometimes called an X-Y chart.
- The purpose of a scatterplot is to make a relationship (or the lack of it) apparent.





3. Relationships among numerical variables....

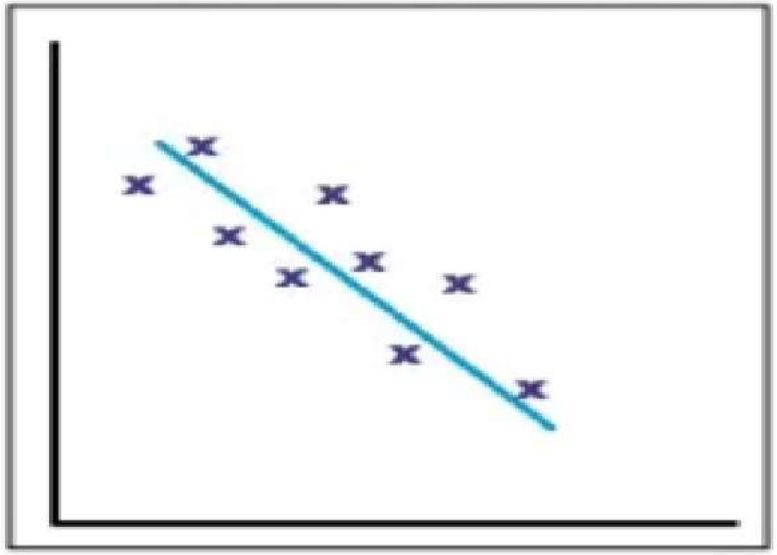
Scatter Plot with Positive correlation



If this straight line rises from left to right, the relationship is positive and the measures will be positive numbers.

This shows that if one variable increases then other increases.

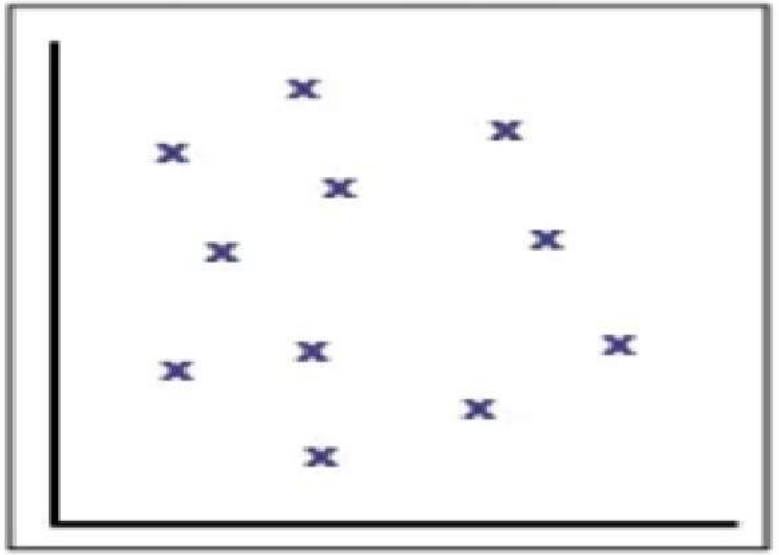
Scatter Plot with Negative correlation



If it falls from left to right, the relationship is negative and the measures will be negative numbers.

This shows that if one variable decreases then other increases.

Scatter Plot with No correlation



There is no pattern to the point.

This shows that there is no connection between the variables ,

3. Relationships among numerical variables...

Covariance: (How 2 things change together)

- A statistical measure that shows whether two variables are related by measuring how the variances change relation to each other. It is essentially an average of products of deviations from means

$$\text{Covar}(X, Y) = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

Here X_i and Y_i: Paired values of i^{th} observation

n: Number of observations.

$Covar(X, Y)$: Covariance.

- If two variables increases/decreases together, covariance will be a large positive value and the relationship is called positive. If one variabe increases and other decreases and vice versa then covariance is large negative value and teh relationship is called negative. If two variables are unrelated the covariance may be a small number.
- However, howmuch large is large and howmuch small is small is undefined, or it is usually difficult to provide any guideline about howmuch large covariance shows a strong relationship and howmuch small covariance shows no relationship.
- Correlation can overcome this drawback to certain extent.

3. Relationships among numerical variables...

Calculate the covariance between two stocks returns (%)

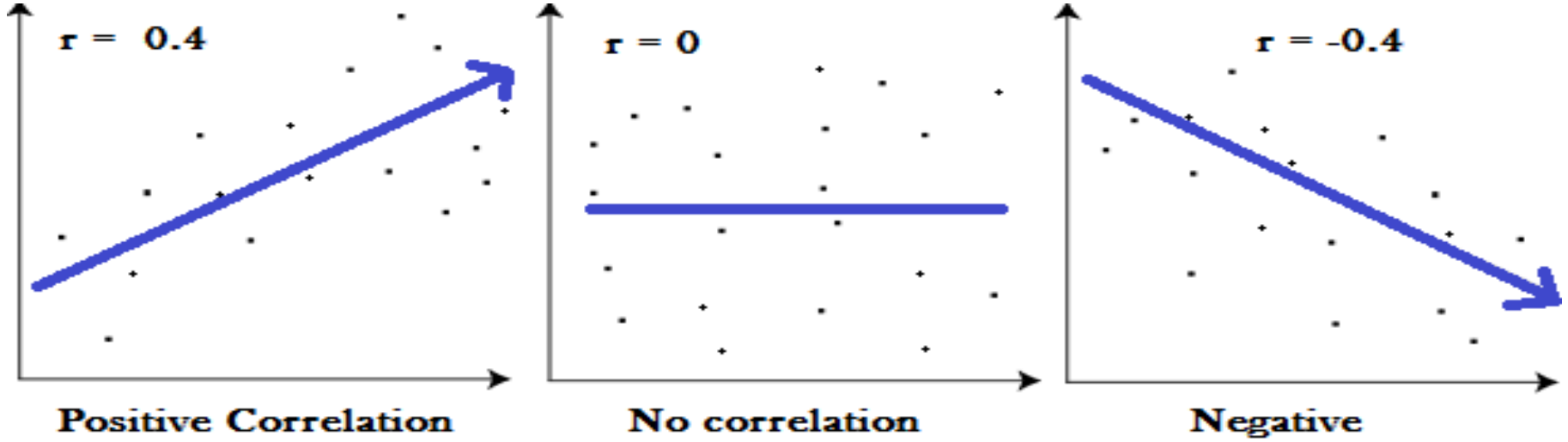
Quarter	Stock X	Stock Y
1	4	6
2	2	3
3	5	7
4	-2	0

Quarter	Stock X	Stock Y	(Xi- $\overline{X}$)	(Yi- $\overline{Y}$)	(Xi- $\overline{X}$) (Yi- $\overline{Y}$)
1	4	6	1.75	2	3.5
2	2	3	-0.25	-1	0.25
3	5	7	2.75	3	8.25
4	-2	0	-4.25	-4	17
Mean=>	2.25	4			∑=29

Covariance = $S_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{29}{4-1} = 9.67$

3. Relationships among numerical variables...

- **Correlation:** a measure of how two variables change in relation to each other, but it goes one step further than covariance in that correlation tells how strong the relationship is.
- For example, height and weight are related; taller people tend to be heavier than shorter people. The relationship isn't perfect. People of the same height vary in weight.
- Correlation can tell you just how much of the variation in peoples' weights is related to their heights.
- The main result of a correlation is called the correlation coefficient (or "r"). It ranges from -1.0 to +1.0. The closer r is to +1 or -1, the more closely the two variables are related.



3. Relationships among numerical variables...

- If **r is close to 0**=> it means there is no relationship between the variables.
- If **r is positive** =>it means that as one variable gets larger the other gets larger.
- If **r is negative** => it means that as one gets larger, the other gets smaller (often called an "inverse" correlation).
- Values between **0.7 and 1.0** (**−0.7 and −1.0**) indicate a strong positive (negative) relationship.

Example

The local ice cream shop keeps track of how much ice cream they sell versus the temperature on that day, here are their figures for the last 12 days.

Ice Cream Sales vs. Temperature												
Temp	14.2	16.4	11.9	15.2	18.5	22.1	19.4	25.1	23.4	18.1	22.6	17.2
Sales	215	325	185	332	406	522	412	614	544	421	445	408

Draw a scatter plot

3. Relationships among numerical variables...

First approach for Correlation Coefficient

$$\text{Correlation (X, Y)} = \frac{\text{COV(x,y)}}{\sqrt{\text{VAR(x)} * \text{VAR(y)}}} \quad \textbf{OR} \quad r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Let us call the two sets of data "x" [Temperature] and "y" [Ice Cream]

1. **Step 1:** Find the mean of x, and the mean of y
2. **Step 2:** Subtract the mean of x from every x value [let "a"], do the same for y [Let "b"]
3. **Step 3:** Calculate: a*b, a^2 and b^2 for every value
4. **Step 4:** Sum up a*b, sum up a^2 and sum up b^2
5. **Step 5:** Divide the sum of a*b by the square root of [(sum of a^2) × (sum of b^2)]

3. Relationships among numerical variables...

SOLUTION:

2 *Subtract Mean*

3 *Calculate ab, a² and b²*

Temp °C	Sales	"a"	"b"	a×b	a²	b²
14.2	$215	-4.5	-$187	842	20.3	34,969
16.4	$325	-2.3	-$77	177	5.3	5,929
11.9	$185	-6.8	-$217	1,476	46.2	47,089
15.2	$332	-3.5	-$70	245	12.3	4,900
18.5	$406	-0.2	$4	-1	0.0	16
22.1	$522	3.4	$120	408	11.6	14,400
19.4	$412	0.7	$10	7	0.5	100
25.1	$614	6.4	$212	1,357	41.0	44,944
23.4	$544	4.7	$142	667	22.1	20,164
18.1	$421	-0.6	$19	-11	0.4	361
22.6	$445	3.9	$43	168	15.2	1,849
17.2	$408	-1.5	$6	-9	2.3	36
18.7	**$402**			**5,325**	**177.0**	**174,757**

1 *Calculate Means*

4 *Sum Up*

5 $$\frac{5{,}325}{\sqrt{177.0 \times 174{,}757}} = 0.9575$$

3. Relationships among numerical variables…

Second approach for Correlation Coefficient

- The following formula is another one (normally used) to estimate the correlation coefficients between two variables X and Y.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{\left[n\Sigma x^2 - (\Sigma x)^2\right]\left[n\Sigma y^2 - (\Sigma y)^2\right]}}$$

- x is the independent variable and y is the dependent variable.
- n is the number of observations
- r, the computed value is known as the correlation coefficients .

3. Relationships among numerical variables...

Example:

Company	Sales in 1000s (Y)	No. of agents in 100s (X)	X^2	Y^2	XY
A	25	8			
B	35	12			
C	29	11			
D	24	5			
E	38	14			
F	12	3			
G	18	6			
H	27	8			
I	17	4			
J	30	9			

$n = 10$

$\sum X = 80$ & $\sum Y = 255$

$\sum XY = 2289$

$\sum X^2 = 756$ & $\sum Y^2 = 7097$

$(\sum X)^2 = 6400$ & $(\sum Y)^2 = 65025$

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{\left[n\Sigma x^2 - (\Sigma x)^2\right]\left[n\Sigma y^2 - (\Sigma y)^2\right]}}$$

Hence, r = 0.95

3. Relationships among numerical variables...

Class Exercise

Find the correlation coefficients of the below sample.

Subject	Age (x)	Glucose Level (y)
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81

3. Relationships among numerical variables...

Third approach for Correlation Coefficient when instead of population size and individual occurance of the (x, y) given is the probability of (x, y) or frequency or percentage of occurance

$$\text{Correlation } (x, y) = \frac{\text{COV}(x,y)}{\sqrt{\text{VAR}(x) * \text{VAR}(y)}}$$

P(x, y)	x	y
0.15	10	2
0.35	6	4
0.25	3	3
0.25	15	5

Here probability of occurance of (x, y) is given instead of individual list of (x, y). Thus n=1

Same way frequency of (x, y) can be given where n will be sum of the total frequency or percentage of (x, y) can be given where n will be 100

3. Relationships among numerical variables...

Third approach for Correlation Coefficient

$$\text{Correlation } (x, y) = \frac{\text{COV(x,y)}}{\sqrt{\text{VAR(x)} * \text{VAR(y)}}}$$

Average value of x and y now can be considered as expected x (**E(x)**) and expected y (**E(y)**)

$$E(x) = \frac{\sum x_i}{n} = \frac{\sum(x_i * (P(x_i, y_i))}{1}$$

$$E(y) = \frac{\sum y_i}{n} = \frac{\sum(y_i * (P(x_i, y_i))}{1}$$

$$Cov(x, y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{\sum(x_i - E(x))(y_i - E(y))P(x_i, y_i)}{1}$$

In probability, n is1

$$\sqrt{var_x} = \sigma_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}} = \sqrt{\frac{\sum(x_i - E(x))^2 * P(x_i, y_i)}{1}}$$

$$\sqrt{var_y} = \sigma_y = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n}} = \sqrt{\frac{\sum(y_i - E(y))^2 * P(x_i, y_i)}{1}}$$

Chapter 2: **Statistical Concepts**

- **Distribution of a single variable:**
  - Basic Concepts (population and sample, data sets, variables, and observations, types of data)
  - Descriptive measures for categorical variables
  - Descriptive measures for numerical variables
  - Outliers and Missing values
- **Finding relationships among variables:**
  - Categorical variables
  - Categorical variables and a Numerical variable
  - Numerical variables
- **Sampling and distributions:**
  - Terminology
  - Estimation
  - Confidence Interval estimation
  - Sampling distributions
  - Confidence interval,
  - Hypothesis testing, Chi-square test for independence

Sampling and distributions

- **Population:** It is the entire group of observations from which one wants to draw the conclusions about. The analysis report of the populations is the true and accurate representation of opinion with no margin of error
- **Sample:** It is the specific group/subset of the population that represent the whole population. The analysis report of the sample has some margin of error and confidence interval, and can be used for population sample's report after further factoring in the margin of error and confidence interval.
- **Sampling:** Sampling is a technique of selecting individual members or a subset of the population to make statistical inferences from them and estimate the characteristics of the whole population.

Sampling and distributions...

Measurable quality: Mean, median, mode, standard deviation, variance, covariance, correlation

The measurable quality of

- of population is called as parameter. [numerical or measurable actual element]
- of sample is called as statistics.[descriptive/probable component]

	PARAMETER	STATISTICS
MEAN	$\mu = \frac{\sum X_i}{N}$	$\bar{x} = \frac{\sum x_i}{n}$
VARIANCE	$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$	$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$
STANDARD DEVIATION	$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$	$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$

Distributions

There are three distinct types of distribution of data

1. **Population Distribution:** It characterizes the distribution of elements of a population.
2. **Sample Distribution:** It characterizes the distribution of elements of a sample drawn from a population.
3. **Sampling Distribution:** It describes the expected behavior of a large number of simple random samples drawn from the same population. or It is a theoritical **probability distribution of a statistics** obtained from large number of samples obtained from a population. Its primary purpose is to establish representative results from small samples of a comparatively larger population. It simplifies the process of making inferences, or conclusions, about large amounts of data.

When the population size is large enough, the value of a statistic from random samples can combiningly inform the statistic value of the entire group or population.

These distributions help you understand how a sample statistic varies from sample to sample.

Sampling Distributions

- Sampling distributions are essential for ***inferential statistics*** because they allow you to understand a specific sample statistic in the broader context of other possible values.
- We are moving **from descriptive statistics to inferential statistics.**
- ***Inferential statistics*** *allow the researcher to come to conclusions about a population on the basis of* ***descriptive statistics about a sample.***

The **sampling distribution** of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population.

For example:

- Your sample says that a candidate gets support from 47%.
- Inferential statistics allow you to say that the candidate gets support from 47% of the population with a margin of **error of +/- 4%**.
- This means that the support in the population is likely somewhere between 43% and 51%.

Factors that influence sampling distribution

- Each random sample selected may have a different value assigned to the statistic being studied. For example, if you randomly sample data three times and determine the mean, or the average, of each sample, all three means are likely to be different and fall somewhere along the graph. **That's variability**.
- There are **three primary factors** that **influence the variability** of a sampling distribution. They are:
  - The number of observation in a population: The symbol for this variable is "N." It is the measure of observed activity in a given group of data.
  - The number of observation in the sample: The symbol for this variable is "n." It is the measure of observed activity in a random sample of data that is part of the larger grouping.
  - The method of choosing the sample: How you chose the samples can account for variability in some cases.

Types of sampling distribution

- There are three standard types of sampling distributions in statistics:

1. Sampling distribution of mean

The most common type of sampling distribution is the mean. It focuses on calculating the mean of every sample group chosen from the population and plotting the data points. The graph shows a normal distribution where the **center** is the mean of the sampling distribution, which represents the mean of the entire population.

2. Sampling distribution of proportion

This sampling distribution focuses on proportions in a population. You select samples and calculate their proportions. The means of the sample proportions from each group represent the proportion of the entire population

3. T-distribution

A T-distribution is a sampling distribution that involves a small population or one where you don't know much about it. It is used to estimate the mean of the population and other statistics such as confidence intervals, statistical differences and linear regression. The T-distribution uses a t-score to evaluate data that wouldn't be appropriate for a normal distribution.

Chapter 2: **Statistical Concepts**

- **Distribution of a single variable:**
  - Basic Concepts (population and sample, data sets, variables, and observations, types of data)
  - Descriptive measures for categorical variables
  - Descriptive measures for numerical variables
  - Outliers and Missing values
- **Finding relationships among variables:**
  - Categorical variables
  - Categorical variables and a Numerical variable
  - Numerical variables
- **Sampling and distributions:**
  - Terminology
  - Estimation
  - Confidence Interval estimation
  - Sampling distributions
  - Confidence interval,
  - Hypothesis testing, Chi-square test for independence

Estimation and Sampling Distributions

- Estimation is the process of determining a likely value for a population parameter (eg, the true population mean or proportion) based on a random sample.
- But, how good are sample statistics at estimating population parameters?
  - How accurately does the sample mean ($\bar{x}$), sample variance (s^2), and sample standard deviation (s) estimate the population mean (μ), population variance (σ^2), and population standard deviation (σ)?
- Statistical estimation procedures provide estimates of population parameter with a desired degree of confidence. The degree of confidence can be controlled in part,
  - **by the size the sample** (larger sample greater accuracy of the estimate)
  - **by the type of the estimate made.**

Estimation and Sampling Distributions...

The statistical estimation of the population parameter is further divided into two types

(i) ***Point Estimation***

(ii) ***Interval Estimation***

Point Estimation

- The objective of point estimation is to obtain a single number from the sample which will represent the unknown value of the population parameter.
- Population parameters (population mean, variance, SD etc let θ is denoted by $\hat{\theta}$) are estimated from the corresponding sample statistics (sample mean, variance, SD etc).
- A statistic used to estimate a parameter is called a point estimator or simply an estimator, the actual numerical value obtained by **estimator** is called an **estimate**.

Interval Estimation

- When estimation of population parameter is done with an interval/range of statistics values with the expected value believe to lie within a certain degree of confidence then it is called Interval Estimation.

Point Estimation of Population Mean from Sampling Distribution

Q: Suppose a population size of N=4, incomes of four business firms and it is required to find the average(mean) return of these firms. The incomes (in Lakhs) are 100,200, 300 and 400. Considering sample size 2 draw the sampling distribution graph for mean and justify relation between population mean (μ) and sampling mean ($\bar{x}$).

Solution: N=4 opulation mean (μ)= (100+200+300+400)/4 = 250

For sample size n, number of possible samples = C(4,2) = 6

Sample	Sample elements X_i	Sample means $\bar{X}$
1	100,200	150
2	100,300	200
3	100,400	250
4	200,300	250
5	200,400	300
6	300,400	350

Now, from the table it can be observed that each sample has a different mean [with the exception of third and fourth samples]. Therefore four of the six samples will result in some error in the estimation process. This sampling error is the difference between the population mean and the sample mean.

'mean of the sample means' or the grand mean =

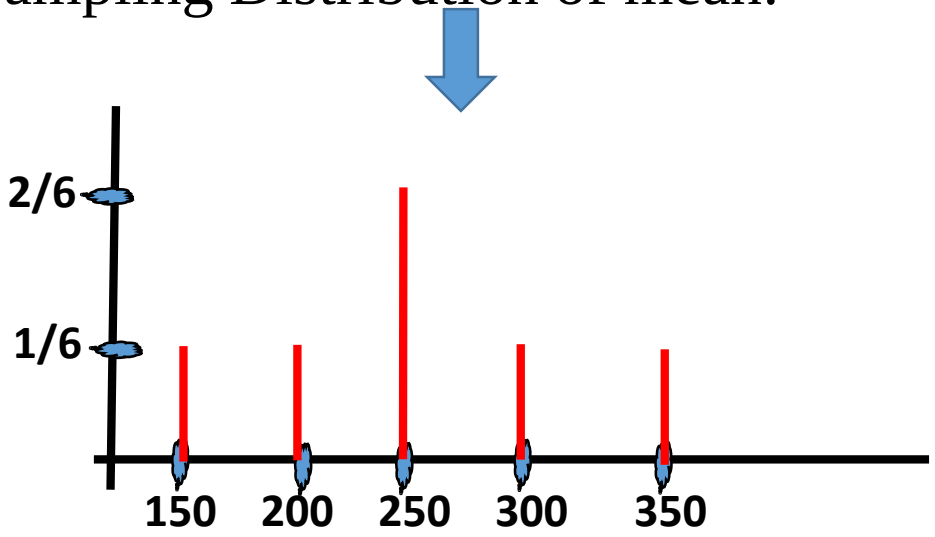
$$\bar{\bar{X}} = \frac{\sum_{i=1}^{k} \bar{x}_i}{k} = (150+200+250+250+300+350)/6 = 250$$

For all possible sample of size n from a population if each sample mean will be calculated, then the mean of those sample means ($\bar{\bar{x}}$) would equal the population mean.

Point Estimation of Population Mean from Sampling Distribution

- To construct sampling distribution of mean consider the possible sample means and calculate with their probability.
- Let assume that each sample is equally likely to be chosen. Then the probability of selecting a sample is 1/6
- Sample means and their respective possibilities:
- Sampling Distribution of mean:

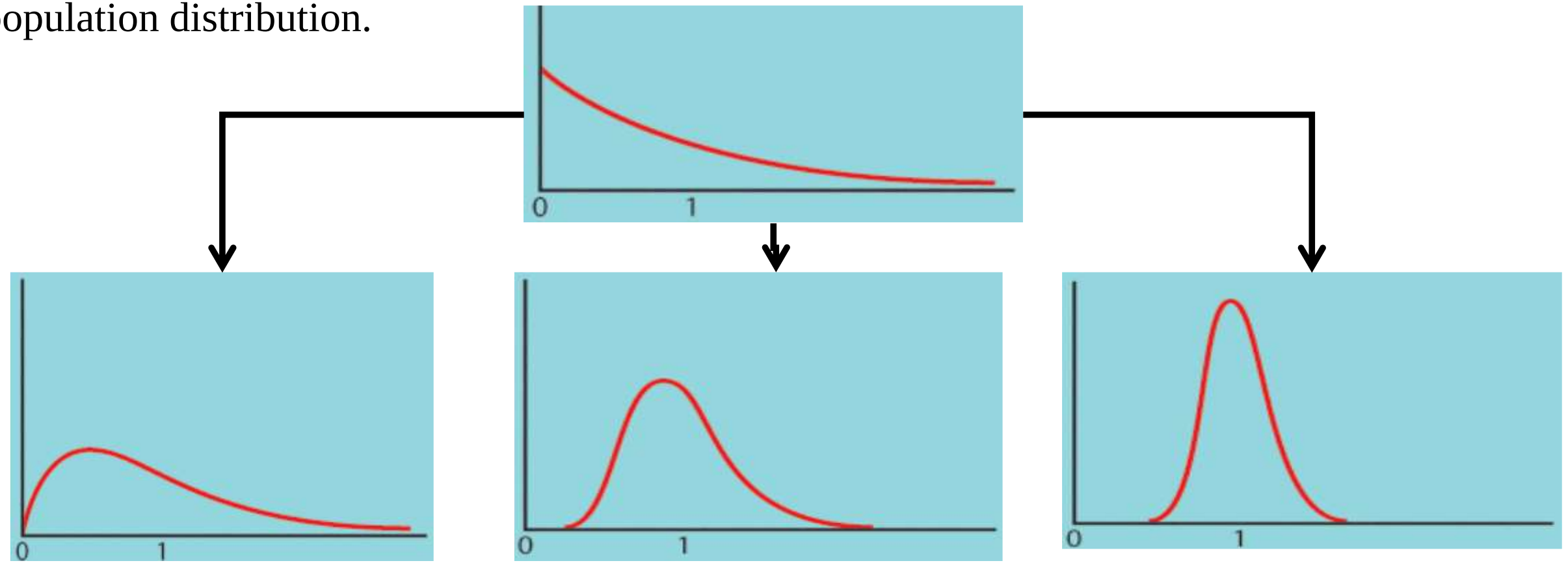
Sample mean $\bar{X}$	Number of samples yielding $\bar{X}$	Probability $P(\bar{X})$
150	1	1/6
200	1	1/6
250	2	2/6
300	1	1/6
350	1	1/6
		Total 1



The difference between sample mean and population mean (i.e. μ-$\bar{x}_i$ 100,50,0,-50-100) are called as **sampling error**. These sampling errors of sample means are not consistently greater than or less than the population mean. **Thus the sample mean is an unbiased estimator of μ.**

Estimating Mean of Sampling Distribution...

- **For normally distributed populations:** When a variable in a **population is normally distributed**, then the sampling distribution of sample mean for all possible samples of size n is also normally distributed
- **The central limit theorem**: When randomly sampling (**when n is large enough**) from **any** population the sampling distribution is approximately normal irrespective of type of population distribution.



Sampling distribution of $\bar{x}$ for n = 2 **Sampling distribution of $\bar{x}$ for n = 10** **Sampling distribution of $\bar{x}$ for n = 25**

Point Estimation of Population Variance and Standard Deviation from Sampling Distribution

Let the parameter values are unknown and it is required to estimate from the sample statistics.

From few sample means the value of parameter mean can be estimated (with very less error).

However, sample variance and sample standard deviation can't guide properly to estimate parameter variance and standard deviation.

Let size of population N= 25 [Quiz mark (FM: 10) of 25 students]

Parameters:

- Population Mean (μ): 5.4
- Population Variance (σ^2): 7.04
- Satandard Deviation (σ): 2.63

Sample Variance and Standard Deviation as Biased Estimators

Let sample size n=5 and 2 samples S1={10, 8, 5, 5, 3} and S2={9, 6, 5, 4, 2}

Let sample mean of S1 is M1 = 6.2

Let sample mean of S2 is M2 = 5.2

Sample Variance: $\frac{\sum(x_i-\bar{x})^2}{n}$

For sample S_1 variance = s_1^2

$= \frac{(10-6.2)^2+(8-6.2)^2+(5-6.2)^2+(5-6.2)^2+(3-6.2)^2}{5}$

= 6.16 (this is < population variance 7.04)

For sample S_2 variance = s_2^2

$= \frac{(9-5.2)^2+(6-5.2)^2+(5-5.2)^2+(4-5.2)^2+(2-5.2)^2}{5}$

=6.36 (this is < population variance7.04)

In both the samples, the variances are less than the population variance. If the process of finding variances of other samples will be tried then the same (smaller valued) will be observed.

Thus, sample variance underestimate the population variance or are biased estimation.

Solution: When sample variances are used to estimate the population variances the correction for the underestimation can be done by increasing the value of sample variance by considering the divisor as n-1 instead of n

Sample Variance and Standard Deviation as Biased Estimators...

Hence, Sample Variance as estimation of population variance is: $\frac{\sum(x_i-\bar{x})^2}{n-1}$

For sample S_1 variance = $\hat{s}_1^2$

$= \frac{(10-6.2)^2+(8-6.2)^2+(5-6.2)^2+(5-6.2)^2+(3-6.2)^2}{5-1}$

= 7.7 (Now this is > 7.04)

For sample S_1 variance = $\hat{s}_2^2$

$= \frac{(9-5.2)^2+(6-5.2)^2+(5-5.2)^2+(4-5.2)^2+(2-5.2)^2}{5-1}$

=7.95 (Now this is > 7.04)

In both the samples, the variances are now larger than the population variance.

But that's OK, the overestimation is better than underestimation.

Sample Variance and Standard Deviation as Biased Estimators....

For the same sample size n=5 and 2 samples S1={10, 8, 5, 5, 3} and S2={9, 6, 5, 4, 2}

Standard Deviation: $\sqrt{\frac{\sum(x_i-\bar{x})^2}{n}}$ OR $\sqrt{variance}$

For sample S_1 Standard Deviation = $s_1 = \sqrt{var}$

$= \sqrt{6.16} = 2.482$

(this is < population SD 2.636)

For sample S_2 Standard Deviation = $s_2 = \sqrt{var}$

$= \sqrt{6.36} = 2.522$

(this is < population SD 2.636)

In both the samples, the SD are less than the population SD. If the process of finding SD of other samples will be tried then the same (smaller valued) will be observed.

Thus, sample SD underestimate the population's SD or are biased estimation.

Solution: When sample SD are used to estimate the population SD the correction for the underestimation can be done by increasing the value of sample SD by considering the divisor as n-1 instead of n

Sample Variance and Standard Deviation as Biased Estimators....

For the same sample size n=5 and 2 samples S1={10, 8, 5, 5, 3} and S2={9, 6, 5, 4, 2}

Standard Deviation: $\sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}$ OR $\sqrt{sample\ variance\ used\ for\ estimation}$

For sample S_1 Standard Deviation = $\widehat{s_1}$

$= \sqrt{7.7} = 2.775$

(this is > population SD 2.636)

For sample S_2 Standard Deviation = $s_2 = \sqrt{var}$

$= \sqrt{7.95} = 2.822$

(this is > population SD 2.636)

In both the samples, the SDs are now larger than the population SDs.

But that's OK, the overestimation is better than underestimation.

Conclusion: If measuring variability is within the sample/within the population => use sample formula & population formula. If measuring variability (vav/SD) of sample is used to estimate population parameter => divide n-1

Standard Error of the Sampling Distribution

- The sampling distribution has a standard deviation/difference from the population distribution and is called as “Standard Error”.
- The mean of the sampling distribution mean is the population mean but the variance and standard deviation are smaller than the population variance and SD.
- *Sampling error or standard error is mainly the variation of sample mean around the population mean.*
- In sampling theory:
  - **68%** of all sample means will lie between + & **- one standard error** from population mean.
  - **95%** of all sample mean will lie betw pulation mean

$$\mathrm{SE} = \frac{\sigma}{\sqrt{n}}$$

SE = standard error of the sample

σ = sample standard deviation

n = number of samples

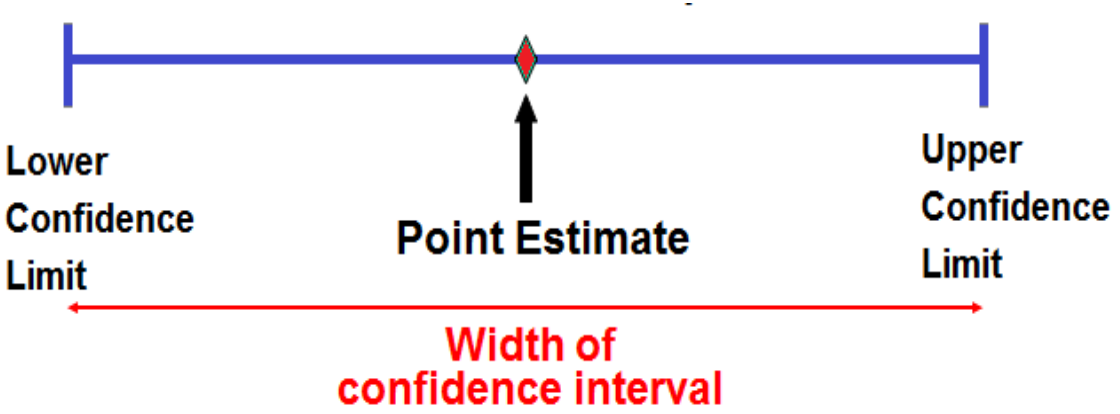
Chapter 2: **Statistical Concepts**

- **Distribution of a single variable:**
  - Basic Concepts (population and sample, data sets, variables, and observations, types of data)
  - Descriptive measures for categorical variables
  - Descriptive measures for numerical variables
  - Outliers and Missing values
- **Finding relationships among variables:**
  - Categorical variables
  - Categorical variables and a Numerical variable
  - Numerical variables
- **Sampling and distributions:**
  - Terminology
  - Estimation
  - Confidence Interval estimation 
  - Sampling distributions
  - Confidence interval,
  - Hypothesis testing, Chi-square test for independence

Interval Estimation for Population Parameter

- A point estimator (such as sample mean), which can not be expected to be exactly equal to the population parameter because the mean of a sample taken from a population may assume different values for different samples.
- Therefore we estimate an interval/ range of values (set of values) within which the population parameter is expected to lie with a certain degree of confidence.
- This range of values used to estimate a population parameter is known as interval estimate or estimate by a confidence interval, and is defined by two numbers, between which a population parameter is expected to lie.
- Note: The bigger your sample size, the more narrow the confidence interval will be.

For example, $a < \bar{x} < b$ is an interval estimate of the population mean μ , indicating that the population mean is greater than a but less than b. The purpose of an interval estimate is to provide information about how close the point estimate is to the true parameter.



Confidence Interval estimation

- The confidence interval is the range of values that you expect your estimate to fall between a certain percentage of the time if you run your experiment again or re-sample the population in the same way.

- **A confidence interval is the mean of your estimate plus and minus the variation in that estimate. This is the range of values you expect your estimate to fall between if you redo your test, within a certain level of confidence.**

$$\text{confidence interval} = \text{sample statistic} \pm \text{a multiple of the standard error of the statistic}$$

- **Confidence, in statistics, is another way to describe probability. For example, if you construct a confidence interval with a 95% confidence level, you are confident that 95 out of 100 times the estimate will fall between the upper and lower values specified by the confidence interval.**

Numeric on Confidence Interval estimation

Q: find the confidence interval of the sample of 10 test scores: {80, 95, 90, 90, 95, 75, 75, 85, 90 and 80} with a 97% confidence level

Solution:

Step-1: Find Sample Mean (μ)

$$\mu = (80+95+90+90+95+75+75+85+90+80)/10 = 855/10 = 85.5$$

Step-2: Find Standard Deviation (σ)

$$\sigma = [(80 - 85.5)^2 + (95 - 85.5)^2 + (90 - 85.5)^2 + (90 - 85.5)^2 + (95 - 85.5)^2 + (75 - 85.5)^2 + (75 - 85.5)^2 + (85 - 85.5)^2 + (90 - 85.5)^2 + (80 - 85.5)^2]/10$$

$$=(30.5 + 90.25 + 20.25 + 20.25 + 90.25 + 90.25 + 90.25 + 0.25 + 20.25)/10 = 45.25$$

Step-3: Find margin Error i.e. a multiple of standard error $\left(z\frac{\sigma}{\sqrt{n}}\right)$

For 97% of confidence interval z value is 0.97.

$$\text{Margin Error} = 0.97*45.25/\sqrt{10} = 13.89$$

Step-4: Confidence Interval

$$\text{Mean} \pm \text{Margin Error} = 85.5 \pm 13.89 = \{99.39, 71.61\}$$

Chapter 2: **Statistical Concepts**

- **Distribution of a single variable:**
  - Basic Concepts (population and sample, data sets, variables, and observations, types of data)
  - Descriptive measures for categorical variables
  - Descriptive measures for numerical variables
  - Outliers and Missing values
- **Finding relationships among variables:**
  - Categorical variables
  - Categorical variables and a Numerical variable
  - Numerical variables
- **Sampling and distributions:**
  - Terminology
  - Estimation
  - Confidence Interval estimation
  - Sampling distributions
  - Confidence interval,
  - Hypothesis testing, Chi-square test for independence

Hypothesis testing

- Hypothesis testing is a formal procedure for investigating our ideas about the world using statistics. It is most often used by scientists to test specific predictions, called hypotheses, that arise from theories.
- ***A hypothesis test evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data. These two statements are called the null hypothesis and the alternative hypothesis.***
- These hypotheses can then be tested to find out whether differences or relationships observed in the sample are statistically significant in terms of the population.
- In order to do this, two complementary, contradictory hypotheses need to be formulated;
- the **null hypothesis** and the **alternative hypothesis** (or *research hypothesis*).

Hypothesis testing (cont..)

Null hypothesis (H_0)

This hypothes is always states that there is ***no difference*** or no relationship between variables in a population. Or a hypothesis that proposes that no statistical significance exists in a set of given observations

Predict there is no relation between two variables:

Example-1: *Application of Bio fertiliser has no effect on plant growth.*

Example-2: There is no association between age group and intent to get vaccinated

Alternative hypothesis (H_1)

Also known as the research hypothesis, this hypothesis always states the opposite of the null hypothesis; i.e. that ***there is a difference*** or relationship between variables in a population.

Predict there is relation between two variables:
Example-1: *Application of Bio fertiliser has effect on plant growth.*

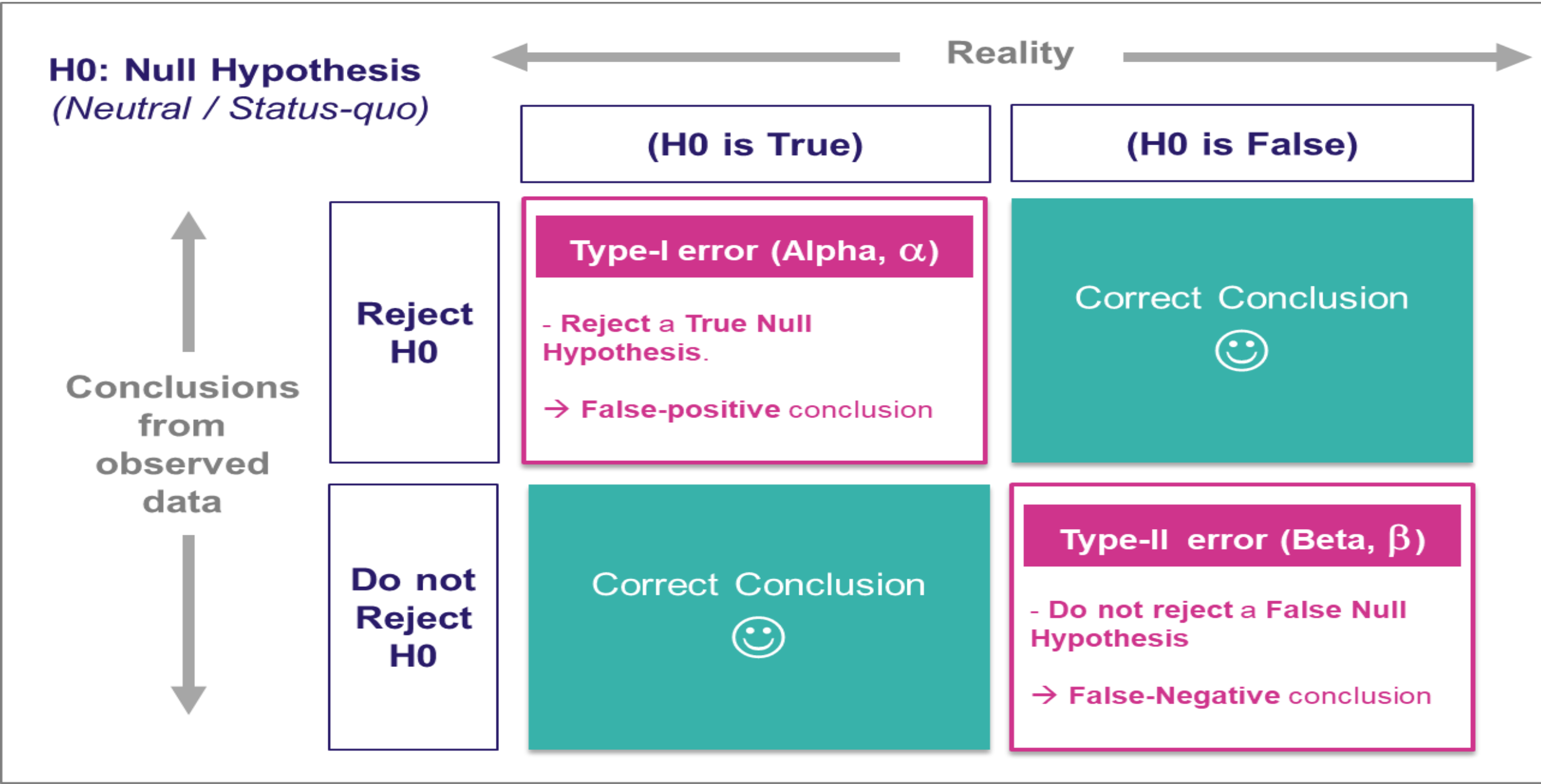
Example-2: There is association between age group and intent to get vaccinated

Testing Hypothesis

- In statistics the alternative hypothesis (H1) is the hypothesis the researchers wish to evaluate
- Since H_o testing is based on sample means, not population means, there is a possibility of making an error or wrong decision in rejecting or failing to reject H_o
  - **Type I error:** When we conclude that there is a relationship or effect but in fact there is not one *(false positive)*.
  - **Type II error:** When we conclude that there is no relationship or effect when in fact there is one *(false negative)*.

If your sample data provide sufficient evidence, you can reject the null hypothesis for the entire population. Your data favor the alternative hypothesis.

Confusion Matrix/Contingency Matrix for Type-I & Type-II Error



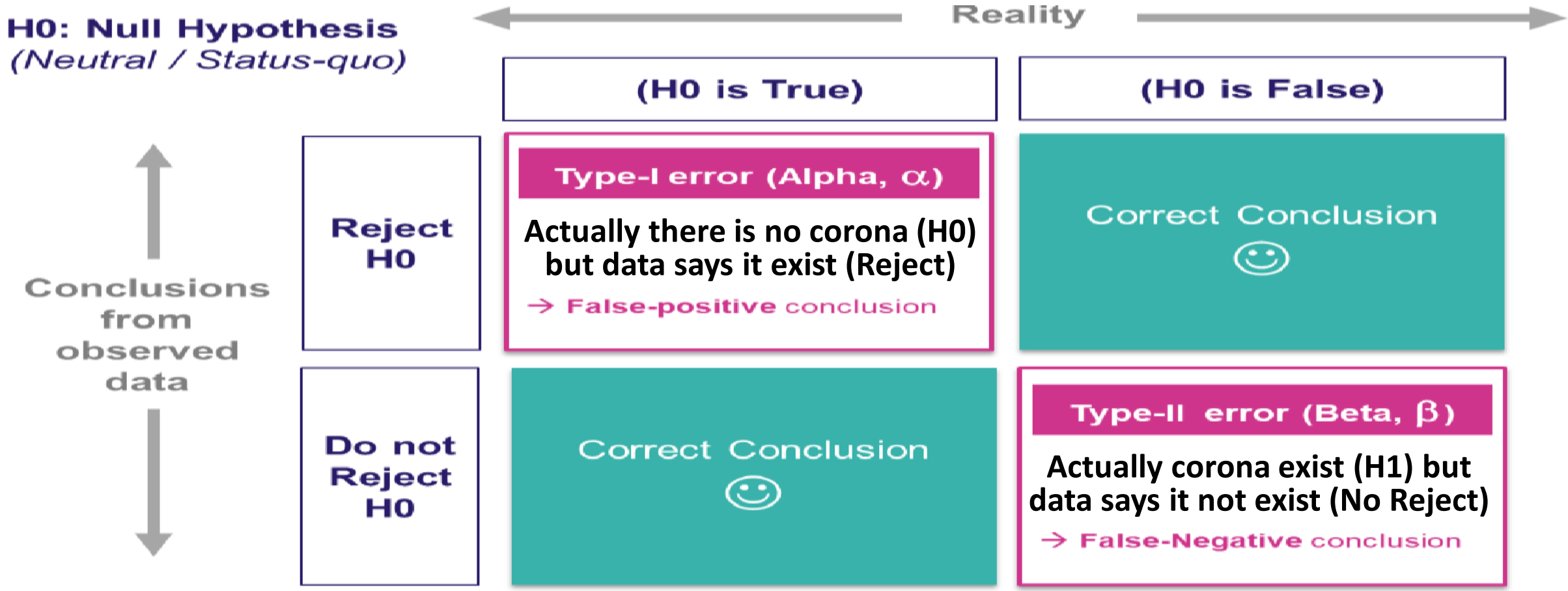
Confusion Matrix/Contingency Matrix for Type-I & Type-II Error

Problem 1: Testing of Corona Virus.

H0: You don't have corona virus. **H1:** You have corona virus.

Reject: The evidence from sample says corona virus exist. So reject H_0.

Don't Reject: There is no sufficient evidence to justify the existance of the corona virus.



Confusion Matrix/Contingency Matrix for Type-I & Type-II Error

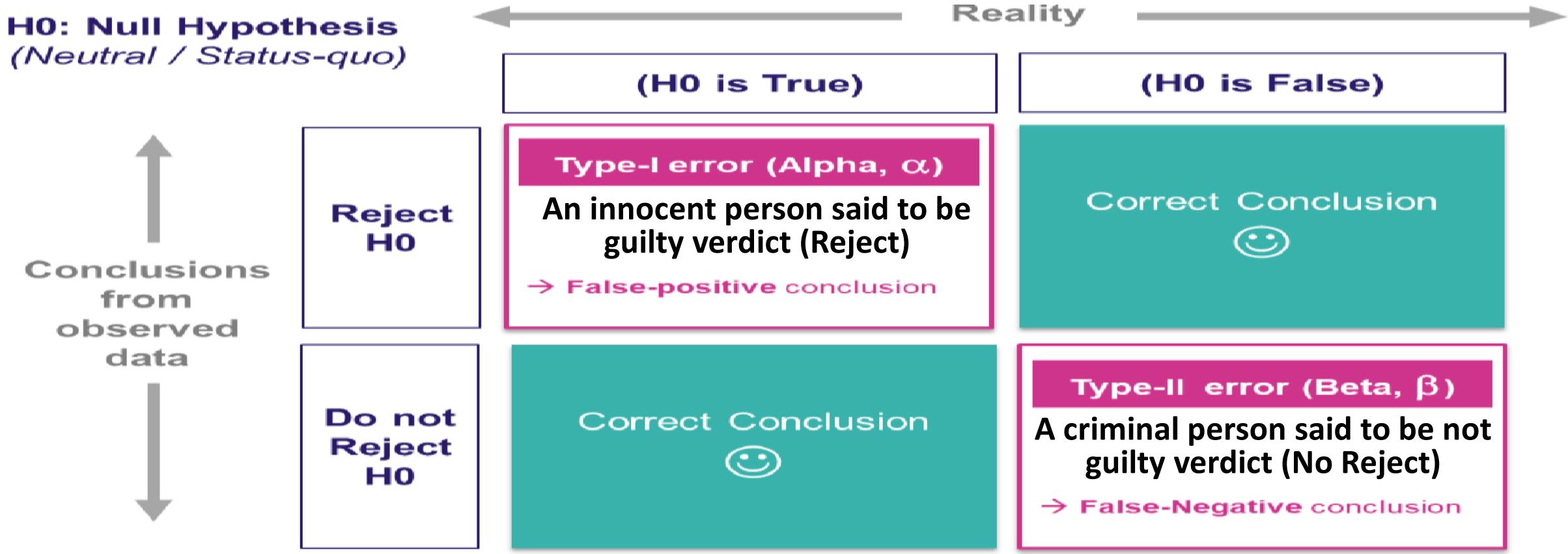
Problem 2: Judgement of arrested person or the defendant.

H0: Defendent is of innocent or unnecessarily the person has arrested.

H1: Defendent is of guilty/criminal otherwise police wouldn't arrest him.

Reject: The evidence says criminal is guilty. So reject H_0.

Don't Reject: There is no sufficient evidence to proof the innocence of the person.



Testing Hypothesis

Type I error

- In a hypothesis test, a type I error occurs when you reject a null hypothesis that is actually true (wrongly reject). In other words, a statistically significant test result suggests that a population effect exists but, in reality, it does not exist.
- The **reson of Type-I Error:** is the difference observed in the sample is the product of random sample error.
- The **probability of committing a type I error equals:** to the significance level (α) that has been set for the hypothesis test. A significance level of 0.05 indicates that there is a 5% chance of being wrongly reject the null hypothesis.
- To lower this risk, a lower value for alpha can be used. However, for lower value for alpha it is less likely to detect the true difference, if it really exists.

Testing Hypothesis

Type II error

- In a hypothesis test, a type II error occurs when one fail to reject a null hypothesis that is actually false.
- In other words, one obtain an **insignificant test result** even though a population effect actually exists.
- Some combination of a small sample size, inherent variability in the data, and bad luck with random sample error might have obscured the population effect.
- **probability of a Type II error = β:** not rejecting the null hypothes when the null hypothesis is false. (1−β) is called the Power of the Test. Type-II error can be minimized by increasing power of the test.

Significance level

- The **significance level (α)** is a measure of the **strength of the evidence** that must be present in the sample before you will reject the null hypothesis and conclude that the effect is statistically significant (difference exist).
- The researcher determines the significance level before conducting the experiment.
- The significance level is the probability of rejecting the null hypothesis when it is true.
- ***For example, a significance level of 0.05 indicates a 5% risk [error] of concluding that a difference exists when there is no actual difference.***
- Lower significance levels indicate that you require stronger evidence before you will reject the null hypothesis.

P-Value

- P-Value is the probability (0 to 1) of NULL hypothesis being true.
- It is a number (statistical value) describing how likely it is that the observed data would have occurred by random chance.
- The P value is the level of statistical significance value at which NULL hypothesis is rejected. The standard significance value is 0.05, i.e. there is 5% chance of being randomness or standard error in the observed data.
- If **P<= 0.05: statistically significant:** It indicates there is strong evidence against the null hypothesis. Therefore, reject the null hypothesis and accept the alternative hypothesis.
- A **P >0.05: not statistically significant:** It indicates there is strong evidence for the null hypothesis. This means retain the null hypothesis and reject the alternative hypothesis.
- **Note 1:** One can “Reject” or “Fail to Reject” the null hypothesis but can’t accept it.
- **Note 2:**The p -value is conditional upon the null hypothesis being true or not but is unrelated to the truth or falsity of the alternative hypothesis.

Degree of Freedom

- The term “Degrees of Freedom” refers to the statistical indicator that shows how many variables in a data set can be changed while restricted by certain constraints. Or the degree of freedom indicates the number of variables that need to be estimated in order to complete a data set.
- It finds extensive use/application in probability distributions, hypothesis testing, and regression analysis.
- The formula for degrees of freedom
  - for single variable samples, such as a 1-sample t-test with sample size N: (N–1)
  - for two-variable samples, such as the Chi-square test with R number of rows and C number of columns: (R-1)*(C-1)

Degree of Freedom...

Problem 1: A sample (data set) with 8 values with the condition that the mean of the data set should be 20. Then find the degree of freedom and if the values for the data set are selected randomly as: 8, 25, 35, 17, 15, 22, 9, then what will be the last value of the data set?

Ans: Degree of freedom= n-1=8-1=7

Last Value= $\frac{\sum x_i}{n}$=20 => 20 * 8 – (8 + 25 + 35 + 17 + 15 + 22 + 9) = 29

Problem 2: For the following 2×2 contigency table with a respective sum for each row and column, calculate its degree of freedom.

From the table, it can be seen that there is only one value in black which is independent and needs to be estimated. Once that value is estimated, then the remaining three values can be derived easily based on the constraints. Therefore,

Degrees of Freedom = (r-1)*(c-1) = (2-1)*(2-1) = 1

	Category A		Total
Category B	22	10	32
	8	18	26
Total	30	28	58

Chi-Square Test (Relationships Among Categorical Variables)

- A Pearson’s chi-square test is a statistical test for categorical data and is denoted by a Greek letter χ^2
- It was developed by Karl Pearson in 1900.
- Chi Square Test is a non parametric test not based on any assumption or distribution of any variable.
- In general, the test is used to measure the difference between what is observed and what is expected according to an assumed hypothesis. It is used to determine whether the data are significantly different from what one expected.
- This non-parametric test is based on frequencies and not on the parameters like mean and SD.
- This test is used for testing the hypothesis and is not useful for estimation.

CONDITIONS FOR THE APPLICATION OF $\chi 2$ TEST

The following conditions should be satisfied before X^2 test can be applied:

1) **The data must be in the form of frequencies**
2) **The frequency data must have a precise numerical value and must be organised into categories or groups.**
3) **Observations recorded and used are collected on a random basis.**
4) **All the itmes in the sample must be independent.**
5) **No group should contain very few items, say less than 10. In case where the frequencies are less than 10, regrouping is done by combining the frequencies of adjoining groups so that the new frequencies become greater than 10. (Some statisticians take this number as 5, but 10 is regarded as better by most of the statisticians.)**
6) **The overall number of items must also be reasonably large. It should normally be at least 50.**

Calculation of Chi Square

$$X^2 = \sum \frac{(o-e)^2}{e}$$

Where,

O :- Observed Frequencies, i.e. the frequency that has actually observed and has given in the contigency table.

E :- Expected Frequencies, i.e. the frequencies that would be expected in the contigency table if the two variables are statistically independent.

$$E = \frac{(Col\ Total) * (Row\ total)}{N}$$

Calculation of Chi Square Problem

Problem:

In an anti-malaria campaign, Quinine was administrated to 500 persons out of a total population of 2000.

The number of fever cases is as follows. Discuss the usefulness of quinine in checking malaria.

Treatment	Fever	No Fever	
Quinine	20	480	500
No Quinine	100	1400	1500
	120	1880	2000

Calculation of Chi Square Problem...

Solution:

- H0: Null Hypothesis: Quinine has no effect in checking malaria
- H1: Alternative Hypothesi: Quinine is effective in checking malaria

Find the expected frequencies from the corresponding observed frequencies.

Treatment	Fever (Observed)	Fever (Expected)	No Fever (Observed)	No Fever (Expected)	
Quinine	20	30	480	470	500
No Quinine	100	90	1400	1410	1500
	120	120	1880	1880	2000

$$(QF)=\frac{120*500}{2000}=30$$

$$E(NQF)=\frac{120*1500}{2000}=90$$

$$E(QNF)=\frac{1880*500}{2000}=470$$

$$E(NQNF)=\frac{1880*1500}{2000}=1410$$

Note: Or find one expected value and then using dgree of freedom concept find the rest entries from the total constraint.

Calculation of Chi Square Problem...

Calculate the value of Chi Square using the formula : $X^2 = \sum \frac{(o-e)^2}{e}$

O	E	(O-E)	$(O-E)^2$	$(O-E)^2/E$
20	30	-10	100	100/30 =3.33
100	90	+10	100	100/90 = 1.11
480	470	+10	100	100/470=0.21
1400	1410	-10	100	100/1410=0.07
				∑=4.72

$$X^2 = \sum \frac{(o-e)^2}{e} = 4.72$$

Calculation of Chi Square Problem...

Now, it is required to determione teh tabulated value of χ^2

- Degree of Freedom = (R-1)*(C-1) = (2-1)*(2-1)=1
- The χ^2 value for 1 degree of freedom and 5% level of significance is 3.84

Calculated value of χ^2 = 4.72 > Tabulated value of χ^2 = 3.84

The calculated value is more than tabulated value and hence reject null hypothesis.

Thus, the inference is quinine is useful in checking malaria.

Chi-Square Distribution Table

TABLE 6-1 Critical Values of the χ^2 Distribution

df \ p	0.995	0.975	0.9	0.5	0.1	0.05	0.025	0.01	0.005	df
1	.000	.000	0.016	0.455	2.706	3.841	5.024	6.635	7.879	1
2	0.010	0.051	0.211	1.386	4.605	5.991	7.378	9.210	10.597	2
3	0.072	0.216	0.584	2.366	6.251	7.815	9.348	11.345	12.838	3
4	0.207	0.484	1.064	3.357	7.779	9.488	11.143	13.277	14.860	4
5	0.412	0.831	1.610	4.351	9.236	11.070	12.832	15.086	16.750	5
6	0.676	1.237	2.204	5.348	10.645	12.592	14.449	16.812	18.548	6
7	0.989	1.690	2.833	6.346	12.017	14.067	16.013	18.475	20.278	7
8	1.344	2.180	3.490	7.344	13.362	15.507	17.535	20.090	21.955	8
9	1.735	2.700	4.168	8.343	14.684	16.919	19.023	21.666	23.589	9
10	2.156	3.247	4.865	9.342	15.987	18.307	20.483	23.209	25.188	10
11	2.603	3.816	5.578	10.341	17.275	19.675	21.920	24.725	26.757	11
12	3.074	4.404	6.304	11.340	18.549	21.026	23.337	26.217	28.300	12
13	3.565	5.009	7.042	12.340	19.812	22.362	24.736	27.688	29.819	13
14	4.075	5.629	7.790	13.339	21.064	23.685	26.119	29.141	31.319	14
15	4.601	6.262	8.547	14.339	22.307	24.996	27.488	30.578	32.801	15

APPLICATIONS OF A CHI SQUARE TEST.

This test can be used in

1) Goodness of fit of distributions

2) test of independence of attributes

3) test of homogenity.

1) TEST OF GOODNESS OF FIT OF DISTRIBUTIONS:

- **This test enables us to see how well does the assumed theoretical distribution (such as Binomial distribution, Poisson distribution or Normal distribution) fit to the observed data.**

- **The $\chi 2$ test formula for goodness of fit is:**

$$X^2 = \sum \frac{(o-e)^2}{e}$$

Where,
o = observed frequency
e = expected frequency

- **If $\chi 2$ (calculated) > $\chi 2$ (tabulated), with (n-1) d.f, then null hypothesis is rejected otherwise accepted.**

- **And if null hypothesis is accepted, then it can be concluded that the given distribution follows theoretical distribution.**

2) TEST OF INDEPENDENCE OF ATTRIBUTES

- Test enables us to explain whether or not two attributes are associated.

- For instance, we may be interested in knowing whether a new medicine is effective in controlling fever or not, $\chi 2$ test is useful.

- In such a situation, we proceed with the null hypothesis that the two attributes (viz., new medicine and control of fever) are independent which means that new medicine is not effective in controlling fever.

- $\chi 2$ (calculated) > $\chi 2$ (tabulated) at a certain level of significancfe for given degrees of freedom, the null hypothesis is rejected, i.e. two variables are dependent.(i.e., the new medicine is effective in controlling the fever) and if, $\chi 2$ (calculated) <$\chi 2$ (tabulated) ,the null hypothesis is accepted, i.e. 2 variables are independent.(i.e., the new medicine is not effective in controlling the fever).

- when null hypothesis is rejected, it can be concluded that there is a significant association between two attributes.

3) TEST OF HOMOGENITY

- This test can also be used to test whether the occurance of events follow uniformity or not e.g. the admission of patients in government hospital in all days of week is uniform or not can be tested with the help of chi square test.

- $\chi 2$ (calculated) < $\chi 2$ (tabulated), then null hypothesis is accepted, and it can be concluded that there is a uniformit in the occurance of the events. (uniformity in the admissio of patients through out the week)