Data Analytics (IT-3006)

Kalinga Institute of Industrial Technology
Deemed to be University
Bhubaneswar-751024

School of Computer Engineering

3 Credit

Lecture Note

Course Contents

Sr #	Major and Detailed Coverage Area	Hrs
2	**Statistical Concepts** **Data Exploration:** Distribution of a single variable, Basic Concepts (populations and samples, 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. **Sampling and distributions:** Terminology, Estimation, Confidence Interval estimation, Sampling distributions, Confidence interval, Hypothesis testing, Chi-square test for independence.	8

Data Exploration

Data Exploration

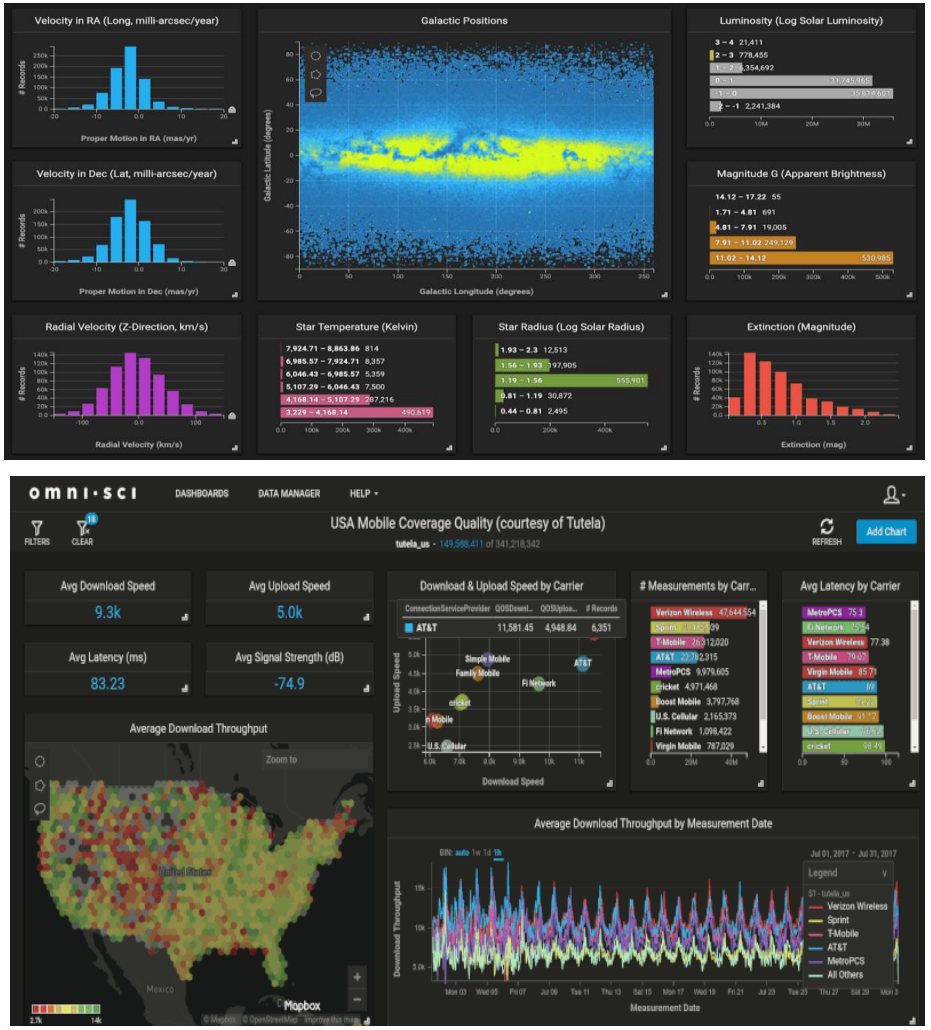
- Data exploration refers to the initial step in data analysis in which data analysts use data visualisation and statistical techniques to describe dataset characterizations, such as size, quantity, and accuracy, in order to better understand the nature of the data.

- Raw data is typically reviewed with a combination of manual workflows and automated data-exploration techniques to visually explore data sets, look for similarities, patterns and outliers and to identify the relationships between different variables.

- This is also sometimes referred to as **exploratory data analysis**, which is a statistical technique employed to analyse raw data sets in search of their broad characteristics.

Importance of Data Exploration

- Starting with data exploration helps users to make better decisions on where to dig deeper into the data and to take a broad understanding of the business when asking more detailed questions later.
- Performing the initial step of data exploration enables data analysts to better understand and visually identify anomalies and relationships that might otherwise go undetected.
- Data exploration tools include data visualization software and business intelligence platforms, such as Microsoft Power BI, Qlik and Tableau.



Exploratory Data Analysis

- 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)** can be classified as **univariate**, **bivariate**, and **multivariate** 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.
- 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

Exploratory Data Analysis cont…

- Raw data are not very informative. Exploratory Data Analysis (EDA) is how we make sense of the data by converting them from their raw form to a more informative one.
- **In particular, EDA consists of:**
  - Organizing and summarizing the raw data
  - Discovering important features and patterns in the data and any striking deviations from those patterns, and then
  - Interpreting our findings in the context of the problem
- **Usefulness:**
  - Describing the distribution of a single variable (center, spread, shape, outliers)
  - Checking data (for errors or other problems)
  - Checking assumptions to more complex statistical analyses
  - Investigating relationships between variables

Univariate data and its analysis

- This type of data consists of only one **variable**. *A **variable** is a characteristic that can be measured and that can assume different values. Height, age, income, province or country of birth, grades obtained at school and type of housing are all examples of variables.*
- The analysis of univariate data is thus 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 example of a univariate data can be height (in cm)

Height	164	168	170	169	173	175	180	175	176

- Suppose that the heights of seven students of a class is recorded, there is only one variable that is height and it is not dealing with any cause or relationship. The description of patterns found in this type of data can be made by drawing conclusions using central tendency measures (mean, median and mode), dispersion or spread of data (range, minimum, maximum, quartiles, variance and standard deviation) and by using frequency distribution tables, histograms, pie charts, frequency polygon and bar charts.

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.
- 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 are directly proportional to each other and thus related because as the temperature increases, the sales also increase.
- 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 of data.

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)**.

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.

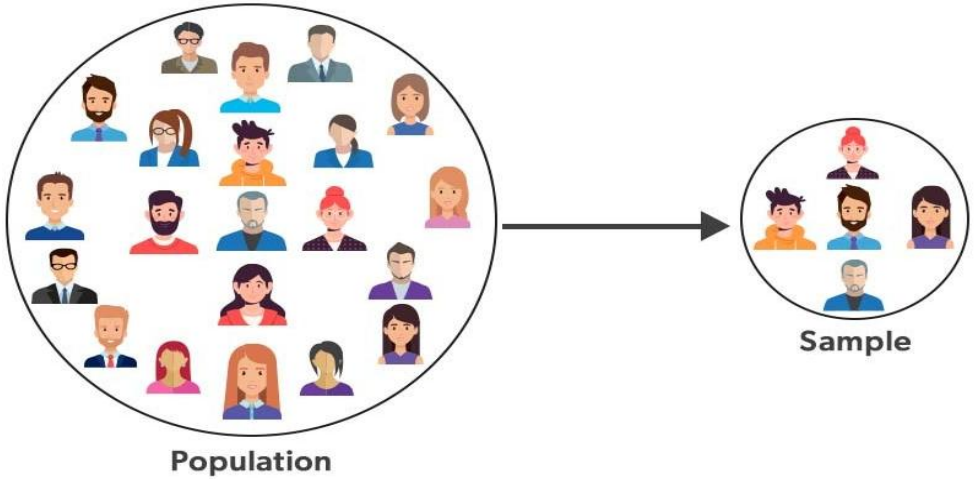
Basic Concepts

Several important concepts:

- Population and sample
- Data sets
- Variables and observations
- Types of data

Population and Sample

- A population includes all of the entities of interest in a study (people, households, machines, etc.) and the examples includes:
  - 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.



Population and sample example

Population

KIIT administrator wants to analyse the final exam scores of all graduating students to see if there is a trend. Since they are interested in applying their findings to all graduating students at KIIT university, they use the whole population dataset.

Sample

KIIT want to study political attitudes in students. KIIT population is around the 30,000 undergraduate students. Because it's not practical to collect data from all of them, so one may use a sample of 300 undergraduate volunteers from different school who meet the inclusion criteria. This is the group who are expected to be part of the survey.

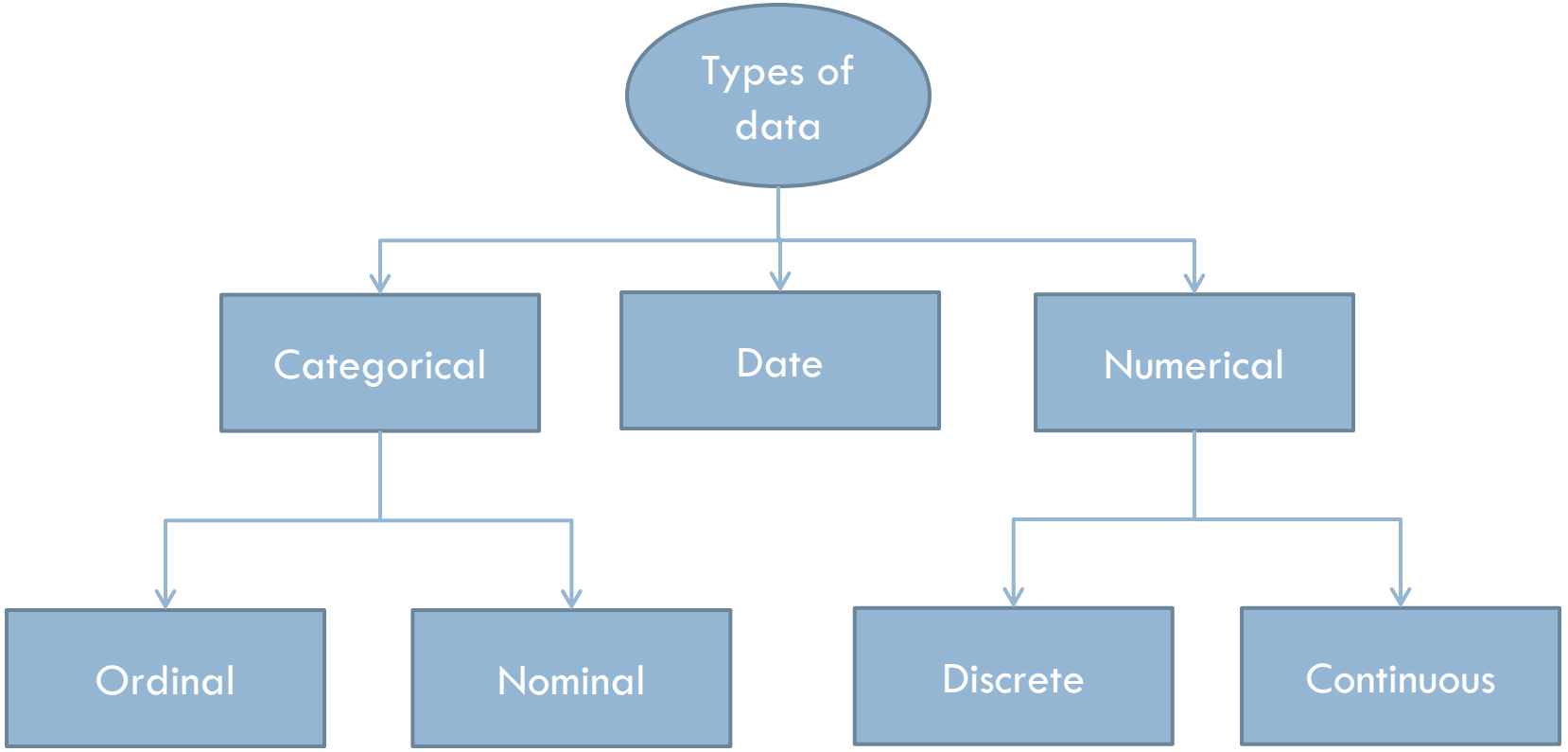
Datasets, Variables, and Observations

- A **dataset** 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.

	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
28	27	27	Male	Illinois	3	$45,400	2
29	28	63	Male	Michigan	2	$53,900	1
30	29	52	Male	California	1	$44,100	3
31	30	48	Female	New York	2	$31,000	4

Illustrate variables and observations

Types of data



Types of data cont…

- A variable is numerical if meaningful arithmetic can be performed on it. Otherwise, the variable is categorical.
- **Age, children**, and **salary** are clearly numerical and for example, it makes perfect sense to sum or average any of these. In contrast, **gender** and **state** are clearly categorical because they are expressed as text, not numbers.
- There is a third data type, a **date** variable. For obvious reasons, dates are treated differently from typical numbers.
- A categorical variable is **ordinal** if there is a natural ordering of its possible categories (e.g., Likert scale). If there is no natural ordering, the variable is **nominal** (e.g., gender, state).
- Categorical variables can be coded numerically. Gender can be coded as 1 for males and 0 for females. This 0–1 coding for a categorical variable is very common. Such a variable is called a **dummy** variable, and it often simplifies the analysis.

Types of data cont…

- In addition, the age variable can be categorized as “young” (34 years or younger), “middle-aged” (from 35 to 59 years), and “elderly” (60 years or older). This method of categorizing a numerical variable is called **binning** (putting the data into discrete bins), and it is also common (It is also called **discretizing**)
- Numerical variables can be classified as **discrete** or **continuous**. The basic distinction is whether the data arise from counts or continuous measurements. The variable children is clearly a count (**discrete**), whereas the variable salary is best treated as **continuous**.
- Data sets can also be categorized as **cross-sectional** or **time series**. Cross-sectional data are data on a cross-section of a population at a distinct point in time and time series data are data collected over time.

Types of data cont…

- The key difference between **time series** and **cross-sectional** data is that the time series data focuses on the **same variable over a period of time** while the cross-sectional data focuses on **several variables at the same point of time**. Furthermore, the time series data consist of observations of a single subject at multiple time intervals whereas, the cross-sectional data consist of observations of many subjects at the same point in time.

Time Series

Year	Profit
2001	50000
2002	60000
2003	70000
2004	80000

Cross-sectional

City	Max Temp	Humidity	Wind Speed
A	29	60%	20 kph
B	27	65%	30 kph
C	32	70%	35 kph
D	35	72%	25 kph

Frequency distribution

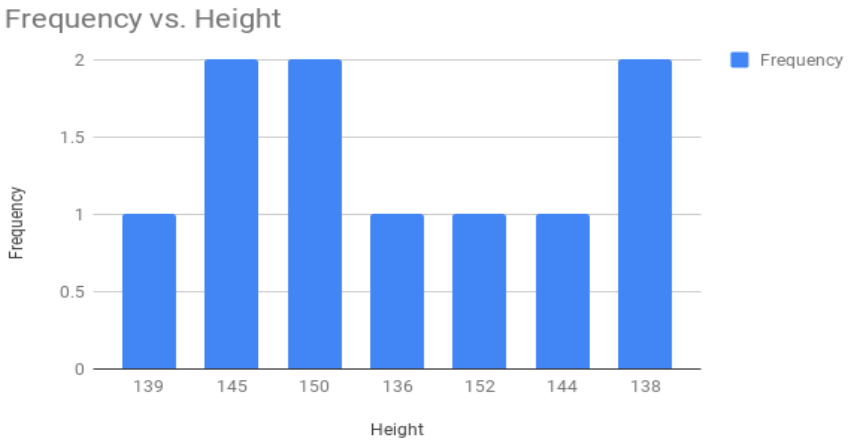
Frequency distribution provides the information of the number of occurrences (frequency) of distinct values distributed within a given period of time or interval, in a list, table, or graphical representation.

Example: Let us take the example of the heights of ten students in cm i.e., 139, 145, 150, 145, 136, 150, 152, 144, 138, 138.

Frequency distribution table

Height	Frequency
139	1
145	2
150	2
136	1
152	1
144	1
138	2

Frequency distribution graph



Descriptive Measures for Categorical Variables

1. **Frequencies:** The number of observations for a particular category
2. **Proportions:** The percent that each category accounts for out of the whole
3. **Marginals:** The totals in a cross tabulation by row or column
4. **Visualization**: Understand the features of the data through statistics and visualization
5. **Dummy variables:** Take values of 0 and 1, where the values indicate the presence or absence of something.

Frequencies

- To produce contingency tables which calculate counts for each combination of categorical variables.
- For instance, the essence is to get the total count of female and male customers.
- If the essence is to understand the number of married and single females and male customers, a cross classification table can be produced.
- If the essence is to produce multidimensional tables based on three or more categorical variables. In this case the count of customers by marital status, gender, and location can be assessed.

Female	Male
7170	6889

	Female	Male
Married	3602	3264
Single	3568	3625

		Place-1	Place-2	Place-3
Married	Female	190	638	188
	Male	197	692	210
Single	Female	183	686	175
	Male	239	717	242

Proportions

- To produce contingency tables which calculate proportions for each combination of categorical variables.
- For instance, the essence is to get the proportions of female and male customers.
- If the essence is to understand the proportions of married and single females and male customers, a cross classification table can be produced.
- If the essence is to produce multidimensional tables based on three or more categorical variables. In this case the proportions of customers by marital status, gender, and location can be assessed.

Female	Male
0.51	0.49

	Female	Male
Married	0.25	0.23
Single	0.27	0.28

		Place-1	Place-2	Place-3
Married	Female	0.014	0.045	0.013
	Male	0.014	0.049	0.015
Single	Female	0.013	0.049	0.012
	Male	0.017	0.051	0.017

Mariginals

- Marginals show the total counts or percentages across columns or rows in a contingency table.
- Marginal frequencies and the percentages for these marginal frequencies can be computed.
- Consider

	Female	Male
Married	3602	3264
Single	3568	3625

Frequency mariginals

Married	Single
6866	7193

Row

Female	Male
7170	6889

Column

Percentage mariginals

Row

	Female	Male
Married	0.52	0.48
Single	0.49	0.51

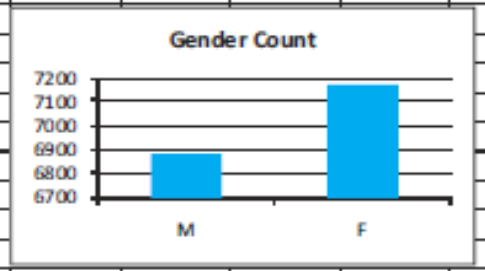
Column

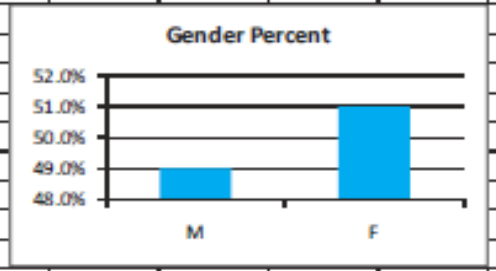
	Female	Male
Married	0.52	0.48
Single	0.49	0.51

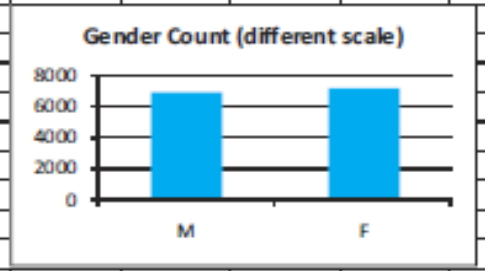
Visualisation

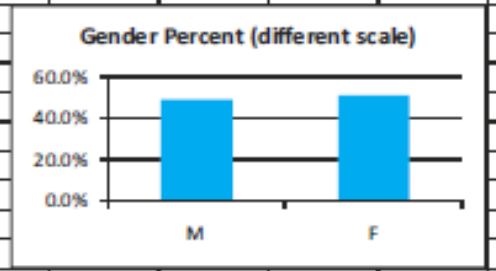
Bar chart and Pie chart are most often used to visualize categorical variables.

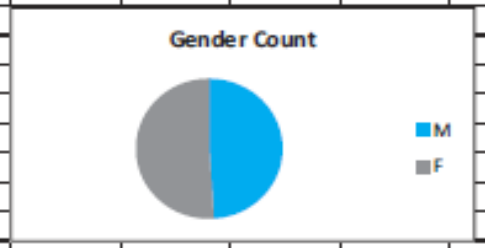
	R	S	T
1	Categorical summaries		
2	Gender	Count	Percent
3	M	6889	49.0%
4	F	7170	51.0%
5			100.0%
6			
7	Marital Status	Count	Percent
8	S	7193	51.2%
9	M	6866	48.8%
10			100.0%
11			
12	Homeowner	Count	Percent
13	Y	8444	60.1%
14	N	5615	39.9%
15			100.0%
16			
17	Annual Income	Count	Percent
18	$10K - $30K	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%
27			
28			
29			
30			

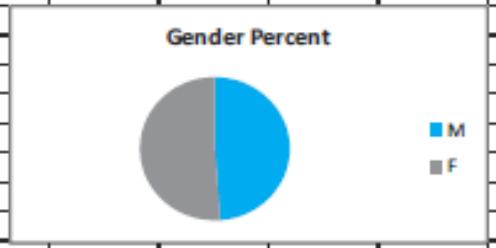












Dummy Variables

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.
- 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	12/18/2014	7223	F	0
3	2	12/20/2014	7841	M	1
4	3	12/21/2014	8374	F	0
5	4	12/21/2014	9619	M	1
6	5	12/22/2014	1900	F	0
7	6	12/22/2014	6696	F	0
8	7	12/23/2014	9673	M	1
9	8	12/25/2014	354	F	0
10	9	12/25/2014	1293	M	1
11	10	12/25/2014	7938	M	1
14055	14054	12/29/2016	2032	F	0
14056	14055	12/29/2016	9102	F	0
14057	14056	12/29/2016	4822	F	0
14058	14057	12/31/2016	250	M	1
14059	14058	12/31/2016	6153	F	0
14060	14059	12/31/2016	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.
- The various numerical summary measures can be categorized into several groups such as
  - Measures of Central Tendency
  - Minimum, Maximum, Percentiles, and Quartiles
  - Measures of Variability
  - Empirical Rules for Interpreting Standard Deviation
  - Measures of Shape

Measures of Central Tendency – Mean, Median & Mode

- The m**ean** is the average of all values. If the data set represents a sample from some larger population, this measure is called the sample mean and is denoted by $\overline{X}$. If the data set represents the entire population, it is called the population mean and is denoted by μ.
- The **median** is the middle observation when the data are sorted from smallest to largest. If the number of observations is odd, the median is literally the middle observation. If the number of observations is even, the median is usually defined as the average of the two middle observations i.e., if there are 10 observations, the median is usually defined as the average of the fifth and sixth smallest values.
- The **mode** is the value that appears most often. In most cases where a variable is essentially continuous, the mode is not very interesting because it is often the result of a few lucky ties.

Minimum, Maximum, Percentiles, and Quartiles

- The **minimum** is the lowest value.
- The **maximum** is the highest value.
- For any percentage p, the p^{th} **percentile** is the value such that a percentage p of all values are less than it.
- The **quartiles** divide the data into four groups, each with (approximately) a quarter of all observations.
  - The first (Q_1), second (Q_2), and third (Q_3) quartiles are the percentiles corresponding to p = 25%, p = 50%, and p = 75%.
  - Q_0 is the smallest value in the data
  - Q_4 is the largest value in the data
  - By definition, Q_2(p = 50%) is equal to the median.
  - Inter Quartile Range (IQR) = Q3 – Q1

Quartile Calculation

Quartile calculation for even number of data

Find Q1, Q2 and Q3 of the set {4, 17, 7, 14, 18, 12, 3, 16, 10, 4, 4, 11}.

1. Put them in order: 3, 4, 4, 4, 7, 10, 11, 12, 14, 16, 17, 18

2. Cut it into halves: 3, 4, 4 4, 7, 10 | 11, 12, 14, 16, 17, 18

In this case all the quartiles are between numbers:

Quartile 1 (Q_1) = (4+4)/2 = 4, Quartile 2 (Q_2) = (10+11)/2 = 10.5

Quartile 3 (Q_3) = (14+16)/2 = 15

Quartile calculation for odd number of data

Find Q1, Q2 and Q3 of the set {12, 11, 4, 5, 9, 7, 6, 2, 1}.

1. Put them in order: 1, 2, 4, 5, 6, 7, 9, 11, 12

2. Calculate the median i.e., the middle most value = 6 which is also Q_2

3. Cut it into halves by excluding median: 1, 2, 4, 5 | 7, 9, 11, 12

In this case all the quartiles are :

Quartile 1 (Q_1) = (2+4)/2 = 3, Quartile 3 (Q_3) = (9 +11)/2 = 10

Percentile Calculation

Definition:

k= the kth percentile. It may or may not be part of the data.

i= the index (ranking or position of a data value)

n= the total number of data

To calculate percentile:

1. Order the data from smallest to largest.

2. Calculate i = (k / 100) * (n + 1)

3. If i is an integer, then the k^{th} percentile is the data value in the i^{th} position in the ordered set of data. If i is not an integer, then round i up and round i down to the nearest integers. Average the two data values in these two positions in the ordered data set.

Percentile Calculation cont…

Listed are 29 ages for academy award winning best actors in order from smallest to largest.

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

Find the 70^{th} percentile:

i = (k / 100) * (n + 1) = (70 / 100) * (29 + 1) = 21

21 is an integer, and the data value in the 21^{st} position in the ordered data set is 64. The 70^{th} percentile is 64 years.

Find the 83^{rd} percentile:

i = (k / 100) * (n + 1) = (83 / 100) * (29 + 1) = 24.9 which is not an integer.

Round it down to 24 and up to 25. The age in the 24^{th} position is 71 and the age in the 25^{th} position is 72. Average 71 and 72. The 83^{rd} percentile is 71.5 years.

Measures of Variability

- Measures of variability provide summary statistics to understand the variety in relation to the midpoint of the data i.e., representation of the amount of dispersion in a dataset.
- Thus, measures of variability includes:
  1. Range
  2. Interquartile range
  3. Variance
  4. Standard deviation
  5. Mean absolute deviation
- **Range:** It is the maximum value minus the minimum value.
- **Interquartile range:** It is the third quartile minus the first quartile.
- **Variance:** It is essentially the average of the squared deviations from the mean.
  - If X_i is a typical observation, its squared deviation from the mean is $(X_i - mean)^2$.
  - The **sample variance** is denoted by s^2, and the **population variance** by σ^2.

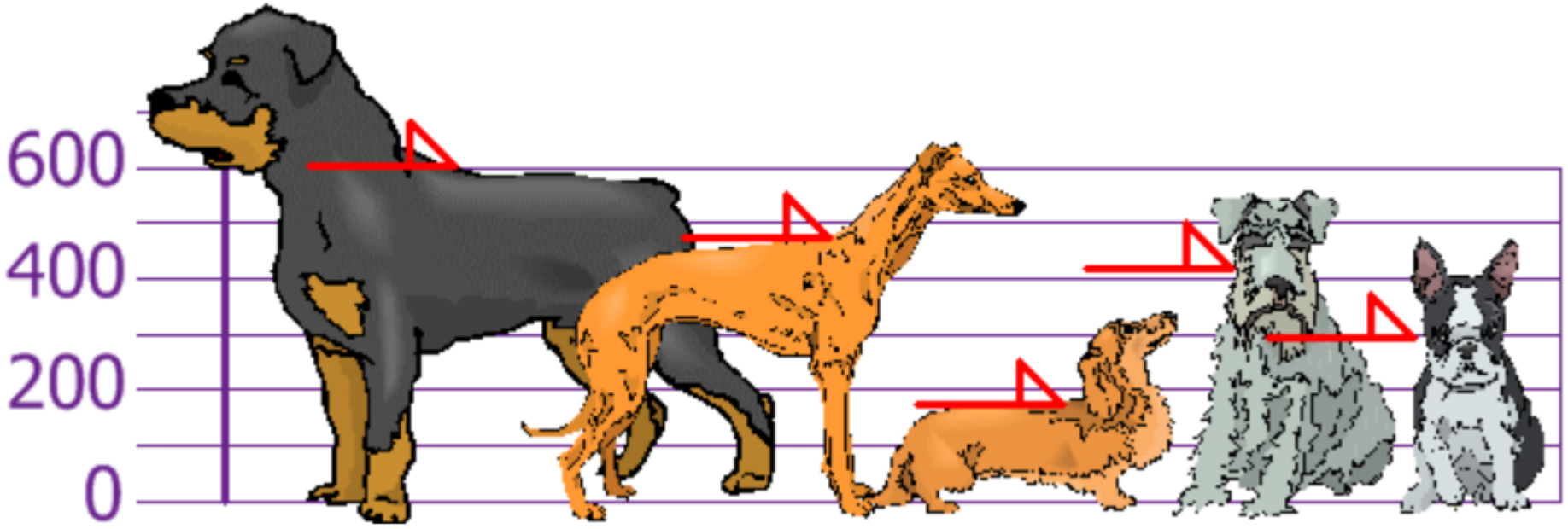
$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \text{mean})^2}{n-1} \qquad \sigma^2 = \frac{\sum_{i=1}^{n} (X_i - \text{mean})^2}{n}$$

Measures of Variability cont…

- **Standard deviation :** A fundamental problem with variance is that it is in squared units. Therefore, a more natural measure is the standard deviation, which is the square root of the variance.
  - The **sample standard deviation**, denoted by s, is the square root of the sample variance.
  - The **population standard deviation**, denoted by σ, is the square root of the population variance.
- **Mean absolute deviation (MAD):** It is the average distance between each data point and the mean.

$$\mathrm{MAD} = \frac{\sum |x_i - \bar{x}|}{n}$$

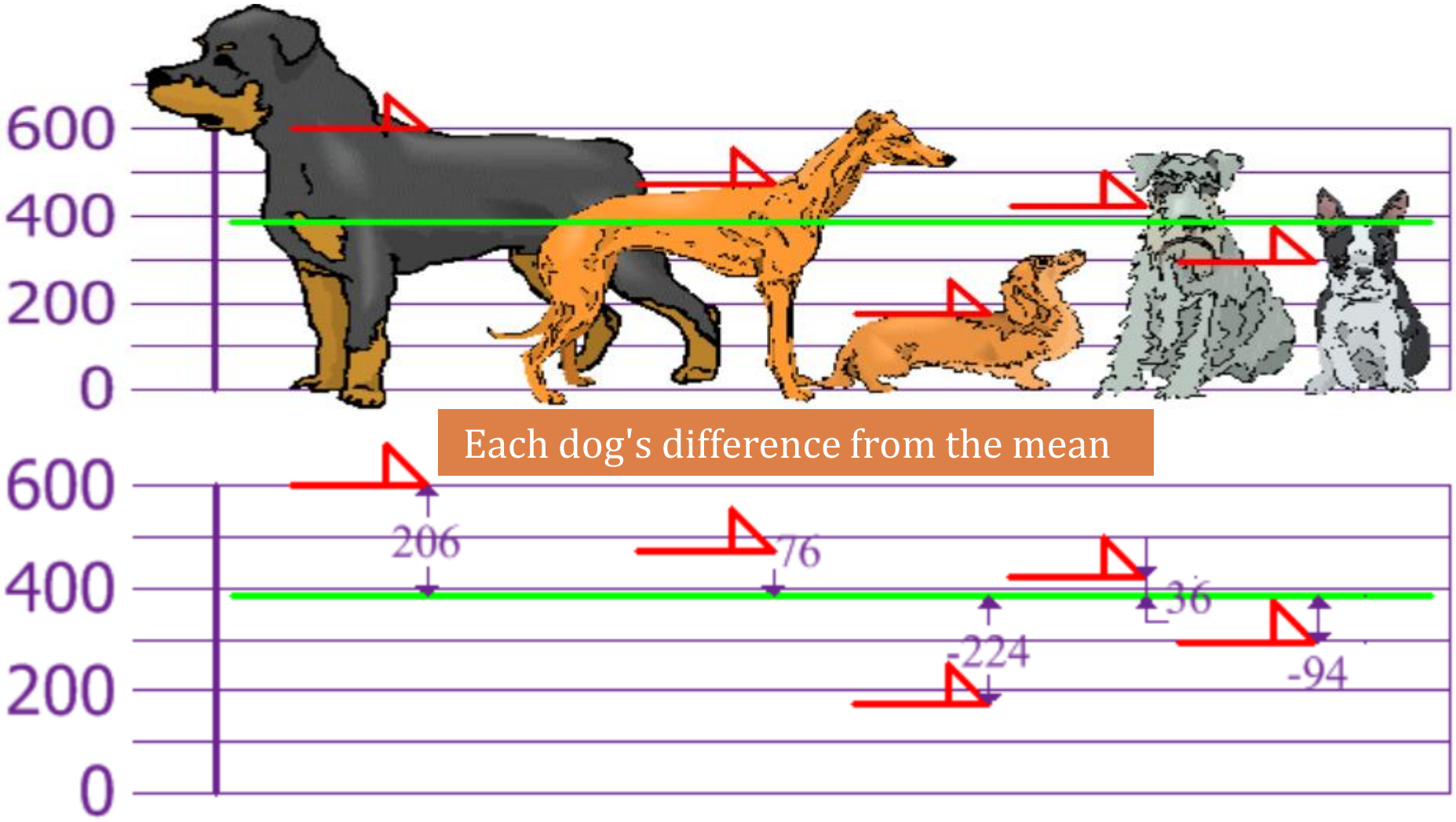
Example



The heights (at the shoulders) are: 600mm, 470mm, 170mm, 430mm and 300mm. Find out the range, the mean, the variance, and the standard deviation.

Range = maximum – minimum = 600 – 170= 430
Mean = (600 + 470 + 170 + 430 + 300) / 5 = 394

Example cont...



Example cont...

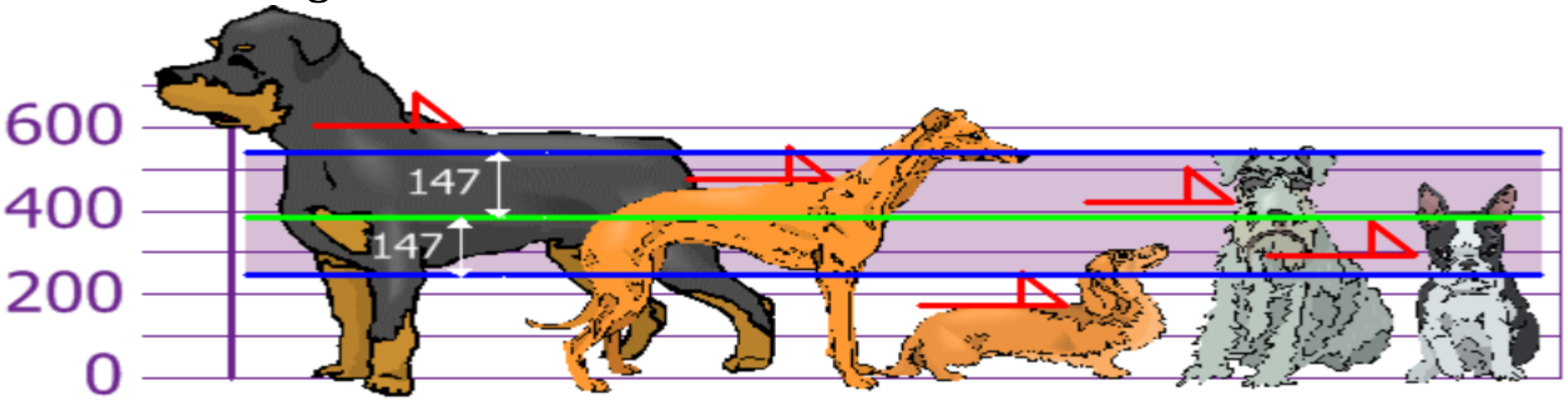
To calculate the variance, take each difference, square it, and then average the result:

$\sigma^2 = (206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2)/5 = 21704$

The standard deviation is just the square root of variance, so:

$\sigma = \sqrt{21704} = 147.32... = 147$

The good thing about the standard deviation is that it is useful. Now, it can be shown which heights are within one standard deviation (147mm) of the mean. So, using the standard deviation, there is a "**standard**" way of knowing what is normal, and what is extra large or extra small. **Rottweilers** are tall dogs and **Dachshunds** are a bit short, **right?**



Class Exercise

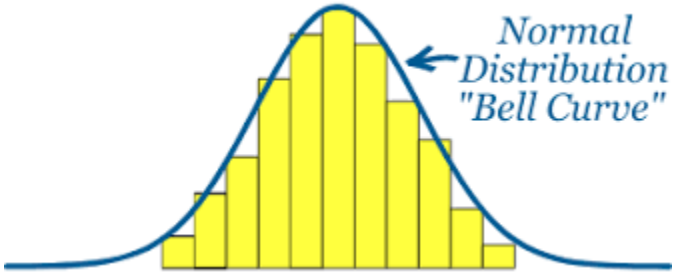
Q1. Consider your travel time in minutes from the hostel to library : 15, 29, 8, 42, 35, 21, 18, 42, 26. Calculate:

- Mean
- Median
- Mode
- Standard deviation
- Variance
- Range
- Quartiles
- Percentile
- Sample size

Q2. The following dollar amounts were the hourly collections from a Salvation Army kettle at a local store one day in December: $12, $12, $12, $12, $12, $12, $12, $12, $12, $12, $12, and $12. Determine the Interquartile Range for the amount collected.

Normal Distribution

- In a normal distribution, data is symmetrically distributed with no skew. When plotted on a graph, the data follows a bell shape, with most values clustering around a central region and tapering off as they go further away from the center. **Note:** If the bell curve is shifted to the left or the right, it is said to be skewed.
- Normal distributions are also called **Gaussian distributions** or **bell curves** because of their shape.



Histogram

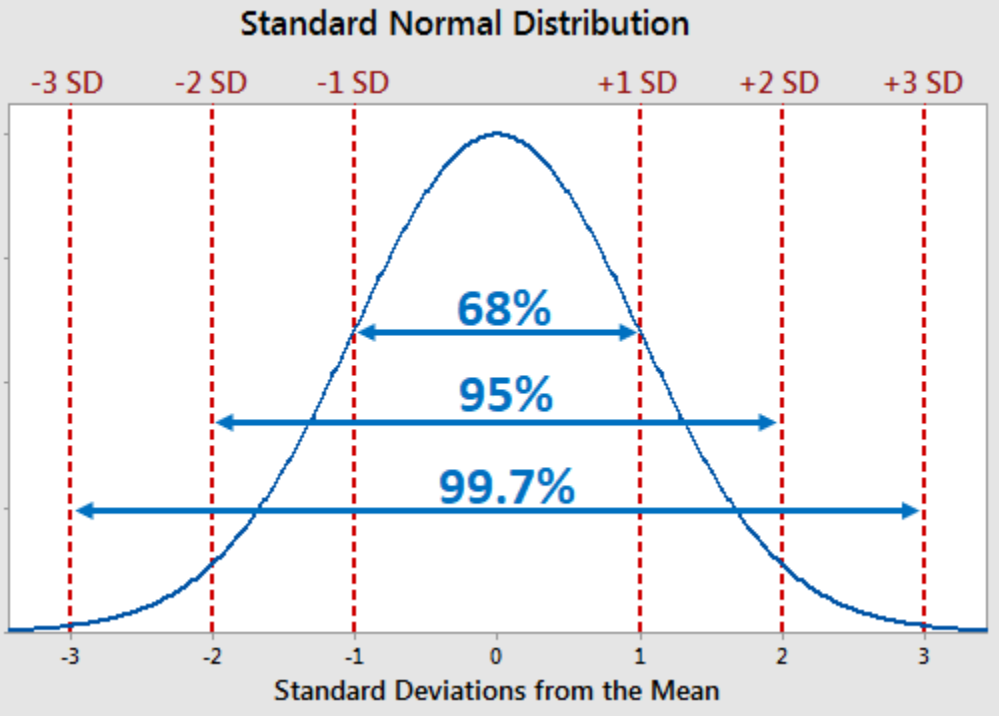
A histogram is an approximate representation of the distribution of numerical data.

Standard Normal Distribution

- The standard normal distribution is one of the forms of the normal distribution.
- It occurs when a normal random variable has a mean equal to zero and a standard deviation equal to one. In other words, a normal distribution with a mean 0 and standard deviation of 1 is called the standard normal distribution.
- Also, the standard normal distribution is centered at zero, and the standard deviation gives the degree to which a given measurement deviates from the mean.

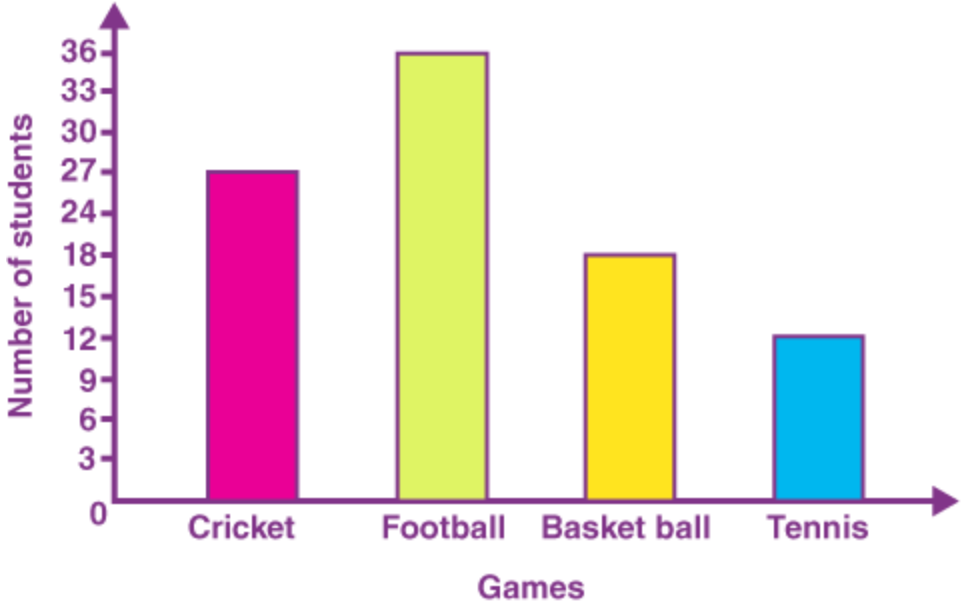
Empirical Rules for Interpreting Standard Deviation

The empirical rule also known as the **68 95 99** rule, states that for normal distributions (***symmetric or bell-shaped***), 68% of observed data points will lie inside one standard deviation of the mean, 95% will fall within two standard deviations, and 99.7% will occur within three standard deviations. ***Empirical means it is grounded in practical reality.*** In general, this limit serves as a valuable way to identify outliers.



Histogram

- A histogram is a graphical representation of a grouped frequency distribution of the data with similar classes.
- It is represented by a set of rectangles, adjacent to each other, where each bar represent a kind of data.
- The heights of rectangles are proportional to corresponding frequencies of similar classes (e.g., cricket or football, etc as per below diagrram)



Data and Distribution

DA Score

Score	Frequency
60	3
65	10
70	12
75	15
80	20
85	15
90	12
95	10
100	3

Frequency = no. of students who received that score

Plot Histogram

Recap on data distribution

Data is normally distributed. Normal distribution curve is a bell-shaped frequency distribution curve. The bell curve gets its name quite simply because its shape resembles that of a bell. Most of the data values tend to cluster around the mean. Right and the left of the distribution are perfect mirror images of one another.

Data and Distribution cont...

Size of Items	Frequency
2-4	5
4-6	8
6-8	5
8-10	5
10-12	8
12-14	3
14-16	1

Size of Items	Frequency
10-15	2
15-20	5
20-25	3
25-30	8
30-35	3
35-40	5
40-45	6

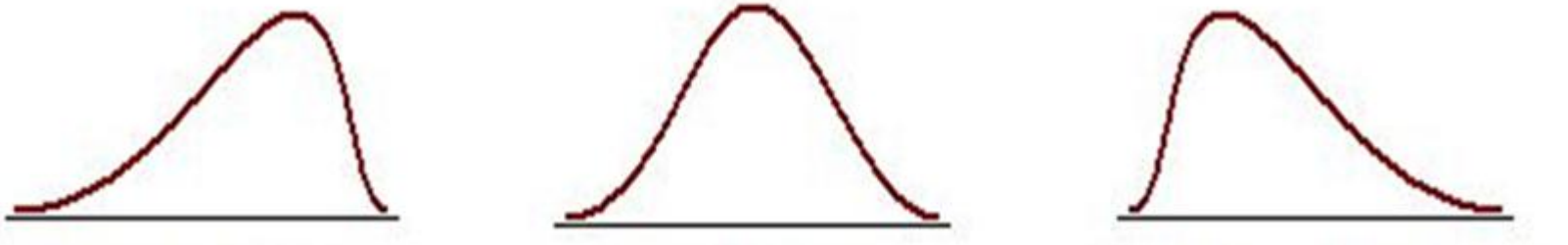
Plot Histogram

Measures of Shape

There are two final measures of a distribution i.e., **skewness** and **kurtosis**.

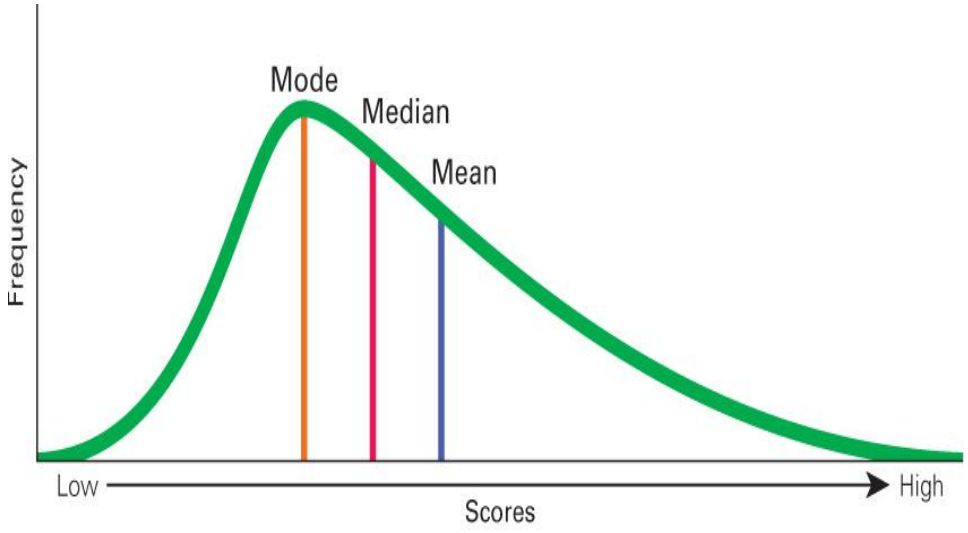
Skewness

- Not every distribution of data is symmetric. Sets of data that are not symmetric are said to be asymmetric. The measure of how asymmetric a distribution can be is called skewness.
- Skewness is a measure that refers to the extent of symmetry or asymmetry in a distribution.
- Skewness is asymmetry in a statistical distribution, in which the curve appears distorted or skewed either to the left or to the right.

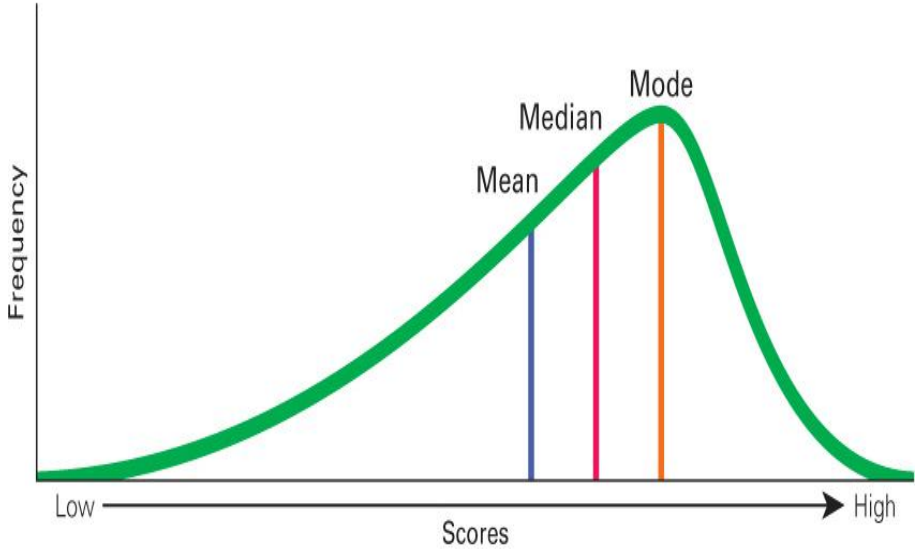


Skewness Types

- Skewed to the Right: Data that are skewed to the right have a long tail that extends to the right. In this situation, the mean and the median are both greater than the mode.
- Skewed to the Left: Data that are skewed to the left have a long tail that extends to the left. In this situation, the mean and the median are both less than the mode.



Positively Skew



Negatively Skew

Measures of Skewness

The measures of skewness can be both absolute and relative.

- Absolute Measure: It tells the extent of asymmetry and whether it is positive or negative.

  Symbolically: absolute skewness = **mean – mode** or **mean – median**

  If the value is > 0, skewness is positive else negative.

- Relative Measure: In order to make comparison between the skewness in two or more distributions, coefficient of skewness is computed for the given series or distribution. The formula used for measuring skewness is

$$\text{skewness} = SK_p = \frac{\textbf{mean – mode}}{\textbf{standard deviation}}$$

This is called **Karl Pearson's coefficient of skewness**

Illustration:

Section A: mean = 46.83, SD (standard deviation) = 14.8, and mode = 51.67

Section B: mean = 47.83, SD = 14.8, and mode = 47.07

Section A: SK_p = -0.327 and Section B: SK_p = 0.051

So, distribution of marks in section A is more skewed (relative). The skewness of Section A is negative , while that of B is positive.

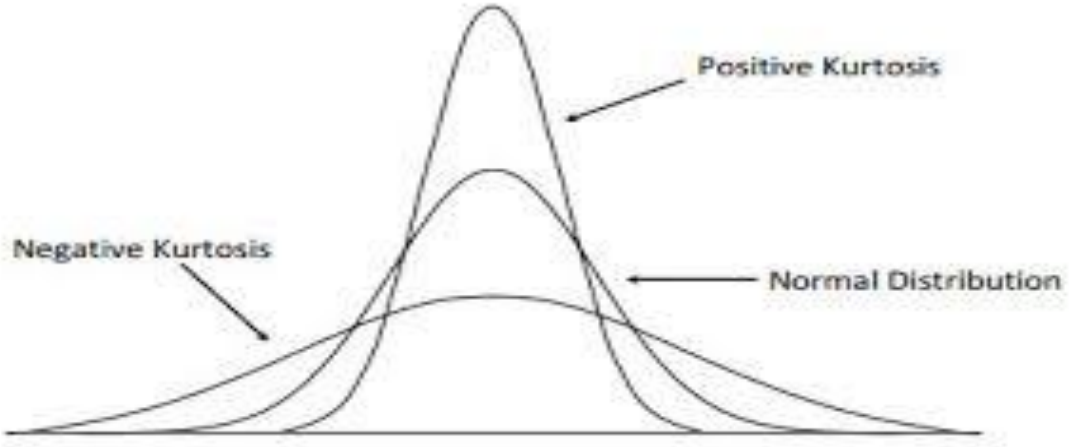
Measures of Skewness cont…

When is the skewness too much? The rule of thumb seems to be:

- If the skewness is between -0.5 and 0.5, the data are fairly symmetrical
- If the skewness is between -1 and – 0.5 or between 0.5 and 1, the data are moderately skewed
- If the skewness is less than -1 or greater than 1, the data are highly skewed

Kurtosis

Along with skewness, kurtosis is an important descriptive statistic of data distribution. Kurtosis measures the degree to which the distribution has either fewer and less extreme outliers, or more and more extreme outliers than the normal distribution. In nutshell, skewness essentially measures the symmetry of the distribution, while kurtosis determines the heaviness of the distribution tails. In general, the higher the kurtosis, the sharper the peak and the longer the tails. This is called leptokurtic, and is indicated by positive kurtosis values. The opposite—platykurtosis—has negative kurtosis values.



Kurtosis cont…

Plot histogram for each dataset and observe the distribution

Score	Frequency
75	3
80	3
85	3
90	4
95	27
100	32
105	27
110	4
115	3
120	3
125	3

Score	Frequency
75	19
80	20
85	27
90	26
95	27
100	27
105	27
110	26
115	27
120	22
125	21

Kurtosis cont...

- **High kurtosis** in a data set is an indicator that data has heavy tails or outliers. If there is a high kurtosis, then, we need to investigate why do we have so many outliers. It indicates a lot of things, maybe wrong data entry or other things. Investigate!
- **Low kurtosis** in a data set is an indicator that data has light tails or lack of outliers. If we get low kurtosis(too good to be true), then also we need to investigate and trim the dataset of unwanted results.

Kurtosis Calculation

$$S_{kr} = \frac{1}{n}\frac{\sum_{i=1}^{n}(X_i - \bar{X})^4}{S^4}$$

Using the data from the example above (12 13 54 56 25), determine kurtosis.

$$X = \frac{(12 + 13 + \ldots + 25)}{5} = \frac{160}{5} = 32$$

$$S^2 = \frac{(12-32)^2 + \ldots + (25-32)^2}{4} = 467.5$$

$$\textit{Therefore,} \qquad S = 467.5^{\frac{1}{2}} = 21.62$$

$$S_{kr} = \frac{1}{n}\frac{\sum_{i=1}^{n}(X_i - \bar{X})^4}{S^4} = \frac{1}{5}\frac{-20^4 + (-19^4) + 22^4 + 24^4 + (-7^4)}{21.62^4} = 0.7861$$

Types of Kurtosis

The types of kurtosis are determined by the excess kurtosis of a particular distribution. The excess kurtosis can take positive or negative values, as well as values close to zero.

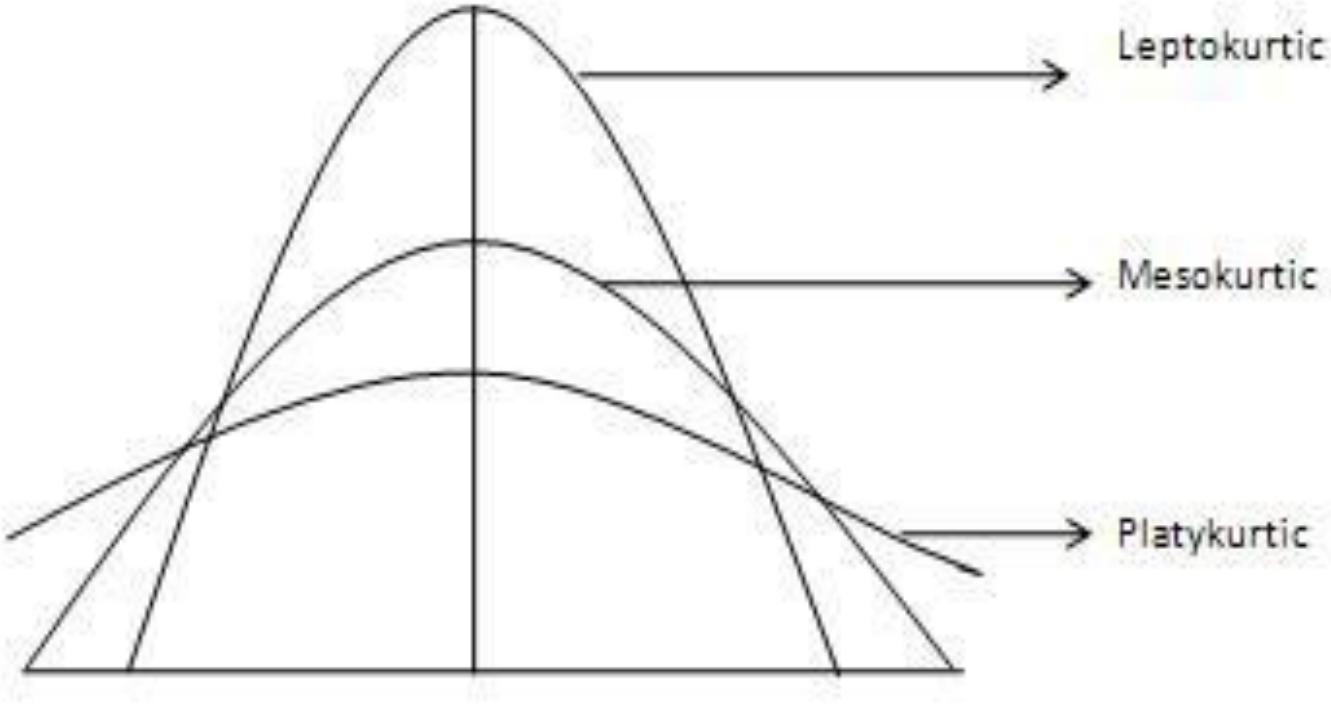
Excess Kurtosis = Kurtosis – 3

1. **Mesokurtic:** Data that follows a mesokurtic distribution shows an excess kurtosis of zero or close to zero. It means that if the data follows a normal distribution, it follows a mesokurtic distribution.

2. **Leptokurtic**: Leptokurtic indicates a positive excess kurtosis. The leptokurtic distribution shows heavy tails on either side, indicating the large outliers. In finance, a leptokurtic distribution shows that the investment returns may be prone to extreme values on either side. Therefore, an investment whose returns follow a leptokurtic distribution is considered to be risky.

3. **Platykurtic:** A platykurtic distribution shows a negative excess kurtosis. The kurtosis reveals a distribution with flat tails. The flat tails indicate the small outliers in a distribution. In the finance context, the platykurtic distribution of the investment returns is desirable for investors because there is a small probability that the investment would experience extreme returns.

Types of Kurtosis cont...



Outliers

- An outlier is a value or an entire observation (row) 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. In specific, 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**.
- When dealing with outliers, it is best to run the analyses two ways: with the outliers and without them.

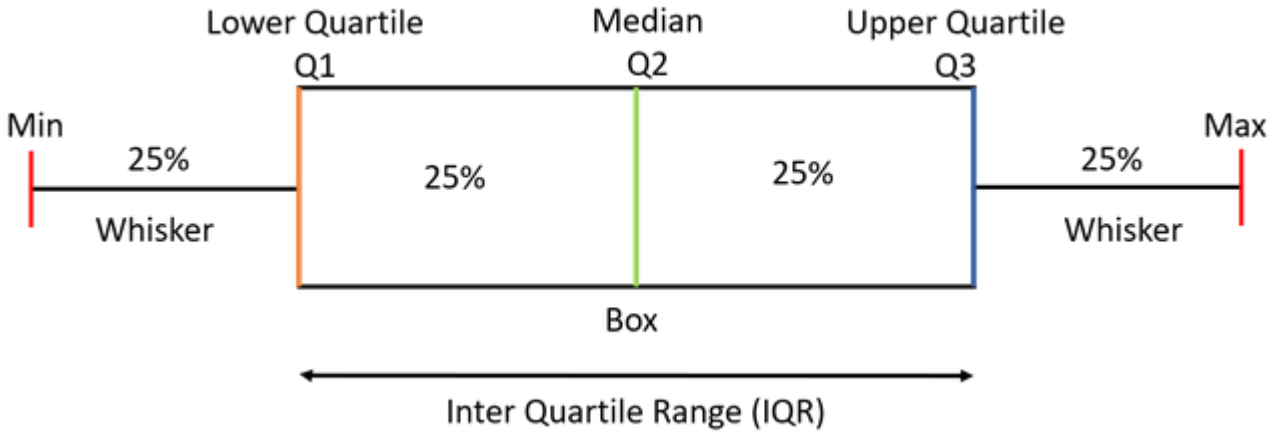
For example, let us consider a row of data [10,15,22,330,30,45,60]. In this dataset, it can be easily concluded that 330 is way off from the rest of the values in the dataset, thus 330 is an outlier. It was easy to figure out the outlier in such a small dataset, but when the dataset is huge, various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

Types of outliers

Outliers are of three types, namely:

- **Type 1: Global (or Point) Outliers -** These are the simplest form of outliers. If, in a given dataset, a data point strongly deviates from all the rest of the data points, it is known as a global outlier. Mostly, all of the outlier detection methods are aimed at finding 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.
- **Type 2: Contextual (or Conditional) Outliers -** A data point is considered a contextual outlier if its value significantly deviates from the rest the data points in the same context. This also means that same value may not be considered an outlier if it occurred in a different context. For example, a temperature reading of 40°C may behave as an outlier in the context of a "winter season" but will behave like a normal data point in the context of a "summer season".
- **Type 3: Collective Outliers** - If in a given dataset, some of the data points, as a whole, deviate significantly from the rest of the dataset, they may be termed as collective outliers. Here, the individual data objects may not be outliers, but when seen as a whole, they may behave as outliers. 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.

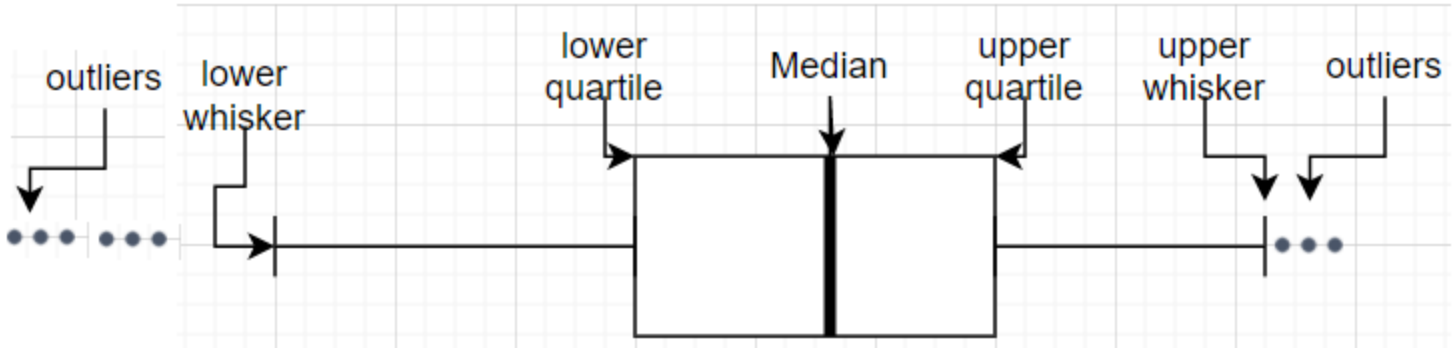
Box Plot



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 dataset excluding the outliers.
- Inter Quartile Range (IQR) is the difference of Q3 and Q1 i.e., Q3 – Q1

Outliers Detection with Box Plot



- Two lines (called whiskers) outside the box extend to the smallest (i.e., lower whisker) and largest (i.e., upper whisker) observations. The upper and lower whiskers can be understood as the boundaries of data, and any data lying outside it are the outliers.
- Lower whisker = Q1-1.5*(Q3-Q1)
- Upper whisker = Q3+1.5*(Q3-Q1)

Outliers Detection with Z score

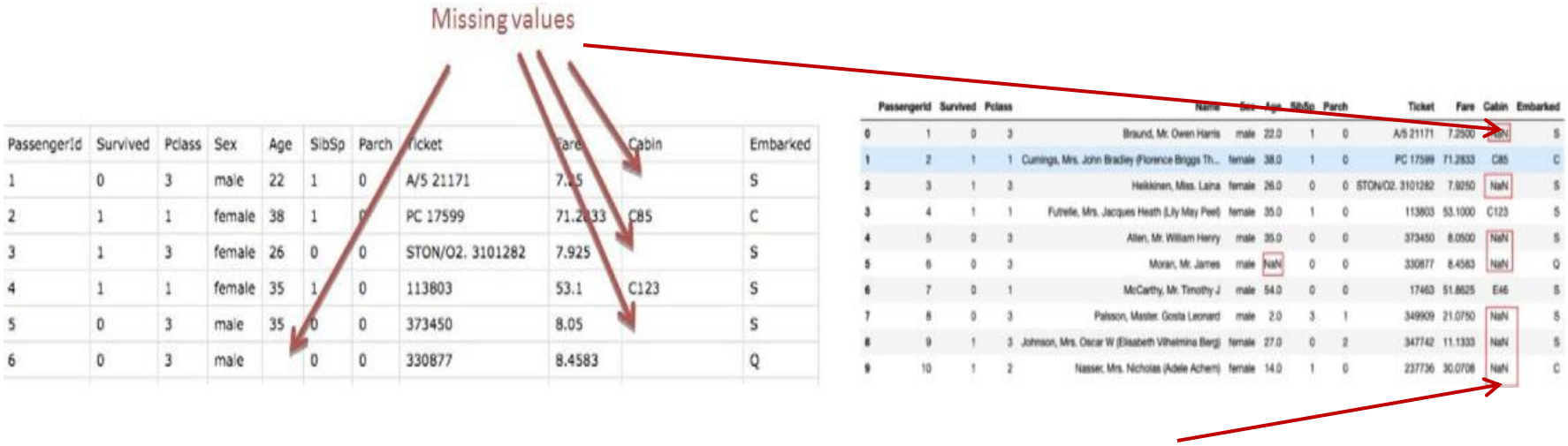
- A z-score tells where the data point lies on a distribution curve i.e., how far from the mean a data point is.

$$Z = \frac{x_i - \bar{x}}{s}$$

- In the above case, x_i is the respective data point, $\bar{x}$ is the mean and s is the standard deviation.
- Usually z-score =3 is considered as a cut-off value to set the limit. Therefore, any z-score greater than +3 or less than -3 is considered as outlier.

Missing Values

- Most real data sets have gaps in the data. There are two issues: how to detect these **missing values** and **what to do about them.**
- **Missing values** occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data.



Missing values are often represented by **NaN**. It stands for **Not a Number.**

Missing Values cont…

What to do about the missing values:

- One option is to simply ignore them. Then you will have to be aware of how the software deals with missing values.
- Another option is to fill in missing values with the average of non missing values, but this isn't usually a very good option.
- A third option is to examine the non missing values in the row of a missing value; these values might provide clues on what the missing value should be.

Types of missing value:

- Missing Completely At Random (MCAR)
- Missing At Random (MAR)
- Missing Not At Random (MNAR)

Missing Completely At Random

- In MCAR, the probability of data being missing is the same for all the observations. In this case, there is no relationship between the missing data and any other values observed or unobserved (the data which is not recorded) within the given dataset. That is, missing values are completely independent of other data. There is no pattern.
- In the case of MCAR data, the value could be missing due to human error, some system/equipment failure, loss of sample, or some unsatisfactory technicalities while recording the values.
- For Example, suppose in a library there are some overdue books. Some values of overdue books in the computer system are missing. The reason might be a human error, like the librarian forgetting to type in the values. So, the missing values of overdue books are not related to any other variable/data in the system. It should not be assumed as it's a rare case.

Missing At Random

- MAR data means that the reason for missing values can be explained by variables on which one have complete information, as there is some relationship between the missing data and other values/data.
- In this case, the data is not missing for all the observations.
- It is missing only within sub-samples of the data, and there is some pattern in the missing values.
- For example, in the survey data, one may find that all the people have answered their 'Gender,' but 'Age' values are mostly missing for people who have answered their 'Gender' as 'female.' (The reason being most of the females don't want to reveal their age.)
- So, the probability of data being missing depends only on the observed value or data. In this case, the variables 'Gender' and 'Age' are related. The reason for missing values of the 'Age' variable can be explained by the 'Gender' variable, but one can not predict the missing value itself.

Missing Not At Random

- If the missing data does not fall under the MCAR or MAR, it can be categorized as MNAR. It can happen due to the reluctance of people to provide the required information. A specific group of respondents may not answer some questions in a survey.
- For example, suppose the name and the number of overdue books are asked in the poll for a library. So most of the people having no overdue books are likely to answer the poll. People having more overdue books are less likely to answer the poll. So, in this case, the missing value of the number of overdue books depends on the people who have more books overdue.
- Another example is that people having less income may refuse to share some information in a survey or questionnaire.

Finding relationships among variables

Introduction

- Fundamentally, it means that the values of one variable correspond to the values of another variable, for each case in the dataset.
- If the variables are perfectly related, then knowing the value of one variable tells exactly what the value of the other variable is.
- Consider a data set with two categorical variables, smoking and drinking. Each person is categorized into one of three smoking categories: nonsmoker (NS), occasional smoker (OS), and heavy smoker (HS). Similarly, each person is categorized into one of three drinking categories: nondrinker (ND), occasional drinker (OD), and heavy drinker (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?
- Therefore, it is very important to understand relationship between variables to draw the right conclusion. Without an understanding of this, wrong results can be inferred from the data.
- Types of relationships among variables:
  - Categorical vs. Categorical
  - Numerical vs. Numerical

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 same is true of examining relationships between two categorical variables.
- One can find the counts of the categories of either variable separately, and more importantly, one can find counts of the joint categories of the two variables.
- It is customary to display all such counts in a table called a **crosstabs** (also sometimes called a **contingency table**).

Example

- **Question:** Is there any indication that smoking and drinking habits are related? If so, how are they related?
- **Objective:** To use a crosstabs to explore the relationship between smoking and drinking.
- **Solution:** Data set lists the smoking and drinking habits of 8761 adults.

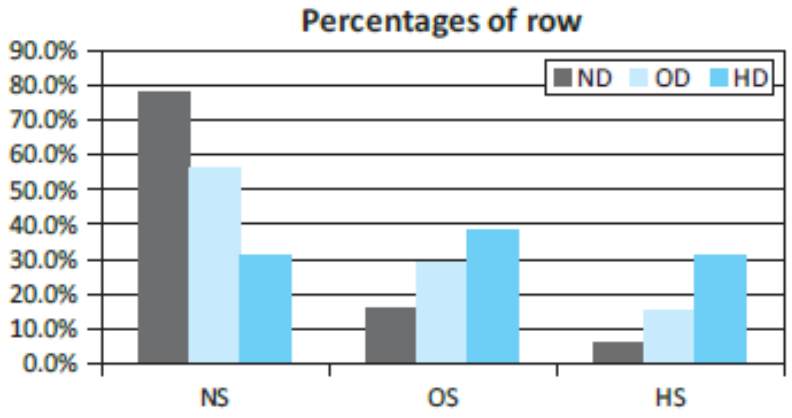
	A	B	C
1	**Person**	**Smoking**	**Drinking**
2	1	NS	OD
3	2	NS	HD
4	3	OS	HD
5	4	HS	ND
6	5	NS	OD
7	6	NS	ND
8	7	NS	OD
9	8	NS	ND
10	9	OS	HD
11	10	HS	HD

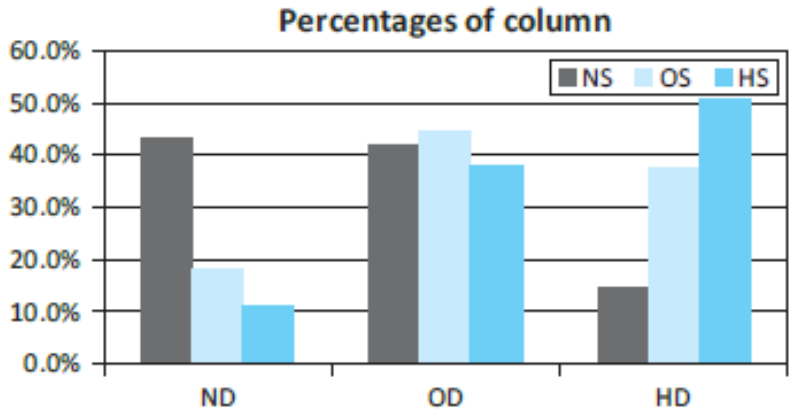
	E	F	G	H	I
1	Crosstabs from COUNTIFS formulas				
2					
3		NS	OS	HS	Total
4	ND	2118	435	163	2716
5	OD	2061	1067	552	3680
6	HD	733	899	733	2365
7	Total	4912	2401	1448	8761
8					
9	Shown as percentages of row				
10		NS	OS	HS	Total
11	ND	78.0%	16.0%	6.0%	100.0%
12	OD	56.0%	29.0%	15.0%	100.0%
13	HD	31.0%	38.0%	31.0%	100.0%

15	Shown as percentages of column			
16		NS	OS	HS
17	ND	43.1%	18.1%	11.3%
18	OD	42.0%	44.4%	38.1%
19	HD	14.9%	37.4%	50.6%
20	Total	100.0%	100.0%	100.0%

Example cont…

- The middle table indicates that only 6% of the nondrinkers are heavy smokers, whereas 31% of the heavy drinkers are heavy smokers.
- Similarly, the bottom table indicates that 43.1% of the nonsmokers are nondrinkers, whereas only 11.3% of the heavy smokers are nondrinkers.
- In short, these tables indicate that smoking and drinking habits tend to go with one another.
- These tendencies are reinforced by the column charts of the two percentage tables.





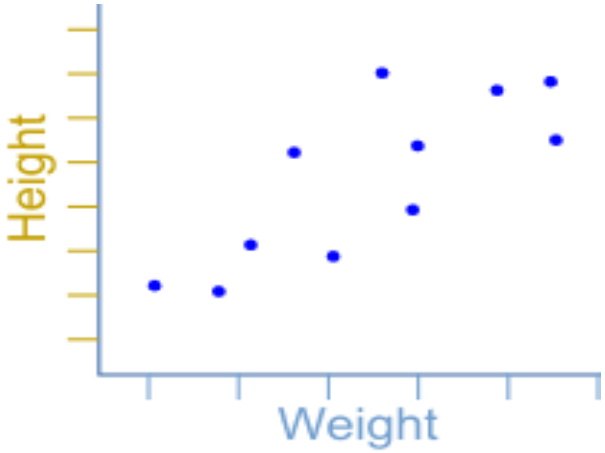
Relationships among numerical variables

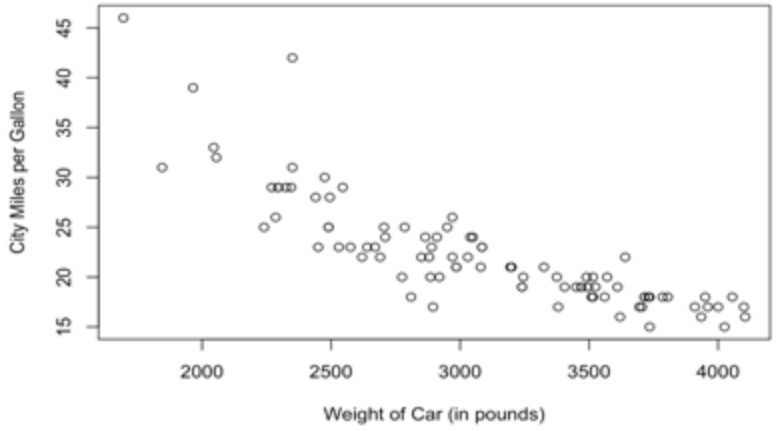
- To study relationships among numerical variables, a new type of chart, called a **scatterplot**, and two new summary measures, **correlation** and **covariance**, are used.
- However, they are appropriate only for truly numerical variables, not for categorical variables that have been coded numerically.

Correlation	Covariance
Indicates the direction and strength	Indicates the direction
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
It can be between – 1 to + 1	It can be between - ∞ to + ∞
There are multiple variations of a correlation coefficient i.e., the **Pearsons correlation coefficient,** the Spearman's Rho and the Kendall's Tau.	

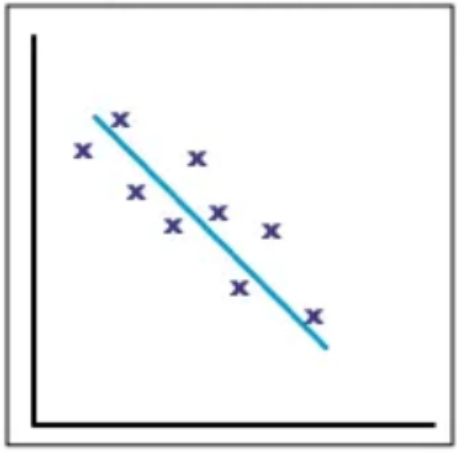
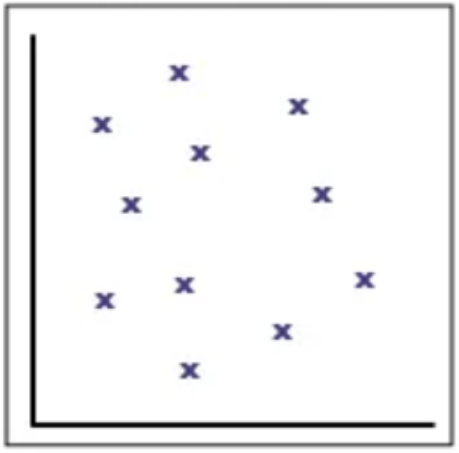
Scatterplots

- A scatterplot is a scatter of points, where each point denotes the values of an observation for two selected variables.
- 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.



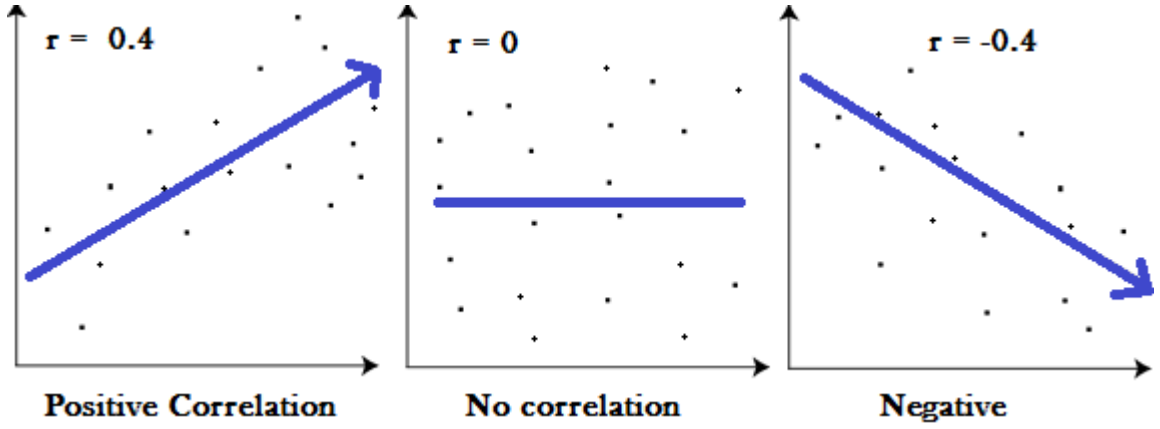


Scatterplots cont…

Positive correlation	Negative correlation	No correlation
		
The points lie close to a straight line, which has a positive gradient. This shows that as one variable **increases** the other **increases**.	The points lie close to a straight line, which has a negative gradient. This shows that as one variable **increases**, the other **decreases**.	There is no pattern to the points. This shows that there is **no connection** between the two variables.

Correlation

Correlation is a technique that can show whether and how strongly pairs of variables are related. 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. Nonetheless, the average weight of people 5'5'' is less than the average weight of people 5'6'', and their average weight is less than that of people 5'7'', etc. 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.



Correlation cont…

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

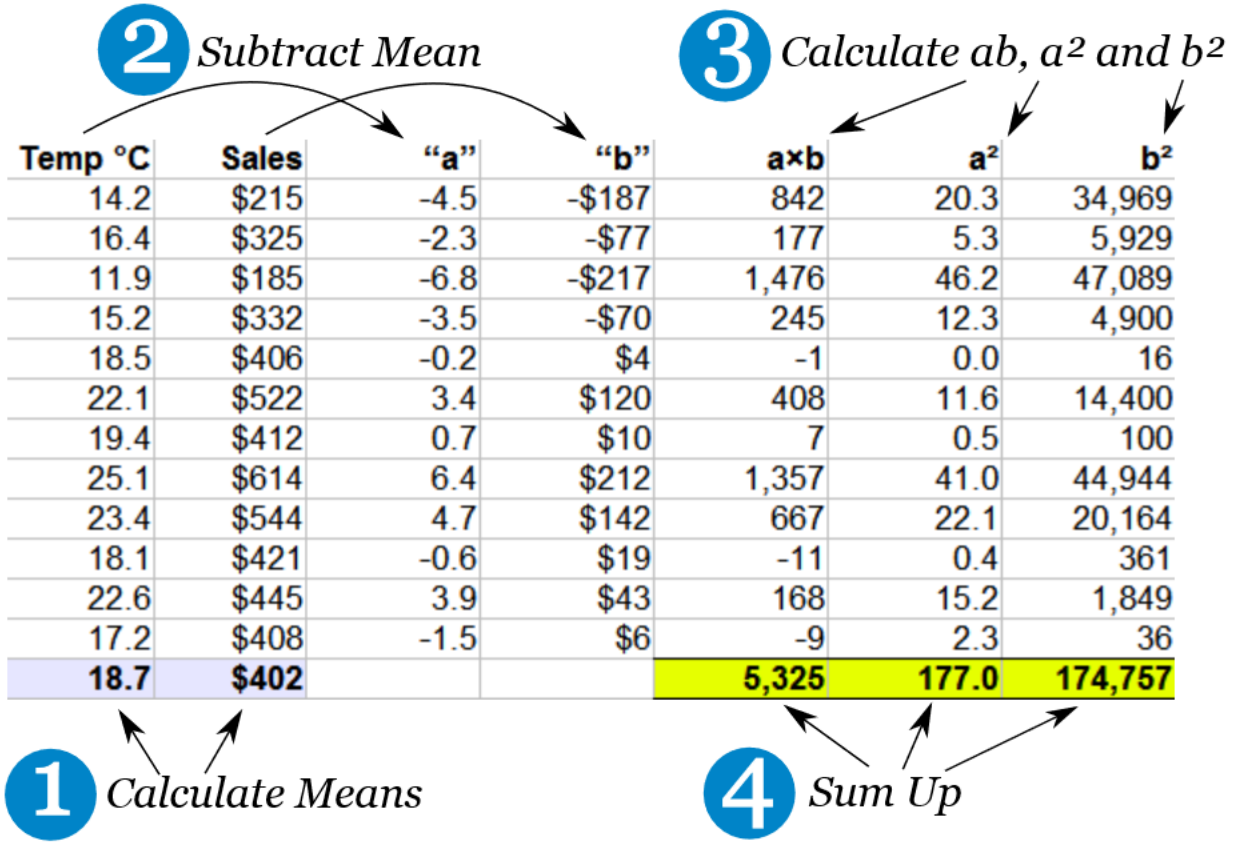
How to calculate Correlation Coefficient?

Let us call the two sets of data "x" and "y" (in our case Temperature is x and Ice Cream Sales is y)

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 (call them "a"), do the same for y (call them "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)]

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Correlation Coefficient Calculation



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**

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

Correlation Coefficients Calculation

Company	Sales in 1000s (Y)	Number of agents in 100s (X)
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

- $n = 10, \sum X = 80, \sum Y = 255, \sum XY = 2289$
- $\sum X^2 = 756, \sum Y^2 = 7097, (\sum X)^2 = 6400, (\sum Y)^2 = 65025, r = 0.95$

Class Exercise

Find the correlation coefficients of the below sample.

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

Covariance

It is a measure of the relationship between two variables.

Example: John is an investor. His portfolio primarily tracks the performance of the S&P 500 and John wants to add the stock of ABC Corp. Before adding the stock of ABC Corp to his portfolio, he wants to assess the directional relationship between the stock of ABC Corp and the S&P 500.

John does not want to increase the unsystematic risk of his portfolio. Thus, he is not interested in owning securities in the portfolio that tend to move in the same direction.

John can calculate the covariance between the stock of ABC Corp. and S&P 500. Unlike the correlation coefficient, covariance is measured in units and it can be positive or negative values. The values are interpreted as follows:

- **Positive covariance:** Indicates that two variables tend to move in the same direction.
- **Negative covariance:** Reveals that two variables tend to move in inverse directions.

Covariance vs. Correlation

Using covariance, we can only gauge the direction of the relationship (whether the variables tend to move in tandem or show an inverse relationship). However, it does not indicate the strength of the relationship, nor the dependency between the variables.

On the other hand, correlation measures the strength of the relationship between variables. Correlation is the scaled measure of covariance. It is dimensionless. In other words, the correlation coefficient is always a pure value and not measured in any units.

The relationship between the two concepts can be expressed using the formula:

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

Where

$\rho(X,Y)$ – the correlation between the variables X and Y

$Cov(X,Y)$ – the covariance between the variables X and Y

σX – the standard deviation of the X-variable

σY – the standard deviation of the Y-variable

Covariance cont…

The formula for calculating covariance of sample data is shown below.

$$\text{Cov}(X, Y) = \frac{\sum(X_i - \bar{X})(Y_j - \bar{Y})}{n - 1}$$

Where:

X_i – the values of the X-variable

Y_j – the values of the Y-variable

$\bar{X}$ – the mean (average) of the X-variable

$\bar{Y}$ – the mean (average) of the Y-variable

n – the number of data points

The prices of ABC Corp. and the S&P 500 are as follows. Find the covariance.

Year	S&P 500	ABC Corp
2013	1692	68
2014	1978	102
2015	1884	110
2016	2151	112
2017	2519	154

Sampling and distributions

Population

In any sampling problem there is a **relevant population**. The population is the set of all members about which a study intends to make inferences, where an inference is a statement about a numerical characteristic of the population, such as an average income or the proportion of incomes below $50,000. The different types of population such as:

- **Finite Population:** When the number of elements of the population is fixed and thus making it possible to enumerate it in totality, the population is said to be finite. Example is the books in a library, as it can be calculated easily and the cars in a town.
- **Infinite Population:** When the number of units in a population are uncountable, and so it is impossible to observe all the items of the universe, then the population is considered as infinite. Example is the births of insect, as one cannot calculate the birth of insects easily.
- **Existent Population:** The population which comprises of objects that exist in reality is called existent population. Examples are books, students etc.
- **Hypothetical Population:** Hypothetical or imaginary population is the population which exists hypothetically. Examples are an outcome of rolling the dice, the outcome of tossing a coin.

Sample

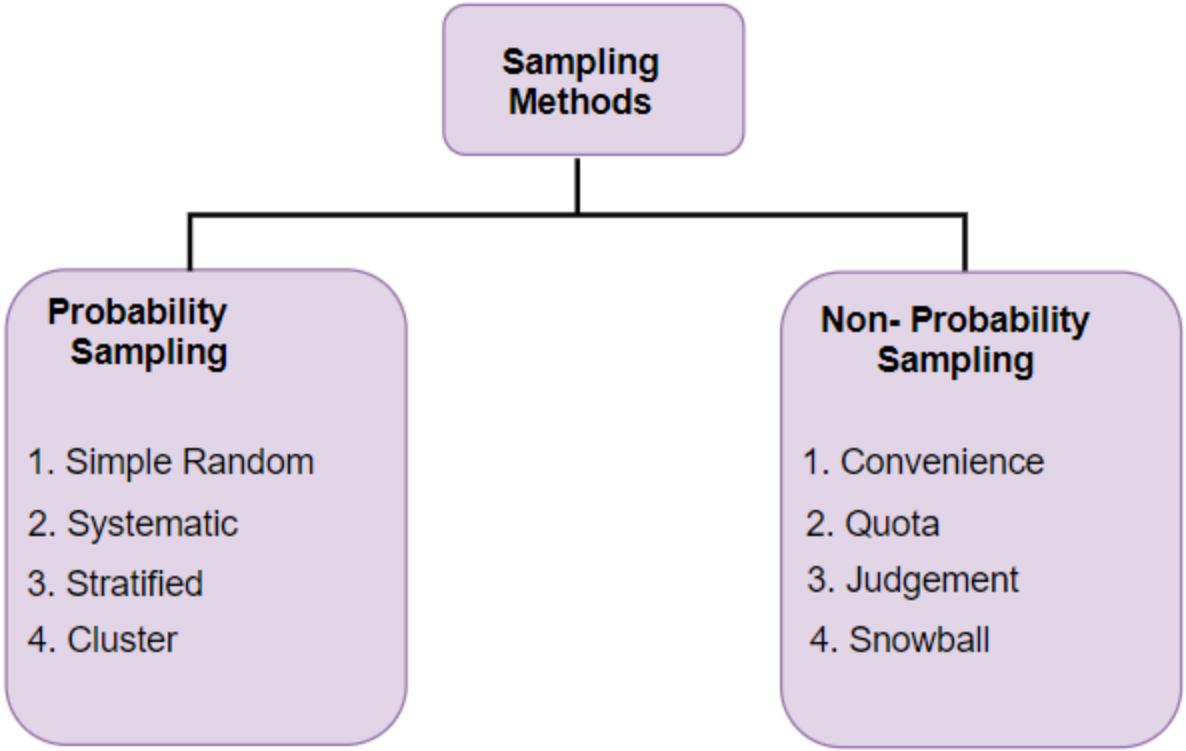
- **Sampling:** The selection of subset of the population
- **Sampling Unit:** The population divided into a finite number of distinct and identifiable units is called sampling units. In other words, the individuals whose characteristics are to be measured in the analysis are called sampling units.
- **Sampling Frame:** It is the list of all the sampling units with a proper identification. The sampling frame or frame should be accurate, free from omission and duplication (overlapping), adequate, and must cover the whole of the population and should be well identified. For example, the sampling frame is a list of all sampling units consisting of :
  - List of villages in a region
  - List of household within a village
  - List of schools in the district

Sampling Method

- To draw valid conclusions from the results, one have to carefully decide how to select a sample that is representative of the group as a whole. This is called a **sampling method**.
- There are two primary types of sampling methods i.e., **probability sampling** and **non-probability sampling**.

Probability sampling	Non-probability sampling
Any element can be chosen randomly from the population. It deals with choosing the sample randomly.	Every element is chosen on the subjective judgment (purposefully/intentionally) from the population on the basis of certain past experience & knowledge rather than random selection.
Every individual element in the population has a known and equal chance of getting selected.	Every single individual elements in the population may not have an opportunity to be chosen as a sample.
Example: When an unbiased coin is thrown (randomly), the probability of getting the head is ½.	**Example:** One person could have a 10% chance of being selected and another person could have a 50% chance of being selected.

Types of Sampling Method

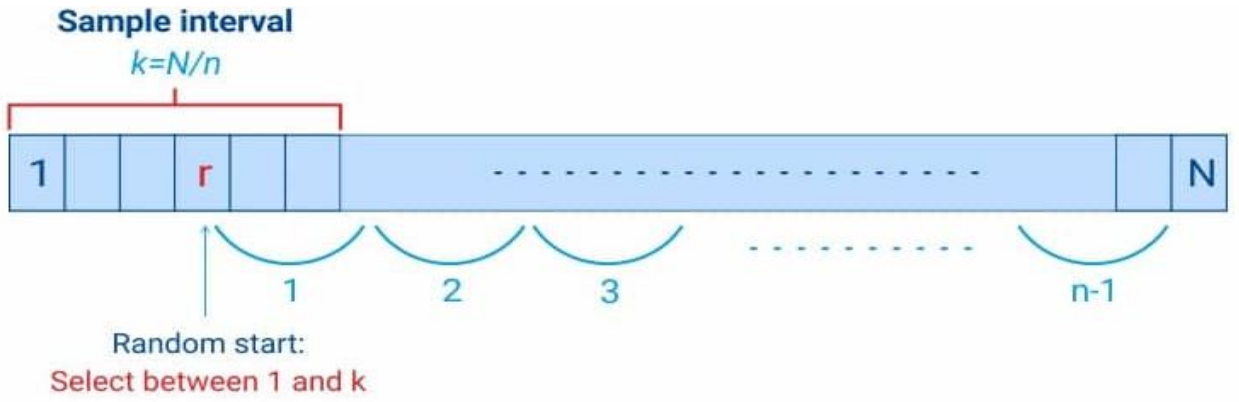


Simple Random Sampling

- It is a randomly selected subset of a population, wherein each member of the population has an exactly equal chance of being selected.
- The samples are determined by assigning sequential values to each item within a population, then randomly selecting those values.
- An example of a simple random sample would be the names of 25 employees being chosen out of a hat from a company of 250 employees. In this case, the population is all 250 employees, and the sample is random because each employee has an equal chance of being chosen.
- Simple random sampling works best if there is a lot of time and resources to conduct the study, or one is studying a limited population that can easily be sampled.

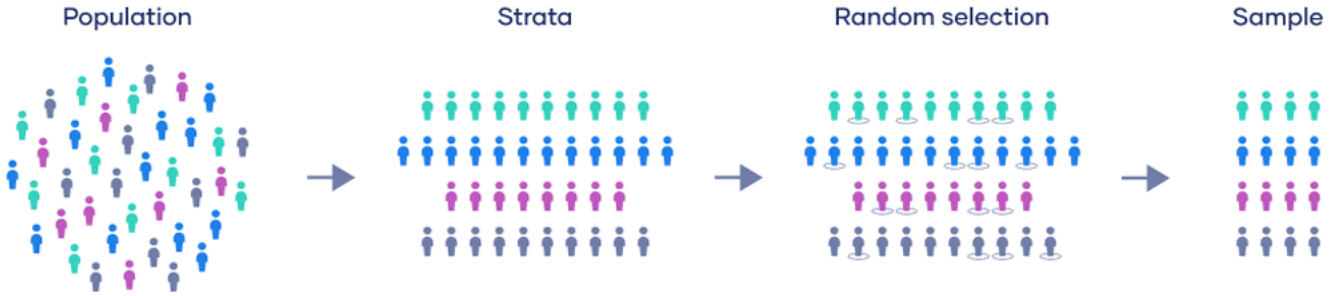
Systematic Sampling

- Sample members from a larger population are selected according to a random starting point but with a fixed, periodic interval.
- This interval, called the **sampling interval**, is calculated by dividing the population size by the desired sample size. The sampling or skip interval (k) = N (total population units)/n (sample size)
- For instance, if a local NGO is seeking to form a systematic sample of 500 volunteers from a population of 5000, they can select every 10th person in the population to build a sample systematically.



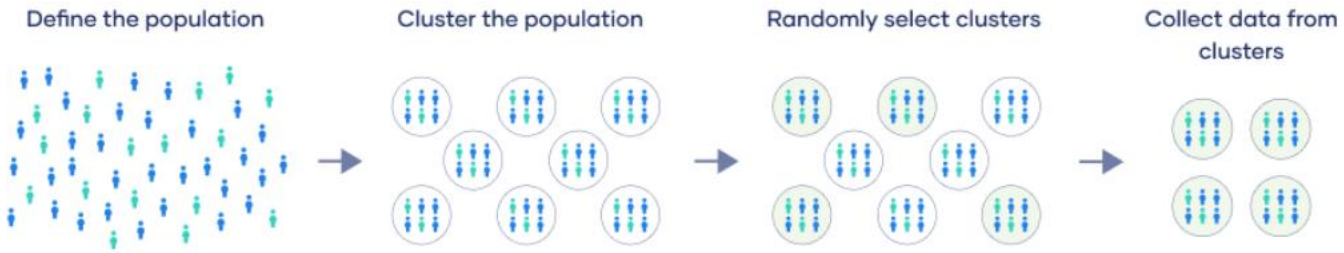
Stratified Sampling

- The population is divided into homogeneous subpopulations called strata (the plural of stratum) based on specific characteristics (e.g., race, gender identity, location, etc.). Every member of the population should be in exactly one stratum.
- Each stratum is then sampled using simple random sampling.
- It is used when a population's characteristics are diverse and it is to ensure that every characteristic is properly represented in the sample.



Cluster Sampling

- The population is divided into internally heterogeneous and externally homogeneous subpopulations known as clusters i.e., clusters may be formed by different cities in a country, different areas in a city, different organizations, different universities, different industrial estates, etc.
- Then, randomly select among these clusters to form a sample.
- It is often used to study large populations, particularly those that are widely geographically dispersed.

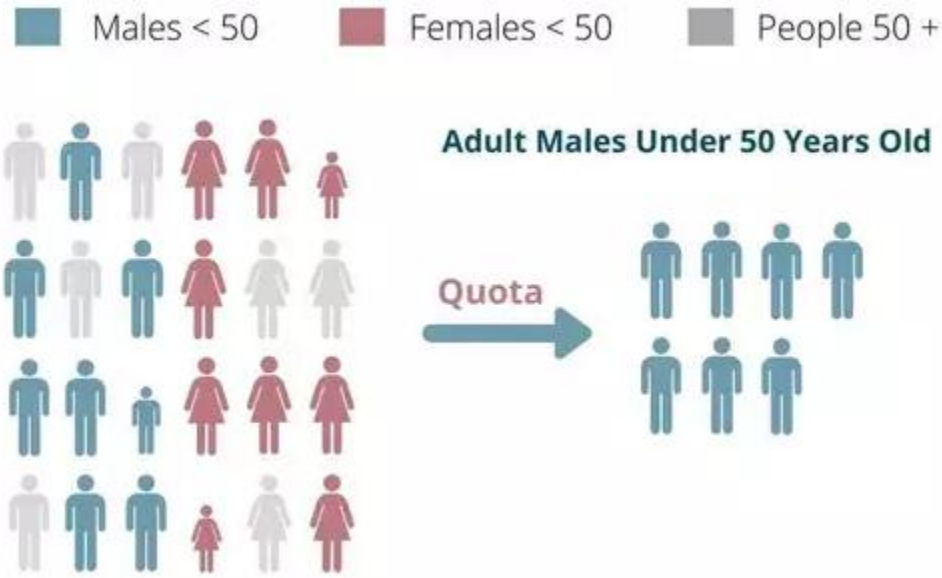


Convenience Sampling

- It involves the sample being drawn from that part of the population that is close to hand i.e., easiest to access. This can be due to geographical proximity, availability at a given time, or willingness to participate.
- Suppose one is researching public perception towards the city of Seattle and determined that a sample of 100 people is sufficient to answer a question. To collect the data, one can stand at a subway station and approach passersby, asking them whether they want to participate in the survey. One should continue to ask until the sample size is reached.
- Another example would be a new NGO wants to establish itself in 20 cities. It selects the top 20 cities to serve based on the proximity to where they are based.

Quota Sampling

- A sample is created according to specific traits or qualities.
- For example, a cigarette company wants to find out what age group prefers what brand of cigarettes in a particular city. They apply survey quota on the age groups of 21-30, 31-40, 41-50, and 51+. From this information, they gauges the smoking trend among the population of the city.

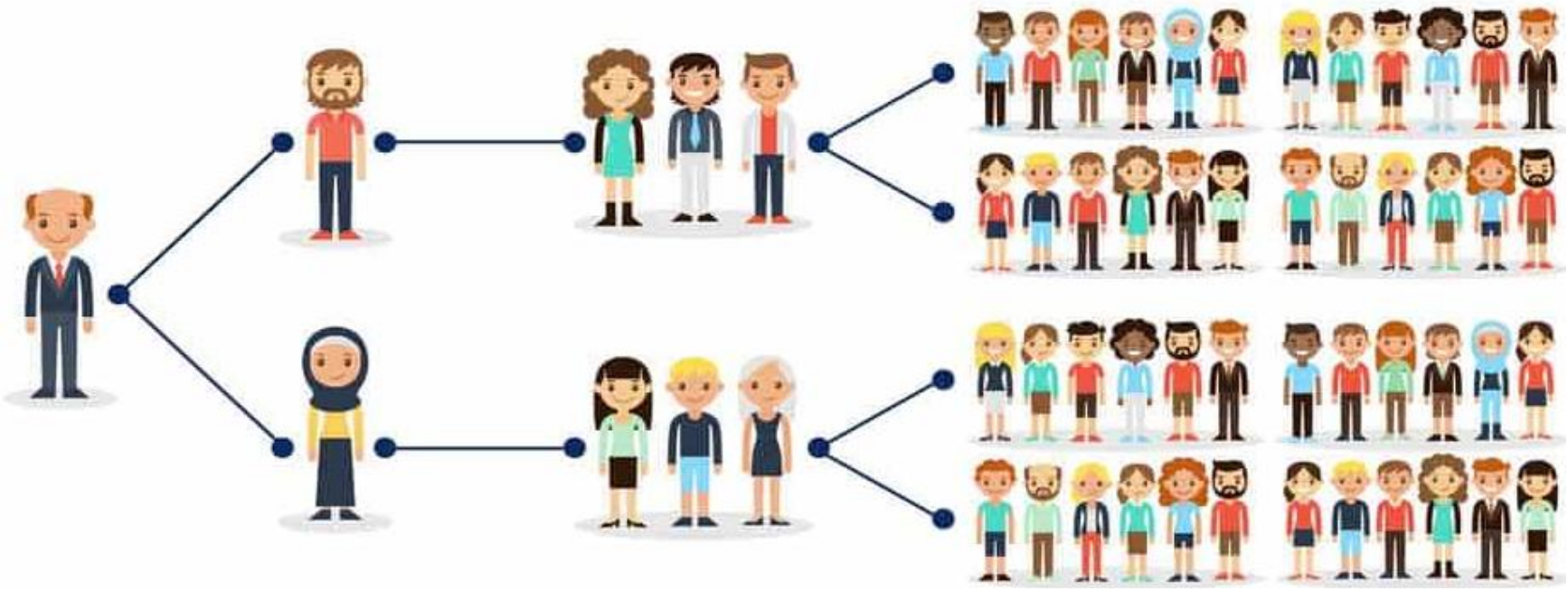


Judgment Sampling

- The sample is selected based on his/her own existing knowledge, or his/her professional judgment.
- It is also called purposive sampling or authoritative sampling.
- It is usually used in situations where the target population comprises of highly intellectual individuals who cannot be chosen by using any other probability or non-probability sampling technique.
- Consider a scenario where a panel decides to understand what are the factors which lead a person to select ethical hacking as a profession. Ethical hacking is a skill which has been recently attracting youth. More and more people are selecting it as a profession. The researchers who understand what ethical hacking is will be able to decide who should form the sample to learn about it as a profession. That is when judgmental sampling is implemented. Researchers can easily filter out those participants who can be eligible to be a part of the research sample.

Snowball Sampling

- New units are recruited by other units to form part of the sample.
- It is also known as chain sampling or network sampling, snowball sampling begins with one or more study participants. It then continues on the basis of referrals from those participants. This process continues until you reach the desired sample, or a saturation point.

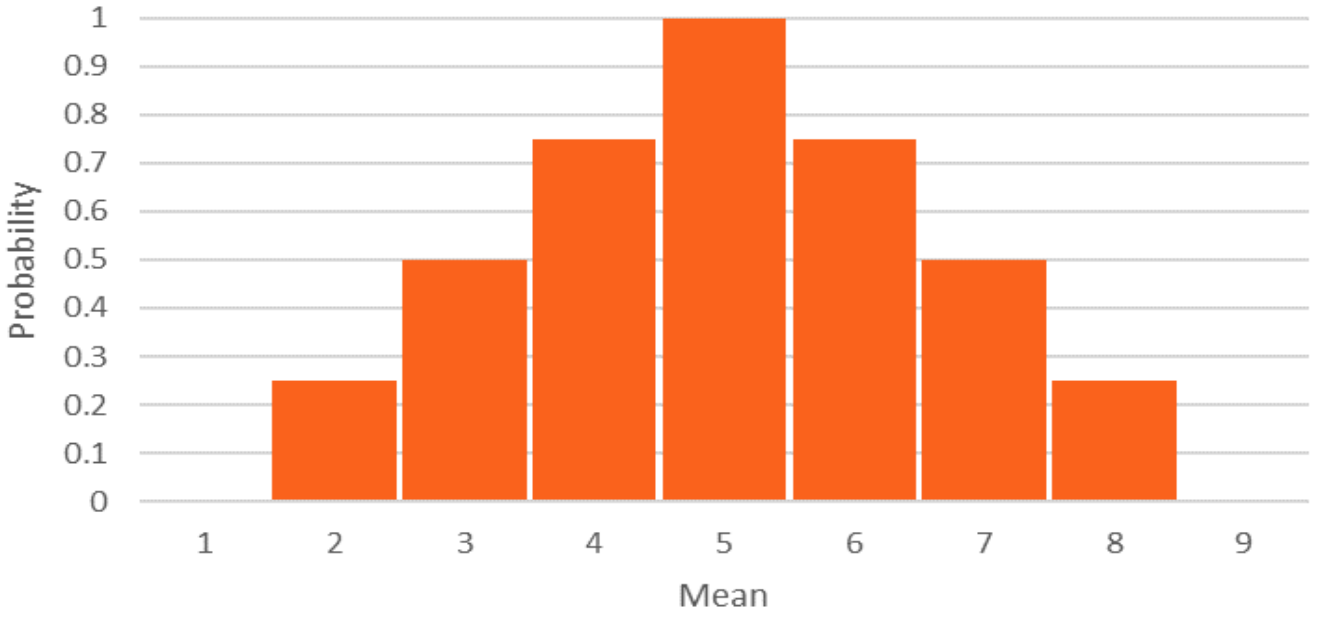


Sampling Error

- Sampling error is defined as the amount of inaccuracy in estimating some value, which occurs due to considering a small section of the population, called the sample, instead of the whole population.
- It occurs when the sample used in the study is not representative of the whole population.
- The formula to find the sampling error is given as follows:
  If N is the sample size and SE is the sampling error, then
  Sampling Error i.e., SE = $(1/\sqrt{N})\ 100$
  For example, a random sample of 1,000 has about a $1/\sqrt{n}$; = 3.2% error
- There are two methods by which this sampling error can be reduced and are:
  - **Increase sample size:** A larger sample size results in a more accurate result because the study gets closer to the actual population size.
  - **Stratification:** The population is classified into different groups called strata, that contain similar units. From each of these strata, a sub-sample is selected in a random manner.

Sampling Distribution

The sampling distribution is a probability distribution based on a large number of samples of size n from a given population. It represents the distribution of frequencies on how spread apart various outcomes will be for a specific population. The probability distribution gives the possibility of each outcome of a random experiment or event. It provides the probabilities of different possible occurrences.



Sampling Distribution cont...

Example: A rowing team consists of four rowers who weigh 152, 156, 160, and 164 pounds. Find all possible random samples with sample of size two and compute the sample mean for each one. Use them to find the probability distribution and plot the sampling distribution.

Solution: The following table shows all possible samples with replacement of size two, along with the mean of each.

Sample	Mean
152,152	152
152,156	154
152,160	156
152,164	158

Sample	Mean
156,152	154
156,156	156
156,160	158
156,164	160

Sample	Mean
160,152	156
160,156	158
160,160	160
106,164	162

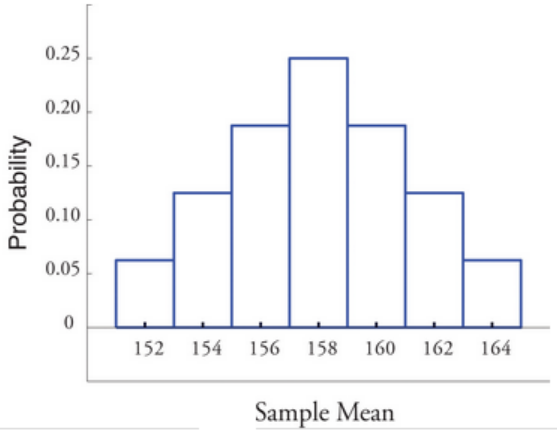
Sample	Mean
164,152	158
164,156	150
164,160	162
164,164	164

Sampling Distribution cont…

- The table (in last page) shows that there are seven possible values of the sample mean $\bar{x}$. The value $\bar{x}$=152 happens only one way (the rower weighing 152 pounds must be selected both times), as does the value $\bar{x}$ =164, but the other values happen more than one way, hence are more likely to be observed than 152 and 164 are. Since the 16 samples are equally likely, we obtain the probability distribution of the sample mean just by counting:

$\bar{x}$	152	154	156	158	160	162	164
P($\bar{x}$)	1/16 =0.0625	2/16 =0.1256	3/16 =0.1875	4/16 =0.25	3/16 =0.1875	2/16 =0.1256	1/16 =0.0625

- Plotting the frequency distribution of each sample statistic and the resulting graph will be the **sampling distribution of mean.**



Sampling Distribution cont…

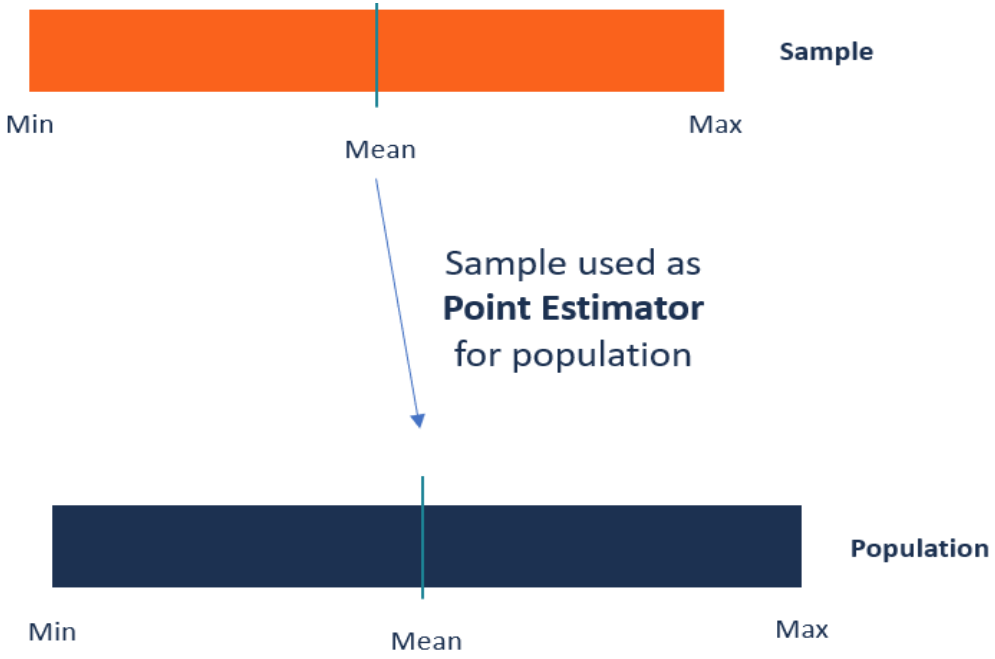
- Its primary purpose is to establish representative results of samples of a comparatively larger population. Since the population is too large to analyze, one can select a smaller group of samples and repeatedly sample or analyze them.
- The gathered data, or statistic, is used to calculate the likely occurrence, or probability, of an event.
- Using a sampling distribution simplifies the process of making inferences, or conclusions, about population.
- Factors that influence the variability of sampling distribution
  - **The number observed in a population:** It is the measure of observed activity in a given group of data.
  - **The number observed in the sample:** 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 to chose the samples can account for variability in some cases.

Estimation

- It is often of interest to learn about the characteristics of a large group of elements such as individuals, households, buildings, products, parts, customers, and so on. All the elements of interest in a particular study form the population. Because of time, cost, and other considerations, data often cannot be collected from every element of the population. In such cases, a subset of the population, called a sample, is used to provide the data.
- Data from the sample are then used to develop **estimates** of the characteristics of the larger population. Therefore, the procedure of making judgment or decision about a population parameter is referred to as **statistical estimation** or **simply estimation**.
- There are two types of estimates namely **point estimation** and **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 such as population mean, population variance, etc are estimated from the corresponding sample statistics such as sample mean, sample variance, etc.



Point Estimation cont…

- Most often, the methods of finding the parameters of large populations are unrealistic. For example, when finding the average age of kids attending kindergarten, it is impossible to collect the exact age of every kindergarten kid in the world. Instead, the point estimator is used to make an estimate of the population parameter.
- It is desirable for a point estimate to be
  - **Consistent:** the larger the sample size, the more accurate the estimate. For the point estimator to be consistent, the expected value should move toward the true value.
  - **Unbiased:** The expectation of the observed values of many samples equals the corresponding population parameter i.e., the sample mean is an unbiased estimator for the population mean.
  - **Most efficient:** The most efficient point estimator is the one with the smallest variance. Generally, the efficiency of the estimator depends on the distribution of the population. For example, in a normal distribution, the mean is considered more efficient than the median, but the same does not apply in asymmetrical distributions.

How to find Point Estimate?

Point Estimate	Population Parameter
S (sample deviation)	σ (population deviation)
$\bar{x}$ (sample mean)	μ (population mean)
S^2 (sample variance)	σ^2 (population variance)

- **Example 1:** A sample of 40 packages of rice has a mean weight of 5.7 kg with a standard deviation of 0.4 kg. Find the best estimate of the population mean?
  Solution: In such a case, the sample mean (i.e., 5.7) is the best point estimate for population mean.

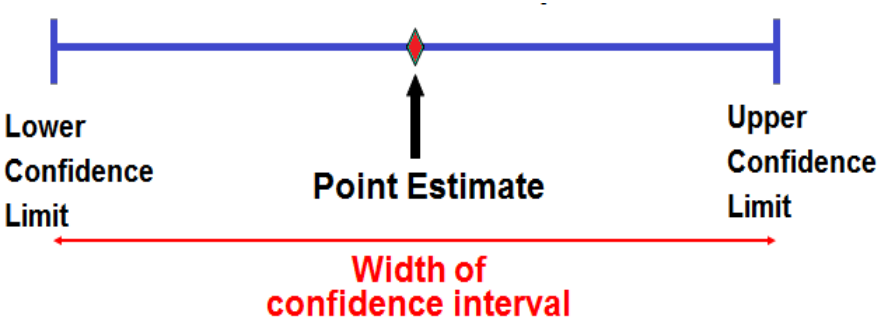
- **Example 2:** calculate the best point estimate from the list of data i.e., 15.22, 14.34, 18.12, 12.61, 15.61, 14.22, 19.41, 12.22, 17.12, 14.22, 12.91 and 18.12.
  Solution: In such a case, the sample mean (i.e., 15.34) is the best point estimate for population mean.

Interval Estimation

- An **interval** is a **range of values**. Let's say we wanted to find out the average cigarette use of senior citizens. We can't survey every senior citizen on the planet (due to time constraints and finances), so we take a sample of 1000 senior citizens and find that 10% of them smoke cigarettes. Although we have only taken a sample, we can use that figure to estimate that "about" 10% of the whole population smoke cigarettes. In reality, it's unlikely to be exactly 10% (as we only sampled a small percentage of people), but it's probably somewhere around there, perhaps between 5 and 15%. That "somewhere between 5 and 15%" is an interval estimate.
- There's nothing wrong with making a good guess at an interval, but sometimes we want to be very confident that our results are sound and repeatable. "Repeatable" means that if we do the whole thing over again, we'll get the same results. One way to do this is to express a **confidence level**. Confidence levels are percentages of certainty. For example, we might say we are 99% confident (i.e., we have a 99% confidence level) that between 5 and 15% of senior citizens smoke cigarettes. When the interval estimate has a confidence level attached, it's called a **confidence interval**.

Confidence Interval Estimation

- The lower bound (in the example, 5%) is called a lower confidence limit and the upper bound (in the example, 15%) is called an upper confidence limit.



- The bigger the sample size, the more narrow the confidence interval will be.
- How to determine the lower and upper confidence limit?

Confidence limit $\longrightarrow \mu = \bar{x} \pm Z \frac{\sigma}{\sqrt{n}}$

Standard deviation

Sample size

Mean

Z-score

A measure of how many standard deviations are below or above the population mean

- Z-Scores for commonly used confidence intervals are as follows:
  - 90% → 1.645
  - 95% → 1.96
  - 99% → 2.576
  - 80% → 1.282
  - 50% → 0.674
  - 98% → 2.326

Refer Appendix for further details

Confidence Interval Estimation cont…

Suppose a student measuring the boiling temperature of a certain liquid observes the readings (in degrees Celsius) 102.5, 101.7, 103.1, 100.9, 100.5, and 102.2 on 6 different samples of the liquid. What is the interval estimation for the population mean at a 95% confidence level?

Solution:

The sample mean of the boiling temperatures to be 101.82, with the standard deviation σ=0.984. The confidence level is 95% and the sample size is 6. The Z-score for 95% confidence level is 1.96.

$\mu = 101.82 \pm 1.96 * (0.984/\sqrt{6}) = 106.62, 97.02$

So, **upper confidence limit**= 106.621 and **lower confidence limit**= 97.019

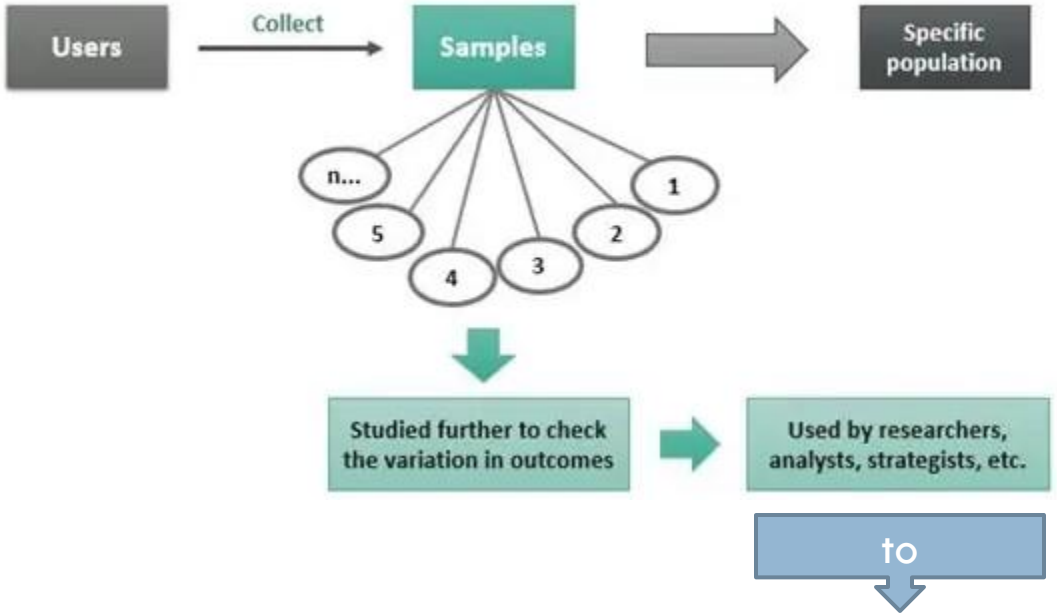
Standard error (SE) = $\sigma / \sqrt{n} = 0.402$ → tells how accurately the sample reflects the total population (measures the preciseness of an estimate of a population mean)

Margin of error = $Z^*(\sigma/\sqrt{n}) = 1.96 * (0.984/2.45) = 0.786$ → number of random sample errors in the data that we are measuring (measures the half-width of a confidence interval for a population mean)

Problem statement: The sample with the test scores in data analytics after end semester examination are 55, 65, 80, 95, 90, 90, 95, 75, 75, 85, 90 and 80. Calculate the confidence limit and margin error. Consider 95% confidence level.

Sampling Distributions

Recap



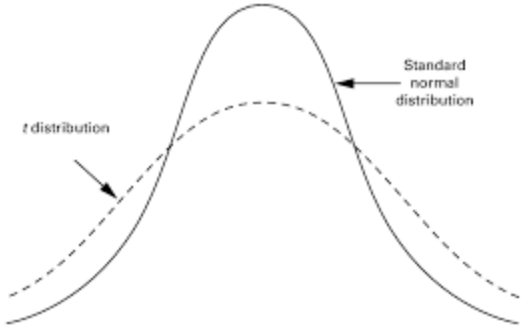
represents the probability of varied outcomes when a study is conducted.

There are 2 types of sampling distributions i.e.,

- Sampling distribution of mean – **[Discussed earlier]**
- T-distribution

T-Distribution

- The t-distribution describes the **standardized distances of sample means** to the **population mean** when the **population standard deviation is not known**, and the observations come from a normally distributed population.
- The t-distribution is similar to normal distribution but flatter and shorter than a normal distribution i.e., it is symmetrical, bell-shaped distribution, similar to the standard normal curve.



- The height of the t-distribution depends on the **degrees of freedom** (df) and refers to the maximum number of logically independent values, which are values that have the freedom to vary, in the sample.

Degree of freedom

The easiest way to understand degrees of freedom conceptually is through several examples.

- Consider a data sample consisting of five positive integers. The values of the five integers must have an average of six. If four of the items within the data set are {3, 8, 5, and 4}, the fifth number must be 10. Because the first four numbers can be chosen at random, the degrees of freedom is four.
- Consider a data sample consisting of one integer. That integer must be odd. Because there are constraints on the single item within the data set, the degrees of freedom is zero.
- The formula to determine degrees of freedom is df = N – 1 where N is sample size.
- For example, imagine a task of selecting 10 baseball players whose bating average must average to .250. The total number of players that will make up our data set is the sample size, so N = 10. In this example, 9 (10 - 1) baseball players can theoretically be picked at random, with the 10th baseball player having to have a specific batting average to adhere to the .250 batting average constraint.

T-Distribution cont…

- As the df increases, the t-distribution will get closer and closer to matching the standard normal distribution.
- The values of the t-statistic is : $t = [\bar{x} - \mu] / [s / \sqrt{n}]$ where,
  t = t score,
  $\bar{x}$ = sample mean,
  μ = population mean,
  s = standard deviation of the sample,
  n = sample size
  Note: A t-score is equivalent to the number of standard deviations away from the mean of the t-distribution.
- A law school claims it's graduates earn an average of $300 per hour. A sample of 15 graduates is selected and found to have a mean salary of $280 with a sample standard deviation of $50. Assuming the school's claim is true, what is the t-score?
  Solution: $t = (280 - 300) / (50/\sqrt{15}) = -20 / 12.909945 = -1.549$.

T-Distribution cont…

- Student's t distribution is used when
  - The sample size must be 30 or less than 30.
  - The population standard deviation(σ) is unknown.
  - The population distribution must be unimodal and skewed.
- **Note:**

  The t-score represents the number of standard errors by which the sample mean differs from the population mean. For example, if a t-score is 2.5, the sample mean is 2.5 standard errors above the population mean. If a t-score is −2.5, the sample mean is 2.5 standard errors below the population mean.

Inferential Statistics

- Statistics can be classified into two different categories i.e., **descriptive statistics** and **inferential statistics.**
- The *descriptive statistics* summarizes the features of the dataset, whereas *inferential statistics* help to make conclusion from the data.
- Inferential statistics is the process of using a sample to infer the properties of a population and allows to generalize the population.
- In general, inference means "guess", which means making inference about something. So, statistical inference means, making inference about the population.
- Let's look at a real flu vaccine study for an example of making a statistical inference. The scientists for this study want to evaluate whether a flu vaccine effectively reduces flu cases in the general population. However, the general population is much too large to include in their study, so they must use a representative sample to make a statistical inference about the vaccine's effectiveness.
- **Hypothesis testing** is one of the type of inferential statistics.

Hypothesis

- A **hypothesis** is defined as a **formal statement**, which gives the explanation about the relationship between the two or more variables of the specified population i.e., it includes components like variables, population and the relation between the variables.
- Hypothesis example:
  - **Two variables** - if you eat more vegetables, you will lose weight faster. Here, eating more vegetables is an independent variable, while losing weight is the dependent variable.
  - **Two or more dependent variables and two or more independent variables** - Eating more vegetables and fruits leads to weight loss, glowing skin, and reduces the risk of many diseases such as heart disease.
  - Consumption of sugary drinks every day leads to obesity
  - If a person gets 7 hours of sleep, then he will feel less fatigue than if he sleeps less.

Hypothesis Testing

- In today's data-driven world, decisions are based on data all the time. Hypothesis plays a crucial role in that process, whether it may be making business decisions, in the health sector, academia, or in quality improvement. Without hypothesis & hypothesis tests, you risk drawing the wrong conclusions and making bad decisions.
- **Hypothesis testing** is a type of statistical analysis in which assumptions are put about a population parameter to the test. It is used to estimate the relationship between variables.
- Examples:
  - A faculty assumes that 60% of his students come from higher-middle-class families.
  - A doctor believes that 3D (Diet, Dose, and Discipline) is 90% effective for diabetic patients.
- It involves setting up a **null hypothesis** and an **alternative hypothesis**. These two hypotheses will always be mutually exclusive. This means that if the null hypothesis is true then the alternative hypothesis is false and vice versa.

Null Hypothesis and Alternate Hypothesis

- The null hypothesis is the assumption that the **event will not occur**. A null hypothesis has no bearing on the study's outcome unless it is rejected.
- **Example:**
  - Smokers are no more susceptible to heart disease than nonsmokers.
  - The new drug has a cure rate no higher than other drugs on the market.
- H_0 is the symbol for it, and it is pronounced **H-naught**.
- Hypothesis testing is used to conclude if the null hypothesis can be rejected or not. Suppose an experiment is conducted to check if girls are shorter than boys at the age of 5. The null hypothesis will say that they are of the same height.
- The alternate hypothesis is the logical opposite of the null hypothesis. The acceptance of the alternative hypothesis follows the rejection of the null hypothesis.
- It indicates that there is a statistical significance between two possible outcomes and can be denoted as H_a.
- For the above-mentioned example, the alternative hypothesis would be that girls are shorter than boys at the age of 5.
- The null hypothesis is usually the current thinking, or status quo. The alternative hypothesis is usually the hypothesis to be proved. The burden of proof is on the alternative hypothesis.

Null Hypothesis and Alternate Hypothesis cont...

- A sanitizer manufacturer claims that its product kills 95 percent of germs on average. To put this company's claim to the test, create a null and alternate hypothesis.
  - H_0 (Null Hypothesis): Average = 95%.
  - Alternative Hypothesis (H_a): The average is less than 95%.

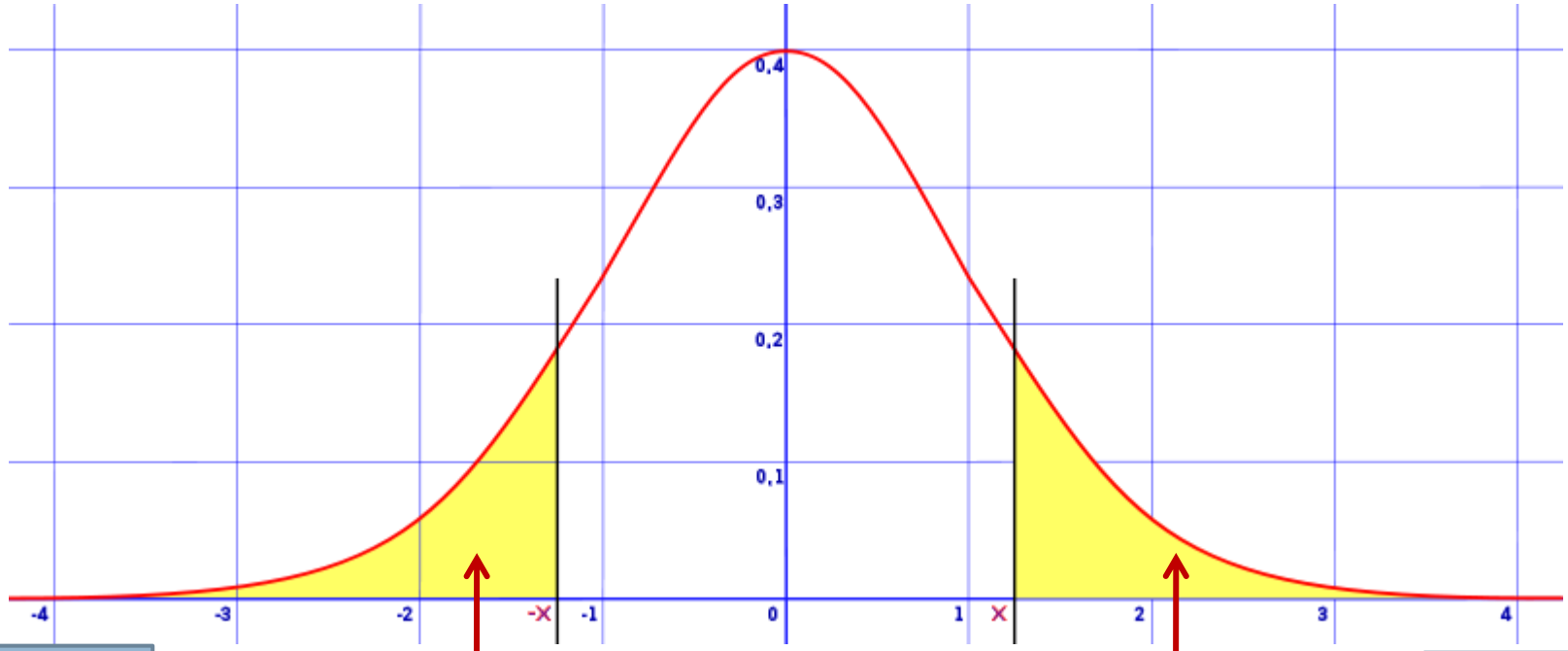
Research question	H_0	H_a
Does tooth flossing affect the number of cavities?	Tooth flossing has no effect on the number of cavities.	Tooth flossing has an effect on the number of cavities.
Does the amount of text highlighted in the textbook affect exam scores?	The amount of text highlighted in the textbook has no effect on exam scores.	The amount of text highlighted in the textbook has an effect on exam scores.
Does daily meditation decrease the incidence of depression?	Daily meditation does not decrease the incidence of depression	Daily meditation decreases the incidence of depression.

Null Hypothesis and Alternate Hypothesis cont…

- How to write null and alternate hypothesis - The only thing to know are the dependent (DV) variables and independent variables (IV). To write null hypothesis, and alternative hypothesis, fill in the following sentences with variables i.e., does independent variable affect dependent variable?
  - Null hypothesis (H_0): **IV** does not affect **DV**.
  - Alternative Hypothesis (H_a): **IV** affects **DV.**
- Characteristics of a Hypothesis
  - It has to be clear and accurate in order to look reliable.
  - It has to be specific.
  - There should be scope for further investigation and experiments.
  - It should be explained in simple language while retaining its significance.
  - IVs and DVs must be included with the relationship between them.

Tails of distributions

The tails of a distribution are the appendages on the side of a distribution. Although it can apply to a set of data, it makes more sense if that data is graphed, because the tails become easily visible.

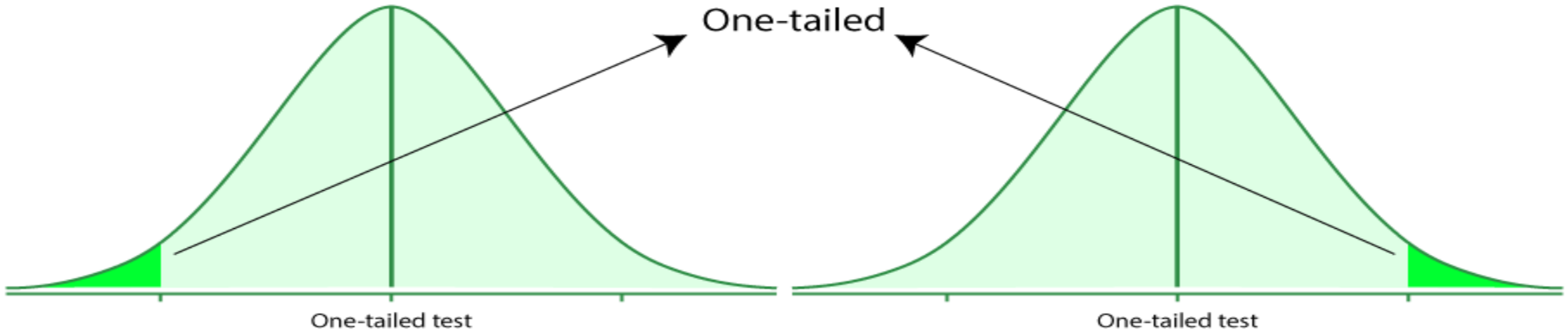


Hypothesis Testing cont…

- The purpose of statistical inference is to draw conclusions about a population on the basis of data obtained from a sample of that population.
- Hypothesis testing is the process used to evaluate the strength of evidence from the sample and provides a framework for making determinations related to the population, i.e, it provides a method for understanding how reliably one can extrapolate observed findings in a sample under study to the larger population from which the sample was drawn.
- The investigator formulates a specific hypothesis, evaluates data from the sample, and uses these data to decide whether they support the specific hypothesis.
- The first step in testing hypotheses is the transformation of the research question into a null hypothesis, and an alternative hypothesis. Subsequently, the hypothesis testing.
- In hypothesis testing, a one-tailed test and a two-tailed test are alternative ways of computing the statistical significance of a parameter inferred from a data set, in terms of a test statistic.

One-Tailed Hypothesis Testing

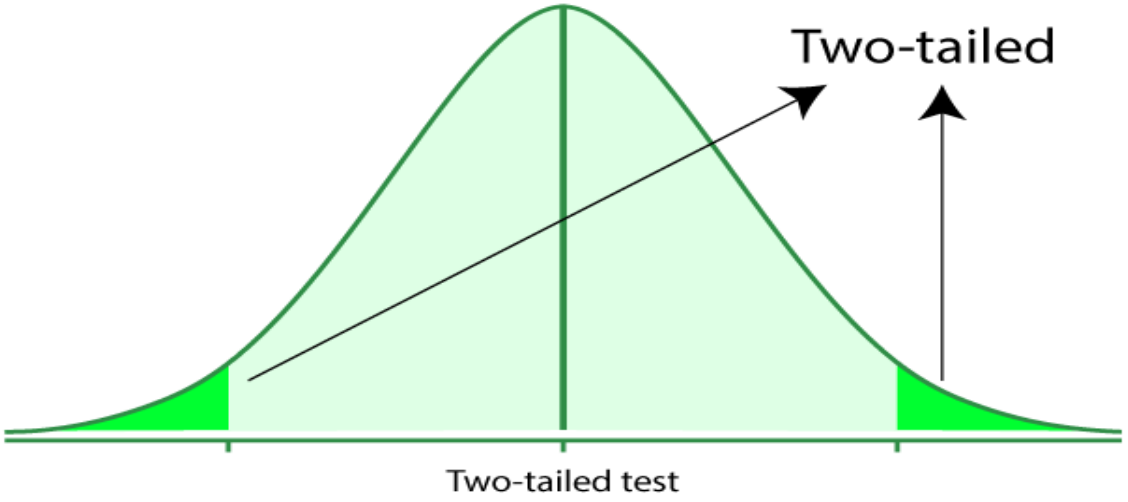
- A one-tailed test is based on a unidirectional hypothesis where the area of rejection is on only one side of the sampling distribution.
- It determines whether a particular population parameter is larger or smaller than the predefined parameter. It uses one single critical value to test the data.



- **Example:** Effect of participants of students in coding competition on their fear level and H_0: There is no important effect of students in coding competition on their fear level. The main intention is to check the **decreased** fear level when students participate in a coding competition.

Two-Tailed Hypothesis Testing

- A two-tailed test is also called a non-directional hypothesis. For checking whether the sample is greater or less than a range of values, the two-tailed is used. It is used for null hypothesis testing.



- **Example:** Effect of new bill pass on the loan of farmers and H_0: There is no significant effect of the new bill passed on loans of farmers. The main intention is to check the new bill passes can affect **in both ways either increase or decrease** the loan of farmers.

Types of Error

- Regardless of whether the investigator decides to accept or reject the null hypothesis, it might be the wrong decision.
- The investigator might incorrectly reject the null hypothesis when it is true, and might incorrectly accept the null hypothesis when it is false.
- In the tradition of hypothesis testing, these two types of errors have acquired the names i.e., **type I** and **type II** errors.
- In general, **commit a type I error** occurs when one incorrectly reject a null hypothesis that is true. On the other hand, **type II error** occurs when you one incorrectly accept a null hypothesis that is false.

		Truth	
		H_0 is true	H_a is true
Decision	Reject H_0	Type I error	No error
	Do not reject H_0	No error	Type II error

Rejection Region

- The question, then, is how strong the evidence in favor of the alternative hypothesis must be to reject the null hypothesis.
- This is done by means of a **p-value**. The p-value is the probability of seeing a random sample at least as extreme as the observed sample, given that the null hypothesis is true. The smaller the p-value, the more evidence there is in favor of the alternative hypothesis.
- The p-values are expressed as decimals and can be converted into percentage. For example, a p-value of 0.0237 is 2.37%, which means there's a 2.37% chance of the results being random or having happened by chance.
- In the hypothesis test, if the value is:
  - A small p value (<=0.05), reject the null hypothesis.
  - A large p value (>0.05), do not reject the null hypothesis
- The p-values are usually calculated using p-value tables, or calculated automatically using statistical software like R, SPSS, Python etc.
- **Note:** *Other way to decide the rejection region is with z-score and it is applicable when the sample size is less than or equal to 30.*

Hypothesis Testing Example

- An investor says that the performance of their investment portfolio is equivalent to that of the Standard & Poor's (S&P) 500 Index. The person performs a two-tailed test to determine this.
- The null hypothesis here says that the portfolio's returns are equivalent to the returns of S&P 500, while the alternative hypothesis says that the returns of the portfolio and the returns of the S&P 500 are not equivalent.
- The p-value hypothesis test gives a measure of how much evidence is present to reject the null hypothesis. The smaller the p value, the higher the evidence against null hypothesis.
- Therefore, if the investor gets a p value of .001, it indicates strong evidence against null hypothesis. So he confidently deduces that the portfolio's returns and the S&P 500's returns are not equivalent.

Hypothesis Testing Numerical

Problem Statement: In the population, the average IQ is 100 with a standard deviation of 15. A team of scientists want to test a new medication to see if it has either a positive or negative effect on intelligence, or not effect at all. A sample of 30 participants who have taken the medication has a mean of 140. Did the medication affect intelligence?

Solution:

Step 1: Set up the null and alternate hypothesis

H_0: medication does not affect intelligence.

H_a: medication affects intelligence.

Step 2: Determine the type of test to use

Since the sample size is 30, the z-test is used.

Step 3: Calculate the tested statistic z using the formula

$$z = \frac{\bar{x}_n - \mu_0}{\sigma} \sqrt{n}$$

Where $\bar{x}_n$ is the mean of the population, μ_0 is the null hypothesis (i.e., the mean) to be tested, σ is the standard deviation, n is the sample size.

Hypothesis Testing Numerical cont...

Using the data given in the equation we would have the following:

$\mu_0 = 100, \sigma = 15, n = 30, \bar{x}_n = 140$

Plugging the values into the formula:

$$z = \frac{140 - 100}{15}\sqrt{30} = 14.606$$

Step 4: Look up the values of z (called the critical value) from statistical table (The table is predefined and should be referred)

From the table, the confidence level value is 1.96 with the confidence interval of 0.95.

Step 5: Draw a conclusion

In this case the tested statistic value of z calculated is more than the critical value obtained from statistical tables (i.e., 14.606 > 1.96). Therefore the null hypothesis is rejected in the favor of the alternative hypothesis.

This means that the medication administered affect intelligence.

Chi-square test for independence

- A chi-square test of independence is to test whether two categorical variables are related to each other or not.
- **Example 1:** we have a list of movie genres; this is the first variable. The second variable is whether or not the patrons of those genres bought snacks at the theater. The idea (or null hypothesis) is that the type of movie and whether or not people bought snacks are unrelated. The owner of the movie theater wants to estimate how many snacks to buy. If movie type and snack purchases are unrelated, estimating will be simpler than if the movie types impact snack sales.
- **Example 2:** a veterinary clinic has a list of dog breeds they see as patients. The second variable is whether owners feed dry food, canned food or a mixture. The idea (or null hypothesis) is that the dog breed and types of food are unrelated. If this is true, then the clinic can order food based only on the total number of dogs, without consideration for the breeds.

Chi-square Test for Independence Example

- Let's take a closer look at the movie snacks example. Suppose we collect data for 600 people at our theater. For each person, we know the type of movie they saw and whether or not they bought snacks.
- For the valid Chi-square test, the following conditions to be satisfied:
  1. Data values that are a simple random sample from the population of interest.
  2. Two categorical or nominal variables.
  3. For each combination of the levels of the two variables, we need at least five expected values. When we have fewer than five for any one combination, the test results are not reliable. To confirm this, we need to know the total counts for each type of movie and the total counts for whether snacks were bought or not. For now, we assume we meet this requirement and will check it later.

Chi-square Test for Independence Example cont…

- The data summarized in a contingency table is as follows:

Type of movie	Snacks	No snacks
Action	50	75
Comedy	125	175
Family	90	30
Horror	45	10

- Before we go any further, let's check the assumption of five expected values in each category. The data has more than five counts in each combination of Movie Type and Snacks.
- To find expected counts for each Movie-Snack combination, we first need the row and column totals, which are shown below:

Type of movie	Snacks	No snacks	Row Totals
Action	50	75	50 + 75 = 125
Comedy	125	175	125 + 175 = 300
Family	90	30	90 + 30 = 120
Horror	45	10	45 + 10 = 55
Column Totals	50+125+90+45 = 310	75+175+30+10 = 290	Grand Total = 600

Chi-square Test for Independence Example cont…

- The expected counts for each Movie-Snack combination are based on the row and column totals. We multiply the row total by the column total and then divide by the grand total. This gives us the expected count for each cell in the table.
- For example, for the Action-Snacks cell: (125 * 310) / 600 = 65. If there is not a relationship between movie type and snack purchasing we would expect 65 people to have watched an action film with snacks.
- For the Action-No Snacks cell: (125 * 290) / 600 = 60. Similarly, it can be counted for others…
- The expected count appears in bold beneath the actual count.

Type of movie	Snacks	No snacks	Row Totals
Action	50 125*310/600 = **65**	75 125*290/600 = **60**	125
Comedy	125 300*310/600 = **155**	175 300*290/600 = **145**	300
Family	90 120*310/600 = **62**	30 120*290/600 = **58**	120
Horror	45 55*310/600 = **28**	10 55*290/600 = **27**	55
Column Totals	310	290	Grand Total = 600

Chi-square Test for Independence Example cont…

- All of the expected counts for our data are larger than five, so we meet the requirement for applying the independence test.
- If we look at each of the cells, we can see that some expected counts are close to the actual counts but most are not.
- If there is no relationship between the movie type and snack purchases, the actual and expected counts will be similar. If there is a relationship, the actual and expected counts will be different.

Performing the Chi-square Test

- The basic idea in calculating the test statistic is to compare actual and expected values, given the row and column totals that we have in the data.
- First, we calculate the difference from actual and expected for each Movie-Snacks combination.
- Next, we square that difference. Squaring gives the same importance to combinations with fewer actual values than expected and combinations with more actual values than expected.
- Next, we divide by the expected value for the combination. We add up these values for each Movie-Snacks combination. This gives the test statistic.

Chi-square Test for Independence Example cont…

Type of movie	Snacks	No snacks	Row Totals
Action	Actual: 50 Expected: **65** Difference: 50 – 65 = -15 Squared Difference = 225 Divide by Expected: 225/65 = 3.46	Actual: 75 Expected: **60** Difference: 75 – 60 = 15 Squared Difference = 225 Divide by Expected: 225/60 = 3.75	125
Comedy	Actual: 125 Expected:**155** Difference: 125– 155 = -30 Squared Difference = 900 Divide by Expected: 900/155 = 5.81	Actual: 175 Expected: **145** Difference: 175– 145 = 30 Squared Difference = 900 Divide by Expected: 900/145 = 6.21	300
Family	Actual: 90 Expected:**62** Difference: 90 – 62 = 28 Squared Difference = 784 Divide by Expected: 784/62 = 12.65	Actual: 30 Expected: **58** Difference: 30 – 58 = -28 Squared Difference = 784 Divide by Expected: 784/58 = 13.52	120
Horror	Actual: 45 Expected:**28** Difference: 45 – 28 = -16 Squared Difference = 256 Divide by Expected: 256/28 = 9.14	Actual: 10 Expected: **27** Difference: 10 – 27 = -17 Squared Difference = 289 Divide by Expected: 289/27 = 10.70	55
Column Totals	310	290	Grand Total = 600

Chi-square Test for Independence Example cont...

- Lastly, to get our test statistic, we add the numbers in the final row for each cell: 3.46 + 3.75 + 5.81 + 6.21 + 12.65 + 13.52 + 9.14 + 10.70 = 65.24
- Now, we need to find the critical value from the Chi-square distribution based on degrees of freedom and significance level. This is the value to expect if the two variables are independent.
- The degrees of freedom depend on how many rows and how many columns we have. The degrees of freedom (df) are calculated as $df=(r-1)\times(c-1)$ where r is the number of rows, and c is the number of columns in the contingency table. From the example, r is 4 and c is 2. Hence, $df = (4-1)\times(2-1)=3\times1=3$.
- The Chi-square value with $\alpha = 0.05$ (it is given and represents the probability of rejecting the null hypothesis when it is true) and three degrees of freedom is 7.815. **Note:** This value of 7.815 to be infer from the Chi-square distribution table. Refer Appendix for further details
- We compare the value of our test statistic (65.24) to the Chi-square value. Since 65.24 > 7.815, we reject the idea that movie type and snack purchases are independent.

Chi-square Test for Independence Example cont...

- Therefore, we conclude that there is some relationship between movie type and snack purchases.
- However, the owner of the movie theater cannot estimate how many snacks to buy regardless of the type of movies being shown. Instead, the owner must think about the type of movies being shown when estimating snack purchases.
- It's important to note that we cannot conclude that the type of movie causes a snack purchase. The independence test tells us only whether there is a relationship or not; it does not tell that one variable causes the other.

Statistical details

- The null hypothesis is that the type of movie and snack purchases are independent. It is written as: H_0:Movie Type and Snack purchases are independent
- The alternative hypothesis is the opposite i.e., H_a: Movie Type and Snack purchases are not independent.


THANK
YOU!

Appendix

Z-values for confidence interval

Confidence Level	Z value
0.70	1.04
0.75	1.15
0.80	1.28
0.85	1.44
0.90	1.64
0.92	1.75
0.95	1.96
0.96	2.05
0.98	2.33
0.99	2.58
0.50	0.674

Appendix

Chi-square Distribution Table

Degrees of freedom (*df*)	Significance level (α)							
	.99	.975	.95	.9	.1	.05	.025	.01
1	-------	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578