BITS Pilani
Pilani | Dubai | Goa | Hyderabad

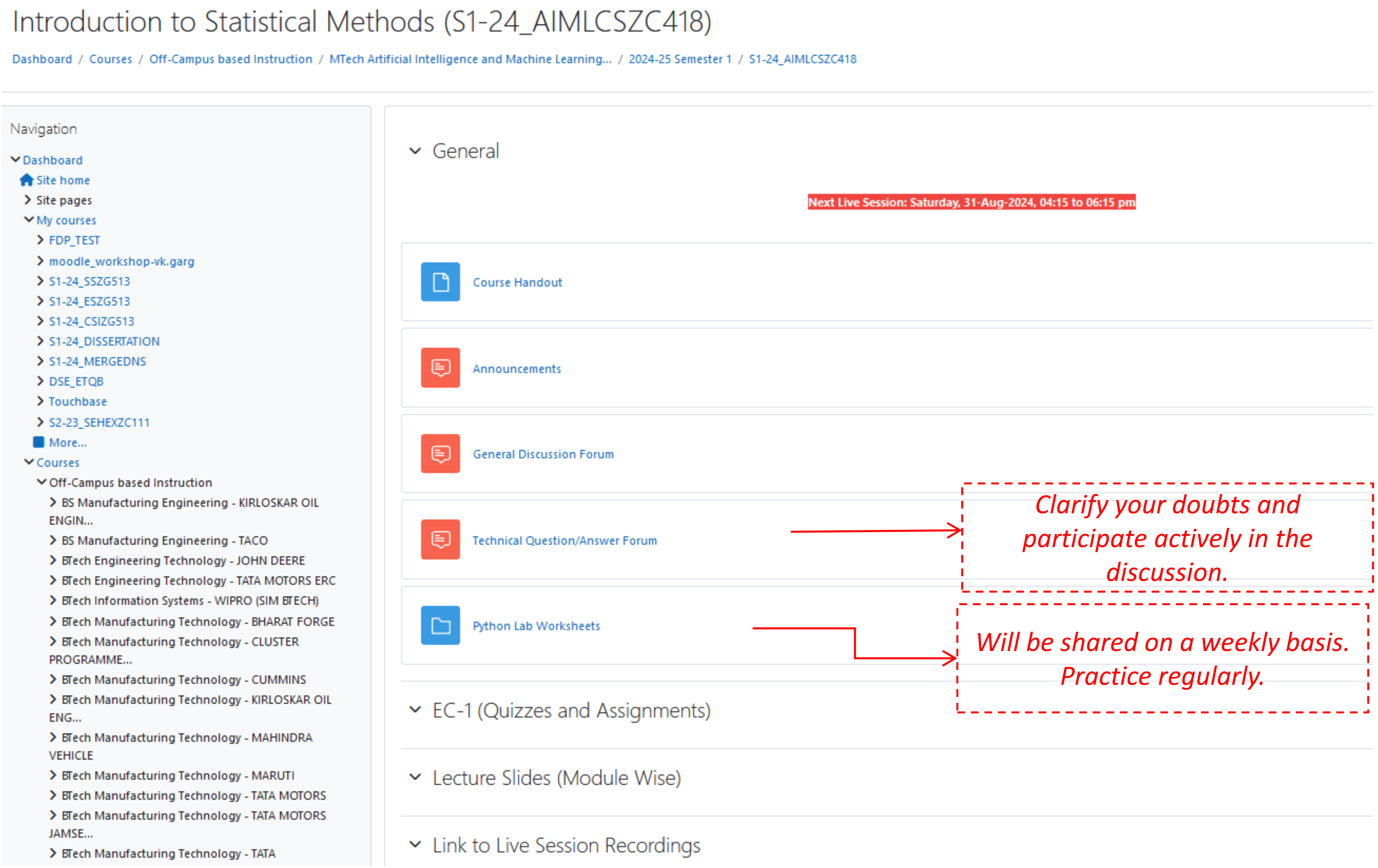
Introduction to Statistical Methods

Introduction, Revision-1.0

Prof Vineet Garg

BITS Pilani Work Integrated Learning Programmes (WILP)
Bangalore Professional Development Center

Course Page on eLearn/Taxila



Introduction to Statistical Methods (S1-24_AIMLCSZC418)

Dashboard / Courses / Off-Campus based Instruction / MTech Artificial Intelligence and Machine Learning... / 2024-25 Semester 1 / S1-24_AIMLCSZC418

Navigation

- Dashboard
  - Site home
  - Site pages
  - My courses
    - FDP_TEST
    - moodle_workshop-vk.garg
    - S1-24_SSZG513
    - S1-24_ESZG513
    - S1-24_CSIZG513
    - S1-24_DISSERTATION
    - S1-24_MERGEDNS
    - DSE_ETQB
    - Touchbase
    - S2-23_SEHEXZC111
    - More...
  - Courses
    - Off-Campus based Instruction
      - BS Manufacturing Engineering - KIRLOSKAR OIL ENGIN...
      - BS Manufacturing Engineering - TACO
      - BTech Engineering Technology - JOHN DEERE
      - BTech Engineering Technology - TATA MOTORS ERC
      - BTech Information Systems - WIPRO (SIM BTECH)
      - BTech Manufacturing Technology - BHARAT FORGE
      - BTech Manufacturing Technology - CLUSTER PROGRAMME...
      - BTech Manufacturing Technology - CUMMINS
      - BTech Manufacturing Technology - KIRLOSKAR OIL ENG...
      - BTech Manufacturing Technology - MAHINDRA VEHICLE
      - BTech Manufacturing Technology - MARUTI
      - BTech Manufacturing Technology - TATA MOTORS
      - BTech Manufacturing Technology - TATA MOTORS JAMSE...
      - BTech Manufacturing Technology - TATA

General

Next Live Session: Saturday, 31-Aug-2024, 04:15 to 06:15 pm

Course Handout

Announcements

General Discussion Forum

Technical Question/Answer Forum

Python Lab Worksheets

EC-1 (Quizzes and Assignments)

Lecture Slides (Module Wise)

Link to Live Session Recordings

Clarify your doubts and participate actively in the discussion.

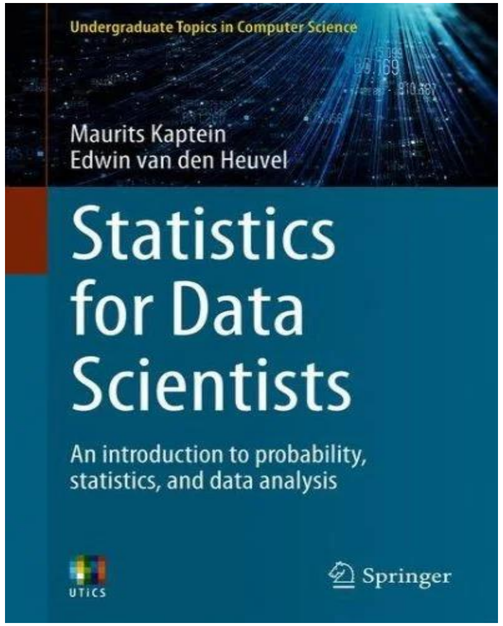
Will be shared on a weekly basis. Practice regularly.

EC-1 Details

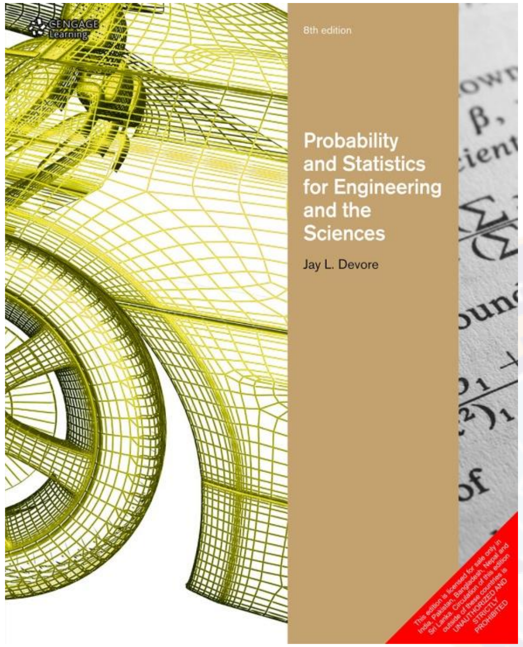
Introduction to Statistical Methods			
EC		Weightage	Dates
EC-1	Theory Quiz-1	7.5%	29-Sep to 06-Oct 2024
	Python Quiz-1	7.5%	20-Oct to 27-Oct 2024
	Theory Quiz-2	7.5%	05-Jan to 12 Jan 2025
	Python Quiz-2	7.5%	12-Jan to 19 Jan 2025

- All EC-1 components will be individual and conducted through eLearn/Taxila.
- The EC-1 components will remain open for the specified 8 days. But a time limit will be applicable when you start attempting it.
- Only 1 attempt will be permitted.
- There will no MAKEUP for EC-1 components.

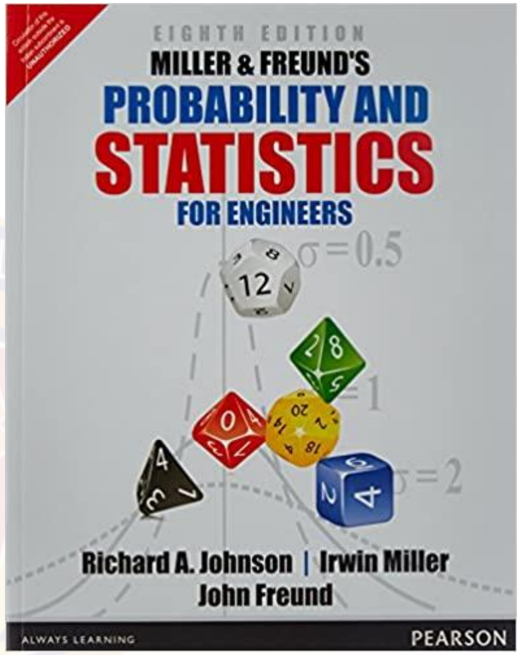
Useful Text and Reference Books


Undergraduate Topics in Computer Science
Maurits Kaptein
Edwin van den Heuvel
Statistics for Data Scientists
An introduction to probability, statistics, and data analysis
UTICS
Springer

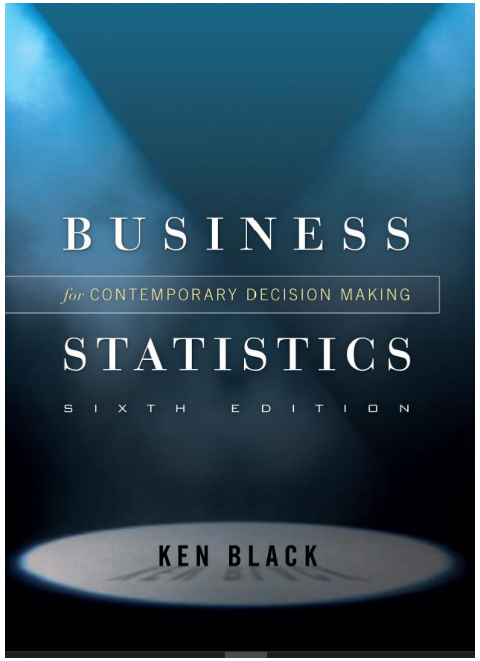
Statistics for Data Scientists: An introduction to probability ,statistics and data analysis by Maurits Kaptein et al, Springer 2022



Probability and Statistics for Engineering and Sciences,8th Edition by Jay L Devore, CENGAGE Learning



Miller and Freund’s Probability and statistics for Engineers, 8th Edition, Pearson

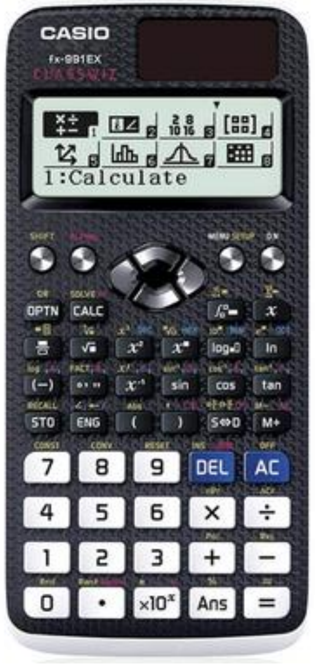


A good and detailed book: Business Statistics for Contemporary Decision Making, 6th Edition by Ken Black, Wiley

Important Points

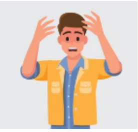
- During the session please keep your notebook, pen and scientific calculator ready. It will help you to solve problems during the sessions. ✓
- Please do not share the screenshots or copied out of MS-Excel, Python, R, Calculators or any other tool or programming language in the eLearn/Taxila technical question answer forum, mid-semester and comprehensive exams. Such responses will not be evaluated. ✗
- A decent scientific calculator (e.g. Casio-991EX) is recommended. We will use different statistics and calculus functions from it. ✓
- Solve slide exercises and leverage Technical Question / Answer Forum from the eLearn/Taxila. It will help you in Mid-Semester (EC-2) and Comprehensive exams (EC-3). ✓
- Please install Anaconda on your laptop (if not using already). We will use Jupyter Notebook in this course. Alternatively, you can use Google Colab. ✓

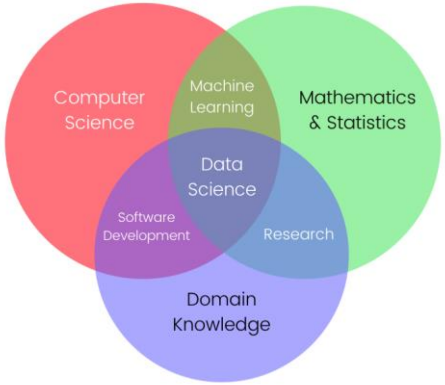
ANACONDA jupyter Google colab

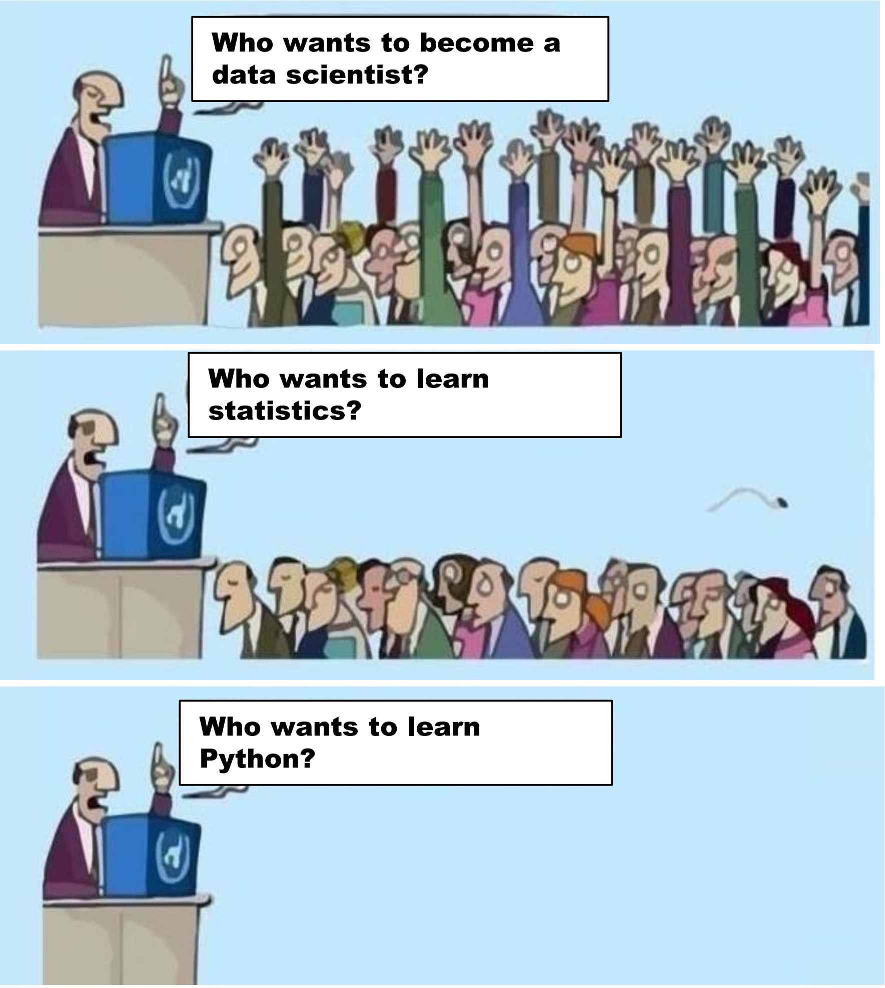


Why Statistics?

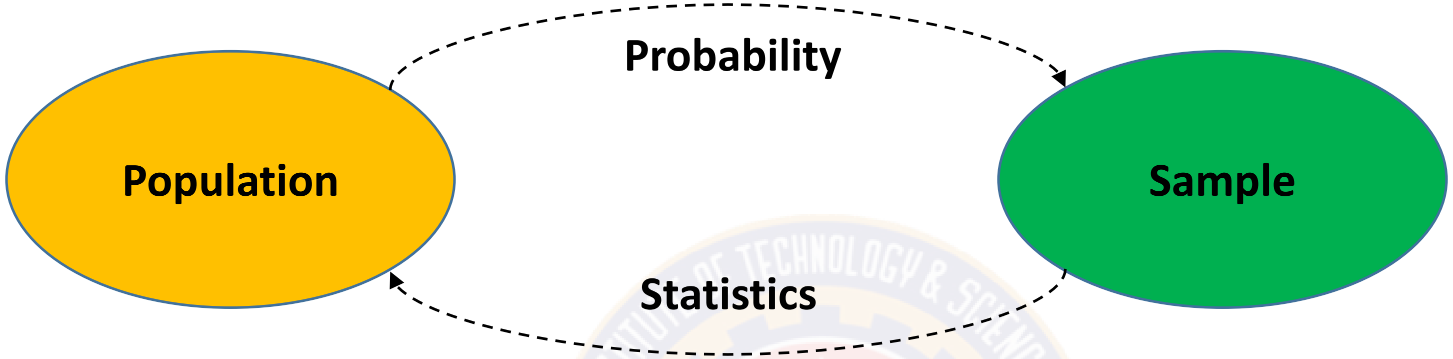
- The discipline of statistics teaches us how to make intelligent judgements and informed decisions in the presence of **uncertainty** and **variation**.
- A highly underestimated area by data science aspirants !
- Lack of programming skills makes the situation even worse !
- An important area of study for the various dimensions of Data Science.
- We do not need to study statistics if there is no uncertainty and variation:
  - Each dose of vaccination provides the same amount of immunity to all.
  - A battery model-xyz has same life time irrespective of the vehicle it is used in.
  - Life of the LED screen is same for all the TV sets a company produces.
- Have you ever come across the statements like the following:
  - 50% of all drivers wear seat belts. How likely is that a sample of 100 drivers will include at least 70 who wear seat belts?
  - A sample of 250 drivers revealed that 180 regularly wear seat belts. Does this provide substantial evidence that more than 50% of all drivers wear seat belts?
- What is the difference between the above two examples of seat-belts?
  - The first statement is to deal with probability.
  - The second is to deal with an area of statistics where **sample** is used to get the estimate of **population**.
- Will we study probability also in detail?
  - Only to the extent that has direct bearing on statistics.






Who wants to become a data scientist?
Who wants to learn statistics?
Who wants to learn Python?

Statistics and Probability



- In a probability problem, properties of the population under study are assumed to be known and then questions regarding a sample taken from the population are answered - **Deductive Reasoning**.
    - **Example:** 50% of all drivers wear seat belts. How likely is that a sample of 100 drivers will include at least 70 who wear seat belts?
- In a statistics problem, characteristics of a sample are known and this information enables to draw the conclusions about the population – **Inductive Reasoning**.
    - **Example:** A sample of 250 drivers revealed that 180 regularly wear seat belts. Does this provide substantial evidence that more than 50% of all drivers wear seat belts?

What is population and what is sample?

Descriptive and Inferential Statistics

- The study of statistics can be divided into two branches:

1. **Descriptive Statistics**: summarization and presentation of the data gathered on a group to describe or reach conclusions about that same group.
    - **Example:** An IT start-up company employs five software engineers at a monthly salary of ₹60,000/- each, one office assistant at ₹4,000/- per month and the founder herself decided to draw ₹300,000/- of salary per month. A tech journalist published an article about this startup and reported the mean salary as ₹100,667/- for its employees. You may or may not agree with the journalist. But she is trying to describe the salary in an IT startup.

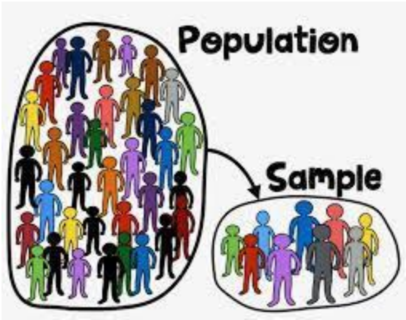
2. **Inferential Statistics:** gather data from a sample and use the statistics generated to reach conclusions about the population from which the sample was taken.
    - **Example:** A soft drink company creates an advertisement. A vending machine that talks to the buyers. Market researchers want to measure the impact of the new advertisement on various age groups. They can sample few people from each age group and measure the effectiveness of the advertisement. E.g. who liked it? Who ignored it? Etc.
    - Why did the soft drink company create the samples of people from different age groups?

- A descriptive measure of a **sample** is called a ***statistic***. Statistics are usually denoted by English/Roman letters. Examples of statistics are sample mean ($\bar{x}$), sample variance (s^2), and sample standard deviation (s).

- A descriptive measure of the **population** is called a ***parameter***. Parameters are usually denoted by Greek letters. Examples of parameters are population mean (μ), population variance (σ^2), and population standard deviation (σ).

- The population could be **real** or **conceptual**. For example all who drive cars (real), all TV that will be produced (in future years - conceptual).

- It is difficult to calculate parameters for the entire population, so inferences about parameters are made using sample statistics. So there is always uncertainty in these inferences. In order to estimate the level of confidence in these uncertainties, different techniques are used that we will also review as part of this course.



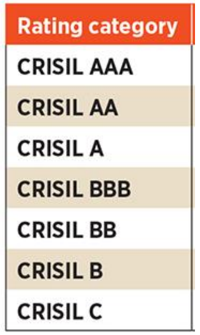
What is Flow of this Course?

- Introduction
- Probability Concepts
- Probability Distributions
  - Random Variables – Discrete and Continuous
  - Discrete Distributions – Binomial, Poisson and Hypergeometric
  - Continuous Distributions – Uniform, Normal and Gamma
- Inferential Statistics / Hypothesis Testing
  - Sampling
  - Central Limit Theorem
  - Estimation and Hypothesis Testing
- Regression and Forecasting
  - Linear, Multiple and Polynomial Regression
  - Time Series

Data Types

- Million and billion storage units of numerical data is gathered everyday. All such data should not be analyzed in the same statistical way because they represent the different entities.
- Example: Two jerseys with numbers 40 and 80. Two shipments weighing 40 and 80 kg etc. One box weighs double the other but what is the meaning of saying a similar statement about jerseys?
- Four common levels of data measurement are as following:

1. **Nominal**: used only to classify or categorize.
   - Examples: Jersey numbers, Employee ID, Gender, PIN Codes etc.
2. **Ordinal**: ordinal-level measurement can be used to rank or order objects. But the difference between the ranks may not be established. Examples:
   - Supervisor rating for the staff are *Excellent*, *Good*, *Fair* and *Poor*. But it can't be said that the difference between Excellent and Good is same as the difference between Fair and Poor.
   - Ratings of the different mutual funds.



Rating category
CRISIL AAA
CRISIL AA
CRISIL A
CRISIL BBB
CRISIL BB
CRISIL B
CRISIL C

3. **Interval**: distances between consecutive numbers have meaning and the data is always numerical. The zero point is a matter of convention or convenience and not a natural or fixed zero point. Examples:
   - Temperatures of 20^o, 21^o and 22^o Fahrenheit. But zero degrees Fahrenheit is not the lowest possible temperature (same is true for degree Celsius as well)
   - Calendar dates.

4. **Ratio**: ratio data have an absolute zero, and the ratio of two numbers is meaningful. It is also used for numerical data. A true zero point exists. Examples:
   - The ratio of the weights of the two boxes is 40:80.
   - Temperature measurement in ^oK.

We will deal primarily with Interval and Ratio numeric data.

Exercise

The following questionnaire is presented to the patients of a hospital on discharge. What different data measurement are involved in it?

1. How long ago were you released from the hospital?
2. Which type of unit were you in for most of your stay?

    o Intensive care o Pediatric unit o Hospital Room

3. In choosing a hospital, how important was the hospital’s location?

    o Very Important o Somewhat Important o Not at all Important

4. How serious was your condition when you were first admitted to the hospital?

    o Critical o Serious o Moderate

5. Rate the skill of your doctor:

    o Excellent o Good o Fair

6. The date of your discharge?
7. The total amount you paid?

Descriptive Statistics: Mean

- It is necessary to have an overall picture of the data. If few integers are given as: 25, 32, 24, 28, 26 and 27, how will you **characterize** and **describe** it?

- Three basic areas of statistical description:
  1. **Measures of Central Tendency**: Measures of central tendency yield information about the center, or middle part of a group of numbers. Mean, mode, median and mid-range are the measures for central tendency.
  2. **Measures of Variability or Dispersion**: describes the spread or the dispersion of a set of data. Variance, standard deviation, range are few measures of variability.
  3. **Visualization**: it includes charts and graphs and attempt to present the data pictorially.

- The most common and efficient numeric measure of the center of a data set is the mean (or arithmetic mean or average). Let N values are x_1, x_2, x_3......x_N. The mean of this data set is given by:

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

- Sometimes, each value x_i may be associated with a weight w_i. This weight reflects the significance, importance, or the occurrence frequency. In that case mean is:

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

The sample mean is represented by (x̄).
The population mean is represented by the Greek letter (μ).

Exercise

1. For a group of employees the salary data in thousands of rupees is 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70 and 110.

   i. Calculate the mean salary.

   ii. Calculate the mean salary using weights

   (Answer: Rs. 58,000 for both sub-parts)

Weighted Mean Approach (values in '000):

$$\bar{x} = \frac{30+36+47+50+(2*52)+56+60+63+(2*70)+110}{1+1+1+1+2+1+1+1+2+1}$$

$$= \frac{696}{12} = 58$$

2. A learning program comprises of four subjects of 5, 4, 4, 3 units (credits) respectively are offered in the two semesters. Subjects 1 and 2 in the first semester and subjects 3 and 4 in the second semester.

   i. In the first semester, a participant received A and A- grades in the first two subjects. Calculate her CGPA if A and A- are equivalent to 10 and 9 respectively in numeric terms.

   (Answer = 9.56)

   i. In the second semester, the same participant received B and B- in the remaining two subjects. Calculate her CGPA after the end of the second semester if B and B- are equivalent to 8 and 7 respectively in numeric terms. Note that CGPA is calculated cumulatively for all the completed courses.

   (Answer = 8.69)

Median, Mode and Midrange

Median:

- A few numbers of extremes can provide the corrupt central tendency.
- These extremes are called outliers. E.g. director's salary may be higher than that of software engineers'. Admin assistant's salary may be lower than that of software engineers'. Mean can not represent the true center.
- Chopping off some small % of low and high data extremes provides *trimmed mean*. But, too large chopping may lose important data. So chopping may not be a preferred method.
- For such skewed data, a better measure is the center of the data or ***median***. It is the middle value in the set of ordered data values after arranging them in non-decreasing order (or even in non-increasing order).
- So median is the middle value if odd number of values are present. Otherwise, average of the middle two values.
- Median is also called the 50th percentile (denoted by Q2). Similarly 25th percentile (Q1) and 75th percentile (Q3) are also used.

Mode:

- Value that occurs most frequently is called the ***mode***.
- It is possible that the highest frequency of a data element is same for several different data elements. So the data sets with one, two and three modes are called uni-modal, bi-modal and tri-modal data sets respectively.

Midrange:

- ***Midrange*** is the average of largest and smallest values in the data set.

Exercise

For a group of employees the salary data in thousands of rupees is 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70 and 110.

i. Calculate the median salary.

(Answer: Rs. 54,000)

ii. Calculate the median salary ignoring the last salary value of Rs. 110,000.

(Answer: Rs. 52,000)

iii. Calculate the mode.

(Answer: Rs. 52,000 and Rs. 70,000)

iv. What modal dataset is this?

(Answer: Bi-modal)

v. Calculate the midrange.

(Answer: Rs. 70,000)

Quartiles & Inter Quartile Range (IQR)

Five Number Summary!

Even Count of Data Elements:

For a group of employees the salary data in thousands of rupees is 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70 and 110. Divide the data into quartiles and find out the IQR.

Step-1: Order the values in ascending order.

Step-2: Identify min and max values.

Step-3: Identify median (Q2).

Step-4: Identify Q1 and Q3. This is similar as identification of median of the leftover elements on the left and the right side of identified Q2 of step-3 (left Q2 element for the left part, right Q2 element for the right part need to be included in these parts).

Step-5: identify IQR as Q3-Q1. It will be required for outliers.

Odd Count of Data Elements:

For a group of employees the salary data in thousands of rupees is 30, 36, 47, 50, 52, 52, 56, 60, 63, 70 and 70. Divide the data into quartiles and find out the IQR.

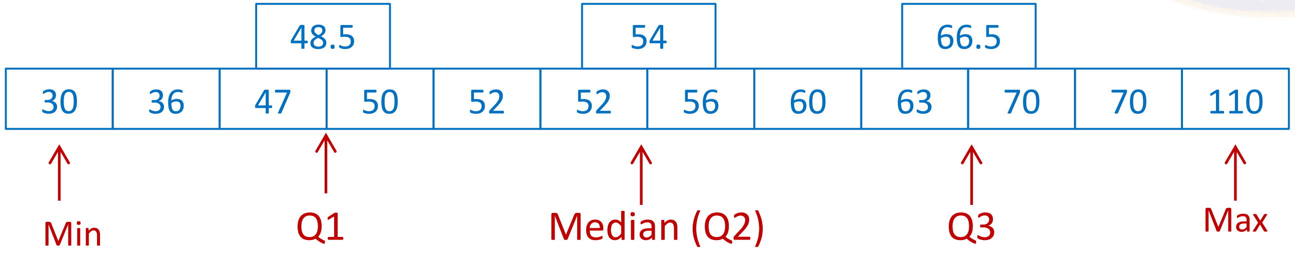
Step-1: Order the values in ascending order.

Step-2: Identify min and max values.

Step-3: Identify median (Q2).

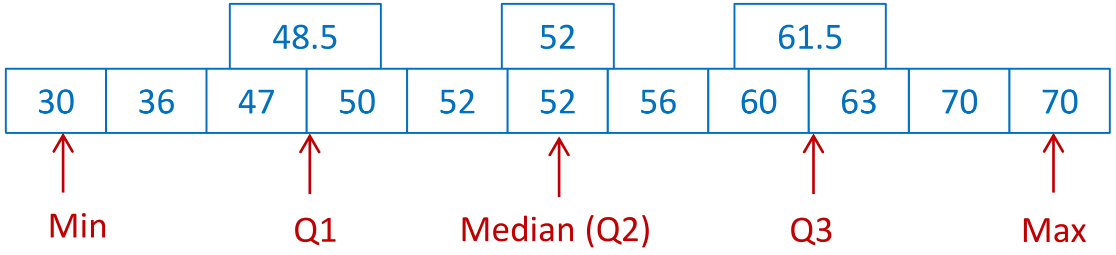
Step-4: Identify Q1 and Q3. This is similar as identification of median of the leftover elements on the left and the right side of identified Q2 of step-3 (including Q2 for both).

Step-5: identify IQR as Q3-Q1. It will be required for outliers.



30	36	47	50	52	52	56	60	63	70	70	110

IQR = 66.5 – 48.5 = 18

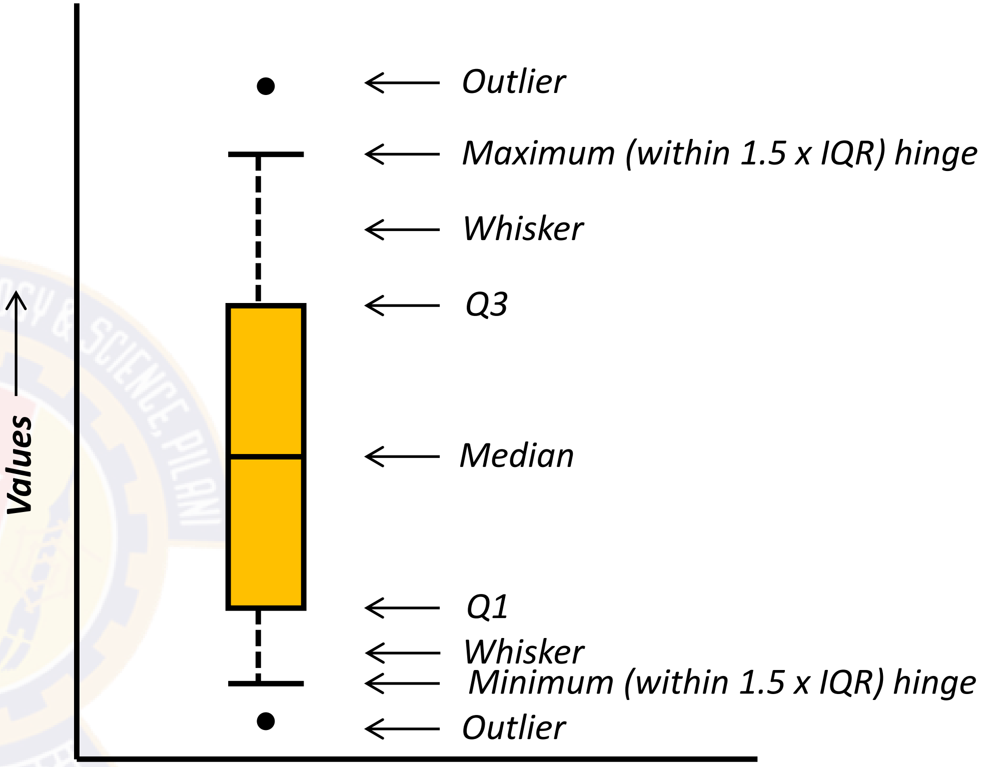


30	36	47	50	52	52	56	60	63	70	70

IQR = 61.5 – 48.5 = 13

Boxplot

- Five-Number Summary consists of - **Minimum, Q1, Median (Q2), Q3, Maximum.**
- BoxPlot is one of the most popular way to summarize the data.
- Data elements which are < **Q1 - (1.5 x IQR**) and which are > **Q3 + (1.5 x IQR)** can be treated as **outliers** or **anomalies**.
- Five-Number Summary is shown using box-plots.
- Whiskers (dashed lines) are drawn up to maximum/minimum data elements excluding outliers. Outliers are shown as dots separately.
- Maximum/Minimum lines are called **hinges** or **fences**. They are to be taken from data elements. They are not (Q3 + 1.5.IQR) or (Q1-1.5.IQR) theoretical values.
- Box plots for multiple data sets can be drawn next to each other for comparison.



Outlier

Maximum (within 1.5 x IQR) hinge

Whisker

Q3

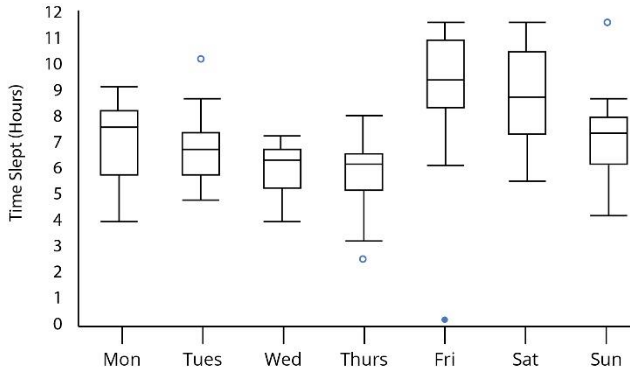
Median

Q1

Whisker

Minimum (within 1.5 x IQR) hinge

Outlier



Exercise

Data analyst of a bank noticed the following credit card transactions of a customer for a year: January to December respectively. All the values are in thousands ₹: 7, 13, 8, 32, 12, 10, 15, 5, 35, 17, 19 and 11. Which two months should he suspect as to have the fraudulent transactions?

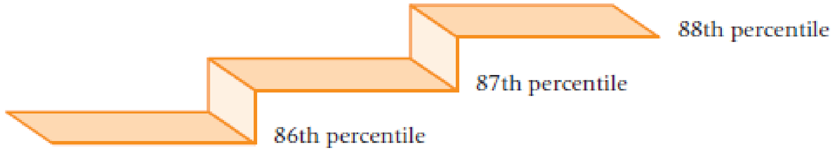
Answer: Q3 = 18 and Q1 = 9. So, IQR = 9. Data points above Q3 with a margin of 1.5xIQR (= 13.5) are 32 and 35. These months are **April** and **September**. There are no data points below Q1 with a margin of 1.5xIQR (= 13.5).

Important Note: In these question, do not just believe your intuitions looking at the data. For example you might say 32 and 35 thousand values are obvious outliers. In data science, inferences need to be numerically justified. In this exercise outliers are identified with a statistics based logic.

👉 Also pay attention to the questions. Values are in thousands and question asks for the months (not the values).

Percentiles

- Percentiles are measures of central tendency that divide a group of data into 100 parts. There are 99 percentiles because it takes 99 dividers to separate a group of data into 100 parts.
- The nth percentile is the value such that at least n percent of the data are below that value and at most (100 - n) percent are above that value.
- Scores of CAT, SAT, GRE, GMAT and many other examinations are reported in percentiles.
- Percentiles are something like “stair-step” as shown in the figure.



Steps to calculate the percentile:

1. Organize the numbers into an ascending-order (non-decreasing) array.
2. Calculate the percentile location (*i*) by:

$$i=\frac{P}{100}(N),\text{where:}$$

$$P\text{:the percentile of interest}$$

$$i\text{ : percentile location}$$

$$N\text{ : size of the dataset}$$

3. Determine the location by either (a) or (b).
   a) If i is a whole number, the Pth percentile is the average of the value at the ith location and the value at the (i+1)st location.
   b) If i is not a whole number, the Pth percentile value is located at the whole number part of (i+1) .

Examples

Example-1: Determine the 30th percentile of the following eight numbers: 14, 12, 19, 23, 5, 13, 28, 17.

Step-1: Organize the data elements in non-decreasing order.

5, 12, 13, 14, 17, 19, 23, 28

Step-2: 30th percentile location (i) = (30/100)*8 = 2.4
Step-3: Since i is not a whole number. The 30th percentile location is the whole part of i+1 that is 2.4+1 = 3.4, that is 3.

The 3rd value is 13. **So 13 is the 30th percentile.**

Example-2: Determine the quartiles (Q1, Q2 and Q3) of the following eight numbers: 106, 109, 114 ,116, 121, 122, 125, 129.

Step-1: Data elements are already in non-decreasing order. **Step-2**: 25th percentile (Q1) location (i) = (25/100)*8 = 2 **Step-3**: Since i is a whole number. The 25th percentile is the average of ith and (i+1)st values, that is the average of 2nd and 3rd values, that is the average of 109 and 114 = **111.5**	**Step-1**: Data elements are already in non-decreasing order. **Step-2**: 50th percentile (Q2) location (i) = (50/100)*8 = 4 **Step-3**: Since i is a whole number. The 50th percentile is the average of ith and (i+1)st values, that is the average of 4th and 5th values, that is the average of 116 and 121 = **118.5**	**Step-1**: Data elements are already in non-decreasing order. **Step-2**: 75th percentile (Q3) location (i) = (75/100)*8 = 6 **Step-3**: Since i is a whole number. The 75th percentile is the average of ith and (i+1)st values, that is the average of 6th and 7th values, that is the average of 122 and 125 = **123.5**

Measures of Variability / Dispersion

Range:

- The Range is the difference between the largest and the smallest value of a data set.

Mean Absolute Deviation (MAD):

- The mean absolute deviation (MAD) is the average of the absolute values of the deviations around the mean for a set of numbers. For a data element x of a data set whose mean is µ, the absolute deviation around mean is |(x- µ)|.

Variance:

- ***Variance*** (σ^2) of N observations (x_1, x_2,....x_N) is calculated as:

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

$$where, \mu \text{ is the mean of the dataset values}(x_i).$$

*The shown formula is for the population. Sample variance and standard deviation are represented by **s²** and **s** respectively and in that case denominator is (N-1) while calculating the variance (Bessel's correction).*

Standard Deviation:

- Square root of the variance is called the ***Standard Deviation*** (σ).
- A low standard deviation means data observation is close to the mean. Otherwise it is spread out over a range of large values. A zero standard deviation means all the observations have the same value.

Exercise

1. Find out the variance and standard deviation for the dataset X = {5, 9, 16, 17, 18}.

2. For a group of employees the salary data in thousands of rupees is 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70 and 110. Calculate the range, variance and standard deviation.

   (Answer: 80, 379.17, 19.47 thousands)

3. There are two datasets X={1, 2, 3, 4} and Y={1, 3, 5, 7}. In which dataset observations are closer to the mean?

   (Answer: standard deviation of X = 1.12 and Y = 2.24 that indicates dataset X is more packed near the mean.)

4. A dataset is given as X = {5, 9, 16, 17, 18}. Find out its Mean Absolute Deviation (MAD)?

Ans:1	X	x-μ	$(x-\mu)^2$
	5	-8	64
	9	-4	16
	16	3	9
	17	4	16
	18	5	25
Sum	**65**		**130**
N	5		
μ	13		
σ^2			**26**
σ			**5.10**

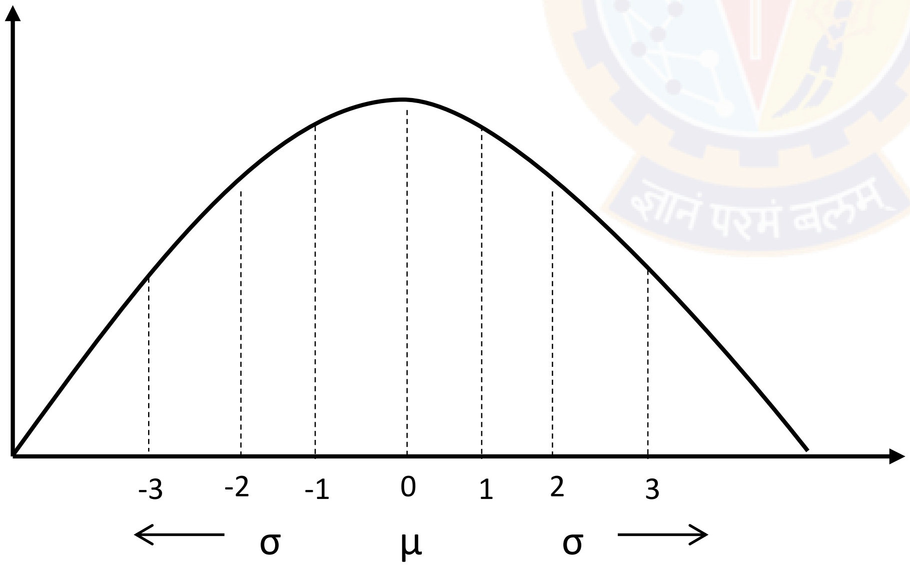
Ans:4	X	|x-μ|
	5	8
	9	4
	16	3
	17	4
	18	5
Sum	**65**	**24**

X = {5, 9, 16, 17, 18}
Mean (μ) = (5+9+16+17+18)/5 = 65/5 = 13
So, MAD(X) = 24/5 = 4.8

Standard Deviation Meaning

For the Normal Distribution (symmetrical) curve where is mean is μ and standard deviation is denoted by σ:

- From (μ–σ) to (μ+σ): contains about 68% of the data
- From (μ–2σ) to (μ+2σ): contains about 95% of the data
- From (μ–3σ) to (μ+3σ): contains about 99.7% of the data



We will revisit normal distribution in the subsequent modules.

Example

A company produces an automobile part that is specified to weigh 1365 grams. Because of imperfections in the manufacturing process not all of the parts produced weigh exactly 1365 grams. The weights of the parts produced are normally distributed with a mean weight of 1365 grams and a standard deviation of 294 grams.

1. What is the range of weights for 95% of the parts weights fall?

Answer: $(\mu - 2\sigma)$ = 1365 – 2*294 = **777** grams to $(\mu + 2\sigma)$ = 1365 + 2*294 = **1953** grams

2. 16% weights are more from what weight value?

Answer: $(\mu + \sigma)$ = 1365 + 294 = **1659** grams

3. 0.15% weights are less from what weight value?

Answer: $(\mu - 3\sigma)$ = 1365 – 3*294 = **483** grams

Note that:

From $(\mu - \sigma)$ to $(\mu + \sigma)$: contains about 68% of the data

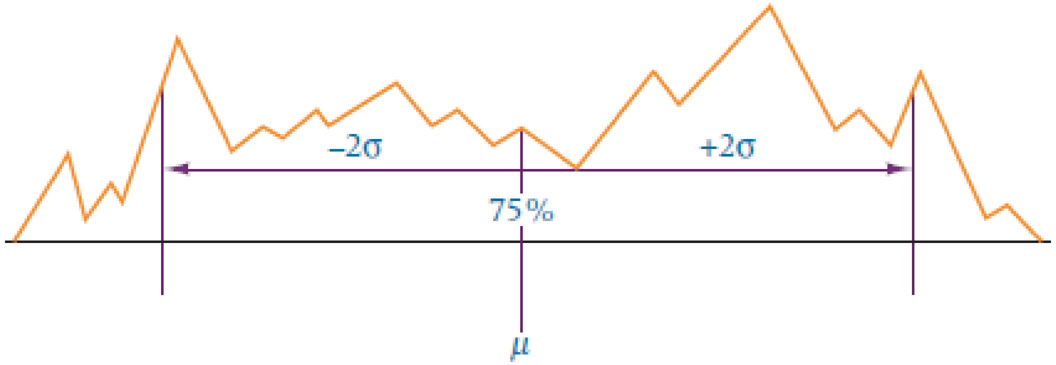
From $(\mu - 2\sigma)$ to $(\mu + 2\sigma)$: contains about 95% of the data

From $(\mu - 3\sigma)$ to $(\mu + 3\sigma)$: contains about 99.7% of the data

Distribution Not Known. Then?

Chebyshev’s Theorem

- Chebyshev’s Theorem applies to all distributions regardless of their shape and thus can be used whenever the data distribution shape is unknown or it is non-normal or even if it is normal.
- Chebyshev’s Theorem states that at least $1 - (1/k^2)$ values will fall within k standard deviations of the mean that is (μ ± k.σ), regardless of the shape of the distribution, where k > 1.
- The values from the Theorem and Empirical values may not exactly match. It is an approximation when distribution of data is not known.
- The shown figure below is not in the normal distribution.
- It shows that for k = 2, 1-(1/2)^2 = 75% values will fall within (μ–2σ) to (μ+2σ).



Example

Let the average age of a professional employed by a particular computer company is 28 with a standard deviation of 6 years. A histogram of this company reveals that the data is **not** normally distributed but rather congregated in the 20s and few workers are over 40. Apply Chebyshev’s Theorem to determine within what range of ages would at least 80% of the workers’ ages fall.

At least 1-(1/k^2)values will fall within k standard deviations, so:

$$1-\frac{1}{k^2}=0.80$$

$$\frac{1}{k^2}=0.20$$

$$k=2.24$$

$$\mu-k.\sigma=28-2.24*6=14.56$$

$$\mu+k.\sigma=28+2.24*6=41.44$$

80% of the employees will fall in the range of 14.56 to 41.44 years.

Exercise

The water consumption in gallons per day is shown by the washing machines of 50 households. It is not in normal distribution.

1. Find the mean.

Answer: 15.48 gallons

2. Find the sample standard deviation.

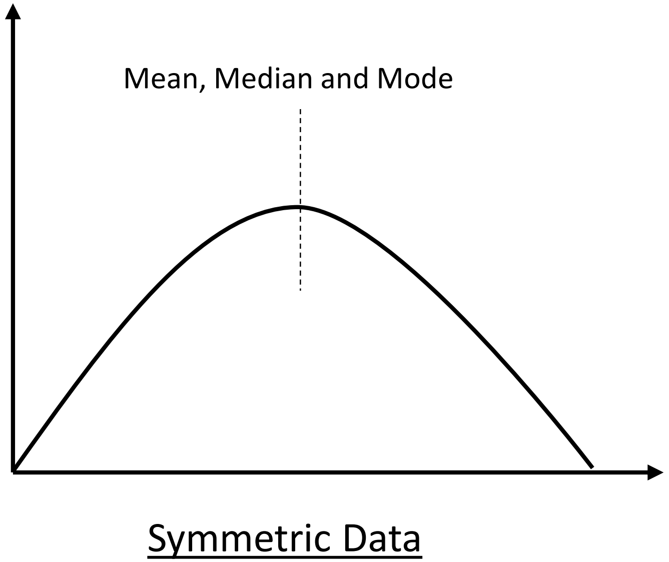
Answer: 1.23

3. What range of water consumption values range within 88.9% of the values?

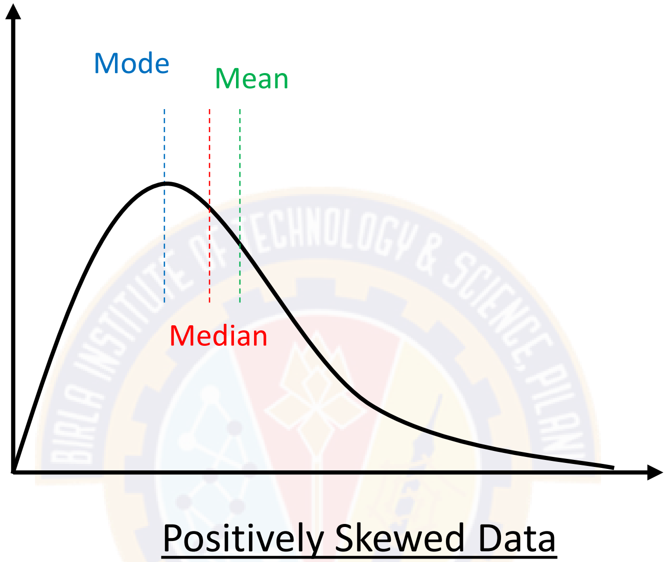
Answer: 11.79 to 19.17 gallons

Gallons of Water									
15	17	16	15	16	17	18	15	14	15
16	16	17	16	15	15	17	14	15	16
16	17	14	15	12	15	16	14	14	16
15	13	16	17	17	15	16	16	16	14
17	16	17	14	16	13	16	15	16	15

Measures of Shape: Skewness



Symmetric Data



Positively Skewed Data

Pearson coefficient of skewness

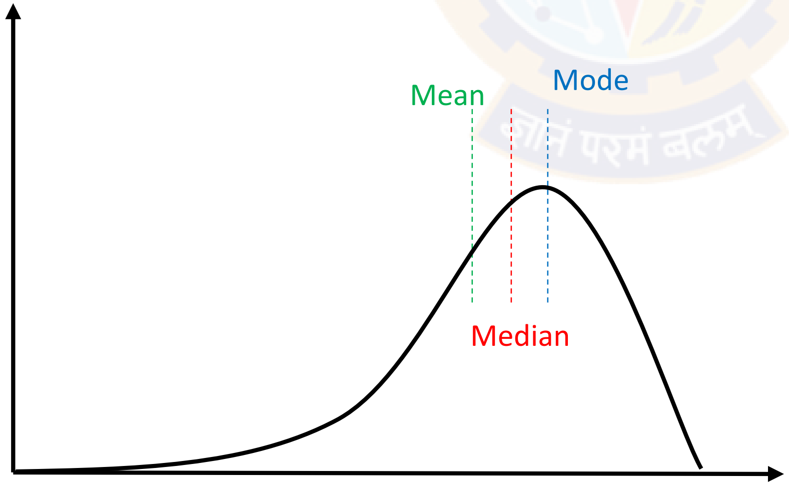
$$S_k = \frac{3.(mean\ -\ median)}{standard\ deviation}$$

Where if,

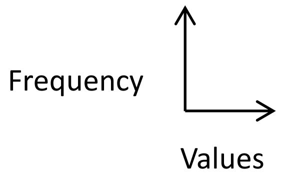
$S_k = 0$, *Symmetrical Distribution*

$S_k > 0$, *Positively Skewed Data*

$S_k < 0$, *Negatively Skewed Data*



Negatively Skewed Data



Exercise

Fill the table for the cells having x and justify the relationship among mean, mode and median as shown in the previous graphs:

Row #	Data Type	Val-1	Val-2	Val-3	Val-4	Val-5	Val-6	Val-7	Val-8	Mean	Median	Mode
1	x	20	25	30	35	35	40	45	50	x	x	x
2	x	10	11	11	13	13	14	30	35	x	x	x
3	x	25	30	35	40	40	45	10	12	x	x	x

Answer				
Row #	Data Type	Mean	Median	Mode
1	Symmetric	35	35	35
2	Positively Skewed	17.13	13.00	11, 13
3	Negatively Skewed	29.63	32.5	40

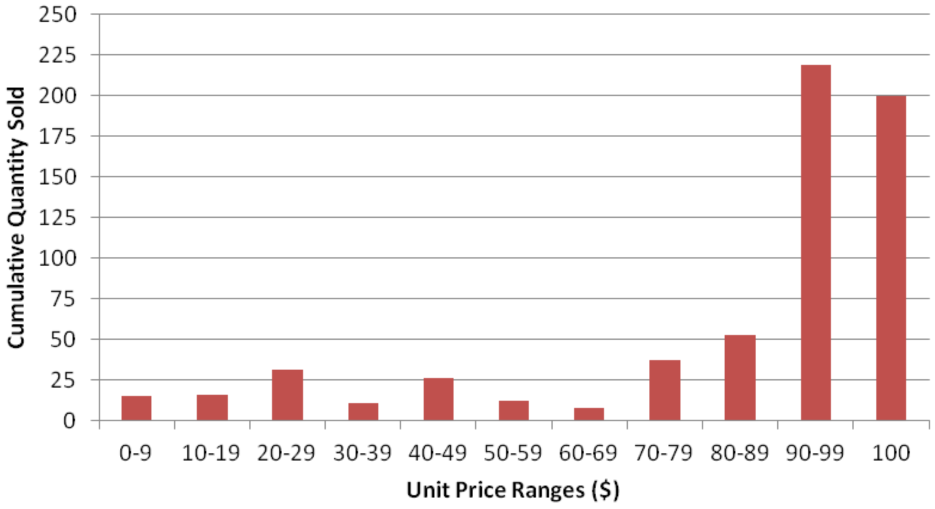
Visualization: Histograms

- Visualization of data is one of the most powerful and appealing techniques for data exploration.
  - Humans have a well developed ability to analyze large amounts of information that is presented visually.
  - Can detect general patterns and trends.
  - Can detect outliers and unusual patterns.
  - Can help to make useful data driven decisions.
- We will review Histograms and few other visualization techniques in this module to group the data and represent them pictorially.

Histograms:

- Divide the values into bins (classes) and show a bar plot of the number of objects in each bin.
- The height of each bar indicates the count of objects.
- **Example**: Data is shown for an electronics store. How to visually analyze the count of items sold in different price ranges?

Unit Price ($)	Quantity Sold
9	15
12	6
14	9
21	4
19	1
23	6
25	5
27	5
28	5
29	6
31	5
32	6
40	2
45	4
48	12
46	8
51	12
65	8
74	23
78	12
79	2
81	8
83	12
85	19
86	12
87	2
91	3
92	21
94	40
95	54
96	12
99	89
100	200



A possible decision to make - can this store plan to remove the items which are less than 90$ per unit?

Ungrouped / Grouped Data

- Let us say there is a raw data set with 60 data elements as shown in the table. This dataset is **ungrouped**.
- The **range** of this data set is the difference between the largest and the smallest that is = 12.0 – 2.3 = 9.7.
- As a rule of thumb 5 to 15 bins are selected. If there are too many bins, the data aggregation will not be helpful. Too few bins will not show any useful distribution. Let us say, 6 bins are selected for the data.
- **Width** of each bin = 9.7/6 ≈ 2 (round up to the next integer)
- The frequency distribution must start at a value equal to or lower than the lowest number of the ungrouped data and end at a value equal to or higher than the highest number. The lowest unemployment rate is 2.3 and the highest is 12.0, so the frequency distribution starts at 1 and ends it at 13.
- The table shows the **frequency distributed** grouped data. We are following a **left-end-inclusion** rule that means 3 is included in the 3-5 bin, 5 is included in the 5-7 bin etc.
- Average of two endpoints of the bin-interval in a bin is called the **mid-point**.
- **Cumulative Frequency** is also shown. Note that the last value (60) is equal to the total count of elements.
- Frequency distributed data is the **grouped** data. Let us review its plots on the next slide.

2.3	2.8	3.6	2.4	2.9	3.0	4.6	4.4	3.4	4.6	6.9	6.0
7.0	7.1	5.9	5.5	4.7	3.9	3.6	4.1	4.8	4.7	5.9	6.4
6.3	5.6	5.4	7.1	7.1	8.0	8.4	7.5	7.5	7.6	11.0	12.0
11.3	10.6	9.7	8.8	7.8	7.5	8.1	10.3	11.2	11.4	10.4	9.5
9.6	9.1	8.3	7.6	6.8	7.2	7.7	7.6	7.2	6.8	6.3	6.0

Ungrouped Data

Bin-Interval	Frequency
1-3	4
3-5	12
5-7	13
7-9	19
9-11	7
11-13	5
Sum	**60**

Mid Point
2
4
6
8
10
12

Cumulative Frequency
4
16
29
48
55
60

Grouped Data

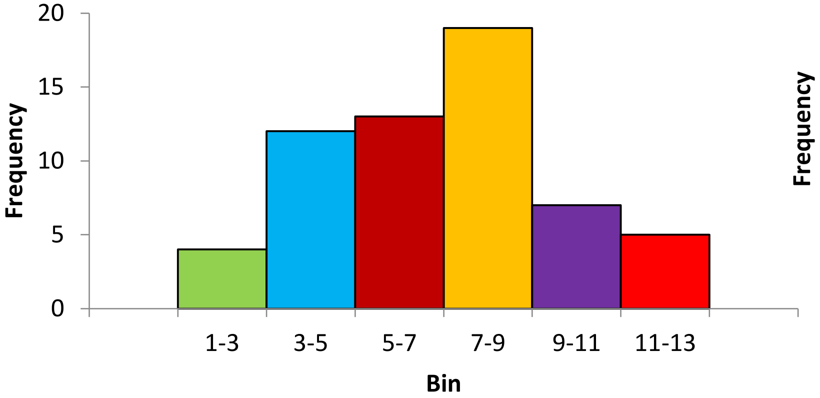
Histograms & Frequency Polygons

Bin-Interval	Frequency
1-3	4
3-5	12
5-7	13
7-9	19
9-11	7
11-13	5

Grouped Data

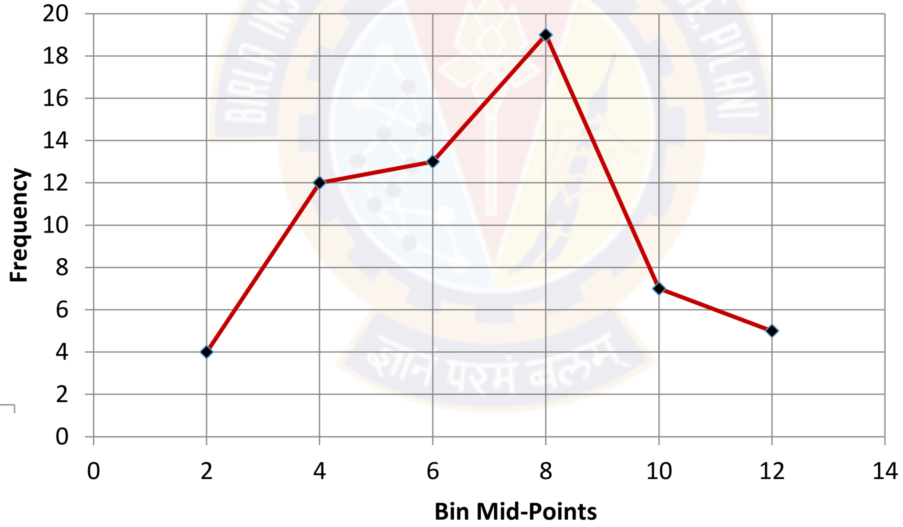
Mid Point
2
4
6
8
10
12

Cumulative Frequency
4
16
29
48
55
60



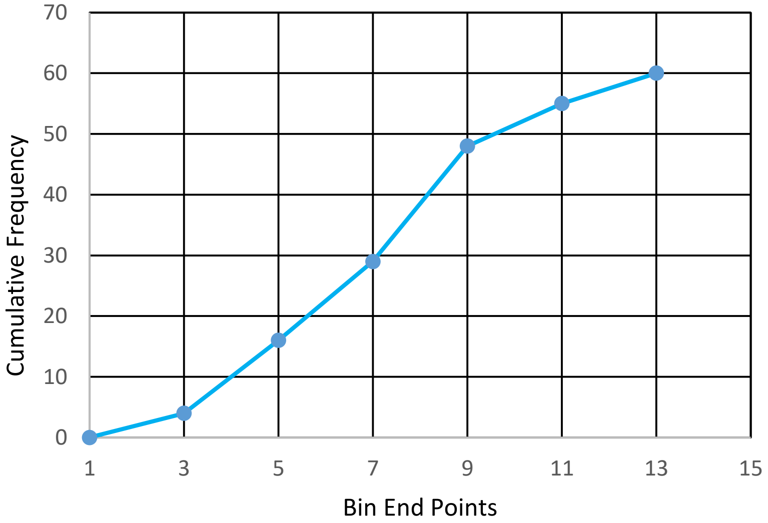
Histogram

In histogram, x-axis represents the bins and y-axis the frequency.



Frequency Polygon

In frequency polygon, x-axis represents the bin mid-points and y-axis the frequency.



Ogives

In ogives, x-axis represents the bin end-points and y-axis the cumulative frequency.

Exercise

Two rival mobile operators (OxyTel and VoxCom) enter into a southern state of India and start offering prepaid mobile packs that have combinations of voice, data, messaging and other services. Their pre-arrival marketing study suggest that college youth normally prefer less than ₹ 100 packs and working professionals prefer higher cost packs for more services. Both the operators target the entire population and all the plans are monthly. Opening day data for the first 2400 customers is collected for both the operators that is shown in the table below.

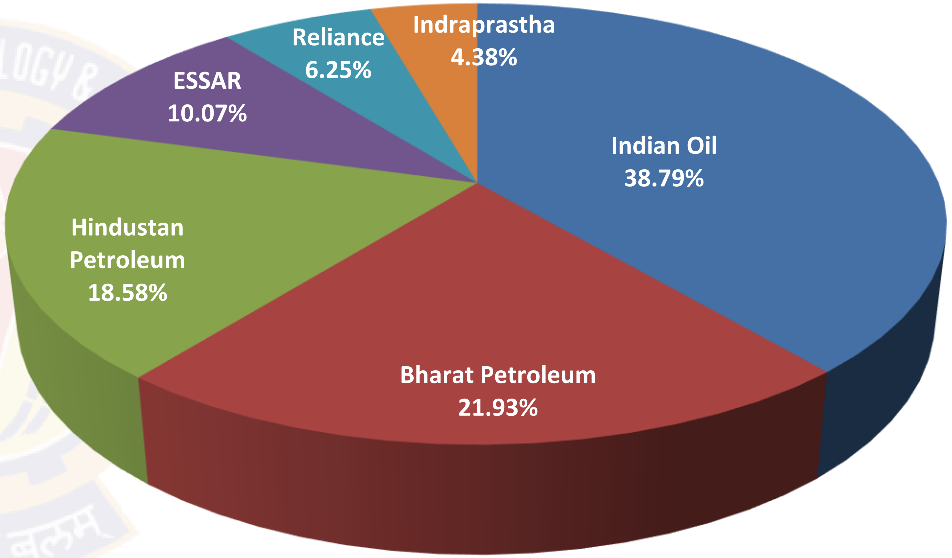
Draw this data using an appropriate visualization technique and derive at least two useful observations which might help the operators to improvise their marketing strategy.

x: Pack Rates (₹)	**0 < x <= 24**	**24 < x <= 49**	**49 < x <= 74**	**74 < x <= 99**	**99 < x <= 124**	**124 < x <= 149**	**149 < x <= 174**	**174 < x <= 199**	**199 < x <= 224**	**224 < x <= 250**
OxyTel	150	250	350	450	400	350	200	150	50	50
VoxCom	50	100	100	150	400	350	150	300	450	350

Pie Charts

- A **pie chart** is a 2D or 3D circular depiction of the data where the area of the whole pie represents 100% of the data.
- **Slices** of the pie represent a percentage breakdown of the data elements.

Sl. No.	Company	Sales (₹ millions)	Proportion %	Degrees
1	Indian Oil	₹ 3,72,824.00	38.79	139.65
2	Bharat Petroleum	₹ 2,10,783.00	21.93	78.96
3	Hindustan Petroleum	₹ 1,78,558.00	18.58	66.88
4	ESSAR	₹ 96,758.00	10.07	36.24
5	Reliance	₹ 60,044.00	6.25	22.49
6	Indraprastha	₹ 42,101.00	4.38	15.77
Sum		**₹ 9,61,068.00**	**100.00**	**360.00**



Examples:

How Proportion is calculated?

Indian oil = 3,72,824 / 9,61,068 = 0.3879 = 38.79%

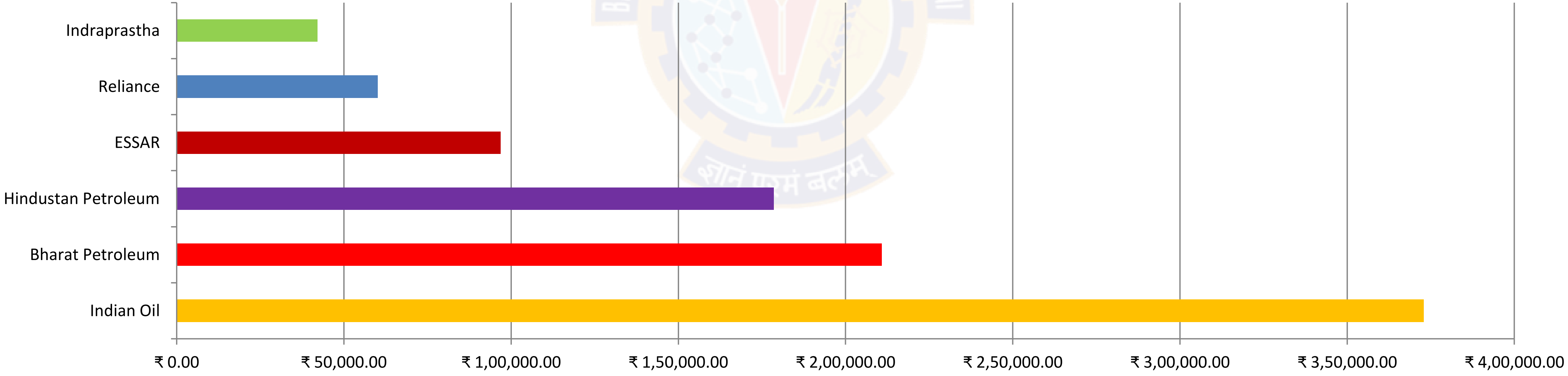
How Degrees is calculated?

Indian Oil = 0.3879 * 360^o = 139.65^o

Bar Charts

- A **bar chart** contains categories along one axis and a series of bars, one for each category, along the other axis.
- Typically, the length of the bar represents the magnitude of the measure (amount, frequency, money, percentage, etc.) for each category.
- The orientation of the bars could be horizontal or vertical.

Sl. No.	Company	Sales (₹ millions)
1	Indian Oil	₹ 3,72,824.00
2	Bharat Petroleum	₹ 2,10,783.00
3	Hindustan Petroleum	₹ 1,78,558.00
4	ESSAR	₹ 96,758.00
5	Reliance	₹ 60,044.00
6	Indraprastha	₹ 42,101.00

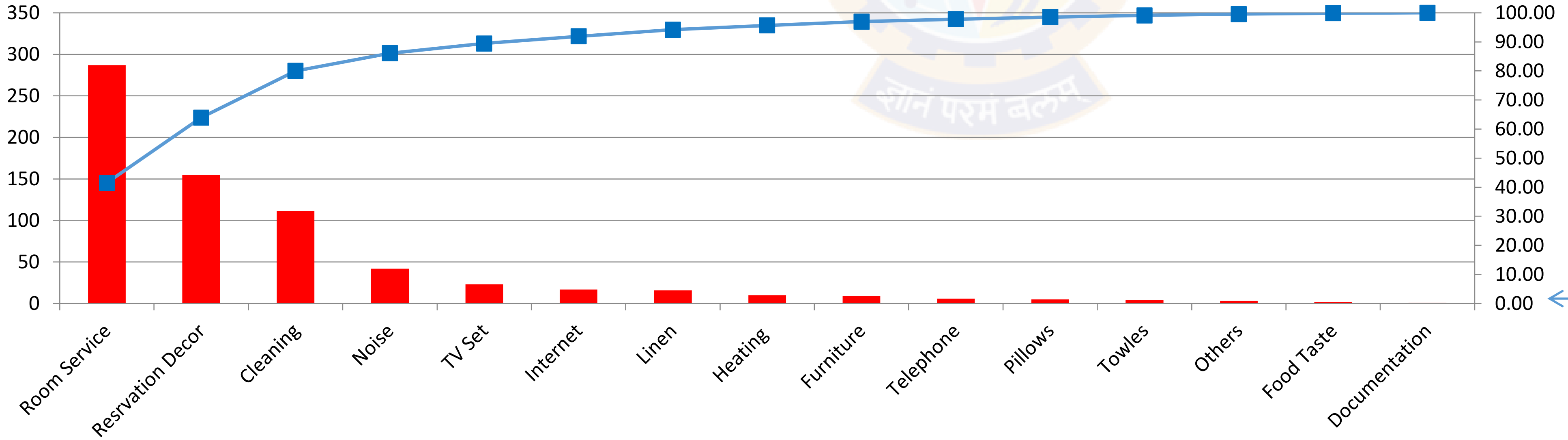


Pareto Charts

Notice 3-axis Chart!

- An important concept and movement in business is total quality management that needs accurate charts and plots.
- Pareto analysis is a quantitative technique tallying the number and types of defects that occur with a product or service.
- Data Analysts use this tally to produce a vertical bar chart that displays the most common types of defects, ranked in order of occurrence from left to right. The chart is called a **Pareto Chart,** which is a type of bar chart.
- The table shows the categorization of complaints and their counts arranged in the decreasing order for a hotel in a year.

Complaint Type	Number of Complaints	Cumulative Number	Cumulative %
Room Service	287	287	41.53
Reservation Decor	155	442	63.97
Cleaning	111	553	80.03
Noise	42	595	86.11
TV Set	23	618	89.44
Internet	17	635	91.90
Linen	16	651	94.21
Heating	10	661	95.66
Furniture	9	670	96.96
Telephone	6	676	97.83
Pillows	5	681	98.55
Towles	4	685	99.13
Others	3	688	99.57
Food Taste	2	690	99.86
Documentation	1	**691**	**100.00**



Visually obvious that first 5 types of issues account for around 90% of the complaints.

Cumulative % of Complaints.

Exercise

1. The following list shows the top six pharmaceutical companies in the United States and their sales figures ($ millions) for a recent year. Use this information to construct a pie and a bar chart to represent these six companies and their sales.

Sl. No.	Company	Sales
1	Pfizer	$ 52,921.00
2	Johnson & Johnson	$ 47,348.00
3	Merck	$ 22,939.00
4	Bristol-Myers	$ 21,886.00
5	Abbott	$ 20,473.00
6	Wyeth	$ 17,358.00

2. A leading mobile operator conducted a survey and received the following complaints as major issues for a period. Draw a Pareto Chart and identify which top issues account for 75% of the complaints.

Sl. No.	Complaint Category	Complaint Count
1	Frequent Call Drop	224
2	Slow Internet	150
3	Coverage Area	210
4	Faulty Billing	110
5	Useless Discount Offers	45
6	Service Issues	85

Thank You

Appendix

(for self study)

Median for Grouped Distribution

It is cumbersome to calculate the median for a large number of observations that can be grouped. In these scenarios, approximate median is calculated by:

Class	5-10	10-15	**15-20**	20-25	25-30	30-35	35-40	40-45
Frequency	5	6	**15**	10	5	4	2	2

Cumulative Frequency	5	11	**26**	36	41	45	47	**N = 49**

$$Median = l + \left\{ \frac{\frac{N}{2} - F}{f} \right\} . h$$

where,

median class the class whose frequencyis just above N/2.

$l =$ *lower limit of the median class*

$f =$ *frequencyof the median class*

$h =$ *width of the median class*

$F =$ *Cumulative frequencyof the class peceding the median class*

$N =$ *Sum of frequencies, the total obsevations*

- N = sum of frequencies = 49
- N/2 = 24.5, so cumulative frequency just above 24.5 is 26
- Corresponding class is: 15-20
- l = 15, f = 15, h = 5, F = 11
- So, median = 15 + ((24.5-11) / 15)* 5 = **19.5**