Objective: Upon completion of this session, you will be able to understand the various central measures for the given data, solve simple problems to understand the application of Probability concepts and interpretations.

Key takeaways:

- Data type classification
- Given the data, compute central measures such as mean, median, mode, quartiles.
- Properties of probability, Joint probability, conditional probability, marginal probability
- Understand the jargon and compute True positives, True negatives, False Positives and False negatives using a tree diagram or tabular format.
- Randomness and Sample space.
- Introduction to R, computing measures, use of set.seed
    - Sample and population and the role of probability in inferential statistics
    - Purpose and usage of randomness in sample

Problem 1: You plan to hire a taxi to commute. When you access the app based on the pick-up and drop point, you get an estimated charge for the travel. Name the factors that the app must be considering to arrive at an estimated cost and respective data types.

Distance – num
Time of the day – num; (If you consider as Morning, Afternoon, Evening then its categorical)
# of cars in proximity - Num
# of requests – Num
Type of the car – categorical
Sharing/individual – binary
Estimated Cost - num

Few more examples that you come across daily:

1. Blood pressure reading
2. Number of stocks traded
3. Education background
4. Type of groceries purchased
5. Price of petrol
6. Rating a Restaurant
7. Buy a car or not
8. Lifetime of a battery

Problem 2: Here is the data of experience of a CPEE class. We have grouped individuals into 6 groups and here is the data. Compute the average and median values for each group and list your observations.

Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
0	3	14	9	16	13
0	12	9.5	4.5	12	10
0	3.5	4.5	9	6.5	10
0	14	7.5	8	1	1
0	3	5	4	11	3
0	2.8	2	6	5	4
1	5	4.8	5	3	6
1	9	3.6	3.5	8	3.8
1	5.5	6	2.8	3	4
12	9	8.5	12	4	8

1. What is the average in each group?
2. What is the median in each group?
3. What is the average experience across all groups?
4. What do you observe?

Problem 3: You and your friends regularly order food online and prefer door delivery services. Each one believes that their respective service providers are very prompt. To understand it better, you have started collecting the time to deliver food in 20 different occasions for all. Here is the data of delivery times.

Time taken to deliver the order in minutes				
EagleBoys	**FoodPanda**	**Swiggy**	**PiazzaHut**	**Dominos**
30	**39**	**33**	**30**	**35**
35	**37**	**31**	**35**	**23**
23	**35**	**25**	**23**	**35**
12	**33**	**37**	**12**	**33**
15	**31**	**28**	**15**	**30**
16	**25**	**36**	**16**	**31**
19	**37**	**20**	**19**	**25**
31	**28**	**30**	**31**	**37**
35	**10**	**35**	**35**	**28**
21	**46**	**23**	**0**	**36**
39	**30**	**12**	**60**	**20**
37	**35**	**15**	**37**	**12**

35	**23**	**16**	**35**	**15**
33	**12**	**19**	**33**	**16**
31	**15**	**31**	**31**	**19**
25	**16**	**35**	**25**	**31**
37	**19**	**0**	**37**	**35**
28	**31**	**60**	**28**	**21**
36	**35**	**37**	**36**	**44**
20	**21**	**35**	**20**	**32**

1. Now that you know central measures help you understand data better you go ahead with computing the central measures. (mean, median, mode, quartiles, range, inter-quartile range, standard deviation)
2. What do you observe?
3. Do you still believe that all the service providers are prompt in their services?

Other applications

1. Performance of batsmen in a cricket team
2. Performance of stocks in market
3. Performance of machines in a manufacturing unit

Problem 3:

1. Two people work in a factory making parts for cars. The table shows how many complete parts they make in one week.

Worker	Mon	Tue	Wed	Thu	Fri
Philip	20	21	22	20	21
Mathews	30	15	12	36	28

(a) Find the mean, median and range for Philip and Mathews.
(b) Who is more consistent?

(a) Philip: Mean= (20+21+22+20+21)/5=20.8, Median=(20 20 21 21 22) = , Range=22-20=2

Matthews: Mean=(30+15+12+36+28)/5=24.2, Median=12 15 28 30 36, Range=36-12=24

(b) Look for least standard deviation/variance/MAD (Mean Absolute Deviation) for the spread.

In our case range also gives a better picture. Hence Philip.

2. Find the mode for 8,6,2,4,6,8,10,8

The frequency for 8 is 3, and all other values occur less frequently. Therefore, the mode is 8

3. Analyze the performance of your class in the first WUQ taken at INSOFE

Scores: 11, 7.5, 8.5, 10, 10, 10.5, 5.5, 10, 9, 9.5, 5.25, 8, 6.5, 10.5, 8.75, 0, 6, 6, 6.75, 8.75, 0, 9.5, 7.5, 8.5, 7

(a) How is the spread of the scores? Compute range, variance & standard deviation

(b) Which central tendency measure is the most represent

tative of the performance of the class?

City 1	29	32	36	40	43	37	36	33	32	37	31	29
City 2	20	24	31	37	40	38	37	34	34	33	28	23
City 3	23	26	32	38	41	40	35	33	35	37	30	25
City 4	20	24	29	34	37	36	32	30	33	32	27	23
City 5	19	24	29	38	43	38	33	34	36	34	29	23

(c) Find the 25th percentile, 50th percentile and 75th percentile for this data.

a) Range = Max-Min;

Variance = $\sum(x- \text{mean})^2/n$;

Stdev = sqrt(Variance)

b) Check for IQR.

If value is outside Whiskers (Q3+1.5*IQR and Q1-1.5*IQR) it has an outlier. Extreme outlier is (Q3+3IQR). Since our data has 0s, mean may not be a good measure. Median is a better measure since outliers doesn't effect it.

4. Temperatures in 5 cities measured on 12 days is given below. The weather department says that two cities have similar weather. Use central tendencies to identify those two cities.

	Mean	Median	Mode	Stdev
City 1	34.58333	34.5	29	4.33712
City 2	31.58333	33.5	37	6.487167
City 3	32.91667	34	35	5.915439
City 4	29.75	31	32	5.327885
City 5	31.66667	33.5	29	7.062492

1. A large retailer store regularly orders cartons of Pineapples. The average weight of the cartons is supposed to be 22 kgs. Random samples of cartons from two suppliers were weighed. The weights in kgs of the cartons were

Supplier – I	17	22	22	22	27
Supplier - II	17	19	20	27	27

a. Compute the range of carton weights from each supplier

Range is the difference between the smallest and the largest values. For both the cases, the range is same (10kgs)

b. Compute the mean weight of cartons from each supplier.

In both cases the mean is 22 kgs.

c. Look at the two samples again. The samples have the same range and mean. How do they differ? The retailer store uses one carton of blueberries in each blueberry muffin recipe. It is important that the cartons be of consistent weight so that the muffins turn out right.

Supplier I provide more cartons that have weights closer to the mean. Or, put another way, the weights of cartons from supplier I are more clustered around the mean. The retailer store might find supplier I might be more satisfactory.

Although, the range tells the difference between the largest and smallest values in a distribution, it does not tell how much other values may vary from one another or from the mean.

2. What is the probability that we get a 5th Tuesday in a 30-day month?

30 days => 4 weeks +2 days;
Last two days can be any one of the following: {Sun-Mon, Mon-Tues, Tues-Wed, Wed-Thurs, Thurs-Fri, Fri-Sat, Sat-Sun}
Of them Tuesday occurs in two cases; Hence probability = 2/7

3. Below is a table of graduates and post graduates

	Graduate	Post Graduate	Total
Male	19	41	**60**
Female	12	28	**40**
Total	**31**	**69**	**100**

a) What is the probability that a randomly selected individual is a male and a graduate? What kind of probability is it (Marginal/ Joint/Conditional)?

Joint Probability. P (Male and Graduate) = 19/100

b) What is the probability that a randomly selected individual is a male?

Marginal Probability: P(Male) = 60/100

c) What is the probability of a randomly selected individual being a graduate? What kind of probability is this?

Marginal Probability. P (Graduate) = 31/100

d) What is the probability that a randomly selected person is a female given that the selected person is a post graduate? What kind of probability is this?

Conditional Probability. P (Female|Post Graduate) = 28/69

4. In a region during a 1 year period, there were 1000 deaths. It was observed that 321 people died of a renal failure and 460 people had at least one parent with renal failure. Of these 460 people, 115 died of renal failure. Calculate the probability of a person that he dies of renal failure if neither of his parents had a renal failure

Ans: Let H=the event that at least one of parents of the randomly selected man die of cause related to renal failure.
D= event that the randomly selected man died of renal failure.

D/H	Parent had RF	Parent ! have RF	Total

People died of RF	115	206	321
People !died of RF	345	334	679
Total	460	540	1000

$$P(D|H') = \frac{P(D \cap H')}{P(H')}$$
$$= 206/540 = 38\%$$

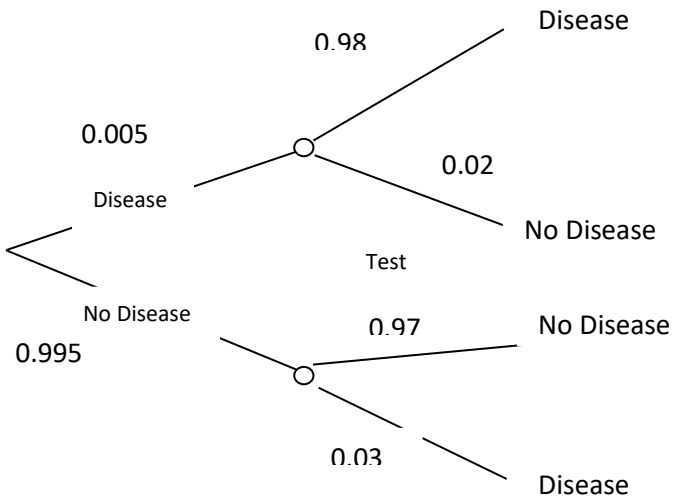
5. 0.5 percent of the population of an area is affected by a particular disease. A test is developed to detect the disease. This test gives a false positive 3% of the time and false negative 2% of the time.

   a) Draw the tree diagram for this problem.

   b) What is the probability that the test gives a positive result?

   c) If a person's test turns out to be positive, what is the probability that he (actually) has the disease

Ans: a)



b) We want to compute P(T). We do so by conditioning on whether or not Joe has the disease:
P(T) = P(T|D)P(D) + P(T|Dc)P(Dc) = (.98)(.005) + (.03)(.995)
By Law of total probability

c) We want to compute P(D|T) = P (D ∩ T)/P(T)
= P(T|D) P(D)/ (P(T|D) P(D) + P(T|Dc) P(Dc))
= (.98) (.005)/ ((.98) (.005) + (.03) (.995)) ≈ .14

6. Let three fair coins be tossed. Let Event A = {all heads or all tails}, Event B = {at least two heads}, and Event C = {at most two tails}. Of the pairs of events, (A, B), (A, C), and (B, C), which are independent and which are dependent? (Justify).

If A and B are independent, then $P(A \cap B) = P(A) * P(B)$.
If A & B are dependent, $P(A \cap B) = P(A)*P(B|A) = P(B)*P(A|B)$ OR $P(A \cap B) \neq P(A)*P(B)$

We write the event space for each of A, B and C.

A = {HHH, TTT},
B = {HHH, HHT, HTH, THH},
C = {HHH, HHT, HTH, THH, HTT, THT, TTH}.

$P(A \cap B) = 1/8$ and $P(A) \cdot P(B) = (2/8)(4/8) = 1/8$ so A and B are independent.
$P(A \cap C) = 1/8$ and $P(A) \cdot P(C) = (2/8)(7/8)$, so A and C are dependent.
$P(B \cap C) = 4/8$ and $P(B) \cdot P(C) = (4/8)(7/8)$, so B and C are dependent

7. A bank has developed an analytical model that helps them assess the credit worthiness of individuals and offer loans accordingly. To validate the performance of the model, they constructed a classification matrix on historical data.

	Predicted as credit worthy	**Predicted as not credit worthy**
Truly credit worthy	8000	900
Truly not credit worthy	100	1000

a. Identify "True Positives, True Negatives, False positives and False Negatives" from the table and compute "Accuracy, Precision, Recall and F1 statistic". (Please write the formula used to calculate each metric and substitute appropriate values to score.)

TP= 8000, TN= 1000, FP= 100 FN=900

Accuracy= (8000+1000)/ (8000+900+100+1000)

Precision= 8000/ (8000+100)

Recall= 8000/ (8000+900)

F1 Statistic= 2*Precision*Recall/ (Precision + Recall)

b. In this analysis, will you be more worried about false positives or false negatives?
In this case, we would be more worried about false positives (i.e. predicting a non-credit worthy person as credit worthy)

8. Do you want to simulate the Monte hall problem and check whether the changing to another door has the higher odds of winning the Car? Please click on the link below and play.

http://www.seas.upenn.edu/~probabil/Monte-HTML/monte4.html