Important Topics for Statistical Learning for Data Science

1) Explain the following terms along with example:
i) Elements; ii) Variables iii)Data iv) sample v)population

2) Consider the experiment of rolling a pair of dice. Suppose that we are interested in the sum of the face values showing on the dice. What is the probability of obtaining a value 7?

3) Calculate sample mean of the following class size data for a sample of five college classes.
46 54 42 46 32

4) Differentiate Categorical and Quantitative data with examples.

5)

Frequency Distribution of soft drink purchases:

Soft Drink	Frequency
Coca-Cola	20
Diet Coke	10
Dr. Pepper	8
Pepsi	15
Sprite	7
total	**60**

Calculate the relative frequency for Coca-Cola and Diet Coke.

6) Differentiate between Population & Sample. What are the benefits of sampling.

Define Covariance & Coefficient of correlation. For below given data:

City	Burger (X)	Movie Ticket (Y)
Tokyo	5.99	32.66
London	7.62	28.41
New york	5.75	20
Sydney	4.45	20.71
Chicago	4.99	18
Boston	4.39	16

Compute Covariance & Coefficient of correlation. State about correlation between them.

7) Construct a frequency distribution for below (20) given data and a relative frequency distribution. Use intervals of 6 days.
(4 12 8 14 11 6 7 13 13 11 11 20 5 19 10 15 24 7 29 6)

8) List the different ways to summarize numerical data? A manufacturer of insulation randomly selects 20 winter days and records the daily high temperature: (24, 35, 17, 21, 24, 37, 26, 46, 58, 30, 32, 13, 12, 38, 41, 43, 44, 27, 53, 27). Plot histogram and Polygon chart for the given data.

9) Short note on various Measures of variations.

10) Use the data given below to construct relative frequency using 13 equal intervals.
83 51 66 61 82 65 54 56 92 60 65 87 68 64 51 70 75 66 74 68 44 55 78 69 98 67 82 77 79 62 38 88 76 99 84 47 60 42 66 74 91 71 83 80 68 65 51 56 73 55

10) Suppose that the data for analysis is salary in thousands of Dollars
(30,36,47,50,52,52,56,60,63,70,70,110) Show the Boxplot of the data.

11) What are scales of measurements? Compare Categorical and Quantitative Data?

12) Short note on Binomial Distribution.

13) Consider the experiment of rolling a pair of dice. Suppose that we are interested in the sum of the face values showing on the dice. What is the probability of obtaining a value 7?

14) Three identical boxes contain red and white balls. The first box contains 3 red and 2 white balls, the second box has 4 red and 5 white balls, and the third box has 2 red and 4 white balls. A box is chosen very randomly and a ball is drawn from it. If the ball that is drawn out is red, what will be the probability that the second box is chosen?

15) Retirement policy is to be presented to top management. To know the support of the policy a manager conducts a poll.

	Machinists	Inspectors
Strongly support	9	10
Mildly support	11	3
Undecided	2	2
Mildly oppose	4	8
Strongly oppose	4	7

a)What is the prob that a machinist randomly selected from the polled group mildly supports the package?
b) What is the prob that an inspector randomly selected from the polled group is undecided? c) What is the prob that a worker (machinist or inspector) randomly selected from the polled group strongly or mildly supports the package.?
d) What types of probability estimates are these?

16) Explain Continuous Probability Distribution & its Types. Find the Normal Probabilities for following cases:
a)Suppose X is normal with mean 18.0 and standard deviation 5.0. Find P(X > 18.6).
b) Suppose X is normal with mean 18.0 and standard deviation 5.0. Find P(18 < X < 18.6).
c) Suppose X is normal with mean 18.0 and standard deviation 5.0. • Now Find P(17.4 < X < 18).

17) Discuss Discrete, Binomial and Poisson probability Distribution?
18) Describe Multiplication Law and Baye's Theorem with example?
19) Discuss Stratified Random Sampling, Cluster Sampling and Systematic Sampling?
20) The mean expenditure per customer at a tire store is $75.00,with a standard deviation of $8.00. If a random sample of 30 customers is taken, what is the probability that the sample average expenditure per customer for this sample will be $77.00 or more?
21) Suppose that during any hour in a large department store, the average number of shoppers is 448, with a standard deviation of 21 shoppers. What is the probability that a random sample of 49 different shopping hours will yield a sample mean between 441 and 446 shoppers?
22) Explain different types of sampling methods with the help of examples.
23) Consider a hypothesis H0 where $\phi 0 = 5$ against H1 where $\phi 1 > 5$. The test is Right tailed or left tailed? Explain.
24) Identify dependent and independent attributes from following statement:

How does the amount of makeup one applies affect how clear their skin is?
25) Consider a hypothesis H0 where $\phi 0 = 5$ against H1 where $\phi 1 > 5$. The test is Right tailed or left tailed? Explain.
26) What is hypothesis testing? What are possible errors in Hypothesis testing? What is the relation between level of significance & the rejection region.
27)

Nadir is testing an octahedral dice to see if it is biased. The results are given in the table below:
Test the hypothesis that the dice is fair.

Score	1	2	3	4	5	6	7	8
Frequency	8	10	11	7	12	14	10	7

28) What is Null and Alternative Hypotheses? Describe Type I and Type II error?
29) Explain Goodness of fit Test?

Find Linear regression equation for the following two sets of data:

x	2	3	5	7	9	3	11	13	8	5
y	4	5	7	10	15	12	4	5	10	9

30) Explain least squares method with example.
31) Differentiate between linear regression and multiple regression.
32) Explain Simple Linear Regression Model.
33) Which tests can be used to determine whether a linear association exists between the dependent and independent variables in a simple linear regression model?
34)

Find Linear regression equation for the following two sets of data:

x	2	4	6	8
y	3	7	5	10

Predict the value for (3,7) and (8,11)

35) Explain the least square method. What are the two basic categories of least-squares problems?
36)

a) Evaluate the following dataset to fit a multiple linear regression model.

y	x1	x2
140	60	22
155	62	25
159	67	24
179	70	20
192	71	15
200	72	14
212	75	14
215	78	11

37)

Find Linear regression equation for the following two sets of data:

x	2	3	5	7	9	3	11	13	8	5
y	4	5	7	10	15	12	4	5	10	9

Predict value for (2, 5) and (6,11)

38)

The following table shows the midterm and final exam grades obtained for students in Database course. Use linear regression to predict the final exam grade of a student who received 86 on the midterm exam.

Midterm Exam(x)	72	80	81	74	94	59	83	65	33	88	81
Final Exam (y)	84	63	77	78	90	49	79	77	52	74	90

39) Assume that you are the new owner of a small ice cream shop in a little village. You noticed that there was more business in the warmer months than the cooler months as shown in the table below. Compute covariance and correlation coefficient to support your assumption.

Temperature (x)	Customer (y)
98	15
87	12
90	10
85	10
95	16
75	7

40) Where we use nonparametric tests? Explain Wilcoxon signed rank test with example.
41) Consider the following set of data:
{23.32 32.33 32.88 28.98 33.16 26.33 29.88 32.69 18.98 21.23 26.66 29.89}
What is the lag-one sample autocorrelation of the time series?
42) Explain any one nonparametric method in time series analysis in detail.
43) List common types of data patterns that can be identified for time-series plot.

44) Consider the following Time series data:

Week	1	2	3	4	5	6
Value	20	18	14	16	11	13

Using the naive method (most recent value) as the forecast for the next week, compute the following measures of forecast accuracy.
a. Mean absolute error.
b. Mean squared error.
c. Mean absolute percentage error.
d. What is the forecast for week 7?

45) Discuss Wilcoxon signed-Rank Test and Mann-Whitney-Wilcoxon Test?