Basic Theory Questions

Question 1

What are some challenges or limitations in the field of data science? Suggest some ways to overcome them.

Question 2

Explain the data science lifecycle or workflow. What steps are crucial towards successful completion of any data science project? Discuss using a sample use case.

Question 3

Using a sample case-study, discuss the different ethical considerations in data science projects. Suggest some ways to handle them.

Question 4

Discuss the scenarios where it becomes important to perform inferential statistics. Can descriptive and inferential statistics co=exist? Justify your answer with suitable answers.

Question 5

Which of the following skills is NOT typically required for a data scientist?

a. Programming

b. Domain knowledge

c. Graphic design

d. Statistics

Question 6

Which data science task involves discovering patterns and relationships in data without specific goals or labels?

a. Regression analysis

b. Clustering

c. Classification

d. Feature selection

Question 7

Discuss some real-world applications where,

1) Mean is preferred over mode.
2) Mode is preferred over median.
3) Median is preferred over mean and mode.

Question 8

Hypothesize certain real-world applications where median can mislead us. Suggest some ways to handle these scenarios.

Question 9

How can interpretability affect the results of data science projects. Justify with the help of one or two examples from the real-world.

Question 10

In context to data science, suggest ways in which the real world differs from ideal world.

Descriptive Statistics

Question 1

Is it possible to convert nominal data into ordinal data, or vice versa? If so, how might you do it? Discuss taking some real-world examples.

Question 2

Imagine a scenario where you are working with a purely ordinal data. Discuss what precautions need to be taken while analyzing this kind of data and why these precautions need to be taken in the first place.

Question 3

Assume that you have been hired for a data collection task; the data that needs to be collected is purely nominal in nature. What factors are supposed to be taken into consideration while collecting this nominal data. Also discuss the ramifications if these considerations are not taken into account.

Question 4

Discuss the inherent challenges involved in filling missing values in an ordinal dataset with the help of relevant examples.

Question 5

You are a data scientist in a reputable company working in the medical domain. An intern approaches you and asks if there are scenarios where nominal and ordinal data can be interchanged without affecting the quality and nature of the medical data collected. Suggest some insights to help your intern.

Question 6

Scenario: Imagine you have monthly sales data for a retail store for the past year.

What is the primary purpose of using a bar chart in this scenario, and how does it compare to other types of charts like line charts or pie charts?

Question 7

Suppose you conducted a survey asking participants to rate their satisfaction with a new product on a scale of 1 to 5 (1 being very unsatisfied, 5 being very satisfied). You want to present the results.

a. How could you use a bar chart to represent the survey responses effectively, and what would the x and y-axes represent in this case?

b. What would the height or length of each bar indicate in the chart, and how would you label the bars to make it clear to the audience?

c. Would you consider using any additional elements, such as error bars or annotations, to provide more context or information in the bar chart?

Question 8

You are analysing the age distribution of a town's population for urban planning purposes.

a. How would you construct a histogram to represent the age distribution effectively? What information would the x-axis and y-axis convey?

b. What might be the implications of observing a multimodal distribution in the age histogram, and how could this information be used for planning services and infrastructure?

c. When dealing with population data, what considerations should you take into account when setting the bin widths to ensure the histogram provides meaningful insights?

Question 9

You are analysing customer purchase amounts in an online store over a month to identify spending patterns.

a. How would you create a histogram to represent the distribution of customer purchase amounts effectively? What would each bar represent?

b. If you observe a long tail on the right side of the histogram, what does that imply about customer spending habits, and how might you use this information for business decisions?

c. What considerations should you keep in mind when labelling the axes and selecting the scale for the histogram when dealing with monetary values

Question 10

How can histograms effectively represent continuous data, and what challenges might arise when attempting to do so with bar charts? Discuss using relevant real-world applications.

Question 11

You are involved in the data collection process where numerical values are being collected. The values also differ significantly in their scale of magnitude. Discuss the limitations of using a bar chart to represent this kind of data.

Question 12

How can histograms be used to identify skewness or asymmetry in data distributions, and why is this challenging to achieve with bar charts?

Question 13

Provide examples of situations where pie charts are frequently misapplied and suggest alternative chart types that might be more appropriate. You can explain with the help of relevant use cases.

Question 14

How does the presence of a cluster or grouping of data points in a scatter plot differ from a more spread-out distribution, and what might these patterns indicate?

Question 15

With reference to scatter plots, answer the following set of questions:

a. Explain why scatter plots are often used to examine correlations between variables, and how this relates to the distinction between correlation and causation.

b. What additional analyses or information would you need to establish causation based on a scatter plot?

Question 16

Is there any way to determine whether a data distribution is skewed to the left or right based on the appearance of a box plot?

Question 17

You have two sets of data, Set A and Set B. Set A has a mean of 50, a median of 55, and a mode of 60. Set B has a mean of 55, a median of 50, and a mode of 50. Which set has the most symmetric distribution, and what can you infer about the shapes of their distributions based on these measures?

Question 18

Explain how measures of central tendency like the mean and median behave in the presence of a bimodal distribution. Provide an example scenario where identifying both modes is crucial.

Question 19

You are a data analyst at an educational institution and have collected test scores for two different tests, Test A and Test B. Test A has a mean score of 75 and a standard deviation of 10, while Test B has a mean score of 75 and a standard deviation of 5.

a. Compare the variability in the scores of Test A and Test B using their standard deviations.

b. Explain how the differences in standard deviation might indicate differences in test performance variability.

Random Variables and Probability Distributions

Question 1

You are a teacher, and you have collected exam scores from a class of 30 students. The scores follow a normal distribution with a mean of 75 and a standard deviation of 10.

a. What is the probability that a randomly selected student scored above 85?

b. If you were to select a random sample of 10 students from this class, what is the probability that their average score is below 70?

Question 2

In a factory, 95% of manufactured items are of acceptable quality (success), while 5% are defective (failure). If you randomly select 10 items for inspection:

a. What is the probability that exactly 3 of them are defective?

b. Calculate the mean and standard deviation of the number of defective items in a sample of 10.

Question 3

Assume that the heights of a population of adults follow a normal distribution with a mean of 170 cm and a standard deviation of 10 cm.

a. What percentage of the population is taller than 180 cm?

b. If you randomly select three individuals from this population, what is the probability that at least one of them is shorter than 160 cm?

Question 4

You are monitoring the time between customer arrivals at a store, and it follows an exponential distribution with an average rate of 4 arrivals per hour.

a. What is the probability that the time between two arrivals is less than 10 minutes (1/6 of an hour)?

b. Calculate the mean waiting time until the next customer arrives.

Question 5

You are conducting a random survey, and participants are asked to pick a random number between 1 and 100.

a. Define a random variable W that represents the number chosen by a participant.

b. What is the probability density function (PDF) for random variable W within the given range?
c. Calculate the probability that a participant selects a number between 30 and 50.

Question 6

You are playing a game of darts, and the probability of hitting the bullseye on each throw is 0.2.

a. Define a random variable Z that represents the number of throws until you hit the bullseye for the first time.

b. Calculate the probability mass function (PMF) for random variable Z.

c. What is the expected number of throws until you hit the bullseye for the first time?

Inferential Statistics

Question 1

Why is it often impractical or impossible to collect data from an entire population, necessitating the use of samples? Discuss with the help of relevant examples. Also illustrate scenarios that allow collection of population data.

Question 2

You are a market researcher conducting a survey to estimate the average monthly spending on a particular brand of smartphones among a population of smartphone users in a city. You randomly select a sample of 100 smartphone users and ask them about their monthly spending on the brand [Assume M = 1.96 for 95 % confidence].

Based on your sample data, you calculate the sample mean spending as $250 with a standard deviation of $30.

a. Calculate a 95% confidence interval for the true population mean monthly spending on this brand of smartphones.

b. Interpret the meaning of the 95% confidence interval. What does it tell you about the range within which you expect the true population mean spending to fall?

c. How would the width of the confidence interval change if you had used a larger sample size of 200 smartphone users instead of 100? Explain the concept of margin of error in this context.

Question 3

As a teacher, you want to determine the variance in exam scores to understand the dispersion of student performance. You collect data from a class of 40 students.

1. Calculate the sample variance without using Bessel's correction.
2. Calculate the sample variance using Bessel's correction.
3. Explain why using Bessel's correction provides a more accurate estimate of the population variance in this context.

Question 4

In a legal case, the null hypothesis (H0) is that the defendant is innocent, while the alternative hypothesis (Ha) is that the defendant is guilty. The court's decision is based on the evidence presented.

a. Explain what a Type I error would mean in this legal context.

b. What are the potential consequences of making a Type I error in a criminal trial?

Question 5

Discuss the following scenarios with the help of relevant real-world scenarios.

a. How does increasing the sample size affect the likelihood of making a Type I error in hypothesis testing?

b. How does increasing the sample size affect the likelihood of making a Type II error in hypothesis testing?

Question 6

In medical testing, a Type I error means incorrectly diagnosing a healthy person as having a disease (false positive), while a Type II error means failing to diagnose a person with the disease (false negative).

a. Imagine a scenario where a new medical test for a serious disease has been developed. Discuss the consequences of making a Type I error in this context.

b. How would the consequences of a Type II error differ from those of a Type I error in this medical testing scenario?

Question 7

A food manufacturer claims that the shelf life of its product is 100 days. To verify this claim, a quality control manager randomly selects a sample of 25 product items and records their shelf life in days. The sample data is as follows:

Sample Mean Shelf Life ($\bar{x}$) = 98 days Sample Standard Deviation (s) = 7 days

The quality control manager wants to perform a one-sample hypothesis test to determine whether there is enough evidence to conclude that the actual shelf life of the product is different from the claimed 100 days.

The hypotheses are as follows:

- Null Hypothesis (H0): The actual shelf life is 100 days ($\mu = 100$).
- Alternative Hypothesis (Ha): The actual shelf life is different from 100 days ($\mu \neq 100$).

Using a significance level (alpha) of 0.05, conduct a one-sample hypothesis test to determine whether there is evidence to support the claim that the actual shelf life is different from 100 days.

Question 8

Suppose you are a teacher and want to determine whether there is a significant difference in the average exam scores between two different classes, Class A and Class B. You collect the following data:

Class A (Sample 1):

- Sample Size (n1) = 30
- Sample Mean Score ($\bar{x}1$) = 85
- Sample Standard Deviation (s1) = 10

Class B (Sample 2):

- Sample Size (n2) = 35
- Sample Mean Score ($\bar{x}2$) = 90
- Sample Standard Deviation (s2) = 12

You want to perform a two-sample hypothesis test to determine whether there is enough evidence to conclude that there is a significant difference in the average exam scores between the two classes.

The hypotheses are as follows:

- Null Hypothesis (H0): The average exam scores in Class A and Class B are the same ($\mu1 = \mu2$).
- Alternative Hypothesis (Ha): The average exam scores in Class A and Class B are different ($\mu1 \neq \mu2$).

Using a significance level (alpha) of 0.05, conduct a two-sample hypothesis test to determine whether there is evidence to support the claim of a significant difference in average exam scores between the two classes.

Question 9

A teacher wants to determine whether there is a significant difference in the average exam scores among three different classes: Class A, Class B, and Class C. The teacher collects the following data:

Class A (Sample 1):

- Sample Size (n1) = 25
- Sample Mean Score ($\bar{x}1$) = 85
- Sample Variance (s1^2) = 64

Class B (Sample 2):

- Sample Size (n2) = 30
- Sample Mean Score ($\bar{x}2$) = 88
- Sample Variance (s2^2) = 72

Class C (Sample 3):

- Sample Size (n3) = 28
- Sample Mean Score ($\bar{x}3$) = 82
- Sample Variance (s3^2) = 68

You want to perform an Analysis of Variance (ANOVA) to determine whether there is enough evidence to conclude that there is a significant difference in the average exam scores among the three classes.

The hypotheses are as follows:

- Null Hypothesis (H0): The average exam scores in the three classes are the same ($\mu1 = \mu2 = \mu3$).
- Alternative Hypothesis (Ha): At least one class has a different average exam score.

Using a significance level (alpha) of 0.05, conduct an ANOVA to determine whether there is evidence to support the claim of a significant difference in average exam scores among the three classes.

Calculate the test statistic (F), critical value (if applicable), and decide regarding the null hypothesis. Additionally, provide a conclusion based on your findings in the context of the exam scores for Classes A, B, and C.

Question 10

A marketing research firm wants to determine whether there is an association between gender (male or female) and product preferences (Product A, Product B, or Product C). They conduct a survey of 300 individuals and record the following data:

- Gender:
    - Male: 120 respondents
    - Female: 180 respondents
- Product Preferences:
    - Product A: 80 males, 120 females
    - Product B: 30 males, 40 females
    - Product C: 10 males, 20 females

The marketing research firm wants to test whether there is a significant association between gender and product preferences. They plan to use the chi-square test for independence.

State the null hypothesis (H0) and the alternative hypothesis (Ha) for this test.

Perform the chi-square test for independence and calculate the chi-square statistic (χ^2) and the degrees of freedom. Use a significance level (alpha) of 0.05.

Based on your calculations, make a decision regarding the null hypothesis. Provide a conclusion based on your findings in the context of the association between gender and product preferences.