



STATISTICS QUESTIONS

SAYAN GHOSH

7 MARCH

Q1. What are the three measures of central tendency?

Q2. What is the difference between the mean, median, and mode? How are they used to measure the central tendency of a dataset?

Q3. Measure the three measures of central tendency for the given height data:

[178,177,176,177,178.2,178,175,179,180,175,178.9,176.2,177,172.5,178,176.5]

Q4. Find the standard deviation for the given data:

[178,177,176,177,178.2,178,175,179,180,175,178.9,176.2,177,172.5,178,176.5]

7 MARCH

Q5. How are measures of dispersion such as range, variance, and standard deviation used to describe the spread of a dataset? Provide an example.

Q6. What is a Venn diagram?

Q7. For the two given sets $A = (2,3,4,5,6,7)$ & $B = (0,2,6,8,10)$. Find:

(i) $A \cap B$

(ii) $A \cup B$

Q8. What do you understand about skewness in data?

Q9. If the data is right skewed then what will be the position of the median with respect to the mean?

7 MARCH

Q10. Explain the difference between covariance and correlation. How are these measures used in statistical analysis?

Q11. What is the formula for calculating the sample mean? Provide an example calculation for a dataset.

Q12. For normal distribution data, what is the relationship between its measure of central tendency?

Q13. How is covariance different from correlation?

Q14. How do outliers affect measures of central tendency and dispersion? Provide an example.

8 MARCH

Q1. What is the Probability density function?

Q2. What are the types of Probability distribution?

Q3. Write a Python function to calculate the probability density function of a normal distribution with a given mean and standard deviation at a given point.

Q4. What are the properties of Binomial distribution? Give two examples of events where the binomial distribution can be applied.

Q5. Generate a random sample of size 1000 from a binomial distribution with a probability of success of 0.4 and plot a histogram of the results using Matplotlib.

8 MARCH

Q6. Write a Python function to calculate the cumulative distribution function of a Poisson distribution with a given mean at a given point.

Q7. How is Binomial distribution different from Poisson distribution?

Q8. Generate a random sample of size 1000 from a Poisson distribution with a mean of 5 and calculate the the sample mean and variance.

Q9. How mean and variance are related in Binomial distribution and Poisson distribution?

Q10. In normal distribution with respect to mean position, where does the least frequent data appear?

9 MARCH

Q1: What are the Probability Mass Function (PMF) and Probability Density Function (PDF)? Explain with an example.

Q2: What is Cumulative Density Function (CDF)? Explain with an example. Why CDF is used?

Q3: What are some examples of situations where the normal distribution might be used as a model? Explain how the parameters of the normal distribution relate to the shape of the distribution.

Q4: Explain the importance of Normal Distribution. Give a few real-life examples of Normal Distribution.

9 MARCH

Q5: What is Bernaulli Distribution? Give an Example. What is the difference between Bernoulli Distribution and Binomial Distribution?

Q6. Consider a dataset with a mean of 50 and a standard deviation of 10. If we assume that the dataset is normally distributed, what is the probability that a randomly selected observation will be greater than 60? Use the appropriate formula and show your calculations.

Q7: Explain uniform Distribution with an example.

Q8: What is the z score? State the importance of the z score.

9 MARCH

Q9: What is Central Limit Theorem? State the significance of the Central Limit Theorem.

Q10: State the assumptions of the Central Limit Theorem.

10 MARCH

Q1: What is Estimation Statistics? Explain point estimate and interval estimate.

Q2. Write a Python function to estimate the population means using a sample mean and standard deviation.

Q3: What is Hypothesis testing? Why is it used? State the importance of Hypothesis testing.

Q4. Create a hypothesis that states whether the average weight of male college students is greater than the average weight of female college students.

Q5. Write a Python script to conduct a hypothesis test on the difference between two population means, given a sample from each population.

10 MARCH

Q6: What is a null and alternative hypothesis? Give some examples.

Q7: Write down the steps involved in hypothesis testing.

Q8. Define the p-value and explain its significance in hypothesis testing.

Q9. Generate a Student's t-distribution plot using Python's Matplotlib library, with the degrees of freedom parameter set to 10.

Q10. Write a Python program to calculate the two-sample t-test for independent samples, given two random samples of equal size and a null hypothesis that the population means are equal.

10 MARCH

Q11: What is Student's t distribution?
When to use the t-Distribution.

Q12: What is t-statistic? State the formula for t-statistic.

Q13. A coffee shop owner wants to estimate the average daily revenue for their shop. They take a random sample of 50 days and find the sample mean revenue to be \$500 with a standard deviation of \$50. Estimate the population means revenue with a 95% confidence interval.

Q14. A researcher hypothesizes that a new drug will decrease blood pressure by 10 mmHg. They conduct a clinical trial with 100 patients and find that the sample mean decrease in blood pressure

10 MARCH

is 8 mmHg with a standard deviation of 3 mmHg. Test the hypothesis with a significance level of 0.05.

Q15. An electronics company produces a certain type of product with a mean weight of 5 pounds and a standard deviation of 0.5 pounds. A random sample of 25 products is taken, and the sample mean weight is found to be 4.8 pounds. Test the hypothesis that the true mean weight of the products is less than 5 pounds with a significance level of 0.01.

Q16. Two groups of students are given different study materials to prepare for a test. The first group ($n_1 = 30$) has a mean score of 80 with a standard deviation of 10, and the second group ($n_2 = 40$) has a

10 MARCH

the mean score of 75 with a standard deviation of 8. Test the hypothesis that the population means for the two groups are equal with a significance level of 0.01.

Q17. A marketing company wants to estimate the average number of ads watched by viewers during a TV program. They take a random sample of 50 viewers and find that the sample mean is 4 with a standard deviation of 1.5. Estimate the population mean with a 99% confidence interval.

11 MARCH

Q1: What is the difference between a t-test and a z-test? Provide an example scenario where you would use each type of test.

Q2: Differentiate between one-tailed and two-tailed tests.

Q3: Explain the concept of Type 1 and Type 2 errors in hypothesis testing. Provide an example scenario for each type of error.

Q4: Explain Bayes's theorem with an example.

Q5: What is a confidence interval? How to calculate the confidence interval, explain with an example.

11 MARCH

Q6. Use Bayes' Theorem to calculate the probability of an event occurring given prior knowledge of the event's probability and new evidence. Provide a sample problem and solution.

Q7. Calculate the 95% confidence interval for a sample of data with a mean of 50 and a standard deviation of 5. Interpret the results.

Q8. What is the margin of error in a confidence interval? How does sample size affect the margin of error? Provide an example of a scenario where a larger sample size would result in a smaller margin of error.

Q9. Calculate the z-score for a data point with a value of 75, a population mean of

11 MARCH

70, and a population standard deviation of 5. Interpret the results.

Q10. In a study of the effectiveness of a new weight loss drug, a sample of 50 participants lost an average of 6 pounds with a standard deviation of 2.5 pounds. Conduct a hypothesis test to determine if the drug is significantly effective at a 95% confidence level using a t-test.

Q11. In a survey of 500 people, 65% reported being satisfied with their current job. Calculate the 95% confidence interval for the true proportion of people who are satisfied with their job.

Q12. A researcher is testing the effectiveness of two different teaching

11 MARCH

methods on student performance. Sample A has a mean score of 85 with a standard deviation of 6, while sample B has a mean score of 82 with a standard deviation of 5. Conduct a hypothesis test to determine if the two teaching methods have a significant difference in student performance using a t-test with a significance level of 0.01.

Q13. A population has a mean of 60 and a standard deviation of 8. A sample of 50 observations has a mean of 65. Calculate the 90% confidence interval for the true population mean.

Q14. In a study of the effects of caffeine on reaction time, a sample of 30 participants had an average reaction

11 MARCH

time of 0.25 seconds with a standard deviation of 0.05 seconds. Conduct a hypothesis test to determine if caffeine has a significant effect on reaction time at a 90% confidence level using a t-test.

12 MARCH

Q1. Calculate the 95% confidence interval for a sample of data with a mean of 50 and a standard deviation of 5 using Python. Interpret the results.

Q2. Conduct a chi-square goodness of fit test to determine if the distribution of colors of M&Ms in a bag matches the expected distribution of 20% blue, 20% orange, 20% green, 10% yellow, 10% red, and 20% brown. Use Python to perform the test with a significance level of 0.05.

Q3. Use Python to calculate the chi-square statistic and p-value for a contingency table with the following data and interpret the results of the test.

12 MARCH

	Group A	Group B
Outcome 1	20	15
Outcome 2	10	25
Outcome 3	15	20

Q4. A study of the prevalence of smoking in a population of 500 individuals found that 60 individuals smoked. Use Python to calculate the 95% confidence interval for the true proportion of individuals in the population who smoke.

Q5. Calculate the 90% confidence interval for a sample of data with a mean of 75 and a standard deviation of 12 using Python. Interpret the results.

12 MARCH

Q6. Use Python to plot the chi-square distribution with 10 degrees of freedom. Label the axes and shade the area corresponding to a chi-square statistic of 15.

Q7. A random sample of 1000 people was asked if they preferred Coke or Pepsi. Of the sample, 520 preferred Coke. Calculate a 99% confidence interval for the true proportion of people in the population who prefer Coke.

Q8. A researcher hypothesizes that a coin is biased toward tails. They flip the coin 100 times and observe 45 tails. Conduct a chi-square goodness of fit test to determine if the observed frequencies match the expected

12 MARCH

frequencies of a fair coin. Use a significance level of 0.05.

Q9. A study was conducted to determine if there is an association between smoking status (smoker or non-smoker) and lung cancer diagnosis (yes or no). The results are shown in the contingency table below. Conduct a chi-square test for independence to determine if there is a significant association between smoking status and lung cancer diagnosis.

	Lung Cancer: Yes	Lung Cancer: No
Smoker	60	140
Non-Smoker	30	170

12 MARCH

Q10. A study was conducted to determine if the proportion of people who prefer milk chocolate, dark chocolate, or white chocolate is different in the U.S. versus the U.K. A random sample of 500 people from the U.S. and a random sample of 500 people from the U.K. were surveyed. The results are shown in the contingency table below. Conduct a chi-square test for independence to determine if there is a significant association between chocolate preference and country of origin. Use a significance level of 0.01.

	Milk Chocolate	Dark Chocolate	Milk Chocolate
U.S (n=500)	200	150	150
U.K (n=500)	225	175	100

12 MARCH

Q11. A random sample of 30 people was selected from a population with an unknown mean and standard deviation. The sample mean was found to be 72 and the sample standard deviation was found to be 10. Conduct a hypothesis test to determine if the population mean is significantly different from 70. Use a significance level of 0.05.

13 MARCH

Q1. Explain the assumptions required to use ANOVA and provide examples of violations that could impact the validity of the results.

Q2. What are the three types of ANOVA, and in what situations would each be used?

Q3. What is the partitioning of variance in ANOVA, and why is it important to understand this concept?

Q4. How would you calculate the total sum of squares (SST), explained sum of squares (SSE), and residual sum of squares (SSR) in a one-way ANOVA using Python?

13 MARCH

Q5. In a two-way ANOVA, how would you calculate the main effects and interaction effects using Python?

Q6. Suppose you conducted a one-way ANOVA and obtained an F-statistic of 5.23 and a p-value of 0.02. What can you conclude about the differences between the groups, and how would you interpret these results?

Q7. In a repeated measures ANOVA, how would you handle missing data, and what are the potential consequences of using different methods to handle missing data?

Q8. What are some common posthoc tests used after ANOVA, and when would

13 MARCH

you use each one? Provide an example of a situation where a post-hoc test might be necessary.

Q9. A researcher wants to compare the mean weight loss of three diets: A, B, and C. They collect data from 50 participants who were randomly assigned to one of the diets. Conduct a one-way ANOVA using Python to determine if there are any significant differences between the mean weight loss of the three diets. Report the F-statistic and p-value, and interpret the results.

Q10. A company wants to know if there are any significant differences in the average time it takes to complete a task using three different software programs:

13 MARCH

Program A, Program B, and Program C. They randomly assign 30 employees to one of the programs and record the time it takes each employee to complete the task. Conduct a two-way ANOVA using Python to determine if there are any main effects or interaction effects between the software programs and employee experience level (novice vs. experienced). Report the F-statistics and p-values, and interpret the results.

Q11. An educational researcher is interested in whether a new teaching method improves student test scores. They randomly assigned 100 students to either the control group (traditional teaching method) or the experimental group (new teaching method) and

13 MARCH

administered a test at the end of the semester. Conduct a two-sample t-test using Python to determine if there are any significant differences in test scores between the two groups. If the results are significant, follow up with a posthoc test to determine which group(s) differ significantly from each other.

Q12. A researcher wants to know if there are significant differences in the average daily sales of three retail stores: Store A, Store B, and Store C. They randomly select 30 days and record the sales for each store on those days. Conduct a repeated measures ANOVA using Python to determine if there are any significant differences in sales between the three stores. If the results are significant,

13 MARCH

follow up with a posthoc test to determine which store(s) differ significantly from each other.

14 MARCH

Q1. Write a Python function that takes in two arrays of data and calculates the F-value for a variance ratio test. The function should return the F-value and the corresponding p-value for the test.

Q2. Given a significance level of 0.05 and the degrees of freedom for the numerator and denominator of an F-distribution, write a Python function that returns the critical F-value for a two-tailed test.

Q3. Write a Python program that generates random samples from two normal distributions with known variances and uses an F-test to determine if the variances are equal. The program should output the F-value,

14 MARCH

degrees of freedom, and p-value for the test.

Q4. The variances of the two populations are known to be 10 and 15. A sample of 12 observations is taken from each population. Conduct an F-test at the 5% significance level to determine if the variances are significantly different.

Q5. A manufacturer claims that the variance of the diameter of a certain product is 0.005. A sample of 25 products is taken, and the sample variance is found to be 0.006. Conduct an F-test at the 1% significance level to determine if the claim is justified.

Q6. Write a Python function that takes in

14 MARCH

the degrees of freedom for the numerator and denominator of an F-distribution and calculates the mean and variance of the distribution. The function should return the mean and variance as a tuple.

Q7. A random sample of 10 measurements is taken from a normal population with unknown variance. The sample variance is found to be 25. Another random sample of 15 measurements is taken from another normal population with unknown variance, and the sample variance is found to be 20. Conduct an F-test at the 10% significance level to determine if the variances are significantly different.

14 MARCH

Q8. The following data represent the waiting times in minutes at two different restaurants on a Saturday night: Restaurant A: 24, 25, 28, 23, 22, 20, 27; Restaurant B: 31, 33, 35, 30, 32, 36. Conduct an F-test at the 5% significance level to determine if the variances are significantly different.

Q9. The following data represent the test scores of two groups of students: Group A: 80, 85, 90, 92, 87, 83; Group B: 75, 78, 82, 79, 81, 84. Conduct an F-test at the 1% significance level to determine if the variances are significantly different.