



Problem Statement - 1

Q. Compare Mean and Median for Age and BMI

Calculate the mean and median for the age and BMI columns. Discuss why they might be different and what this implies about the data distribution. Is either of these distributions likely skewed?

Summary Statistics



Problem Statement - 2

Q. Quartiles and Percentiles of Medical Charges

Calculate the 10th, 25th, 50th, 75th, and 90th percentiles for charges. What do these percentiles tell you about the spread of medical charges? Identify which range (e.g., between 25th and 75th percentile) includes the majority of the data.



Problem Statement - 3

Q. Analyzing Variance and Standard Deviation by Smoking Status

Group the data by smoker status and calculate the variance and standard deviation of charges for smokers and non-smokers separately. What does this tell you about the variability of medical charges within each group?



Problem Statement - 4

Q. Relationship Between Age and Charges

Calculate the correlation between age and charges. What does the correlation value suggest about the relationship between age and medical charges? Is it a strong, weak, positive, or negative relationship?



Problem Statement - 5

Q. Outlier Analysis Using Box Plot

Create a box plot for the charges variable. Identify any potential outliers. How do outliers affect the mean and median of charges? Discuss how you might handle these outliers if you were preparing data for a machine learning model.

Normal Probability Plots



Problem Statement - 1

Q. Generate a normal probability plot for a dataset of your choice. What does the plot indicate about the normality of the data? Are there any noticeable deviations from the straight line?



Poisson Distribution



Problem Statement - 1

Q. Simulate a Poisson distribution with an average rate (λ) of 4. Generate a histogram of the simulated data and overlay the theoretical Poisson probability mass function. How does the simulated data compare to the theoretical distribution?

Poisson Distribution



Problem Statement - 2

Q. Collect a dataset (e.g., the number of daily customer visits to a store) and calculate the average number of events (λ). Use a Poisson distribution to predict the probability of having 0, 1, or 2 visits on a given day. How accurate are these predictions based on your dataset?

Poisson Distribution



Problem Statement - 3

Q. Compare the characteristics of the Poisson distribution with the normal distribution using a dataset of medical visits. Under what conditions would you choose to model the data using the Poisson distribution instead of the normal distribution?

Central Limit Theorem & Sampling Mean



Problem Statement - 1

Q. Effect of Sample Size on Distribution Shape

Draw samples of sizes 10, 50, 100, and 200 from the charges variable. For each sample size, plot the sampling distribution of the mean. Describe how the shape of the sampling distribution changes as the sample size increases. Why does this happen?

Central Limit Theorem & Sampling Mean



Problem Statement - 2

Q. Comparing Sample Mean to Population Mean

For a sample size of 50, draw 1000 samples from the charges variable and calculate the mean of the sampling distribution of the sample mean. Compare this mean with the population mean of charges. What do you observe, and how does this support the CLT?

Central Limit Theorem & Sampling Mean



Problem Statement - 3

Q. Standard Deviation of Sampling Distributions

Draw samples with sizes 30, 60, and 120 from the charges variable. Calculate the standard deviation of the sampling distribution of the sample mean for each sample size. How does the standard deviation change with increasing sample size, and why?

Central Limit Theorem & Sampling Mean



Problem Statement - 4

Q. CLT with Non-Normal Data

Test the CLT with a skewed variable, such as BMI, which may not follow a normal distribution. Draw samples of increasing sizes (e.g., 10, 30, 100) and plot the sampling distribution of the sample mean for each size. Does the CLT still apply? How does the distribution of the sample means compare to the population distribution of BMI?