Biostatistics (BIO545)

Quiz 2
Duration: 30 min
Each question carries 5 marks

Question 1: A researcher is studying the average commute time of employees at a large company. They randomly sample 100 employees and calculate the sample mean commute time. They repeat this process many times, each time drawing a new random sample of 100 employees and calculating the sample mean. Which of the following best describes the distribution of these sample means?

A) It will be identical to the distribution of individual commute times.

B) It will be approximately normal, regardless of the shape of the distribution of individual commute times, as long as the sample size is sufficiently large.

C) It will be skewed in the same direction as the distribution of individual commute times. D)

It will be uniform, as the sampling process eliminates any underlying patterns in the data.

Correct Answer: B

Explanation: This question directly tests understanding of the core concept of the CLT. The CLT states that the distribution of sample means will approach a normal distribution as the sample size increases, regardless of the shape of the original population distribution. 1 Option B captures this key idea.

Question 2: You are analyzing the weights of apples from an orchard. The distribution of individual apple weights is known to be slightly skewed to the right. You take multiple random samples of different sizes from the apple population and calculate the mean weight for each sample. Which of the following statements is TRUE regarding the distribution of these sample means?

A) The distribution of sample means will become more skewed to the right as the sample size increases.

B) The distribution of sample means will become more similar to the original skewed distribution as the sample size increases.

C) The distribution of sample means will become less skewed and more symmetrical as the sample size increases, eventually approximating a normal distribution.

D) The shape of the distribution of sample means will remain the same regardless of the sample

size. **Correct Answer: C**

Explanation: This question explores a nuance of the CLT – how the distribution of sample means changes with increasing sample size. While the CLT guarantees eventual normality, this question probes the *process* of convergence. As the sample size increases, the distribution of sample means becomes progressively less skewed and more normally distributed, converging towards a symmetrical bell curve. Option C correctly captures this dynamic.

Question 3: You are evaluating the performance of two different machine learning models designed to predict customer churn. Model A has an AUC-ROC of 0.85, while Model B has an AUC-ROC of 0.92. Which of the following is the *most accurate* interpretation of these results?

A) Model B is definitively better than Model A for all possible business scenarios.

B) Model B is better than Model A at all possible classification thresholds.
C) Model B is generally better than Model A, but the optimal model choice might depend on the specific costs associated with false positives and false negatives.

D) Model A is likely to have fewer false positives than Model B.

Correct Answer: C

Explanation: This question probes the limitations of AUC-ROC. While a higher AUC-ROC *generally* indicates better performance, it doesn't tell the whole story. The optimal model choice can depend on the specific business context, particularly the relative costs of false positives and false negatives. For example, if the cost of a false negative (failing to identify a churning customer) is very high, a model with a slightly lower AUC-ROC but better performance at a specific threshold might be preferred. Option C correctly highlights this nuance. Options A and B are too strong of a claim, as AUC-ROC is an aggregate measure and doesn't guarantee superiority at all thresholds. Option D is incorrect; a higher AUC-ROC doesn't necessarily mean fewer false positives.

Question 4: A model predicts the likelihood of a patient having a specific disease. The AUC-ROC for this model is 0.75. Which of the following statements is the *most accurate* interpretation of this value?

A) The model has a 75% chance of correctly classifying a patient.

B) The model can distinguish between patients with and without the disease 75% of the time.

C) If you randomly select one patient with the disease and one patient without the disease, the model will assign a higher probability of having the disease to the patient who actually has it 75% of the time.

D) The model is 75% accurate in its predictions.

Correct Answer: C

Explanation: This question targets the precise definition of AUC-ROC. The AUC-ROC represents the probability that a randomly chosen positive instance will be ranked higher by the classifier than a randomly chosen negative instance. Option C accurately captures this interpretation. The other options are common misinterpretations. AUC-ROC is not a measure of overall accuracy (Option A and D) or a direct measure of separability in a general sense (Option B). It's specifically about the ranking of positive vs. negative instances.

Question 5: A researcher is studying the prevalence of a genetic mutation in a population. They estimate that the probability of an individual having the mutation is 0.05. If two individuals are randomly selected from the population, what is the probability that *at least one* of them has the mutation? Assume the presence of the mutation in one individual does not affect the probability of it in another.

A) 0.0025 **B) 0.095** C) 0.9025 D) 0.10

Correct Answer: B

Explanation: This question applies the complement rule and the concept of independent events. It's easier to calculate the probability that neither individual has the mutation (0.95 * 0.95 = 0.9025). Then, using the complement rule, the probability that at least one has the mutation is 1 - 0.9025 = 0.0975. This is closest to option B

Question 6: A diagnostic test for a disease has a sensitivity of 90% and a specificity of 80%. The prevalence of the disease in the population is 10%. If a randomly selected individual tests positive, what is the probability that they actually have the disease?

A) ~ 78% B) ~ 36% C) ~ 90% D) ~ 10%

Correct Answer: B

Explanation: This question requires application of Bayes' theorem. We are looking for P(Disease | Positive Test). Bayes' theorem states:
P(Disease | Positive Test) = [P(Positive Test | Disease) * P(Disease)] / P(Positive Test)

We know:

● P(Disease) = 0.10 (prevalence)

● P(Positive Test | Disease) = 0.90 (sensitivity)

● P(Positive Test) = P(Positive Test | Disease) * P(Disease) + P(Positive Test | No Disease) * P(No Disease) = (0.90 * 0.10) + (0.20 * 0.90) = 0.09 + 0.18 = 0.27

Therefore, P(Disease | Positive Test) = (0.90 * 0.10) / 0.27 = 0.09 / 0.27 = 1/3 ≈ 0.33 which is closest to 36% in the options. This question tests the understanding of conditional probability and the importance of prevalence when interpreting diagnostic test results.

Question 7: A researcher is studying the occurrence of a rare disease in a population. They find that the *prevalence* of the disease is relatively low, but the *incidence* is relatively high. Which of the following conclusions is *most* likely true?

A) The disease is likely chronic and incurable.

B) The disease is likely acute and has a short duration, either through recovery or death.

C) The diagnostic criteria for the disease are likely very broad, leading to overestimation of both prevalence and incidence.

D) The disease is likely easily transmitted from person to person.

Correct Answer: B

Explanation: This question tests the relationship between prevalence, incidence, and disease duration. A low prevalence coupled with high incidence suggests that people are getting the disease relatively frequently, but they don't have it for very long. This points towards a short disease duration, characteristic of acute illnesses where individuals either recover or the disease is fatal quickly. If it were chronic (Option A), the prevalence would be expected to be higher. Options C and D, while potentially influencing prevalence and incidence, are not the most likely explanations for the observed pattern.

Question 8: A researcher is analyzing the distribution of blood pressure measurements in a sample population. They create a box plot of the data and observe that the box is relatively short, but the whiskers are quite long. What is the most likely interpretation of this observation?

A) The data is symmetrically distributed with few outliers. B) The data is skewed with many outliers. C)

The data is symmetrically distributed with many outliers. D) The data is skewed with few outliers. **Correct**

Answer: C

Explanation: A short box indicates that the interquartile range (IQR, containing the middle 50% of the data) is small, suggesting data clustered around the median. Long whiskers, however, imply a wider range of values beyond the IQR, indicating the presence of outliers. Since the box itself is short, the bulk of the data within the IQR is concentrated, but the long whiskers show that extreme values (outliers) are present. This combination points towards a symmetrical distribution with outliers. Skewness would typically be indicated by a longer box or differently sized

whiskers.

Question 9: You are comparing the distributions of cholesterol levels in two different groups of individuals using box plots. Group A has a median cholesterol level of 180 mg/dL and an IQR of 30 mg/dL. Group B has a median cholesterol level of 200 mg/dL and an IQR of 20 mg/dL. Which of the following is the most accurate interpretation?
A) Group B has a higher average cholesterol level and more variable cholesterol levels.

B) Group B has a higher average cholesterol level and less variable cholesterol

levels. C) Group A has a higher average cholesterol level and more variable cholesterol

levels.

D) Group A has a higher average cholesterol level and less variable cholesterol levels.

Correct Answer: B

Explanation: The median represents the "middle" value and is a good indicator of central tendency (often thought of as the "average"). Group B's median (200) is higher than Group A's (180), suggesting a higher average cholesterol. The IQR represents the spread or variability of the middle 50% of the data. Group B's IQR (20) is smaller than Group A's (30), indicating less variability in the central portion of Group B's cholesterol levels. Therefore, Group B has a higher average and less variability.

Question 10: A researcher wants to study the health habits of all adults in a large city. Due to resource constraints, they cannot survey everyone. Which sampling method would be *most appropriate* to ensure the sample is representative of the entire population and minimizes bias?

A) Convenience sampling by interviewing people at a local gym.

B) Snowball sampling by asking participants to refer their friends.

C) Stratified random sampling by dividing the city into neighborhoods and randomly selecting individuals from each neighborhood.

D) Cluster sampling by randomly selecting a few city blocks and surveying all adults on those

blocks. **Correct Answer: C**

Explanation: Stratified random sampling is the best option here. It divides the population into subgroups (strata) based on relevant characteristics (neighborhoods in this case) and then randomly samples from each stratum. This ensures representation from all parts of the city and reduces the risk of over- or under-representing specific groups. The other options have significant biases: convenience sampling is limited to gym-goers, snowball sampling is biased towards interconnected individuals, and while cluster sampling is a valid method, stratified sampling is generally better for ensuring representation when you have identifiable subgroups.