Biostatistics (BIO545)

Quiz 1

Each question carries 10 marks

1. Describe how an extreme outlier can influence the interpretation of centrality measures in a dataset and identify which centrality measures are more resistant to such outliers. Explain a graphical technique for detecting outliers.

Extreme outliers can significantly impact measures of central tendency:

- **Mean**: Highly sensitive to outliers since it is calculated by summing all values and dividing by the total number. A single extreme value can distort the mean. ● **Median**: More resistant to outliers since it is the middle value of the sorted dataset. ● **Mode**: Typically unaffected by outliers unless the outlier appears frequently.

Resistant Measures: The median and mode are more robust to extreme values.

Graphical Technique for Outlier Detection: A **box plot** is a useful tool for detecting outliers. In a box plot:

BOX PLOT

A box plot uses the relationships among the median, upper quartile, and lower quartile to describe the skewness of a distribution

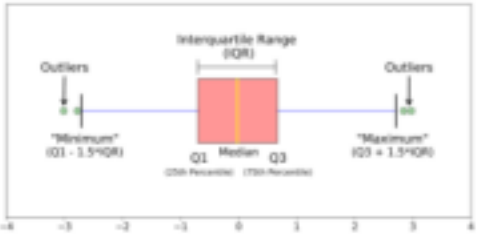
The upper and lower quartiles can be thought of conceptually as the approximate 75th and 25th percentiles of the sample—that is, the points 3/4 and 1/4 along the way in the ordered sample.

How can the median, upper quartile, and lower quartile be used to judge the symmetry of a distribution?

(1) If the distribution is symmetric, then the upper and lower quartiles should be approximately equally spaced from the median.

(2) If the upper quartile is farther from the median than the lower quartile, then the distribution is positively skewed.

(3) If the lower quartile is farther from the median than the upper quartile, then the distribution is negatively skewed.



- The **whiskers** extend up to 1.5 times the **interquartile range (IQR)**. ● **Outliers** are typically plotted as individual points beyond the whiskers.

2**.** Consider a box plot depicting the distribution of exam scores for Classes A and B. If Class A has a larger interquartile range (IQR) than Class B, does this conclusively indicate greater overall score variability in Class A? Discuss the limitations of relying solely on the IQR as a measure of diversity and explore alternative statistical measures or explanations that should be considered when comparing the distributions of the two classes.

A larger IQR in Class A suggests that the middle 50% of scores are more spread out, but it does **not** conclusively indicate greater overall variability.

Limitations of Relying Solely on IQR:

- **Does not account for full variability**: The IQR ignores the tails of the distribution. ● **Cannot detect extreme outliers**: Two distributions can have the same IQR but different variances due to extreme values.

Alternative Measures:

- **Standard Deviation**: Considers the entire dataset and captures the spread more comprehensively.
- **Range**: Measures the difference between the maximum and minimum values.
- **Variance**: Provides a measure of overall dispersion.
- **Box Plot Comparison**: Visualizing outliers and whisker lengths helps understand total variability.

3. A study comparing two vaccines, A and B, found that Vaccine A has a mean reduction of 45% with a standard deviation of 5%, and Vaccine B has a mean reduction of 55% with a standard deviation of 12%. Calculate the coefficient of variation (CV) for both vaccines and identify which vaccine shows more consistency. If the standard deviation for Vaccine B is reduced to 7%, how would this affect the CV?

Ans: A study comparing two vaccines, Vaccine A and Vaccine B, measured their effectiveness in terms of the percentage reduction in disease cases. The mean and standard deviation for each vaccine are provided:

- Vaccine A: Mean reduction = 45%, Standard deviation = 5%
- Vaccine B: Mean reduction = 55%, Standard deviation = 12%

The coefficient of variation (CV) is used to assess the relative variability of each vaccine. A lower CV indicates greater consistency in effectiveness, while a higher CV suggests more variability.

Coefficient of Variation Calculation

The coefficient of variation (CV) is calculated using the formula:

$$CV = (\text{Standard Deviation} / \text{Mean}) \times 100\%$$

Vaccine A:

$$CV_A = (5 / 45) \times 100$$
$$CV_A = (0.1111) \times 100$$
$$CV_A = 11.11\%$$

Vaccine B:

$CV_B = (12 / 55) \times 100$
$CV_B = (0.2182) \times 100$

$CV_B = 21.82\%$

Comparing Consistency

The lower the CV, the more consistent the vaccine performance. Based on the calculations:

- Vaccine A has a CV of 11.11%, which indicates a relatively stable performance. -
Vaccine B has a CV of 21.82%, which shows greater variability in its effectiveness.

Since Vaccine A has a lower CV than Vaccine B, it is more consistent in its disease-reduction effect.

Effect of Reducing the Standard Deviation of Vaccine B

Now, we consider the scenario where the standard deviation of Vaccine B is reduced from 12% to 7%. The new CV for Vaccine B is:

$CV_{B\_new} = (7 / 55) \times 100$
$CV_{B\_new} = (0.1273) \times 100$
$CV_{B\_new} = 12.73\%$

Comparing the Updated CV of Vaccine B with Vaccine A

With the reduced standard deviation, Vaccine B's CV decreases from 21.82% to 12.73%, making it more consistent. However, Vaccine A still has a slightly lower CV (11.11%) compared to the updated Vaccine B (12.73%), indicating that Vaccine A remains the more consistent option.

4. A researcher is studying the average blood pressure of a population. The population has a mean blood pressure of 90 mmHg and a standard deviation of 18 mmHg. If the researcher takes a random sample of 36 individuals and calculates the sample mean blood pressure, explain how the Central Limit Theorem applies. What will be the mean and standard deviation of the sampling distribution of the sample mean?
The Central Limit Theorem (CLT) states that if we take sufficiently large random samples from a population, the distribution of the sample mean will be approximately normal, regardless of the shape of the original population distribution. This holds true as long as the sample size is sufficiently large (typically $n \geq 30$ is considered adequate).

In this case, the researcher is taking a random sample of 36 individuals, which is a sufficiently large sample. As a result, the sampling distribution of the sample mean will be approximately

normal, even if the original population distribution is skewed or non-normal.

Mean of the Sampling Distribution

According to the CLT, the mean of the sampling distribution of the sample mean (denoted as $\mu_{\bar{x}}$) is equal to the population mean (μ). This means that the expected value of the sample mean is the same as the population mean:

$\mu_{\bar{x}} = \mu = 90$ mmHg

Standard Deviation of the Sampling Distribution (Standard Error)

The standard deviation of the sample mean, also known as the standard error of the mean (SEM), is calculated using the formula:

$\sigma_{\bar{x}} = \sigma / \sqrt{n}$

where:
- $\sigma = 18$ mmHg (population standard deviation)
- $n = 36$ (sample size)

Substituting the values:

$\sigma_{\bar{x}} = 18 / \sqrt{36}$
$\sigma_{\bar{x}} = 18 / 6$
$\sigma_{\bar{x}} = 3$ mmHg

This means that the sample means will typically vary by about 3 mmHg from the true population mean.

Conclusion

By applying the Central Limit Theorem, we conclude that the sampling distribution of the sample mean:

- Follows an approximately normal distribution
- Has a mean of 90 mmHg
- Has a standard deviation (standard error) of 3 mmHg

Thus, if the researcher repeatedly takes samples of size 36 and calculates the sample mean each time, those sample means will form a normal distribution centered at 90 mmHg with a spread of 3 mmHg.

5. In a clinical trial evaluating the effectiveness of a new drug, the distribution of patients' symptom reduction times is positively skewed. The 25th percentile is 2 weeks, and the 75th

percentile is 8 weeks. How would you interpret these percentiles in the context of the treatment's effectiveness? What does the positive skewness of the distribution imply about the variability of symptom reduction times? Additionally, suggest a statistical measure to summarize the typical time for symptom reduction in this case.

- **25th percentile (Q1) = 2 weeks**: 25% of patients see symptom reduction within **2 weeks**.
- **75th percentile (Q3) = 8 weeks**: 75% of patients see symptom reduction within **8 weeks**.

Since the distribution is **positively skewed**:

- More patients take **longer than the median** to see results.
- There is a **long tail on the right**, indicating some patients experience **much longer** symptom reduction times.

Alternative Measure for Typical Reduction Time:

- **Median**: More appropriate than the mean in skewed distributions, as it better represents the central tendency.
- **Interquartile Range (IQR)**: Measures the spread of the middle 50% of patients.

6. In a clinical trial evaluating a new vaccine, patients are classified into two groups: those with a specific biomarker (Group X) and those without the biomarker (Group Y). The probability of a patient in Group X responding positively to the vaccine is 0.9, while in Group Y, it is 0.5. If the overall probability of a patient responding positively to the vaccine is 0.7, what is the probability that a randomly selected patient has the biomarker, given that they responded positively to the vaccine?

Using **Bayes' Theorem**:

$P(X \mid R) = (P(R \mid X) * P(X)) / P(R)$

Given:
- $P(R \mid X) = 0.9$ (Probability of a positive response given biomarker X)

- $P(R \mid Y) = 0.5$

- $P(R) = 0.7$ (Overall probability of a positive response)

$P(R) = P(R \mid X) * P(X) + P(R \mid Y) * P(Y)$

Let $P(X) = p$ and $P(Y) = 1 - p$. Then:

$0.7 = (0.9 * p) + (0.5 * (1 - p))$

Solving for p: $p = 0.50$ (50%)

In a clinical trial evaluating a new vaccine, patients are classified into two groups:

- Group X: Patients who have a specific biomarker.
- Group Y: Patients who do not have the biomarker.

The probability of a patient in Group X responding positively to the vaccine is 0.9, while in Group Y, it is 0.5. The overall probability of a patient responding positively to the vaccine is 0.7. We aim to determine the probability that a randomly selected patient has the biomarker, given that they responded positively to the vaccine.