# Quiz 1- Rubrics

Course: Biostatistics
Instructor: Dr. Gaurav Ahuja
Duration: 60 minutes

**Question 1: Explain how the presence of an extreme outlier can impact the interpretation of centrality measures in a dataset, and discuss which centrality measure might be more robust or less influenced by such outliers. Also, discuss graphically a potent means to detect outliers.**

**Answer:** The presence of an extreme outlier can heavily skew centrality measures such as the mean (average), making them less representative of the typical data. Outliers disproportionately influence the mean, pulling it towards extreme values. In contrast, the median, which represents the middle value of the dataset, is less affected by outliers. A boxplot is a potent graphical method to detect outliers. Outliers are visualized as individual points beyond the whiskers of the boxplot, providing a clear indication of their presence in the dataset.

**Explanation with reasonings**

The presence of an extreme outlier can significantly impact the interpretation of centrality measures in a dataset by skewing the results and distorting the overall distribution. Centrality measures such as the mean (average) and median are particularly susceptible to the influence of outliers.

- **Mean (Average):** The mean is highly influenced by outliers because it incorporates all data points equally, including extreme values. When an outlier is present, it can pull the mean towards itself, leading to a misleading representation of the central tendency of the data. As a result, the mean may not accurately reflect the typical value in the dataset.

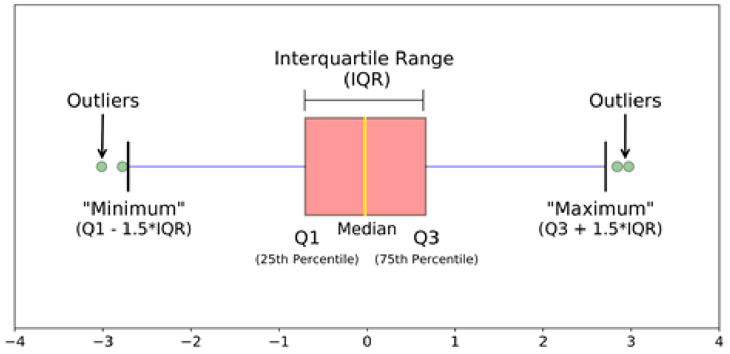
- **Median:** Unlike the mean, the median is less affected by outliers because it is determined by the middle value of the dataset when arranged in ascending or descending order. Since the median only considers the position of values rather than their magnitude, extreme outliers have less impact on its calculation. Therefore, the median can be a more robust measure of centrality in the presence of outliers.

A potent means to detect outliers graphically is by using boxplots or scatterplots:

- **Boxplot**: A boxplot visually displays the distribution of a dataset, highlighting the median, quartiles, and potential outliers. Outliers are typically represented as individual points outside the whiskers of the boxplot, which extend to a certain multiple of the interquartile range (IQR) from the first and third quartiles. Observations beyond these whiskers are considered potential outliers and warrant further investigation.

- **Scatterplot:** In a scatterplot, outliers can be identified as data points that deviate significantly from the general pattern or trend of the data. By visualizing the relationship between variables, scatterplots allow for the identification of any data points that lie far away from the bulk of the data. Outliers in a scatterplot may indicate measurement errors, data entry mistakes, or genuine extreme values that must be examined closely.

Overall, while both mean and median are common measures of centrality, the median is often more robust in the presence of outliers. However, it's essential to consider the data's context and the analysis's specific goals when choosing the most appropriate centrality measure. Additionally, graphical methods such as boxplots and scatterplots provide effective tools for detecting outliers and understanding their impact on the dataset.



**Question 2: Consider a box plot representing the distribution of exam scores for classes A and B. If Class A has a longer interquartile range (IQR) than Class B, can you infer with certainty that Class A has a more diverse range of scores? Explain the potential limitations or alternative explanations that should be considered when interpreting the IQR differences between the two classes.**

**Answer**: No, a longer interquartile range (IQR) in Class A compared to Class B does not necessarily indicate that Class A has a more diverse range of scores. Alternative explanations include differences in sample size, the presence of outliers, or variations in the distribution shape within each class. Therefore, caution should be exercised when interpreting IQR differences, as they may not always reflect true differences in the diversity of score ranges between classes.

**Explanation and reasoning :**

No, you cannot infer with certainty that Class A has a more diverse range of scores solely based on the interquartile range (IQR) differences between Class A and Class B. While a longer IQR can suggest a wider spread of scores within a dataset, it does not necessarily indicate greater diversity or variability in the scores.

There are several potential limitations and alternative explanations that should be considered when interpreting the IQR differences between the two classes:

- **Sample Size**: The size of the classes can influence the variability observed in the scores. A larger sample size generally leads to a more accurate representation of the underlying population distribution. If Class A has a significantly larger sample size than Class B, it might exhibit a wider range of scores and, consequently a longer IQR.

- **Homogeneity within Classes:** The composition of students within each class can impact the variability of scores. If Class A is composed of students with a wider range of abilities or backgrounds compared to Class B, it could lead to a broader spread of scores within Class A, resulting in a longer IQR.

- **Outliers**: Outliers, which are extreme values in the dataset, can affect the IQR calculation and skew the variability interpretation. If one class has more outliers than the other, it could contribute to differences in the IQR between the two classes, even if the majority of scores within each class are relatively similar.

- **Measurement Scales:** The scales on which the exam scores are measured can also influence the interpretation of variability. For example, if Class A's exam scores are measured on a wider scale (e.g., 0-100) compared to Class B's scores (e.g., 0-50), Class A might naturally exhibit a broader range of scores and a longer IQR, even if the underlying variability in performance is similar.

- **Sampling Bias:** If there are systematic differences in how students are selected or assigned to Class A and Class B, it could introduce bias into the comparison of variability between the two classes.

Therefore, while differences in IQR between Class A and Class B may indicate potential differences in the spread of exam scores, it is essential to consider these limitations and alternative explanations before making any definitive conclusions about the diversity of score ranges between the two classes. Additional analyses and contextual information may be necessary to provide a more comprehensive understanding of the variability in exam performance.

**Question 3: In a comparative analysis of two datasets, Dataset X and Dataset Y, both with the same mean and standard deviation, it is observed that Dataset X has a wider range of values and a higher variance. Explain this apparent contradiction, discussing the factors influencing variance and standard deviation, and elaborate on how such differences can emerge even when central tendencies and standard deviations are identical.**

**Answer**: The apparent contradiction arises because variance and range measure different aspects of data variability. While standard deviation is a measure of the average distance of data points from the mean, the range reflects the spread between the maximum and minimum values. Even when mean and standard deviation are identical between datasets, differences in the range and variance can occur due to variations in the distribution shape, presence of outliers, or differences in sample size. For example, Dataset X may have a wider spread of values or more extreme outliers, resulting in a higher variance and wider range despite having the same mean and standard deviation as Dataset Y.

**Explanation with reasonings**

In a comparative analysis where Dataset X and Dataset Y have the same mean and standard deviation but differ in variance and range, the apparent contradiction can be explained by understanding the factors influencing variance and standard deviation.

***Factors Influencing Variance and Standard Deviation:***

- **Sample Size:** Larger sample sizes tend to produce more reliable estimates of variability, resulting in lower variance and standard deviation. If Dataset X has a larger sample size than Dataset Y, it could lead to a narrower spread of values and lower variability in Dataset Y despite having the same mean and standard deviation.

- **Distribution Shape:** The shape of the distribution can affect the spread of values and variability. For example, datasets with more extreme values or a flatter distribution tend to have higher variance and wider ranges, even if the mean and standard deviation are the same.

- **Outliers:** Outliers, or extreme values, can disproportionately affect variance and range, leading to wider spreads and higher variability. If Dataset X contains more outliers or extreme values compared to Dataset Y, it could result in a wider range of values and higher variance in Dataset X, even with identical means and standard deviations.

**How Differences Can Emerge:**

Even when central tendencies (mean) and standard deviations are identical, differences in variance and range can emerge due to the factors mentioned above. For example:

- Dataset X may have a larger sample size than Dataset Y, resulting in a narrower spread of values and lower variability in Dataset Y despite having the same mean and standard deviation.
- Dataset X may have a distribution with more extreme values or a flatter shape, leading to higher variability and a wider range of values compared to Dataset Y.
- Dataset X may contain more outliers or extreme values, contributing to a wider spread and higher variance, even with identical means and standard deviations.

In summary, differences in variance and range between datasets with identical means and standard deviations can arise due to variations in sample size, distribution shape, and the presence of outliers. These differences highlight the importance of considering additional measures of variability beyond just the mean and standard deviation when comparing datasets.

**Question 4: Consider a discrete random variable X with a probability mass function (PMF) P(X). Which of the following statements is true?**
**a) The PMF P(X) must always sum to 1 for all possible values of X.**
**b) The PMF P(X) represents the range of possible values that X can take.**
**c) The PMF P(X) can take negative values for certain outcomes.**
**d) The PMF P(X) is only defined for continuous random variables.**
Answer: a)  The PMF P(X) must always sum to 1 for all possible values of X.

**Question 5: In a clinical trial assessing the efficacy of a new drug, patients are categorized into two groups: those with a specific genetic marker (Group A) and those without the marker (Group B). The probability of a patient responding positively to the drug in Group A is 0.8, while in Group B, it is 0.6. If the overall probability of a patient responding positively to the drug is 0.7, what is the probability of a randomly selected patient having the genetic marker, given that they responded positively to the drug?**

Answer

**Answer : 5** Probability of randomly selected patients having genetic marker (Group A) given that they responded +vely to drug.

A → patient belong to Group A (has genetic marker)

B → patient belong to Group B (doesnot have genetic marker)

R → event that responds positively to the drug

$$P(R|A) = 0.8$$

$$P(R|B) = 0.6$$

$$P(R) = 0.7$$

$$P(A|R) = \frac{P(R|A).P(A)}{P(R)}$$

$$P(A) + P(B) = 1$$

$$P(R) = P(R|A).P(A) + P(R|B).P(B)$$

$$0.7 = (0.8 \cdot P(A)) + (0.6 \cdot (1 - P(A)))$$

$$0.7 = 0.8P(A) + 0.6 - 0.6P(A)$$

$$0.1 = 0.2\,P(A)$$

$$P(A) = 0.1/0.2 = 0.5$$

$P(A) = 0.5$ , $P(R|A) = 0.8$, $P(R) = 0.7$

Use to find $P(A|R)$

$$P(A|R) = \frac{0.8 \times 0.5}{0.7} = \frac{0.4}{0.7} \simeq 0.5714$$

Probability of a randomly selected patient having the genetic marker, given they responded positively to the drug, is approximately 57.14 %.

**Question 6: In a diagnostic study, two different biomarkers (A and B) are evaluated for their performance in distinguishing between healthy and diseased individuals. The Area Under the Receiver Operating Characteristic curve (AUC-ROC) for Biomarker A is 0.85, while for Biomarker B, it is 0.75. Biomarker A has a sensitivity of 0.90, and Biomarker B has a specificity of 0.80. Discuss the implications of these findings and the potential trade-offs between sensitivity and specificity in the context of selecting the most appropriate biomarker for the given diagnostic task.**

**Answer:** Biomarker A has a higher AUC-ROC (0.85) than Biomarker B (0.75), indicating better overall performance distinguishing between healthy and diseased individuals. However, Biomarker A's higher sensitivity (0.90) may lead to more false positives, while Biomarker B's higher specificity (0.80) may result in more false negatives. The choice between them depends on the importance of minimizing false positives or false negatives in the specific diagnostic task.

**Explanation with reasoning**

The findings from the diagnostic study provide valuable information regarding the performance of Biomarkers A and B in distinguishing between healthy and diseased individuals. Here are the implications of these findings and the potential trade-offs between sensitivity and specificity:

- **AUC-ROC Values**: The AUC-ROC values indicate the overall discriminative ability of each biomarker. A higher AUC-ROC value suggests better overall performance in correctly classifying individuals as healthy or diseased. In this case, Biomarker A has a higher AUC-ROC value (0.85) compared to Biomarker B (0.75), indicating that Biomarker A is more effective in distinguishing between healthy and diseased individuals overall.

- **Sensitivity and Specificity**: Sensitivity refers to the ability of a biomarker to correctly identify individuals with the disease (true positive rate), while specificity refers to the ability to correctly identify individuals without the disease (true negative rate).

Biomarker A has a sensitivity of 0.90, indicating that it correctly identifies 90% of individuals with the disease.
Biomarker B has a specificity of 0.80, meaning it correctly identifies 80% of individuals without the disease.

- **Trade-offs between Sensitivity and Specificity**: There is often a trade-off between sensitivity and specificity. Increasing sensitivity may lead to a decrease in specificity and vice versa.

Biomarker A, with a higher sensitivity (0.90), is more likely to identify individuals with the disease correctly. However, this might come at the expense of specificity, meaning it may also incorrectly classify some healthy individuals as diseased (false positives).
Biomarker B, with a higher specificity (0.80), is better at correctly identifying individuals without the disease. However, it may miss some individuals who actually have the disease (false negatives), resulting in lower sensitivity.

- **Selecting the Most Appropriate Biomarker**: The choice of the most appropriate biomarker depends on the specific requirements of the diagnostic task and the consequences of false positives and false negatives.

If the primary concern is to minimize false negatives (missing individuals with the disease), Biomarker A, with its higher sensitivity, would be preferred.
If the primary concern is to minimize false positives (incorrectly diagnosing healthy individuals as diseased), Biomarker B, with its higher specificity, would be preferred.
Ultimately, the decision should consider the clinical context, the consequences of misclassification, and the relative importance of sensitivity and specificity for the specific diagnostic task at hand.

**Question 7: In a clinical trial, the distribution of recovery times for patients undergoing a new treatment follows a skewed distribution. The 25th percentile of recovery times is found to be 7 days, while the 75th percentile is 20 days. Explain how you would interpret**

**these quantiles in the context of patient recovery and discuss the potential implications of the observed skewness on the treatment's effectiveness. Additionally, propose a statistical measure to summarize the central tendency of recovery times in this scenario.**
**Answer:** The 25th percentile (7 days) suggests that 25% of patients recover within 7 days or less, indicating relatively quick recovery for some. The 75th percentile (20 days) indicates that 75% recover within 20 days, with a majority taking longer. The observed skewness implies variability in recovery times, possibly due to treatment response or patient factors. Skewed distributions may complicate assessing treatment effectiveness, making the median a more robust measure of central tendency to summarize recovery times.

**Explanation with reasoning**

In the context of patient recovery, the 25th and 75th percentiles of recovery times provide valuable information about the distribution of recovery durations among patients undergoing the new treatment.

The 25th percentile, which is 7 days in this case, indicates that 25% of patients recover within 7 days or less. A quarter of patients experience relatively quick recovery times with the treatment.

The 75th percentile, which is 20 days, means that 75% of patients recover within 20 days or less. This indicates that the majority of patients take longer to recover, with three-quarters of them seeing improvement within 20 days.

The observed skewness in the distribution of recovery times implies that the data is not symmetrically distributed. Skewed distributions typically have a long tail on one side, indicating that some patients may take significantly longer to recover than the majority. In this case, the skewness suggests that a subset of patients may experience prolonged recovery times, potentially indicating variability in treatment response or underlying patient characteristics.

The implications of skewness on the treatment's effectiveness are twofold:

- **Effectiveness Assessment:** Skewed distributions can complicate the assessment of treatment effectiveness because traditional measures of central tendency, such as the mean, may be influenced by extreme values in the tail of the distribution. Therefore, relying solely on the mean may not accurately reflect the typical recovery time experienced by most patients.

- **Tailored Treatment Approaches:** Understanding the skewness can help healthcare providers tailor treatment approaches based on patient characteristics. For example, patients who are likely to experience prolonged recovery times may benefit from additional interventions or closer monitoring during the treatment process.

A robust measure such as the median would be appropriate to summarize the central tendency of recovery times in this scenario while accounting for skewness. The median represents the middle value of the data when arranged in ascending order and is less affected by extreme values compared to the mean. Therefore, it provides a more robust estimate of the typical recovery time experienced by patients undergoing the new treatment.