

# Mid Semester Exam



INDRAPRASTHA INSTITUTE of  
INFORMATION TECHNOLOGY DELHI

**Course: Biostatistics | Instructor: Dr. Gaurav Ahuja | Duration: 60 minutes | Each question weights Five Marks**

**Attempt Only FIVE (5) Questions of your choice**

**Question 1:** Here is some information about the length of time a patient spent in an ICU at Apollo Hospital during the heatwave season of June and July.

- Shortest time 30 minutes | Longest time 5 hours | Lower quartile 90 minutes | Interquartile range 2 hours | Median time 2 hours 30 minutes

Draw a box plot to show this information.

**Answer:** five summary statistics: minimum, lower quartile (Q1), median (Q2), upper quartile (Q3), and maximum. Here's how we can plot it:

Minimum: 30 minutes

Lower Quartile (Q1): 90 minutes

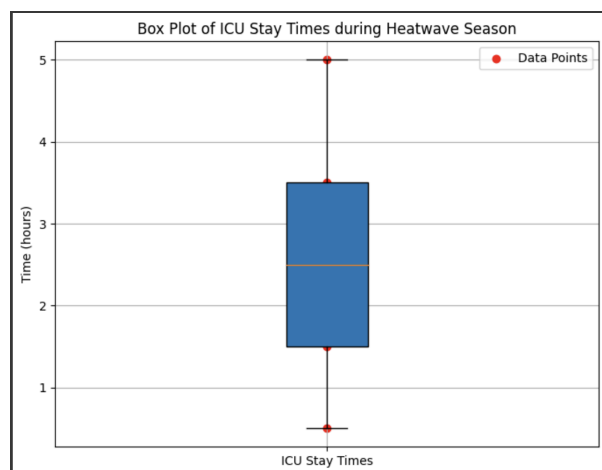
Median (Q2): 2 hours 30 minutes

Upper Quartile (Q3):  $Q1 + \text{Interquartile Range} = 90 \text{ minutes} + 2 \text{ hours} = 3 \text{ hours } 30 \text{ minutes}$

Maximum: 5 hours

Step 1: # Convert minutes to hours

Step 2 : Plot



**Question 2:** Suppose that 28 cancer deaths are noted among workers exposed to asbestos in a building materials plant from 1981 to 1985. Only 20.5 cancer deaths are expected from statewide mortality rates. Suppose that we want to know if there is a significant excess of cancer deaths among these workers. What is the null hypothesis? Is a one- or two-sided test appropriate here?

**Answer :** In this scenario, the null hypothesis ( $H_0$ ) would typically state that there is no significant difference between the observed number of cancer deaths among workers exposed to asbestos and the expected number of cancer deaths based on statewide mortality rates.

The null hypothesis ( $H_0$ ) can be stated as:

$H_0$ : There is no significant excess of cancer deaths among workers exposed to asbestos in the building materials plant compared to statewide mortality rates.

Regarding the type of test appropriate for this scenario, since we're interested in determining if there is a significant excess of cancer deaths (i.e., if the observed number is greater than expected), a one-sided test is appropriate. We're only interested in whether the observed number is significantly greater than expected, not if it's significantly different in either direction. Therefore, a one-sided test would be appropriate to test whether the observed number of cancer deaths among workers exposed to asbestos is significantly higher than expected.

**Question 3:** How do parametric and non-parametric tests differ in analyzing dataset characteristics? Please provide examples with the name of the statistical test under each case.

**Answer :** The main difference lies in the assumptions they make about the underlying distribution of the data.

### **Parametric Tests:**

Parametric tests assume that the data are drawn from a specific probability distribution, usually the normal distribution. These tests are more powerful when their assumptions are met but can be sensitive to violations of those assumptions.

Parametric tests are restricted to data that:

- 1) show a normal distribution
- 2) \* are independent of one another
- 3) \* are on the same continuous scale of measurement

Examples of parametric tests include:

**t-test:** Used to compare means between two groups. Assumptions include normality of data and homogeneity of variances.

**ANOVA (Analysis of Variance):** Used to compare means among three or more groups. Assumptions include normality of data and homogeneity of variances.

**Linear Regression:** Used to model the relationship between a dependent variable and one or more independent variables. Assumptions include linearity, normality, homoscedasticity, and independence of residuals.

## Non-parametric Tests:

Non-parametric tests make fewer assumptions about the underlying distribution of the data, making them more robust in certain situations where parametric assumptions are violated. They are based on ranks or other non-numerical aspects of the data.

Non-parametric tests are used on data that:

- 1) show an other-than normal distribution
- 2) are dependent or conditional on one another
- 3) in general, do not have a continuous scale of measurement

Examples of non-parametric tests include:

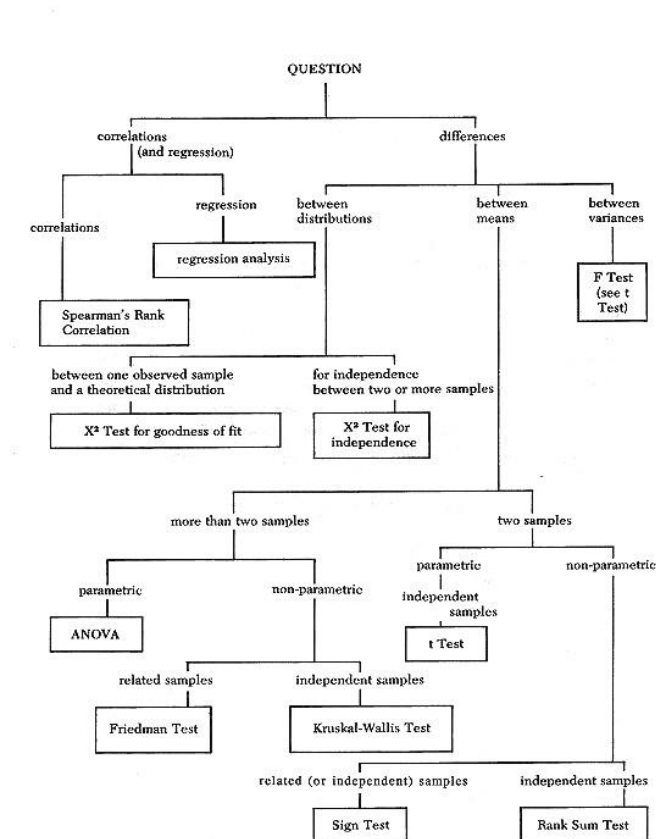
**Mann-Whitney U test (Wilcoxon rank-sum test):** Non-parametric alternative to the independent samples t-test for comparing means between two groups. It does not assume normality.

**Kruskal-Wallis test:** Non-parametric alternative to ANOVA for comparing means among three or more groups. It does not assume normality.

**Spearman's rank correlation coefficient:** Non-parametric alternative to Pearson correlation coefficient for assessing the strength and direction of association between two variables. It does not assume linearity.

The choice between parametric and non-parametric tests depends on the characteristics of the data and the assumptions that can be reasonably made about its distribution. If the data meet the assumptions of parametric tests, they tend to be more powerful. However, if these assumptions are violated, non-parametric tests provide a more robust alternative.

## References for test



**Question 4:** How does the Coefficient of Variation (CV) offer a nuanced perspective on data variability, and in what practical scenarios does its application provide more insightful interpretations compared to other measures of dispersion, such as standard deviation or range?

Answer : The Coefficient of Variation (CV) is a statistical measure used to assess the relative variability of a dataset, particularly in comparison to its mean. It is calculated as the ratio of the standard deviation to the mean, expressed as a percentage:

$$CV = \left( \frac{\text{Standard Deviation}}{\text{Mean}} \right) \times 100\%$$

The CV offers a perspective on data variability by standardizing the measure of dispersion relative to the mean.

## **1 Comparing Variability Across Different Scales:**

The CV allows for the comparison of variability across datasets with different scales and units. For instance, if you're comparing the variability of income levels in different countries or the variability of test scores in different subjects, using CV ensures that the comparison is meaningful regardless of the scale of the data.

## **2. Interpreting Variability Relative to the Mean:**

Unlike measures of dispersion such as standard deviation or range, which provide absolute measures of variability, CV provides a relative measure. This means that the variability is interpreted in the context of the mean. A high CV indicates high variability relative to the mean, whereas a low CV indicates low variability relative to the mean.

## **3. Standardizing Variability for Comparisons:**

When comparing the variability of different datasets, the CV can provide a standardized measure, facilitating easier comparison. For example, if you're comparing the variability of stock returns across different companies or the variability of rainfall across different regions, CV can help standardize these comparisons.

## **4. Assessing Stability in Processes:**

In quality control or process improvement scenarios, CV can be used to assess the stability of a process. A high CV may indicate that the process is unstable or unpredictable, whereas a low CV suggests that the process is consistent and predictable.

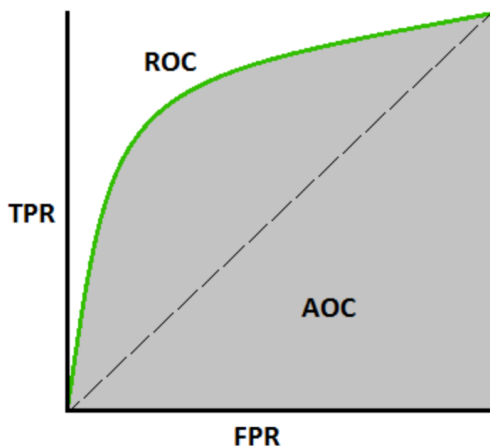
## 5. Identifying Outliers:

In some cases, outliers can significantly affect measures of dispersion such as standard deviation. The CV, being a relative measure, is less affected by outliers and may provide a more robust assessment of variability, particularly in skewed datasets.

In summary, the Coefficient of Variation offers a perspective on data variability by standardizing it relative to the mean. Its application can provide insightful interpretations in scenarios where comparing variability across different scales, interpreting variability relative to the mean, standardizing variability for comparisons, assessing stability in processes, and identifying outliers are important considerations.

**Question 5:** How does the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) serve as a comprehensive metric for evaluating the performance of diagnostic tests in biomedical research?

Answer : Here's how AUC-ROC serves as a comprehensive metric:



---

A receiver operating characteristic (ROC) curve is a plot of the sensitivity (on the y-axis) versus  $(1 - \text{specificity})$  (on the x-axis) of a screening test, where the different points on the curve correspond to different cutoff points used to designate test-positive.

---

## Measures Discrimination Ability:

AUC-ROC evaluates the ability of a diagnostic test to discriminate between two groups, typically diseased and non-diseased individuals. It plots the true positive rate (sensitivity) against the false positive rate ( $1 - \text{specificity}$ ) across various threshold values. A higher AUC value indicates better discrimination ability of the test, with a perfect test achieving an AUC of 1.0.

Model Performance check :

- An excellent model has AUC near to the 1 which means it has a good measure of separability.
- A poor model has AUC near to the 0 which means it has the worst measure of separability.
- In fact, it means it is reciprocating the result. It is predicting 0s as 1s and 1s as 0s.
- And when AUC is 0.5, it means the model has no class separation capacity whatsoever.

**Threshold Independence:** AUC-ROC is threshold-independent, meaning it evaluates the overall discriminatory power of the test across all possible cutoff values. This is crucial in biomedical research as different thresholds may be appropriate depending on the specific clinical context or trade-offs between sensitivity and specificity.

**Robustness to Class Imbalance:** AUC-ROC is robust to class imbalance, which is common in biomedical datasets where the number of diseased and non-diseased individuals may vary significantly. Unlike metrics like accuracy, which can be misleading in imbalanced datasets, AUC-ROC provides a reliable assessment of model performance.

**Interpretability:** The interpretation of AUC-ROC is intuitive: a higher AUC value indicates better discrimination ability, while a value of 0.5 suggests performance no better than random chance..

**Comparison Across Models:** AUC-ROC enables direct comparison of the performance of different diagnostic tests or predictive models. Researchers can evaluate and select the most effective model based on its AUC value, providing valuable insights for clinical decision-making.

**Assessment of Model Calibration:** In addition to discrimination ability, AUC-ROC can also provide insights into model calibration, i.e., the agreement between predicted probabilities and observed outcomes. Deviations from the diagonal line (the line of no-discrimination) on the ROC curve may indicate issues with model calibration.

**Insensitivity to Prevalence:** AUC-ROC is not affected by the prevalence of the disease in the population, making it applicable across different settings and populations without requiring adjustments.

Overall, AUC-ROC serves as a comprehensive metric for evaluating the performance of diagnostic tests in biomedical research by assessing discrimination ability, threshold independence, robustness to class imbalance, interpretability, comparability across models, assessment of model calibration, and insensitivity to disease prevalence.

**Question 6:** List all factors that influence the sample size estimation using the power analysis. Please provide examples by drawing the “Null” and “Alternative” hypotheses for each of the factors.

Answer : Factors influencing sample size estimation using power analysis include:

**1 Effect Size (ES):**

*Null Hypothesis ( $H_0$ ):* There is no difference between the groups (e.g., mean difference = 0).

*Alternative Hypothesis ( $H_1$ ):* There is a difference between the groups (e.g., mean difference  $\neq 0$ ).



*Example:* In a study comparing the effectiveness of two treatments, the null hypothesis might be that there is no difference in the mean outcomes between the treatments, while the alternative hypothesis would be that there is a difference.

## 2. Significance Level ( $\alpha$ ):

*Null Hypothesis ( $H_0$ ):* The true effect size is equal to or smaller than the defined threshold.

*Alternative Hypothesis ( $H_1$ ):* The true effect size is larger than the defined threshold.

*Example:* If  $\alpha = 0.05$ , the null hypothesis would be that the effect size is equal to or smaller than what would occur by chance alone, while the alternative hypothesis would be that the effect size is larger than expected by chance.

## 3 . Statistical Power ( $1-\beta$ ):

*Null Hypothesis ( $H_0$ ):* The true effect size is equal to the minimum detectable effect size.

*Alternative Hypothesis ( $H_1$ ):* The true effect size is larger than the minimum detectable effect size.

*Example:* If the desired power is 0.80, the null hypothesis would be that the effect size is equal to the minimum detectable effect size necessary to achieve 80% power, while the alternative hypothesis would be that the true effect size is larger than this minimum.

## 4.Variability of the Outcome (Standard Deviation):

*Null Hypothesis ( $H_0$ ):* The standard deviation of the outcome variable is equal to a specified value.

*Alternative Hypothesis ( $H_1$ ):* The standard deviation of the outcome variable is different from the specified value.

*Example:* In a study assessing the variability of blood pressure measurements, the null hypothesis might be that the standard deviation is equal to a certain value, while the alternative hypothesis would be that it differs.

## 5. Desired Power:

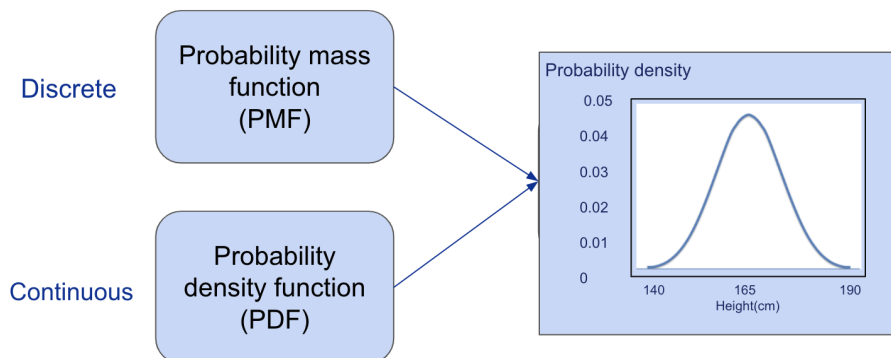
*Null Hypothesis ( $H_0$ ):* The desired statistical power is equal to or smaller than the specified threshold.

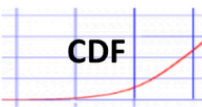
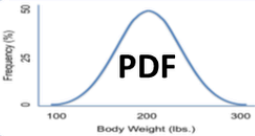
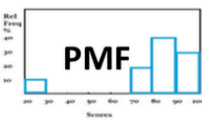
*Alternative Hypothesis ( $H_1$ ):* The desired statistical power is larger than the specified threshold.

*Example:* If the researcher aims for 90% power to detect an effect, the null hypothesis would be that the desired power is equal to or smaller than 90%, while the alternative hypothesis would be that it exceeds 90%.

These factors interact to determine the necessary sample size for a study to achieve adequate power to detect the hypothesized effect.

**Question 7:** Explain Probability Mass Function, Probability Distribution Function, and Cumulative Distribution Function.



			
	<b>Cumulative Density Function</b>	<b>Probability Density Function</b>	<b>Probability Mass Function</b>
Purpose	Cumulative probability associated with a function.	Probabilities for continuous random variables.	Probabilities for discrete random variables.
Example	Cumulative value from negative infinity up to a random variable X (i.e. $x < 10$ )	Probability of a range of outcomes (e.g. $X = 5$ to $6$ )	Probability of a certain outcome (e.g. $X = 6$ )
Properties	Integral of the PDF. A CDF has [2]: a/ Left limit = 0, right limit = 1 b/ Nondecreasing c/ Right continuous (defined up to a point) [3].	Derivative of the CDF. A PDF satisfies the following [4]: a/ It is positive everywhere b/ $AUC = 1$ c/ Total probability = integral of $f(x)$	Satisfies the following[4]: a/ It is positive everywhere b/ $AUC = 1$ c/ Total probability = summations of individual probabilities.

## Probability Mass Function (PMF):

- A Probability Mass Function (PMF) is a function that gives the probability that a discrete random variable is exactly equal to a certain value.
- It maps each possible value of the random variable to the probability of that value occurring.
- For a discrete random variable  $X$ , the PMF is denoted by  $P(X = x)$ , where  $x$  represents the possible values that  $X$  can take.
- The PMF satisfies two properties:
  - Each probability value must be between 0 and 1.
  - The sum of probabilities over all possible values must equal 1.
- Example: Consider rolling a fair six-sided die. The PMF for this scenario would assign a probability of  $1/6$  to each possible outcome (1, 2, 3, 4, 5, or 6), as each outcome has an equal chance of occurring.

## Probability Density Function (PDF):

- A Probability Density Function (PDF) is used to specify the probability of a continuous random variable falling within a particular range of values.
- Unlike the PMF, which is for discrete random variables, the PDF is for continuous random variables.
- The PDF does not directly give the probability of a specific outcome but rather the probability density at a given point.
- The area under the PDF curve over a given interval represents the probability of the random variable falling within that interval.
- The integral of the PDF over its entire range equals 1.
- Example: The PDF of a standard normal distribution (bell-shaped curve) represents the likelihood of observing different values of a standard normally distributed variable.

## Cumulative Distribution Function (CDF):

- A Cumulative Distribution Function (CDF) gives the probability that a random variable  $X$  takes on a value less than or equal to a given value  $x$ .
- It provides a cumulative view of the probability distribution, accumulating probabilities as we move along the range of possible values.
- For a discrete random variable, the CDF is obtained by summing up the probabilities of all values less than or equal to  $x$ . For a continuous random variable, it's obtained by integrating the PDF up to  $x$ .
- The CDF ranges from 0 to 1, inclusive.
- Example: In a binomial distribution, the CDF would give the probability of observing  $k$  or fewer successes in  $n$  independent Bernoulli trials, where  $k$  is a non-negative integer less than or equal to  $n$ .

**Question 8:** Explain Conditional probability with a biomedical example.

**Answer :** Conditional probability is the probability of an event occurring given that another event has already occurred. It's represented as  $P(A|B)$ , where  $A$  is the event of interest and  $B$  is the event that has already occurred.

The quantity  $Pr(A \cap B)/Pr(A)$  is defined as the **conditional probability of  $B$  given  $A$** , which is written  $Pr(B|A)$ .

- (1) If  $A$  and  $B$  are independent events, then  $Pr(B|A) = Pr(B) = Pr(B|\bar{A})$ .
- (2) If two events  $A, B$  are dependent, then  $Pr(B|A) \neq Pr(B) \neq Pr(B|\bar{A})$  and  $Pr(A \cap B) \neq Pr(A) \times Pr(B)$ .

### *Any example related to biomedical Studies*

Consider a diagnostic test for a particular disease, such as HIV. Let's define the events as follows:

Event A: A person tests positive for HIV.

Event B: The person belongs to a high-risk group, such as intravenous drug users.

Now, the conditional probability  $P(A|B)$  would represent the probability that a person tests positive for HIV given that they belong to the high-risk group.

Suppose data shows that among individuals belonging to the high-risk group, 10% test positive for HIV. This would mean:

$$P(A|B) = 0.10$$

This implies that given someone is from the high-risk group (event B), there's a 10% chance they will test positive for HIV (event A).

Conditional probability becomes particularly important in interpreting diagnostic test results in clinical practice. Knowing the conditional probability of a positive test result given certain patient characteristics, such as belonging to a high-risk group, can help healthcare providers better interpret test results and make informed decisions regarding further testing or treatment. Conditional probability in a biomedical context helps quantify the likelihood of certain events, such as a positive test result, given specific conditions or characteristics of individuals, aiding in medical decision-making and risk assessment.

# Mid Semester Exam



INDRAPRASTHA INSTITUTE *of*  
INFORMATION TECHNOLOGY **DELHI**