1. **Which of the following statements about pdf and cdf?**
2. **The cumulative distribution function is the probability that a random variable X, will take a value exactly equal to x.**
3. **The probability density function is the probability that a random variable X, will take a value equal to or less than the specified value.**
4. **The probability density function is the probability that a random variable X, will take a value exactly equal to x.**
5. **The cumulative distribution function is the probability that a random variable X, will take a value equal to or less than the specified value.**

The correct answer iv.

1. **What are bins in histogram? How will changing the number of bins modify the plot?**

The bins in histogram represent the number of points or value falls in the given interval or range. Here the bins represent the continuous interval.

Changing the number of bins will affect the plot in different ways. If the bin is very small or less, then it will be difficult to refine the data points and get better information. If there are a huge number of bins, then many of the bins will not be occupied.

1. **How do you interpret box plot results? What do different boundaries of boxplot signify? What is the significance of whiskers in boxplot?**

The boxplot interprets the result in the for quantiles like. There are three boundaries in the box plot which represents the three percentiles i.e., 25 percentile, 50 percentile, 75 percentile. The lower boundary is 25th percentile, middle boundary is 50th percentile and the upper boundary is 75th percentile.Chart, box and whisker chart

Description automatically generated

Consider the above plot survival vs Age. Here the lower boundary which is 25th percentile represents that there are 25% of people who survived who age around 42 years. Similarly, we also calculate the 50th and 75th percentile.

Whiskers are used to signify the variability of data outside the box or quantiles. It also helps us to decide the maximum and minimum data points. Whiskers also help to find out the outliers.

1. **What are different ways in which we can handle NaN values for**

* **Categorical variable.**
* **Numerical variable.**
* For Categorical Variable:
* The first method to handle the NaN values is to delete the null records. We apply this technique only when the dataset is huge. If the dataset is small, then it is not recommended to delete the values as it may affect the model performance.
* Second method is to check which category has occurred most of the time in the data and then replace the null values with that category.
* We can apply a classifier algorithm to predict the missing value.
* For Numerical Variable:
* For Numerical Variable:
* To handle the missing value for numerical data we can use the technique like mean, median, mode.
* For numerical data also if the dataset is huge then we can drop that row.
* The other method is to fill the missing values using mean, median or mode.