

STATISTICS WORKSHEET- 6

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

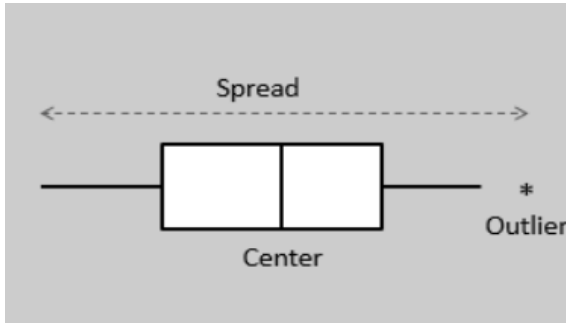
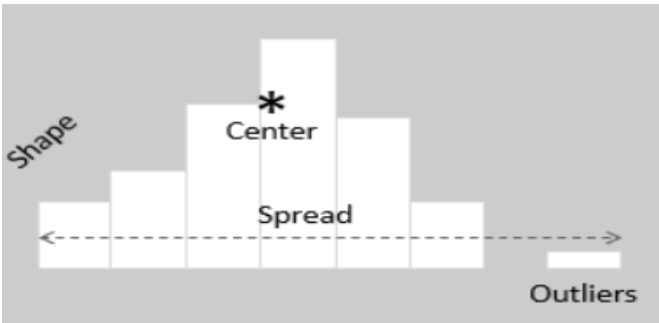
1. Which of the following can be considered as random variable?
 - a) The outcome from the roll of a die
 - b) The outcome of flip of a coin
 - c) The outcome of exam
 - d) All of the mentioned
 2. Which of the following random variable that take on only a countable number of possibilities?
 - a) Discrete
 - b) Non-Discrete
 - c) Continuous
 - d) All of the mentioned
 3. Which of the following function is associated with a continuous random variable?
 - a) pdf
 - b) pmv
 - c) pmf
 - d) all of the mentioned
 4. The expected value or _____ of a random variable is the center of its distribution.
 - a) mode
 - b) median
 - c) mean
 - d) bayesian inference
 5. Which of the following of a random variable is not a measure of spread?
 - a) variance
 - b) standard deviation
 - c) empirical mean
 - d) all of the mentioned
 6. The _____ of the Chi-squared distribution is twice the degrees of freedom.
 - a) variance
 - b) standard deviation
 - c) mode
 - d) none of the mentioned
 7. The beta distribution is the default prior for parameters between _____.
 - a) 0 and 10
 - b) 1 and 2
 - c) 0 and 1
 - d) None of the mentioned
 8. Which of the following tool is used for constructing confidence intervals and calculating standard errors for difficult statistics?
 - a) baggyer
 - b) bootstrap
 - c) jackknife
 - d) none of the mentioned
-

9. Data that summarize all observations in a category are called _____ data.
- frequency
 - summarized**
 - raw
 - none of the mentioned

Q10 and Q15 are subjective answer type questions, Answer them in your own words briefly.

10. What is the difference between a boxplot and histogram?

Ans: The difference between a boxplot and histogram is given below:

Box plot	Histogram
It is easier to read minimum value, median, outliers, quantiles, and maximum value.	It divides the numeric data into uniform intervals and displays the number of data values falling within each bin. It gives only the count.
It is a good way to summarize large amounts of data.	They group data into a small chunk. They are useful for summarizing numeric data in that they show the rough distribution of values.
Use the boxplot to display the range and distribution of data.	Use the histogram to display the number of values within an interval.
box plots allow you to compare multiple data sets better than histograms as they are less detailed and take up less space.	histograms are better in determining the underlying distribution of the data.
	

11. How to select metrics?

Ans: First of all, metrics which we optimise tweaking a model and performance evaluation metrics in machine learning are not typically the same.

- For performance evaluation, initial business metrics can be used.
- Based on prerequisites, we need to understand what kind of problems we are trying to solve. Here is a list of some common problems in machine learning:

Classification: This algorithm will predict data type from defined data arrays. For example, it may respond with yes/no/not sure.

Regression: The algorithm will predict some values. For example, weather forecast for tomorrow.

Ranking: The model will predict an order of items. For example, we have a student group and need to rank all the students depending on their height from the tallest to the shortest.

The number of instances per class: A lot depends on the number of instances per class. One needs to check if it's a class imbalance dataset (some classes having much more data than others) or a balanced dataset i.e. classes roughly having the same number of instances.

The Business use-case to solve: Understanding the business needs whether to give every class equal importance or give more importance to some classes than rest. This also gives the direction around the right metric to use.

12. How do you assess the statistical significance of an insight?

Ans: Statistical significance is often calculated with statistical hypothesis testing, which tests the validity of a hypothesis by figuring out the probability that your results have happened by chance. To assess statistical significance, you would use hypothesis testing. The null hypothesis and alternate hypothesis would be stated first. Second, you'd calculate the p-value, which is the likelihood of getting the test's observed findings if the null hypothesis is true. Finally, you would select the threshold of significance (alpha) and reject the null hypothesis if the p-value is smaller than the alpha — in other words, the result is statistically significant.

13. Give examples of data that does not have a Gaussian distribution, nor log-normal.

Ans: Any type of categorical data won't have a gaussian distribution or lognormal distribution. Exponential distributions - eg. the amount of time that a car battery lasts or the amount of time until an earthquake occurs, distributions of income; distributions of house prices; distributions of bets placed on a sporting event.

14. Give an example where the median is a better measure than the mean.

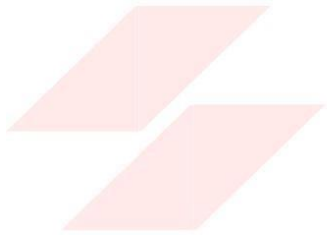
Ans: Income is the classic example of when to use the median instead of the mean because its distribution tends to be skewed.

15. What is the Likelihood?

Ans: The term Likelihood refers to the process of determining the best data distribution given a specific situation in the data. When we calculate the likelihood, you're attempting to determine whether the parameters in a model can be trusted based on the sample data you have observed. Likelihood function is a fundamental concept in statistical inference. It indicates how likely a particular population is to produce an observed sample. the likelihood is a quantity proportional to the probability that, from a population having a particular value of θ , a sample having the observed value x_0 , should be obtained. Likelihood, being the outcome of a likelihood function thus defined, describes the plausibility, under a certain statistical model (the null hypothesis in hypothesis testing), of a certain parameter value after observing a particular outcome. Formally: $L(\theta; x_0) \propto f(x_0; \theta), \forall \theta \in \Theta$.

Likelihood is central to parametric statistical inference. The likelihood is a basis for the likelihood ratio test: a uniformly most powerful test for comparing two point hypotheses. It is also the basis for the maximum likelihood estimate.

In practice one often calculates the natural logarithm of the likelihood function (log-likelihood) as being more convenient (easier to differentiate). The fact that a logarithm is strictly increasing is useful when calculating maximum likelihood: log-likelihood reaches the maximum at the same point as the likelihood.



FLIP ROBO