

STATISTICS WORKSHEET- 6

Q1. 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

→D

Q2. 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

→A

Q3. Which of the following function is associated with a continuous random variable?

a) pdf b) pmv c) pmf d) all of the mentioned

→A

Q4. The expected value or _____ of a random variable is the center of its distribution.

a) mode b) median c) mean d) bayesian inference

→C

Q5. 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

→C

Q6. The _____ of the Chi-squared distribution is twice the degrees of freedom.

a) variance b) standard deviation c) mode d) none of the mentioned

→A

Q7. 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

→C

Q8. 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

→B

Q9. Data that summarize all observations in a category are called _____ data.

a) frequency b) summarized c) raw d) none of the mentioned

→B

Q10. What is the difference between a boxplot and histogram?

→ In the univariate case, box-plots do provide some information that the histogram does not. That is, it typically provides the median, 25th and 75th percentile, min/max that is not an outlier and explicitly separates the points that are considered outliers. In box plots we can better visualize the outliers with respect to the inferential statistics of the feature. Histograms only give a measure of density of the feature values

Q11. How to select metrics?

→ The metrics are chosen on terms of nature of the problem. Classification, Regression and unsupervised learning all have different metrics. Also based on the problem given to decide if we want specificity or sensitivity also where and how the results would be applied in real world. To know distribution of the target variable.

Q12. How do you assess the statistical significance of an insight?

→ Statistical significance can be accessed using hypothesis testing: – Stating a null hypothesis which is usually the opposite of what we wish to test. We choose a suitable statistical test and statistics used to reject the null hypothesis and choose a critical region for the statistics to lie in that is extreme enough for the null hypothesis to be rejected (p-value)

then calculate the observed test statistics from the data and check whether it lies in the critical region. There are multiple tests we performed based on the nature of the problem and features of our dataset.

Q13. Give examples of data that does not have a Gaussian distribution, nor log-normal.

→ Types of distribution that are non-Gaussian or non-log normal are the skewed distributions, discrete distributions and binomial distribution.

Q14. Give an example where the median is a better measure than the mean.

→

When there are way too many outliers, in those cases if we use mean. We will be way off as mean is drastically affected by outliers. Thus, in such cases it is preferable to use median as the metric for central tendency than mean. Another time when we usually prefer the median over the mean (or mode) is when our data is skewed

Q15. What is the Likelihood?

→ A likelihood function takes the data set as a given, and represents the likeliness of different parameters for your distribution. The Likelihood function gives us an idea of how well the data summarizes these parameters.

