Statistics– WORKSHEET Solutions

1. a
2. a
3. b
4. d
5. c
6. b
7. b
8. a
9. c

10. Normal distribution is when the pdf of a function looks like a bell curve. An Ideal Normal distribution has same mean, median and mode. Data scientist perform various transformations of the data just so they can get the data in a normal distribution. Normal distribution is also known as the Gaussian distribution. Normal distribution is the ideal distribution a random variable should follow to get best results in analytics.

11. Handling of missing data is a subjective process performed in guidance of the domain expert. In general, if we have less than 5 percent of data missing we could drop those. WE can use various imputations techniques to impute data. We can use mean, median or mode imputations. We can use ffill and bfill (pandas functions) to perform imputation. In some cases we could use random variable imputation (a function from feature-engine library,

python) as it gives better results if the data is skewed. We can also perform end tail imputations, where the data is replace with the tail values. Lastly we can also use Bayesian imputation is its suitable to the problem

12. A/B testing is a popular way to test your products. A/B testing is an experiment Show your current experience to half your visitors and offer an alternative experience to the other half; observe differences in performance, then either continue with the old one or switch all traffic to the new one. We created null hypothesis and an alternate hypothesis to counter it. We have to collect enough evidence through our tests to reject the null hypothesis.

13. No, mean data imputation considered as bad practice in general, but there could be some use cases where it could come in handy. One of the main reason mean imputations is not used as it does not preserve the relationship among the variable. Like if you are doing mean imputation and the data is missing complete at random (MCR), mean imputation will not bias you estimate. One more reason is Mean Imputation Leads to An Underestimate of Standard Errors. Actually any data doing imputation will lead to an underestimate of standard errors, but other techniques lave very low error than mean imputation.

14. Linear regression is a predictive analysis technique to find the relation between 2 or more variable by fitting linear equation to the samples. The variable we want to predict is called the dependent variable. Least Squares Method is one of the method used to perform linear regression.

15. Statistics have majorly categorized into two types:

Descriptive statistics

Inferential statistics

Descriptive statistics: In this type of statistics, the data is summarized through the given observations. Descriptive statistics uses parameters such as the mean or standard deviation. Descriptive statistics is a way to organize, represent and describe a collection of data using tables, graphs, and summary measures.

Inferential Statistics: Inferential statistics deals with techniques used for the analysis of data, making estimates and drawing conclusions from limited information obtained through sampling and testing the reliability of the estimates. We perform hypothesis testing in for inferential statistics.