STATISTICS WORKSHEET-1

1. A. True
2. A. The Central limit theorm
3. B. Modeling bounded count data
4. D. All of the above mentioned
5. C. Poisson
6. B. False
7. B. Hypothesis
8. A. 0
9. C. Outliers can not conform regression relationship
10. A normal distribution or Gaussian distribution refers to a probability distribution where the values of a random variable are distributed symmetrically. These values are equally distributed on the left and right side of the central tendency. Then a bell shaped curve is formed.
    The maximum number of values appears close to the mean, the tail consist of only a few values. The empirical rule applies to such probability functions. Therefore, 68% of the values lie with in one standard deviation range. 95% of observations lie within two standard deviations, and 99.7% of the values appear within three standard deviations.

11. A common technique is to use the mean or median of the non-missing observations. This can be useful in cases where the number of missing observations is low. However, for large number of missing values, using mean or median can result in loss of variation in data and it is better to use imputations.

    Techniques for Handling the Missing Data

    List wise or case deletion.
    Pair wise deletion.
    Mean substitution.
    Regression imputation.
    Last observation carried forward.
    Maximum likelihood.
    Expectation-Maximization.
    Multiple imputation.

Multiple Imputation is always the best way to deal with missing data.

Multiple Imputation

Multiple imputation is considered a good approach for data sets with a large amount of missing data. Instead of substituting a single value for each missing data point, the missing values are exchanged for values that encompass the natural variability and uncertainity of the right values .Using the imputed data, the process is repeated to make multiple imputed data sets. Each set is then analyzed using the standard analytical procedures, and the multiple analysis results are combined to produce an overall result. Multiple imputations can produce statistically valid results even when there is a small sample size or a large amount of missing data.

12. A/B testing, also known as split testing or bucket testing, is essentially an experiment where two or more variants of an ad, marketing email, or web page are shown to users at random, and then different statistical analysis methods are used to determine which variant drives more conversions.Typically in A/B testing, the variant that gives higher conversions is the winning one, and that variant can help you optimize your site for better results.

13. Mean imputation reduces the variance of the imputed variables.

Mean imputation shrinks standard errors, which invalidates most hypothesis tests and the calculation of confidence interval.

Mean imputation does not preserve relationships between variables such as correlations.

imputing the mean preserves the mean of the observed data. So if the data are missing completely at random, the estimate of the mean remains unbiased. Since most research studies are interested in the relationship among variables, mean imputation is not a good solution.

14. Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the

regression equation with one dependent and one independent variable is defined by the formula

$$Y=a+bx+e$$

Y is the dependent variable and x is the independent variable ,a is called the intercept b is the slope of the equation. The slope is the amount by which y increases when x increases by 1 unit.e be the error term

15. The two main branches of statistics are
    1. Descriptive statistics
    2. Inferential statistics.

    Both of these are employed in scientific analysis of data and both are equally important

    Descriptive Statistics is a method of organizing, summarizing, and presenting data in a convenient and informative way.

    Inferential statistics helps in finding the conclusion regarding the population after analysis on the sample drawn from it.