PYTHON – WORKSHEET 1

Answers:

1. C (%)
2. B (0)
3. C (24)
4. A (2)
5. D (6)
6. C (the finally block will be executed no matter if the try block raises an error or not.)
7. A (It is used to raise an exception.)
8. C (In defining a generator)
9. A (_abc)
   C (abc2)
10. A (yield)
    B (raise)

STATISTICS WORKSHEET-1

1. A (True)
2. A (Central Limit Theorem)
3. B (Modeling bounded count data)
4. D (All of the mentioned)
5. C (Poisson)
6. B (False)
7. B (Hypothesis)
8. A (0)
9. C (Outliers cannot conform to the regression relationship)
10. Normal Distribution is also known as Gaussian Distribution. It also called as bell curve some time. Normal distribution has positive standard deviation and this positive standard deviation tell us how far the data is spread out. The graph will become narrow when the standard deviation will be small and the graph will become wider when the standard deviation will be larger.

    Approximately 68% od data falls within one standard deviation of mean

    Approximately 95% od data falls within two standard deviation of mean.

    Approximately 99.7% od data falls within three standard deviation of mean

11. There are many ways to handle missing data. Most of the time application will remove data in listwise sequence. Depending on why and how much data is gone, listwise

deletion may not be a good idea. There is another method called **Imputation.** **Imputation** is the process of substituting an estimate of missing values and analysing the data set as if the imputed values were true observed values. It includes following methods:

a. mean substitution
b. regression imputation
c. last observation carried forward
d. maximum likelihood
e. expectation-maximization
f. multiple imputation

12. A/B test is also known as split test. It is effective in learning of 2 versions of something and compare which one is more effective. It eliminates all the guess-work and enables experience optimizers to make data-backed decisions. A refers to "control" or original testing variable and B refers to "variation" or new version of testing variable. A/B testing works on:

a. Make a hypothesis: it is an assumption or an idea proposed to see if it is true.
   i. **Null hypothesis H0:** it states that sample observations result are purely from chance, there is no difference between control and variant group.
   ii. Alternative hypothesis H1: it states that the sample observations are influenced by some non-random clause, which states there is difference between control and variant group.
   After completing null and alternative hypothesis, decide which sample has to participate in test by creating two groups.

b. Control Group and Test Group: Here, random sampling will be done. It is needed as it removes bias, because we want results of test to be representative instead of sample itself. Another important thing is sample size. It is needed to describe the sample size before conducting test so we can eliminate under coverage bias. Then conduct test and collect data.

13. Mean imputation in missing data is a acceptable but using mean in the missing data can reduce accuracy of model and bias the result. Even inserting 0's can impact on accuracy of model. There are few advantages and disadvantages of mean imputation.
Advantages: missing values doesn't reduce sample size, simple to apply and understand
Disadvantages: it leads to bias in multivariate estimation such as correlation or regression, standard errors and variance of imputed variables are biased

14. Linear regression used to predict the value of variable based on another variable values. There are two variables- dependant and independent. The variable which we want to predict values of called dependent variable and the variables which we are using to predict the value of other variable called independent variable.
The equation of linear regression is: $y=c+b*x$

Where, y= estimated dependent variable score
c= constant
b= regression coefficient
x= score on the dependent variable

Uses of regression- determining the strength of predictors, furcating an effect, trend forecasting.

15. Statistics has two main branches and many sub-branches as follow:

A. Descriptive statistics: focuses on collecting, summarizing and presenting set of data
- **i. Measure of frequency**: it includes count, percent and frequency
- **ii. Measure of central tendency:** it includes mean, mode and median
- **iii. Measure of dispersion or variation:** it includes range, variance or standard deviation
- **iv. Measure of position:** includes percentile ranks and quartile ranks.

B. Inferential statistics: analyses sample data to draw conclusions.
- **i. One sample hypothesis test:** it is used to determine weather an unknown sample mean is different from specific value.
- **ii. Confidence interval:** it shows the probability that a parameter will fall within a pair of values around the mean.
- **iii. Contingency tables and chi-test:** used to determine whether there is difference between the expected frequencies and observed frequencies in pone or more categories.
- **iv. T-test or ANOVA**: use to compare means of two samples. ANOVA is used to compare the means among three or more groups.
- **v. Pearson correlation:** it represents the relationship between two variables that are measured on the same interval or ratio scale.
- **vi. Bi-variate regression:** is simple linear regression which is used to predict one variable from another variable.
- **vii. Multivariate regression:** is used when more than one variable is used to predict variation in another variable.

MACHINE LEARNING WORKSHEET-1

Answers:

1. A (Least square error)
2. A (Linear regression is sensitive to outliers)
3. B (Negative)
4. B (Correlation)
5. C (Low bias and high variance)
6. B (Predictive modal)

7. D (Regularization)
8. D (SMOTE)
9. C (Sensitivity and Specificity)
10. B (False)
11. B (Apply PCA to project high dimensional data)
12. A (We don't have to choose the learning rate)
B (It becomes slow when number of features is very large)
13. Regularization in machine learning is form of regression, that constrains, regularizes or shrinks the coefficient estimates towards zero. This technique discourages learning a more complex or flexible model, so as to avoid the risk of overfitting.
14. In regularization, there are few algorithms like LASSO, Ridge and Elastic-Net regression.
A. LASSO: also known as Least Absolute Shrinkage and Selection Operator. It uses L1 Regularization technique. It adds "Absolute value of magnitude" of coefficient, as penalty term to the loss function
B. Ridge regression: It uses L2 regularization. This regularization adds the penalty as model complexity increases. Ridge regression adds "Squared magnitude of the coefficient" as penalty term to the loss function.
15. An error term essentially means that the model is not completely accurate ad the result in differing results during real-world applications. It refers to the sum of the deviations within the regression line, which provides an explanation for the difference between the theoretical value of the model and the actual observed results.