# STATISTICS WORKSHEET 1

**Question 1** : Bernoulli random variables take (only) the values 1 and 0.

**Answer** : a) True

**Question 2** : Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases ?

**Answer** : a) Central Limit Theorem

**Question 3** : Which of the following is incorrect with respect to use of Poisson distribution?

**Answer** : b) Modeling bounded count data

**Question 4** : Point out correct statement.

**Answer** : d) All of the mentioned.

**Question 5** : _____ random variables are used to model rates.

**Answer** : c) Poisson

**Question 6** : Usually replacing the standard error by its estimated value does change the CLT.

**Answer** : b) False

**Question 7** : Which of the following testing is concerned with making decisions using data?

**Answer** : b) Hypothesis

**Question 8** : Normalized data are centered at _____ and have units equal to Standard deviations of the original data.

**Answer** : a) 0

**Question 9** : Which of the following statement is incorrect with respect to outliers ?

**Answer** : c) Outliers cannot conform to the regression relationship.

## Answer in brief

**Question 10** : What do you understand by the term Normal Distribution ?

**Answer** : In probability theory and Statistics, the Normal Distribution, also called the Gaussian Distribution, is the most significant continuous probability distribution. Sometimes it is also called bell shaped curve.

Normal Distribution Properties :

- In a normal distribution, the mean, median and mode are equal (i.e. Mean = Median = Mode).
- The total area under the curve should be equal to one.
- The normally distributed curve should be symmetric at the center.
- There should be exactly half of the values are to the right of the center and exactly half of the values are to the left of the center.
- The normal distribution should be defined by the mean and standard deviation.
- The normal distribution curve should have only one peak. (i.e. Unimodal)
- The curve approaches the x-axis, but it never touches and it extends farther away from the mean.

**Question 11**: How do you handle missing data ? What imputation techniques do you recommend ?

**Answer** :

- Many real world datasets may contain missing values for various reasons. They are often encoded as NaNs, blanks or any other placeholders.
- Training a model with a dataset that has a lot of missing values can drastically impact the machine learning model's quality. Some algorithms such as scikit-learn estimators assume that all values are numerical and have and hold meaningful value.
- One way to handle this problem is to get rid of observations that have missing data. However, you will risk losing data points with valuable information. A better strategy would be to impute the missing value.

Following are some imputation techniques :

1) Imputation Using (Mean / Median) Values :

This works by calculating the mean / median of the non-missing values in a column and then replacing the missing values within each column separately and independently from the others. It can only be used with numeric data.

Pros :

- Easy and fast.
- Works well with small numerical datasets.

Cons :

- Doesn't factor the correlation between features. It only works on the column level.
- It will give poor results on encoded categorical features.
- Doesn't account for the uncertainty in the imputations.

2) Imputation Using Most Frequent or Zero/Constant Values :

- It works with categorical features (strings) by replacing missing data with the most frequent values within each column.

Pros :

- Works well with categorical features.

Cons :

- It also doesn't factor the correlations between features.
- It can introduce bias in the data.

3) Imputation Using K-NN :

- The K nearest algorithm that is used for simple classification. It creates a basic mean impute then uses the resulting complete list to construct a KD Tree to compute nearest neighbors (NN). After it finds the K-NN, it takes the weighted average of them.

Pros :

- Can be much more accurate than the mean, median or most frequent imputation methods. (It depends on the dataset)

Cons :

- Computationally expensive. KNN works by storing the whole training dataset in memory.

- KNN is quite sensitive to outliers in the data (unlike SVM).

4) Extrapolation and Interpolation :

- It tries to estimate values from other observations within the range of a discrete set of known data points.

5) Hot – Deck Imputation :

- Works by randomly choosing the missing value from a set of related and similar variables.

Conclusion : There is no perfect way to compensate for the missing values in a dataset. Each strategy can perform better for certain datasets and missing data types but many perform much worse on other types of datasets.

**Question 12** : What is A / B testing ?

**Answer** : A / B testing - also called split testing or bucket testing – compares the performance of 2 versions of content to see which one appeals more to visitors / viewers.

A / B testing provides the most benefits when it operates continuously. A regular flow of tests can deliver a stream of recommendations on how to fine tune performance and continuous testing is possible because the available options for testing are nearly unlimited.

As noted above, A / B testing can be used to evaluate just about any digital marketing asset including :

- Emails
- Newsletters
- Advertisements
- Mobile apps
- Text messages
- Website pages
- Components on web pages

A / B testing examples :

A list of digital marketing elements that can be tested includes one or more of the items below :

- Navigation links
- Calls to action (CTAs)
- Design / Layout
- Copy
- Content offer
- Headline
- Logos and taglines / slogans
- Email subject line
- Friendly email “from” address
- Images
- Social media buttons

The role of analytics in website A / B testing :

- Throughout the lifecycle of any A / B test, analytics is at the heart of planning, execution and performance recommendations.

- The development of a test hypothesis requires a strong foundation in analytics. One need to understand current performance and traffic levels.

Some key data points of web analytics :

- Traffic (Page views, unique visitors) to the page, component or other element being reviewed for the test scenarios.
- Engagement (time spent, pages per visit, bounce rate)
- Performance trended over time.

Once an A / B test launches, analytics also plays a central role. A dashboard is used to monitor performance metrics in real time, to validate the test is operating as expected and to respond to any anomalies or unexpected results. This can include stopping the test, making adjustments and restarting, and ensuring performance data reflects any changes as well as the timing of those changes. The performance dashboard helps determine how long to keep the test running and to ensuring that statistical significance is achieved.

**Question 13** : Is mean imputation of missing data acceptable practice ?

**Answer** : Mean imputation is the practice of replacing null values in a dataset with the mean of the data.

There are demerits of using it viz are:

- It doesn't take into account feature correlation.
- It will give poor result on encoded categorical features. (Do not use it on categorical features.
- It doesn't account for the uncertainty in the imputations.

So, it is generally a bad practice.

**Question 14** : What is linear regression in Statistics ?

**Answer** : 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 dependent variable.
2) Which variables in particular are significant predictors of the outcome variable and in what way do they indicated by the sign of the beta estimates.

The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b * x

Where, y= estimated dependent variable score
c = constant / intercept / bias
b = regression coefficient / slope
x = score on the independent variable.

Three major uses for regression analysis are:

1) Determining the strength of predictors,
2) Forecasting an effect and,
3) Trend forecasting.

**Question 15** : What are the various branches of Statistics ?

**Answer** : Statistics is the application of Mathematics, which was basically considered as the Science of the different types of stats. For example, the collection and interpretation of data about a nation like its economy and population, military, literacy, etc. In terms of mathematical analysis, the Statistics include linear algebra, Stochastic study, differential equation and measure theoretic probability theory.

Branches of Statistics :

1) Descriptive Statistics 2) Inferential Statistics

**1) Descriptive Statistics** :

- In this type of Statistics, the data is summarised through the given observations. The summarization is one from a sample population using parameters such as the mean or Standard deviation.

- DS are also categorised into 4 different categories.

  - Measure of frequency
  - Measure of dispersion
  - Measure of central tendency
  - Measure of position.

  - The frequency measurement displays the no. of times a particular data occurs. Range, Variance, Standard Deviation are measures of dispersion. It identifies the spread of the data. Central tendencies are the mean, median and mode of the data and the measures of position describes the percentile and quartile ranks.

**2) Inferential Statistics** :

- It is used to interpret the meaning of Descriptive Statistics. That means once the data has been collected, analysed and summarised then we use these stats to describe the meaning of the collected data or it is used to draw conclusions from the data that depends on random variations such as observational errors, sampling variation, etc.

- IS is a method that allows us to use information collected from a sample to make decisions, predictions or inferences from a population. It grants us permission to give statements that goes beyond the available data or information. For example, Deriving estimates from hypothetical research.

**********