Worksheet Set-1 (Statistics)

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

   Answer: A) True

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

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

   Answer: b) Modeling bounded count data

4. **Point out the correct statement.**

   Answer: d) All of the mentioned

5. **______ random variables are used to model rates.**

   Answer: c) Poisson

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

   Answer: B)False

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

   Answer: b) Hypothesis

8. **Normalized data are centered at_____and have units equal to standard deviations of the original data.**

   Answer: a) 0

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

   Answer: c) Outliers cannot conform to the regression relationship

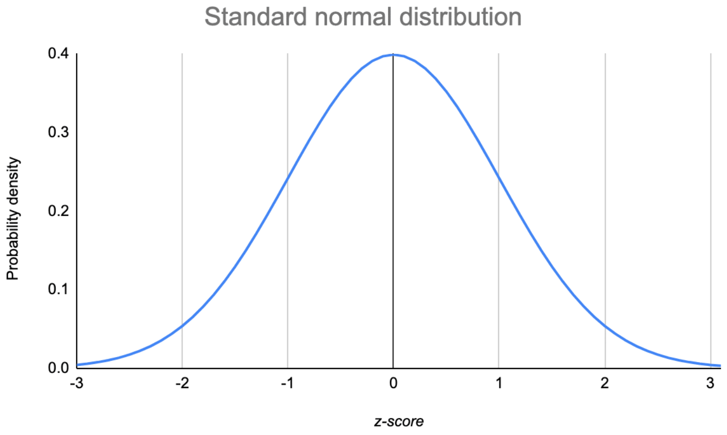
10. **What do you understand by the term Normal Distribution?**

    Normal Distribution also called Gaussian distribution, most common distribution function for independent , randomly generated variables. In simpler words, it is an arrangement of a data set in which most values cluster in the middle range and rest taper off symmetrically towards either extreme.

    The graph of the normal distribution is characterized by two parameters: the mean, or average, which is the maximum of the graph and about which the graph is always symmetric; and the standard deviation, which determines the amount of dispersion away from the mean.

    Graphical representation of normal distribution is called a bell-curve . In normal-distribution, the mean, median and mode are all the same.

    Example: Height is one example that follow normal distribution, Most people are of average height, no of people who are taller or shorter than the average height are fairly equal and very small no of people are either extremely short or tall.



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

Data may be missing due to different reasons, Missing at Random(MAR), Missing completely at Random(MCAR), Missing Not at Random(MNAR).

We can handle the missing data in two ways:

- Deletion
- Imputation

Deletion may not be the most effective method to handle missing data, if too much data is discarded it might not be possible to deliver the best results. Imputation methods can deliver reliable results.

The different Imputation techniques are as follows:

Single Imputation techniques:

Mean, Median, Mode:

This technique is one of the most common methods of imputing values when dealing with missing data. When there is small no of missing observations, we can calculate the mean and median of the existing observations. However, when there is large no of missing data, mean and median results can result in a loss of variation in the data

Time- Series Specific Methods:

The time series methods of imputation assume the adjacent observations will be like the missing data. These methods work well when that assumption is valid. However, these methods won't always produce reasonable results, particularly in the case of strong seasonality.

- No trend or seasonality.
- Trend, but no seasonality.
- Seasonality, but no trend.
- Both trend and seasonality.

Last Observation Carried Forward (LOCF) & Next Observation Carried Backward (NOCB)

In this method, every missing value is replaced with the last observed value. Longitudinal data track the same instance at different points along a timeline. This method is easy to understand and implement. However, this method may introduce bias when data has a visible trend. It assumes the value is unchanged by the missing data.

Linear Interpolation

Linear interpolation is often used to approximate a value of some function by using two known values of that function at other points. This formula can also be understood as a weighted average. The weights are inversely related to the distance from the end points to the unknown point. The closer point has more influence than the farther point.

Seasonal Adjustment with Linear Interpolation

We need perform the seasonal adjustment by computing a centred moving average or taking the average of multiple averages – say, two one-year averages – that are offset by one period relative to another. You can then complete data smoothing with linear interpolation.

Multiple Imputation Techniques:

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 uncertainty 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.

K Nearest Neighbors

We choose a distance measure for k neighbors, and the average is used to impute an estimate and then we must select the number of nearest neighbors and the distance metric. KNN can identify the most frequent value among the neighbors and the mean among the nearest neighbors.

12. What is A/B testing?

A/B testing, also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable (web page, page element, etc.) are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drive business metrics. It eliminates all the guesswork out of website optimization.

In A/B testing, A refers to ‘control’ or the original testing variable. Whereas B refers to ‘variation’ or a new version of the original testing variable.

A/B testing is one of the components of the overarching process of Conversion Rate Optimization (CRO), using which we can gather both qualitative and quantitative user insights. We can use this collected data to understand user behavior, engagement rate, pain points, and even satisfaction with website features, including new features, revamped page sections, etc.

We should do A/B testing for various reasons:

- Solve visitor pain points
- Get better ROI from existing traffic
- Reduce bounce rate
- Make low-risk modifications
- Achieve statistically significant improvements
- Redesign website to increase future business gains

13. Is mean imputation of missing data acceptable practice?

Mean Imputation is a commonly used method but in case of large missing data, it is unreliable. There are different problems with mean imputation for missing data.

Bad Practice in general

As mean imputation is the replacement of the missing observation with the mean of the non-missing observations of that variable, it can have a negative impact on the accuracy when training a ML model.

Mean Imputation does not preserve the relationship between variables:

It dosen't take into account the fact that variables are correlated to each other.

Reduces the variance of the data:

When comparining variance on real data and missing data, we tend to get reduced variance which leads to narrower confidence interval in probability distribution. As variance is getting effected, it underestimates the standard deviation as well.

14.What is linear regression in statistics?

Linear regression is the commonly used technique in predictive analysis. It is the relationship between one or more predictor variables and one outcome variable.

Idea of regression is to examine two things:

1) Does the predictor variable provides an accurate outcome?
2) Which variables are significant predictors of the outcome variable and In what way do they estimates the outcome.

Major uses of Regression Analysis are:

1) Determining strength of predictors

2) Forecasing as effect

3) Trend Forecasting

Types of Linear Regression:

- **Multiple linear regression:**

  1 dependent variable (interval or ratio) , 2+ independent variables (interval or ratio or dichotomous)

- **Logistic regression:**

  1 dependent variable (dichotomous), 2+ independent variable(s) (interval or ratio or dichotomous)

- **Ordinal regression**

  1 dependent variable (ordinal), 1+ independent variable(s) (nominal or dichotomous)

- **Multinomial regression**

  1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio or dichotomous)

- **Discriminant analysis**

  1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio)

We need to consider model fitting when selecting a model for analysis. Adding independent variables to a linear regression model will always increase the explained variance of the model. However, overfitting can occur by adding too many variables to the model, which reduces model generalizability. Statistically, if a model includes a large number of variables, some of the variables will be statistically significant due to chance alone.

15. What are the various branches of statistics?

There are 2 Branches of Statistics:

- Descriptive Statistics

- Inferential Statistics

Both descriptive and inferential statistics go hand in hand and one cannot exist without the other.

Descriptive Statistics:

It deals with the presentation and collection of data. This is usually the first part of a statistical analysis.

It is important to be aware of designing experiments, choosing the right focus group, and avoid biases that can alter the result of the experiment.

Different areas of study require different kinds of analysis using descriptive statistics. For example, a physicist studying turbulence in the laboratory needs the average quantities that vary over small intervals of time. The nature of this problem requires that physical quantities be averaged from a host of data collected through the experiment.

Inferential Statistics:

It involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics.

Most social sciences experiments deal with studying a small sample population that helps determine how the population in general behaves. It is to provide prediction of future and generalization about a population by studying a smaller sample.

We need to be careful while drawing conclusions as to not draw biased or wrong conclusions. There are ways in which one can manipulate studies and results through various means. For example, data dredging is increasingly becoming a problem as computers hold loads of information and it is easy, either intentionally or unintentionally, to use the wrong inferential methods.