STATISTICS WORKSHEET-1

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

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

a) True

b) False

Answer : A) True

A Bernoulli random variable is the simplest kind of random variable. It can take on two values, 1 and 0. It takes on a 1 if an experiment with probability p resulted in success and a 0 otherwise

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?

a) Central Limit Theorem

b) Central Mean Theorem

c) Centroid Limit Theorem

d) All of the mentioned

Ans : A) Central Limit Theorem

the CENTRAL LIMIT THEOREM (CLT) - one of the most important theorems in all of statistics. It states that the distribution of averages of iid variables (properly normalized) becomes that of a standard normal as the sample size increases.

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

a) Modeling event/time data

b) Modeling bounded count data

c) Modeling contingency tables

d) All of the mentioned

Ans : b) Modeling bounded count data

Poisson distribution is used for modeling unbounded count data.

4. Point out the correct statement.

a) The exponent of a normally distributed random variables follows what is called the log- normal distribution

b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent

c) The square of a standard normal random variable follows what is called chi-squared distribution

d) All of the mentioned

Ans : d) All of the mentioned.

Many random variables, properly normalized, limit to a normal distribution.

5. ______ random variables are used to model rates.

a) Empirical
b) Binomial
c) Poisson
d) All of the mentioned

Ans : c) Poisson

Poisson distribution is used to model counts.

6. Usually replacing the standard error by its estimated value does change the CLT.
a) True
b) False

Ans : b) False

7. Which of the following testing is concerned with making decisions using data?
a) Probability
b) Hypothesis
c) Causal
d) None of the mentioned

Ans : b) Hypothesis
Hypothesis testing is a statistical method that is used in making statistical decisions using experimental data. Hypothesis Testing is basically an assumption that we make about the population parameter.

8.. Normalized data are centered at ______ and have units equal to standard deviations of the original data.
a) 0
b) 5
c) 1
d) 10
Ans : a) 0

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

a) Outliers can have varying degrees of influence

b) Outliers can be the result of spurious or real processes

c) Outliers cannot conform to the regression relationship

d) None of the mentioned

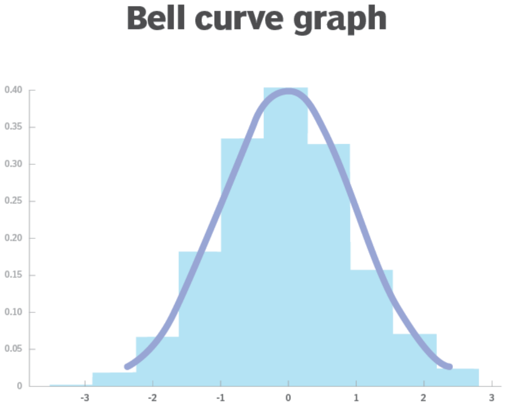
Ans : c)

Outliers can conform to the regression relationship.

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

- A normal distribution is a type of **continuous probability distribution in which most data points cluster toward the middle of the range, while the rest taper off symmetrically toward either extreme.**
- **The middle of the range is also known as the *mean* of the distribution.**
- Graphically, a normal distribution is a **bell curve because of its flared shape.**
- All kinds of variables in natural and social sciences are normally or approximately normally distributed. **Height, birth weight, reading ability, job satisfaction, or SAT scores** are just a few examples of such variables.
- **Some of the important properties of the normal distribution are listed below:**
    - 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 1.**
    - The normally distributed **curve should be symmetric at the centre.**
    - There should be **exactly half of the values are to the right of the centre and exactly half of the values are to the left of the centre.**
    - The normal distribution should be defined by **the mean and standard deviation.**
    - **The mean ,** is the central highest value of the bell curve. All other values in the distribution either cluster around it or are at some distance away from it. Changing the mean on a graph will shift the entire curve along the x-axis, either toward the left or toward the right.
    - **The Standard Deviation:** it represents the typical distance between the average and the observations. A smaller deviation will reduce the spread -- tightening the distribution -- while a larger deviation will increase the spread and produce a wider distribution.
    - The normal distribution curve **must 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.



Mathematically , it is represented as following :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Where,

x **is value of the variable**

f(x) **represents the probability density function**

μ (mu) **is the mean**

σ (sigma) **is the standard deviation.**

The empirical rule for normal distributions describes where most of the data in a normal distribution will appear, and it states the following:

- 68.2% of the observations will appear within +/-1 standard deviation of the mean;
- 95.4% of the observations will fall within +/-2 standard deviations; and
- 99.7% of the observations will fall within +/-3 standard deviations.

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

An Incomplete data can bias the results of the machine learning models and/or reduce the accuracy of the model.
Missing data are the values that are not stored for some variable/s in the given dataset, there can be multiple reasons why we have missing data. There are two ways to Handle Missing data :

1. **Deleting the Missing data**

2. **Imputing the Missing Data**

In order to get the effective approach of handling missing datas it's necessary to understand why the data could be missing, the different types of missing dataset :

1. **Missing Completely at Random (MCAR)** - The probability of data being missing is the same for all the observations.Tn this case, there's no relation between the missing data and other values observed in given dataset, i.e., the data missing is completely random of missing because of human error. Now, this type of datas can be deleted.
2. **Missing At Random (MAR)** : The reason for missing values can be explained by variables, there can be some relation between the missing and other values in dataset. In this case the data is not missing for all observation but, only within sub samples of the data, and there is a pattern in the missing values. This type of data's can be deleted or imputed.

3. **Missing NOT At Random** : Missing values depend on the unobserved data. If there is some structure/pattern in missing data and other observed data **can not explain** it, then it is considered to be Missing Not At Random (MNAR). If the missing data does not fall under the MCAR or MAR, it can be categorized as MNAR. These type of Missing values should not be deleted , instead we can replace these missing values.

Data Imputation : The replacement of missing or inconsistent data elements with approximated values is known as imputation in data. It is intended for the substituted values to produce a data record that passes edits on Numerical or Categorial missing data.

Following are the imputation techniques :

- **Arbitrary Value** :By making an educated guess about the missing value, then you can replace it with some arbitrary value using ".fillna(0)". Works best for Numerical Variables.
- **Replacing with the mean/median :** This is the most common method of imputing missing values of numeric columns. If there are outliers, then the mean will not be appropriate. The median is the middlemost value. It's better to use the median value for imputation in the case of outliers.
- **K Nearest Neighbors**: The objective is to find the k nearest examples in the data where the value in the relevant feature is not absent and then substitute the value of the feature that occurs most frequently in the group.The algorithm uses '**feature similarity**' to predict the values of any new data points. This means that the new point is assigned a value based on how closely it resembles the points in the training set. This can be very useful in making predictions about the missing values by finding the *k's* closest neighbours to the observation with missing data and then imputing them based on the non-missing values in the

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

- **Most Frequent Value/ Replacing with Mode** : The most frequent value in the column is used to replace the missing values in another popular technique that is effective for both nominal and numerical features.
- **Imputation Using Deep Learning (Datawig)**: This method works very well with categorical and non-numerical features. It is a library that learns Machine Learning models using Deep Neural Networks to impute missing values in a dataframe. It also supports both CPU and GPU for training.

In 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 may perform much worse on other types of datasets.

12. What is A/B testing?

An AB test is an example of **statistical hypothesis testing**, a process whereby a hypothesis is made about the relationship between two data sets and those data sets are then compared against each other to determine if there is a statistically significant relationship or not.

AB Testing Procedure :

- **Identify goals:** Conversion goals are the metrics that you are using to determine whether or not the variation is more successful than the original version. Goals can be anything from clicking a button or link to product purchases.
- **Make a hypothesis:** Once we've identified a goal we can begin generating A/B testing ideas and test hypotheses for why we think they will be better than the current version. Once we have a list of ideas, prioritize them in terms of expected impact and difficulty of implementation. In Hypothesis testing we have two types of testing :
  1. **Null hypothesis or H0:**
  The **null hypothesis** is the one that states that sample observations result purely from chance. From an A/B test perspective, the null hypothesis states that there is **no** difference between the control and variant groups. It states the default position to be tested or the situation as it is now, i.e. the status quo. Here our H0 is " there is no difference in the conversion rate in customers receiving newsletter A and B".
  2. **The alternative hypothesis H1** : challenges the null hypothesis and is basically a hypothesis that the researcher believes to be true. The alternative hypothesis is what you might hope that your A/B test will prove to be true.
- **Create different variations:** Using A/B testing software (like Optimizely Experiment), make the desired changes to an element of your website or mobile app. This might be changing the color of a button, swapping the order of elements on the page template, hiding navigation elements, or something entirely custom. Many leading A/B testing tools have a visual editor that will make these changes easy.
- **Run experiment:** Run the experiment and wait for visitors to participate. At this point, every visitors interaction with each experience is measured, counted and compared against the baseline to determine how each performs.
- **Wait for the test results:** Depending on how big the the target audience is, it can take a while to achieve a satisfactory result. Good experiment results will tell us when the results are

statistically significant and trustworthy. Otherwise it would be hard to tell if the change truly made an impact.

- **Analyze results:** Once the experiment is complete, it's time to analyze the results. A/B testing software will present the data from the experiment and show you the difference between how the two versions of your page performed .It is important to achieve statistically significant results.

A/B use in different fields :

A media company might want to increase readership, increase the amount of time readers spend on their site, and amplify their articles with social sharing. To achieve these goals, they might test variations on:

- Email sign-up modals
- Recommended content
- Social sharing buttons

A travel company may want to increase the number of successful bookings are completed on their website or mobile app, or may want to increase revenue from ancillary purchases. To improve these metrics, they may test variations of:

- Homepage search modals
- Search results page
- Ancillary product presentation

An e-commerce company might want to improve their customer experience, resulting in an increase in the number of completed checkouts, the average order value, or increase holiday sales. To accomplish this, they may A/B test:

- Homepage promotions
- Navigation elements
- Checkout funnel components

A technology company might want to increase the number of high-quality leads for their sales team, increase the number of free trial users, or attract a specific type of buyer. They might test:

- Lead form fields
- Free trial signup flow

13. Is mean imputation of missing data acceptable practice?

The process of replacing null values in a data collection with the data's mean is known as mean imputation. Even Though Mean Imputation is Easy and fast imputation technique , it only works well for small numerical datasets, it has more cons to pros.
Cons:

- Doesn't factor the correlations between features.
- It only works on the column level.
- Will give poor results on encoded categorical features (do NOT use it on categorical features).
- Not very accurate.
- Doesn't account for the uncertainty in the imputations.
- It decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower

Hence, Mean imputation is typically considered terrible practice since it ignores feature correlation.

14. What is linear regression in statistics?

Linear regression is a type of statistical analysis used to predict the relationship between two variables. It assumes a linear relationship between the independent variable and the dependent variable, and aims to find the best-fitting line that describes the relationship. The line is determined by minimizing the sum of the squared differences between the predicted values and the actual values.Linear regression is commonly used in many fields, including economics, finance, and social sciences, to analyze and predict trends in data. It can also be extended to multiple linear regression, where there are multiple independent variables, and logistic regression, which is used for binary classification problems.

Linear regression shows the linear relationship between the independent(predictor) variable i.e. X-axis and the dependent(output) variable i.e. Y-axis, called linear regression. If there is a single input variable **X**(independent variable), such linear regression is called **simple linear regression**.

15. What are the various branches of statistics?

Statistics is a method of interpreting, analysing and summarising the data. Hence, the types of statistics are categorised based on these features: Descriptive and inferential statistics. Based on the representation of data such as using pie charts, bar graphs, or tables, we analyse and interpret it. Statistics have majorly categorised into two types:

1. **Descriptive statistics**
2. **Inferential statistics**

Descriptive Statistics

In this type of statistics, the data is summarised through the given observations. The summarisation is one from a sample of population using parameters such as the mean or standard deviation. Descriptive statistics is a way to organise, represent and describe a collection of data using tables, graphs, and summary measures. For example, the collection of people in a city using the internet or using Television.

Descriptive statistics are also categorised into four different categories:

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

The frequency measurement displays the number of times a particular data occurs. Range, Variance, Standard Deviation are measures of dispersion. It identifies the spread of data. Central tendencies are the mean, median and mode of the data. And the measure of position describes the percentile and quartile ranks.

Inferential Statistics

This type of statistics 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 we can say, it is used to draw conclusions from the data that depends on random variations such as observational errors, sampling variation, etc.

Inferential Statistics 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.