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
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
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
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
5. _____ random variables are used to model rates.
   a) Empirical
   b) Binomial
   c) Poisson
   d) All of the mentioned
6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False
7. 1. Which of the following testing is concerned with making decisions using data?
   a) Probability
   b) Hypothesis
   c) Causal
   d) None of the mentioned
8. 4. Normalized data are centered at_____and have units equal to standard deviations of the original data.
   a) 0
   b) 5
   c) 1
   d) 10
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

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

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

   The probability density function for a continuous random variable in a system defines the Normal Distribution. Let's assume that X is the random variable and that f(x) is the probability density function. In order to determine the probability of the random variable X, it specifies a function that is integrated across the range or interval (x to x + dx) while taking into account values between x and x+dx.

$$f(x) \geq 0 \ \forall \ x \in (-\infty, +\infty)$$

$$\text{And } \int_{-\infty}^{+\infty} f(x) = 1$$

Normal Distribution Formula

The probability density function of normal or gaussian distribution is given by;

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

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

   When one or more of an individual's variables have no values, there are missing data. The statistical power of the analysis may decline as a result of missing data, which may affect the reliability of the findings.
   There are several causes of missing data. The data is gathered from many sources, and there is a danger that the data might be lost during mining. However, the most frequent reason for missing data is item nonresponse, which refers to persons who are afraid to respond to sensitive issues like age, salary, or gender or who are unwilling to do so due to a lack of understanding of the topic.

Types of Missing data

There are 3 major categories of missing values which are given as follows:

Missing Completely at Random(MCAR)

If the missing values on a particular variable (Y) don't have a link with other variables in a given data set or with the variable (Y) itself, the variable is missing completely at random (MCAR). To put it another way, when data is MCAR, neither a link between the

data missing and any values nor a specific cause for the missing values exists.

Missing at Random(MAR)

When there is a systematic association between missing values and other observable data but not the missing data, MAR develops because the missingness is not random.

Missing Not at Random(MNAR)

The final and most difficult situation of missingness. MNAR occurs when the missingness is not random, and there is a systematic relationship between missing value, observed value, and missing itself. To make sure, If the missingness is in 2 or more variables holding the same pattern, you can sort the data with one variable and visualize it.

Imputation techniques:

The imputation technique replaces missing values with substituted values. The missing values can be imputed in many ways depending upon the nature of the data and its problem. Imputation techniques can be broadly they can be classified as follows:

Imputation with constant value:

As the title hints — it replaces the missing values with either zero or any constant value.

We will use the SimpleImputer class from sklearn.

```
from sklearn.impute import SimpleImputer
train_constant = train.copy()
#setting strategy to 'constant'
mean_imputer = SimpleImputer(strategy='constant') # imputing using constant value
train_constant.iloc[:,:] = mean_imputer.fit_transform(train_constant)
train_constant.isnull().sum()
```

```
Item_Identifier            0
Item_Weight                0
Item_Fat_Content           0
Item_Visibility            0
Item_Type                  0
Item_MRP                   0
Outlet_Identifier          0
Outlet_Establishment_Year  0
Outlet_Size                0
Outlet_Location_Type       0
Outlet_Type                0
Item_Outlet_Sales          0
dtype: int64
```

Imputation using Statistics

Only the Simple Imputer approach will change; the syntax is the same as imputation with constant. You can choose from "Mean," "Median," or "Most Frequent."

The missing values will be replaced by "Mean" using the mean of each column. Data that is numerical and not skewed is preferable.

The word "Median" will use the median in each column to fill in any missing values. Data that is numerical and skewed is desirable.

“Most frequent” will replace missing values using the most frequent in each column. It is preferred if data is a string(object) or numeric.

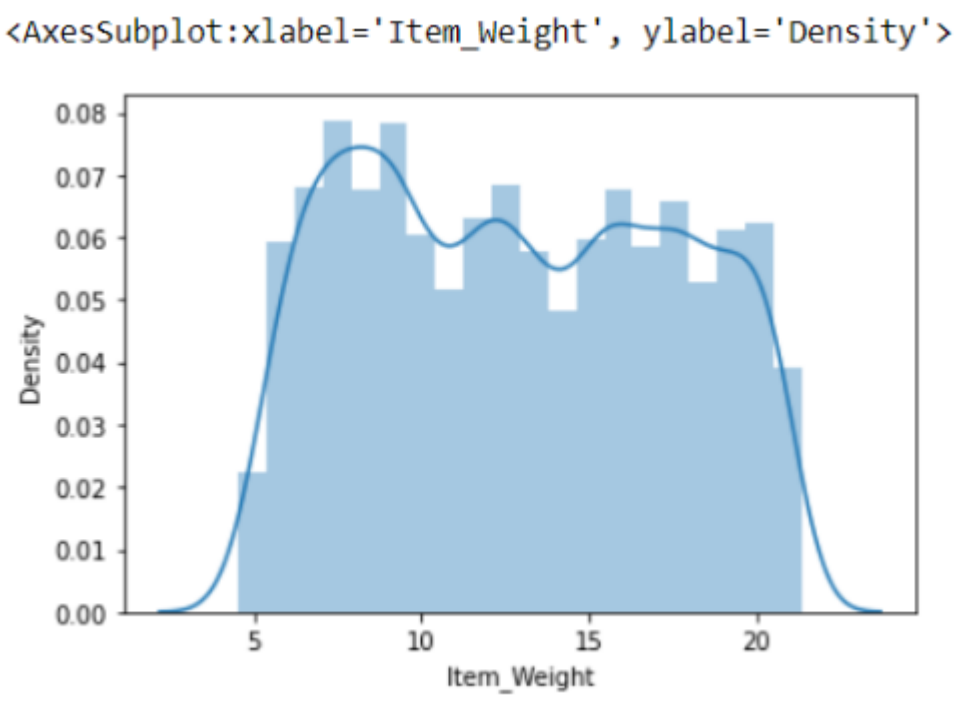
Before using any strategy, the foremost step is to check the type of data and distribution of features (if numeric).

```
train['Item_Weight'].dtype
```

```
dtype('float64')
```

```
sns.distplot(train['Item_Weight'])
```

```
<AxesSubplot:xlabel='Item_Weight', ylabel='Density'>
```



Item Weight column satisfying both conditions numeric type and doesn't have skewed (follow Gaussian distribution). here, we can use any strategy.

```
from sklearn.impute import SimpleImputer
train_most_frequent = train.copy()
#setting strategy to 'mean' to impute by the mean
mean_imputer = SimpleImputer(strategy='most_frequent')# strategy can also be mean or median
train_most_frequent.iloc[:,:] = mean_imputer.fit_transform(train_most_frequent)
train_most_frequent.isnull().sum()
```

```
Item_Identifier            0
Item_Weight                0
Item_Fat_Content           0
Item_Visibility            0
Item_Type                  0
Item_MRP                   0
Outlet_Identifier          0
Outlet_Establishment_Year  0
Outlet_Size                0
Outlet_Location_Type       0
Outlet_Type                0
Item_Outlet_Sales          0
dtype: int64
```

Advanced Imputation Technique:
Unlike the previous techniques, Advanced imputation techniques adopt machine learning algorithms to impute the missing values in a dataset. Followings are the machine learning algorithms that help to impute missing values.

K_Nearest Neighbor Imputation:

The KNN algorithm helps to impute missing data by finding the closest neighbors using the Euclidean distance metric to the observation with missing data and imputing them based on the non-missing values in the neighbors.

```
train_knn = train.copy(deep=True)
from sklearn.impute import KNNImputer
knn_imputer = KNNImputer(n_neighbors=2, weights="uniform")
train_knn['Item_Weight'] = knn_imputer.fit_transform(train_knn[['Item_Weight']])
train_knn['Item_Weight'].isnull().sum()
```

The fundamental weakness of KNN doesn't work on categorical features. We need to convert them into numeric using any encoding method. It requires normalizing data as KNN Imputer is a distance-based imputation method and different scales of data generate

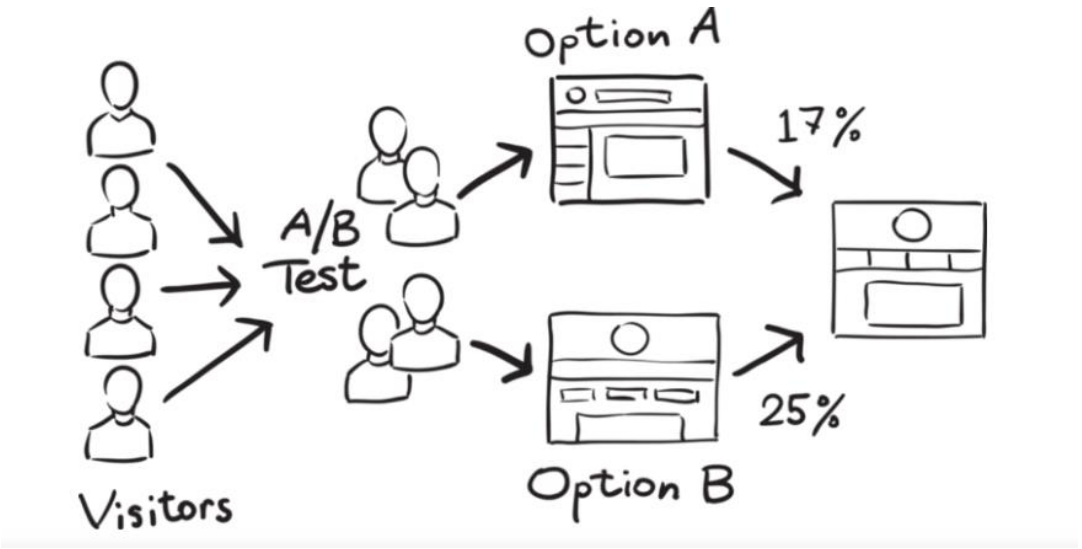
biased replacements for the missing values.

3. What is A/B testing?

An elementary randomised control experiment is A/B testing. It is a method for contrasting two variations of a variable to see which performs better in a regulated setting.

Let's imagine, for example, that you own a business and wish to boost product sales. Either random experimentation or scientific and statistical techniques can be used in this situation. One of the most well-known and frequently employed statistical tools is A/B testing.

For e.g. Here A will remain unchanged while you make significant changes in B's packaging. Now, on the basis of the response from customer groups who used A and B respectively, you try to decide which is performing better.



It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. The population refers to all the customers buying your product, while the sample refers to the number of customers that participated in the test.

4. Is mean imputation of missing data acceptable practice?

Mean imputation is the process of replacing null values in a data collection with the mean of the data.

Mean imputation is frequently seen as a bad approach since it disregards feature correlation. Think about the following situation: we have a table containing age and fitness scores, but the fitness score for an eight-year-old is missing. The elderly person would appear to be far more fit than he actually is if we average the fitness ratings of those between the ages of 15 and 80.

Second, mean imputation increases bias while reducing the variance of our data. The model is less accurate and the confidence interval is smaller as a result of the lower variance.

5. What is linear regression in statistics?

The associations between at least one explanatory variable and an outcome variable are modelled using linear regression. The independent and dependent variables are recognised as just that—variables. The process is referred to as simple linear regression when there is just one independent variable (IV). Multiple regression is the statistical term for a situation when there are more IVs.

This flexible analysis allows you to separate the effects of complicated research questions by modeling and controlling all relevant variables. It lets you isolate the role that each variable plays. This procedure uses sample data to estimate the population parameters. The regression coefficients in your statistical output are the parameter estimates.

Understanding the correlations between variables and predicting are the two main applications of linear regression.

1) The estimated strength and polarity (positive/negative) of the link between each independent variable and the dependent variable are represented by the coefficients.
2) Given the values of the independent variables you define, a linear regression equation enables you to predict the mean value of the dependent variable.

6. What are the various branches of statistics?

Data collection, descriptive statistics, and inferential statistics are the three main branches of statistics.

1) Data collection

The process of gathering data is what matters most. In terms of mathematics, this generally doesn't need to worry us too much, but there are important factors to take into account when gathering data.

You must be selective about where you obtain data if you are gathering it. Assume, for instance, that you wish to survey individuals about their election preferences. You must select a representative sample of people because it is impractical to ask the entire nation (the population). It's not as simple as it seems. For instance, polls were occasionally conducted in the middle of the 20th century by phoning random numbers in a telephone directory.

This seems representative, but only the wealthy had telephones in those days, so you were only polling a small portion of society—a group that would be more likely to support one party than the other.

2) Descriptive statistics

The area of statistics known as descriptive statistics deals with how we portray the data we have. Basically, this may be done in one of two ways: either visually (using graphs, charts, etc.) or quantitatively (via averages and so on).

Descriptive statistics fundamental goal is to "display the data" in an intelligible manner. Every bit of information needs to be summarized because if you just list it out in its entirety, no one will understand it.
Imagine if every single individual surveyed by a polling organization had their votes posted on the TV news; it would be a massive list of parties, and you couldn't draw any conclusions.
n. Rather, you are given visual representations of the data (a bar chart, for example) that may show the percentage of the vote that each party received. In the general election of 2010, close to 30 million people cast ballots.
You would be completely bewildered if every vote were simply listed and displayed one after the other; instead, a summary of votes is given (for example, as percentages: Conservative 36%, Labour 29%, Liberal Democrat 23%, Others 12%). This is an illustration of descriptive statistics, which "describe" or "summarise" the total data such that it is understandable to individuals.

3) Inferential statistics

The part of statistics that deals with drawing inferences from the data is called inferential statistics.
After a series of incidents, a council, for instance, could be contemplating lowering the speed limit on a major route. To determine if the speed limit needs to be decreased, they may survey vehicle speeds (gather data) to make this determination.

For instance, several vehicles are moving too quickly). Be aware, however, that this may not be the case; everyone may be travelling at a pace that is totally appropriate, and the incidents may not be related to speed at all (a blind spot or a pothole, for example). It's called inferential statistics when you take your available data and draw a "inference" or "conclusion" from it. When we talk about topics like hypothesis testing, where we check to see whether the data backs up a claim we make, we'll see much more of this in the future.