STATISTICS WORKSHEET-1

1. True
2. Central Limit Theorem
3. Modeling Contingency tables
4. All of the mentioned
5. Poisson
6. False
7. Hypothesis
8. 0
9. Outliers cannot conform to the regression relationship

WORKSHEET

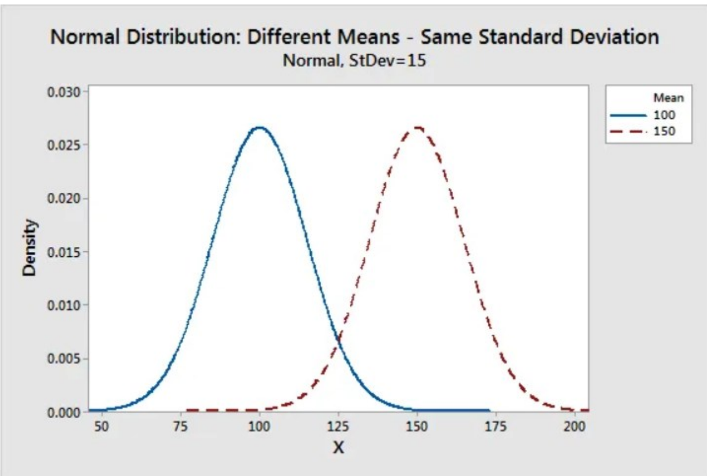
Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

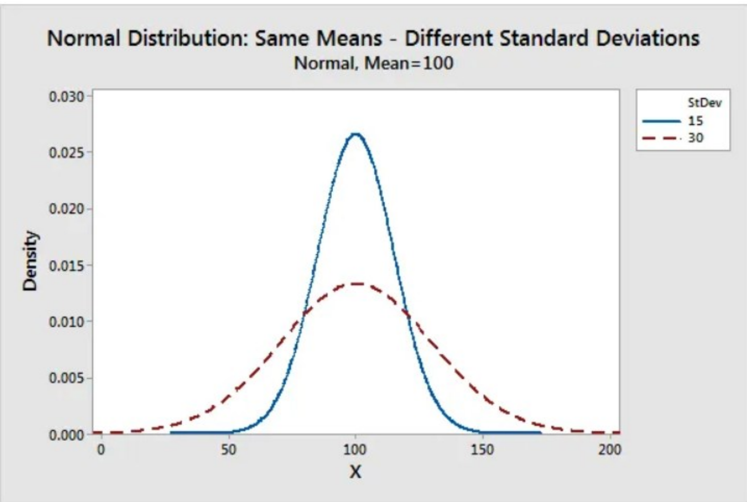
The Normal distribution is also known as **Gaussian** or **Gauss distribution**. Many groups follow this type of pattern. That's why it's widely used in business, statistics, and in government bodies. For instance, we can use it to measure heights of people, Measurement errors, blood pressure, points on a test, IQ scores and salaries.

There are two main parameters of a normal distribution- the **mean** and **standard deviation**. With the help of these parameters, we can decide the shape and probabilities of the distribution wrt our problem statement. As the parameter value changes, the shape of the distribution changes.

- The mean determines the location of the peak, and most of the data points are clustered around the mean in a normal distribution graph.
- If we change the value of the mean, then the curve of normal distribution moves either to the left or right along the X-axis.



- The standard deviation measures how the data points are dispersed relative to the mean.
- It determines how far the data points are away from the mean and represents the distance between the mean and the data points.
- The standard deviation defines the width of the graph. As a result, changing the value of standard deviation tightens or expands the width of the distribution along the x-axis.
- Usually, a smaller standard deviation wrt to the mean results in a steep curve while a larger standard deviation results in a flatter curve.



The normal distribution is a continuous probability distribution that is symmetrical on both sides of the mean, that gives you a symmetrical bell curve, so the right side of the center is a mirror image of the left side. The area under the normal distribution curve represents probability and the total area under the curve sums to one.

Most of the continuous data values in a normal distribution tend to cluster around the mean, and the further a value is from the mean, the less likely it is to occur. The tails are asymptotic, which means that they approach but never quite meet the horizon. For a perfectly normal distribution the mean, median and mode will be the same value, visually represented by the peak of the curve.

Properties of Normal Distribution:

- It is symmetric
- The mean, median and mode are equal
- Empirical rule
- The total area under the curve is unity(=1)

---

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

The first step in handling missing values is to look at the data carefully and find out the missing values. To find the missing values in entire data set.

```
IN: import pandas as pd
    train_df= pd.read_csv("train.csv")
#Find the missing values from each column
train_df.isnull().sum()
```

This gives the output of the missing values.
Analyze each column with missing values carefully to understand the reasons behind the missing values as it is crucial to find out the strategy for handling the missing values.

There are 2 primary ways of handling missing values:

1. Deleting the Missing values
2. Imputing the Missing Values

Deleting the Missing value

Generally, this approach is not recommended. It is one of the quick and dirty techniques one can use to deal with missing values. If the missing value is of the type Missing Not At Random (MNAR), then it should not be deleted.

If the missing value is of type Missing At Random (MAR) or Missing Completely At Random (MCAR) then it can be deleted.

The disadvantage of this method is one might end up deleting some useful data from the dataset.There are 2 ways one can delete the missing values:

Deleting the entire row

If a row has many missing values then you can choose to drop the entire row. If every row has some (column) value missing then you might end up deleting the whole data.

Code to drop the entire row is as follows:

```
IN: df=train_df.dropna(axis=0)
    df.isnull().sum()
```

Deleting the entire column
If a certain column has many missing values, then we can choose to drop the entire column.

Code to drop the entire column is as follows:

IN:

```
df=train_df.drop([‘Dependents‘],axis=1)

df.isnull().sum()
```

Imputing the Missing values:There are different ways of replacing the missing values.We can use the python libraries Pandas and Sci-kit learn as follows:

Replacing with Arbitrary Value

Ex. In the following code, we are replacing the missing values of the ‘Dependents’ column with ‘0’.

IN:

```
#Replace the missing value with ‘0’ using ‘fiilna’ method

train_df[‘Dependents’]=train_df[‘Dependents’].fillna(0)

train_df[‘Dependents’].isnull().sum()
```

Replacing With Mean

This is the most common method of imputing missing values of numeric columns. If there are outliers then the mean will not be appropriate. In such cases, outliers need to be treated first.

You can use the ‘fillna’ method for imputing the columns ‘LoanAmount’ and ‘Credit_History’ with the mean of the respective column values.

IN:

```
# Replace the missing values for numerical columns with mean

train_df[‘LoanAmount’]=train_df[‘LoanAmount’].fillna(train_df[‘LoanAmount’].mean())

train.df[‘Credit_History’]=train_df[‘Credit_History’].fillna(train_df[‘Credit_History’].mean()
```

Replacing With Mode

Mode is the most frequently occurring value. It is used in the case of categorical features. We can use the ‘fillna’ method for imputing the categorical columns ‘Gender’, ‘Married’, and ‘Self_Employed’.

```
# Replace the missing values for numerical columns with mode
train_df[‘LoanAmount’]=train_df[‘LoanAmount’].fillna(train_df[‘LoanAmount’].mean())

train.df[‘Credit_History’]=train_df[‘Credit_History’].fillna(train_df[‘Credit_History’].mean()
```

---

12. What is A/B testing?

A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment. A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.
A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

In hypothesis testing, we have to make two hypotheses i.e Null hypothesis and the alternative hypothesis. Let's have a look at both.

1. **Null hypothesis or H_0:**

   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 H_0 is " there is no difference in the conversion rate in customers receiving newsletter A and B".

2. Alternative Hypothesis or **H_0**:

The alternative hypothesis 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.

Once we are ready with our null and alternative hypothesis, the next step is to decide the group of customers that will participate in the test. Here we have two groups – **The Control group**, and **the Test (variant) group**.
Randomly selecting the sample from the population is called **random sampling**. It is a technique where each sample in a population has an equal chance of being chosen. Random sampling is important in hypothesis testing because it eliminates sampling bias, and **it's important to eliminate bias because you want the results of your A/B test to be representative of the entire population rather than the sample itself.**
Another important aspect we must take care of is **the Sample size.** It is required that we determine the minimum sample size for our A/B test before conducting it so that we can eliminate **under coverage bias.** It is the bias from sampling too few observations.

One way to perform the test is to calculate **daily conversion rates** for both the treatment and the control groups. Since the conversion rate in a group on a certain day represents a single data point, the sample size is actually the number of days. Thus, we will be testing the difference between the mean of daily conversion rates in each group across the testing period.

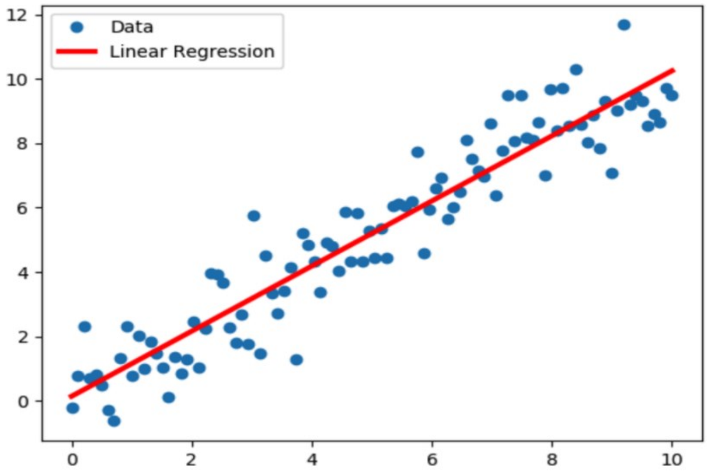
---

13. Is mean imputation of missing data acceptable practice?
Mean imputation. Perhaps the easiest way to impute is to replace each missing value with the mean of the observed values for that variable. Unfortunately, this strategy can severely distort the distribution for this variable, leading to complications with summary measures including, notably, underestimates of the standard deviation. Moreover, mean imputation distorts relationships between variables by "pulling" estimates of the correlation toward zero.

---

14. What is linear regression in statistics?

The most extensively used modelling technique is linear regression, which assumes a linear connection between a dependent variable (Y) and an independent variable (X). It employs a regression line, also known as a best-fit line. The linear connection is defined as Y = c+m*X + e, where 'c' denotes the intercept, 'm' denotes the slope of the line, and 'e' is the error term.The linear regression model can be simple (with only one dependent and one independent variable) or complex (with numerous dependent and independent variables) (with one dependent variable and more than one independent variable).



---

15. What are the various branches of statistics?

The two main branches of statistics are **descriptive statistics** and **inferential statistics**. Both of these are employed in scientific analysis of data and both are equally important for the student of statistics.

Descriptive Statistics

Descriptive statistics deals with the presentation and collection of data. This is usually the first part of a statistical analysis. It is usually not as simple as it sounds, and the statistician needs to be aware of designing experiments, choosing the right focus group and avoid biases that are so easy to creep into 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

Inferential statistics, as the name suggests, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics.
Most predictions of the future and generalizations about a population by studying a smaller sample come under the purview of inferential statistics. Most social sciences experiments deal with studying a small sample population that helps determine how the population in general behaves. By designing the right experiment, the researcher is able to draw conclusions relevant to his study.
While drawing conclusions, one needs to be very careful so as not to draw the wrong or biased conclusions. Even though this appears like a science, 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.
Both descriptive and inferential statistics go hand in hand and one cannot exist without the other. Good scientific methodology needs to be followed in both these steps of statistical analysis and both these branches of statistics are equally important for a researcher.
