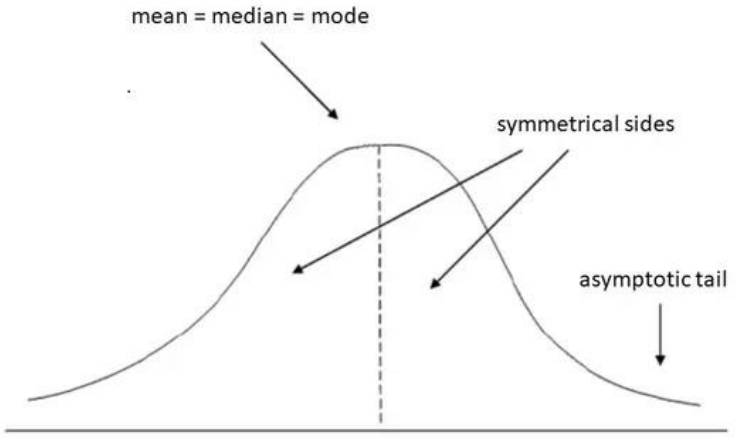
STATISTICS WORKSHEET-1

QUESTIONS	ANSWERS
1	a
2	a
3	b
4	d
5	c
6	b
7	b
8	a
9	c

Answer 10:

Normal Distribution : It is also known as called Gaussian distribution. It is a continuous probability distribution which is symmetrical on both sides of the mean, so the right side of the center is a mirror image of the left side. It appears like bell curve (diagram below). Here mean=mode=median. The total area under the curve sums to 1. The tails are asymptotic i.e they approach but never quite meet the horizon .



Answer 11.

Handling missing data & imputation techniques:

Missing values occurs when no data is stored for a variable(feature) in an observation. It appears as ?,N.A,0,just blank cell.

There are many methods to handle the missing data which is as follows.

1.First one should **check with the data collection** source and can go back and find what should be the acutal value.

2.**Dropping the missing value** either by dropping the whole variable or dropping the single data entry with missing value. When there are less observations with missing value, then dropping the particular entry is best choice.

3.**Replacing the missing value** is best as no data is wasted either by average(mean) of the similar datapoints or replacing with most frequently occurring value.

Imputation techniques:

Imputation with constant value: lt replaces the missing values with either zero or any constant value.

Imputation using Statistics:(Mean or median or most frequent(mode):

Mean" will replace missing values with the mean value from the rest of the column. It is preferred if data is numeric and not skewed.

Median" will replace missing values using the median in each column. It is preferred if data is numeric and skewed (when outliers are present).

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

Answer 12.

A/B Testing:

A/B testing is a statistical way of comparing two or more versions such as Version A or Version B to determine not only which version performs better but also to understand if a difference between the two version is statistically significant. It includes application of statistical hypothesis testing or "two-sample hypothesis testing" as used in the field of statistics. Large social media sites like LinkedIn, Facebook, and Instagram use A/B testing to make user experiences more successful and as a way to streamline their services

Answer 13.

Linear regression: -

Linear regression is a statistical tool to define the relationship between one or more independent variable and dependent variable. It describes the relationship between two or more variable. It is expressed as y=B0+B1x+e where

Y: is the dependent variable, the variable you want to predict
X:is the independent variable, the variable we are using to predict y
B0:is the intercept
B1:is the slope/co-efficient
e: the regression residual error

Answers 14.

The mean imputation of missing data is not an acceptable practice as it generally works on column level, misses correlations between features and can't be use on categorical features (though imputing with most frequent works here) and also Outlier's data points will have a significant impact on the mean. It decrease the variance of the data while increasing bias. It doesn't give very accurate result.

Answer 15
There are mainly two branches of statistics
1. Descriptive Statistics 2. Inferential Statistics

Descriptive statistics: As the name suggest, it describes and summarizes the characteristics or feature of the data using numbers and graphs. Example marks in class, weight of the student. It can be used to describe either the entire population or the individual data. It uses Central tendency, Dispersion.

Inferential Statistics: Here we use sample data to make an inference about the population. It usually occurs when the data is big. Hypothesis testing and confidence intervals are the applications of the statistical inference. It is a method of making decisions about the parameters of a population, based on random sampling. It is used in wide range of application like share market, fraud detection.