STATISTICS WORKSHEET-1

1.a

2.a

3.b

4.d

5.c

6.b

7.b

8.a

9.c

10. Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.

Normal distributions are symmetrical, but not all symmetrical distributions are normal.

Many naturally-occurring phenomena tend to approximate the normal distribution.
In finance, most pricing distributions are not, however, perfectly normal

11. Real-world data is messy and usually holds a lot of missing values. Missing data can skew anything for data scientists and, A data scientist doesn't want to design biased estimates that point to invalid results. Behind, any analysis is only as great as the data**.** Missing data appear when no value is available in one or more variables of an individual**.** Due to Missing data, the statistical power of the analysis can reduce, which can impact the validity of the results.

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:

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.

12. When conducting a test, you are making an assumption about a population parameter and a numerical value. This is your hypothesis (corresponds to Step 9 of conversion optimization system).

In a simplified example, your hypothesis could look like this:

By adding reviews on the product pages, you will increase social proof and trust and confidence in the product, thus increase the number of micro conversions on the page resulting in an overall increase in conversion rates.

This is your hypothesis in "normal words.". But how would it look like in statistics?

In statistics your hypothesis breaks down into:

- Null hypothesis
- Alternative hypothesis

The null hypothesis states the default position to be tested or the situation as it is (assumed to be) now, i.e. the status quo.

The alternative hypothesis challenges the status quo (the null hypothesis) and is basically a hypothesis that the researcher (you) believes to be true. The alternative hypothesis is what you might hope that your A/B test will prove to be true.

Let’s look at an example:

Conversion rate on product pages of Acme.Inc is equal to 8%. One of the problems that they revealed during the heuristic evaluation was there were simply no product reviews on the product pages. They believe that adding reviews would help visitors make a decision thus increasing flow to cart page and conversions.

The null hypothesis here would be: *no reviews generates a conversion rate equal to 8% (the status quo)*

The alternative hypothesis here would be: *adding reviews will cause conversion rate to be more than 8%.*

Now, the researcher, namely you, will have to collect enough evidence to reject null hypothesis and prove that the alternative hypothesis is true.

13. The process of replacing null values in a data collection with the data’s mean is known as mean imputation.

Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does.

Second, mean imputation 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.

14. Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a “least squares”

method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).

You can perform the linear regression method in a variety of programs and environments, including:

- R linear regression
- MATLAB linear regression
- Sklearn linear regression
- Linear regression Python
- Excel linear regression

15. There are three real branches of statistics: **data collection, descriptive statistics and inferential statistics**.

Data collection:= is all about how the actual data is collected. For the most part, this needn't concern us too much in terms of the mathematics (we just work with what we are given), but there are significant issues to consider when actually collecting data.
For data such as marks in a class test, this is fairly straightforward. Each student has a defined mark associated with them, so the marks are simply collected together to make the data set.
Sometimes, data is harder to collect. Counting the number of bees in a colony isn't easy, because they move and fly around; you may have to approximate in such cases.
Also, if you are collecting data, you need to be careful where you get it from. For example, suppose you want to conduct a poll on who people plan to vote for in an election. You can't realistically ask everyone in the whole country (the *population*), so you have to choose a representative *sample* of people. This isn't as easy as it sounds. In the mid 20th century, for example, polls were sometimes carried out by randomly calling people in the telephone directory. This sounds representative, but in those days only the richer people had telephones, and so you were asking only a particular section of society, who might well be more inclined to vote for one party rather than other. The same issue may apply with doing a poll by email today.
B.**Descriptive statistics**:- is the part of statistics that deals with presenting the data we have. This can take two basic forms – presenting aspects of the data either visually (via graphs, charts, etc.) or numerically (via averages and so on).

Common visual techniques that we shall discuss in Chapter 2 include graphs, bar charts, pie charts and more, but we shall focus mainly on numerical techniques such as averages and spreads.
The basic aim of descriptive statistics is to 'present the data' in an understandable way. If you simply write down every piece of data, it means little to someone who sees it; it needs to be summarised. Imagine if, on the TV news, they listed on the screen the votes of every single person interviewed by a polling company; it would just be a huge list of parties, and you couldn't arrive at any meaningful conclusion. Instead, you are presented with visual charts (a bar chart, say) to give, perhaps, the percentage of the vote each party has. In the 2010 General Election almost 30 million people voted. If each vote was simply written down and displayed, one after the other, you'd be totally lost; what happens is that a summary of votes is presented (for example as percentages: Conservative 36%, Labour 29%, Liberal Democrat 23%, Others 12%). This is an example of descriptive statistics – 'describing' or 'summarising' the overall data for people to understand.

C. ***Inferential statistics*** is the aspect that deals with making conclusions about the data.This is quite a wide area; essentially you are asking 'What is this data telling us, and what should we do?'
For example, a council might be considering altering the speed limit on a main road, after a number of accidents. They might do this by surveying the speeds of cars (data collection) and then arrive at a conclusion as to whether the speed limit needs to be lowered (if, for example, a number of cars are driving too fast). Note, though, that this may not be the case; everyone might be driving at a perfectly acceptable speed, and the accidents are down to something other than speed (a blind spot or a pothole, for example). This is inferential statistics: take the data you have and make an 'inference' or 'conclusion' from it. We shall see much more of this later when we discuss things such as *hypothesis testing*, where we test to see whether the data supports a belief that we have.

,