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. In graphical form, the normal distribution appears as a "bell curve". In a normal distribution, the mean is zero and the standard deviation is 1.

11. Whenever the data is missing in any data set even a single missing value can greatly affect your model and it won’t give the desired result. In machine learning, there are some imputation techniques that we use to fill in the missing values, some of them are Simple Imputer, KNN Imputer, and Iterative Imputer. Simple imputer uses the fillna().mean() to fill the missing values. The KNN imputer works similarly to the KNN model which uses the nearest neighbors to calculate the missing value , lastly the Iterative Imputer works like the linear regression model and tries to fill the missing values by iterating and using rest of the data is features and label.

12. A/B testing, also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable (web page, page element, etc.) are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drives business metrics. Essentially, A/B testing eliminates all the guesswork and enables experience optimizers to make data-backed decisions. In A/B testing, A refers to ‘control’ or the original testing variable. Whereas B refers to ‘variation’ or a new version of the original testing variable. The version that moves your business metric(s) in the positive direction is known as the ‘winner.’ Implementing the changes of this winning variation on your tested page(s) / element(s) can help optimize your website and increase business ROI.

13. Generally mean imputation is not advisable as it has 2 major flows. Firstly Mean imputation does not preserve the relationships among variables. True, imputing the

mean preserves the mean of the observed data. So if the data are missing completely at random, the estimate of the mean remains unbiased. That’s a good thing. Plus, by imputing the mean, you are able to keep your sample size up to the full sample size. That’s good too. This is the original logic involved in mean imputation. If all you are doing is estimating means (which is rarely the point of research studies), and if the data are missing completely at random, mean imputation will not bias your parameter estimate. It will still bias your standard error. Since most research studies are interested in the relationship among variables, mean imputation is not a good solution. Secondly Mean Imputation Leads to An Underestimate of Standard Errors. A second reason is applies to any type of single imputation. Any statistic that uses the imputed data will have a standard error that’s too low. In other words, yes, you get the same mean from mean-imputed data that you would have gotten without the imputations. And yes, there are circumstances where the mean is unbiased. Even so, the standard error of that mean will be too small. Because the imputations are themselves estimates, there is some error associated with them. But your statistical software doesn’t know that. It treats it as real data. Ultimately, because your standard errors are too low, so are your p-values. Now you’re making Type I errors without realizing it. That’s not good.

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).

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 whom people plan to vote for in an election. You can't really 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 another. The same issue may apply to doing a poll by email today. So there are issues in the collection of the data; you need to make sure that the data has been collected fairly before you go on to deal with it, and try to present it and make conclusions.The words population and sample are used in general in statistics. The population is the entire set of data, and a sample is a (hopefully representative) subset of the population – so just 'some of' the data values. Why would we need a sample? As indicated above, it's usually unrealistic to get data from the entire population (it would incredibly expensive and time-consuming, and also the population may be changing as you collect the data), so often we need to simply take a sample.

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). 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 ever 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.

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.