**1. Bernoulli random variables take (only) the values 1 and 0.   
a) True b) False   
Answer = True**

**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   
Answer = a) Central Limir Theorem**

**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   
Answer = b) Modeling bounded count data**

**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   
Answer = d) All of the mentioned**

**5. \_\_\_\_\_\_ random variables are used to model rates.   
a) Empirical b) Binomial c) Poisson d) All of the mentioned   
Answer = c) Poisson**

**6. Usually replacing the standard error by its estimated value does change the CLT.   
a) True b) False   
Answer = 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   
Answer = b) Hypothesis**

**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   
Answer = a) 0**

**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  
Answer = c) 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?   
Answer -   
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 graph form, normal distribution will appear as a bell curve. The mean, mode and median are all equal. The curve is symmetric at the centre around the mean. Exactly half of the values are to the left of centre and exactly half the values are to the right. The total area under the curve is 1.**

**11. How do you handle missing data? What imputation techniques do you recommend?   
Answer -   
In a data set if there are missing values present, we can either remove the observations having missing data or we can use imputation technique. If there are a smaller number of missing data present in our data set then it is better to remove the observations having missing data (i.e., 15-20% are missing data in a feature), however if the number of missing data is huge, we should try to impute with some technique. For continuous data we can use median or mean imputation and for categorical data we can use mode imputation. We can also use KNN imputation, forward fill, backward fill. Imputation a missing value depends upon the type of data we have; we can also impute the missing values with the mean with respect to any class of a column.**

**12. What is A/B testing?   
Answer -  
A/B testing is one of the main objectives in data science, it is a way to compare the two versions of a variable to find out which performs better in a controlled environment.  
For eg, lets put a question here.  
A company has stated that their straw machine makes straw that are 4mm in diameter. A worker believes that the machine no longer makes the straws of this size and to check this a sample of 100 straws is taken to test if it is true or not?  
Now in above question there a situation of change where the company needs to decides if the machine is working properly or not.  
To find out this A/B testing is done. In A/B testing we usually test the hypothesis. Hypothesis is an assumption about a parameter of the population.  
Steps of A/B testing:  
1- Create Hypothesis. – here we will decide 2 assumption/hypothesis. First is that the machine is working fine and making straws of 4mm only (mean of the diameter of straws is 4mm). Second is that the machine is not working fine (mean of the diameter of the straws is not equal to 4mm.)  
2 – Create 2 groups, in this we will randomly distribute 100 samples in 2 groups.  
3 – Now, For rejecting our null hypothesis we have to prove the Statistical significance of our test.   
4 – Significance test includes these 3 terms which are needed for calculation: a) significance level, b) p-value, c)Confidence level.  
5 – Now we will compute the p-value using the** [**formula**](https://www.google.com/search?safe=active&sxsrf=ALeKk03YmWJsHEXlKguplOW879kev1vCBw%3A1615203916073&ei=TA5GYL6CBJGH4-EPif2g2Ak&q=formula+of+t+test&oq=formula+of+t+test&gs_lcp=Cgdnd3Mtd2l6EAMyBwgAELADEEMyBwgAELADEEMyBwgAELADEEMyBwgAELADEEMyBwgAELADEEMyBwgAELADEEMyBwgAEEcQsAMyBwgAEEcQsAMyBwgAEEcQsAMyBwgAEEcQsANQ7FBY21FgvFJoAXACeACAAaABiAGQApIBAzAuMpgBAKABAaoBB2d3cy13aXrIAQrAAQE&sclient=gws-wiz&ved=0ahUKEwi-2KC9z6DvAhWRwzgGHYk-CJsQ4dUDCA4&uact=5)**.  
6 – Now if our p-value is > significance value then we can accept the null hypothesis.  
  
All this procedure is A/B testing to check what will be the performance if there is any change.**

**13. Is mean imputation of missing data acceptable practice?   
Answer -   
Mean imputation is not always a good practice. For eg. Suppose we have a data column where all the values are descrete and not categorical, let’s say age. If there are any missing values present in the column and if we impute them with the mean of that column, then it might gives us a float (decimal) value which is not acceptable for age. Same goes if we have a column with number of days. A day cannot be in decimal values it is a whole discrete values, that is why mean imputation of missing data is not always acceptable. Mean also reduced the variance of data, which leads to introducing a bias data.**

**14. What is linear regression in statistics?   
Answer -   
Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an independent variable, and the other is considered to be a dependent variable.**

**15. What are the various branches of statistics?  
Answer-  
There are 2 ranches of statistics. 1st is Descriptive statistics – It includes the understanding of raw data and covering that raw data into a better form. There are 2 methods of Descriptive statistics. A) Numerical = It consists of Mean, median, mode, standard deviation, etc which all comes in CLT.  
B) Graphical = it consists of the visualisation method of the data.  
  
2nd is Inferential statistics- In this we use sample data to take some decision or measures to improve the business estimates, decision making. This is like the A/B testing.**