Statistics worksheet 1:

1) a- True

2) a- Central Limit Theorem

3) c- Modelling contingency tables

4) c- The square of a standard normal random variable follows what is called chi-squared distribution

5) c- Poisson

6) a- True

7) b- Hypothesis

8) a- 0

9) c- Outliers cannot conform to the regression relationship

10) *Normal distribution:*  
Normal distribution, or Gaussian distribution, is a type of distribution that can be easily identified by its bell-shaped curve. When plotted in a graph, the data creates a symmetrical bell shape where the mean, median and mode are at the center of the distribution. You can use the standard deviation to help identify the dispersion of the data.

11) *‘How do you handle missing data? What imputation techniques do you recommend?’:*  
I believe the imputation method a data scientist chooses is relative to the nature of the dataset as well as the objective/goal they are trying to achieve. By this, I think it is important to consider the impact of removing or manipulating data as you don’t want to distort your findings. During my practise, I have deleted rows where the missing data cannot be replaced and has no meaningful impact on the dataset. I have also used mean imputation to replace a missing value with an average where appropriate. We have also explored KNN imputation.

12) ‘*What is A/B testing?’*  
A/B testing, or split testing, is a method of comparing the sample of a dataset to the population data to determine whether there is a significant statistical difference in performance against a given metric.

13) ‘*Is mean imputation of missing data acceptable practice?’*

Mean imputation can be useful as it means that the sample size isn’t reduced and is an easy method to implement. However, it can distort the data and lead to inaccurate results. Therefore, although it can be useful, I would consider whether there is a more suitable imputation method such as KNN, which may yield more accurate conclusions.

14) ‘*What is linear regression in statistics?*’   
Linear regression is used to show the relationship between a dependent variable (label) and one or more independent variables. Linear regression aims to find the best-fitting linear relationship that describes how changes in the independent variables are associated with changes in the dependent variable. Eg. Number of bedrooms (independent variable) has a positive linear relationship to the value of the house (dependent variable).

15) ‘*What are the various branches of statistics?*’

The two main branches of statistics are descriptive and inferential statistics.  
**Descriptive:** Descriptive statistics uses measures such as mean, mode, median, standard deviation, variance etc. to present data and show data patterns.

**Inferential**: Inferential statistics uses methods such as regression analysis and hypothesis testing to make predictions about the population data based on a sample.