STATISTICS

1) Bernoulli random variables take (only) the values 1 and 0.

Answer: (a)

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?

Answer: (a)

3) Which of the following is incorrect with respect to the use of Poisson distribution?

Answer: (b)

4) Point out the correct statement.

Answer: (d)

5) _____ random variables are used to model rates.

Answer: (c)

6) Usually replacing the standard error by its estimated value does change the CLT.

Answer: (b) False

7) Which of the following testing is concerned with making decisions using data?

Answer: (b)Hypothesis

8) Normalized data are centered at _____and have units equal to standard deviations of the original data.

Answer: (a)

9) Which of the following statement is incorrect with respect to outliers?

Answer: (c)

10) What do you understand by the term Normal Distribution?

Answer: In probability theory and statistics, the Normal Distribution, also called the Gaussian Distribution, is the most significant continuous probability distribution. Sometimes it is also called a bell curve. A large number of random variables are either nearly or exactly represented by the normal distribution, in every physical science and economics. Furthermore, it can be used to approximate other probability distributions, therefore supporting the usage of the word “normal” as in about the one, mostly used.

Also, the normal distribution is defined by the probability density function for a continuous random variable in a system.

And also, the normal distribution is a core concept in statistics, the backbone of data science. While performing exploratory data analysis, we first explore the data and aim to find its probability distribution, right? And guess what – the most common probability distribution is **Normal Distribution.**

11) How do you handle missing data? What imputation techniques do you recommend?

Answer: Missing data can be dealt with in a variety of ways. I believe the most common reaction is to ignore it. Choosing to make no decision, on the other hand, indicates that your statistical programmes will make the decision for you.

Your application will remove things in a listwise sequence most of the time. Depending on why and how much data is gone, listwise deletion may or may not be a good idea.

Another common strategy among those who pay attention is imputation. Imputation is the process of substituting an estimate for missing values and analysing the entire data set as if the imputed values were the true observed values.

Model-based imputation

This is an interesting way of handling missing data. We take feature *f1* as the class and all the remaining columns as features. Then we train our data with any model and predict the missing values.

Or

Hot deck imputation

A value picked at random from a sample member who has comparable values on other variables. To put it another way, select all the sample participants who are comparable on other factors, then choose one of their missing variable values at random.

One benefit is that you are limited to just feasible values. In other words, if age is only allowed to be between 5 and 10 in your research, you will always obtain a value between 5 and 10. Another factor is the random element, which introduces some variation. For exact standard errors, this is crucial.

12) What is A/B testing?

Answer: An AB test is an example of statistical hypothesis testing, a process whereby a hypothesis is made about the relationship between two data sets and those data sets are then compared against each other to determine if there is a statistically significant relationship or not.

13) Is mean imputation of missing data acceptable practice?

Answer: Mean imputation so simple and yet so dangerous. Perhaps that's a bit dramatic, but mean imputation (also called mean substitution) really ought to be a last resort.

It's a popular solution to missing data, despite its drawbacks. Mainly because it's easy. It can be really painful; to lose a large part of the sample so carefully collected, only to have little power. But that doesn't make it a good solution, and it may not help you find relationships with strong parameter estimates. Even if they exist in the population. On the other hand, there are many alternatives to mean imputation that provide much more accurate estimates and standard errors, so there really is no excuse to use it.

14) What is linear regression in statistics?

Answer: Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables, in particular, are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables.

15) What are the various branches of statistics?

Answer: Descriptive and Inferential

Descriptive

Central tendency

1)Mean

2)Median

3)Mode

Dispersion of Data

1) Range
2) Percentile
3) Standard deviation
4) SKEW
5) Variant

Inferential Statistics

1) Zscore (Value of outlier)
2) Hypothesis testing
   a) T-Test
   b) Chi-Square test
   c) Student T test
   d) Paired T – test
   e) Pearson co-relation test
   f) Spearman co-relation test