Statistic work sheet 1

ANSWERS

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

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 Limit 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 lognormal

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: All of the mentioned

5. _____ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

Answer: Poisson

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

a) True

b) False

Answer: False

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

a) Probability

b) Hypothesis

c) Causal

d) None of the mentioned

Answer : Hypothesis

8. 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 : 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

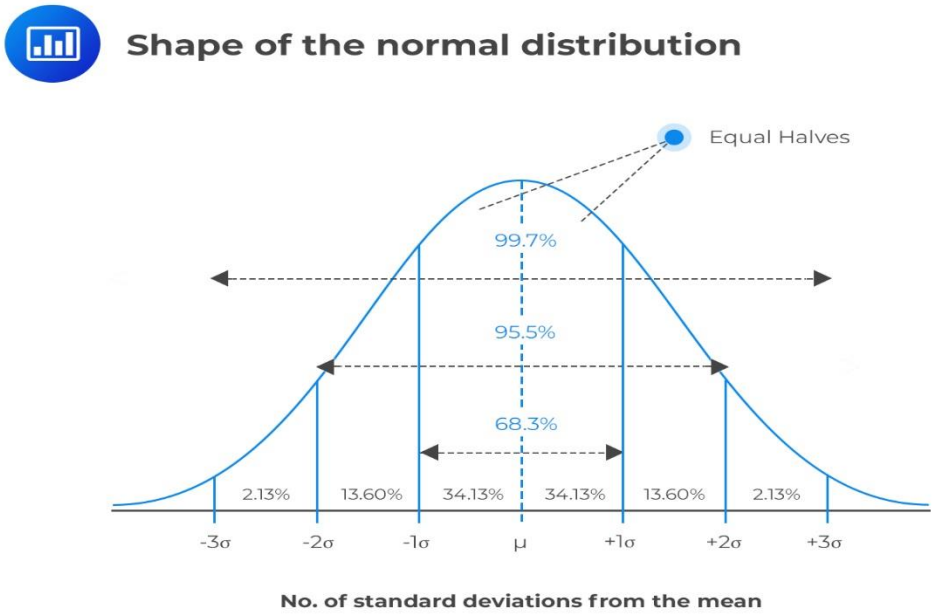
d) None of the mentioned

Answer: Outliers cannot conform to the regression relationship

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 is a bell-shaped frequency distribution curve which helps describe all the possible values a random variable can take within a given range with most of the distribution area is in the middle and few are in the tails, at the extremes. This distribution has two key parameters: the mean (μ) and the standard deviation (σ) which plays key role in assets return calculation and in risk management strategy.



The above figure shows that the statistical normal distribution is a bell-shaped curve. The range of possible outcomes of this distribution is the whole real numbers lying between -∞ to +∞. The tails of the bell curve extend on both sides of the chart (+/-) without limits.

Approximately 68% of all observation fall within +/- one standard deviation(σ)

Approximately 95% of all observation fall within +/- two standard deviations (σ)

Approximately 99% of all observation fall within +/- three standard deviations (σ)

It has a skewness of zero (symmetry of a distribution). If the distribution of data is

asymmetric, then the distribution is uneven if the data set has skewness greater than zero or positive skewness. Then, the right tail of the distribution is more prolonged than the left, and for negative skewness (less than zero) left tail will be longer than the right tail. It has a deviation of 3 (measures peakedness of a distribution), which indicates distribution is neither too peaked nor too thin tails. If the deviation is more than three than distribution is more peaked with fatter tails, and if the kurtosis is less than three, then it has thin tails, and the peak point is lower than the normal distribution.

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

Answer: Missing data is a huge problem for data analysis because it distorts findings. It's difficult to be fully confident in the insights when you know that some entries are missing values. Hence, why they must be addressed. According to data scientists, there are three types of missing data. These are Missing Completely at Random (MCAR) – when data is completely missing at random across the dataset with no discernable pattern. There is also Missing At Random (MAR) – when data is not missing randomly, but only within sub-samples of data. Finally, there is Not Missing at Random (NMAR), when there is a noticeable trend in the way data is missing. Data scientists use two data imputation techniques to handle missing data: Average imputation and common-point imputation. Average imputation uses the average value of the responses from other data entries to fill out missing values. However, a word of caution when using this method – it can artificially reduce the variability of the dataset. Common-point imputation, on the other hand, is when the data scientists utilise the middle point or the most commonly chosen value. For example, on a five-point scale, the substitute value will be 3. Something to keep in mind when utilising this method is the three types of middle values: mean, median and mode, which is valid for numerical data (it should be noted that for nonnumerical data only the median and mean are relevant).

12. What is A/B testing?

Answer: A/B testing (also known as split testing or bucket testing) is a method of comparing two versions of a webpage or app against each other to determine which one performs better. A/B testing is essentially an experiment where two or more variants of a page are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal. Running an A/B test that directly compares a variation against a current experience lets you ask focused questions about changes to your website or app and then collect data about the impact of that change. Testing takes the guesswork out of website optimization and enables data-informed decisions that shift business conversations from "we think" to "we know." By measuring the impact that changes have on your metrics, you can ensure that every change produces positive results.

13. Is mean imputation of missing data acceptable practice?

Answer: Mean imputation is the practice of replacing null values in a data set with the mean of the data.Mean imputation is generally bad practice because it doesn't take into account feature correlation. For example, imagine we have a table showing age and fitness score and imagine that an eighty-year-old has a missing fitness score. If we took the average fitness scorefrom an age range of 15 to 80, then the eighty-year-old will appear to have a much higher fitness score that he actually should.Second, mean imputation reduces the variance of the data and increases bias in our data. This leads to a less accurate model and a narrower confidence interval due to a smaller variance.

14. What is linear regression in statistics?

Answer: Linear regression is an algorithm used to predict, or visualize, a relationship between two different features/variables. In linear regression tasks, there are two kinds of variables being examined: the dependent variable and the independent variable. The independent variable is the variable that stands by itself, not impacted by the other variable. As the independent variable is adjusted, the levels of the dependent variable will fluctuate. The dependent variable is the variable that is being studied, and it is what the regression model solves for/attempts to predict. In linear regression tasks, every observation/instance is comprised of both the dependent variable value and the independent variable value.

15. What are the various branches of statistics?

Answer:

- The two main branches of statistics are descriptive statistics and inferential statistics. Both of these are employed in scientific analysis of data and both are equally important for the student of statistics.
- Descriptive statistics deals with the presentation and collection of data. This is usually the first part of a statistical analysis. It is usually not as simple as it sounds, and the statistician needs to be aware of designing experiments, choosing the right focus group and avoid biases that are so easy to creep into the experiment.
- Different areas of study require different kinds of analysis using descriptive statistics. For example, a physicist studying turbulence in the laboratory needs the average quantities that vary over small intervals of time. The nature of this problem requires that physical quantities be averaged from a host of data collected through the experiment.

- Inferential statistics, as the name suggests, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics.
- Most predictions of the future and generalizations about a population by studying a smaller sample come under the preview of inferential statistics. Most social sciences experiments deal with studying a small sample population that helps determine how the population in general behaves. By designing the right experiment, the researcher is able to draw conclusions relevant to his study.
- While drawing conclusions, one needs to be very careful so as not to draw the wrong or biased conclusions. Even though this appears like a science, there are ways in which one can manipulate studies and results through various means. For example, data dredging is increasingly becoming a problem as computers hold loads of information and it is easy, either intentionally or unintentionally, to use the wrong inferential methods.
- Both descriptive and inferential statistics go hand in hand and one cannot exist without the other. Good scientific methodology needs to be followed in both these steps of statistical analysis and both these branches of statistics are equally important for a researcher