STATISTICS WORKSHEET- 6

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

1. Which of the following can be considered as random variable? a) The outcome from the roll of a die b) The outcome of flip of a coin c) The outcome of exam d) All of the mentioned

Ans = d) All of the mentioned

2. Which of the following random variable that take on only a countable number of possibilities? a) Discrete b) Non Discrete c) Continuous d) All of the mentioned

Ans = a) Discrete

3. Which of the following function is associated with a continuous random variable? a) pdf b) pmv c) pmf d) all of the mentioned

Ans = a) pdf

4. The expected value or _____ of a random variable is the center of its distribution. a) mode b) median c) mean d) bayesian inference

Ans = c) mean

5. Which of the following of a random variable is not a measure of spread? a) variance b) standard deviation c) empirical mean d) all of the mentioned

Ans = a) variance

6. The _____ of the Chi-squared distribution is twice the degrees of freedom. a) variance b) standard deviation c) mode d) none of the mentioned

Ans = a) variance

7. The beta distribution is the default prior for parameters between _____ a) 0 and 10 b) 1 and 2 c) 0 and 1 d) None of the mentioned

Ans = c) 0 and 1

8. Which of the following tool is used for constructing confidence intervals and calculating standard errors for difficult statistics? a) baggyer b) bootstrap c) jacknife d) none of the mentioned

Ans = b) bootstrap

9. Data that summarize all observations in a category are called _____ data. a) frequency b) summarized c) raw d) none of the mentioned

Ans = b) summarized

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

10. What is the difference between a boxplot and histogram?

Ans = Histograms and box plots are graphical representations for the frequency of numeric data values. They aim to describe the data and explore the central tendency and variability before using advanced statistical analysis techniques. Both histograms and box plots allow to visually assess the central tendency, the amount of variation in the data as well as the presence of gaps, outliers or unusual data points.

Both histograms and box plots are used to explore and present the data in an easy and understandable manner. Histograms are preferred to determine the underlying probability distribution of a data. Box plots on the other hand are more useful when comparing between several data sets. They are less detailed than histograms and take up less space.

11. How to select metrics?

Ans = metrics Performance metrics are defined as information and project-specific data used to characterize and assess an organization's quality, capabilities, and skills. Performance metrics are defined differently in every industry and can change based on a company's services and products. Common performance metrics include sales, return on investment, customer satisfaction, industry and consumer reviews, and a company's reputation with its consumers. The value of optimizing your customer experience is clear to most brands and marketers. Increasing loyalty, reducing customer service costs and increasing revenue growth from retained customers are three big reasons, in addition to many others.

While it may be easier to come to the conclusion that you *should* measure your customer experience efforts, a more difficult task is to decide exactly *what* you should measure. To do this, you need to determine what you want your measurement to achieve.

We won’t discuss specific metrics in this article, because that can be completely dependent on the type of business you belong to, as well as the nuances of your customer experience. But as you’ll see, some general rules apply, regardless of the specifics.

12. How do you assess the statistical significance of an insight?

Ans = Statistical significance is a measure of reliability in the result of an analysis that allows you to be confident in your decision making. Statistical significance is the likelihood that a relationship between two or more variables in an analysis is not purely coincidental, but is actually caused by another factor. In other words, statistical significance is a way of mathematically proving that a certain statistic is reliable. In the real world, businesses use statistical significance to understand how strongly the results of their surveys, polls, should influence their decisions.

Statistical significance is important because it gives you confidence in your analysis and its resulting insights. There is no business value in taking actions on insights that are misleading or incorrect, and what’s more, taking action on misleading information can also keep you from investing resources correctly.

13. Give examples of data that doesnot have a Gaussian distribution, nor log-normal.

Ans = We will briefly look at the definition of the log-normal and then go onto calculate the distribution’s parameters μ and σ from simple data. We will then have a look at how to calculate the mean, mode, median and variance from this probability

distribution. The log-normal distribution is a right skewed continuous probability distribution, meaning it has a long tail towards the right. It is used for modelling various natural phenomena such as income distributions, the length of chess games or the time to repair a maintainable system and more. The name of the “log-normal” distribution reveals that it relates to logarithms as well as the normal distribution. How? Let’s say your data fits a log-normal distribution. If you then take the logarithm of all your data points, the newly transformed points will now fit a normal distribution. This simply means that when you take the log of your log-normal data you end up with a normal distribution. See figure below.

14. Give an example where the median is a better measure than the mean.

Ans = Measures of central tendency are summary statistics that represent the center point or typical value of a dataset. Examples of these measures include the mean, median, and mode. These statistics indicate where most values in a distribution fall and are also referred to as the central location of a distribution. You can think of central tendency as the propensity for data points to cluster around a middle value.

In statistics, the mean, median, and mode are the three most common measures of central tendency. Each one calculates the central point using a different method. Choosing the best measure of central tendency depends on the type of data you have. In this post, I explore the mean, median, and mode as measures of central tendency, show you how to calculate them, and how to determine which one is best for your data.

Most articles about the mean, median, and mode focus on how you calculate these measures of central tendency. I'll certainly to that, but I'm going to start with a slightly different approach. My philosophy throughout my blog is to help you intuitively grasp statistics by focusing on concepts. Consequently, I'm going to start by illustrating the central point of several datasets graphically—so you understand the goal. Then, we'll move on to choosing the best measure of central tendency for your data and the calculations. The three distributions below represent different data conditions. In each distribution, look for the region where the most common values fall. Even though the shapes and type of data are different, you can find that central tendency. That's the area in the distribution where the most common values are located. These examples cover the mean, median, and mode.

15. What is the Likelihood?

Ans = Likelihood is a confusing term. Likelihood is not a probability, but is proportional to a probability; the two terms can't be used interchangeably. In this post, we will be dissecting likelihood as a concept and understand it's importance in machine learning. Let us understand likelihood and how it is different from a probability distribution with an imaginary city, Databerg (a cringe name, but bear with me). Let's also imagine we have access to the pricing data of all houses in this city. I don't know exactly how this distribution looks since Databerg isn't a real city, but intuitively I'd say we would notice many houses that are moderately priced and a few houses that are very expensive. If one were to plot a distribution of these prices, it might look something like this.