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

   Answer: 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

   Answer: 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

   Answer: 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

   Answer: 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

   Answer: 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

   Answer: 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

Answer: 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

Answer: 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

Answer: 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?

Answer:

Comparing Histograms and Box Plots

Although histograms and box plots are collectively part of the chart aid category, they do represent very different types of charts. Both charts effectively represent different data sets; however, in certain situations, one chart may be superior to the other in achieving the goal of identifying **variances among data**. The type of chart aid chosen depends on the type of data collected, rough analysis of data trends, and project goals.

A histogram is highly useful when wide variances exist among the observed frequencies for a particular data set. As seen in the two graphs to the left, the histogram shows that there are three peaks within the data, indicating it is tri-modal (three commonly recurring groups of numbers). This is important because to improve processes, it is critical to understand what is causing these three modes. Had this data simply been graphed using a box plot, the values would average one another out, causing the distribution to look roughly normal.

Another instance when a histogram is preferable over a box plot is when there is very little variance among the observed frequencies. The histogram displayed to the right shows that there is little variance across the groups of data; however, when the same data points are graphed on a box plot, the distribution looks roughly normal with a high portion of the values falling below six.

The final set of graphs shows how a box plot can be more useful than a histogram. This occurs when there is moderate variation among the observed frequencies, which causes the histogram to look ragged and non-symmetrical due to the way the data is grouped. This may lead one to assume the data is slightly skewed. However, when a box plot is used to graph the same data points, the chart indicates a perfect normal distribution.

11. How to select metrics?

Answer:

Choosing the right metrics

You may be thinking: what the heck is a 'good' metric? You see, all metrics are certainly not created equal. We've all been guilty of slipping into the trap of vanity metrics at some point, but we'll try and help you avoid that mistake.
'Good' can be broadly defined as metrics that show if you're achieving your objectives (the ones you prioritized before). Fundamentally, good metrics have three characteristics.

1.Good metrics are important to your company growth and objectives. Your key metrics should always be closely tied to your primary objective. A good metric example might be month-on-month revenue growth or LTV:CAC ratio. 'Important' is somewhat subjective since growth for one company may be centered around revenue while another company may focus more on user growth. The key point is to choose metrics that clearly indicate where you are now in relation to your goals.

2.Good metrics can be improved. Good metrics measure progress, which means there needs to be room for improvement. For example, reducing churn by 0.8% or increasing your activation rate by 3%. One exception to this might be customer satisfaction - if you're already at 100%, your team will be focused on maintaining that level instead of improving it.

3.Good metrics inspire action. When your metrics are important and can be improved, you and your team will immediately know what to do or what questions to ask. For example, why has our conversion rate dropped? Did we make site changes or test a new acquisition channel? Why is churn increasing? By asking questions you can determine possible causes and work to resolve them right away.

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

Answer:

Calculating the statistical significance is rather extensive if you calculate it by hand and this is why it's typically calculated using a calculator. When you calculate it by hand, however, it will help you more fully understand the concept. Here are the steps for calculating statistical significance:
1. Create a null hypothesis.
2. Create an alternative hypothesis.

3. Determine the significance level.
4. Decide on the type of test you'll use.
5. Perform a power analysis to find out your sample size.
6. Calculate the standard deviation.
7. Use the standard error formula.
8. Determine the t-score.
9. Find the degrees of freedom.
10. Use a t-table.

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

Answer:

Exponential distributions do not have a log-normal distribution or a Gaussian distribution. In fact, any type of data that is categorical will not have these distributions as well. Example: **Duration of a phone car, time until the next earthquake**, etc.

Thus we can apply some non-Gaussian distributions, e.g., **the beta distribution, the Dirichlet distribution**, to model the distribution of this type of data. The choice of a suitable distribution is favorable for modeling efficiency.

A log-normal distribution is a continuous distribution of random variable y whose natural logarithm is normally distributed. For example, **if random variable y = exp { y } has log-normal distribution then x = log (y) has normal distribution**.

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

Answer:

When you have a symmetrical distribution for continuous data, the mean, median, and mode is equal. In this case, analysts tend to use the mean because it includes all of the data in the calculations. However, if you have a skewed distribution, the median is often the best measure of central tendency.

When you have ordinal data, the median or mode is usually the best choice. For categorical data, you have to use the mode.
In cases where you are deciding between the mean and median as the better measure of central tendency, you are also determining which types of statistical hypothesis tests are appropriate for your data—if that is your ultimate goal. I have written an article that discusses when to use parametric (mean) and nonparametric (median) hypothesis tests along with the advantages and disadvantages of each type.

15. What is the Likelihood?

Answer:
In statistics, the likelihood function (often simply called the likelihood) measures the goodness of fit of a statistical model to a sample of data for given values of the unknown parameters. It is formed from the joint probability distribution of the sample, but viewed and used as a function of the parameters only, thus treating the random variables as fixed at the observed values