STATISTICS WORKSHEET- 6

1. Which of the following can be considered as random variable?

Ans : All of the mentioned

2. Which of the following random variable that take on only a countable number of possibilities?

Ans : Discrete

3. Which of the following function is associated with a continuous random variable?

Ans : pdf

4. The expected value or _____ of a random variable is the center of its distribution.

Ans : mean

5. Which of the following of a random variable is not a measure of spread?

Ans : variance

6. The _____ of the Chi-squared distribution is twice the degrees of freedom.

Ans : standard deviation

7. The beta distribution is the default prior for parameters between ____

Ans : 0 and 1

8. Which of the following tool is used for constructing confidence intervals and calculating standard errors for difficult statistics?

Ans : bootstrap

9. Data that summarize all observations in a category are called _____ data.

Ans : summarized

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

Ans : A histogram is a type of bar chart that graphically displays the frequencies of a data set. Similar to a bar chart, a histogram plots the frequency, or raw count, on the Y-axis (vertical) and the variable being measured on the X-axis (horizontal).

A box plot, also called a box-and-whisker plot, is a chart that graphically represents the five most important descriptive values for a data set. These values include the minimum value, the first quartile, the median, the third quartile, and the maximum value. When graphing this five-number summary, only the horizontal axis displays values. Within the quadrant, a vertical line is placed above each of the summary numbers. A box is drawn around the middle three lines (first quartile, median, and third quartile) and two lines are drawn from the box's edges to the two endpoints (minimum and maximum).

Comparing Histograms and Box Plots

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.

1)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.

2)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.

3)Although histograms are better in displaying the distribution of data, you can use a box plot to tell if the distribution is symmetric or skewed. In a symmetric distribution, the mean and median are nearly the same, and the two whiskers has almost the same length.

11. How to select metrics?

Ans : 1. Define your primary objective

Before you even begin to sift through the various metrics and statistics available to you, it is essential that your company's governing objectives have been clearly established. As a B2B finance company, a primary objective could be to increase market share by 3% before the end of the year.

While an overarching goal such as this may seem somewhat abstract, if marketing metrics aren't considered with this objective in mind, you're bound to pick increasingly irrelevant ones over time.

2. Choose your metric(s) - determine cause and effect

Once a clear, overarching objective has been established, most marketing companies look to major metrics to determine their success—factors such as the generation of sales and leads.

But these metrics aren't the only indicator of a company's success. Less easily quantifiable factors such as customer satisfaction and brand loyalty also play a significant role in the ability to achieve overall marketing objectives, especially in the long term.

Examining the relationship between these metrics can allow marketers and others to develop a cause-and-effect theory to determine what drives the end results.

3. Create relevant activities

Digital technology has made it easier than ever to track the engagement of various types of marketing materials, be they a video, article, or even a podcast. Let's look at how a marketing agency could help its clients to improve performance.

Once a marketing agency has determined that engaging content is what drives sales and leads for their clients, the agency must determine which types of content reliably generates that engagement.

4. Evaluate periodically

The metrics and statistics that drive value for your clients can change over time, especially as new technologies emerge and target demographics shift.

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

Ans : Statistical significance can be accessed using hypothesis testing:

– Stating a null hypothesis which is usually the opposite of what we wish to test (classifiers A and B perform equivalently, Treatment A is equal of treatment B)

– Then, we choose a suitable statistical test and statistics used to reject the null hypothesis

– Also, we choose a critical region for the statistics to lie in that is extreme enough for the null hypothesis to be rejected (p-value)

– We calculate the observed test statistics from the data and check whether it lies in the critical region

Common tests:

– One sample Z test

– Two-sample Z test

– One sample t-test

– paired t-test

– Two sample pooled equal variances t-test

– Two sample unpooled unequal variances t-test and unequal sample sizes (Welch's t-test)

– Chi-squared test for variances

– Chi-squared test for goodness of fit

– Anova (for instance: are the two regression models equals? F-test)

– Regression F-test (i.e: is at least one of the predictor useful in predicting the response?)

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

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

Ans : Income is the classic example of when to use the median instead of the mean because its distribution tends to be skewed. The median indicates that half of all incomes fall below 27581, and half are above it. For these data, the mean overestimates where most household incomes fall.

15. What is the Likelihood?

Ans : Likelihood function is a fundamental concept in statistical inference. It indicates how likely a particular population is to produce an observed sample.

Let P(X; T) be the distribution of a random vector X, where T is the vector of parameters of the distribution. If Xo is the observed realization of vector X, an outcome of an experiment, then the function L(T | Xo) = P(Xo | T)

The likelihood function itself is not probability (nor density) because its argument is the parameter T of the distribution, not the random (vector) variable X itself. For example, the sum (or integral) of the likelihood function over all possible values of T should not be equal to 1.

Even if the set of all possible values of the vector T is discrete, the likelihood function still may be continuous (as far as the set of parameters T is continuous).