STATISTICS

WORKSHEET- 6

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

d) All of the mentioned

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

a) Discrete

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

a) pdf

**4. The expected value or \_\_\_\_\_\_\_ of a random variable is the center of its distribution.**

b) median

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

c) empirical mean

**6. The \_\_\_\_\_\_\_\_\_ of the Chi-squared distribution is twice the degrees of freedom.**

a) variance

**7. The beta distribution is the default prior for parameters between \_\_\_\_\_\_\_\_\_\_\_\_**

c) 0 and 1

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

b) bootstrap

**9. Data that summarize all observations in a category are called \_\_\_\_\_\_\_\_\_\_ data.**

d) none of the mentioned

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

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

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

**11. How to select metrics?**

The metrics selection depends upon the the target variable, if it is categorical or continuous, then in categorical what is more important to us TPR, FPR and in regression if data distribution.

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

Statistical significance is often calculated with statistical hypothesis testing, which tests the validity of a hypothesis by figuring out the probability that your results have happened by chance.

Here, a “hypothesis” is an assumption or belief about the relationship between your datasets. The result of a hypothesis test allows us to see whether this assumption holds under scrutiny or not.

A standard hypothesis test relies on two hypotheses.

Null hypothesis: The default assumption of a statistical test that you’re attempting to disprove (e.g., an increase in cost won’t affect the number of purchases).

Alternative hypothesis: An alternate theory that contradicts your null hypothesis (e.g., an increase in cost will reduce the number of purchases). This is the hypothesis you hope to prove.

**13. Give examples of data that does not have a Gaussian distribution, nor log-normal.**

The Skewed  
Discrete Distribution

Binomial Distribution

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

Let’s say you run a customer satisfaction survey with a sample of 9 and rate their overall satisfaction scores on a scale of 1 to 10. You get an average of 5.22. You know that in general, you tend to retain customers with a score over 3, so you’re satisfied, because this indicates that you’re still above where you want to be. But then, suddenly, you lose 6 of those 9 customers. You go back to look at your data, and you find these scores:

1, 3, 3, 3, 3, 5, 9, 10, 10

The median of this group is a 3, indicating that at least half of your customers or more were unhappy. The scores became lopsided because of the unexpected 10’s, and you missed out on an important part of your data – the midpoint that indicated that as many as half of your customers or more were dissatisfied with your company.

Median can play a major role in things like income level research as well, because a few millionaires may make it look like the socio-economic status of your sample is higher than it really is.

Whenever a graph falls on a normal distribution, using the mean is a good choice. But if your data has extreme scores (such as the difference between a millionaire and someone making 30,000 a year), you will need to look at median, because you’ll find a much more representative number for your sample.

**15. What is the Likelihood?**

Probability attaches to possible results; likelihood attaches to hypotheses. Explaining this distinction is the purpose of this first column.