**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

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

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

a) pdf b) pmv c) pmf

d) all of the mentioned

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

b) median c) mean

d) bayesian inference

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

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

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

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

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

b) summarized c) raw

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 and box plots are very similar in that they both help to visualize and describe numeric data. Although histograms are better in determining the underlying distribution of the data, box plots allow you to compare multiple data sets better than histograms as they are less detailed and take up less space. It is recommended that you plot your data graphically before proceeding with further statistical analysis.

11. How to select metrics?

The five easy steps listed below will enable them to systematically arrive at the appropriate metrics.

Step 1 Why is the measurement required?

Step 2 What needs to be measured?

Step 3 What is the precision of measurement required?

Step 4 How will it be measured?

Step 5 What use will the measurement be put to? By whom?

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

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:

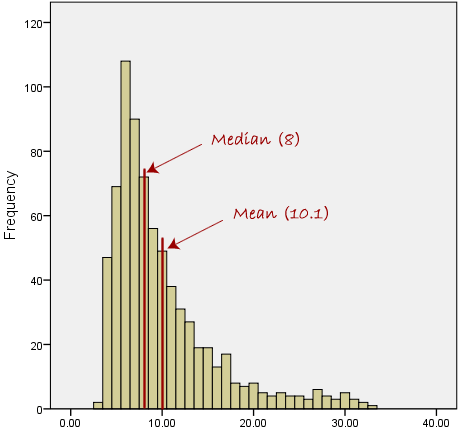
* One sample Z test
* Two-sample Z test
* One sample t-test
* paired t-test
* Two sample pooled equal variances t-test
* Two sample unpooled unequal variances t-test and unequal sample sizes (Welch’s t-test)
* Chi-squared test for variances
* Chi-squared test for goodness of fit
* Anova (for instance: 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.



When our data is skewed, for example, as with the right-skewed data set below:



We find that the mean is being dragged in the direct of the skew. In these situations, the median is generally considered to be the best representative of the central location of the data. The more skewed the distribution, the greater the difference between the median and mean, and the greater emphasis should be placed on using the median as opposed to the mean. A classic example of the above right-skewed distribution is income (salary), where higher-earners provide a false representation of the typical income if expressed as a mean and not a median.

15. What is the Likelihood?

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. The likelihood function describes a hypersurface whose peak, if it exists, represents the combination of model parameter values that maximize the probability of drawing the sample obtained. The procedure for obtaining these arguments of the maximum of the likelihood function is known as maximum likelihood estimation, which for computational convenience is usually done using the natural logarithm of the likelihood, known as the log-likelihood function. Additionally, the shape and curvature of the likelihood surface represent information about the stability of the estimates, which is why the likelihood function is often plotted as part of a statistical analysis.