STATISTICS WORKSHEET- 6

1. d) All of the mentioned
2. a) Discrete
3. a) pdf
4. c) mean
5. c) empirical mean
6. a) variance
7. c) 0 and 1
8. b) bootstrap
9. b) summarized
10. What is the difference between a boxplot and histogram?

* A [histogram](https://www.brighthubpm.com/six-sigma/13307-what-is-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 on the Y-axis (vertical) and the variable being measured on the X-axis (horizontal). A box plots a chart that graphically represents the five point summary 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.

1. How to select metrics?

* One of the standard methods for selecting metric is it should meet the quality criteria for the evaluation.
* Also, it should look for customer feedback and how it that impacting the overall business.
* The other selection criteria is while meeting the end objective it should follow process design for a particular task that will ensure robustness of product/service design.

1. 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 unpoled 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.

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

* Life table is example of exponential distribution, wind speed is Weibull distribution, surgery patient's stay in hospital is gamma distribution.

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

* Let’s say we are finding out the salary for software engineers between 3 to 5 years of experience across the country. In this case there will be outliers present in the dataset as the salary range varies based on the skills and also based on states. So, in this case median salary will give better idea to understand the overall salary data rather than mean.

1. What is the Likelihood?

* Likelihood is a method that underlies most common statistical methods used in psychology. It is the basis of classical methods of maximum likelihood estimation, and it plays a key role in Bayesian inference.  The likelihood of a hypothesis, *L*(*H*), is conditioned on the data, as if they are fixed while the hypothesis can vary based on the data.