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.

c) mean

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

a) variance

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.

b) summarized

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

Histograms are bar charts that show the frequency of a numerical variable's values and are used to approximate the probability distribution of the given variable. It allows you to quickly understand the shape of the distribution, the variation, and potential outliers.

Boxplots communicate different aspects of the distribution of data. While you can’t see the shape of the distribution through a box plot, you can gather other information like the quartiles, the range, and outliers. Boxplots are especially useful when you want to compare multiple charts at the same time because they take up less space than histograms

11. How to select metrics?

A proper metric has an owner, it is meaningful to your customer (whether internal or external) and it has a proper definition. A metric like predictability never really improves and it becomes a curse to anyone that touches it. (My nick-name was ‘mister predictability’).

Nowadays I take a different approach on metrics. I use three basic rules in selecting metrics:

1. Use standards. I prefer metrics that have been tested by others;

2. Measure yourself the way your customer measures you

3. Only measure metrics that have an owner

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

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

Many random variables have distributions that are *asymptotically* Gaussian but may be significantly non-Gaussian for small numbers. For example the Poisson Distribution, which describes (among other things) the number of unlikely events occurring after providing a sufficient opportunity for a few events to occur. It is pretty non-Gaussian unless the mean number of events is very large. The mathematical form of the distribution is still Poisson, but a histogram of the number of events after many trials with a large average number of events eventually looks fairly Gaussian. . For example, something that comes up all the time is that we detect stars in astronomical images and solve for their celestial coordinates. My current project uses images about 1.5 degrees on a side and typically detects 60 to 80 thousand stars per image, with the number well modeled as a Poisson Distribution, assuming that the image is not of a star cluster surrounded by mostly empty space. That's about 8 or 9 stars per square arcminute. If we cut out "postage stamps"from the image that are half an arcminute per side, then the mean number of detected stars in them is about 2. If we do that for (say) 1000 postage stamps and make a histogram of the number of detected stars in them, it will not look very Gaussian, but as we increase the size of the postage stamps, it becomes asymptotically Gaussian.

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

Mean vs. Median

Mean is simply another term for "Average." It takes all of the numbers in the dataset, adds them together, and divides them by the total number of entries. Median, on the other hand, is the 50% point in the data, regardless of the rest of the data. For example, if you have the following data:

1, 1, 1, 1, 1, 1, 2, 2, 4

The median is just "1" since that is the middle number in the dataset, while the mean (average) is 1.56. For a lot of analysis, the mean is very useful. Indeed, if you’re trying to understand data that falls under a normal curve, the mean can tell you a lot of information, because it helps remove some statistical noise from the data and gives you an overall average score for the group.

But the mean is far too often overused, because when it comes to collecting data, it’s not uncommon to find that there are extreme scores that may be altering the final results of your analysis

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.