Solved Statistics Worksheet

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
   Ans- 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
   Ans- a) Discrete

3. Which of the following function is associated with a continuous random variable?
   a) pdf
   b) pmv
   c) pmf
   d) all of the mentioned
   Ans- a) pdf

4. The expected value or ________ of a random variable is the center of its distribution.
   a) mode
   b) median
   c) mean
   d) bayesian inference
   Ans- c) mean

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
   Ans- c) empirical mean

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
   Ans- a) variance

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
   Ans- c) 0 and 1

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
   Ans- b) bootstrap

9. Data that summarize all observations in a category are called __________ data.
   a) frequency
   b) summarized
   c) raw
   d) none of the mentioned
   Ans- b) summarized

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

Ans.-

Boxplot	Histogram
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).	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).
Box plot gives the quartiles and indicate the median data to compare easily	Histogram gives only the count
Boxplot are less detailed than histograms and take up less space.	Histograms are more detailed and take up more space.
Box plots are more useful when comparing between several data sets.	Histograms are preferred to determine the underlying probability distribution of a data
A boxplot is useful when there is moderate variation among the observed frequencies	A histogram is highly useful when wide variances exist among the observed frequencies for a particular data set.

11. How to select metrics?

Ans.

The regression task is the prediction of the state of an outcome variable at a particular timepoint with the help of other correlated independent variables. The regression task, unlike the classification task, outputs continuous values within a given range.

The various metrics used to evaluate the results of the prediction are :

- Mean Squared Error: MSE or Mean Squared Error is one of the most preferred metrics for regression tasks. It is simply the average of the squared difference between the target value and the value predicted by the regression model. As it squares the differences, it penalizes even a small error which leads to over-estimation of how bad the model is. It is preferred more than other metrics because it is differentiable and hence can be optimized better.
- Root Mean Squared Error: RMSE is the most widely used metric for regression tasks and is the square root of the averaged squared difference between the target value and the value predicted by the model. It is preferred more in some cases because the errors are first squared before averaging which poses a high penalty on large errors. This implies that RMSE is useful when large errors are undesired.
- Mean Absolute Error: MAE is the absolute difference between the target value and the value predicted by the model. The MAE is more robust to outliers and does not penalize the errors as extremely as mse. MAE is a linear score which means all the individual differences are weighted equally. It is not suitable for applications where you want to pay more attention to the outliers.
- R^2 Error: Coefficient of Determination or R^2 is another metric used for evaluating the performance of a regression model. The metric helps us to compare our current model with a constant baseline and tells us how much our model is better. The constant baseline is chosen by taking the mean of the data and drawing a line at the mean. R^2 is a scale-free score that implies it doesn't matter whether the values are too large or too small, the R^2 will always be less than or equal to 1.
- Adjusted R^2: Adjusted R^2 depicts the same meaning as R^2 but is an improvement of it. R^2 suffers from the problem that the scores improve on increasing terms even though the model is not improving which may misguide the researcher. Adjusted R^2 is always lower than R^2 as it adjusts for the increasing predictors and only shows improvement if there is a real improvement.

Classification is the task of assigning a new observation to one of the categories or classes using the learning from an existing labeled dataset called training data. The factors to consider while choosing a classification metric. There are mainly two factors that decide the choice of classification.

1. The number of instances per class: A lot depends on the number of instances per class. One needs to check if it's a class imbalance dataset (some classes having much more data than others) or a balanced dataset i.e. classes roughly having the same number of instances.
2. The Business use-case to solve: Understanding the business needs whether to give every class equal importance or give more importance to some classes than rest. This also gives the direction around the right metric to use.

The various classification metrics –

- Accuracy- Accuracy is the total number of correctly labeled instances (doesn't matter positive or negative) to that of all the instances. Accuracy is a good metric only in the following cases.
    1. The classes or categories have evenly distributed instances i.e. It is a balanced dataset.
    2. The cost of false positives is the same as the cost of false negatives.
- Precision - Precision is the number of correctly positively labeled instances to that of all positively labeled instances (no matter those instances are actually positive or not) by the ML model. Precision is a metric that shows how confident we can be of positive prediction of a model even though possibly some actual positive instances are missed out.
- Recall or Sensitivity - The Recall or Sensitivity is the number of correctly positively labeled instances by the ML model to that of all the actual positive instances (no matter the ML model can detect correctly or not). Recall as a metric focuses on the fact that false negatives are costlier and no actual positives should be missed out.
- F1 score - F1 Score considers the goodness of both precision and recall. It is taken as the harmonic mean of the Precision and Recall.
- AUC_ROC - he Receiver Operator Characteristic (ROC) is a probability curve that plots the TPR(True Positive Rate) against the FPR(False Positive Rate) at various threshold values and separates the 'signal' from the 'noise'. The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes.

Metrics like accuracy, precision, recall are good ways to evaluate classification models for balanced datasets, but if the data is imbalanced then other methods like ROC/AUC perform better in evaluating the model performance.

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

Ans.

Statistical significance refers to the likelihood that a relationship between two or more variables is not caused by random chance. In essence, it's a way of proving the reliability of a certain statistic. Its two main components are sample size and effect size. In the use of statistical hypothesis testing, a data set's result can be deemed statistically significant if you have reached a certain level of confidence in the result. In statistical hypothesis testing, this means the hypothesis is unlikely to have occurred given the null hypothesis. According to a null hypothesis, there is no relationship between the variables in question.

The following is the process to assess the statistical significance of an insight-

- Create null hypothesis- The first step in calculating statistical significance is to determine your null hypothesis. Your null hypothesis should state that there is no significant difference between the sets of data you're using. Keep in mind that you don't need to believe the null hypothesis.
- Create alternative hypothesis- Next, create an alternative hypothesis. Typically, your alternative hypothesis is the opposite of your null hypothesis since it'll state that there is, in fact, a statistically significant relationship between your data sets
- Determine the significance level- Your next step involves determining the significance level or rather, the alpha. This refers to the likelihood of rejecting the null hypothesis even when it's true. A common alpha is 0.05 or five percent.
- Decide on the type of test- Next, you'll need to determine if you'll use a one-tailed test or a two-tailed test. Whereas the critical area of distribution is one-sided in a one-tailed test, it's two-sided in a two-tailed test. In other words, one-tailed tests analyze the relationship between two variables in one direction and two-tailed tests analyze the relationship between two variables in two directions. If the sample you're using lands within the one-sided critical area, the alternative hypothesis is considered true.
- Perform power analysis to find out sample size- You'll then need to do a power analysis to determine your sample size. A power analysis involves the effect size, sample size, significance level and statistical power. For this step, consider using a calculator. This type of analysis allows you to

see the sample size you'll need to determine the effect of a given test within a degree of confidence. In other words, it'll let you know what sample size is suitable to determine statistical significance. For example, if your sample size ends up being too small, it won't give you an accurate result.

- Calculate standard deviation and standard error- Next, you'll need to calculate the standard deviation and standard error
- Determine t- score
- Find the degrees of freedom
- Use t table- Finally, you'll calculate the statistical significance using a t-table. Start by looking at the left side of your degrees of freedom and find your variance. Then, go upward to see the p-values. Compare the p-value to the significance level or rather, the alpha. Remember that a p-value less than 0.05 is considered statistically significant.

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

Ans.

The examples of data that doesnot have a Gaussian distribution, nor log-normal are-

- Allocation of wealth among individuals
- Values of oil reserves among oil fields (many small ones, a small number of large ones)

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

Ans.

Median is a better measure when the data is skewed. Income is the classic example of when to use the median instead of the mean because its distribution tends to be skewed.

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.