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?

Ans -> d) All of the Mentioned

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

Ans -> a) Discrete

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

Ans-> a) pdf

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

Ans -> c) Mean

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

Ans -> a) Variance

6. The _________ of the Chi-squared distribution is twice the degrees of freedom

Ans -> a) Variance

7. The beta distribution is the default prior for parameters between ____________

Ans -> c) 0 and 1

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

Ans -> b) Bootstrap

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

Ans -> b) Summarized

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

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

Ans -> Histogram -> Used to plot the frequency distribution of quantitative variables

- Hist plot showing group frequence of continuous data

Boxplot -> Used for checking the outliers .

- Its show in quantile .(25%,50%,75%,100%)

11, How to select metrics?

Ans -> Lets assume that 100 people visited hospital you have give a test . so you have given a blood sample .for Glucose Level, BP,Sugar.

And you see some numbers for everything . our body contains number only . with those numbers doctor predicted he looked the data of all 100 people . and he predicted report,report has not come yet . before report doctor has predicted positive , negative .

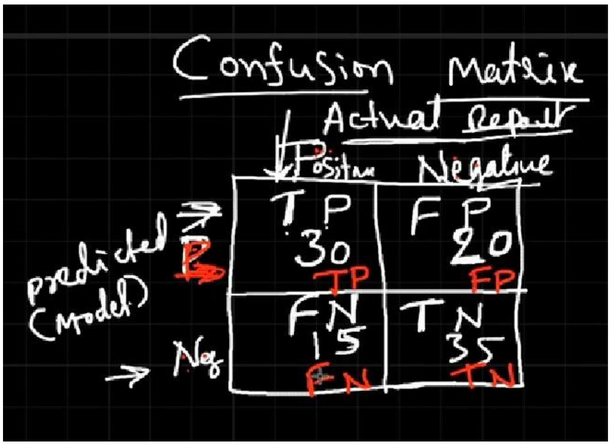
After two days report has come of 100 people report can never lie prediction can wrong . report have also positive ,negative list.

So what we can do compare actual with predicted . we we compare beacause we need to know how good is the doctor .

Overall i have to tell how good is the doctor . so we have to know how many doctor can predict correct and how many doctor can predict wrong .

Correctly Predicted means 100% wrong predicted means 0%

How are going to put in terms of a table to understand this . so that is what we are going to do that were your confusion metrics comes into picture.



Confusion Matrix

		Actual Result	
		Positive	Negative
predicted (Model)	P	TP 30 TP	FP 20 FP
	Neg	FN 15 FN	TN 35 TN

Where the terms have the meaning:

TP - A result that was predicted as positve by the classification model and also is positive .

TN - A result that was predicted as negative by the classification model and also is negative

FP - A result that was predicted as positive by the classification model but actually is negative

FN - A result was predicted as negative bu the classification model but actually is positive.

The Creadibility of the model is based on how many correct prediction did the model do .

--> So , Now This is the confusion metric . this is how we need to interpret . Now based on the confusion metric we are going to discuss metric.

The Creadibility of the model is based on how many correct prediction did the model do .

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

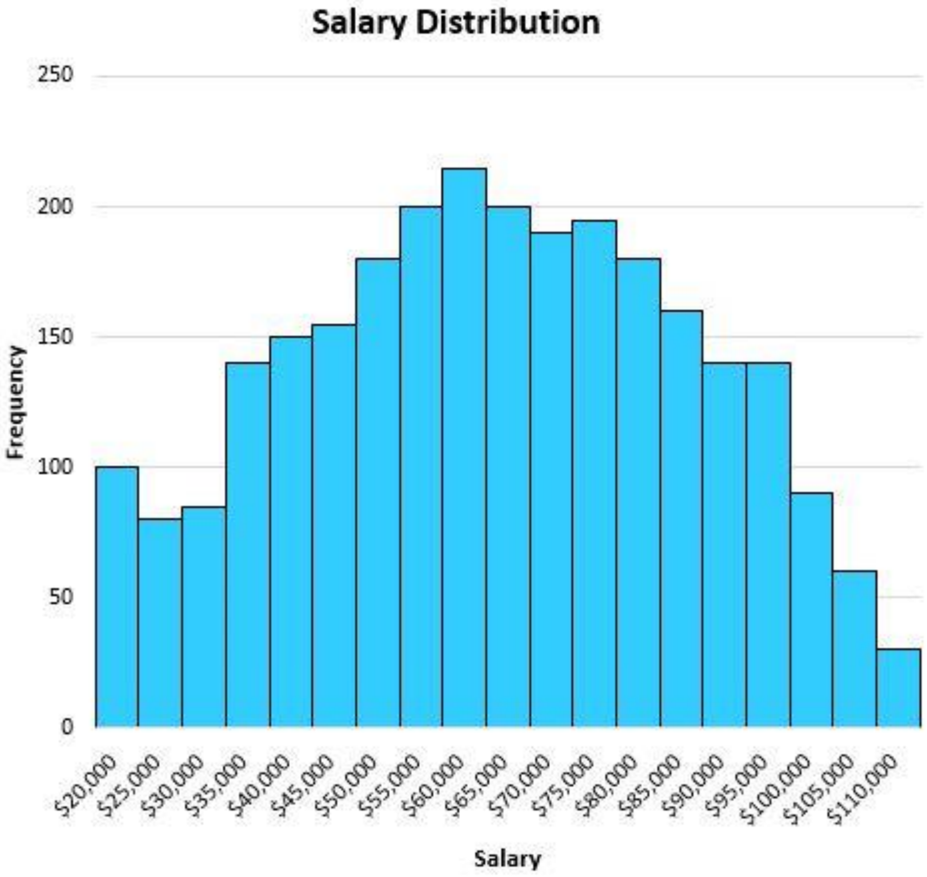
Ans -> Statistical significance can be accessed using hypothesis testing: – Stating a null hypothesis which is usually the opposite of what we wish to test. we choose a suitable statistical test and statistics used to reject the null hypothesis and choose a critical region for the statistics to lie in that is extreme enough for the null hypothesis to be rejected (p-value)

We then calculate the observed test statistics from the data and check whether it lies in the critical region. There are multiple test we performed based on the nature of the problem and features of our dataset.

13. Give examples of data that doesnot have a Gaussian distribution, nor log-normal.
Ans -> Types of distribution that are non-Gaussian or non-log normal are the skewed distributions, discrete distributions and binomial distribution

14. What is likelihood ?
Ans -. When there are way to many outliers, in those cases if we use mean. We will be way off as mean is drastically affected by outliers. Thus, in such cases it is preferable to use mean as the metric for central tendency than mean. . Another time when we usually prefer the median over the mean (or mode) is when our data is skewed

15. Give an example where the median is a better measure than the mean.
Ans -> When a distribution is skewed, the median does a better job of describing the center of the distribution than the mean. For example, consider the following distribution of salaries for residents in a certain city: The median does a better job of capturing the “typical” salary of a resident than the mean.


Salary Distribution
Frequency
$20,000
$25,000
$30,000
$35,000
$40,000
$45,000
$50,000
$55,000
$60,000
$65,000
$70,000
$75,000
$80,000
$85,000
$90,000
$95,000
$100,000
$105,000
$110,000
Salary