STATISTICS WORKSHEET-4

**1. What is central limit theorem and why is it important?**

The Central Limit Theorem tells us that as sample sizes get larger, the sampling distribution of the mean will become normally distributed, even if the data within each sample are not normally distributed.it is important for statistics because it allows us to safely assume that the sampling distribution of the mean will be normal in most cases. This means that we can take advantage of statistical techniques that assume a normal distribution, as we will see in the next section.

**2. What is sampling? How many sampling methods do you know?**

Sampling is a technique of selecting individual members or a subset of the population to make statistical inferences from them and estimate characteristics of the whole population.

1. Probability sampling involves random selection, allowing you to make strong statistical inferences about the whole group.
2. Non-probability sampling involves non-random selection based on convenience or other criteria, allowing you to easily collect data

**3. What is the difference between type1 and typeII error?**

Type I errors happen when we reject a true null hypothesis

Type II errors happen when we fail to reject a false null hypothesis

**4. What do you understand by the term Normal distribution?**

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve.

**5. What is correlation and covariance in statistics?**

|  |  |
| --- | --- |
| Covariance | Correlation |
| Covariance is a measure of how much two random variables vary together | Correlation is a statistical measure that indicates how strongly two variables are related |
| involve the relationship between two variables or data sets | involve the relationship between multiple variables as well |
| Lie between -infinity and +infinity | Lie between -1 and +1 |
| Measure of correlation | Scaled version of covariance |
| provide direction of relationship | provide direction and strength of relationship |
| dependent on scale of variable | independent on scale of variable |
| have dimensions | dimensionless |

**6. Differentiate between univariate ,Biavariate,and multivariate analysis.**

* Univariate statistics summarize only one variable at a time.
* Bivariate statistics compare two variables.
* Multivariate statistics compare more than two variables.

**7. What do you understand by sensitivity and how would you calculate it?**

Sensitivity (True Positive rate) measures the proportion of positives that are correctly identified (i.e. the proportion of those who have some condition (affected) who are correctly identified as having the condition).

To calculate sensitivity, we'll need:

Number of true positive cases (TP); and

Number of false negative cases (FN).

And the following sensitivity equation:

Sensitivity = TP / (TP + FN)

**8. What is hypothesis testing? What is H0 and H1? What is H0 and H1 for two-tail test?**

Hypothesis testing is a statistical method that is used in making statistical decisions using experimental data. Hypothesis Testing is basically an assumption that we make about the population parameter.

Hypothesis testing is formulated in terms of two hypotheses:

• H0: the null hypothesis;

• H1: the alternate hypothesis.

In a two-tailed test

• H0: θ = θ’

• H1: θ != θ’

**9. What is quantitative data and qualitative data?**

Qualitative data is a set of information which can not be measured using numbers. It generally consists of words, subjective narratives.

Quantitative data is a set of numbers collected from a group of people and involves statistical analysis.For example if you conduct a satisfaction survey from participants and ask them to rate their experience on a scale of 1 to 5

**10. How to calculate range and interquartile range?**

The formula to calculate the range is:

R= H - L

Formula to find the range

R = range

H = highest value

L = lowest value

The interquartile range formula is the first quartile subtracted from the third quartile:

IQR = Q3 – Q1.

**11. What do you understand by bell curve distribution ?**

Bell curve refers to the bell shape that is created when a line is plotted using the data points for an item that meets the criteria of normal distribution, the center contains the greatest number of a value and, therefore, it is the highest point on the arc of the line. This point is referred to the mean

**12. Mention one method to find outliers.**

The most effective way to find all of your outliers is by using the interquartile range (IQR). The IQR contains the middle bulk of your data, so outliers can be easily found once you know the IQR.

**13. What is p-value in hypothesis testing?**

A p value is used in hypothesis testing to help you support or reject the null hypothesis. The p value is the evidence against a null hypothesis. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis.

**14. What is the Binomial Probability Formula?**

Binomial probability refers to the probability of exactly x successes on n repeated trials in an experiment which has two possible outcomes (commonly called a binomial experiment).

If the probability of success on an individual trial is p , then the binomial probability is nCx⋅px⋅(1−p)n−x .

**15. Explain ANOVA and it’s applications.**

Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples.

ANOVA to help you understand how your different groups respond, with a null hypothesis for the test that the means of the different groups are equal. If there is a statistically significant result, then it means that the two populations are unequal (or different).