Statistics Worksheet 4

1. The central limit theorem states that the sampling distribution of the mean approaches a normal distribution, as the sample size increases. Therefore, as a sample size increases, the sample mean and standard deviation will be closer in value to the population mean μ and standard deviation σ .
   The central limit theorem tells us that no matter what the distribution of the population is, the shape of the sampling distribution will approach normality as the sample size (N) increases.

2. 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. Different sampling methods are widely used by researchers so that they do not need to research the entire population to collect actionable insights. Following are a few methods:
   - Simple Random Sampling
   - Cluster Sampling
   - Systematic Sampling
   - Stratified Random Sampling
   - Snowball Sampling

3. **Type I Error**
   The first kind of error that is possible involves the rejection of a null hypothesis that is actually true. This kind of error is called a type I error and is sometimes called an error of the first kind. Type I errors are equivalent to false positives.

   Type II Error
   The other kind of error that is possible occurs when we do not reject a null hypothesis that is false. This sort of error is called a type II error and is also referred to as an error of the second kind. Type II errors are equivalent to false negatives.

4. 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. **Correlation** is a statistical term describing the degree to which two variables move in coordination with one-another. If the two variables move in the same direction, then those variables are said to have a positive correlation. If they move in opposite directions, then they have a negative correlation.
   Covariance is a measure of the joint variability of two random variables. If the greater values of one variable mainly correspond with the greater values of the other variable, and the same holds for the lesser values, the covariance is positive.

6. **Univariate Analysis**

   Univariate analysis is the simplest form of data analysis where the data being analyzed contains only one variable. Since it's a single variable it doesn't deal with causes or relationships. The main purpose of univariate analysis is to describe the data and find patterns that exist within it.

   Bivariate Analysis

   Bivariate analysis is used to find out if there is a relationship between two different variables. For example, creating a scatterplot by plotting one variable against another on X and Y axis. It can sometimes give us a picture of what the data is trying to tell us. If the data seems to fit a line or curve then there is a relationship or correlation between the two variables.

   Multivariate Analysis

   Multivariate analysis is the analysis of three or more variables.

7. A sensitivity analysis determines how different values of an independent variable affect a particular dependent variable under a given set of assumptions.
   It is calculated by the following formula:
   $\mathbf{Z = X^2 + Y^2}$

8. Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter. The methodology employed by the analyst depends on the nature of the data used and the reason for the analysis.
   The null hypothesis would be represented as Ho: P = 0.8. The alternative hypothesis would be denoted as "H1" and be identical to the null hypothesis.
   For a two tailed test, the null hypothesis (H0) should be rejected when the test value is in either of two critical regions on either side of the distribution of the test value and vice versa for alternate hypothesis.

9. **Quantitative data** is information about quantities, and therefore numbers, **examples** are length, mass, temperature, and time whereas **qualitative data** is descriptive, and regards phenomenon which can be observed but not measured, such as language.

10. The **Range** is the difference between the lowest and highest values. Example: In {67, 6, 9, 3, 24} the lowest value is 3, and the highest is 67. So the **range** is 67 − 3 = 64.
    We can find the interquartile range or IQR in four simple steps:

1. Order the data from least to greatest
2. Find the median
3. Calculate the median of both the lower and upper half of the data
4. The IQR is the difference between the upper and lower medians

11. The term "bell curve" is used to describe a graphical depiction of a normal probability distribution, whose underlying standard deviations from the mean create the curved bell shape. A standard deviation is a measurement used to quantify the variability of data dispersion, in a set of given values around the mean. The mean, in turn, refers to the average of all data points in the data set or sequence and will be found at the highest point on the bell curve.

12. Following are a few methods for outliers:

    1) Z score method
    2) IQR method
    3) Cook's Distance method

13. 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. **P values** are expressed as decimals although it may be easier to understand what they are if you convert them to a percentage. For example, **a p value** of 0.0254 is 2.54%.

14. The binomial distribution formula is:
    $b(x; n, P) = nCx * Px * (1 – P)n – x$
    Where:
    b = binomial probability
    x = total number of "successes" (pass or fail, heads or tails etc.)
    P = probability of a success on an individual trial
    n = number of trials

15. 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.
    There are many industries that can use the ANOVA test to identify issues or variances between samples. The ANOVA is a good statistical technique for testing. Businesses that might consider the use of the ANOVA include manufacturing, healthcare, service, food, and more.

.