STATISTICS WORKSHEET-1 ANSWERS

1. A – True

2. A Central Limit Theorem

3. B Modeling bounded count data

4. D All of the mentioned

5. C Poisson

6. B False

7. B Hypothesis

8. A 0

9. C Outliers cannot conform to the regression relationship

SUBJECTIVE ANSWERS

10. What do you understand by the term Normal Distribution?

Answer:

- **Normal Distribution:** Normal Distribution is the most common or normal form of distribution of Random Variables, hence the name "normal distribution." It is also called Gaussian Distribution in Statistics or Probability. We use this distribution to represent a large number of random variables.

- A probability distribution determines the probability of all the outcomes a random variable takes. The distribution can either be continuous or discrete distribution depending upon the values that a random variable takes.

- The normal distribution is a continuous probability distribution function also known as Gaussian distribution which is symmetric about its mean and has a bell-shaped curve. It is one of the most used probability distributions. Two parameters characterize it

  Mean(μ)- It represents the center of the distribution

  Standard Deviation(σ) – It represents the spread in the curve

The formula for normal distribution:

$$f(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Where x is the random variable

Properties Of Normal Distribution:

- Symmetric distribution – The normal distribution is symmetric about its mean point. It means the distribution is perfectly balanced toward its mean point with half of the data on either side.

- Bell-Shaped curve – The graph of a normal distribution takes the form bell-shaped curve with most of the points accumulated at its mean position. The shape of this curve is determined by the mean and standard deviation of the distribution

- Empirical Rule – The normal distribution curve follows the empirical rule where 68% of the data lies within 1 standard deviation from the mean of the graph, 95% of the data lies within 2 standard deviations from the mean and 99.7% of the data lies within 3 standard deviations from the mean.

- Additive Rule – The sum of two or more normal distributions will always be a normal distribution.

- Central Limit Theorem – It states if we take the mean of large no data points collected from independent and identical distributed random variables then this mean will follow a normal distribution regardless of their original distribution.

11. How do you handle missing data? What imputation techniques do you recommend?

Answer:

Handling missing data is a common challenge in data analysis, especially when dealing with large datasets. Let's explore some techniques to efficiently handle missing values:

1. **Accepting Missing Data:**
    a. Leave the blank cells in your dataset and analyze them as-is.
    b. This approach assumes that the missing data does not significantly impact the overall analysis.

2. **Deletion**:
    - Remove data points or records that have missing values.
    - Listwise deletion: Remove entire rows with missing values. Simple but can reduce sample size significantly.
    - Pairwise deletion: Use available data for each analysis, ignoring missing values for specific variables.
    - Column-wise deletion: Remove entire columns with missing values.
3. **Imputation**:
    - Fill in missing data with estimated values.
    - Simple imputation methods include replacing absent values with the variable's mean, median, or mode.

Handling missing data is crucial for accurate statistical analyses. Here are some common imputation techniques:

1. **Hot Deck and Cold Deck Imputation:** In hot deck imputation, missing values are replaced with values from similar records, while cold deck imputation uses fixed external data. Both methods aim to maintain similarity between records.

2. **Listwise and Pairwise Deletion:** Listwise deletion removes entire cases with missing data, while pairwise deletion retains cases with at least some valid data. However, these methods can reduce sample size and bias results.

3. **Mean Imputation:** Replace missing values with the mean of the observed values for that variable. Simple but may distort variance and correlations.

4. **Regression Imputation:** Predict missing values using regression models based on other variables. Useful when relationships exist between variables.

5. **Multiple Imputation:** Generate multiple datasets with imputed values, accounting for uncertainty. Analyze each dataset separately and combine results. Widely used for robustness.

6. **Stochastic Imputation:** Randomly draw values from a distribution to replace missing data. Useful when the missingness mechanism is complex.

Remember that the choice of method depends on the missing-data mechanism and the specific context of your analysis2. Always assess the impact of imputation on your results and consider sensitivity analyses to evaluate robustness.

12. What is A/B testing?

Answer:

- In statistical terms, A/B testing is a method of two-sample hypothesis testing. This means comparing the outcomes of two different choices (A and B) by running a controlled mini-experiment. This method is also sometimes referred to as split testing.

- A/B testing is often discussed in the context of user experience (UX), conversion rate optimization (CRO), and other marketing and technology-focused applications; however, it can be valuable in other situations as well.

A/B testing or split testing, in a nutshell, is a means to compare two iterations of an email, website, or other marketing asset and assess the performance differences between them.

To accomplish this, distribute one version to one group and the other to a separate group. The effectiveness of every version may then be seen. It's best to approach it as a competition where two versions of your assets are pitted against one another to determine which will prevail.

The Process Of A/B Testing:

Several actions that you may see in a scientific study must be taken to conduct an A/B test. You'll choose one variable to test, just as in any scientific experiment. Whatever you choose can be your variable: the placement of a navigation menu or the color of a banner advertisement.

A/B testing gives you the ability to fully realize the potential of your business and marketing strategy. The steps for carrying out an insightful A/B test are listed below.

After selecting your variable, you should:

- Build on your hypothesis. (What do you expect the result should become?)

- Based on the chosen criteria, create a “control” group and a “challenger” group.

- Divide your sample groups at random into subgroups of the same size.

- Decide on the sample size (if applicable to your test).

- Specify what constitutes a statistically significant outcome.

- Make sure that each campaign is only having one test running at a time. (Doing many tests at once risks compromising results and invalidating your test.)

13. Is mean imputation of missing data acceptable practice?

Answer:

Mean imputation of missing data is a common approach, but its acceptability depends on the context. Here are some considerations:

1. Missing at Random (MAR): Mean imputation is acceptable when missing values occur randomly (MAR). In this case, imputing the mean preserves the overall distribution and does not introduce bias.

2. Biased Results: However, if missingness is not at random, mean imputation can lead to biased results. It assumes that missing values have the same distribution as observed values, which may not be true.

3. Standard Deviation Underestimation: Mean imputation tends to underestimate the standard deviation because it artificially reduces variability.

4. Distorted Relationships: It can distort relationships between variables by pulling correlation estimates toward zero.

5. Documentation: Always document the imputation process and the percentage of missing values in your dataset.

In summary, consider the context, explore other imputation methods (like multiple imputation), and document your choices carefully.

14. What is linear regression in statistics?

Answer:

Linear regression stands as a fundamental and widely utilized form of predictive analysis. It primarily seeks to address two critical questions:

- Firstly, how effectively can a set of predictor variables forecast an outcome (dependent or criterion) variable?
- Secondly, which specific variables emerge as significant predictors of the outcome variable, and how do their beta estimates—reflecting both magnitude and direction—affect this outcome?

Linear regression employs these estimates to describe the dynamics between one dependent variable and one or more independent variables. The most straightforward regression model, featuring one dependent and one independent variable, is encapsulated by the equation

$$y = c + b*x,$$

where: y represents the predicted score of the dependent variable,

c is the constant,

b denotes the regression coefficient, and

x is the score on the independent variable.

Key Applications of Regression Analysis

Determining the Strength of Predictors: This involves assessing the influence of independent variables on a dependent variable. Common inquiries include examining the relationship between variables such as dose and effect, sales and marketing expenditure, or age and income.

Forecasting Effects: Regression helps predict how changes in independent variables impact the dependent variable. A typical question

might be, “What is the expected increase in sales revenue for every additional $1000 spent on marketing?”

Trend Forecasting: It is used for predicting future trends and values, providing point estimates. For instance, one might ask, “What will the price of gold be in 6 months?”

Types of Linear Regression:

Simple linear regression
Involves one dependent variable (interval or ratio) and one independent variable (interval or ratio or dichotomous).

Multiple linear regression
Features one dependent variable (interval or ratio) and two or more independent variables (interval or ratio or dichotomous).

Logistic regression
Deals with one dependent variable (dichotomous) and two or more independent variables (interval or ratio or dichotomous).

Ordinal regression
Comprises one dependent variable (ordinal) and one or more independent variables (nominal or dichotomous).

Multinomial regression
Includes one dependent variable (nominal) and one or more independent variables (interval or ratio or dichotomous).

15. What are the various branches of statistics?

Answer:

Descriptive statistics and **inferential statistics** are the two main branches of statistics. Both the statistics branches are used in scientific data analysis and are equally significant.

Descriptive Statistics

The first aspect of statistics is descriptive statistics, which deals with the presentation and collection of data. It is not as simple as it appears. The statistician must know how to design and experiment, select the appropriate focus group, and prevent biases that are too easy to introduce into the experiment.

Generally, descriptive statistics can be categorized into

- Measures of central tendency
- Measures of variability

To understand both measures of tendency and variability, easily use graphs, tables, and general discussions.

Measures of Central Tendency

Measures of central tendency are used by statisticians to examine the value distribution center. These are the measures of tendency:

Mean

A mean is a common approach for describing the central tendency. To calculate the average of several values, count them all and divide them by the number of possible values.

Median

It is an outcome found in the middle of a set of values. In numerical journals, edit the results, and the result that is in the center of the distributed sample finds that one is an easy technique to get the median.

Mode

In the given data set, the value which occurs most frequently is the mode.

Measures Of Variability

The measure of variability helps the statisticians analyze the distribution from a particular data set. Quartiles, ranges, variances, and standard deviations are the variability variables.

Inferential Statistics :

Inference statistics (statistics branch) are statistical techniques that allow statisticians to utilize data from a sample to conclude, predict the behavior of a given population, and make judgments or decisions.

Using descriptive statistics, inference statistics frequently talk in terms of probability. Furthermore, a statistician uses these techniques mainly for data analysis, writing, and drawing conclusions from the limited data. This is accomplished by taking samples and determining their reliability.

Most future predictions and generalizations based on a population study of a smaller specimen are covered by inference statistics. Furthermore, the majority of social science experiments involve the investigation of a small sample population, and that helps in determining community behavior.

There are some different types of inferential statistics, which include the following, which are shown below:

- Regression analysis
- Analysis of variance (ANOVA)
- Analysis of covariance (ANCOVA)
- Statistical significance (t-test)
- Correlation analysis

You can measure inferential statistics in a variety of ways, including:

Hypothesis tests: Hypothesis tests determine whether your population is worth more than a data point in your analysis. It can also conclude if people differ, which is based on the results of several experiments.

Confidence intervals: Confidence intervals Determine the margin of error in your research and whether or not it affects what you're testing for. For mean and median calculations, you'll primarily need to estimate the range of a population's possible values.

Regression analysis: A regression analysis is a relationship between an experiment's independent and dependent variables. After you know the hypothesis test results, you can perform a regression analysis to determine the relationship of the subject matter. You can test for things like the difference in height and weight between two populations or the height and weight of different genders.