Statistics Worksheet 1 Answer

1. a) True
2. a) Central Limit Theorem
3. c) Modeling contingency tables
4. d) All of the mentioned
5. c) Poisson
6. b) False
7. b) Hypothesis
8. a) 0
9. d) None of the mentioned

10. Answer

The Normal distribution is also known as Gaussian or Gauss distribution. Normal Distribution is a probability function used in statistics that tells about how the data values are distributed. It is the most important probability distribution function used in statistics because of its advantages in real case scenarios. For example, the height of the population, shoe size, IQ level, rolling a dice, and Data Science etc.

Parameter of Normal Distribution:

a) Mean
mean or average value as a measure of central tendency. It can be used to describe the distribution of variables that are measured as ratios or intervals.
The mean determines the location of the peak, and most of the data points are clustered around the mean in a normal distribution graph.
If we change the value of the mean, then the curve of normal distribution moves either to the left or right along the X-axis.

b) Standard Deviation
The standard deviation measures how the data points are dispersed relative to the mean.
It determines how far the data points are away from the mean and represents the distance between the mean and the data points.
The standard deviation defines the width of the graph. As a result, changing the value of standard deviation tightens or expands the width of the distribution along the x-axis.
Usually, a smaller standard deviation wrt to the mean results in a steep curve while a larger standard deviation results in a flatter curve.

Properties of Normal Distribution:

a) It is symmetric

Statistics Worksheet 1 Answer

The shape of the normal distribution is perfectly symmetrical.
This means that the curve of the normal distribution can be divided from the middle and we can produce two equal halves. Moreover, the symmetric shape exists when an equal number of observations lie on each side of the curve.

b) The mean, median, and mode are equal

The midpoint of normal distribution refers to the point with maximum frequency i.e., it consists of most observations of the variable.
The midpoint is also the point where all three measures of central tendency fall. These measures are usually equal in a perfectly shaped normal distribution.

c) Empirical Rule

In normally distributed data, there is a constant proportion of data points lying under the curve between the mean and a specific number of standard deviations from the mean.
Thus, for a normal distribution, almost all values lie within 3 standard deviations of the mean.
These check buttons of normal distribution will help you realize the appropriate percentages of the area under the curve.
Remember that this empirical rule applies to all normal distributions. Also, note that these rules are applied only to the normal distributions.

d) Skewness and kurtosis

Skewness and kurtosis are coefficients that measure how different the distribution is from a normal distribution.
It measures the symmetry of the normal distribution while kurtosis measures the thickness of the tail distribution relative to that of normal distribution.

e) Area under the curve

The total area under the curve is unity(=1)

11. Answer

Following ways to handle missing values/data in the dataset:

a) Deleting Rows with missing values
b) Impute missing values for continuous variable
c) Impute missing values for categorical variable
d) Other Imputation Methods
e) Prediction of missing values
f) Imputation using Deep Learning Library — Datawig

some imputation techniques I would like recommend

a) Hot deck imputation

A randomly chosen value from an individual in the sample who has similar values on other variables.

In other words, find all the sample subjects who are similar on other variables, then randomly choose one of their values on the missing variable.

Statistics Worksheet 1 Answer

One advantage is you are constrained to only possible values. In other words, if Age in your study is restricted to being between 5 and 10, you will always get a value between 5 and 10 this way.

Another is the random component, which adds in some variability. This is important for accurate standard errors.

b) Cold deck imputation

A systematically chosen value from an individual who has similar values on other variables.

This is similar to Hot Deck in most ways, but removes the random variation. So for example, you may always choose the third individual in the same experimental condition and block.

c) Regression imputation

The predicted value is obtained by regressing the missing variable on other variables.

So instead of just taking the mean, you're taking the predicted value, based on other variables. This preserves relationships among variables involved in the imputation model, but not variability around predicted values.

d) Stochastic regression imputation

The predicted value from a regression plus a random residual value.

This has all the advantages of regression imputation but adds in the advantages of the random component.

Most multiple imputation is based off of some form of stochastic regression imputation.

e) Interpolation and extrapolation

An estimated value from other observations from the same individual. It usually only works in longitudinal data.

Use caution, though. Interpolation, for example, might make more sense for a variable like height in children–one that can't go back down over time. Extrapolation means you're estimating beyond the actual range of the data and that requires making more assumptions than you should.

Statistics Worksheet 1 Answer

12. Answer:

A/B testing is a type of split testing and is commonly used to drive improvements to specific variables or elements by measuring user or audience engagement. The approach is commonly used to optimize marketing campaigns or digital assets like websites. In A/B testing a specific variable is altered such as a title, image, or element layout. A sample of the audience is shown the control version and the altered version in a 50/50 split. Half traffic will interact with the original version, the other half will interact with the newer version. Engagement or the completion of a defined goal is the metric that is compared between the versions after a set period of time.

For instance, let's say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.

In the above scenario, you may divide the products into two parts – A and B. Here A will remain unchanged while you make significant changes in B's packaging. Now, on the basis of the response from customer groups who used A and B respectively, you try to decide which is performing better.

A/B testing can be used to:

Refine marketing campaign messaging and design.
Improve conversion rates through enhancements to user experience.
Continuously optimise assets like web pages by considering user engagement

13. Answer

Mean imputation means process of replacing null values in a dataset with the data's mean

Mean imputation is typically considered bad practice since it ignores feature correlation.

Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does.

Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate, and the confidence interval is narrower.

Statistics Worksheet 1 Answer

14. Answer

Linear regression models the relationships between at least one explanatory variable and an outcome variable. These variables are known as the independent and dependent variables, respectively. When there is one independent variable, the procedure is known as simple linear regression. When there are more IVs, statisticians refer to it as multiple regression.

Linear regression has two primary purposes—understanding the relationships between variables and forecasting.

a) The coefficients represent the estimated magnitude and direction (positive/negative) of the relationship between each independent variable and the dependent variable.
b) A linear regression equation allows you to predict the mean value of the dependent variable given values of the independent variables that you specify.

regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula

$y = c + b*x$

where y = estimated dependent variable score,

c = constant,

b = regression coefficient, and

x = score on the independent variable.

Naming the Variables. There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand. The independent variables can be called exogenous variables, predictor variables, or regressors.

Three major uses for regression analysis are:

(a) determining the strength of predictors,

Statistics Worksheet 1 Answer

(b) forecasting an effect, and

(c) trend forecasting.

15. Answer

Branches of statistics:

a) Descriptive Statistics

Descriptive statistics deals with the collection of data, its presentation in various forms, such as tables, graphs and diagrams and finding averages and other measures which would describe the data.

For example: Industrial statistics, population statistics, trade statistics, etc. Businessmen make use of descriptive statistics in presenting their annual reports, final accounts, and bank statements.

Descriptive statistics can be categorized into

Measures of central tendency
Measures of variability

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

Measures of Central Tendency

Measures of central tendency specifically help the statisticians to estimate the center of values distribution. These measures of tendency are:

Mean

This is the conventional method used in describing central tendency. Usually, to compute an average of values, you add up all the values and then divide them with the number of values available.

Median

This is the score found at the middle of a set of values. A simple way to calculate a median is to arrange the scores in numerical orders and then locate the score which is at the center of the arranged sample.

Mode

This is the frequently occurring value in a given set of scores.

b) Inferential Statistics

Inferential statistics deals with techniques used for the analysis of data, making estimates and drawing conclusions from limited information obtained through sampling and testing the reliability of the estimates.

Statistics Worksheet 1 Answer

For example: Suppose we want to have an idea about the percentage of the illiterate population of our country. We take a sample from the population and find the proportion of illiterate individuals in the sample. With the help of probability, this sample proportion enables us to make some inferences about the population proportion.

The different types of calculation of inferential statistics include:

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