

# SOLUTIONS

**Answer 1.** (A) True

**Answer 2.** (A) Central Limit Theorem

**Answer 3.** (B) Modeling bounded count data.

**Answer 4.** (C) The square of a standard normal random variable follows what is called chi-squared distribution

**Answer 5.** (C) Poisson

**Answer 6.** (B) False

**Answer 7.** (B) Hypothesis

**Answer 8.** (A) 0

**Answer 9.** (c) Outliers cannot conform to the regression relationship

**Answer 10.** The normal distribution, often known as the Gaussian distribution, is a probability distribution that is commonly used in statistics to characterise regularly distributed continuous random variables. The normal distribution is a bell-shaped curve that is symmetrical around the mean, with the bulk of data points clustered around the curve's centre and fewer points in the distribution's tails. A normal distribution's mean, median, and mode are all equal and centred in the centre of the curve.

The dispersion of the data around the mean is determined by the standard deviation of the normal distribution. The normal distribution's features are widely known, and it is frequently used as a reference distribution in statistical analysis. Many natural occurrences, such as people's heights, object weights, and measurement errors, are roughly regularly distributed. The assumption of normalcy is significant in statistical inference since many statistical tests and procedures are dependent on it.

**Answer 11.** Missing data is a critical issue in data analysis since it can lead to biased conclusions and lower statistical power. There are several approaches to dealing with missing data, including:

**Complete case analysis:** In this method, only cases with complete data are used for analysis; instances with missing data are not included. However, this might lead to information loss and a smaller sample size.

**Imputation:** It is the process of replacing missing data with approximated values based on observable data. Imputation procedures include the following:

Mean imputation: Using the observed data's mean value to fill in missing values.

Regression imputation is the process of extrapolating missing values from observable data using a regression model.

Multiple imputation involves the creation of several plausible imputed datasets and the independent analysis of each dataset to get pooled estimates.

K-nearest neighbour imputation: Using the values of the nearest neighbours in the observed data to impute missing values.

**Answer 12.** A/B testing, commonly referred to as split testing, is a technique for evaluating the performance of two variants of a website, commercial, or marketing campaign.

An A/B test involves dividing the website or campaign into two groups and displaying a different webpage or advertisement to each group. The only difference between the two versions will be in the call to action, image, or headline.

The marketer can evaluate whether version of the webpage or advertisement is more effective in reaching the target result by monitoring the responses of each group, such as click-through rates, conversions, or engagement. This data-driven testing methodology can assist companies in making wise judgements about their marketing plans, boost conversion rates, and eventually boost income.

**Answer 13.** A common technique for filling in missing values in a dataset is mean imputation, which substitutes the missing values with the mean value of the remaining data. Mean imputation is a quick and simple method to deal with missing data, however it might not always be the optimal strategy.

Mean imputation has the drawback of perhaps introducing bias into the data. Since the mean is used to fill in missing values, the imputed data will have the

same mean as the observed data, which can lead to over- or under-estimation of the missing data's actual values. As a result, predictions and conclusions drawn from the imputed data may be incorrect.

Additionally, mean imputation makes the assumption that the missing data is absent at random and that its absence has nothing to do with its values. Mean imputation may introduce additional bias into the data and result in false findings if the missing data is not absent at random.

Therefore, even though mean imputation is a typical technique for handling missing data, it should only be used after carefully evaluating the type of missing data and any potential for bias. In other circumstances, alternative techniques, like multiple imputation, which assigns missing values based on the distribution of the observed data, may be more appropriate.

**Answer 14.** A dependent variable (also known as the response variable) and one or more independent variables (also known as predictor variables) are modelled using the statistical technique of linear regression.

Finding the best-fitting line (or plane, if there are several independent variables) that illustrates the linear connection between the dependent variable and the independent variable(s) is the aim of linear regression. By minimising the sum of the squared differences between the dependent variable's actual values and predicted values, the best-fitting line is found.

One independent variable in a simple linear regression model results in the equation:

$$y = \beta_0 + \beta_1 x + \epsilon$$

Where:

The dependent variable is  $y$ .

The independent variable is  $x$ .

The intercept, or constant term, is 0.

The regression coefficient, often known as the line's slope, is 1.

The residual or error term is

Both simple and complicated linear regression models can be created using linear regression. The equation becomes: when multiple linear regression is used.

**Answer 15.** Descriptive statistics and inferential statistics are the two primary branches of the large field of statistics.

**Descriptive Statistics:** The collection, organisation, analysis, and presentation of data are all topics covered by descriptive statistics. It covers techniques for enumerating and describing data sets, including graphical representations (histograms, box plots, scatter plots), measurements of variability (standard deviation, variance, range), and measures of central tendency (mean, median, mode).

Making predictions or **inferences** about a wider population based on a sample of data is the focus of inferential statistics. It contains techniques for doing regression analysis, parameter estimation, and hypothesis testing. Techniques for estimating the degree of confidence or uncertainty in the results are also included in inferential statistics.

Statistics can be further broken down into a number of subfields or specialisations in addition to these two basic divisions, including

**Biostatistics:** The use of statistical techniques in the study of biological and medical data is the focus of biostatistics.

**Business statistics:** The use of statistical techniques to address issues in management and business, such as market research, quality assurance, and financial analysis, is the focus of this discipline.

Applying statistical techniques to economic data, such as examining the correlation between economic variables or predicting future economic trends, is the subject of econometrics.

