Statistics Solution

1)a

2)a

3)a

4)d

5)c

6)b

7)b

8)a

9)d

10) 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.

Properties of a normal distribution:

- The mean, mode and median are all equal.
- The curve is symmetric at the center (i.e. around the mean, μ).
- Exactly half of the values are to the left of center and exactly half the values are to the right.
- The total area under the curve is 1.

A normal distribution is the proper term for a probability bell curve.

In a normal distribution the mean is zero and the standard deviation is 1. It has zero skewness.

Normal distributions are symmetrical, but not all symmetrical distributions are normal.

11)Understanding the nature of missing data is critical in determining what treatments can be applied to overcome the lack of data. Data can be missing in the following ways:

Missing Completely At Random (MCAR): When missing values are randomly distributed across all observations, then we consider the data to be missing completely at random. A quick check for this is to compare two parts of data – one with missing observations and the other without missing observations. On a t-test, if we do not find any difference in means between the two samples of data, we can assume the data to be MCAR.

Missing At Random (MAR): The key difference between MCAR and MAR is that under MAR the data is not missing randomly across all observations, but is missing randomly only within sub-samples of data. For example, if high school GPA data is missing randomly across all schools in a district, that data will be considered MCAR. However, if data is randomly missing for students in specific schools of the district, then the data is MAR.

Not Missing At Random (NMAR): When the missing data has a structure to it, we cannot treat it as missing at random. In the above example, if the data was missing for all students from specific schools, then the data cannot be treated as MAR.

Common Methods to Imputing Missing Data:

Mean imputation:

Simply calculate the mean of the observed values for that variable for all individuals who are non-missing.

It has the advantage of keeping the same mean and the same sample size, but many, many disadvantages.

Substitution:

Impute the value from a new individual who was not selected to be in the sample.

In other words, go find a new subject and use their value instead.

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.

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.

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.

Regression imputation:

The predicted value 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.

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.

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 that you should.

As per me, Stochastic regression imputation will give best result as it has all the advantages of regression imputation but adds in the advantages of the random component.

12)A/B test is the shorthand for a simple controlled experiment.[1] As the name implies, two versions (A and B) of a single variable are compared, which are identical except for one variation that might affect a user's behavior. A/B tests are widely considered the simplest form of controlled experiment. However, by adding more variants to the test, this becomes more complex.

A/B tests are useful for understanding user engagement and satisfaction of online features, such as a new feature or product. Large social media sites like LinkedIn, Facebook, and Instagram use A/B testing to make user experiences more successful and as a way to streamline their services.

Like any type of scientific testing, A/B testing is basically statistical hypothesis testing, or, in other words, statistical inference. It is an analytical method for making decisions that estimates population parameters based on sample statistics.

13)No, mean imputation of missing data is not acceptable practice because of below 2 points:

1)**Mean imputation does not preserve the relationships among variables:**

If the data are missing completely at random, mean imputation will not bias your parameter estimate but it will still bias your standard error which is not a good pratcise.

2)**Mean Imputation Leads to An Underestimate of Standard Errors:**

A second reason is applies to any type of single imputation. Any statistic that uses the imputed data will have a standard error that's too low.

In other words, yes, you get the same mean from mean-imputed data that you would have gotten without the imputations. And yes, there are circumstances where that mean is unbiased. Even so, the standard error of that mean will be too small.

Because the imputations are themselves estimates, there is some error associated with them. But your statistical software doesn't know that. It treats it as real data.

14)Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? These 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.

Types of Linear Regression

Simple linear regression

1 dependent variable (interval or ratio), 1 independent variable (interval or ratio or dichotomous)

Multiple linear regression

1 dependent variable (interval or ratio), 2+ independent variables (interval or ratio or dichotomous)

Logistic regression

1 dependent variable (dichotomous), 2+ independent variable(s) (interval or ratio or dichotomous)

Ordinal regression

1 dependent variable (ordinal), 1+ independent variable(s) (nominal or dichotomous)

Multinomial regression

1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio or dichotomous)

Discriminant analysis

1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio)

15)Statistics is a branch of applied mathematics that involves the collection, description, analysis, and inference of conclusions from quantitative data. The mathematical theories behind statistics rely heavily on differential and integral calculus, linear algebra, and probability theory. Statisticians, people who do statistics, are particularly concerned with determining how to draw reliable conclusions about large groups and general phenomena from the observable characteristics of small samples that represent only a small portion of the large group or a limited number of instances of a general phenomenon.

The two major areas of statistics are known as descriptive statistics, which describes the properties of sample and population data, and inferential statistics, which uses those properties to test hypotheses and draw conclusions.

Descriptive Statistics:

In this type of statistics, the data is summarized through the given observations. The summarization is one from a sample of population using parameters such as the mean or standard deviation.

Descriptive statistics is a way to organize, represent and describe a collection of data using tables, graphs, and summary measures. For example, the collection of people in a city using the internet or using Television.

Descriptive statistics are also categorized into four different categories:

1. Measure of frequency
2. Measure of dispersion
3. Measure of central tendency
4. Measure of position

Inferential Statistics:

This type of statistics is used to interpret the meaning of Descriptive statistics. That means once the data has been collected, analyzed and summarized then we use these stats to describe the meaning of the collected data. Or we can say, it is used to draw conclusions from the data that depends on random variations such as observational errors, sampling variation, etc.

Inferential Statistics is a method that allows us to use information collected from a sample to make decisions, predictions or inferences from a population. It grants us permission to give statements that goes beyond the available data or information. For example, deriving estimates from hypothetical research.