Statistics worksheet

|  |
| --- |
|  |
|  |  |
|  | 1) a |
|  |  |
|  | 2) a |
|  |  |
|  | 3) b |
|  |  |
|  | 4) d |
|  |  |
|  | 5) c |
|  |  |
|  | 6) b |
|  |  |
|  | 7) b |
|  |  |
|  | 8) a |
|  |  |
|  | 9) c |
|  |  |
|  | 10) The normal distribution is a continuous probability distribution that is symmetrical on both sides of the mean, so the right side of the center is a mirror image of the left side. |
|  | The area under the normal distribution curve represents probability and the total area under the curve sums to one. |
|  | Most of the continuous data values in a normal distribution tend to cluster around the mean, and the further a value is from the mean, the less likely it is to occur. The tails are asymptotic, which means that they approach but never quite meet the horizon (i.e. x-axis). |
|  | For a perfectly normal distribution the mean, median and mode will be the same value, visually represented by the peak of the curve. |
|  |  |
|  | 11) Missing data (or missing values) is defined as the data value that is not stored for a variable in the observation of interest. The problem of missing data is relatively common in almost all research and can have a significant effect on the conclusions that can be drawn from the data. Accordingly, some studies have focused on handling the missing data, problems caused by missing data, and the methods to avoid or minimize such in medical research |
|  |  |
|  | And here are seven things you can do about that missing data: |
|  |  |
|  | a) Listwise Deletion: Delete all data from any participant with missing values. If your sample is large enough, then you likely can drop data without substantial loss of statistical power. Be sure that the values are missing at random and that you are not inadvertently removing a class of participants. |
|  | b) Recover the Values: You can sometimes contact the participants and ask them to fill out the missing values. For in-person studies, we’ve found having an additional check for missing values before the participant leaves helps. |
|  | c) Educated Guessing: It sounds arbitrary and isn’t your preferred course of action, but you can often infer a missing value. For related questions, for example, like those often presented in a matrix, if the participant responds with all “4s”, assume that the missing value is a 4. |
|  | d) Average Imputation: Use the average value of the responses from the other participants to fill in the missing value. If the average of the 30 responses on the question is a 4.1, use a 4.1 as the imputed value. This choice is not always recommended because it can artificially reduce the variability of your data but in some cases makes sense. |
|  | e) Common-Point Imputation: For a rating scale, using the middle point or most commonly chosen value. For example, on a five-point scale, substitute a 3, the midpoint, or a 4, the most common value (in many cases). This is a bit more structured than guessing, but it’s still among the more risky options. Use caution unless you have good reason and data to support using the substitute value. |
|  | f) Regression Substitution: You can use multiple-regression analysis to estimate a missing value. We use this technique to deal with missing SUS scores. Regression substitution predicts the missing value from the other values. In the case of missing SUS data, we had enough data to create stable regression equations and predict the missing values automatically in the calculator. |
|  | g) Multiple Imputation: The most sophisticated and, currently, most popular approach is to take the regression idea further and take advantage of correlations between responses. In multiple imputation [pdf], software creates plausible values based on the correlations for the missing data and then averages the simulated datasets by incorporating random errors in your predictions. It is one of a number of examples where computers continue to change the statistical landscape. Most statistical packages like SPSS come with a multiple-imputation feature. More on multiple imputation. |
|  |  |
|  | 12) A/B testing is a method of gathering insight to aid in optimization. It involves testing an original design (A) against an alternate version of that design (B) to see which performs better. That original design is also known as “the control” and the alternate version is known as a “variation.” |
|  | This guide will discuss A/B testing as it relates to post-click landing page optimization, but you can use the method to compare and improve other web pages, emails, ads, and more. |
|  |  |
|  | 13) Definition: Mean imputation is the replacement of a missing observation with the mean of the non-missing observations for that variable. |
|  | Mean imputation of missing data is not a acceptable practice. |
|  | It’s a popular solution to missing data, despite its drawbacks. Mainly because it’s easy. It can be really painful to lose a large part of the sample you so carefully collected, only to have little power, mean imputation (also called mean substitution) really ought to be a last resort. |
|  | But that doesn’t make it a good solution, and it may not help you find relationships with strong parameter estimates. Even if they exist in the population. |
|  | On the other hand, there are many alternatives to mean imputation that provide much more accurate estimates and standard errors, so there really is no excuse to use it. |
|  | Major drawbacks of mean imputation are: |
|  | a) Bad practice in general |
|  | b) If just estimating means: mean imputation preserves the mean of the observed data |
|  | c) Leads to an underestimate of the standard deviation |
|  | d) Distorts relationships between variables by “pulling” estimates of the correlation toward zero |
|  |  |
|  | 14) Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables: |
|  | a) One variable, denoted x, is regarded as the predictor, explanatory, or independent variable. |
|  | b) The other variable, denoted y, is regarded as the response, outcome, or dependent variable. |
|  | Because the other terms are used less frequently today, we'll use the "predictor" and "response" terms to refer to the variables encountered in this course. The other terms are mentioned only to make you aware of them should you encounter them. Simple linear regression gets its adjective "simple," because it concerns the study of only one predictor variable. In contrast, multiple linear regression, which we study later in this course, gets its adjective "multiple," because it concerns the study of two or more predictor variables. |
|  |  |
|  | 15) There are two branches of Statistics: |
|  |  |
|  | a) Descriptive Statistics: |
|  | Descriptive statistics is considered as the first part of statistical analysis which deals with collection and presentation of data. Scientifically, descriptive statistics can be defined as brief explanatory coefficients that are used by statisticians to summarize a given data set. Generally, a data set can either represent a sample of a population or the entire populations. Descriptive statistics can be categorized into: |
|  | i) Measures of central tendency |
|  | ii) Measures of variability |
|  | To easily understand the analyzed data, both measures of tendency and measures of variability use tables, general discussions, and graphs. |
|  |  |
|  | b) Inferential/Inductive Statistics: |
|  | Inferential statistics are techniques that enable statisticians to use the gathered information from a sample to make inferences, decisions or predictions about a given population. Inferential statistics often talks in probability terms by using descriptive statistics. These techniques are majorly used by statisticians to analyze data, make estimates and draw conclusions from the limited information which is obtained by sampling and testing how reliable the estimates are. |
|  | The different types of calculation of inferential statistics include: |
|  | i) Regression analysis |
|  | ii) Analysis of variance (ANOVA) |
|  | iii) Analysis of covariance (ANCOVA) |
|  | iv) Statistical significance (t-test) |
|  | v) Correlation analysis |