Report on data due diligence

Technology

In order to draw relevant conclusions from any given data set, the researcher needs to address the following points: 1) There must be clear delineation between continuous and discrete variables. Continuous variables have values that can be measured in intervals. Discrete variables, on the other hand, have values that cannot be measured in intervals (in whole numbers or integers). Examples of continuous variables are temperature, weight, income, height, and depth. Examples of discrete variables are age, number of months, and number of children; 2) The second consideration is related to the type of statistical tests used in specific analysis.

The Student’s t-test is used for comparing the variance of ‘ two’ sample groups. The ANOVA or the Analysis of Variance is used for more than two sample groups (it can also be used in analyzing the variance of two sample groups. Regression analysis provides a synthetic equation which describes the relationship between one or more variables; 3) In group analysis, correlation is of little value. The important thing is the perpetuity of the variance or the existence of disparity; 4) It would be irresponsible for the researcher to conduct multiple statistical tests when only one or two is required to establish the validity of any given hypothesis;

5) The researcher should always consider the normality of any given data set. If the data set is normally distributed, he/she may use the ANOVA or the t-test. If, however, the data set is not normally distributed, he/she should use nonparametric tests like the Kruskal-Wallis Test. 6) Lastly, the researcher must maintain internal validity and reliability in his/her methodologies. This is important because research is grounded on validity and reliability.

Manipulation or undue ‘ dropping’ of variables is always the ‘ mortal sin’ of quantitative research.

In the data set provided, note that continuous and discrete variables are not grouped. The researcher should therefore provide a necessary format for an easier analysis of data. Items with blank ‘ content’ are automatically assigned a ‘ 0’ value. If the associated variable is an ordinal one, retaining the blank item is acceptable. If, however, the associated variable is a nominal one, then the researcher should remove that sample from the data set. Retaining ‘ empty’ values in rank analysis often leads to confusion and misinterpretation of data. The size of any sample always follows from the size of the universe or the population.

If the population is about 120 000, then the sample size should be around 4 to 5 percent. If the population is more or less 1000, then the sample size should be around 40 to 100. Note that if the sample size is equal or less than 25, then the researcher should use the Student’s t-test. Suppose the sample size is greater than 25 (congruent to the population), the researcher may use the Z-test (assuming that the data is normally distributed). Now, the so-called p-value should not be confused with the P(X) – the former indicates the critical area in a distribution while the latter the probability of any given event.

Finally, data analysis is insufficient if not supplemented by discourses on related literature, theories, and hypotheses. Hypothesis Testing Hypothesis I: The age of residents of region I is greater than the age of residents of region VII. Null Hypothesis: There is no difference in the age of residents of region I and the age of residents of region VII. Alternative Hypothesis: The age of

residents of region I is greater than the age of residents of region VII. Preliminary Analysis The variables ‘ age’ and ‘ region’ are discrete variables. It is better to plot discrete variables in bar graphs than in line graphs.

ANOVA may be used in the analysis. Note that analysis is one-tailed. Statistical Test The Student’s t-test is used to determine the mean variance of the two groups. Results The resulting p-value is equal to 0. 191056. Because the p-value is greater than ? = 0. 05, we fail to reject the null hypothesis. In short, there is no difference in the age of residents of region I and the age of residents of region VII. Hypothesis II. The wealth score of mail donors is greater than the wealth score of mail non-donors. Null Hypothesis: There is no difference in the wealth score of mail donors and mail non-donors.

Alternative Hypothesis: The wealth score of mail donors is greater than the wealth score of mail non donors. Preliminary Analysis The variables ‘ age’ and ‘ region’ are discrete variables. It is better to plot discrete variables in bar graphs than in line graphs. ANOVA may be used in the analysis. Note that analysis is one-tailed. Statistical Test The Kruskal-Wallis test is used to determine the mean variance of the two groups. This test is used because the values of the data set are not normally distributed. Results The resulting p-value is equal to 0. 20158. Because the p-value is greater than ? = 0. 05, we fail to reject the null hypothesis.

In short, there is no difference in the wealth score of mail donors and mail non-donors (consequently, the wealth score of mail donors is not greater than the wealth score of mail non-donors). Continuous and Discrete Variables Discrete Variables Some of the discrete variables in the data set are age,

region, and number of children. Age Measures of Central Tendency The mean is equal to 46. 202. In short, the average age of most respondents is 46. The median age is 44. The mode is 46. The largest number of respondents is aged 46. The distribution is relatively skewed to the left. Region Measures of Central Tendency

The mean is equal to 4. 69. In short, most of the respondents came from regions IV and V. The mode is equal to 7. The largest number of respondents came from region VII. The median is equal to 4. The distribution is relatively skewed to the right. Number of Children Measures of Central Tendency The mean is equal to . 586. In short, most of the respondents have at most one child The mode is equal to 0. The largest number of respondents has no children. The median is equal to 2. The distribution is relatively skewed to the right. Continuous Variables Some of the continuous variables in the data set are income and wealth score. Income

The mean is equal to USD 25000- 43, 000 In short, on the average, most of the respondents have incomes ranging from 25000 to 43000. The mode is equal to USD 40000. The largest number of respondents has income of about USD 40000. The median is equal to USD 50000. The distribution is relatively skewed to the left. Wealth Score Measures of Central Tendency The mean is equal to 301. 84. In short, on the average, most of the respondents have wealth score equal to 301. The mode is equal to 235. The largest number of respondents have wealth score equal to 235. The median is equal to 252. The distribution is relatively skewed to the right.