SCMA 524 Term Project, definition and analysis section

R.L. Andrews

This project is designed to be done as a two or three person team. **5 point penalty if done individually without permission**. The first part involves data collection and belief statements about relationships between variables, then the second part involves analysis of the collected data and reporting the findings. The project purpose is to make sure that you understand how conceptual variables are measured either qualitatively or quantitatively and to give you practical experience in

1. acquiring data for a prescribed analysis,

2. expressing belief statements about relationships between the variables, converting them into statistical hypotheses, and

3. using the tools studied in this class to analyze the data for the purpose of identifying if the data support the existence or nonexistence of the hypothesized relationships between the different types of variables in the data set.

For this project you are to obtain and use a data set with **five different variables** (C1, C2, Q1, Q2 and R) that meet the following specifications. Note that you are to have **at least 5 observations for each category of a categorical variable** and **the stipulated minimum number of possible values for the quantitative variables**.

**C1 Categorical Variable 1:** measured with **2 categories** (≥5 per category). (Example: in-state or out-of-state student)

**C2 Categorical Variable 2:** different fromC1 & measured with **3 categories** (5 or more per each category group).

 (Example: freshman/sophomore, junior/senior, or graduate student)

**Q1 Quantitative Variable 1:** measured with a **quantitative measurement** (a number that truly measures a quantity).

**Q2 Quantitative Variable 2:** different from Q1 & measured with a **quantitative measurement**.

**R Response Variable**: like the two above, it is to have a quantitative measurement. Select **a variable different from C1, C2, Q1 & Q2** and one whose value might be related in some way to the values of the other variables.

The quantitative variables should have many possible values (**at least 20 for R, at least 8 for Q1 and at least 8 for Q2**) and cannot be a cumulative measure relative to time or another characteristic. The variables **C1, C2, Q1, Q2 & R must be measurements of separate variables (Q1 & Q2 can’t be a subdivision of R or R a subdivision of Q1 or Q2)**.

**1.** Clearly indicate the entity (object, person, group, case or activity that is the sampling unit) for which measurements are obtained. What do the rows of the data represent?

**2. For each of the variables C1, C2, Q1, Q2 & R, provide an operational definition for the recorded measurement.** An **Operational Definition** is a definition that clearly defines a measurement, item or element for the particular situation being considered. This definition must be clear so that all persons involved will have exactly the same understanding of what is being defined. It clearly describes the units of measurement, such as inches. For the qualitative or categorical variables provide clear definitions for each of the categories.

**3. Tell how the data were gathered.** If you collect the sample directly, describe your sampling procedure. **Is there a phenomenon** (process or population) **for which your data would be a representative sample? Give a reason to support your answer explaining why it is or is not a representative sample of a process or a population.**

**4. Collect and report the values for at least 50 observations for each of the five variables. You may not use any of the data sets from the text or the CD.** If you collect data from a process over time, try to collect the data as close together as possible to remove the possibility of the parameters for the process changing over time.

**MAKE SURE THAT THE DATA ON ALL FIVE VARIABLES MEET THE ABOVE CRITERIA!**

***Have at least 5 observations for each category of a categorical variable.***

Below is a sample for spreadsheet column headings and three observations.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| C1.Gender | C2. Class standing | Q1. Current GPA | Q2. Credit Hours this Semester | R. Amount spent on books & supplies for the semester |
| MALE | GRAD | 3.4 | 6 | $425 |
| FEMALE | FR/SO | 2.5 | 12 | $563 |
| FEMALE | JR/SR | 2.1 | 15 | $637 |

**Save the data in a secure place! Do not save all your data & analysis on only one storage device.**

**5. State what you believe will be the specific anticipated relationships between each of these pairs of variables.**

 **a. C1 vs. R, b. C2 vs. R, c. Q1 vs. R, d. Q2 vs. R, e. Q1 vs. Q2, and f. C1 vs. C2**

For example: a. We think that R, expenses, will not be related in any way with C1, gender.

 d. We think that R, expenses, will increase as Q2, the number of credit hours, increases,.

**(Continues on the next sheet)**

**SCMA 524 Term Project, report guide & analysis section**

**E-mail one Excel file to scma.stat@gmail.com by 9 AM Monday, December 11**

**Copy all partners with the submission e-mail. *I suggest that you do each item as we cover the material.***

**Prepare a report as an Excel file**. The report will have 15 items (plus the bonus if completed). Use the item numbers 1-15 **(**1-5are on the previous page) to **organize and order the report with each tab labeled by the item number(s) on the tab.** **Assemble charts, calculations and conclusions together in a coherent presentation on a worksheet (Use adjacent sheets if more than one is needed). DO NOT place these at the end as appendices.** Insert sentences and longer statements into a textbox. Give support for each answer. No credit will be given for answers without a supporting reason. You are to do the work for this project. You may obtain technical assistance and advice on any portion of the analyses, but **you must do the work.**  (**You may use consultants but may not use subcontractors**.) **Work submitted with your name(s) on it is a pledge that all non-referenced work was done by you or a teammate**. Use graphic illustrations, if possible, along with the analysis procedures we have learned in this class to answer questions. Some of the things asked may not be totally appropriate for your data, but perform the analysis anyway and explain why it would not be appropriate for your data. If you would be willing to allow me to use your data in the future, please indicate that I have permission to use your data. You will be credited for collecting the data if I do use your data in a future quiz, exam or example.

**6 Did time have an effect on the data?** If the data were collected over time or from different time frames, then do a run chart or plot for each quantitative variable and examine the categorical variables to look for time patterns.

**7. Are there any invalid data points in the original data that you are discarding for further analysis because they are not indicative of the phenomenon identified in number 3?** If you have any such points, list the points and for each point tell why you think it is not indicative and should not be retained in the data set for further analysis. D**o not use the discarded points for the remaining work for the parts from 8 through the bonus.**

**8. Is the mean of the response variable R the same for both values of C1?** (sections 13.1-5)Include your believed relationship statement 5a, appropriate null and alternate hypotheses to test 5a, appropriate graph(s), analysis results, and a written statement with your conclusions about the relationship between R and C1.

**9. Is the mean of the response variable R the same for all three values of C2? (**sections 20.6&7) Include your believed relationship statement 5b, appropriate null and alternate hypotheses, appropriate graph(s), analysis results, and a written statement with your conclusions about the relationship between R and C2. **If you conclude there is a statistically significant difference in the means, tell which of the three you think are different and why.**

**10. Create a cross-classification table to determine if there is a relationship between the two categorical variables. (**section 14.6) Give the 5f believed relationship statement, state appropriate null and alternate hypotheses to statistically test for a relationship between C1 and C2. Test this hypothesis with =.10, show analysis results and describe the nature of any significant relationship. Check for any violation of assumptions for hypothesis testing with this procedure.

**11. Is there a linear relationship between R and Q1?** (ch. 4 & 15) Include your believed relationship statement 5c, the appropriate null and alternate hypotheses to test 5c, XY scatter with fitted line, analysis results with a 90% confidence interval for the fitted slope with Y=R & X=Q1, and a written statement with your conclusions about your 5c belief.

**12. Is there a linear relationship between R and Q2?** (ch. 4 & 15)Include your believed relationship statement 5d, the appropriate null and alternate hypotheses to test 5d, XY scatter with fitted line, analysis results with a 90% confidence interval for the fitted slope with Y=R & X=Q2, and a written statement with your conclusions about your 5d belief.

**13. Select the quantitative variable, Q1 or Q2, with the stronger linear relation with the response variable R.** **Take the average of the first two values for this variable and predict the mean value of the response variable using a 90% confidence interval if the quantitative variable has the value of the average you calculated** (section 15.3 & 4)**.**

**14. Find the correlations between all three of the quantitative variables.** Give your believed relationship statement 5e. Is there a significant linear relationship between Q1 and Q2? Does the correlation support 5e? (4.3 & 15.2)

**15. Give multiple regression output for a response variable R using the information in C1, C2, Q1 and Q2 as predictors.** Create a dummy variable (section 18.1) to include the categorical variable C1 and two dummy variables to include C2. Give the regression results using five independent variables (three dummy variables, Q1 and Q2). Next give the results for the model you believe would be the best regression model (section 18.4) for predicting R and give the reasoning to support your choice for this being the best linear model.

**(3 point bonus) Use JMP to build a logistic regression model [17.6] for predicting C1 using Q1, Q2 and R as predictor variables. Tell which category of C1 is being predicted** (the category of C1 for which p^ is being calculated), **give the coefficient & conclusion about the significance of each predictor variable. Use the values of Q1, Q2 & R for the first observation and use the model to calculate p^. Tell why you think this model is useful or not.**