

Quant I - Problem Set II

Univariate and Bivariate Data

Kenneth Benoit

Assigned: Wednesday, October 13th, 2010

Due: Wednesday, October 27th, 2010

1. Load the campaign spending dataset with the command `load('dail2002.RData')` and attach it using `attach(dail2002)`. You can use `names(dail2002)` to see the variable names in the original data frame. After attaching the dataset, you can access copies of the variables directly by typing their names.
 - (a) Produce a side-by-side boxplot of the number of first preference votes received (variable *votes1st*) by incumbency status (variable *incumbf*). Add a label to the axis showing the number of votes. Briefly discuss the graph, explaining what the figure tells you about the distribution of the variables. (7 points)
 - (b) Calculate the variance of the number of first preference votes in R using the respective formula (use the formula for the sample variance with $n - 1$ as denominator). Calculate the sample standard deviation from the variance. Then verify the two results using the R functions `var` and `sd`. Note: Remove the missing values from *votes1st* before you start, e.g. with `votes1st.comp1 <- votes1st[is.na(votes1st)==FALSE]`, which creates a new vector containing only the cases where *votes1st* is not missing. (5 points)
2. Load the R Dataset EES04Trust and attach it using `attach(d)`. Similarly as in Task 1, you can use `names(d)` to inspect the names in the original dataset. Again, after attaching the dataset, you can access copies of the variables directly by typing their names. The dataset is based on the European Election Study from 2004. One of the questions in this survey was as follows: “Now I would like to ask you a question about how much trust you have in people from various countries. Can you please tell me for each, whether you have a lot of trust of [sic!] them or not very much trust.” The variable *trustire* represents the proportion of respondents answering that they have a lot of trust in people from Ireland.

- (a) Use the variable *country* to attribute names to the *trustire* variable. (1 point)
- (b) Use one or several R functions to calculate the minimum, maximum, mean, median, as well as the first and third quartile of the *trustire* variable. Briefly describe what these figures tell you (in statistical terms). (6 points)
- (c) Using an R command that includes a logical expression, calculate the proportion of countries in which respondents express more trust in Irish people than in Italy. (2 points)
- (d) Using an R command that includes a logical expression, calculate the proportion of countries in which respondents express less trust in Irish people than in Austria. (2 points)
- (e) Make R show you the *names* of the countries where respondents trust Irish people more than the Irish trust their own people. (2 points)
- (f) Using R, draw a histogram of the *trustire* variable. Use the options of the command to set the number of boxes to 10, and to make the histogram show probability densities, not frequencies. Overlay the histogram with a density plot. Briefly discuss the graph. (5 points)
- (g) In R, calculate z-scores for the *trustire* variable. Do **not** use the `scale()` command, but write your own code using the respective formula. (2 points)
- (h) Make R show you the *names* of the countries with a z-score smaller than -1. Using exact numeric statements, what does a z-score smaller than -1 mean in this example? (4 points)
- (i) Draw a dot chart for the *trustire* variable, which shows the names of the countries and where the data are sorted by their values. (2 points)
- (j) The variable *trustgb* is equivalent to *trustire*, but refers to trust in people from Great Britain. Produce a scatterplot of *trustgb* (on the x-axis) and *trustire* (on the y-axis). Using the `abline()` command, add a horizontal line (option h) at the mean of the *trustire* variable and a vertical line (option v) at the mean of the *trustgb* variable. Briefly discuss what this graph can tell you about the correlation of the two variables. (6 points)
- (k) Calculate the correlation between the two trust variables step-by-step with one of the formulae. Then verify your result with the

`cor()` function. (5 points)

3. The Irish National Election Study asked people about their satisfaction with the democratic process. The survey also includes information on respondents' area of residence.

Among people living in rural areas or villages, 120 said that they were "very satisfied", 844 said that they were "fairly satisfied", 147 said that they were "not very satisfied", and 31 said that they were "not satisfied at all". With regard to citizens from small or middle-sized towns, respective numbers were 71, 373, 73, and 15. In suburbs of large towns or cities, the figures were 16, 75, 25, and 3. Finally, for respondents from large towns respective numbers were 47, 326, 100, and 21.

- (a) Enter the data into R as a matrix. Add appropriate row and column names and show the table. (4 points)
- (b) Add margins to the table. (1 point)
- (c) Suppose you are interested in where people showing a certain level of satisfaction with democracy live. For this purpose, present the table with column percentages. (2 points)

Total: 56 points