1/15/14 Class Notes

Wednesday's lecture introduced some basic commands and graphing capabilities of R. With the Titanic example, *categorical data* showed how the survivors and non-survivors of the Titanic were categorized into groups while the UT 2000 example showed *quantitative data* including students' GPAs.

The concept of *bivariate data* was explained through *box plots, dot plots, and scatter plots* to show a relationship between two variables. For instance, the graphs easily portrayed a relationship between each college at UT (i.e. Business) and students' GPAs. The within-group variability showed how the GPAs of Business students related to one another within the Business school and between-group variability showed how GPAs of Business students in their respective college compared with that of other colleges at UT.

I. Step-by-Step: Using R

a. Open the pre-downloaded files: File → Open File

i. Make sure that the file is an R file (Titanic.R)

ii. R files are scripts that may contain directions and commands to be used in the assignment

b. Install R packages ("effects" & "mosaic"). Packages need to only be installed once. Once installed, they need to be loaded each time you open R

i. At the right-hand side of R Studio, look for the tab "Packages"

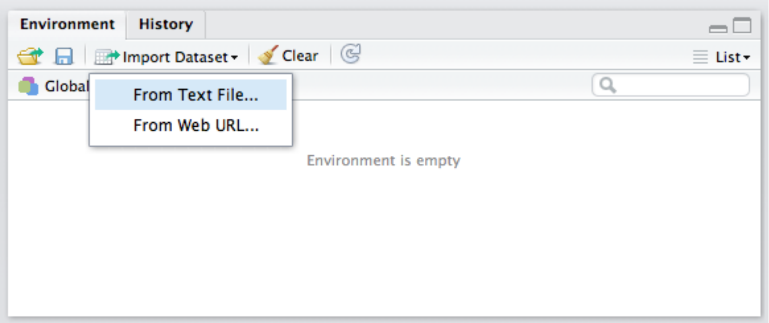
ii. Choose "Install Packages"

iii. In "Packages" section, type "mosaic"

iv. To reload, type >library(mosaic) in the console

v. Type the above command into the console OR highlight this command in the script and press CTRL+Enter (for PC) or Command+Enter (for MACs)

c. To import data to R



Environment is empty

i. If dataset can be downloaded in .csv format:

1. Environment→Import Dataset→From Text File

ii. If dataset is accessed through a web link:

1. Environment→Import Dataset→From Web URL

II. Categorical v. Quantitative Data

a. **Categorical Data**

   i. TitanicSurvival.R: Who survived and who died during the sinking of the Titanic?

      1. Data can be grouped into those who "survived" and "died"

b. **Quantitative/Numerical Data**

   i. Data that can be measured numerically (height, length, etc.)

III. Basic Commands (This example uses the TitanicSurvival.R file)

a. >names(TitanicSurvival) returns variable names. This reminds you of the spelling and capitalization of each variable within the dataset

b. >head(TitanicSurvival) returns the first 6 lines of the data set

c. >tail(TitanicSurvival) returns the last 6 lines of the data set

d. >xtabs(~survived + sex, data=TitanicSurvival) gives you a cross tabulation of variables "survived" and "sex" from the data retrieved by TitanicSurvival.csv file

   i. "~" or the tilde symbol means stratify by or model by

e. >tally(~survived + sex, data(TitanicSurvival) gives you the row and column totals for each variable

f. >tally(~survived + sex:passengerClass, data=TitanicSurvival) is used to stratify by two categories

   i. ":" is used for interaction or cross between factors

IV. Simple Summaries and Graphics (This example uses the ut2000.R file)

a. Good graphs

   i. Facilitate comparison

   ii. Are multivariate

   iii. Are truthful about magnitude

   iv. Usually not for small data

b. >summary(ut2000) returns summary statistics. Includes minimum, 1st quartile, median, 3rd quartile, maximum, and mean for quantitative data. For categorical data, this command will return counts for each variable.

c. >histogram(ut2000$GPA) returns density histogram with x axis bins using the GPA variable and y axis being the density

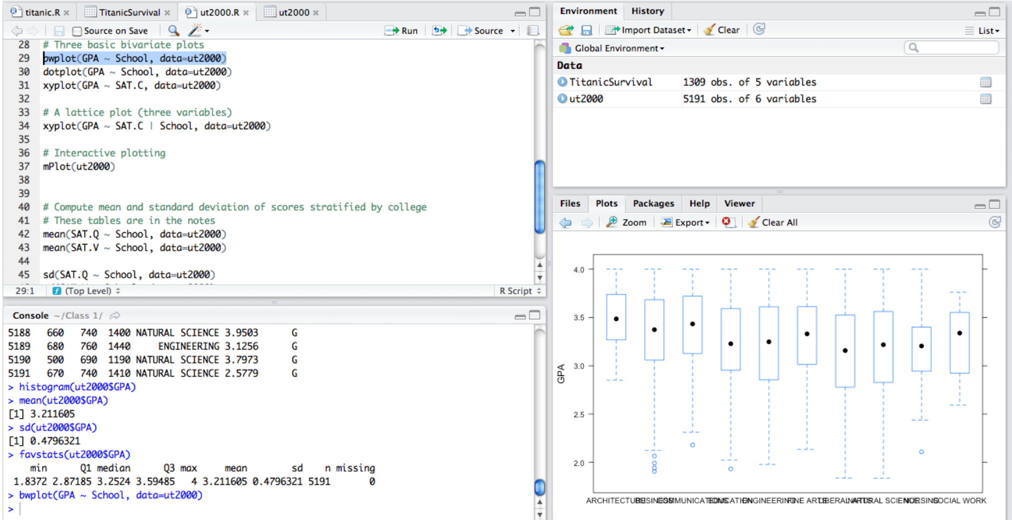
   i. Histograms help you understand the variability within a data set

d. >mean(ut2000$GPA) returns the mean GPA from the ut2000 data set

e. >sd(ut2000$GPA) returns the standard deviation for GPA from the ut2000 data set

f. >favstats(ut2000$GPA) returns the minimum, quartile 1, quartile 3, maximum, standard deviation, and n missing for the GPA variable from the ut2000 data set

V. Graphs



```
# Three basic bivariate plots
bwplot(GPA ~ School, data=ut2000)
dotplot(GPA ~ School, data=ut2000)
xyplot(GPA ~ SAT.C, data=ut2000)

# A lattice plot (three variables)
xyplot(GPA ~ SAT.C | School, data=ut2000)

# Interactive plotting
mPlot(ut2000)


# Compute mean and standard deviation of scores stratified by college
# These tables are in the notes
mean(SAT.Q ~ School, data=ut2000)
mean(SAT.V ~ School, data=ut2000)

sd(SAT.Q ~ School, data=ut2000)
```

```
5188   660   740  1400 NATURAL SCIENCE 3.9503      G
5189   680   760  1440     ENGINEERING 3.1256      G
5190   500   690  1190 NATURAL SCIENCE 3.7973      G
5191   670   740  1410 NATURAL SCIENCE 2.5779      G
> histogram(ut2000$GPA)
> mean(ut2000$GPA)
[1] 3.211605
> sd(ut2000$GPA)
[1] 0.4796321
> favstats(ut2000$GPA)
    min      Q1 median      Q3 max     mean        sd    n missing
 1.8372 2.87185 3.2524 3.59485   4 3.211605 0.4796321 5191       0
> bwplot(GPA ~ School, data=ut2000)
> 
```

```
Data
TitanicSurvival    1309 obs. of 5 variables
ut2000             5191 obs. of 6 variables
```

a. **Box plots**: bivariate graph that can compare within-group variability and between-group variability
   i. >bwplot(GPA ~ school, data=ut2000) returns a plot for each school with the GPA on the Y axis

b. **Dot plots**: bivariate graph with each data point representing a dot. Dot plots are better for smaller data sets so you can distinguish one dot from another
   i. >dotplot(GPA ~ school, data=ut2000) returns a dot plot for each school with GPA on the Y axis

c. **Scatter plot**: bivariate graph that can accommodate large sets of data
   i. >xyplot(GPA ~ SAT.C, data=ut2000) returns a scatter plot with SAT.C on the x axis and GPA on the Y axis

d. **Lattice plot**: used for 3 pieces of information
   i. >xyplot(GPA ~ SAT.C|school, data=ut2000) returns scatter plots stratified by college

e. **Interactive plotting**:
   i. mPlot(ut2000) → must install another package

VI. Means & Standard Deviations

a. >mean(SAT.Q ~ school, data=ut2000) computes the mean SAT.Q stratified by school

b. >sd(SAT.Q ~ school, data=ut2000) computes the standard deviation of SAT.Q stratified by school

c. Satqmeans=mean(SAT.Q ~ school, data=ut2000) stores the results of computations in other variables.
   i. In this case, "satqmeans" will return the computation of mean SAT.Q stratified by school. When you enter this command, nothing new will

appear in the console box. However, if you type >satqmeans the computation will appear

ii. Useful in intermediate computation to be used in subsequent computation

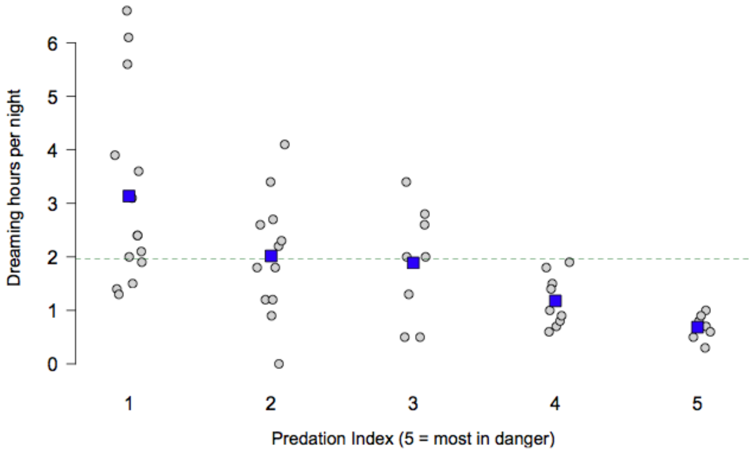
VII. Fitted Values and Residuals

a. Actual value = fitted value (predicted by the model) + deviation of that case (from the prediction)

b. Observed value = model value + residual

c. $Y_i = Y_i$ (hat) + e_i

i. "Hat" is this symbol: ^. It is the generic notation for "predicted"



d. In the graph above, the blue dot is the group mean (fitted value) and the deviation is the distance between the gray data point and the blue dot