Math 130, Section 1 Fall 2011

Lab 1

Goal: To start becoming familiar with the R commander (Rcmdr) interface, and to use it read in data and create some graphical summaries of data.

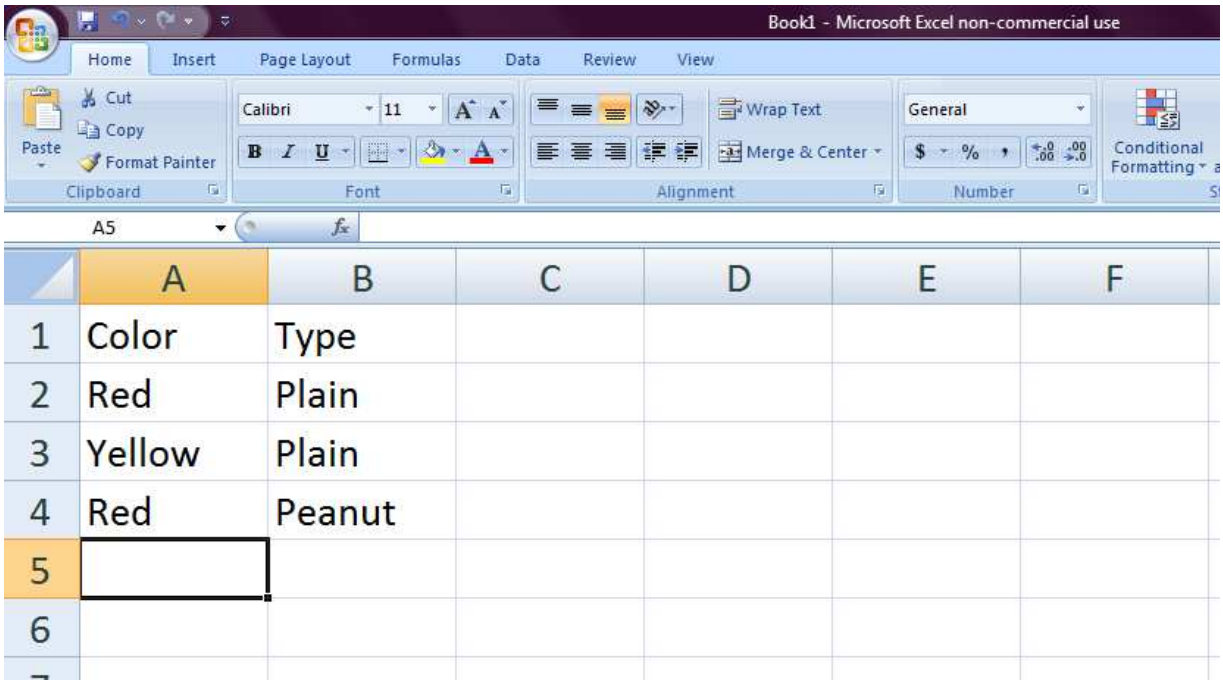
In this semester, you will learn how to use the statistical software **R** and its package **Rcmdr** to implement most of the statistical techniques you learn from lectures and readings. Currently the top four most popular statistical programs (based on my knowledge) are **R,** SAS, SPSS, and Minitab. **R**, however, is the only one among the four which can be used freely, and has become extremely popular. The only concern with **R** is that it might not be user-friendly enough to new users like most of you.v To accommodate this shortcoming of **R**, we will also introduce an **R** package, named **Rcmdr** (R commander), which uses a simple and easy-to-use menu/dialog-box interface. Both R and Rcmdr are installed on computers on campus for your use. If you want, you may try to install them on your laptop today in lab or on your personal desktop later. Brief installation instructions are posted at the course website. Do not hesitate to ask for my help at any time!

For more details about **R**, see http://www.r-project.org/.

To open **R**, click on **Start > All Programs > Course-related > R > R 2.9.1**. Next, to start Rcmdr, select **Packages** at the top menus, click **Load package**….and then select **Rcmdr** in the list. Click OK. Then a separate window titled "R Commander" will open. For today's lab activity, you will need to click back to the original R window to see most graphs (except Stem and Leaf plots), while the numerical outputs will often show in the Output Window under R Commander.

PART 1: Collect and Analyze M&M data

We'll start with a lunch-related question: Is the *color* distribution of M&Ms independent of *type* of candy? First of all, find a partner to work with (at least for Part 1). Second, open Excel. Next, break open bags of plain and peanut M&Ms, count the colors and enter your data in the Excel spreadsheet, which should look like:



[Warning!] **R/Rcmdr is case sensitive**, so when entering your data in the Excel spreadsheet, be careful! Also make sure that you don't give additional space at the end of each word.
i.e. **each row represents one observation** (one M&M candy), except the first row which specifies the names of variables.

Q1. Is the variable **Color** a categorical or quantitative variable?

Please save your M&M data set somewhere (Desktop, U drive, your USB, etc; you may want to save it in your U drive for future use).

To open it in **Rcmdr**, follow the following steps:
1. Click Rcmdr Window, and select **Data > Import data > from Excel, Access, or dBase data set…**.
2. In the window that opens, enter a **name** you would like to call your data set, e.g. *MMdata*. Click OK.
3. Find the data set file from wherever you saved it, click **Open**. In the window that opens, select the spreadsheet you saved your data (if you didn't change anything when saving your data in Excel, you would probably select "Sheet1"), and then click OK.

Now that the data has been collected, feel free to dispose of it an appropriate way.

Recall that an easy way to describe a single categorical variable is to report what is the *sample proportion* (or percent) that fall into each category. For example, suppose we are interested in the variable ***Color***. In Rcmdr, select **Statistics > Summaries > Frequency distributions…** and select the variable *Color* in the "Variables" field. Click OK. This gives us a frequency table and a relative frequency table of the variable ***Color*** in the *Output Window* under Rcmdr.

Q2. Give the frequency table you got as the answer to this part.
How many candies in your bag are *yellow*?

Q3. Check the relative frequency table. What percent of candies in your bags are *red*?

Now let us make a graphical display of the data for the Color variable. For a single categorical variable, a **Bar Graph** is a good graphical summary. Again using Rcmdr, select **Graphs > Bar Graph…**. In the "Variable" field, select the variable *Color*. Then click **OK**. A bar graph should appear in your original **R** window (instead of in the window of Rcmdr). The bar graph makes the color percentages in the previous table easier to visualize.

Q4. Which color was the most common in bags you got?

Instead of describing a single variable, we are often interested in a relationship between two categorical variables. For example, we might want to see if there is a relationship/association between M&Ms' color and type/flavor. To do so, we need to construct a **two-way contingency table**. Using Rcmdr, select **Statistics > Contingency tables > Two-way table**…. In the "Row variable" field select *Color*, and in the "Column variable" field select *Type*. Select *No percentages* and **uncheck** *Chi-square test of independence* under Hypothesis tests. Click OK.

Q5. Give the two-way table shown in the Output Window of Rcmdr as the answer to this part.

The table given above contains the counts for each of all possible combinations of the variables. Use this two-way table to answer the following questions:

Q6. What percent of peanut M&Ms are yellow?

Q7. What percent of yellow M&Ms are peanut?

Q8. What percent of your M&Ms are yellow peanut candies?

Repeat above procedure but select *Row percentages / Column percentages / Percentages of tota*l instead. Use the corresponding three tables to verify your answers to Q6 – Q8 above.

Q9. Give the *conditional distribution* of **Color** for a given **Type** as the answer to this part.

Q10. Is the color distribution of M&Ms independent of type of candy? Explain.

PART 2: Describing Two Categorical Variables (Another example)

200 adults shopping at a supermarket were asked about the highest level of education they had completed and whether or not they smoke cigarettes. Results are summarized in the table below. Is there an association between education level and smoking? Why do you think so?

	Smoker	Nonsmoker	Total
High School	32	61	93
2-yr College	5	17	22
4-yr College	18	72	90
Total	55	150	205

PART 3: Describing some Variables in the First Class Survey

Following the instructions in Part 1, import the data set *ClassData*. (The original file is available on our course website

Q1. Pick one Categorical variable and display its distribution with a Bar Graph: Select **Graphs > Bar Graph…**. Find the graph in the original R window. Use one or two sentences to describe the distribution of this Categorical variable.

Q2. By hand, draw a stem-and-leaf plot of the "Height" variable. Include that hand-drawn graph here, and compare your work with the plot from R: Select **Graphs > Stem-and-leaf display…**.Select the corresponding variable in the "Variable" field. Then click OK. Use one or two sentences to describe the distribution of this variable.

Q3. Draw a histogram of "Height" by hand, and include it here. Compare your work with the histogram from R: Select **Graphs > Histogram…**. Select the corresponding variable in the "Variable" field. Then click OK. Use one or two sentences to describe the distribution of this variable. The histogram generated by R should look like the one you drew by hand. If not, explain the possible reason(s).

Q4. Give a two-way table of Gender vs. Politics. Is there an association between these two categorical variables? Why or why not?