Guide to Analysis with Rcmdr for Math 17

Amy Wagaman, Last Edit: 9/4/09

As you proceed through Introduction to Statistics, you will learn statistical techniques, which will be implemented via the software R and package Rcmdr. These programs are available on computers on campus and are also free if you want to install them on a personal computer. Detailed installation instructions are on the course CMS website.

1 R/Rcmdr: General Introduction and Getting Started

R is one of many programs that can be used for data analysis. Others include Minitab, SPSS, Excel, and SAS (among many others). R, however, is free and has the benefit that people write packages as they develop new techniques for analysis to allow others to implement their methods in R. The main problem with R is that it is not user-friendly to new users since it doesn't have a good interface. To alleviate that problem, we will be using Rcmdr, which is an R package that gives R an appearance much like Minitab, Excel, and SPSS, with pull-down menus.

Open R on your computer from under Programs and the logical submenu for Math 17. It will likely be version 2.9.1 (or 2.9.0). Once R opens you won't see much at all. There will be some text and a $>$ prompt at the left side of the screen. Now, you'll want to start Rcmdr. From the menus at the top, select packages and then load package and find Rcmdr in the list and select it. It will open a separate window for R Commander (Rcmdr, abbreviated throughout). You may have to click back to the original R window occasionally to see graphs that are generated, but the rest of the output will appear in the Rcmdr screen.

2 Handling Data

Data management is an important skill for a statistician. While most of the data sets for the course have undergone substantial pre-processing to make them easy to deal with (few if any missing values, already well set up, etc.), you will still need to read in data and do a few data management tasks at various times as outlined below.

2.1 Opening a data set saved somewhere

Most data sets for the course will be provided on the course website. You should save the data set somewhere (U drive, Desktop, etc.) and then open it from inside Rcmdr as follows:

1. From the Data menu, select Import Data and from the submenu, select from file, clipboard or URL.
2. In the window that opens, change the data set name to whatever you would like to call it.
3. The other options in the window will not need changed for data sets provided in class. If you enter your own data, you will not need the "variable names in file" if you do not start off the data set with the variable names.
4. Clicking Okay will bring up a directory browser for you to find the file where you saved it.
5. Find the file and then click Open in order to load the data set into Rcmdr.

2.2 Opening a data set already in R

A few times during the semester we will work with a data set that is provided already in R and hence in Rcmdr. Opening these data sets can be accomplished as follows:

1. From the Data menu, select Data in Packages and from the submenu, select Read data set from an attached package.
2. In the window that opens, find the box right after the OR that says "Enter name of data set" and enter in the name provided (examples: airquality, iris).
3. Clicking Okay will load the data set into Rcmdr.

2.3 Entering a data set

There may be times (especially when it comes time for your projects) when you will want to enter your own data into Rcmdr. This is very easy to do.

1. Under the Data menu, simply select New data set.
2. In the window that opens, type in the name you want to use for the data set. Then hit Okay.
3. A new window will open which is the blank data set. You may enter data like it was an Excel spreadsheet, and enter variable names, etc. in the boxes at the top. When you enter in variable names, you will be able to set the variable as either numeric or character (meaning categorical represented with letters).
4. Once you have finished entering the data, simply hit the "X" button on the data spreadsheet. This uploads what you have entered as the data set with the name you provided. If you need to edit any entry, follow the directions below on editing.

2.4 Editing/Viewing data

Once a data set is loaded into Rcmdr, you can view the data set by hitting the View data set button in the top middle of the Rcmdr interface. It will appear as a spreadsheet, but you will be unable to edit the entries. If you wish to edit the data, hit the Edit data set button to the left of the View data button (alternatively, you could also just hit edit to start with, assuming you don't change anything accidentally). Editing the data basically works like an Excel spreadsheet. When you are finished editing, simply hit the "X" to close and save the changes. Please note that "undo" does not undo editing changes, so anything you change will stay until you change it back.

2.5 Swapping between data sets

If you have loaded several data sets, you may swap between them at any time. To do so, simply:
1. From the Data menu, select Active data set.
2. From the submenu that appears, choose "Select active data set".
3. A window will appear with a list of all currently loaded data sets. Click on the one you want to use and then click Ok.

This is especially helpful if you have stacked some variables (see below) and want to go back to the original data set to make certain graphs (see graphs below).

2.6 Converting variable types

Statisticians often use a numerical coding (0/1 or 1/2/3) for levels of categorical variables (yes/no, categories 1,2,3, etc.). Rcmdr needs to know what variables you want treated as categorical if they have a numerical coding. Note this issue will only occur if (for example) gender was coded as 1/2 instead of m/f. To set a numerical variable as a categorical variable, follow these steps:
1. From the Data menu, select Manage variables in active data set.
2. From the submenu, select Convert numeric variables to factors.
3. In the new window that opens, select the variable you are swapping to a factor.
4. You may elect to have the factor version of the variable have a new name, in which case you should change the name in the small text box.
5. You also can supply the level names (i.e. maybe you want the levels to be male/female) or use numbers (i.e. it will keep the current coding as the names for the categories). The default is supply level names.
6. Click Ok to do the conversion.
7. If you choose to supply level names, a new window will open when you hit Ok and prompt you to enter the different level names. Enter them and click Ok and it will complete the conversion.

2.7 Stacking

Stacking is a necessary data management action. Certain functions will only work when data is supplied in a certain form. For example, when comparing two populations via two samples, you may need to have the data in one of these 2 common formats: two columns - one containing the data from sample 1 and the second with the data from sample 2 OR two columns - one containing the variable of interest and the second containing an indicator variable for whether that value is from sample 1 or sample 2. From here on, we will call having group data in separate columns Format 1 and the second format will be denoted Format 2. Stacking is when you start from Format 1 and convert to Format 2. Since you basically "stack" the group data and create a new variable to denote which group it came from originally. To stack group data, follow these steps:
1. You need to be sure your data is in Format 1 (or at least the variables you want to stack should be in this format).
2. Under the Data menu, select Active data set.
3. In the submenu, select Stack variables in active data set.
4. A new window will open. In that window, select the variables you want to stack (select more than one by holding down ctrl as you click).
5. Now, set the name for a new data set where the stacked variable will be created, as well as a name for the variable and for the groups (this is the factor name: Type, Group, or just leaving it as factor are acceptable, but you might have something more specific, like gender or species).
6. Click Ok to perform the stacking.
7. The stacked variable data set will now be your active data set. Remember you can swap back to the original data set by following the steps above.

2.8 Compute New Variable

At times you may want to compute new variables. This is especially important for when we get to paired data, where you are interested in the difference between 2 variables (for example, posttest-pretest scores).

1. Computing new variables is done using the Data menu.
2. Select Manage variables in active data set, and then Compute new variable.
3. A new window will open where you can enter the name for the variable you want to compute and then the expression for how it is computed based on pre-existing variables.
4. Clicking Ok will add the new variable to your data set.

2.9 Miscellaneous Data Management Actions

Briefly, here are a list of other data management actions you may wish to use at some point. If you want to do any of these, you may be able to figure it out on your own, but feel free to ask me for help at any time.
1. Subsetting - can undo stacking basically but not as effectively as you may like.
2. Remove outliers - basically remove rows (since rows are observations) - after you figure out which the outliers are and have a justifiable reason for removing them.
3. Remove cases with missing data - sometimes data sets have missing data and while there are some methods to handle this and impute missing values, you may just want to remove the cases with missing data and continue with your analysis. Some methods will not run if the data set has missing values.
4. Set case names - if you have unique row names (for example, data by country), and you want to set the row names, they are called "case" names.
5. Bin numeric variable - basically converts a numeric variable to a categorical one where you get to specify how many categories you want.

3 Graphs and Descriptive Statistics

A preliminary analysis for a data set involves obtaining basic graphs and summary statistics for the variables of interest. The graphs discussed below are for quantitative variables, while summaries can be obtained for both quantitative and categorical variables.

3.1 Stem and Leaf

Stem and leaf plots are created under the Graphs menu by selecting Stem and leaf display. In the window that opens, there are many options to customize the graph. However, just selecting a variable and clicking Ok will generate a basic stem and leaf graph in the Rcmdr window. Of the options, the main one you may want to uncheck is trim outliers. The rest allow you to play with the settings for how the stem and leaf is displayed (leaf units, how many leafs per stem, etc.). Note that there is no option to plot by groups, so you will need your data to be in Format 1 if you want to obtain stem and leaf plots for the same variable for different groups.

3.2 Histogram

Histograms are created under the Graphs menu by selecting Histogram. In the window that opens, you may set a certain number of bins or leave it as automatic (recommended unless you are doing a comparison where you want to control bin size across multiple groups). You can also set the y-axis to be frequency counts, percentages, or densities. Frequency counts make the most sense as a default, but statisticians often use density as well. You may pick either, it simply changes the

scaling of the y-axis but not the shape of the graph. Note that there is no option to plot by groups, so you will need your data to be in Format 1 if you want to obtain histograms for the same variable for different groups.

3.3 Boxplot

Boxplots are created under the Graphs menu by selecting Boxplot. Select the variable of interest and click Ok. The only option available is if you want to be able to identify outliers by clicking your mouse on them (it will display the observation number in the data set for the outlier). Note that there IS an option to plot by groups, so you will need your data to be in Format 2 if you want to obtain comparative boxplots all in the same plot. You can still obtain boxplots for each group separately if your data is in Format 1. If you select the plot by groups option, a small window will appear where you pick the group variable (in case you have several). Then click Ok to return to the boxplot window and Ok again to generate the comparative boxplot.

3.4 QQ Plot

QQ plots are created under the Graphs menu by selecting Quantile-comparison plot. In the window that opens, select the variable you want for the plot. You can opt to be able to identify outliers with the mouse (it will display the observation number in the data set for the selected points). Finally, you must choose which distribution to compare to. The default is the normal distribution, which is what you want for a QQ plot when checking normality. Note that there is no option to plot by groups, so you will need your data to be in Format 1 if you want to obtain QQ-plots for the same variable for different groups.

3.5 Descriptive Statistics

There are 2 ways of getting descriptive statistics depending on which ones you want for a given variable.

Method 1:
Under the Statistics menu, select Summaries and then select Active data set. This provides a basic description of all variables in a data set (both quantitative and categorical). For quantitative variables, you will get the min, Q1, mean, median, Q3, and max. For categorical variables, it will output a brief frequency table (which is shortened if too long). Note this method does not give the standard deviation, but you could compute the IQR and range from the given values.

Method 2:
Under the Statistics menu, select Summaries and then Numerical Summaries (for a quantitative variable) OR Frequency distributions (for a categorical variable). For frequency distributions, a new window opens where you select the categorical variable of interest (and leave the option for a test unchecked) and click Ok to generate a table of counts and another table of the percentages in each category. For numerical summaries, a new window opens where you select the variable(s) you want as well as the statistics you want. Note that standard deviation is provided by default and the rest of the statistics are the same as you would obtain with Method 1. Note that there is a summarize by groups option which works the same way as the boxplot plot by groups option and would require the data to be in Format 2 to use.

4 Hypothesis Testing and Confidence Intervals

Rcmdr will generate hypothesis test output and confidence intervals for all the scenarios we have described. In this section, when I refer to "tests" I mean "tests and CIs". The confidence level can be set in the test window by changing the confidence level from the default of .95 in each of the cases below.

4.1 Proportions

Rcmdr can do tests on proportions if you have the data entered as 0s and 1s and then convert it to a factor variable. However, this is not how most proportion problems will be summarized for you, so you should continue doing these by hand. If you determine that you need to run a proportion test and have data in this format, and need help to get it to run, please ask. This may come up if you are doing supplemental tests as part of your project. For a 2-sample proportion test, the data would need to be in Format 2 (one variable is 0/1 for the yes/no responses, and the other variable denotes the 2 groups).

4.2 One Mean

To perform a one-sample t-test, under the Statistics menu, select Means and then select Single-sample t-test. In the window that opens, select the variable of interest. Then, you need to set several options: the value of μ_0 in the null hypothesis, whether your test is 2-sided (first option) or one-sided to a certain direction (second and third options), and what confidence level you want (for CIs). When you have selected all options and a variable, click Ok. Note that there are defaults for the procedure, namely it will test $H_0 : \mu = 0$ vs. $H_A : \mu \neq 0$ and use 95% confidence for CIs.

4.3 Two Means

For the 2-sample t-test (test for 2 means), the data needs to be in Format 2 (one variable is the quantitative response and the second variable denotes the groups). Note that you probably start off with the data in Format 1 for checking assumptions then stack it into Format 2.

To perform a two-sample t-test, under the Statistics menu, select Means and then select Independent samples t-test. In the window that opens, select the variable of interest as the response variable and the group variable (first box). Then, you need to set several options: whether your test is 2-sided (first option) or one-sided to a certain direction (second and third options), what confidence level you want (for CIs), and whether or not to assume equal variances (usually this is not assumed). When you have selected all options and both variables, click Ok. Note that there are defaults for the procedure, namely it will test $H_0 : \mu_1 - \mu_2 = 0$ vs. $H_A : \mu_1 - \mu_2 \neq 0$ and use 95% confidence for CIs with equal variances not assumed.

4.4 Paired Mean

For the paired t-test, the data needs to be in Format 1 (example: one column for the "before", one column for the "after" values), or you need to have already computed the differences and can run a one-sample t-test on them. You will need to compute the differences anyway in order to check

some assumptions.

To perform a paired t-test, under the Statistics menu, select Means and then select Paired t-test. In the window that opens, you will need to select the two variables. The test is done on the first variable minus the second (so you may prefer a certain order of subtraction). After setting that up, you need to set several options: whether your test is 2-sided (first option) or one-sided to a certain direction (second and third options), and what confidence level you want (for CIs). When you have selected all options and both variables in the order you want, click Ok. Note that there are defaults for the procedure, namely it will test $H_0 : \mu_d = 0$ vs. $H_A : \mu_d \neq 0$ and use 95% confidence for CIs.

5 ANOVA

For an ANOVA, the data needs to be in Format 2 (one variable is the quantitative response and the second variable denotes the groups). Note that you probably start off with the data in Format 1 for checking assumptions then stack it into Format 2.

To perform an ANOVA, under the Statistics menu, select Means and then select One-Way ANOVA. In the window that opens, you can change the name of the model (or leave it). Then you get to select the groups variable from the list of categorical variables and the response variable from the list of quantitative variables. The only other option is for pairwise comparisons of means (these are multiple comparisons). You can either request them when you run the ANOVA or come back and rerun the ANOVA to obtain them once you have determined they are appropriate to have. It does not hurt to generate them and then ignore them if you find you cannot reject the ANOVA null hypothesis. Clicking Ok will put the ANOVA output in the Rcmdr window. If you did select the multiple comparisons output you will also see that in the Rcmdr window and in a separate graph that plots the pairwise confidence intervals.

6 Regression

For linear regression, you can obtain scatterplots, correlations, the regression output, and the residuals to help perform your assumption checks following the steps below.

6.1 Scatterplots and Correlations

For a scatterplot, simply select Scatterplot from under the Graphs menu. You can then select which variable is on which axis in the window. You will see three options checked. You will want to uncheck the "smooth line" option, but can leave the other two. The least-squares line option will print the regression line on the plot for you (even before you fit the model for it), and the boxplot option simply puts boxplots on the edges for each variable. Note there is a plot by groups option if you want to know if you have different regression lines for two or more groups based on some categorical variable.

For correlations, you can either obtain the correlation between 2 variables or a correlation matrix showing the correlations between many variables. Both are done using Statistics then Summaries then selecting Correlation matrix. In the window, picking only two variables will give you the correlation between just those 2. Picking 3 or more will result in a matrix format for the output

with more correlations. Remember to hold control to select several variables. You will want to leave the option on Pearson, though Spearman correlations are similar (partial is something else entirely).

6.2 Regression Output

A linear regression is performed by selecting Fit models under the Statistics menu and then selecting the first option of Linear regression. A window will open with your regression options. The model will have a default name (you can change if you want). You may select only one response variable but one or more explanatory variables (i.e. it goes beyond simple linear regression). Select these, then select Ok. The subset option is there if you want to run a regression on different groups only.

The regression output will appear in your Rcmdr window. You have now successfully fit the regression model, and can do several additional things.

You may want to be able to repeat the model summary (if you do other output and want to see it again without scrolling). To obtain it again, simply select Summarize model from the Models menu (if you have multiple models fit you may need to select your active model first).

You may want to obtain confidence intervals for the slope (and intercept). To obtain these, select Confidence Intervals from the Models menu and then select the confidence level desired in the window.

There are many other options in the regression Models menu, but we turn our attention now to checking assumptions.

6.3 Assumption Checks

For most of the assumption checks, we will need access to the residuals. You should note that Rcmdr has some basic diagnostics built in. In fact, if you select Graphs under the Models menu, you can see many options (feel free to see what these are on your own). However, we can also just save the residuals to the data set and construct the QQ plot and residual plot using the Graphs menu.

To save the residuals to the data set, under the Models menu select Add observation statistics to data. A window will open with many options checked. You only need either the residuals or standardized residuals (called studentized), so you can uncheck many of these. If you save fitted values, these will give you the predicted y's for the x's in the data set. Once you click Ok, the selected statistics will be added to the data set, and you can check by viewing it if you like.

To check for normality of the population error terms, we make a QQ plot of the residuals. Under Graphs, select Quantile-comparison plot and then select the residuals variable. Click Ok to obtain the plot.

Finally, to make a residual plot, use the Scatterplot option under the Graphs menu and select the residuals as your Y axis and the predictor variable as your X axis.

7 Chi-square

7.1 Goodness of Fit

Rcmdr does not do this easily. Continue doing it by hand.

7.2 Homogeneity and Independence

Both can be done using the same procedure. Under the Statistics menu, select Contingency tables and then the last option of Enter and analyze two-way table. A window opens. Select the appropriate number of rows and columns for your table, then enter the observed counts you have. You can ask for percentages to be computed or leave that option as no percentages. Leave the Chi-square test of independence selected, as this provides the test statistic for both procedures. You should check the print expected frequencies to obtain the expected counts in order to check your assumptions. Finally, you may want the components of the chi-square statistic so you can see which cells contribute the most to the statistic, but this is not strictly necessary. Click Ok and the output will show in the Rcmdr window. You may need to scroll a bit to get to the parts you want.