R for SPSS Users: A User Friendly Introduction Using R Commander

AAPOR Webinar Series
April 27, 2016

Trent D. Buskirk, Ph.D.
Marketing Systems Group
Instructor

Outline for the Short Course

- A quick overview of R and R Commander
- Loading R Commander into an R Session
- Exploring the R Commander Graphical User Interface
- Introduction to Basic Data Formatting/Editing using R Commander
- Introduction to Descriptive Statistics Using R Commander
  - Means, proportions
  - Contingency Tables
- Introduction to basic graphics with R Commander
What is R?

Besides being the 18th letter in the English Alphabet, R is a language and environment for statistical computing and graphics. (i.e. free version of S)

R commands are executed via scripts/syntax in an interactive environment.

R functionality is extended by packages that are written for specific statistical/data management/machine learning and modeling tasks.

What is RCommander?

R Commander is a bundle of R packages that essentially overlays a graphical user interface (GUI) on top of the R environment.
Some Key Online Resources

**R-Seek (www.rseek.org):** Think Google specifically for R.
- Can search for program syntax, packages, R reference cards and more specifically related to R.

**Quick R (http://www.statmethods.net):** A very top-level synopsis of R commands organized by function:
- Stats, Input, Graphics, etc.
- Contains both package information (i.e. names, descriptions) as well as code snippets.

**R Blogspot (http://www.r-bloggers.com):** Meta Blogging site specifically dedicated to R users.

---

Working with R

Master Toolbar
Menu and Shortcut icons

R Console
Window

R Command Prompt

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale.
R is a collaborative project with many contributors.
Type 'contributors()' for more information and 'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or 'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

 Revolution R Community version 6.2 (64-bit): an enhanced distribution of R Revolution Analytics packages Copyright (C) 2013 Revolution Analytics, Inc.
Type 'revo()' to visit www.revolutionanalytics.com for the latest Revolution R news, 'forum()' for the community forum, or 'readme()' for release notes.

[Previously saved workspace restored]

>
Steps 1 and 2: Launch R and Select Install Packages

Steps 3 and 4: Select Comprehensive R Archive Network (CRAN) and then the Packages you need (use ctrl to select multiple packages)

We want: Rcmdr
Step 5: Load/Launch Packages (Rcmdr)

This list contains all of the packages that you have previously installed (or are installed by default in R)

Once you install a package, you do not need to re-install it. You just need to load it (launch) when you start R.

A Short Cut for Launching Installed Packages

After you have installed a package in R it stays in memory from one launch of R to another.

However, generally each time you launch R, you need to re-load packages (such as R commander).

You can select Packages → Load package... and select the package name you want to launch (i.e. Rcmdr) ... OR...

You can type library(PackageName) at the command prompt (>) [e.g. > library(Rcmdr)]
A Quick Overview of R Commander

R Commander allows users who are familiar with the GUI of SPSS and EXCEL and JMP to access the functionality of R via a window driven graphic user interface.

While R scripts are generated that correspond to commands accessed and selected via a series of drop-down menu selections (similar to paste syntax in SPSS).

The script versions of the commands appear in the Script window and can be copied, edited and rerun by highlighting and submitting (like SPSS syntax window).

Orientation to the R Commander GUI

1. Toolbars and Data Sets Index
2. Script Pane (commands pasted here or typed) To execute commands entered in this section, highlight them and then click on the submit button
3. Output Pane Commands pasted in Script Window are executed here. If using R studio, the output window is not included, but is part of the R Studio Console Window
4. Messages Pane (Warnings/Errors posted here)
Exploring the R Commander Toolbars

The File Menu Options For:
- Loading/Saving Script files
- Saving Output and R workspaces
- Exiting

Edit Menu Options include:
- Copy, cut, paste options for both script and output window contents
- Right click options are also available within each of the panes

Exploring the R Commander Toolbars, Cont.

The Data menu contains options for:
- Creating new R data sets
- Loading data sets that exist on your computer already as well as data that are contained in other R packages
- Importing Data from other Databases/Statistical packages
Exploring the R Commander Toolbars, Cont.

The Statistics menu contains options for:
- Summary Statistics for continuous and categorical data
- Means, Proportions, and Variance Summaries/Tests
- Various Nonparametric statistical tests, cluster analysis and linear models

The Graphs menu contains options for:
- Generating various types of graphs including:
  - Histograms, Boxplots, QQ plots, 3-D plots, Scatterplots, Bar graphs, Pie Charts.
- Exporting/Saving graphical output
- Controlling basic colors of the plots

The Models menu contains options for:
- Model summaries, diagnostic statistics, hypothesis testing, residual plots and confidence/prediction intervals
Exploring the R Commander Toolbars

The Distributions Menu Contains Options for:

- Generating random observations, probabilities and quantiles for various continuous and discrete probability distributions

The Tools Menu Contains Options for:

- Loading additional packages into R
- Additional R Commander options (e.g. font and output controls)

Example 1 – Importing Excel Files Into R

The Excel 2007 data file called R4SPSSData contains data from 15 hypothetical households including information on:

- household size (Hhsize)
- rental status (OwnRent)
- internet connectivity (InternetHH)
- time in media survey panel (YrsinPanel)
- total TV time (in hours) from the past week (TotTVLSW)
- Age of panelist (AgeofPanelist)

For this example, we will demonstrate how to import this Excel file into R for further processing and analysis and call it “Rtest”
Example 1, Continued

To Import data from Excel, (and similarly for SPSS, Minitab, Stata or other source) follow the menu sequence:

Data ➔ Import data ➔ from Excel, Access,...

Enter the name of the data set – this will be the name you use to reference the data set in R (case matters in R; e.g. SPSS != spss)

An Open Documents window will appear- in the lower right hand corner be sure to select .XLSX as the extension (default is .XLS)

- Once you have selected the file and clicked the “Open” tab, Select the Excel sheet to import.
Example 1, Continued

If successful, the imported data set name will appear in the tool bar and the View and Edit data set tabs will become active.

Example 1 – DEMO

Import the dataset R4SPSSData into R via R Commander and name the dataset Rtest.
Categorical Variables and R Commander

Variable type attribution is extremely critical for maximizing the potential of R commander!

Specifying nominal categorical variables as factors will allow them to be accessed as grouping/stratification variables, for use in boxplots, to define groups for an ANOVA or to control plotting symbols, for example.

Binary variables coded with numbers will not automatically be interpreted as a factor (e.g. 1=Male; 2=Female) so you will need to specify such variables as factors and use the numbers or specify the values for the levels.

Example 2

In the R4SPSSData.xlsx file, Internet Connectivity (InternetHH) within a sampled household is coded as a binary variable with 1 implying internet connection and 0, no connection. OwnRent is also a binary variable with 1=Own and 2=Rent.

(A) Using R commander, convert the numeric variable InternetHH to a factor called Internet with values “YES” and “NO”.

(B) Using Rcommander create a 0/1 variable called Own that is 1 for Owners and 0 for Renters. (in this case we will use a short cut Own=2-OwnRent).

(C) Create a new variable called TotTWQ to represent the quartile in which the case’s TotTWLSW value falls.
Example 2(A) - Solution

Step 1: In R commander, Follow this menu sequence: Data → Manage Variables in Active Data Set → Convert Numeric Variables to Factors

Select Variable to convert then specify level names or numbers and the new name, if applicable.

- If levels are selected, then a prompt will appear with as many levels as unique values for the variable.
Example 2(A) Solutions Continued

In the next lecture we will learn how to generate basic tables to check answers. But another way we can do a simple check is simply to view the data –

Example 2(B)

(B) Using Rcommander create a 0/1 variable called Own that is 1 for Owners and 0 for Renters.

- We want the value of 2 (renter) to be 0
- We want the value of 1 (owner) to be 1

A simple “equation” to convert a 1/2 variable to a 1/0 variable is:

\[
\text{new-1-0-var} = 2 - \text{old-2-1-var}
\]
Example 2(B) Solutions Continued

Here’s what the data set now looks like after adding the variable “Own”
Example 2(C) Solutions

(C) Create a new variable called TotTWQ to represent the quartile in which the case’s TotTWLSW value falls.

Recall quartiles represent values that evenly split up a distribution (dataset) into four parts – each part containing a quarter (approximately) of the data...

Example 2(C) Solutions, Continued

Choose Equal Count Bins for Quantiles

Variable you want to BIN!

How do you want the new variable to be labelled?

Number of BINS to USE in creating the new version of the variable

Name of New Variable
Example 2 (C), Solutions Continued

To confirm the conversion, you should now see the new variable Internet appear as an eligible variable for computing frequencies, for example (i.e. can apply the table function to the variable Internet)

Following this menu sequence:
Statistics ➔ Summaries ➔ Frequency Distributions
Example 2(C), Solutions Continued

The R Commander Script Window now shows the `.Table` command sequence to generate the frequencies and distribution (i.e. counts & percentages) of sampled households who do and do not have internet (which are depicted in the output window below).

Example 3

Part 1: Using R commander, determine how many missing values (coded in R as NA) there are for each of the variables in the `Rtest` data set.

Part 2: Using R commander, compute the mean and standard deviation for number of hours of TV viewing for the last week for sampled households.

Part 3: Compute the average TV viewing for the last week for households that have internet and those that do not.
Example 3, Part 1 - Solutions

Counts of missing values for every variable in the data set (NAs) can be computed by following the menu:

Statistics ➔ Summaries ➔ Count of Missing Values...

Example 3, Part 2 – Solutions

Part 2: Using R commander, compute the mean and standard deviation for number of hours of TV viewing for the last week for sampled households.

Step 1: Follow the menu path:
Statistics ➔ Summaries ➔ Numerical Summaries
Example 3 – Solutions, Continued

• Step 2: Select the (numeric) variable(s) of interest from the Numerical Summaries Pane; Also select the appropriate options including computation of:
  – the Mean, Standard Deviation and Quartiles.

Notice that the corresponding R command "numSummary" appears with the selected options in the R Commander Script Pane.

Notice also that the Mean and standard deviation appear in the R Commander Output window.
Example 3, Part 3 – Solutions

Follow the same path to obtain the Numerical Summaries Pane and select the numerical variable (TotTVLSW) and Summary Statistics Required. Then click on the “Summarize by Groups Tab” and then select the grouping variable...

Caution: If the grouping variable you need does not appear in the list, check to see that that variable was declared a factor as described in Example 2.

Example 3, Part 2 – Solutions, Continued

Notice that the script window has an additional numSummary call with a “groups” option added.
Example 4 – Contingency Tables!

Do a higher proportion of renting HH have internet compared to owning HHs?

Is there a significant association between Household Ownership Status and Internet Connectivity?

Example 4 - Solutions

These questions can be answered by computing a 2-by-2 contingency table using OwnRent and InternetHH.

We will also compute a Fisher’s Exact test of association (or Chi squared test of association) between these two variables.

In R Commander, follow the menu path:
Statistics → Contingency tables → Two way tables...
Example 4 – Solutions, Continued

- Step 2: Select the row and column variables as well as any percentages or statistical tests that are desired.

- Step 3: Here we want to know an attribute of Owners/Renters, so I have selected this variable as the row and InternetHH as the column;

Consistent with this logic, I have also requested row percentages. Chi squared and Fischer’s Exact tests have also been checked

Example 4, Solutions Continued

```r
# Table: xtabs(~OwnRent+InternetHH, data=Rtest)
# Row Percentages
# Chi-square test of independence
# Components of chi-square statistic
# Print expected frequencies
# Fisher's exact test

> .Table <- xtabs(~OwnRent+InternetHH, data=Rtest)
> .Table

<table>
<thead>
<tr>
<th></th>
<th>Own</th>
<th>Rent</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>YES</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

> rowPercents(.Table) # Row Percentages
> chisq.test(.Table, correct=FALSE)
> fisher.test(.Table)
```

```
> .Table

<table>
<thead>
<tr>
<th></th>
<th>Own</th>
<th>Rent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Own</td>
<td>30</td>
<td>70</td>
</tr>
<tr>
<td>Rent</td>
<td>60</td>
<td>10</td>
</tr>
</tbody>
</table>

X-squared = 1.25, df = 1, p-value = 0.2636
```

Fisher’s Exact Test for Count Data

- Fisher's Exact Test for Count Data
  - data: .Table
  - p-value = 0.3287
  - alternative hypothesis: true odds ratio is not equal to 1
  - 95 percent confidence interval: 0.01692169 4.29385832
  - sample estimates:
    - Odds ratio: 0.3129331
Graphics in Rcommander

R in general is known for its extensive graphical capabilities and language (ggplot) that provide an unprecedented level of customization (and coding).

Rcommander has a moderate number of supported graphical procedures that are built in. These graphs can be customized with edits you can provide to the general syntax that is generated by Rcommander.

- The color palettes can be customized for many of the plots.
- Generally, plotting symbols, lines types and other plot features have predefined defaults built in.

Graphics in Rcommander, Continued

Reference for Colors in R (i.e. color names/numbers and RGB's)
Graphics in Rcommander: Histograms

Rcommander can generate histograms for any continuous variable in your dataset. Some customization is available through the Rcommander window interface and other customization is easily available through revising the generated syntax.

Both frequency and relative frequency histograms are available.

R Script (R Markdown)

```r
with(Rtest, Hist(Hsize, scale="percent", breaks=Sturges, col="darkgray", xlab="Household Size"))
with(Rtest, Hist(Hsize, scale="percent", breaks=Sturges, col=1, xlab="Household Size"))
with(Rtest, Hist(Hsize, scale="percent", breaks=Sturges, col=4, xlab="Household Size"))
```

Histograms, Continued
Bar graphs with Rcommander

Scatterplots in Rcommander

Plots can add Least squares regression lines or a smoothed, loess line. If you choose smooth line, you can adjust the “bandwidth” in the span slider.
Saving Graphs: Method 1 – Inside R Console

In general, Rcommander produces no graphical output inside its own output window. Instead you can find all graphs within the main R console window.

Click on the boarder to make this panel “Active”

Saving Graphs – Method 2: Via Rcommander
Example 5

(a) Using the Rtest dataset plot side-by-side boxplots for the ages of panelists by whether or not they live in internet households. Make the boxplots “purple”.

(b) Using this same dataset, plot a scatterplot of the total television viewed in the last sweeps week by the number of years they panelist has been on the panel.
   - For this scatterplot be sure the to include marginal boxplots
   - Make the plotted points sized at 1.8
   - Add a title to the plot ‘TV viewing by Years in Panel’

Example 3.1(d) Solutions

Follow this menu sequence/steps:
- Graphs → Boxplot...
- Then select “AgeofPanlist” in the Variable Window
- Then click “Plot by Groups” and select InternetHH, then OK.
- In the Rcommander script window, add, `col='purple'` inside the Boxplot syntax. Highlight the entire statement and click submit.
Example 5(a) Solutions Continued

Follow this menu sequence/steps:
1. Graphs → Scatterplot...
2. Then select “YrsinPanel” for x
3. Select “TotWLSW” for y
4. The select options to select marginal boxplots on the left, enter the title ‘TV viewing by Years in Panel’ and change point size to 1.8 (both on the right).

Example 5 (b) Solutions

Follow this menu sequence/steps:
1. Graphs → Scatterplot...
2. Then select “YrsinPanel” for x
3. Select “TotWLSW” for y
4. The select options to select marginal boxplots on the left, enter the title ‘TV viewing by Years in Panel’ and change point size to 1.8 (both on the right).
Example 5(b) solutions Cont.

```
TV viewing by Years in Panel
```

Thank You!

TBuskirk@m-s-g.com
tdbuskirk@gmail.com

314-695-1378

www.m-s-g.com
Appendix

A0 – Installing R

A1 – Take home example for variable conversion and basic frequency tables.

A2 – Take home example for frequency distributions

A3 – Graphical plotting symbols and lines reference

A4 – Other R packages for Descriptive Statistics

A5 – Linear Regression using Rcommander

A0: Installing R

R

http://cran.r-project.org/bin/windows/base/
  ✤ Current version as of June 2014:
  ✤ http://cran.r-project.org/bin/windows/base/

RStudio

http://rstudio.org/download/desktop
  ✤ Current version as of June 2012:
  ✤ http://download1.rstudio.org/RStudio-0.96.231.exe
A0: Download R  http://cran.r-project.org/bin/windows/base/

R-3.2.5 for Windows (32/64 bit)

Download R 3.2.5 for Windows (62 megabytes, 32/64 bit)
Installation and other instructions
New features in this version

If you want to double-check that the package you have downloaded exactly matches the package distributed by R, you can compare the fingerprints of the two. You will need a
version of Rtools for Windows: both graphical and command line versions are available.

Frequently asked questions

- How do I install R when using Windows Vista?
- How do I update packages in my personal version of R?
- Should I run 32-bit or 64-bit R?

---

A0: Install R

![Image of installation warning]

Do you want to run or save this file?
Name: Rtools.35.exe
Type: Application, 46 KB
From: cran.r-project.org

Run
Save
Cancel

While files from the Internet can be useful, this file type can potentially harm your computer. If you do not trust the source, do not run or save the software. What's the risk?
A1– Basic Recoding and Frequencies

_convert the OwnRent variable into a factor with the levels 1 corresponding to OWN and 2 to RENT._

What proportion of sampled households are renters?

Appendix A1– Demo Video

Roll over the Graphical Image at the bottom and then press the play triangle in lower left.
Appendix 2: Working more with Frequency Distributions

Are there any outliers in the Number of TVs per household? Don’t cheat – use an interactive boxplot here.

What is the 75\textsuperscript{th} Percentile (i.e. 3\textsuperscript{rd} quartile) for the number of TVs owned?

Which type of household – owner or renter – has a longer tenure on the panel, on average?

A2– Video Solutions

- Are there any outliers in the Number of TVs per household

Roll over the Graphical Image at the bottom and then press the play triangle in lower left.
A2, Solutions Cont.

• What is the 75th Percentile (i.e. 3rd quartile) for the number of TVs owned?

```r
numSummary(Rtest[, "NumTVs"], statistics=c("mean", "sd", "quantiles"), quantiles=c(.75))
```

Output Window

```r
> library(abind, pos=4)
> library(e1071, pos=4)
> numSummary(Rtest[, "NumTVs"], statistics=c("mean", "sd", "IQR", "quantiles"), quantiles=c(.75))
mean  sd  IQR  n
3.333333 3.754363 3.15
> numSummary(Rtest[, "NumTVs"], statistics=c("mean", "sd", "quantiles"), quantiles=c(.25,.50,.75))
mean  sd  25%  50%  75%  n
3.333333 3.754363 1  2  4  15
> numSummary(Rtest[, "NumTVs"], statistics=c("mean", "sd", "quantiles"), quantiles=c(.75))
mean  sd  n
3.333333 3.754363 15
```

A2, Cont.

• Which type of household – owner or renter – has a longer tenure on the panel, on average?

```r
t.test(YrsInPanel~OwnRent, alternative='two.sided', conf.level=.95, var.equal=FALSE, data=Rtest)
```

Output Window

```r
Welch Two Sample t-test
data:  YrsInPanel by OwnRent
t = -0.4024, df = 5.587, p-value = 0.7023
alternative hypothesis: true difference in means is not equal to
95 percent confidence interval:
-2.397217  1.697217
sample estimates:
mean in group OWN  mean in group RENT
1.775  2.100
```

Roll over the Graphical Image at the bottom and then press the play triangle in lower left.
A2, Cont.

- Which type of household – owner or renter – has a longer tenure on the panel, on average?
A3: Basic plotting symbols in R

Plotting Symbols other Parameters in R plots: http://bit.ly/1Mcxape

Color Chart for R plots: http://bit.ly/1RKe7m6

A4: Some Other R packages for Descriptive Statistics

There are hundreds of R packages that extend the analytic and graphics capabilities of R.

These packages are like “mini” SPSS modules.

Some popular examples include:

- lessR – streamlined and simplified syntax for General Data Analysis
- Boot – bootstrap statistics
- NLME – nonlinear models
- RandomForest – random forest classification models
- Lattice – scatterplot matrix and other graphical enhancements
- Matchit – statistical matching models
- Ggplot 2 – enhanced graphics capabilities/forms
- Survey – design and analysis of survey samples
- PASWR – Basic statistics package that contains functions for nonparametric analysis and Exploratory Data Analysis.
A4: Additional Examples of Packages focused on Descriptive Statistics

- The `Hmisc` package:
  ```r
  > library(Hmisc)
  > describe(mydata)
  ```
  REQUESTS: n, nmiss, unique, mean, 5,10,25,50,75,90,95th percentiles, 5 lowest and 5 highest scores

- The `pastecs` package:
  ```r
  > library(pastecs)
  > stat.desc(mydata)
  ```
  REQUESTS: nbr.val, nbr.null, nbr.na, min max, range, sum, median, mean, SE.mean, CI.mean, var, std.dev, coef.var

- The `psych` package:
  ```r
  > library(psych)
  > describe(mydata)
  > describe.by(mydata, grouping variable)
  ```
  REQUESTS: item name, item number, nvalid, mean, sd, median, mad, min, max, skew, kurtosis, se

(Summary statistics as a whole or by the grouping variable)

A5: Regression Modeling in Rcommander

- Step 1: Select the type of model under the Statistics Menu:
  ```r
  Statistics Menu → Fit Model → Linear Regression
  ```

  R names models as Objects and stores model results in these objects.
A5: Linear Regression in Rcommander

Step 2: Select the form of the model – predictors and outcome variables

For Multiple Linear Regression, select multiple predictors using the “CTRL” key

Step 3: Summarize model and explore model fit diagnostics
A5: Linear Regression in Rcommander

**Summarize Function presents overall model fit and coefficient estimates table**

- Provides choices for model fit and summary plots

---

A5: R, Rcommander Regression Output

**Script Window**

```r
RegModel.2 <- lm(TotTWSW ~ AgeofPanelist + YrsInPanel, data=Rtest)
summary(RegModel.2)
anova(RegModel.2)
```

**Output Window**

*Call*

```
Call:
  lm(formula = TotTWSW ~ AgeofPanelist + YrsInPanel, data = Rtest)
```

*Residuals*

```
  Min      1Q  Median      3Q     Max
```

*Coefficients*

```
Estimate  Std. Error t value Pr(>|t|)
(Intercept) -12.2067  12.2495  -0.997  0.33864
AgeofPanelist 1.2183   0.3878   3.141  0.00851
YrsInPanel   -2.9021   3.9446  -0.728  0.48038
```

*Residual standard error: 12.05 on 12 degrees of freedom*

*Multiple R-squared:  0.5436,  Adjusted R-squared:  0.4675*

*p-value: 0.00094*

*Analysis of Variance Table*

```
  Response: TotTWSW
            Df  Sum Sq Mean Sq F value Pr(>F)
AgeofPanelist 1 2271.58 2271.58 13.7613 0.002983
YrsInPanel    1   87.57   87.57  0.5303 0.480370
Residuals     12 1996.84   166.40
```

---