Chapter 1
Variables, Data and Graphs

Section 1.2
Data Analysis with the Computer

David W. Gerbing

The School of Business
Portland State University

- Data Analysis with the Computer
  - R for Data Analysis
  - Read Data

1.2a
R for Data Analysis

Many Applications Exist for Data Analysis

Consider two applications: Excel and R

- MS Excel, the standard application for business calculations
  - Limited statistical capability even for Windows version
  - No `Linux` version
- R
  - Open source, free, and emerging world standard for statistics
  - Runs identically on Windows, Mac and Linux
- **Key Concept**: Excel and R complement each other, so selectively utilize the accessible strengths of each
  - Excel best for data entry, also useful for data manipulation
  - Statistical analysis capabilities of R are world class
  - Gerbing's enhancements to R called `lessR` simplify using R

Get R to Analyze Data

Always free and open source

- Available at `http://cloud.r-project.org/`
- Choose an operating system: Linux, MacOS X or Windows
- `Windows`: Click the `Download R for Windows` link near the top of the page. Click `base` at the top of the new page, then on another new page click the first link on the page, which begins with `Download R`, followed by the version number.
- `Mac`: Click the `Download R for (Mac) OS X` link near the top of the page. On the resulting new page, click the first file to download, under the heading of `Files:`, which lists the version number followed by `(latest version)`.
- Respond with a `y`, for yes, is asked the following question

```
Would you like to use a personal library instead?
```


Functions in Excel and R

Analyze the data with a specified function

- **Function**: Procedure that performs a specific task such as the calculations of an analysis for the data values of one or more variables
- Ex: Function max identifies a variable's largest data value
- To reference the data values of a variable for analysis, provide
  - Excel: Range of cells that contain the data for the variable
  - R: Name of the variable for a specific column of data values
- Ex: Obtain the maximum salary of the nine employees from the previous data table
  - Excel: Locate the relevant cell on the worksheet to display the result, enter an =, then the function name and the cell range of the data, `=max(E2:E8)`
  - R: After the provided prompt, >, enter the function name and variable name, `> max(Salary)`

R Input and Output

R input

- R provides many hundreds of functions for data analysis
- Enter a call to a function in R after the provided command prompt, >, as shown with an excerpt from the start-up screen

```
R version 3.6.2 (2019-12-12) -- "Dark and Stormy Night"
Copyright (C) 2019 The R Foundation ...
...
[R.app GUI 1.70 (7735) x86_64-apple-darwin15.6.0]

>
```

- If a function call is continued to a new line with an Enter or Return before completion, R uses the continuation prompt, +

Enhance R with the `lessR` Functions

Anyone can extend R by adding more functions

- R organizes its large collection of functions into groups called `packages`, such as the `stats` package for statistical analysis
- Gerbing's `lessR` contributed package provides functions for
  Less is More: Less coding, more results
- Download: *One time only*, when running R, enter
  `> install.packages("lessR")`
- *Each* time when beginning a new R session, to access the `lessR` functions, first enter
  `> library("lessR")`
- Or, store this `library` instruction in a text file called `.Rprofile` that automatically runs when an R session begins
  - Windows: Place in the top level of the `Documents` folder
  - Mac, Linux: Place in the top level of the user's home folder

R Output

Distinguish between text and graphic R output

- All R output by default goes to one of two windows
  - Text output to the *console window*, the same window for which to enter R instructions
  - Graphics output to a *graphics window*
- Output in any window can be saved to a file at any time with the usual File ▷ Save menu sequence
- Or, output can be copied directly from the console as straight text or copied from a graphics window as a graphic, and then pasted into any application that can receive text or graphics

lessR Help Menu

- > `Help()`
- The overview help screen indicates how to obtain further help for specific analyses, such as > `Help(Histogram)`
- To get detailed help, place a ? in front of the function name, such as,

  > `?Histogram`

Help Topics for lessR

Help(data) Create a data file from Excel or similar application.
Help(Read) and **Help(Write)** Read or write data to or from a file.
Help(library) Access libraries of functions called packages.
Help(edit) Edit data and create new variables from existing variables.
Help(system) System level settings, such as a color theme for graphics.

Help(Histogram) Histogram, box plot, dot plot, density curve.
Help(BarChart) Bar chart, pie chart.
Help(LineChart) Line chart, such as a run chart or time series chart.
Help(ScatterPlot) Scatterplot for one or two variables, a function plot.

Help(SummaryStats) Summary statistics for one or two variables.
Help(one.sample) Analysis of a single sample of data.
Help(ttest) Compare two groups by their mean difference.
Help(ANOVA) Compare mean differences for many groups.
Help(power) Power analysis for the t-test.
Help(Correlation) Correlation analysis.
Help(Regression) and **Help(Logit)** Regression analysis, logit analysis.
Help(factor.analysis) Confirmatory and exploratory factor analysis.

Help(prob) Probabilities for normal and t-distributions.
Help(random) and **Help(sample)** Create random numbers or samples.

Help(lessR) lessR manual and list of updates to current version.

Figure: General lessR help menu

Incorporate R Output into an Analysis Report

All aspects of a statistical analysis require interpretation

- **Key Concept**: Write the interpretative report of the analysis with a word processor, so integrate with computer output (*though better with R Markdown from RStudio source window*)
- Text integration
  - Copy text output from R and paste into a word processor
  - To line up the columns and to separate computer output from interpretation, display the pasted R output in a monospaced font, usually Courier New with size 9 or 10
- For graphics, can do the usual Copy from R graphics window and Paste into the word processor, or File → Save to a file
- For a pdf file, insert into MS Word
  - Windows: Insert → Object → From File... and then Create from File
  - Mac: Insert → Photo → Picture From File...

1.2b
Read Data for Analysis

Read Data

Read data into an R data frame with the `lessR` `Read` function

- **Data frame**: Data table that exists within an R session
- Both the R data table (frame) and the Excel worksheet have their own names, distinct from the names of the corresponding variables that define the data table
- Read data from an Excel, csv (tab or comma delimited), R, SPSS or SAS file into an R data table named `d` with the `lessR` `Read` function, abbreviated `rd`

> `d <- Read("")` browse the computer for the data file located on the computer's file system or on a local network

> `d <- Read("http://`*web address*`")` web data file

> `d <- Read("`*path name*`")` directly specify data file

- Usually assign the name `d` to the resulting data frame
- Must capitalize `Read` to differentiate from the family of standard R functions that begin with `read`

Ex: Read a Data File into an R Data Frame

Read data directly from the web

```
> library("lessR")
> d <- Read("http://lessRstats.com/data/example.csv")
```

- This `Read` function call reads data in `csv` format from the specified file into an R data frame called `d`, the default name for lessR data frames in the corresponding data analysis functions
- The `data` are now ready for analysis

Read output

- `Read` also provides
  - the names of the variables in the `d` data table, which are the names referenced by the subsequent analyses
  - the type of variable, such as numeric vs. non-numeric

Types of Variables in the Data Table

Different types of variables incur different types of storage

- The properties of the data values for each variable as stored within R within the data fame should correspond to the conceptual meaning of the variable
- A numeric variable, for which R has several storage types
  - **integer**: values that have no decimal digits
  - **numeric**: values that have decimal digits
- **character variable**: An R storage type for non-numeric variables
- The Read function determines how the data values for each variable are stored
- If a variable in a text data file has any data value with a non-numeric character, such as a $, R by default stores the resulting values as a factor, as the nominal data of a categorical variable, but not a problem reading Excel files

Structure and Contents of the Data Frame

Read output shows the variables in the analysis

```
Data Types
-----------------------------------------------------------------
character: Non-numeric data values
integer: Numeric data values, integers only
numeric: Numeric data values with decimal digits
-----------------------------------------------------------------


Variable                 Missing Unique
    Name      Type Values  Values Values   First and last values
-----------------------------------------------------------------
Name      character      7       0      7   Ritchie, Denise ...
Years       integer      6       1      5   7  NA  15 ... 6  18
Gender    character      7       0      2   M  M  M ... F  F  M
Dept      character      7       0      4   ADMN  SALE  ...  MKTG
Salary       double      7       0      7   3788.26  94494.58 ...
-----------------------------------------------------------------
```


Option to Specify the ID Field

The unique ID of each row of data is not a variable to analyze

- The first column of data in the data file `example.csv` is an ID field, which contains the employee names
- Inform R of the ID field with the `row.names` option

```
> d <- Read(".../example.csv", row.names=1)
```

- R uses this ID information in other analyses, such as labeling each point in a graph or labeling each row of output

```
Number of Variables in d:     4
Variable                 Missing Unique
    Name      Type Values  Values Values   First and last values
-----------------------------------------------------------------
Name      character      7       0      7   Ritchie, Denise ...
Years       integer      6       1      5   7  NA  15 ... 6  18
Gender    character      7       0      2   M  M  M ... F  F  M
Dept      character      7       0      4   ADMN  SALE  ...  MKTG
Salary       double      7       0      7   3788.26  94494.58 ...
```


Index Subtract 2 from each listed value to get the Slide

- The End