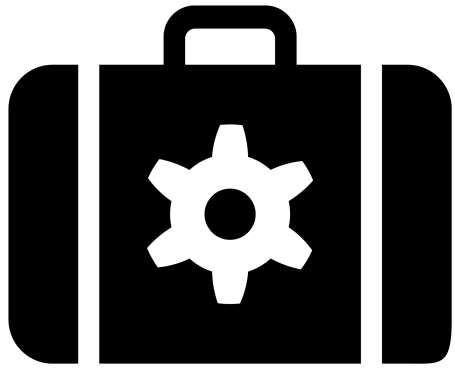

R FOR BUSINESS ADMINISTRATION

R for Business Administration

20160816

Acknowledgment

Special thanks to Dr. Anthony Hayter, who encouraged and supported the original authors of this work to use R in his Professional Master of Business Administration Statistics course (STAT 4610-17) at the University of Denver Daniels College of Business in Denver, Colorado, USA in Spring 2016.

Contents

List of Figures

Introduction

Purpose of this Work

What This Is (And Is Not)

As defined on Wikipedia, "Data is a set of values of qualitative or quantitative variables; restated, pieces of data are individual pieces of information." The article continues, "Data is the least abstract, information the next least, and knowledge the most.

Data becomes information by interpretation; e.g., the height of Mt. Everest is generally considered 'data', a book on Mt. Everest geological characteristics may be considered 'information', and a report containing practical information on the best way to reach Mt. Everest's peak may be considered 'knowledge'...

Some complement the series 'data', 'information' and 'knowledge' with 'wisdom', which would mean the status of a person in possession of a certain 'knowledge' who also knows under which circumstances is good to use it."[1]

This work strives to be a practical guide to using R for business administration. In other words, to teach business professionals and students how to use R to convert data into information.

This work cannot provide the knowledge on the best way to release a new product, market to customers, etc. This work also cannot provide the wisdom to understand the depths of applying the field of statistics to business decision-making. Instead, this work strives to be a starting point to catalyze further inquiry.

Many others have written helpful books, guides, blog posts, videos, articles, and tutorials on statistics, using R, communicating with numbers, and more. This a complementary effort and if using the techniques outlined herein is valuable, you will most likely require additional resources to get the most value for your organization.

The following example from Cincinnati, Ohio, USA will hopefully illustrate this point, which I learned about through an episode of FiveThirtyEight's What's the Point podcast entitled, "Charles Duhigg: Data Can Help Us Change, But We Still Have To Do The Hard Work" hosted by Jody Avirgan.

[1]Source: https://en.wikipedia.org/w/index.php?title=Dataoldid=733102554

Cognitive Disfluency in Cincinnati

This anecdote is included to demonstrate an example when data alone was insufficient in answering an organization's questions.[2] Similarly, data and the techniques outlined in this work are unlikely to answer your organization's questions.

In 2007, South Avondale Elementary School was ranked as one of the worst schools in Cincinnati, Ohio, USA. Situated in a low income neighborhood with high unemployment and crime; students, parents, teachers, and administrators were facing difficult odds and the school had been declared an academic emergency.

In response, officials and corporate partners invested heavily to be on the leading edge of using data and technology to improve performance. Using a variety of tools, data on things like attendance, homework, test scores, and participation were all rolled up into memoranda, dashboards, and reports detailing a variety of metrics.

The experiment was failing, with 90 percent of educators admitting they did not review the data being collected on their students and sent to them. In response, they created the Elementary Initiative, which improved performance from only 37 percent of students meeting education standards to over 80 percent of students.

This program leveraged a concept known as cognitive disfluency by having teachers use a literal data room to physically meet and manually transcribe data, draw graphs, and plan classroom experiments. By disrupting their established pattern of consuming information through dashboards, the teachers developed a deeper understanding of the data and were able to leverage insights in successful ways.

The takeaway is not that you must replicate this process for your organization (although it may be helpful to do so). Instead, recognize that integrating data into decision-making for your organization is a process that will take time, effort, trial, and error.

Given the challenge ahead, consider imitating successful Free Software and Open Source Hardware communities by sharing your questions, answers, and lessons learned with others.

Thank you,

Harris Kenny

[2]This story is summarized from Charles Duhigg's 2016 book Smarter Better Faster: The Secrets of Being Productive in Life and Business.

Installation

Getting Ready

1.1 Your Computing Environment

R is a Free Software[1] project shared under the terms of the Free Software Foundation's GNU General Public License. Similarly, this work recommends you use GNU/Linux for your computing environment.

Not sure where to start? Learn more about some of the leading GNU/Linux distributions:

- Debian: `http://www.debian.org/`
- Fedora: `https://getfedora.org/`
- Ubuntu: `http://www.ubuntu.com/`

1.2 Installing R

R is a software environment that is capable of data manipulation, calculation, and graphical display. The term "environment" is used to communicate that it is a fully planned and coherent system, within which statistical techniques are implemented.

R is an implementation of the S programming language, created by John Chambers while at Bell Labs, combined with lexical scoping semantics inspired by Scheme. R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand and an initial version was released in 1994.[2]

This work is using R 3.1.1 base.

To install R:

1. Go to `https://cran.r-project.org/mirrors.html`.
2. Select a location closest to you.
3. Click on your operating system and follow the appropriate directions.

R is part of many GNU/Linux distributions, you should check with your GNU/Linux package management system in addition to the link above.

[1] Free software is software that gives you the user the freedom to share, study and modify it. We call this free software because the user is free. Learn more at `http://www.fsf.org/about/what-is-free-software`.

[2] Learn more at: `https://www.r-project.org/`.

1.3 Installing Sample Datasets

R has an included package with sample datasets that are used throughout this work to demonstrate concepts and applications for R.[3] This work will teach you various techniques using this data, that you can then apply to your own data.

To load a sample dataset, for example named foobar, enter the following command in R Console:

data(foobar)

You could also enter:

data("foobar")

Note that these are case sensitive.[4]

1.4 Installing R Commander

R has a fairly high learning threshold; it requires learning a programming language and provides relatively little visual feedback. Thankfully, there is a freely licensed tool called R-Commander which helps solve both of these problems by creating a graphical interface for R.

Combined, they make a compelling solution for those who only need basic analysis, while also allowing greater configuration and depth for those who need more sophisticated analysis.[5]

To install R-Commander:

1. Once you have installed R, open it by double-clicking on the icon or opening through a terminal emulator.

2. A window called "R Console" will open.

3. At the prompt (the > symbol), type the following command exactly and then press enter:

   install.packages("Rcmdr", dependencies = TRUE)

[3] See the full list of datasets included in R 3.1.1: `http://www.rdocumentation.org/packages/datasets/versions/3.3.1`.

[4] For more information on how the data command works, visit: `https://stat.ethz.ch/R-manual/R-devel/library/utils/html/data.html`.

[5] Learn more at: `http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/`.

4. R should respond by asking you to select a mirror site, and listing them in a pop-up box.

5. Choose a nearby location.

6. Depending on your connection speed, the installation may take a while. Be patient and wait until you see the prompt again before you do anything.

1.5 Starting R Commander

If R is not already open, open it by clicking on its icon or through a terminal emulator. To open R Commander, at the prompt enter the following command:

```
library(Rcmdr)
```

You should see a large new window pop up, labeled R Commander.[6] You are now ready to analyze your data with R Commander. If you close this window while R is still open, you can start R Commander again by entering the command Commander() in R Console. Entering library(Rcmdr) in this situation will not work unless you close R and open it again.

Congratulations! Now let's get to work.

[6] Find detailed installation instructions online: `http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/installation-notes.html`

Basics

Getting Started

2.1 User Interface

R

R has a commmand-line interface, however multiple projects have developed a graphical user interface (GUI) for R. A GUI is not required to use R and these packages may still require the user to use the command-line interface for specific functions.

R-Commander

The R-Commander GUI is a simple additional window that contains commonly structured menus, buttons, information fields, and dialogue boxes. It also contains script and output text windows. Commands generated through dialogs will automatically post to the output window with their corresponding printed output, and also post to the script window. Lines in the script menu can be edited and resubmitted. Finally, errors, warnings, and notes appear in the messages window at the bottom of the interface.

2.2 Datasets

Importing Data

Using Rcmdr, select Data then drop down to Import Data to import a dataset based (categorized by file format). If you are using R Script, read.table is the most common command to do this. There are many compatible file formats and ways to import data into R, find support for specific common file types below, or online.

Importing CSV Files

Using R Script and read.table, note whether your CSV has a header or not. If it does, mark TRUE. If not, mark FALSE.

Dataset <- read.table("/filename.csv", header = FALSE, sep = ",")

Importing ODS Files

It is possible to import ODS file (an Open Document Format spreadsheet) directly into R after installing a separate package called readODS. Once installed, this is an example of an R Script command to import an ODS file:

read.ods(file = NULL, sheet = NULL, formulaAsFormula = FALSE)

There are several arguments you can use with this command, find a more detailed walkthrough online here: `https://cran.r-project.org/web/packages/readODS/readODS.pdf`

Viewing Datasets

There are multiple ways to view data you have uploaded into R. This is valuable because it is a way to confirm that you properly imported the data. You should always verify that the data imported as you intended before running commands.

View dataset print(Dataset)
View number of records in dataset nrow(Dataset)
View variable names in dataset names(Dataset)
View data types of variables in dataset str(Dataset)
View last records in dataset tail(Dataset)

Active Dataset

Confirm which is your active dataset in R-Commander with the dataset indicator in the top of the main GUI. You can change the dataset by clicking this window to view a dropdown list. It is important to verify you have the right active data before running commands.

Editing Datasets

It is possible to edit data within R using either commands or R-Commander.

Surveys

Understanding Collected Data

3.1 Types of Surveys

Surveys are a valuable way to learn stakeholders' perceptions of your organization's past, present, and future performance. Below are examples of survey groups that may be relevant for your organization:

- Advisors
- Channel Partners
- Customers
- Developers
- Employees
- Journalists
- Investors
- Policymakers
- Suppliers
- Students
- Tradeshow Attendees

Let's try exercises using data that could be collected through surveys.

3.2 Summarizing Data

Averages and Quartiles

Averages and quartiles are a simple way to capture a snapshot of the data you have collected. Recall there are two types of data (continuous and categorical), so R will generate two types of sets of summary statistics.

The exercises in this section will use the "warpbreaks" dataset (The Number of Breaks in Yarn during Weaving) that is included with R. Load this data set into R following the instructions outlined in the Installation section of this work.

First, view the data by entering the following command:

print(warpbreaks)

Or, click View data set in R-Commander in the center of the user interface window.

This yields the following results:

Consider the context. The dataset includes three variables:

1. Breaks - The number of reported breaks for each sample of yarn.
2. Wool - The type of wool for each sample of yarn. A categorial variable, either A or B.
3. Tension - The level of tension for each sample of yard. A categorical variable, either L (low), M (medium), or H (high).

Assume further that yarn breaking during weaving is problematic. In other words, the higher the number in the Breaks column for each row, the worse that sample of yarn has performed.

Second, calculate summary statistics by entering the following command:

summary(warpbreaks)

Or, select in R-Commander by going to Statistics > Summaries > Active Dataset.

This yields the following results:

What is the significance of these results? Let's break them down one-by-one:

- Min. - The minimum data point in the dataset. For column "breaks" this number is 10. In other words, of the surveyed samples of yarn, the lowest number of observed breaks in the dataset is 10. Stated differently, the best performing yarn had 10 breaks.
- 1st Qu. - The first quartile of data in the dataset. For column "breaks" this number is 18.25. In other words, of the surveyed samples of yarn, the bottom quarter of yarn surveyed had 18.25 breaks.
- Median - The median datapoint in the dataset. For column "breaks" this number is 26. In other words, of the surveyed samples of yarn, the middle data point is 26. This is calculated by cross-secting the data to the middle and selecting the single point, or adding the two

middle points and dividing by two if the dataset has an even number. Median calculations are especially valuable for datasets that have outliers that may be skewing mean averages (more on this below).

- Mean - The mean datapoint for the dataset. For column "breaks" this number is 28.15. This is calculated by summing all of the data points and dividing by the total number (or count) of data points. Mean is what most people are referring to when they use the term average, and is the most common form of average used.

- 3rd Qu. - The third quartile of data in the dataset. For column "breaks" this number is 34. In other words, of the surveyed samples of yarn, the top quarter of yarn surveyed had 34 breaks.

- Max.

What else can we conclude from summary statistics of the "breaks" column? The mean is larger than the median, indicating a possible skew in the data towards the samples of yarn with higher numbers of breaks.

There are two other columns of summary statistics calculated by R: wool and tension. This provides a count summary of these categorical variables. Note that the dataset includes an even number of both types of wool (27 of both A and B). The dataset also includes an even number of all three levels of tension (18 Low, Medium, and High).

There is a third less common measure of average called mode[1], the value that has the highest number of occurrences in the dataset. Calculating mode takes several steps, outlined below.

3.3 Proportion Tests

Use proportion tests to compare the proportion or mean from one group to a specified value. These will generate a p-value that can be used to evaluate the statistical significance of the comparison.[2]

The exercises in this section will use the "warpbreaks" dataset (The Number of Breaks in Yarn during Weaving) that is included with R. Load

[1] In R, mode is more importantly an object characteristic in indicating how the object is stored in memory (e.g. as a number, as a character string, as a function).

[2] Learn more at: `https://stat.ethz.ch/R-manual/R-devel/library/stats/html/prop.test.htm`

this data set into R following the instructions outlined in the Installation section of this work.

One Sample Proportion Test

prop.test(<count>, <total>) prop.test(384,5000) prop.test(384, 5000, conf.level=0.95)

Two Sample Proportion Test

prop.test(x = c(388, 545), n = c(443, 564), correct = FALSE)

2-sample test for equality of proportions without continuity correction

data: c(388, 545) out of c(443, 564) X-squared = 29.824, df = 1, p-value = 4.731e-08 alternative hypothesis: two.sided 95 percent confidence interval: -0.12459255 -0.05633856 sample estimates: prop 1 prop 2 0.8758465 0.9663121

3.4 Generating Metadata from Open Responses

It is possible to create metadata from open-ended responses in survey questions that you evaluate through statistical techniques.

For example, say a survey asked customers if there is anything your organization could do to better serve them in the future. Have someone from your organization review the responses and generate a list based on topics submitted by customers.

Next, count each time on appears in the customer responses. This can be done manually or through a formula in spreadsheet software like LibreOffice Calc or Gnumeric.[3] This new set of metadata can now be leveraged in proportion tests or other testing methods.

[3]Learn more about these projects at `https://www.libreoffice.org/` and `http://gnumeric.org/`

Contact

Getting in Touch

This work is currently hosted on GitHub. The source code is available at the following location: `https://github.com/kenhara/r-for-business-administration/`
Interested in getting more involved in the R community? Consider attending the annual useR! conference, the main meeting of the international R user and developer community. Learn more at: `https://www.r-project.org/conferences.html`.
Learn more about the Free Software communities whose efforts are essential to the creation of this work:

- The R Project for Statistical Computing: `https://www.r-project.org/`
- Rcmdr: `http://rcommander.com/`
- LaTeX Memoir: `http://www.latex-project.org/`
- gedit: `https://wiki.gnome.org/Apps/Gedit`
- Free Software Foundation: `http://www.fsf.org/`
- Aleph Objects, Inc.: `https://alephobjects.com/`

Colophon

Created with 100% Free Software

GNU/Linux

LaTeX Memoir

gedit