R Practical: Case Studies of Survey Data

Ian Crandell

April 23rd, 2015

Outline

1. Introduction
2. Survey Concepts
3. Case Study 1: Survey and Graphics
4. Case Study 2: Searching for Patterns with grep()

Introduction: LISAC Collaboration

- One of LISAC's primary mission is to collaborate with researchers from all faculties on campus to enable and accelerate research.
- A popular form of study at OAU is survey sampling. Many of our clients from our inaugural semester brought us data of this kind.

Introduction: Goals

1. To introduce some concepts in the analysis of survey data and warn of some pitfalls
2. To demonstrate many capabilities in R which are helpful for analyses of this kind.

This module will cover a broad swath of functions in R, from text recognition to apply() to plotting and loops.

Survey Concepts: Likert Scale

- A Likert scale is 5 or 7 point scale used to assess the level of agreement a respondent feels about a question.
- Often times, multiple questions will be written to address the same topic, or theme. Such questions can be analyzed on their own or averaged.
- Averaged questions must go in the 'same direction.'

Survey Concepts: Scoring

- When multiple questions can be interpreted as addressing a single issue, the Likert scales can be averaged, a process called scoring.
- This takes the data from an ordinal type to a pseudo-continuous type, allowing us to use regression techniques.
- This requires that the questions thusly grouped have a meaningful interpretation in aggregate. Such aggregates are often called *constructs*.

Case Study

These example data came from a marketing survey distributed by an electronics company. The researchers wanted to determine if a product sensitization class would improve customers' appreciation for some key products.

Data Description

- sex, secondary, class are categorical grouping variables. sex = 1 for male, 0 for female. Secondary = 1 if the person has a secondary education, 0 if not. class = 1 if that person had been exposed to a product promotion class, 0 if not.
- The rest of the data are 3 sets of questions: cam, cell, and tab. The cam set asks subjects how much they like a company's camera product, cell asks about that company's cell phone, tab asks about the company's tablet. These are all on 5 point Likert scales with 1 = strongly disagree and 5 = strongly agree.
- Since each set of questions asks about the same product, it's natural to score them. We will score them into groups called *camera*, *phone*, and *tablet*.

Research Questions

1. Does the sex of the customer have an effect on their opinion of our products?
2. Does the level of education of the customer have an effect on their opinion of our products?
3. Does our sensitization class have an effect on the customer's opinion of our products?

Data Description: Sample questions

cama) Our camera is superior to our competitors' cameras.

camb) I would recommend our camera to my friends.

camc) Our camera is difficult to use.

Notice that camc goes in the opposite direction as cama and camb. A strongly agree for cama and camb indicates a favorable opinion of the camera, but an unfavorable one for camc. We must reorient camc.

Case Study: transform()

transform(data object, commands,...)

- transform() lets you modify data objects directly.
- The first argument is the data object you want to modify. The rest are commands to modify specific columns of the data object.
- This is useful for doing things like standardizing variables or converting numerical types to factors.

Case Study: grep()

grep(pattern, character vector, value = FALSE)

- grep() searches for a provided pattern in each element of a character vector.
- The pattern can be whatever you like and R will find it, even if it's in the middle of a word.
- If value = F, it will return the positions in the vector where the matches were found. If value = T it will return the words themselves.

Case Study: text()

text(coordinates, text, pos)

- *coordinates* determine where the text will be placed. You can either give 2 arguments, one for x and one for y, or you can use the locator function.
- *text* is a character vector of whatever you want printed on the graph
- For *pos*, values of 1, 2, 3 and 4, respectively indicate positions below, to the left of, above and to the right of the specified coordinates.

Case Study: jpeg(), pdf(), etc...

jpeg(filename, width=480, height = 480, units = 'px')

- *jpeg()* is used for exporting pictures. It opens a file which R will write to instead of writing to the *default device*.
- Any graphics will be sent to this file. jpeg() makes jpegs, pdf() makes pdfs, etc.
- When the picture is finished, close this with dev.off()

Case Study: paste()

paste(characters, sep = ' ')

- *paste()* will combine the word you give it into one text string. This is useful for creating many very similar strings.
- If you put a character vector as one of the entries paste() will loop through it.
- sep is a character that will be placed between each given character input. Setting sep = '' will place nothing between the words.

Case Study: mtext()

mtext(text, side, line, at)

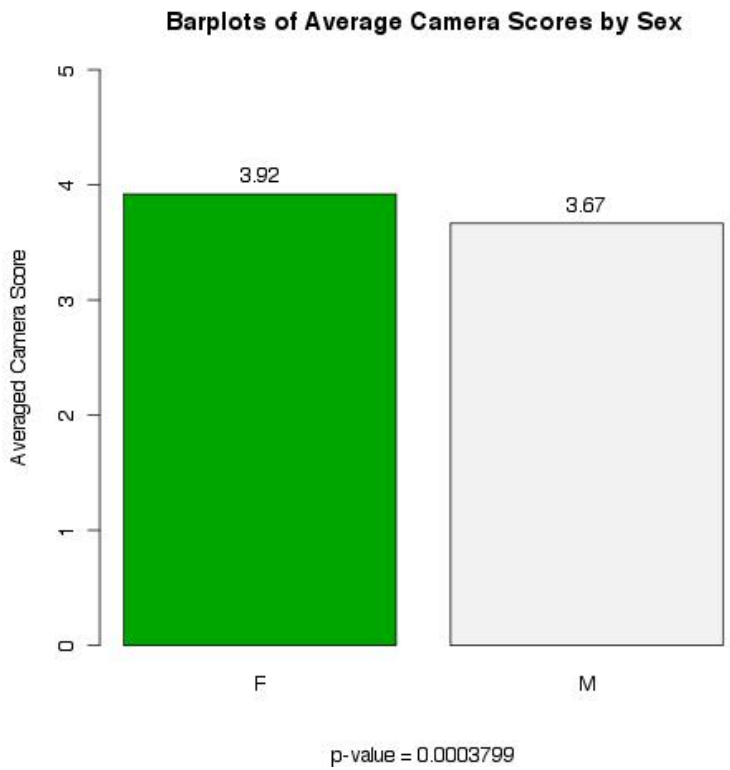
- *mtext()* stands for margin text. *side* is one of 1, 2, 3, 4 for bottom, left, top, right
- *line* is the number of lines out from the main plotting window
- *at* gives the points along the appropriate line where the elements of text will be placed.

Case Study: t.test()

t.test(data, group)

- *t.test()* performs a t test on some vector, *data* broken apart by the values in *group*.
- This implementation will perform the two sample t test.
- Unfortunately, for these kinds of data hypothesis tests aren't very meaningful.

Significant vs Practical Difference



- It's a provable fact that, for any test, p values converge to 0 as the sample size increases.
- For problems with interpretable units this isn't a big problem because we can ask about practical significance.
- This is not easy for Likert scale problems. Care must be taken when interpreting these results.

Case Study Two: Word Search with grep()

- A second data set is from a survey concerning a new health center.
- One of the questions asked respondents to give a reason the new health center might fail.
- The questions was free response, so traditional techniques fail.

Case Study Two: Word Search with grep()

- One method to summarize this kind of data is for the researchers to group the responses into similar categories and give counts.
- For example, how many responses had something to do with the political situation? How many had to do with finance? With lack of trained professionals?
- grep() can be used to find when key words appear in the answers and help the analyst group accordingly.