R: learn by the exercise

Myriam Luce

September 9, 2018

Contents

Chapter 1

Descriptive statistics

1.1 Tables and figures

1.1.1 Frequency table (1D) or contingency table (2D)

If you feel the need to make a table with your data, use a spreadsheet software (Microsoft Excel, LibreOffice Calc, Google Sheets). ;) R is superior in statistics and (arguably) in figures, but spreadsheets definitely have their uses when it comes to tables.

1.1.2 Pie chart

A pie chart is a graph that can be used to visually represent proportions of a *discrete variable*[1]. Note that they have their critics, who recommend never using them for more than two slices, as our brain is bad at comparing the size of slices [1].

As an example data set, let's use eye color in Pensylvania caucasians [3]. An excerpt giving the source data is shown in figure 1.1. Enter the data in your favorite spreadsheet software and save it as a csv (thankfully for you English speakers, there is no need to fiddle with decimal symbol (is it a dot or a comma?) and whether the data is really comma-separated). You should get the following:

```
blue,green,brown
255,170,204
```

R offers various data import options. The most useful I have found were `read.csv`[2] to import csv data and `read.fwf` to import fixed-width data. To demonstrate, figure 1.2 shows what csv (delimited) and fixed-width data look like side by side.

[1] Words in italics are defined in the glossary.
[2] Words in monospace font refer to R commands. The cheat sheet at the end of the tutorial lists most of those used in this document.

Eye color was determined upon clinical examination by a single research nurse using the following categories: blue; gray; green; hazel; light brown; dark brown; and black. Participants also completed a standardized questionnaire that asked self-assessed eye color. The correlation between self-assessed and clinician-assessed eye color was 93%, and there was 100% concordance between self- and clinician-assessed eye color once categorizations were made for analyses. Therefore, analyses were undertaken using categorized exam-determined eye color. Categories of eye color used in analysis were blue/gray (*n* = 255, 40.5%), green/hazel (*n* = 170, 27.0%), and brown/black (*n* = 204, 32.4%). Contingency table analysis using the χ^2 test was undertaken to evaluate the relationship of *P* gene variants and eye color. Unconditional logistic regression analysis was used to estimate the OR effect of *P* gene variants on eye color adjusted for diagnostic outcomes of DN and/or MM. These adjustments were undertaken to remove potential confounding relationships between disease status, *P* gene variants, and eye color.

Figure 1.1: Excerpt from [3].

```
Infant mortality rate,1950,1951,1952,1953,1954,1955,1956,1957,1958,1959,1960,1961,1962
Abkhazia,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
Afghanistan,,,,,,,,,,"245,00","240,50","236,30","232,30","228,50","224,60","220,70","2
Akrotiri and Dhekelia,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
Albania,,,,,,,,,,"120,10","115,40","110,80","106,50","101,80","97,90","94,10","90,20","
Algeria,,,,"150,40","149,90","149,50","149,00","148,80","148,40","148,20","148,10","1
American Samoa,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
Andorra,,,,,,,,,,-,,,,,,,,,,-,,,,,,,,,,-,,,,,,,,,7,"7,5",7,"6,5","6,1","5,6","5,2",5,"
Angola,,,,,,,,,,208,,,,,,,,,,180,,,,,,,,,,"138,30","137,50","136,80","136,00","135,30"
Anguilla,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
Antigua and Barbuda,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,24,"22,6","21,5","20,3","1
Argentina,68,67,65,63,60,62,57,68,61,59,"59,87","59,73","59,59","59,39","59,25","59,11"
Armenia,,,,,,,,,,,,,,,,,,,,,,,,,,,,"70,8","68,3","65,9","63,5","61,30","59,20","57,10","
Aruba,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
Australia,"25,00","24,60","24,10","23,60","23,00","22,60","22,10","21,60","21,20","20,
Austria,66,61,52,50,"48,10","46,60","45,20","43,60","41,70","39,60","37,30","35,00","3
Azerbaijan,,,,,,,,,,,-,,,,,,,,,,,,,,,,,,,,,95,,"84,9","83,1","81,5","79,9","78,5","77,3"
Bahamas,,,,,,,,,,,51,,,,,,,,25,"25,5",26,"26,4","26,8","27,2","27,2","27,2","26,9","26,
```

```
                 Nino1+2      Nino3        Nino34        Nino4
Week             SST SSTA     SST SSTA     SST SSTA     SST SSTA
03JAN1990        23.4-0.4     25.1-0.3     26.6 0.0     28.6 0.3
10JAN1990        23.4-0.8     25.2-0.3     26.6 0.1     28.6 0.3
17JAN1990        24.2-0.3     25.3-0.3     26.5-0.1     28.6 0.3
24JAN1990        24.4-0.5     25.5-0.4     26.5-0.1     28.4 0.2
31JAN1990        25.1-0.2     25.8-0.2     26.7 0.1     28.4 0.2
07FEB1990        25.8 0.2     26.1-0.1     26.8 0.1     28.4 0.3
14FEB1990        25.9-0.1     26.4 0.0     26.9 0.2     28.5 0.4
21FEB1990        26.1-0.1     26.7 0.2     27.1 0.3     28.9 0.8
28FEB1990        26.1-0.2     26.7-0.1     27.2 0.3     29.0 0.8
07MAR1990        26.7 0.3     26.7-0.2     27.3 0.2     28.9 0.7
14MAR1990        26.1-0.4     26.9-0.2     27.3 0.1     28.6 0.4
21MAR1990        26.1-0.2     27.2 0.0     27.6 0.3     28.7 0.5
28MAR1990        25.7-0.4     27.5 0.2     27.8 0.3     28.8 0.5
04APR1990        25.6-0.3     27.6 0.3     27.9 0.4     28.8 0.4
11APR1990        25.1-0.6     27.6 0.2     27.9 0.2     28.8 0.3
18APR1990        25.3 0.0     27.7 0.2     28.0 0.2     28.9 0.4
25APR1990        25.1 0.0     27.7 0.4     28.2 0.4     29.2 0.6
02MAY1990        24.6-0.2     27.6 0.3     28.1 0.3     29.0 0.4
09MAY1990        24.2-0.2     27.5 0.3     28.1 0.3     28.9 0.2
16MAY1990        24.3 0.1     27.4 0.3     28.0 0.2     28.8 0.1
23MAY1990        23.7-0.2     27.2 0.2     28.1 0.3     29.0 0.2
30MAY1990        23.4-0.1     27.1 0.3     27.9 0.2     28.9 0.1
```

Figure 1.2: Delimited data (left) and fixed-width data (right).

Go ahead and load your small csv into R with `read.csv('C:/......data.csv', header=TRUE)`. To avoid messing with default working folder in R settings, I recommend always using the full absolute file path (i.e. starting with C:). Note that you should use the *forward slash* "/" as a path separator, even on Windows. The second parameter, `header=TRUE`, tells R that the first line in your file corresponds to the column headers, not actual data. You can then use the function `pie(counts, labels)` to produce a pie chart. However, as shown below, a naive approach might displease.

```
> color = read.csv('C:/.../r-tutorial/eyecolor.csv', header=TRUE)
> color
  blue green brown
1  255   170   204
> pie(color, colnames(color))
Error in pie(color, colnames(color)) : 'x' values must be positive.
```

You might be scratching your head and wondering which part of 255, 270 or 204 is not positive, and you'd be justified to do so. Here, one must dive into computer programming concerns to understand what is going on. The "not positive" message hints at a problem with the format of the input data. Let's demonstrate:

```
> values = c(255, 170, 204)
> labels = colnames(color)
> pie(values, labels)        # works! produces figure 1.3
> typeof(color)
[1] "list"
> typeof(values)
[1] "double"
> typeof(as.integer(color))
[1] "integer"
> pie(as.integer(color), colnames(color))        # works too!
```

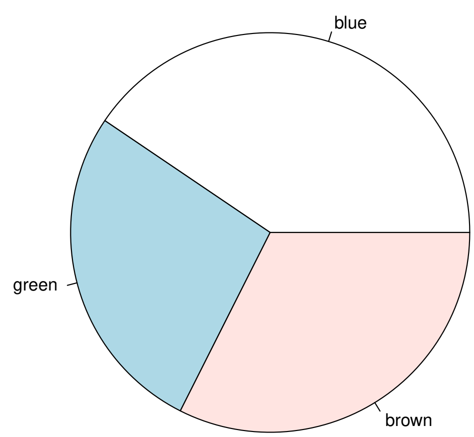

Figure 1.3: Eye color among Pennsylvania caucasians.

Technically, `read.csv` returns a `data.frame`, while `pie` only accepts numbers. You can convert your data frame contents to anything reasonable (R will turn "2" into an integer, but not "abc") using the host of `as.xyz` functions. Let's take a painful tangent into R types that will hopefully help you later on.

R types

Logical:	TRUE or FALSE
Numeric:	real, by the math definition (ex. 12.3). Double is a numeric with better precision.
Integer:	integer, by the math definition (ex. 12).
Character:	text of any length
Factor:	a type that represents a discrete variable
Ordered:	a type that represents an ordinal variable
List:	a 1D collection of "things" (may be strings, numbers, or a mix of them)
Vector:	a 1D collection of things of *one type*
Matrix:	a 2D collection of things of *one type*
Array:	a nD collection of things of *one type*
Data Frame:	a (mostly) 2D collection of things, where each column can be of a different type

For future reference, Quick R gives an excellent introduction on the subject [2].

1.1.3 Bar graph

1.1.4 Histogram

1.1.5 Line graph

1.1.6 Scatter graph

1.1.7 Box and whiskers graph

1.2 Numbers

1.2.1 Center

Mean

Median

Mode

1.2.2 Dispersion

Range

Variance

Standard deviation

Coefficient of variation

Quartiles and percentiles

1.2.3 Shape

Skewness

Kurtosis

L-moments

Chapter 2

Probabilities

2.1 Factorial

2.2 Combinations

2.3 Permutations

2.4 Probability Mass/Density Function

Chapter 3

Statistics

3.1 Binomial distribution

3.2 Multinomial distribution

3.3 Poisson distribution

3.4 Inverse binomial distribution

3.5 Hypergeometric distribution

3.6 Normal distribution

3.7 Exponential distribution

3.8 Gamma distribution

3.9 c2 distribution

3.10 Fisher-Snedecor distribution

3.11 Student's law

Chapter 4

Inferential statistics

4.1 Student's test

4.2 Student's paired test

4.3 Bartlett's test

4.4 Single-factor ANOVA

4.5 c2 test

4.6 Wilcoxon-Mann-Whitney test

4.7 Kolmogorov-Smirnov test

4.8 Kruskal-Wallis test

4.9 Pearson's test

4.10 Spearman's test

4.11 Kendall's test

4.12 Simple linear regression

4.13 Multiple linear regression

Chapter 5

Cheat sheet

5.1 Plumbing

?	?exact_function_name
??	??keyword
typeof	typeof(R_variable)
class	class(R_variable)
str	str(R_variable)
colnames	colnames(R_variable)
as.integer	as.integer(R_variable)

5.2 Data import and export

read.csv	read.csv(’delimited_data.csv’, header=TRUE, sep=”,”, dec=”.”)
read.fwf	read.fwf(’fixed_width_data.txt’, widths=c(10, 5, 4), header=TRUE, skip=2)
write.csv	write.csv(R_variable, file=’desired_file_name.csv’, append=FALSE)

Bibliography

[1] *Pie chart*. en. Page Version ID: 856409948. Aug. 2018. URL: https://en.wikipedia.org/w/index.php?title=Pie_chart&oldid=856409948 (visited on 09/09/2018).

[2] *Quick-R: Data Types*. URL: https://www.statmethods.net/input/datatypes.html (visited on 09/10/2018).

[3] Timothy R. Rebbeck et al. "P Gene as an Inherited Biomarker of Human Eye Color". en. In: *Cancer Epidemiology and Prevention Biomarkers* 11.8 (Aug. 2002), pp. 782–784. ISSN: 1055-9965, 1538-7755. URL: http://cebp.aacrjournals.org/content/11/8/782 (visited on 09/09/2018).

Glossary

discrete variable A variable that refers to categorical data (ex. color of eyes), as opposed to continuous data (ex. height in mm). 3, 6

ordinal variable A variable that refers to categorical data, but where the categories can be ordered (ex. small, medium, large). 6