Using R to analyze data

William Revelle
Northwestern University

Abstract

This is a very brief guide to help students in a research methods course make use of the R statistical language to analyze some of the data they have collected. For more extensive tutorials of R in psychology, see my short and somewhat longer tutorials.

9 simple steps

1. Install R on your computer or go to a machine that has it
2. Enter your data using a text editor and save as a text file (perhaps comma delimited if using Excel). Note that the computer simulation we did has already stored it in a readable format.
3. Start R, read (or paste) the data file
4. Find basic descriptive statistics (e.g., means, standard deviations, minimum and maxima)
5. Prepare a simple descriptive graph (e.g, a box plot) of your variables
6. Find the correlation matrix to give an overview of relationships
7. If you have categorical experimental variables, you can do an Analysis of Variance (ANOVA)
8. If you have a continuous independent variable, do the appropriate multiple regression using standardized scores.
9. Graph the results (advanced)

Preliminaries: Installing R

However, to do analyses on your own machine, you will need to download the R program from the R project and install it yourself. Go to the R home page and then

contact: William Revelle revelle@northwestern.edu
Draft version of May 3, 2021

choose the Download from CRAN (Comprehensive R Archive Network) option. This will take you to list of mirror sites around the world. The most convenient mirror is simply the cloud. You may download the Windows, Linux, or Mac versions at this site. Installation differs across platforms but, at least for a Mac, is very easy.

Many people find RStudio a convenient interface.

Run R with a text editor to save your commands for later use. Or, use the edit window in R to write out your commands as you work. Save this file for later use.

R is case sensitive and does not give overly useful diagnostic messages. If you get an error message, don't be flustered but rather be patient and try the command again using the correct spelling for the command.

Help and guidance

All packages and core-R have help pages. When in doubt, use them. Either by entering **help()** or by ?what you need help for.

Install psych

The *psych* package has been developed specifically to help students and researchers in psychology. Think of it as a Swiss Army knife. Not the optimal tool for anything, but useful for most things.

To install psych, you can simply enter two commands in R

R code

```
install.packages("psych", dependencies=TRUE)
install.packages)("psychTools")
```

You can also go to the install packages menu and do the same thing.

Entering the data

For most data analysis, rather than manually enter the data into R, it is probably more convenient to use a spreadsheet (e.g., Excel or OpenOffice) as a data editor, save as a tab or comma delimited file, and then read the data from that file or read from the clipboard using the `read.clipboard()` a command. (found in the psych package).

For analyzing the data from the computer simulation, the data are already in a readable form, with labels for the variables in the first row. If you have run multiple experiments, there are some comment lines that need to be removed before reading into R. Most of the examples in this tutorial assume that the data have been entered this way. Many of the examples in the help menus have small data sets which are included in the *psychTools* package.

For the example, we read data from the local computer for 100 cases of simulated data with 5 independent variables and 3 dependent variables.

To read a file from your local machine, change the datafilename to specify the path to the data. A simple trick on the Mac is to store the datafile on your desktop. Thus the path is "Desktop/myfilename".

If you have opened the file using a text editor or Excel, you can just copy the appropriate rows into the clipboard and then use the "read.clipboard()" command.

Here we read the data in from an external file.

First, make sure that you have made the *psych* active by using the (library) command.

After reading the data, to make sure that you have read the data in correctly, ask for the dimensions of the data, and he first and last few lines,

R code

```
library(psych)  #this makes it active.  Do this once per session.
my.data.file <- "https://personality-project.org/courses/205/sim.data.txt"
my.data <- read.file(my.data.file)
#or if you have the data in the clipboard
my.data <- read.clipboard()

dim(my.data)
headTail(my.data)
```

```
dim(my.data)
[1] 100  10
>    headTail(my.data)
    snum sex drug time anxiety impulsivity arousal tension performance cost
1      1   2    0   14       6           8      30      21          55    1
2      2   1    1   15       6           3      99      48          77    1
3      3   1    1   19       5           5      97      39          79    1
4      4   1    1   21       6           1      57      41          66    1
...  ... ...  ...  ...     ...         ...     ...     ...         ...  ...
97    97   1    1   16       3           7      87      25          68    1
98    98   1    1    9       5           4      56      43          63    1
99    99   2    0   21       4           9      66      14          66    1
100  100   2    1   19       5           7      99      33          51    1
```

Descriptive statistics

It is always important to find the basic descriptive statistics of your data. the `desribe` function will do this. You can also use `describe` to break the data down by groups by specifying the grouping variables.

R code

```
describe(my.data)  #basic descriptive statistics
```

```
>    describe(my.data)  #basic descriptive statistics
            vars   n  mean    sd median trimmed   mad min max range  skew kurtosis   se
snum           1 100 50.50 29.01   50.5   50.50 37.06   1 100    99  0.00    -1.24 2.90
sex            2 100  1.45  0.50    1.0    1.44  0.00   1   2     1  0.20    -1.98 0.05
drug           3 100  0.56  0.50    1.0    0.58  0.00   0   1     1 -0.24    -1.96 0.05
time           4 100 15.15  4.19   15.0   15.22  5.93   8  22    14 -0.07    -1.10 0.42
anxiety        5 100  3.96  1.89    4.0    3.99  1.48   0   9     9 -0.08    -0.49 0.19
impulsivity    6 100  5.15  2.08    5.0    5.20  1.48   0  11    11 -0.09    -0.13 0.21
arousal        7 100 64.56 27.59   64.0   66.26 36.32   1  99    98 -0.33    -1.11 2.76
tension        8 100 24.53 13.23   22.0   23.90 14.08   3  61    58  0.44    -0.44 1.32
performance    9 100 64.97  9.09   66.0   65.25  8.90  41  83    42 -0.27    -0.59 0.91
cost          10 100  1.00  0.00    1.0    1.00  0.00   1   1     0   NaN      NaN 0.00
```

Describe the data by levels of caffeine and by gender:

R code

```
describe(my.data~drug + sex)
```

```
 Descriptive statistics by group
drug: 0
sex: 1
            vars  n  mean    sd median trimmed   mad min max range  skew kurtosis   se
snum           1 24 44.21 24.98   40.0   42.45 25.95  11  96    85  0.60    -0.69 5.10
sex            2 24  1.00  0.00    1.0    1.00  0.00   1   1     0   NaN      NaN 0.00
drug           3 24  0.00  0.00    0.0    0.00  0.00   0   0     0   NaN      NaN 0.00
time           4 24 14.92  5.12   14.0   14.90  8.15   8  22    14 -0.02    -1.62 1.05
anxiety        5 24  4.17  2.39    5.0    4.15  2.97   0   9     9 -0.06    -1.05 0.49
impulsivity    6 24  5.38  1.91    6.0    5.45  2.22   2   8     6 -0.23    -1.18 0.39
arousal        7 24 42.38 18.70   43.0   43.45 22.24   1  66    65 -0.35    -1.11 3.82
tension        8 24 16.21 10.35   15.5   15.15 10.38   3  47    44  1.03     1.02 2.11
performance    9 24 60.54  8.41   60.5   60.75  8.15  41  77    36 -0.31    -0.24 1.72
cost          10 24  1.00  0.00    1.0    1.00  0.00   1   1     0   NaN      NaN 0.00
------------------------------------------------------------------------------------------
drug: 1
sex: 1
            vars  n  mean    sd median trimmed   mad min max range  skew kurtosis   se
snum           1 31 48.71 30.74     46   48.68 43.00   2  98    96 -0.02    -1.41 5.52
sex            2 31  1.00  0.00      1    1.00  0.00   1   1     0   NaN      NaN 0.00
drug           3 31  1.00  0.00      1    1.00  0.00   1   1     0   NaN      NaN 0.00
time           4 31 15.65  4.03     16   15.80  4.45   8  22    14 -0.21    -1.12 0.72
anxiety        5 31  4.10  1.68      4    4.08  1.48   0   7     7 -0.19    -0.45 0.30
impulsivity    6 31  5.19  1.76      5    5.20  1.48   1   8     7 -0.07    -0.62 0.32
arousal        7 31 82.10 19.53     87   85.56 14.83  33  99    66 -1.27     0.44 3.51
tension        8 31 32.71 12.14     35   32.36 11.86   8  61    53  0.20    -0.38 2.18
performance    9 31 70.13  7.41     69   70.56  8.90  49  83    34 -0.57     0.34 1.33
cost          10 31  1.00  0.00      1    1.00  0.00   1   1     0   NaN      NaN 0.00
------------------------------------------------------------------------------------------
drug: 0
sex: 2
            vars  n  mean    sd median trimmed   mad min max range  skew kurtosis   se
snum           1 20 51.45 33.17   57.0   51.88 45.22   1  99    98 -0.17    -1.56 7.42
sex            2 20  2.00  0.00    2.0    2.00  0.00   2   2     0   NaN      NaN 0.00
drug           3 20  0.00  0.00    0.0    0.00  0.00   0   0     0   NaN      NaN 0.00
time           4 20 14.50  3.91   14.0   14.38  4.45   8  22    14  0.32    -0.95 0.88
```

```
anxiety        5 20  3.55  1.88    3.0     3.56  1.48   0   7     7  0.18    -0.94 0.42
impulsivity    6 20  5.35  2.64    4.5     5.25  2.22   0  11    11  0.31    -0.52 0.59
arousal        7 20 41.00 17.95   43.0    42.25 19.27   4  66    62 -0.41    -0.82 4.01
tension        8 20 13.85  6.36   13.0    13.69  5.93   3  28    25  0.31    -0.68 1.42
performance    9 20 58.70  7.60   57.5    58.31  9.64  49  72    23  0.27    -1.46 1.70
cost          10 20  1.00  0.00    1.0     1.00  0.00   1   1     0   NaN      NaN 0.00
---------------------------------------------------------------------------------------------
drug: 1
sex: 2
            vars  n  mean    sd median trimmed   mad min max range  skew kurtosis   se
snum           1 25 58.00 26.81     61   58.76 31.13   8 100    92 -0.23    -1.15 5.36
sex            2 25  2.00  0.00      2    2.00  0.00   2   2     0   NaN      NaN 0.00
drug           3 25  1.00  0.00      1    1.00  0.00   1   1     0   NaN      NaN 0.00
time           4 25 15.28  3.75     15   15.38  4.45   8  22    14 -0.20    -0.77 0.75
anxiety        5 25  3.92  1.66      4    4.00  1.48   0   7     7 -0.46    -0.39 0.33
impulsivity    6 25  4.72  2.15      5    4.86  1.48   0   8     8 -0.59    -0.84 0.43
arousal        7 25 82.96 19.20     88   85.57 13.34  37  99    62 -1.29     0.19 3.84
tension        8 25 30.92 10.50     33   30.76  7.41  13  58    45  0.13    -0.02 2.10
performance    9 25 67.84  7.97     68   68.33  7.41  51  80    29 -0.42    -0.59 1.59
cost          10 25  1.00  0.00      1    1.00  0.00   1   1     0   NaN      NaN 0.00
```

Correlations and scatter plots

To find the correlation between a set of variables, use the `corr.test` or `lowerCor` functions. These are merely adaptations of the basic `cor` function with a few additions to make them more convenient. Note how `corr.test` supplies statistical tests for the correlations and how `lowerCor` makes more pleasing output.

To show the resulting correlations as scatter plots, use the `pairs.panels` function.

R code

```
corr.test(my.data[2:9] )# or

lowerCor(my.data[2:9] ) #for prettier output

pairs.panels(my.data[2:9] ) #drop the first and last variables
pairs.panels(my.data[  ]
```

```
corr.test(my.data[2:9] )# or
Call:corr.test(x = my.data[2:9])
Correlation matrix
              sex  drug  time anxiety impulsivity arousal tension performance
sex          1.00 -0.01 -0.05   -0.10       -0.07   -0.01   -0.08       -0.12
drug        -0.01  1.00  0.09    0.03       -0.09    0.74    0.63        0.52
time        -0.05  0.09  1.00    0.00        0.05    0.51    0.08        0.34
anxiety     -0.10  0.03  0.00    1.00        0.04    0.01    0.71        0.02
impulsivity -0.07 -0.09  0.05    0.04        1.00   -0.18   -0.01       -0.05
arousal     -0.01  0.74  0.51    0.01       -0.18    1.00    0.45        0.64
tension     -0.08  0.63  0.08    0.71       -0.01    0.45    1.00        0.34
performance -0.12  0.52  0.34    0.02       -0.05    0.64    0.34        1.00
Sample Size
[1] 100
```

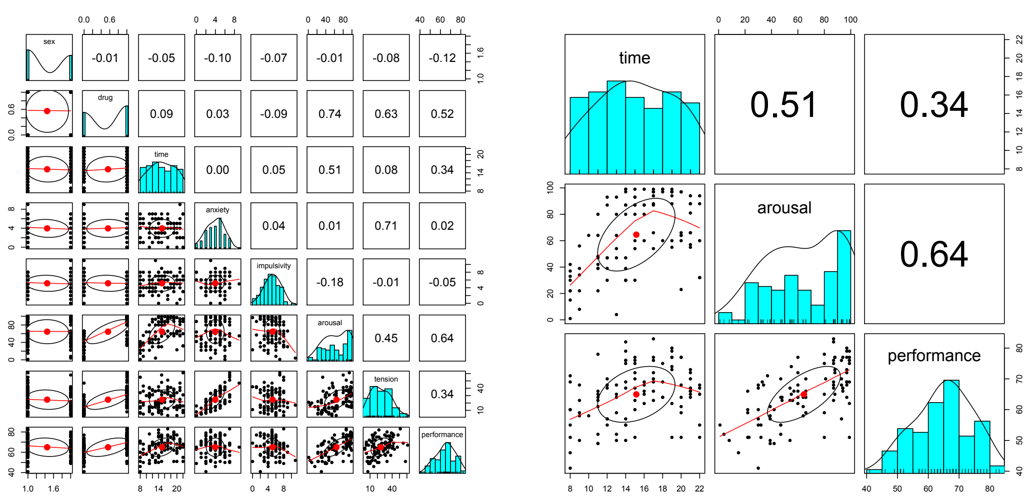
```
Probability values (Entries above the diagonal are adjusted for multiple tests.)
             sex drug time anxiety impulsivity arousal tension performance
sex          0.00 1.00 1.00    1.00        1.00       1       1        1.00
drug         0.94 0.00 1.00    1.00        1.00       0       0        0.00
time         0.64 0.37 0.00    1.00        1.00       0       1        0.01
anxiety      0.33 0.73 0.98    0.00        1.00       1       0        1.00
impulsivity  0.52 0.36 0.60    0.67        0.00       1       1        1.00
arousal      0.94 0.00 0.00    0.88        0.08       0       0        0.00
tension      0.42 0.00 0.41    0.00        0.95       0       0        0.01
performance  0.24 0.00 0.00    0.85        0.59       0       0        0.00
 To see confidence intervals of the correlations, print with the short=FALSE option

>
> lowerCor(my.data[2:9] ) #for prettier output
            sex   drug  time  anxty impls arosl tensn prfrm
sex          1.00
drug        -0.01  1.00
time        -0.05  0.09  1.00
anxiety     -0.10  0.03  0.00  1.00
impulsivity -0.07 -0.09  0.05  0.04  1.00
arousal     -0.01  0.74  0.51  0.01 -0.18  1.00
tension     -0.08  0.63  0.08  0.71 -0.01  0.45  1.00
performance -0.12  0.52  0.34  0.02 -0.05  0.64  0.34  1.00
```

Graphical displays of the data

A scatter plot matrix (SPLOM) shows the correlations as well as the pairwise scatter plots. We can specify which variables to plot.

R code

```
pairs.panels(my.data[2:9] ) #drop the first and last variables
pairs.panels(my.data[cs(time,arousal,performance)])  #just show certain variables
```



The plotting is done in the graphic window and can be saved as a pdf.

Inferential statistics

Are the results that we observe unlikely due to chance. Can we summarize the results in terms of slopes of regressions between the variables? All of this is done using the `aov` function for Analysis of Variance, or the `lm` function (linear model for regression) or the `setCor` function. The default for `setCor` is to standardize the variables, so we need to set this to be FALSE.

R code

```
model.aov <- aov(arousal ~ drug, data=my.data)  #The ANOVA approach
summary(model.aov)  #show the output
print(model.tables(model.aov,"means",digits=3) )#get the means

summary(lm(arousal ~ drug, data=my.data))  # or
setCor(arousal ~ drug , data=my.data, std= FALSE) #don't standardize
```

These three approaches produce the same results (they just look a little different)

```
The aov output and a table of the means

> model.aov <- aov(arousal ~ drug, data=my.data)  #The ANOVA approach
> summary(model.aov)  #show the output
            Df Sum Sq Mean Sq F value Pr(>F)
drug         1  40880   40880   116.2 <2e-16 ***
Residuals   98  34488     352
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


> print(model.tables(model.aov,"means",digits=3) )#get the means
Tables of means
Grand mean

64.56

 drug
drug
    0     1
41.75 82.48
Warning message:
In replications(paste("~", xx), data = mf) : non-factors ignored: drug




The linear model (lm) output
> summary(lm(arousal ~ drug, data=my.data))

Call:
lm(formula = arousal ~ drug, data = my.data)

Residuals:
```

```
   Min     1Q Median     3Q    Max 
-49.48 -11.75   5.25  14.52  24.25 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   41.750      2.828   14.76   <2e-16 ***
drug          40.732      3.779   10.78   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.76 on 98 degrees of freedom
Multiple R-squared:  0.5424,	Adjusted R-squared:  0.5377 
F-statistic: 116.2 on 1 and 98 DF,  p-value: < 2.2e-16


The setCor output

> setCor(arousal ~ drug , data=my.data, std= FALSE)
Call: setCor(y = arousal ~ drug, data = my.data, std = FALSE)

Multiple Regression from raw data 

 DV =  arousal 
            slope   se     t       p lower.ci upper.ci  VIF
(Intercept) 41.75 2.83 14.76 1.2e-26    36.14    47.36 2.27
drug        40.73 3.78 10.78 2.5e-18    33.23    48.23 2.27

Residual Standard Error =  18.76  with  98  degrees of freedom

 Multiple Regression
           R   R2  Ruw R2uw Shrunken R2 SE of R2 overall F df1 df2       p
arousal 0.74 0.54 0.52 0.27        0.54     0.07    116.16   1  98 2.5e-18
```

Graphing results

For categorical variables (drug, sex) we could use an error bars plot. For continuous variables, we draw scatter plots with regression lines.

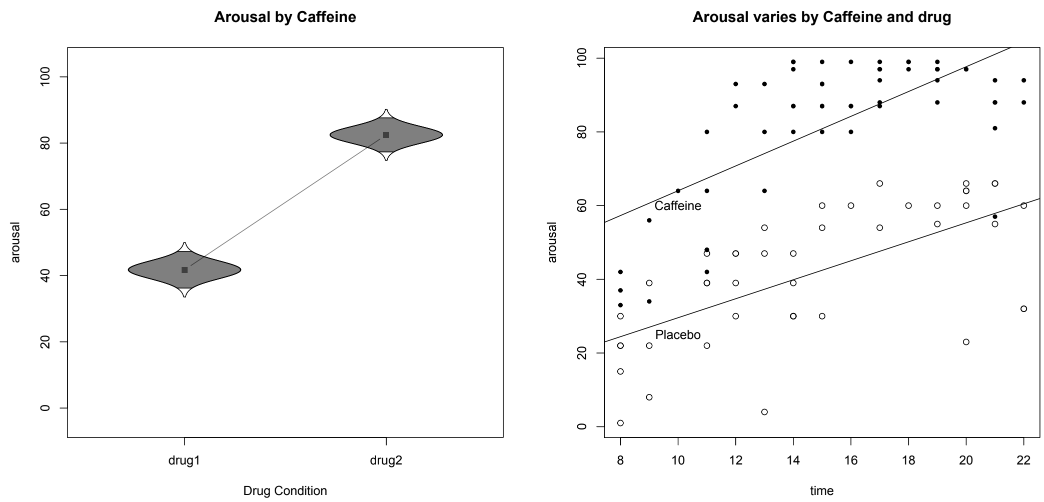
R code

```
error.bars(arousal ~ drug,data=my.data)
#improve this graphic by setting various parameters
error.bars(arousal ~ drug,data=my.data, ylab="arousal", xlab="Drug Condition", main="Arousal by Caffeine")
```

A more complicated graph is to plot a continuous variable (say time of day) and plot arousal as function of the continuous variable as well as the drug conditions.

R code

```
plot(arousal ~ time,data=my.data, pch=c(1,16)[my.data$drug + 1], ylab="Arousal", xlab="Time of Day",main="
by(my.data,my.data$drug, function(x) abline(lm(arousal ~ time,data=x)))  #add the lines
text(10,24,"Placebo")   #Add labels to the lines
text(10,60,"Caffeine")
```



Interactions

Does the effect of any of our independent variables depend upon another independent variable? This is a question of interactions. So, for instance, do the drug effects interact with impulsivity?

This is done in `lm` by examining the centered variables. `setCor` automatically centers variables when examining interactions.

R code

```
setCor(arousal ~ time * impulsivity,data= my.data)
```

```
setCor(arousal ~ time * impulsivity,data= my.data)
Call: setCor(y = arousal ~ time * impulsivity, data = my.data)

Multiple Regression from raw data

 DV =  arousal
                  slope   se     t       p lower.ci upper.ci  VIF
(Intercept)        0.00 0.08  0.00 1.0e+00    -0.16     0.16 1.00
time               0.54 0.08  6.56 2.7e-09     0.38     0.70 1.01
impulsivity       -0.20 0.08 -2.47 1.5e-02    -0.37    -0.04 1.00
time*impulsivity   0.23 0.08  2.77 6.8e-03     0.06     0.39 1.01

Residual Standard Error =  0.82  with  96  degrees of freedom

 Multiple Regression
         R   R2  Ruw R2uw Shrunken R2 SE of R2 overall F df1 df2        p
arousal 0.6 0.35 0.53 0.28        0.33     0.07     17.55   3  96 3.65e-09
```

It is hard to visualize an interaction, and so it is better to plot it. That will be coming soon.