Quick Introduction to R

Stats 500, Fall 2017
Brian Thelen, University of Michigan
443 West Hall, bjthelen@umich.edu

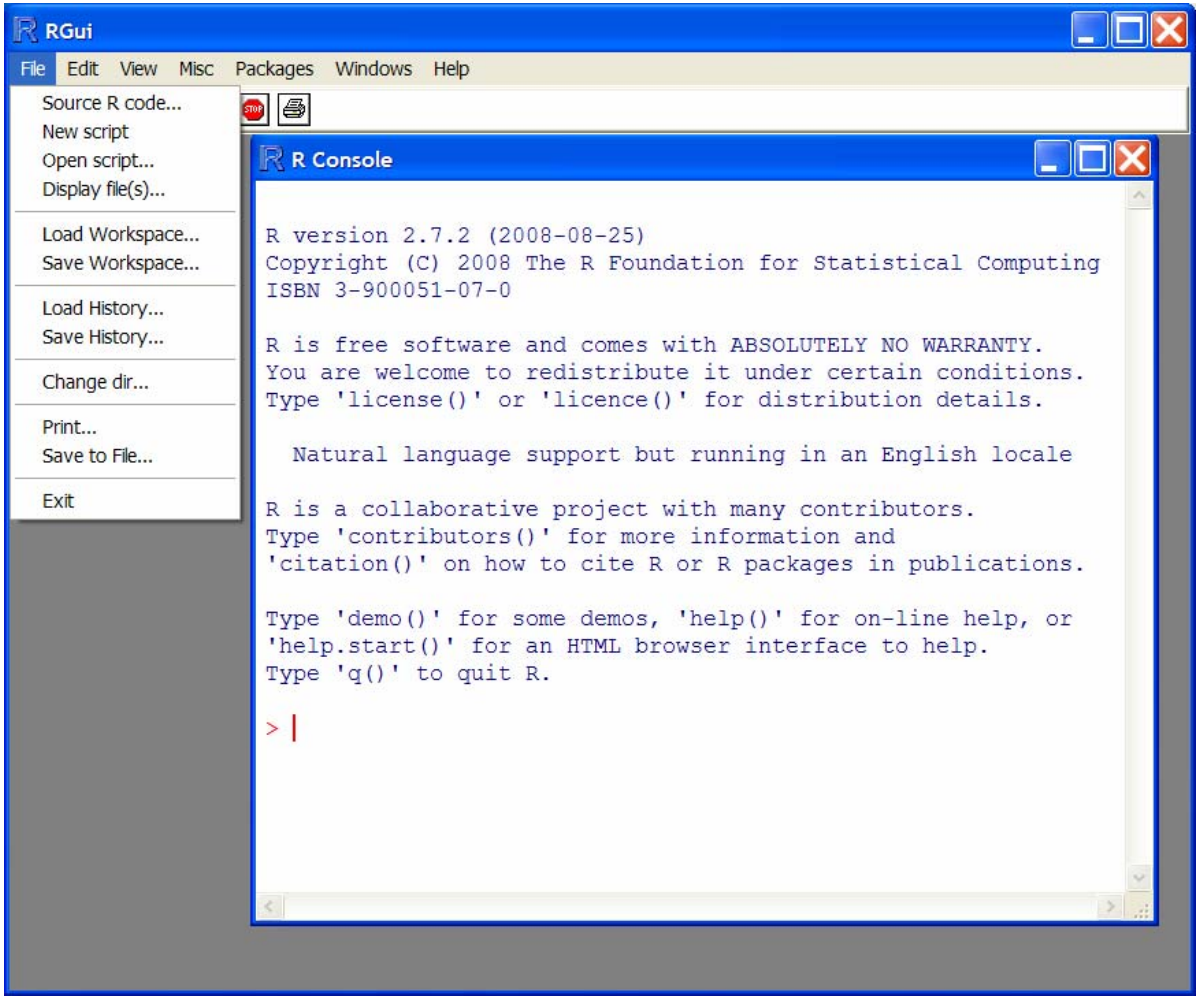
R: Overview

- R is a free software environment for statistical computing and graphics
- Runs on UNIX platforms, Windows and Mac
- Website is http://www.r-project.org/
  - Downloadable precompiled binary distribution (for all of the platforms) available on Comprehensive R Archive Network (CRAN) sites
  - Example CRAN site is http://www.biometrics.mtu.edu/CRAN/
- Freeware version of S (S-Plus) developed at Bell Labs
- Many add-on packages avaiable via CRAN

R References and Resources on Canvas

- **R-refcard** – quick reference card for many R-commands
  - Quick high-level list/description
- **R-intro** – brief introduction to structure of R and how to do most basic things
- **R-debuts** – another brief introduction to structure of R and how to do most basic things
- Under "Help Button" in RGui, there are a lot of different aids as well
  - R-reference manual is the most authoritative and complete (relatively easy to search)

RGui for Windows



Remark. R is an object-oriented language – objects are variables, data, functions, results, etc,. These are stored in the active memory of the computer in the form of objects which have a name. Below are simple data definition and arithmetic operations.

```
> n <-10
> x <-c(1,3,5)
> n
[1] 10
> x
[1] 1 3 5
> xx <-c(1,3,5)+1
> xx
[1] 2 4 6
> ls()
[1] "n" "x" "xx"
> rm(n,x)
> ls()
[1] "xx"
> y <- 2*xx
> y
```

Pima Data Example: Exploratory Data Analysis

```
## Load the library
> library(faraway)
## Read in the data
> data(pima)
> pima
    pregnant glucose diastolic triceps insulin  bmi ...
1          6     148        72      35       0 33.6 ...
2          1      85        66      29       0 26.6 ...
3          8     183        64       0       0 23.3 ...
...
767        1     126        60       0       0 30.1 ...
768        1      93        70      31       0 30.4 ...
```

```
> help(pima)
 The dataset contains the following variables
     'pregnant' Number of times pregnant
     'glucose' Plasma glucose concentration at 2 hours
               in an oral glucose tolerance test
     'diastolic' Diastolic blood pressure (mm Hg)
     'triceps' Triceps skin fold thickness (mm)
     'insulin' 2-Hour serum insulin (mu U/ml)
     'bmi' Body mass index (weight in kg/(height in m)^2)
     'diabetes' Diabetes pedigree function
     'age' Age (years)
     'test' test whether the patient shows signs of
            diabetes (coded 0 if negative, 1 if positive)
```

```
## Dimension of the data
> dim(pima)
[1] 768   9

## Numerical Summaries
> summary(pima)
    pregnant        glucose      diastolic        triceps
 Min.   : 0.0   Min.   :  0   Min.   :  0   Min.   : 0
 1st Qu.: 1.0   1st Qu.: 99   1st Qu.: 62   1st Qu.: 0
 Median : 3.0   Median :117   Median : 72   Median :23
 Mean   : 3.9   Mean   :121   Mean   : 69   Mean   :21
 3rd Qu.: 6.0   3rd Qu.:140   3rd Qu.: 80   3rd Qu.:32
 Max.   :17.0   Max.   :199   Max.   :122   Max.   :99
```

```
    insulin          bmi             diabetes          age       
 Min.   :  0   Min.   : 0.0   Min.   :0.08   Min.   :21  
 1st Qu.:  0   1st Qu.:27.3   1st Qu.:0.24   1st Qu.:24  
 Median : 31   Median :32.0   Median :0.37   Median :29  
 Mean   : 80   Mean   :32.0   Mean   :0.47   Mean   :33  
 3rd Qu.:127   3rd Qu.:36.6   3rd Qu.:0.63   3rd Qu.:41  
 Max.   :846   Max.   :67.1   Max.   :2.42   Max.   :81  
      test       
 Min.   :0.000  
 1st Qu.:0.000  
 Median :0.000  
 Mean   :0.349  
 3rd Qu.:1.000  
 Max.   :1.000  
```

```
## Missing Values
> sort(pima$diastolic)
  [1]   0   0   0   0   0   0   0   0   0   0   0   0
 [13]   0   0   0   0   0   0   0   0   0   0   0   0
 [25]   0   0   0   0   0   0   0   0   0   0   0  24
 [37]  30  30  38  40  44  44  44  44 ...
> pima$diastolic[pima$diastolic == 0] = NA
> pima$glucose[pima$glucose == 0] = NA
> pima$triceps[pima$triceps == 0] = NA
> pima$insulin[pima$insulin == 0] = NA
> pima$bmi[pima$bmi == 0] =NA
```

```
## Categorical Variable
> pima$test = factor(pima$test)
> summary(pima$test)
  0   1
500 268
## Shortcut notation - can skip typing the dataset name
> attach(pima)
> summary(test)
  0   1
500 268
> levels(pima$test) = c("negative", "positive")
```

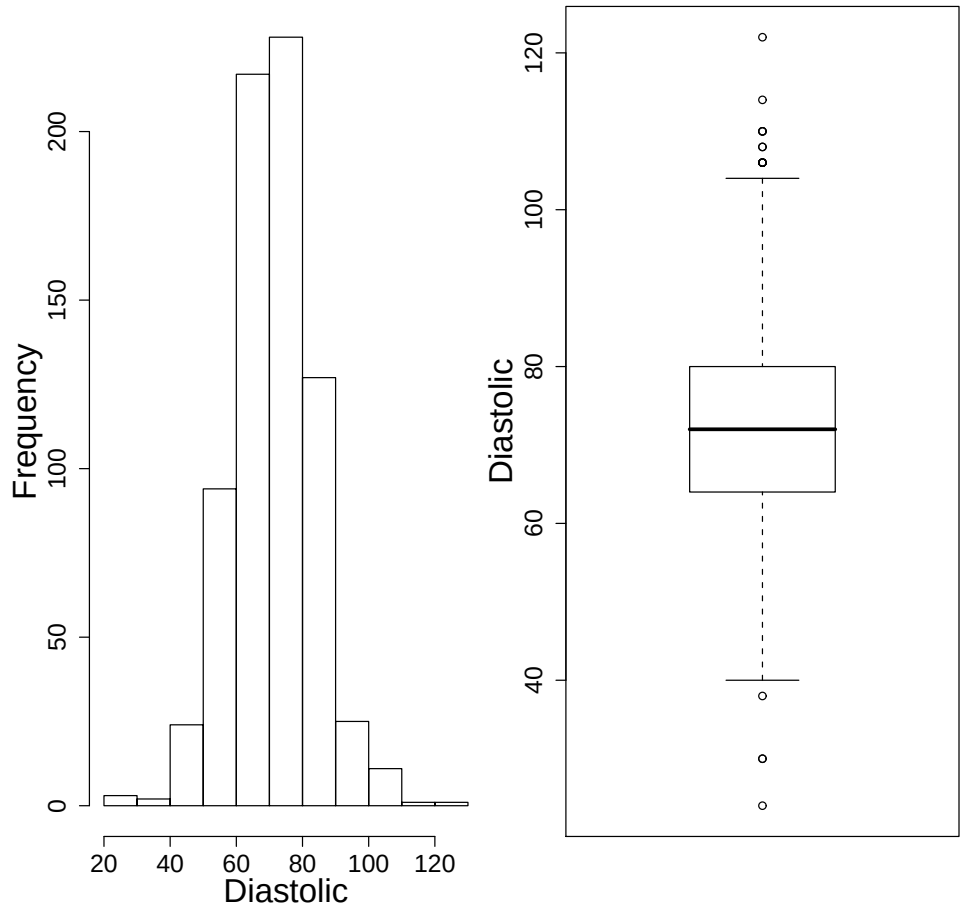
```
## New Summary
> summary(pima)
    pregnant         glucose       diastolic       triceps
 Min.   : 0.0   Min.   : 44   Min.   : 24   Min.   :  7
 1st Qu.: 1.0   1st Qu.: 99   1st Qu.: 64   1st Qu.: 22
 Median : 3.0   Median :117   Median : 72   Median : 29
 Mean   : 3.8   Mean   :122   Mean   : 72   Mean   : 29
 3rd Qu.: 6.0   3rd Qu.:141   3rd Qu.: 80   3rd Qu.: 36
 Max.   :17.0   Max.   :199   Max.   :122   Max.   : 99
                NA's   :  5   NA's   : 35   NA's   :227
```

```
    insulin          bmi             diabetes          age       
 Min.   : 14   Min.   :18.2   Min.   :0.08   Min.   :21  
 1st Qu.: 76   1st Qu.:27.5   1st Qu.:0.24   1st Qu.:24  
 Median :125   Median :32.3   Median :0.37   Median :29  
 Mean   :156   Mean   :32.5   Mean   :0.47   Mean   :33  
 3rd Qu.:190   3rd Qu.:36.6   3rd Qu.:0.63   3rd Qu.:41  
 Max.   :846   Max.   :67.1   Max.   :2.42   Max.   :81  
 NA's   :374   NA's   :11.0                              
      test     
 negative:500  
 positive:268  
```

```
## Individual summary functions
> mean(pima$diastolic, na.rm=T)
[1] 72.40518
# get all numeric variables at once
> mean(pima, na.rm=T)
> median(pima$diastolic, na.rm=T)
[1] 72
> range(pima$diastolic, na.rm=T)
[1]  24 122
> quantile(pima$diastolic, na.rm=T)
  0%  25%  50%  75% 100%
  24   64   72   80  122
## Other functions:  var(), sd()
```

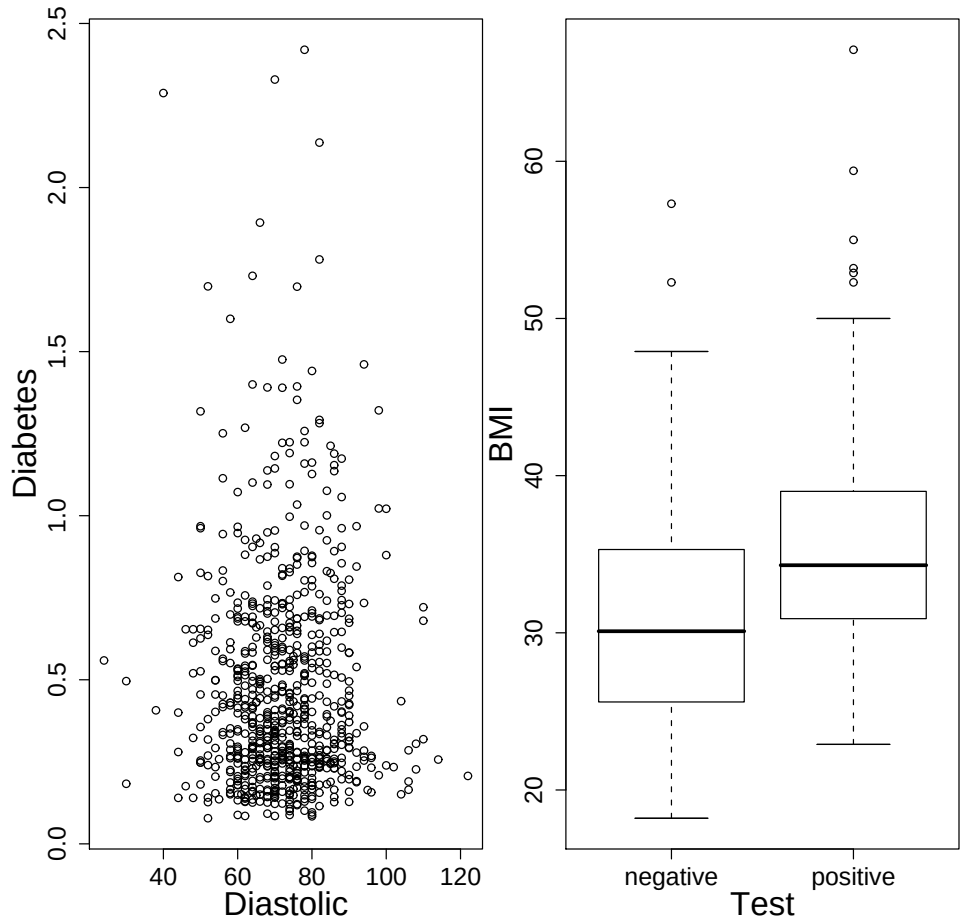
Graphical Summaries: single variable

```
> hist(pima$diastolic)
> boxplot(pima$diastolic)
```



Graphical Summaries: two variables

```
> plot(pima$diastolic, pima$diabetes)
> plot(pima$test, pima$bmi)
```



```
## Selecting Subsets of the Data
## The second row
> pima[2,]
  pregnant glucose diastolic triceps insulin
2        1      85        66      29      NA
       bmi diabetes age     test
2     26.6    0.351  31 negative
## The third column
> pima[,3]
  [1]  72  66  64  66  40  74  50  NA  70 ...
## The (2,3) element
> pima[2,3]
[1] 66
```

```
## The first, second and fourth row
> pima[c(1,2,4), ]
  pregnant glucose diastolic triceps insulin ...
1        6     148        72      35      NA ...
2        1      85        66      29      NA ...
4        1      89        66      23      94 ...
## The third through sixth rows
> pima[3:6, ]
  pregnant glucose diastolic triceps insulin ...
3        8     183        64      NA      NA ...
4        1      89        66      23      94 ...
5        0     137        40      35     168 ...
6        5     116        74      NA      NA ...
```

```
## "Everything but"
> pima[, -c(1,2)]
    diastolic triceps insulin  bmi diabetes age     test
1          72      35      NA 33.6    0.627  50 positive
2          66      29      NA 26.6    0.351  31 negative
3          64      NA      NA 23.3    0.672  32 positive
...

## Cases which have pregnant greater than 14
> pima[pima$pregnant > 14, ]
    pregnant glucose diastolic triceps insulin ...
89        15     136        70      32     110 ...
160       17     163        72      41     114 ...

## Help
> help(boxplot)
> ?boxplot
> help('*')
> help.start()
```