1 Overview

Start R Your first task is to start R. Simple?

Run a command Try to run the following command:

```
> 2 + 2
```

```
[1] 4
```

Now try to edit the command – use the arrow keys

Create a data set Create the following data vectors: The even numbers 2 through 20 stored as `evens`; The prime numbers from 2 through 29 stored as primes.

Apply a function to the data set For the `evens` data set find the mean and variance. For the `primes` data set find the median and IQR.

For the `evens` data set what is done by `cumsum(evens)`?

Look up help on a function The `mean` function has an option for trimming. Read the manual page and compute the 10% trimmed mean for the primes.

Load a built-in data set Load the built-in data set `rivers`. Call the `stem` function on the data set. Is the data set skewed?

Install and load an external package Install and load the `UsingR` data set. For its `babies` data set answer the following:

1. How many variables are there? What types are they? What are there names? How many subjects are there?
2. Find numeric summaries of the weight of the baby (`wt`). Do a stem and leaf plot to see if there are any outliers, if so find the trimmed mean (10%).
3. Repeat for the Mother's weight `wt1`.

Find graphical summaries of a data set Again for the `babies` data set: Find a histogram of `wt` and `wt1`.

Make a density plot of the `ht` variable.

Make a boxplot of the `gestation` variable. Look into the manual page to see how to make it go left to right.

Bivariate summaries Find the correlation between `wt` and `mpg` for the built in `mtcars` data set.

What does the command `cor(mtcars)` do?

What does the command `pairs(mtcars)` do?

Make a scatter plot of the two variables.

Load the data set `Cars93` from the `MASS` package. Make a scatterplot of `Weight` on the x axis and `MPG.city`on the y axis. Colour according to `Cylinder` and adjust the size based on `Type`.

This webpage has a short bit on visualzing correlation matrices: `http://blog.revolution-computing.com/2009/03/visualizing-correlation-matrices.html` The

On the webpage: `http://labs.dataspora.com/gameday/pitcher/johan-santana/276371` we have a very interesting pair of graphics to baseball fans. These graphics are produced using R and the `RApache` package. Here we are interested in looking the the data which can be downloaded with:

```
> JS <- read.csv("http://www.math.csi.cuny.edu/verzani/tmp/276371.csv")
```

Using qplot, make a scatter plot of the x and z variables using colour, size or facets to show the other variables.

Head to the website `http://learnr.wordpress.com/` and see if you can make any of the graphics.

From the mailing list:

```
hello there,

Is there a way of truncating in the opposite direction so as to retain only
the values to the right of the decimal??

i.e. rather than:
> trunc(39.5)
[1] 39

i would get something like:
> revtrunc(39.5)
[1] 0.5

I've been searching to no avail but I imagine there is a very simple
solution!
Tyler
```

Any help? (just to show what trunc does)

2 Data Manipulation

Understand structure of R objects Load the `Aids2` data set in the `MASS` package. Describe the structure of the data set: how many variables, cases, types of variables, ...

Use editor and read.table Enter the following values into R using an editor and `read.table`

```
Age   Height
10      48
10      52
11      54
11      60
12      51
```

Use write.csv, read.csv Open Excel, enter the same data, save as a "csv" file and then read into R via `read.csv`.

Creating sequences, random data Create the following sequences using `:` or `seq` or `rep` or some other means.

1. 5,4,3,2,1
2. 2,4,6,8,10
3. 1,1,1,1,1
4. 1,1,1,1,1,2,2,2,2,2

The sample function How many times on average would you run `sample(1:3)` before the output is 3,2,1? Try it

Subscripting: vectors, matrices, data frames, lists Assign `x` to be `1:10`. Find commands that return: all but the first element, all but the last element, all elements more than 5, the index of the elements that are even (`%%`).

Which of these (any, some, all?) will remove a row from a data frame?

```
> require(MASS)
> m <- Cars93[1:5, 1:6]
> names(m)

[1] "Manufacturer" "Model"        "Type"         "Min.Price"    "Price"
[6] "Max.Price"

> byNegativeIndex <- m[, -2]
> usingLogical <- m[, names(m) != "Model"]
> usingListInfo <- m
> usingListInfo$Model <- NULL
> usingSubset <- subset(m, select = names(m) != "Model")
```

From the R mailinglist:

So how do I select all temperatures of 90 and 80 ie Temp = c(80,90) from the `airquality` data set?

From the R mailinglist

I am trying to extract one particular piece of data(Float) from all the data returned by yahooKeystats, but thus far I'm having no luck. This is what I've got so far:

```
> library(fImport)              # must install
Loading required package: timeSeries
Loading required package: timeDate
> data<-yahooKeystats("IBM")
```

Is there a way to just pull out the number of float shares? Something along the lines of:
`data$Float` or `data[Float]`?

From the R mailinglist

Hello all,
I have to import numeric data from file but found it contains Infinite values which need to be eliminated. I tried to replace them in this way:

```
data[which(data=="-Inf")] <- -0.3
data[which(data=="+Inf")] <-  0.3
```

Start with

```
> x <- c(Inf, 1, 2)
```

and replace the "Inf" with 0.3.

Working with factors

Reshaping data 1. From the mailing list:

Hi. I have a 925 by 925 correlation matrix corM. I want to identify all variables that have correlation greater than 0.9. Can anyone suggest an "R way" of doing this?

Have a go:

```
> n <- 10
> m <- matrix(runif(n * n), nrow = n)
```

Some solutions:

```
> data.frame(m) > 0.9
```

```
         X1    X2    X3    X4    X5    X6    X7    X8    X9   X10
 [1,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [2,] FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
 [3,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
 [4,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
```

```
 [5,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
 [6,] FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
 [7,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
 [8,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [9,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[10,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

Or keeping as a matrix

```
> matrix(m > 0.9, nrow = n)
```

```
       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10]
 [1,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [2,] FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
 [3,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
 [4,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
 [5,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
 [6,] FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
 [7,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
 [8,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [9,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[10,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

Or (Uwe Ligges), the following

```
> rownames(m) <- colnames(m) <- letters[1:n]
> iind <- row(m)[m > 0.9]
> jind <- col(m)[m > 0.9]
> cbind(rownames(m)[iind], colnames(m)[jind])
```

```
     [,1] [,2]
[1,] "f"  "b"
[2,] "b"  "c"
[3,] "b"  "g"
[4,] "f"  "g"
[5,] "c"  "i"
[6,] "d"  "i"
[7,] "e"  "i"
[8,] "g"  "i"
```

Using `which` (Dimitris):

```
> which(m > 0.9, arr.ind = TRUE)
```

```
  row col
f   6   2
b   2   3
b   2   7
f   6   7
c   3   9
d   4   9
e   5   9
g   7   9
```

2. Is there a way to omit empty cells?

```
> a <- c(1.1, 2.1, 9.1)
> b <- cut(a, 0:10)
> c <- data.frame(b, b)
> d <- table(c)
> dim(d)
[1] 10 10
```

Using `interaction`

```
> table(interaction(c, drop = T))

  (1,2].(1,2]   (2,3].(2,3] (9,10].(9,10]
            1             1             1
```

Hadley W. Solution:

```
       b    b.1 V1
1  (1,2]  (1,2]  1
2  (2,3]  (2,3]  1
3 (9,10] (9,10]  1
```

3. Again from the mailing list:

   How can I delete both rows and columns that do not meet a particular cut off value. Example:

```
> d <- rbind(c(0, 1, 6, 4), c(2, 5, 7, 5), c(3, 6, 1, 6), c(4,
+     4, 4, 4))
> f <- as.matrix(d)
> f
     [,1] [,2] [,3] [,4]
[1,]    0    1    6    4
[2,]    2    5    7    5
[3,]    3    6    1    6
[4,]    4    4    4    4
```

   I would like to delete all rows and columns that do not contain at least one element with a value less than 1. So I'd end up with:

```
> f
     [,1] [,2] [,3]
[1,]    0    1    6
[3,]    3    6    1
```

   Several solutions for you to try to understand: Stephan Kolassa:

```
> f[rowSums(f <= 1) > 0, colSums(f <= 1) > 0]

     [,1] [,2] [,3]
[1,]    0    1    6
[2,]    3    6    1
```

   Wacek Kusnierczyk has this elegant method:

```
> dd = d <= 1
> d[apply(d, 1, any), apply(d, 2, any)]
     [,1] [,2] [,3] [,4]
[1,]    0    1    6    4
[2,]    2    5    7    5
[3,]    3    6    1    6
[4,]    4    4    4    4
```

Rolf Turner gives similar using a function (functions to come later)

```
> d[apply(d, 1, function(x) {
+     any(x <= 1)
+ }), apply(d, 2, function(x) {
+     any(x <= 1)
+ })]
     [,1] [,2] [,3]
[1,]    0    1    6
[2,]    3    6    1
```

3 Inference

The t-test (MASS) The `shoes` data set in `MASS` has amount of shoe wear on R and L shoes of 10 boys. Do a one sample test of `A` with a mean of 10 for H_0.

Do two sample paired and unpaired tests for equivalence of means. Repeat with the `wilcox.test` function.

Lattice graphics (MASS) For the `whiteside` data set in the `MASS` package, make an `xyplot` with Gas as a dependent variable and Temp a dependent variables. Include two panels base on the value of Insul.

(HSAUR) The `plasma` data set from the `HSAUR` package has variables `ESR`, `fibrinogen` and globulin. `ESR` is a thresholded sedimentation rate, and is binary.

A plot of the conditional distribution is given by `cdplot`. What does this graphic show: `cdplot(ESR ~ fibrinogen, data=plasma)`? Now fit a logistic regression model, storing the model fit in the variable `res`. What does this command return: `confint(res, parm='fibrinogen')`? (The CI for log-odds increase by stepping 1 unit.)

Simple regression Fit a linear model for `Gas` with `Temp` as a predictor for each level of `Insul`.

Assessment of model fit Look at the diagnostic plots for the two models you just made.

Multiple regression Use `anova` to compare the models `Gas ~ Temp` and `Gas ~ Temp + Insul`.

(MASS) For the `hills` data set in the `MASS` package fit the model `time ~ dist + climb`. Look at the diagnostic plots to investigate outliers in the linear model.
Can you refit the above, excluding the "Knock Hill" data point? Excluding both "Knock Hill" and "Ben of Jura"?
The `weight` argument allows for weighted regression. Fit the model with `weight=1/dist ^2`.

Binary logistic regression (Neely, http://www.stat.wisc.edu/ neely/Site/CDE

The `Chile` data set in the `car` package are from a nation survey in Chile. There are missing values in the data set. (See help page).

Make an exploratory plot of `vote` versus `statusquo` for the votes which were Y or N (The variable has 4 levels, we restrict to two for logistic regression purposes.)

Fit the logistic regression model to predict `vote` based on `statusquo` for those who voted Y or N.

4 Programming

For loops Use a for loop to compute $\binom{52}{5}$ writing this as

$$\binom{52}{5} = \frac{52}{5} \cdot \frac{51}{4} \cdot \frac{50}{3} \cdot \frac{49}{2} \cdot \frac{48}{1}$$

((The `choose` function is a better alternative.)

if the matrix `m` is generated by the command `matrix(rnorm(100), nrow=10)`, use a for loop to find the median of each column. (An alternative to `apply(m,2,median)`.

From the mailing list:

> Hi everyone, I am trying to accomplish a small task that is giving me quite a headache. I would like to automatically generate a series of matrices and give them successive names. Here is what I thought at first:
>
> ```
> t1<-matrix(0, nrow=250, ncol=1)
>
> for(i in 1:10){
> t1[i]<-rnorm(250)
> }
> ```
>
> What I intended was that the loop would create 10 different matrices with a single column of 250 values randomly selected from a normal distribution, and that they would be labeled t11, t12, t13, t14 etc.
> Can anyone steer me in the right direction with this one?
>
> Thanks!
> Brendan

Solution by Jorge Ivan Velez:

```
> bigt <- sapply(1:10, function(x) rnorm(5))
> colnames(bigt) <- paste("t", 1:10, sep = "")
> bigt
```

```
             t1          t2          t3         t4          t5         t6
[1,] -1.03358946  0.49541963  0.08239991  0.9333845  1.47921195 -0.6188329
[2,] -0.05923464 -0.02887876 -0.23569494 -1.3016770  0.08373286 -1.3251205
[3,]  0.18630620 -1.08695434  1.91712716 -0.4095879  1.37963134  0.2870747
[4,] -0.17340744 -1.14212533 -1.11947089  0.1360216 -0.17080010 -0.5865833
[5,]  0.05581644 -0.31621291 -0.89486570 -2.0215362 -0.07912180  1.3948760
            t7         t8         t9         t10
[1,]  1.1431901 -0.3121529 -1.8006786  0.85110089
[2,]  0.1567099  0.6563149  1.8107636  1.01191529
[3,] -0.4295507 -0.5949102  2.3688001 -0.07232937
[4,]  0.8276459 -0.2627074 -1.2046398 -0.23821795
[5,] -1.7580465 -1.0337979  0.5236957  1.09290875
```

From the mailing list:

```
I would like to create a matrix in R that looks similar to this:

     [,1] [,2] [,3] [,4]
[1,]  NaN  1  2  3
[2,]  NaN  1  2  4
[3,]  NaN   1  2  5
[4,]  NaN  2  3  4
[5,]  NaN  2  3  5
[6,]  NaN    3    4    5

I have the loop below:

where A for example is 5

matrixx<-function(A){
B=matrix(NaN,nrow=(A+1),ncol=4)
        for(k in 1:(A+1)){
        for(i in 1:(A-2)){
     for(j in (i+2):A){
     }
}
       }
B[k,]=c(NaN,i,(i+1),j)
print(B)
}

But it only prints the final line in:

> matrixx(5)
     [,1] [,2] [,3] [,4]
[1,]  NaN  NaN  NaN  NaN
[2,]  NaN  NaN  NaN  NaN
[3,]  NaN  NaN  NaN  NaN
[4,]  NaN  NaN  NaN  NaN
[5,]  NaN  NaN  NaN  NaN
[6,]  NaN    3    4    5

Could anyone give me a hand? Would be much appreciated.

Thanks Emma
```

Hadley Wickam offers

```
> candidates <- t(combn(5, 3))
> firstdiff <- candidates[, 2] - candidates[, 1]
```

```
> cbind(NaN, candidates[firstdiff == 1, ])

     [,1] [,2] [,3] [,4]
[1,]  NaN    1    2    3
[2,]  NaN    1    2    4
[3,]  NaN    1    2    5
[4,]  NaN    2    3    4
[5,]  NaN    2    3    5
[6,]  NaN    3    4    5
```

Define a function Define your own functions to do find the mean and standard deviation of a variable. How would you write them so that the input could be a vector, matrix or data frame?

Can you write a function to compute $\binom{n}{k}$ writing this as

$$\binom{n}{k} = \prod \frac{n}{k} \cdot \frac{n-1}{k-1} \cdots \frac{n-k+1}{k-k+1}$$

The negative log-likelihood function for n normal data points is a function of parameters μ, σ that depends on data x defined by

$$L(\mu, \sigma; x) = -\sum_i \log \phi(x_i; \mu, \sigma).$$

Write a function that produces this function that depends on x, mu and sigma.

The book *Bayesian Computation with R* by Jim Albert has code to compute the log posterior density, similar to the last question. In that text, the function has one argument for the parameters (`theta`) and one for the data (`data`). An example computation is if one places a Normal(10, 20) prior on μ and flat prior on $\log\sigma$ then the log posterior has the form

$$\log g(\theta|x) = \log \phi(\mu; 10, 20) + \sum_i \log \phi(x_i; \mu, \sigma)$$

Implement this in a function. Can you write the function for the case that `theta` is a n by 2 matrix of values and the return value is an n by 1 matrix?

The `optim` function will find a minimum of a function. It uses a function for the second argument. Here is a silly way to solve for the minimum of a quadratic

```
> f <- function(x) x^2 - 3 * x + 10
> out <- optim(c(x = 1), f)
> f(out$par) - out$value

x
0
```

Write a function to evaluate $f(x) = x^6 - 3x + 12$ and use `optim` to solve for the point of minimum.

Simulation Write a simulation to study the central limit theorem when the population is the t distribution with small degrees of freedom.

From the mailinglist

> Dear R People:
> Has anyone produced code for a Julia set, please?
> It's not all that tough to do, just thought I'd check before re-inventing the wheel.

Recursion Can you write a recursive function to find $n!$ for positive integer n?

Aggregation For the `iris` data set, find numeric summaries of each of the first 4 variables for each level of `Species`.