Basic Statistics in R

Debolina Ghatak

July 26, 2020

1. Data Input
2. Some Matrix Operations
3. Descriptive Statistics
4. Simulation and Distribution Functions
5. Plots

Direct Input

- *Data in Vector form*: Say height of 10 students in class.
  - Height values in cm: 157, 162, 166, 175, 182, 177, 164, 159, 174, 179.
  - R Code:
    $> c(157, 162, 166, 175, 182, 177, 164, 159, 174, 179)$
  - Output:[1] 157 162 166 175 182 177 164 159 174 179
- *Data in Matrix form*: Say height, weight and sex of 10 students:
  - Height values in cm as before. Weight values in kg: 62, 65, 73, 63, 82, 85, 72, 55, 64, 65.
    Sex: F M M M M M F F F M
  - R Code for direct matrix entry:
    $> matrix(c(157, 162, 166, 175, 182, 177, 164, 159, 174, 179, 62, 65, 73, 63, 82, 85, 72, 55, 64, 65, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0), nrow = 10)$
  - Output:

Continued

```
      [,1] [,2] [,3]
 [1,]  157   62    1
 [2,]  162   65    0
 [3,]  166   73    0
 [4,]  175   63    0
 [5,]  182   82    0
 [6,]  177   85    0
 [7,]  164   72    1
 [8,]  159   55    1
 [9,]  174   64    1
[10,]  179   65    0
```

- However there is a different method to perform the same activity. To label the vectors and join them.

Continued

- At first we need to name the variabes, then use R codes to join them.
- R Code:
  $> Heights = c(157, 162, 166, 175, 182, 177, 164, 159, 174, 179)$
  $> Weights = c(62, 65, 73, 63, 82, 85, 72, 55, 64, 65)$
  $> Sex = c(1, 0, 0, 0, 0, 0, 1, 1, 1, 0)$
  $rbind(Heights, Weights, Sex)$
  $cbind(Heights, Weights, Sex)$
- Note that the cbind code will give us the same result as before while rbind will give us it's transpose matrix.

Reading a file

- Usually files can be saved in either .txt or .csv format to be able to read in R. The two files were stored in "Home" Dirctory in Ubuntu. For Windows the files can be saved either in "My Documents" or the location can be added in R code.
- $read.table("DATA\_SIM.txt", header = FALSE)$
- $read.csv("cerealsugar.csv")$

Dimension, Transpose and Sum

- *Length of a vector* A can be determined in R.
  Rcode: length(A)
- *Dimension of a matrix* A can be determined in R.
  Rcode: dim(A)
- *Transpose of a matrix* A can be determined in R.
  Rcode: t(A)
- *Sum of two matrices* A and B can be determined in R.
  Rcode: A+B

Multiplication and Inverse

- For two real numbers a and b their sum, difference, product and ratio are given respectively by the codes
  Rcode: $a + b, a - b, a * b, a/b$
- Unlike multiplication of numbers, *matrix multiplication* has a different style. For two matrices A and B their matrix product is given by,
  Rcode: $A\% * \%B$
- The *Determinant of a matrix* A can be found using
  Rcode: det(A)
- The *Inverse of a matrix* A can be found using
  Rcode: solve(A)

Rank and Eigenvalues

- To find rank of matrix A,
  Rcode: *qr(A)$rank*
- To find eigen values or vectors
  Rcode: *eigen(A)$values* , *eigen(A)$vectors*

Measures of Location

- Sum of a elements of a vector A:
  Rcode: $sum(A)$
- Mean of a vector:
  Rcode: $mean(A)$
- Product of elements of a vector:
  Rcode: $prod(A)$
- Geometric mean of a vector:
  Rcode: $(prod(A))\hat{}(1/length(A))$
- Median of a vector:
  Rcode: $median(A)$

Measures of Dispersion

- Variance of a vector A:
  Rcode: $var(A)$
- Standard Deviation of a vector:
  Rcode: $sd(A)$ or $sqrt(var(A))$
- Range of a vector:
  Rcode: $max(A) - min(A)$
- Quartiles of a vector:
  Rcode: $quantile(A, 0.25), quantile(A, 0.75)$
- If no other analysis is required and one wants to see the values of Maximum, Minimum and quartiles,
  Rcode: $summary(A)$

Sampling from a set of elements

From a given set of elements, say, Space, one can sample a few elements with or without replacement. Varying the probability of selction of each element is also allowed.

Rcode: sample(Space,10,prob=probv,replace=TRUE)

Here 10 is the sample size, probv denotes the probability vector of selection of each element in Space.

Sampling from distributions

- A sample of size 10 from Binomial distribution with parameters 10 and 0.5:
  Rcode: $rbinom(10, 10, .5)$
- A sample of size 10 from Poisson distribution with parameter 5:
  Rcode: $rpois(10, 5)$
- A sample of size 10 from Normal distribution with parameters 2 (mean) and 4 (sd):
  Rcode: $rnorm(10, 2, 4)$
- A sample of size 10 from Exponential distribution with parameter $2(1/mean)$:
  Rcode: $rexp(10, 2)$

Distribution functions

- CDF and PMF of a Binomial random variable with parameters 10 and 0.5 at the point 1:
  Rcode: $pbinom(1, 10, 0.5)$; $dbinom(1, 10, 0.5)$
- CDF and PMF of a Poisson random variable with parameter 5 at the point 1:
  Rcode: $ppois(1, 5)$; $dpois(1, 5)$
- CDF and PDF of a Normal random variable with parameters 2 and 4 at the point 0:
  Rcode: $pnorm(0, 2, 4)$; $dnorm(0, 2, 4)$
- CDF and PDF of an Exponential random variable with parameter 2 at the point 2:
  Rcode: $pexp(2, 2)$; $dexp(2, 2)$

Plotting Data

- One can draw a simple *histogram* of data vector A in R and can also specify break points if needed.
  Rcode: $hist(A)$
- Corresponding to some individuals if we have data values for two attributes A and B, their values can be plotted against each other.
  Rcode: $plot(A\ B, main = "Plot", xlab = "B", ylab = "A", type = "p", col = 2, pch = 7)$
  Note that,
  - *main* denotes the title of the plot to be printed.
  - *xlab and ylab* denotes the name of the attributes to be printed.
  - *type* denotes whether the graph will be point graph or lines will be drawn.

Drawing a mathematical function in R

My code for drawing a mathematical function in R will look like:

```
> x=seq(-10,10,0.01) #Specifying points
> y=exp(x) #Calculating the function values
> plot(y~x,main="Exponential Function",xlab="x",
ylab="e^x",col=3,type="l",lty=2,xlim=c(0,10),ylim=c(0,10000))
```

Assignment

- Extract the data vector from the cerealsugar.csv file and calculate their mean,median,standard deviation and print the summary statistics.
- Write the above vector as a 10×10 matrix and find it's rank.
- Plot the values of the DATA_SIM.txt file and their emperical cdf using R code plot.ecdf. Present nicely.