SOLUTIONS MANUAL FOR

Data Analysis and Statistics for Geography, Environmental Science & Engineering

by

Miguel F. Acevedo

CRC Press
Taylor & Francis Group

SOLUTIONS MANUAL FOR

Data Analysis and Statistics for Geography, Environmental Science & Engineering

by

Miguel F. Acevedo

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an informa business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed on acid-free paper
Version Date: 20121219

International Standard Book Number: 978-1-4398-9097-4 (Paperback)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

To the instructor

This manual is supplementary material to the book Acevedo M.F. 2012. *Data Analysis and Statistics for Geography, Environmental Science, and Engineering*. CRC Press. 2013. It provides solutions to the exercises in all chapters of that book. I have included solutions to all exercises in the first nine chapters since these are most likely to be used in introductory courses. For the remaining chapters, I provided solutions for most exercises but excluded some because the exercises are numerous and likely be part of more advanced courses. Solutions to the excluded exercises are provided to adopting instructors upon request.

Contents

Chapter 1 Introduction

Exercise 1-1

Use a variable X to denote human population on Earth. Explain why it varies in time and space and give examples of a value at a particular location or region and time.

Solution:

Human population can be considered a variable that takes discrete values and it varies in continuous time due to deaths and births. An example is $X(t)$=6,23 × 10^9 people or 6,23 billion people at time t.

Exercise 1-2

Suppose you build a model of light transmission through a forest canopy using measured light (treated as dependent variable) at various heights (treated as independent variable) and use it to predict light at those heights where is not measured. Would this be a process-based model or an empirical model?

Solution:

This will be an empirical model because it relates all pairs of sunlight and height values and using a predictor equation (for example, regression) will develop a prediction of sunlight at each height.

Exercise 1-3

Extend exercise 1-2 to use the concept from physics that light is attenuated as it goes through a medium. Propose that attenuation is proportional to the density of foliage at various heights, and then propose a model based on an equation before you collect data. Would this be a process-based model or an empirical model?

Solution:

This will be a process-based model. We state that difference in light intensity L(h) –L(h-dh) for a difference in height dh is proportional to foliage density F(h) via an attenuation coefficient k. This is to say $L(h)-L(h-dh)=kF(h)$.

Exercise 1-4

To make sure you understand the workspace. Save your workspace **.Rdata** file. Then close R and start a new R session, Load the workspace, make sure you have the objects created before.

Solution:

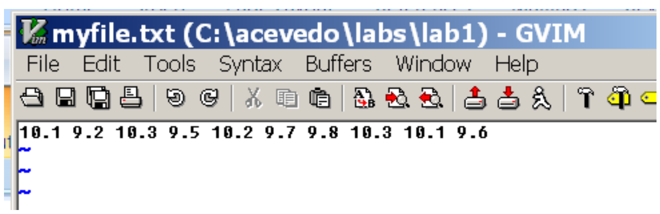
Just proceed as indicated.

Exercise 1-5

Use the notepad or Vim to create a simple text file **myfile.txt.** Type 10 numbers in a row separated by a blank space, trying to type numbers around a value of 10. Save in folder **lab1**. Now read the file using scan, calculate sample mean, variance, and standard deviation, plot a stem-and-leaf diagram and a histogram and discuss.

Solution:

The solution details vary according to the numbers created by each student. Suppose one possible solution is



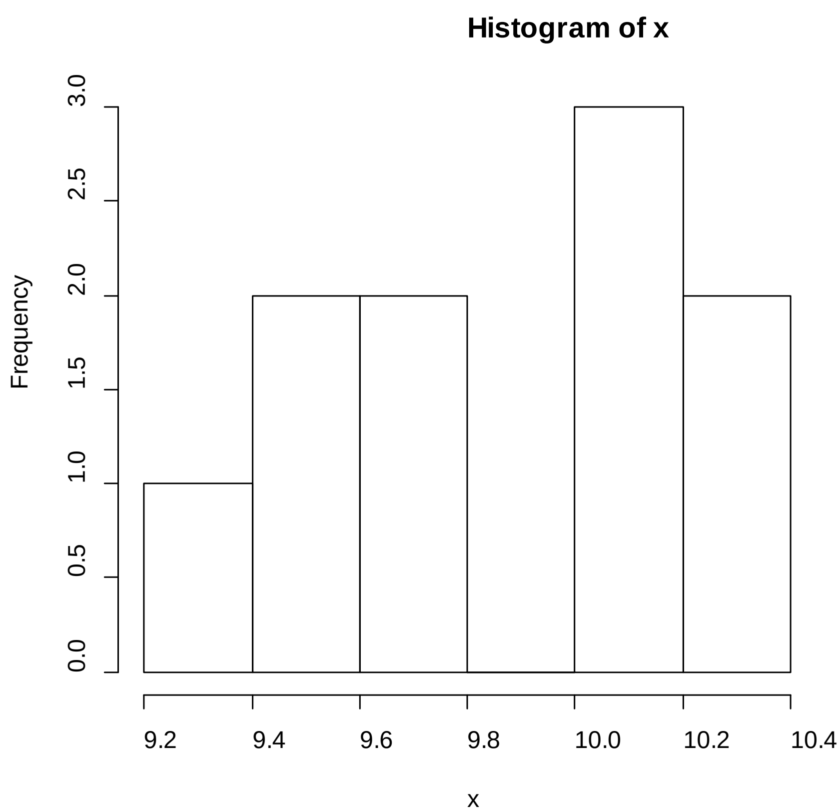

```
x <- scan("lab1/myfile.txt")
length(x)
stem(x)
hist(x)
round(mean(x),1)
round(var(x),1)
round(sd(x),1)
```

On the console obtain

```
> x <- scan("lab1/myfile.txt")
Read 10 items
> length(x)
[1] 10
> stem(x)

  The decimal point is at the |

   9 | 2
   9 | 5678
```

```
  10 | 11233

> hist(x)
> round(mean(x),1)
[1] 9.9
> round(var(x),1)
[1] 0.1
> round(sd(x),1)
[1] 0.4
>
```

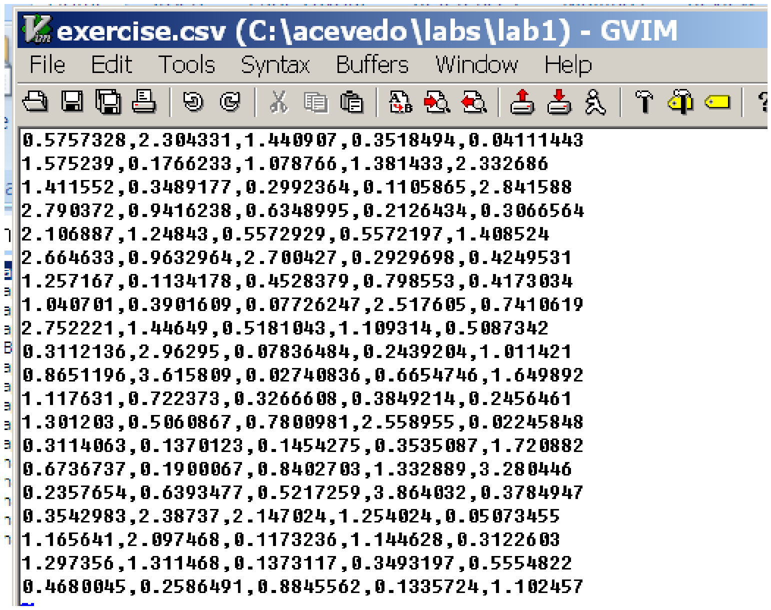
On the graphics window obtain



Exercise 1-6

Use file **lab1\exercise.csv**. Examine the file contents using the notepad or Vim. Read the file, list numbers on the R Console rounding to 2 decimals. Calculate sample mean, variance, and standard deviation, plot a stem-and-leaf diagram and a histogram and discuss.

Solution:



exercise.csv (C:\acevedo\labs\lab1) - GVIM

File Edit Tools Syntax Buffers Window Help

```
0.5757328,2.304331,1.440907,0.3518494,0.04111443
1.575239,0.1766233,1.078766,1.381433,2.332686
1.411552,0.3489177,0.2992364,0.1105865,2.841588
2.790372,0.9416238,0.6348995,0.2126434,0.3066564
2.106887,1.24843,0.5572929,0.5572197,1.408524
2.664633,0.9632964,2.700427,0.2929698,0.4249531
1.257167,0.1134178,0.4528379,0.798553,0.4173034
1.040701,0.3901609,0.07726247,2.517605,0.7410619
2.752221,1.44649,0.5181043,1.109314,0.5087342
0.3112136,2.96295,0.07836484,0.2439204,1.011421
0.8651196,3.615809,0.02740836,0.6654746,1.649892
1.117631,0.722373,0.3266608,0.3849214,0.2456461
1.301203,0.5060867,0.7800981,2.558955,0.02245848
0.3114063,0.1370123,0.1454275,0.3535087,1.720882
0.6736737,0.1900067,0.8402703,1.332889,3.280446
0.2357654,0.6393477,0.5217259,3.864032,0.3784947
0.3542983,2.38737,2.147024,1.254024,0.05073455
1.165641,2.097468,0.1173236,1.144628,0.3122603
1.297356,1.311468,0.1373117,0.3493197,0.5554822
0.4680045,0.2586491,0.8845562,0.1335724,1.102457
```

Then

```
> x.ex <- scan("lab1/exercise.csv",sep=",")
Read 100 items
> round(x.ex,2)
  [1] 0.58 2.30 1.44 0.35 0.04 1.58 0.18 1.08 1.38 2.33 1.41 0.35 0.30 0.11 2.84
 [16] 2.79 0.94 0.63 0.21 0.31 2.11 1.25 0.56 0.56 1.41 2.66 0.96 2.70 0.29 0.42
 [31] 1.26 0.11 0.45 0.80 0.42 1.04 0.39 0.08 2.52 0.74 2.75 1.45 0.52 1.11 0.51
 [46] 0.31 2.96 0.08 0.24 1.01 0.87 3.62 0.03 0.67 1.65 1.12 0.72 0.33 0.38 0.25
 [61] 1.30 0.51 0.78 2.56 0.02 0.31 0.14 0.15 0.35 1.72 0.67 0.19 0.84 1.33 3.28
 [76] 0.24 0.64 0.52 3.86 0.38 0.35 2.39 2.15 1.25 0.05 1.17 2.10 0.12 1.14 0.31
 [91] 1.30 1.31 0.14 0.35 0.56 0.47 0.26 0.88 0.13 1.10
>
```

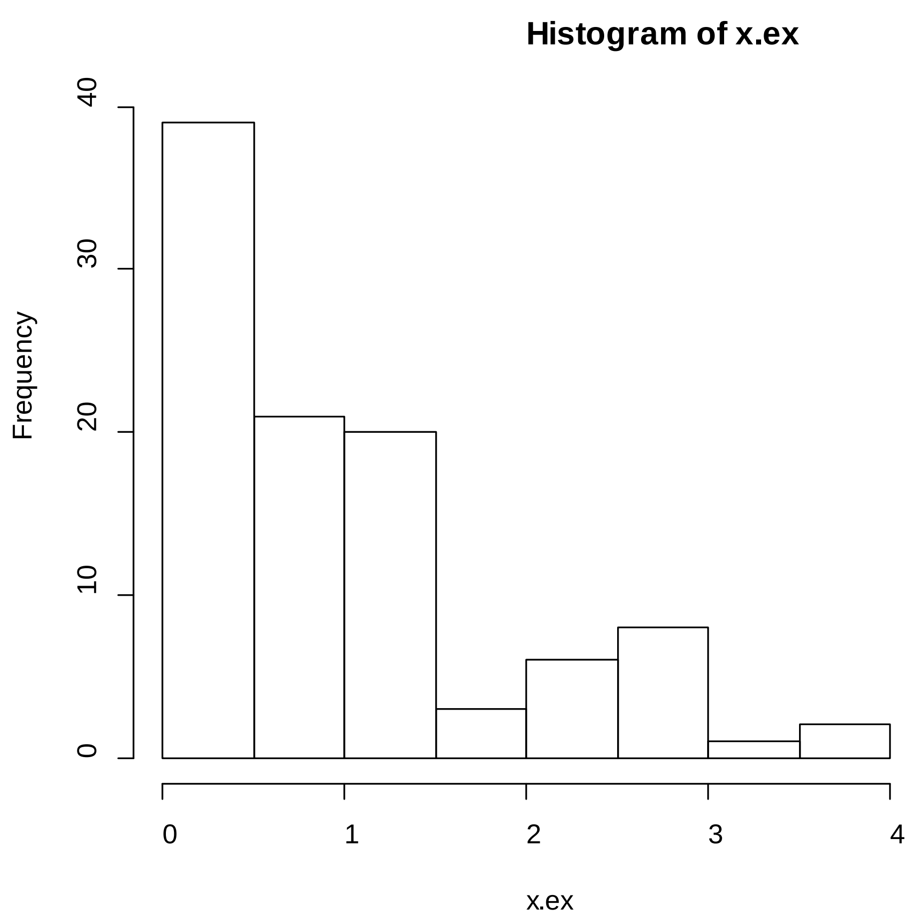
Now

```
> round(mean(x.ex),2)
[1] 1
> round(var(x.ex),2)
[1] 0.82
> round(sd(x.ex),2)
[1] 0.91
> stem(x.ex)

  The decimal point is at the |

  0 | 00011111111112222223333333334444444
  0 | 55555666666777788899
  1 | 000111112233333344444
  1 | 667
  2 | 111334
```

```
  2 | 5677888
  3 | 03
  3 | 69

>
```

Finally

```
hist(x.ex)
```



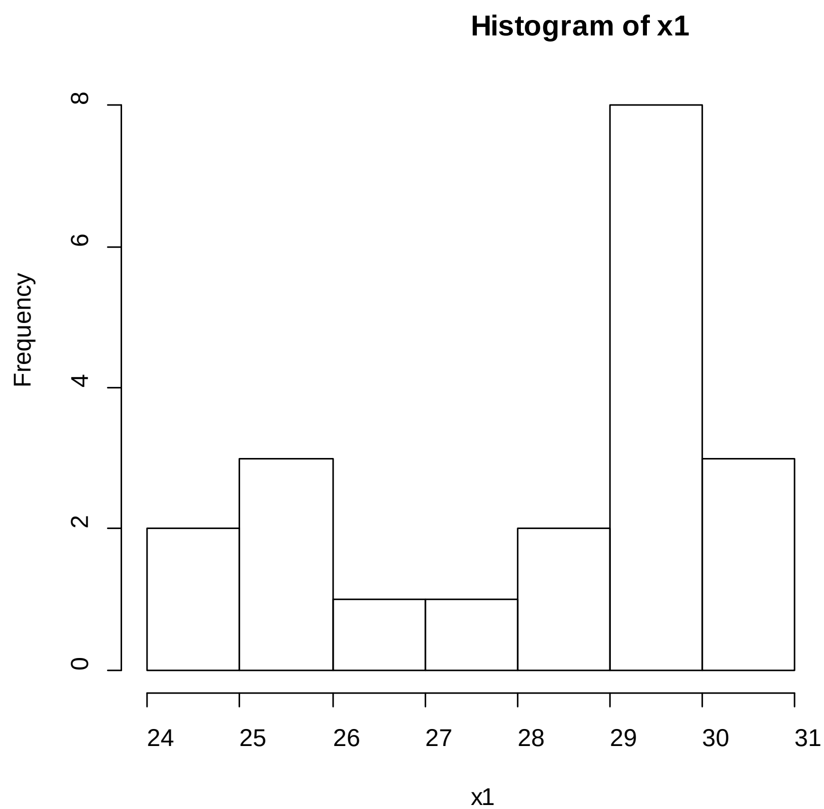
Exercise 1-7

Separate the first 20 and last 20 elements of salinity x array into two objects. Plot a stem-and-leaf plot and a histogram for each.

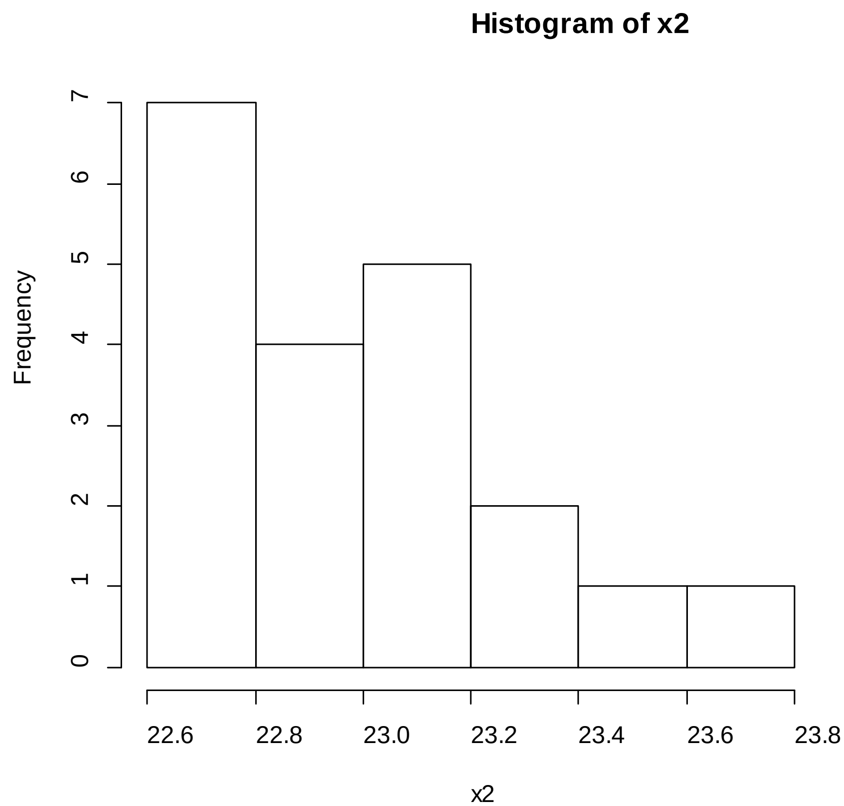
Solution:

```
>  x1<- x[1:20]
```

```
> x2<- x[21:40]
> hist(x1)
> stem(x1)

  The decimal point is at the |

  24 | 26467
  26 | 53
  28 | 792334677
  30 | 0789
```




```
> stem(x2)
  The decimal point is at the |
  22 | 77777789
  23 | 0001112234
  23 | 57

> hist(x2)
```


Histogram of x2
Frequency
0
1
2
3
4
5
6
7
22.6
22.8
23.0
23.2
23.4
23.6
23.8
x2

Chapter 2 Probability Theory

Exercise 2-1

Suppose we flip a fair coin to obtain heads or tails. Define the sample space and the possible outcomes. Define events and the probabilities of each.

Solution:

Sample space U={heads, tails}, Event A ={side facing up is heads}, then P[A] = ½, or 1 out of two outcomes. Event B ={side facing up is tails}, then P[B] = ½, or 1 out of two outcomes.

Exercise 2-2

Define event *A*={rain today} with probability 0.2. Define the complement of event *A*. What is the probability of the complement?

Solution:

Complement is B={does not rain today}; P(B)=1-P(A)=1-0.2=0.8

Exercise 2-3

Define *A*={rains less than 1 inch } *B*={rains more than 0.5 inches}. What is the intersection event *C*?

Solution:

Event C={rains less than 1 inch and more than 0.5 inch} this is to say C={rain in between 0.5 and 1 inch}.

Exercise 2-4

A pixel of a remote sensing image can be classified as grassland, forest or residential. Define *A*={land cover is grassland} *B*={land cover is forest}. What is the union event *C*? What is *D*= the complement of *C*?

Solution:

Event C={land cover is grass or forest}, Event D={land cover is residential}

Exercise 2-5

Assume we flip a coin three times in sequence. The outcome of a toss is independent of the others. Calculate and enumerate the possible combinations and their probabilities.

Solution:

Possible outcomes $n=2^3=8$, Sample space U={HHH, HHT, HTH, HTT, THH, THT, TTH, TTT} Each outcome is equally likely with probability 1/8, obtained by $(½)^3$.

Exercise 2-6

Assume we take water samples from water wells to determine if the well is contaminated. Assume we sample four wells and that they are independent. Calculate the number and enumerate the possible events of contamination results. Calculate the number and enumerate those that would have exactly two contaminated wells in the four trials.

Solution:

$n=2^4=16$, Sample space U={NNNN,CNNN,NCNN, etc} where C=contaminated, N= not contaminated. Of these $\binom{4}{2}=6$ include exactly two contaminated, these are

{CCNN,CNCN,CNNC,NCCN,NCNC,NNCC}

Exercise 2-7

Using the tree of Figure 2-8 What is the total probability of the test is in error? Hint: *BD* **or** *AC*. What is the probability that the test is correct?

Solution:

P[BD]= 0.056, P[AC]= 0.006

Test is in error: P[BD]+P[AC]=0.056+0.006=0.062

Test is correct 1-(P[BD]+P[AC])=1-0.062=0.938 (could also sum P(AD)+P(BC))

Exercise 2-8

Using Figure 2-8 and Bayes' theorem: what is the probability that the water is contaminated given a positive test result? Hint: calculate *P*[*A*|*D*].

Solution:

$$P[A\,|\,D]=\frac{P[AD]}{P[D]}=\frac{P[D\,|\,A]P[A]}{P[D\,|\,A]P[A]+P[D\,|\,B]P[B]}$$

P(D) = 0.2(1-0.03) + 0.8(0.07) = 0.25

P(A|D) = 0.2(0.97)/(0.25) = 0.776

Exercise 2-9

Assume 20% of an area is grassland. We have a remote sensing image of the area. An image classification method yields correct grass class with probability=0.9 and correct non-grass class with probability=0.9. What is the probability that the true vegetation of a pixel classified as grass is grass? Repeat assuming that grasslands is 50% of the area? Which one is higher and why?

Solution:

Apply Bayes Theorem as above. Probability of grass P(G)=0.2 then probability of non grass P(NG)=0.8. Probability of grass class given that is grass is P(g|G)=0.9 so P(ng|G)=0.1. Probability of non grass class given it is non grass is P(ng|NG)=0.9 so P(g|NG)=0.1.

We want P(G|g)

$$P[G\,|\,g] = \frac{P[Gg]}{P[g]} = \frac{P[g\,|\,G]P[G]}{P[g\,|\,G]P[G] + P[g\,|\,NG]P[NG]} =$$
$$\frac{0.9\times 0.2}{0.9\times 0.2 + 0.1\times 0.8} = \frac{0.18}{0.18+0.08} = \frac{0.18}{0.26} = 0.69$$

Now if P(G)=0.5

$$P[G\,|\,g] = \frac{0.9\times 0.5}{0.9\times 0.5 + 0.1\times 0.5} = \frac{0.45}{0.45+0.05} = \frac{0.45}{0.5} = 0.9$$

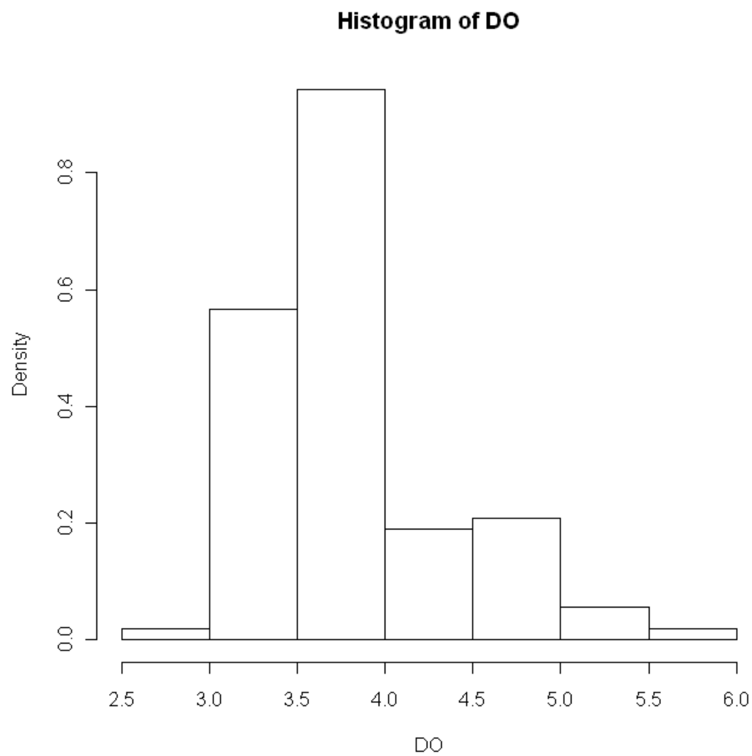
The result is higher for P(G)=0.5. This makes sense because higher P(G) increases P(Gg).

Exercise 2-10

Plot a histogram in probability density scale for DO variable of the x object from datasonde.csv. Save the graph as a jpeg file. Insert to an application.

Solution:

```
hist(DO,prob=T)
```



Exercise 2-11

Read file lab2/lake-lewisville.csv to a data frame. Use both Rcmdr and Rconsole.

Solution:

```
x <- read.table("lab2/lake-lewisville.csv",header=T,sep=",")
> x
       Date     Time Temp SpCond   TDS Salinity DOsat    DO Depth   pH Turbid IBatt
1  1/1/2010  0:00:00 7.59  328.3 213.4     0.16 109.2 13.06 0.826 8.59    6.8  10.7
2  1/1/2010  0:30:00 7.59  328.3 213.4     0.16 109.7 13.12 0.829 8.59    6.5  10.7
3  1/1/2010  1:00:00 7.57  328.2 213.3     0.16 109.3 13.07 0.830 8.59    6.4  10.8
4  1/1/2010  1:30:00 7.55  328.2 213.3     0.16 109.3 13.07 0.831 8.59    6.5  10.8
5  1/1/2010  2:00:00 7.55  328.2 213.3     0.16 109.0 13.04 0.828 8.60    6.7  10.8
6  1/1/2010  2:30:00 7.51  328.3 213.4     0.16 109.0 13.05 0.829 8.59    6.7  10.7
7  1/1/2010  3:00:00 7.53  328.0 213.2     0.16 109.0 13.05 0.831 8.59    6.7  10.8
8  1/1/2010  3:30:00 7.50  328.2 213.3     0.16 108.9 13.04 0.831 8.59    6.8  10.8
9  1/1/2010  4:00:00 7.50  328.2 213.3     0.16 108.7 13.02 0.824 8.59    6.3  10.8
10 1/1/2010  4:30:00 7.50  328.3 213.4     0.16 108.6 13.00 0.827 8.59    6.2  10.7
➢ etc
```

Exercise 2-12

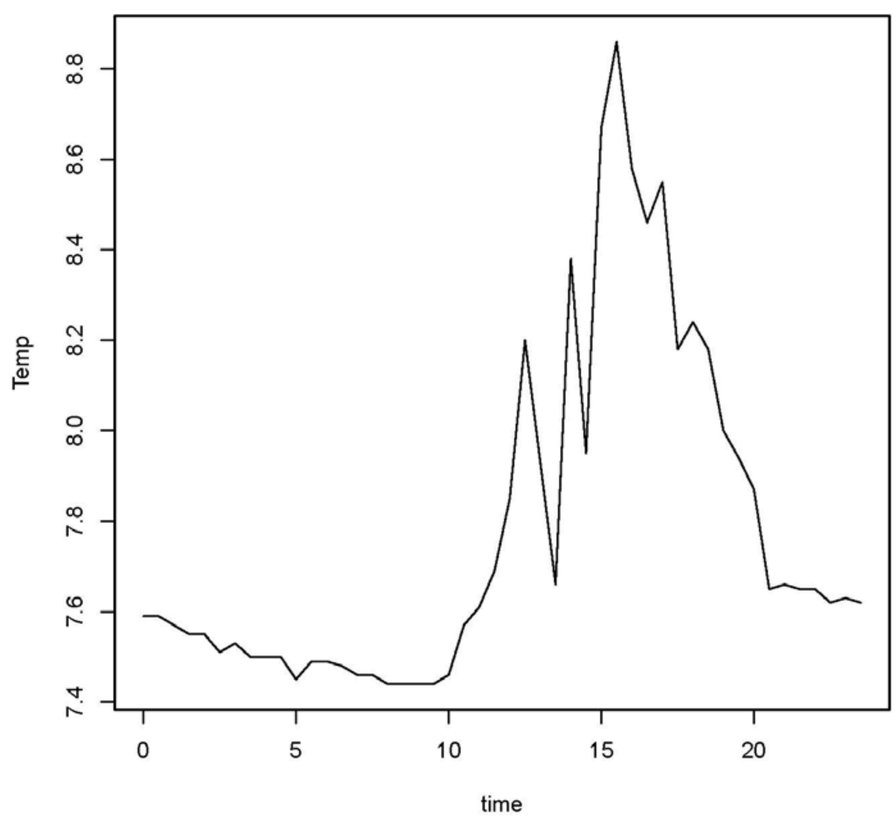
Plot variables of data frame created in exercise 2-11.

Solution:

Time should be converted to a sequence of real numbers from hour 0 to hour 23.5. It is convenient to write a loop and plot each variable.

```
attach(x)
time <- seq(0,23.5,0.5)
pdf(file="lab2/lakelewisville.pdf")
for(i in 3:12)
plot(time,x[,i], type="l", col=1,ylab=names(x)[i])
dev.off()
```

The PDF contains one page per variable. For example

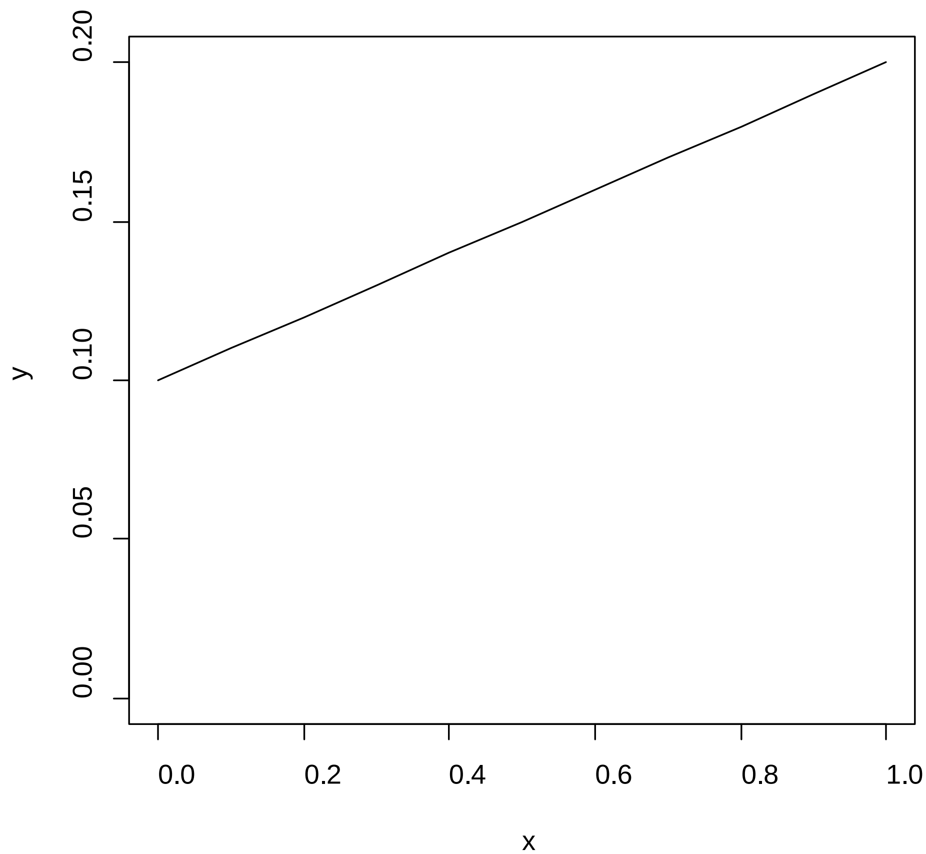


Exercise 2-13

Generate a linear function $y = ax + b$. Using a=0.1, b=0.1. Plot y for values of x in 0 to 1. Limit y-axis to go from 0 to the maximum of y.

Solution:

```
> a=0.1;b=0.1; x=seq(0,1,0.1)
> y <- a*x+b; plot(x,y,type="l",ylim=c(0,max(y)))
```



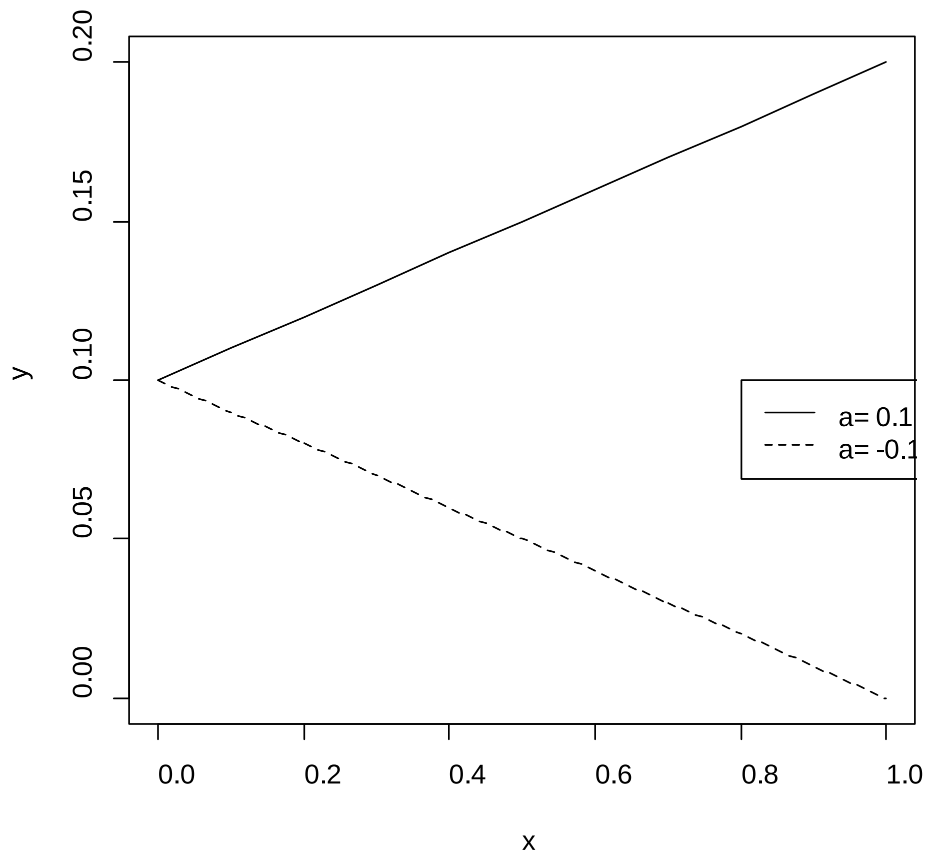
Exercise 2-14

Generate a linear function $y = ax + b$ Using b=0.1 and two values of a, a=0.1 and a=-0.1

Plot y for values of x in the interval [0,1]. Limit the y-axis to the interval [minimum of y, maximum of y]. Place a legend.

Solution:

```
a=c(0.1,-0.1); b=0.1; x=seq(0,1,0.1)
y <- matrix(nrow=length(x), ncol=length(a))
for(i in 1:2) y[,i] <- a[i]*x+b
matplot(x,y,type="l",ylim=c(min(y),max(y)), col=1)
legend(0.8,b,paste("a=",as.character(a)), lty=c(1:length(a)))
```



Exercise 2-15

This exercise refers to the Bayes' rule script. Change probability of contamination $P[A]$ to 0.3. Plot the probability of contamination given that a test is negative $P[A|C]$ vs. false negative error with false positive error as a parameter. Hint: modify the script given above for Bayes' rule to reverse the roles of Fneg and Fpos.

Solution:

```
# pA =contamination p[A]
```

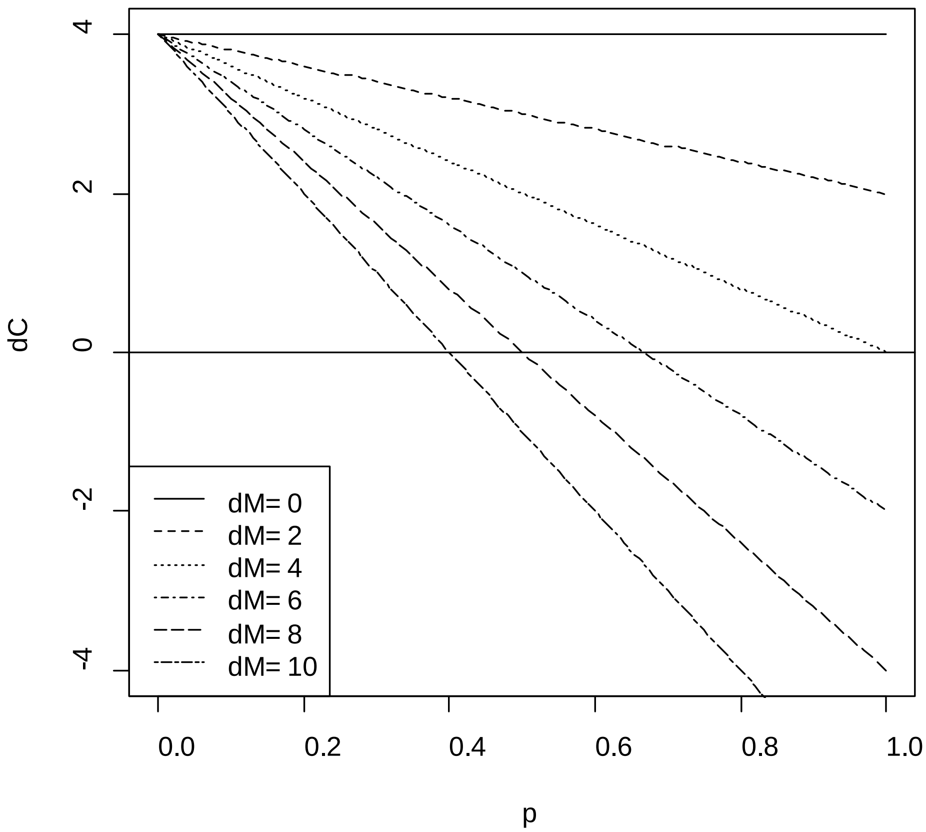
```
# Fneg = false negative p[C|A]
# Fpos = false positive p[D|B]
# fix pA and explore changes of p[A|C]
# as we vary Fpos and Fneg
# fix pA
pA=0.3
# sequence of values
Fneg <- seq(0,1,0.05); Fpos <- seq(0,1,0.2)
# array to store results
Cont.neg <- matrix(nrow=length(Fneg),ncol=length(Fpos))
# Bayes theorem
for(i in 1:length(Fpos))
Cont.neg[,i] <- Fneg*pA/(Fneg*pA + (1-Fpos[i])*(1-pA))
# plot
matplot(Fneg,Cont.neg, type="l",lty=1:length(Fpos), col=1,
xlab="False Negative Error", ylab="Prob(Contaminated | test negative)")
legend(0,1, paste("Fpos=",as.character(Fpos)), lty=1:length(Fpos), col=1)
```

Exercise 2-16

On the decision making script. Change ΔI to 4 and plot again. Discuss the changes obtained for the values of p at which we would decide for alternative A_1.

Solution:

```
# fix delta I
dI <- 4
# sequences for delta M and p
dM <- seq(0,10,2); nM <- length(dM)
p <- seq(0,1,0.01); np <- length(p)
# prepare a 2D array to store results
C <- matrix(nrow=np, ncol=nM)
# loop to calculate C for various dM
for(i in 1:nM) C[,i] <- dI-dM[i]*p
# plot the family of lines
matplot(p,C,type="l",lty=1:nM,col=1,ylim=c(−dI,dI))
# draw horizontal line at 0 to visualize crossover
abline(h=0)
# legend to identify the lines, use a keyword to position it
legend("bottomleft",leg=paste("dM=",dM),lty=1:nM,col=1)
```



The values of p have increased by a factor of 2.