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)

$$P[G \mid g] = \frac{P[Gg]}{P[g]} = \frac{P[g \mid G]P[G]}{P[g \mid G]P[G] + P[g \mid NG]P[NG]} =$$

$$\frac{0.9 \times 0.2}{0.9 \times 0.2 + 0.1 \times 0.8} = \frac{0.18}{0.18 + 0.08} = \frac{0.18}{0.26} = 0.69$$

Now if P(G)=0.5

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

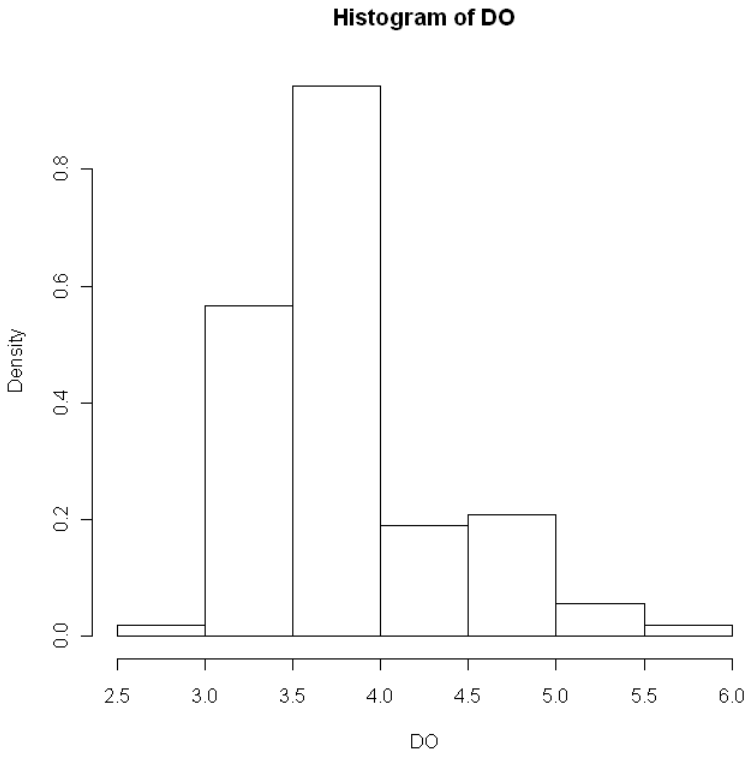
The result is higher for P(G)=0.5. This makes sense because higher P(G) increases P(Gg).

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

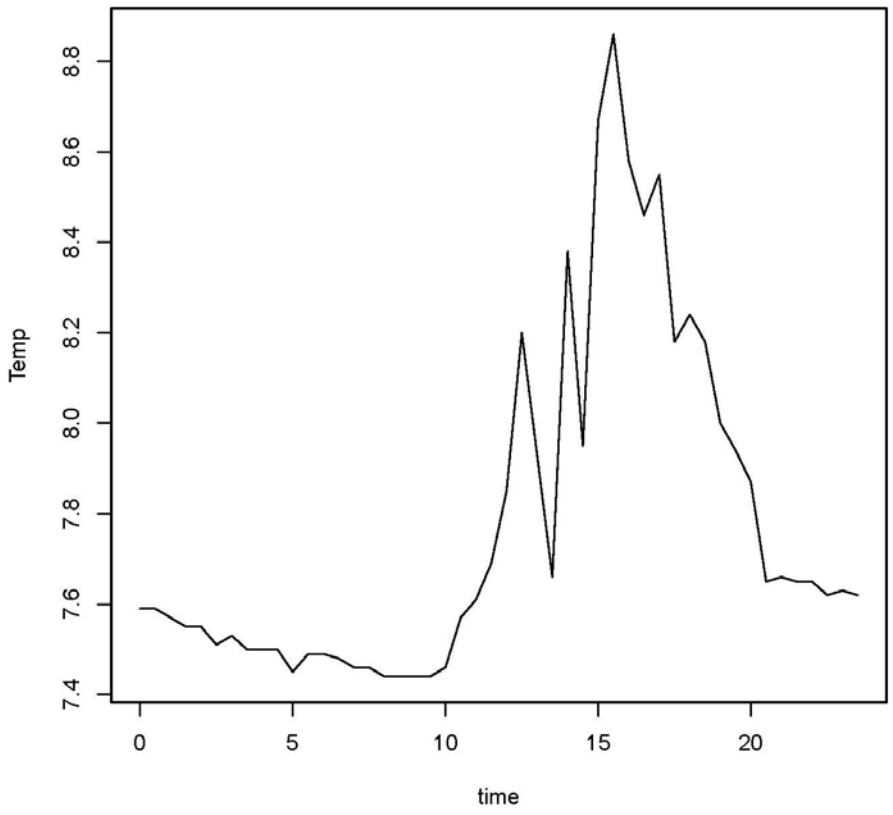
Plot variables of data frame created in exercise 2-11.

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

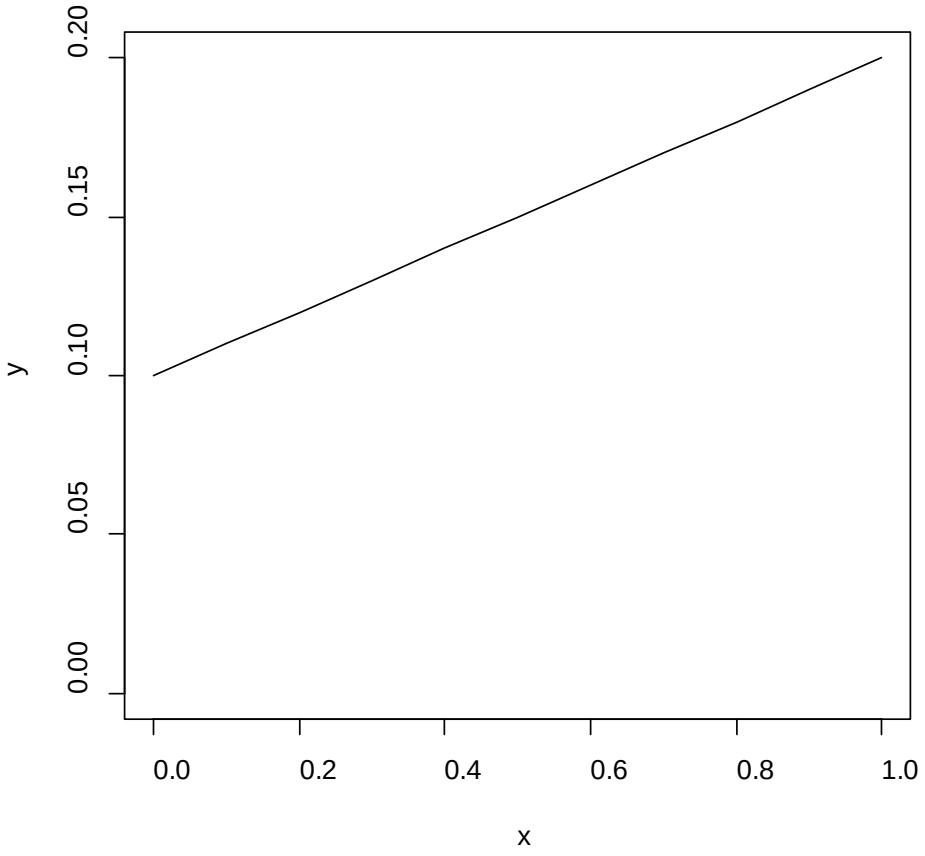


Exercise 2-13

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)))
```



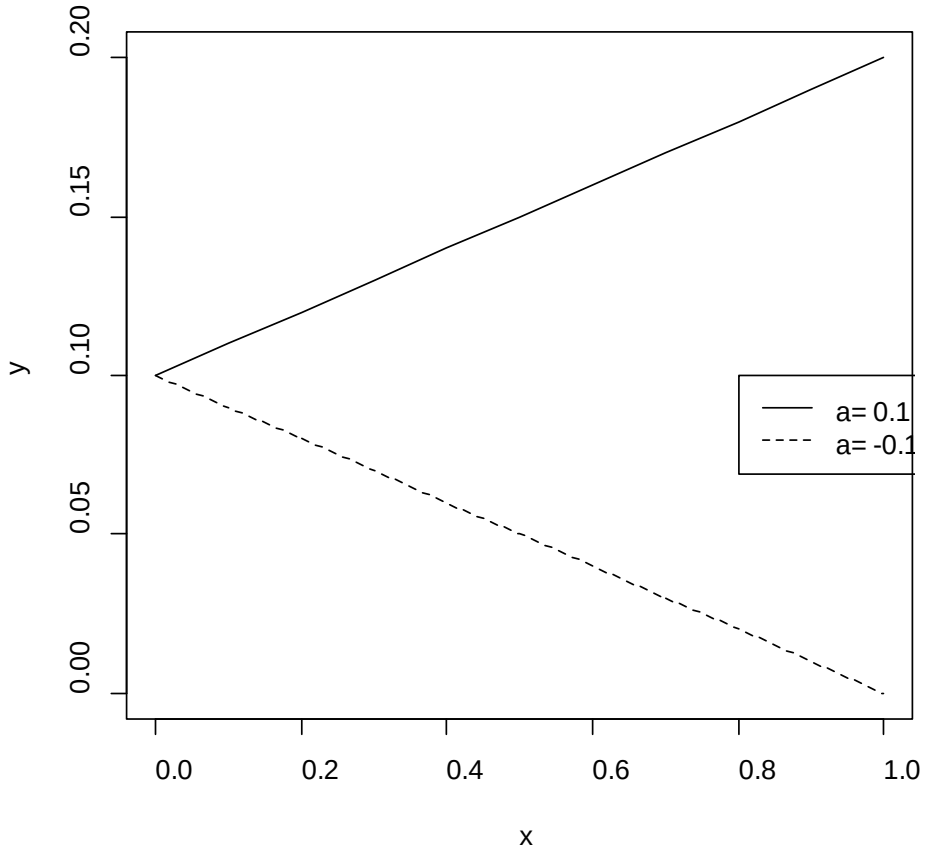
Exercise 2-14

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]
```

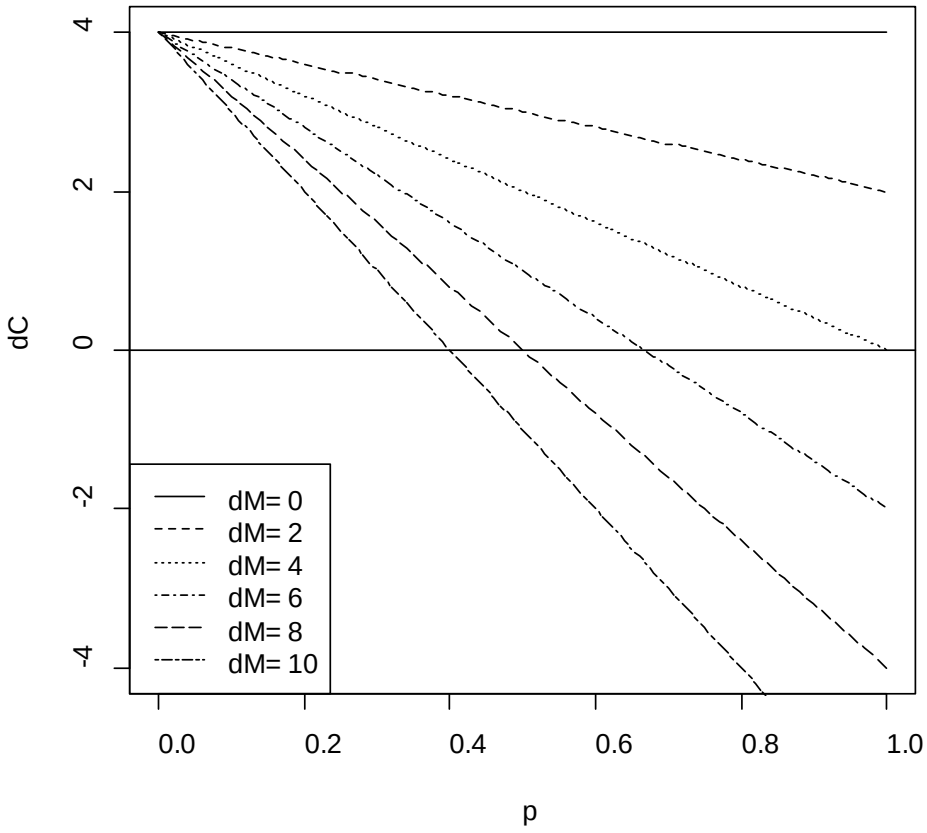
```
# Fneg = false negative p[C|A]
# Fpos = false positive p[D|B]
# fix pA and explore changes of p[A|C]
# as we vary Fpos and Fneg
# fix pA
pA=0.3
# sequence of values
Fneg <- seq(0,1,0.05); Fpos <- seq(0,1,0.2)
# array to store results
Cont.neg <- matrix(nrow=length(Fneg),ncol=length(Fpos))
# Bayes 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.