quantBioiHW1

Carlee Bettler

2/1/2019

2.a. These are the 16 probabilities in table form

0.3\*0.3

## [1] 0.09

0.2\*0.2

## [1] 0.04

0.3\*0.2

## [1] 0.06

#Math for probabilities   
  
  
Q1table <- matrix(c("0.04","0.06","0.06","0.04","0.06","0.09","0.09","0.06","0.06","0.09","0.09","0.06","0.04","0.06","0.06","0.04"),ncol=4,byrow=TRUE)  
colnames(Q1table) <- c("A","C", "G","T")  
rownames(Q1table) <- c("A","C","G","T")  
Q1table <- as.table(Q1table)  
Q1table

## A C G T   
## A 0.04 0.06 0.06 0.04  
## C 0.06 0.09 0.09 0.06  
## G 0.06 0.09 0.09 0.06  
## T 0.04 0.06 0.06 0.04

#Table of probabilities of base combinations

2.b.i P(C)=0.3

P(T)=0.2

answer:P(E)= 0.3+0.2 = 0.5

2.b.ii

P(A)=0.2

P(C)=0.3

P(T)= 0.2

answer:P(F) = 0.2+0.3+0.2 = 07

2.b.iii.

P(E∪F) = P(E)+P(F)-P(E∩F)

P(E)= 0.5

P(F)= 0.7

answer: P(E∪F) = 0.85

.5+.7 -(.5\*.7)

## [1] 0.85

2.b.iv.

P(E∩F) = P(E)\*P(F)

P(E)= 0.5

P(F)= 0.7

answer: P(E∩F) = 0.35

.5\*.7

## [1] 0.35

2.b.v. P(F^c) = 1-P(F)

P(F)= 0.7

answer: P(F^c) = 1-0.7 = 0.3

2.c.i. P(G|E)= P(G∩E)/P(E)

P(G) = 2/16 = 1/8 = 0.125, because it’s a set covering 2/16 nucleotide combinations

P(E)= 0.5

P(G∩E) = 0.0625

answer: P(G|E)= 0.125

1/8

## [1] 0.125

#P(G) decimal  
  
0.5\*.125

## [1] 0.0625

#P(GintersectionE)   
  
#P(G|E)  
0.0625/0.5

## [1] 0.125

2.c.ii.

P(F|G∪E) = (P(F)∩P(G∪E))/P(G∪E)

answer: P(F|G∪E) = 0.7

#P(GunionE):  
.125+0.5-(.125\*.5)

## [1] 0.5625

(0.5625\*.7)/0.5625

## [1] 0.7

#(P(F)intersectionP(GunionE))/P(GunionE)

2.c.iii.

P(F∪G|E)

P(G|E) = P(G)∪P(E)/P(E)

answer: P(F∪G|E) = 0.7375

.7+.125-(.7\*.125)

## [1] 0.7375

#P(FunionG)   
  
0.7375\*.5/.5

## [1] 0.7375

#P(FunionG|E)

3.a.i.

Since I am trying to find an equal to probability, I will use dbinom(x,n,p)

x=0, n=10, p=0.3

answer: P(N=0) = 0.02824752

dbinom(0,10,0.3)

## [1] 0.02824752

3.a.ii. Since I am trying to find a less than or equal to probability, I will use pbinom(x,n,p)

x=0, n=10, p=0.3

answer: P(N<=3) = 0.6496107

pbinom(3,10,0.3)

## [1] 0.6496107

3.a.iii.

E[n]= n\*p (theoretical mean equation)

n = 10

p = 0.3

answer: E[N]= 3

10\*0.3

## [1] 3

3.a.iv.

Var[N]= n*p*(1-p) (theoretical variance equation)

n=10, p=0.3

answer: Var[N] = 2.1

10\*0.3\*(1-0.3)

## [1] 2.1

3.b. Sample mean = 3.0041

Sample variance = 2.052488

P(N <= 3) = 0.6489

P(N = 0) = 0.027

set.seed(1234)  
x <- rbinom(10000,10,0.3)  
#Simulating 10000 observations with n = 10 and p= 0.3  
  
mean(x)

## [1] 3.0041

#Sample mean   
  
var(x)

## [1] 2.052488

#Sample variance  
  
sum(x<=3)/10000

## [1] 0.6489

#P(N<=3)  
  
sum(x==0)/10000

## [1] 0.027

#P(N = 0)

In comparing the theoretical values with my real values, we see that with a large number of data points (in this case 10000) the sample values are close to and approaching the theoretical values. For example, the theoretical mean is 3, while the sample mean closely approaches 3 at 3.0041, but is not exactly 3.

4.a.

a expected frequency = 246.5

t expected frequency= 246.5

c expected frequency = 253.5

g expected frequency = 253.5

a observed frequency = 266

t observed frequency = 227

c observed frequency = 260

g observed frequency = 247

(.507/2)\*1000

## [1] 253.5

#C/G expected probabilities   
((1-.507)/2)\*1000

## [1] 246.5

#A/T expected probabilites   
  
library(ade4)  
library(seqinr)  
x<-read.fasta('/Users/carleebettler/Downloads/Ecoli\_Data/AE005174v2.fas')   
ecoli<-c(x[[1]],x[[2]])  
obsFr<-table(ecoli[1:1000])[c("a","c","t","g")]  
obsFr

##   
## a c t g   
## 266 260 227 247

4.b. Ho:%(G+C) = 50.7%

H1:%(G+C) != 50.7%

The p-value = 1, which is greater than an alpha value of 0.05. Therefore we fail to reject the null hypothesis.

library(ade4)  
library(seqinr)  
x<-read.fasta('/Users/carleebettler/Downloads/Ecoli\_Data/AE005174v2.fas')   
  
ecoli<-c(x[[1]],x[[2]])  
  
1-.507

## [1] 0.493

p <- c(0.507,0.493)  
#This gives you the combined probabilities of C+G, and A+T  
  
obsFr<-table(ecoli[1:1000])[c("a","c","t","g")]  
  
obsFr

##   
## a c t g   
## 266 260 227 247

(obsFr[2]+obsFr[4])\*2

## c   
## 1014

#Doubled c+g percentages, doubled to include both strands  
  
(obsFr[1]+obsFr[3])\*2

## a   
## 986

#Doubled a+t percentages, doubled to include both strands  
  
  
obsFr2 <- c(1014,986)  
  
chisq.test(obsFr2,p=p)

##   
## Chi-squared test for given probabilities  
##   
## data: obsFr2  
## X-squared = 0, df = 1, p-value = 1

#Does chi squared goodness of fit test

Question 5.a. Below are the marginal probabilities for various X and Y values

X=2 X=3 X=7

0.44 0.32 0.24

Y=1 Y=3 Y=6 Y=9

0.31 0.12 0.30 0.27

#add probabilities to get to get z = 18   
  
data<-matrix(c(0.11,0.20,0,0.05,0.02,0.05,0.20,0,0.1,0.08,0.1,0.09),nrow=3)  
row.names(data)<-c("X=2","X=3","X=7")  
colnames(data)<-c("Y=1","Y=3","Y=6","Y=9")  
  
x <- rowSums(data)  
#Sums up the rows  
y <- colSums(data)  
#Sums up the columns   
  
x

## X=2 X=3 X=7   
## 0.44 0.32 0.24

y

## Y=1 Y=3 Y=6 Y=9   
## 0.31 0.12 0.30 0.27

5.b. This is my table of the probability distribution Z=XY

Z P(Z)

2 0.11

3 0.20

6 0.05

7 0.00

9 0.02

12 0.2

18 0.08

21 0.05

27 0.1

42 0.1

63 0.09

Note that the only P(Z) value we have to add for was the P(Z=18)

data<-matrix(c(0.11,0.20,0,0.05,0.02,0.05,0.20,0,0.1,0.08,0.1,0.09),nrow=3)  
row.names(data)<-c("X=2","X=3","X=7")  
colnames(data)<-c("Y=1","Y=3","Y=6","Y=9")  
data

## Y=1 Y=3 Y=6 Y=9  
## X=2 0.11 0.05 0.2 0.08  
## X=3 0.20 0.02 0.0 0.10  
## X=7 0.00 0.05 0.1 0.09

#This is my original joint probability table for refernce   
  
yvals <- c(1,3,6,9)  
rbind(yvals\*2, yvals\*3, yvals\*7)

## [,1] [,2] [,3] [,4]  
## [1,] 2 6 12 18  
## [2,] 3 9 18 27  
## [3,] 7 21 42 63

#These are my XY values for reference   
  
  
Q4table <- matrix(c("2","3","6","7","9","12","18","21","27","42","63","0.11","0.20","0.05","0.00","0.02","0.2","0.08","0.05","0.1","0.1","0.09"),ncol=2,byrow=FALSE)  
colnames(Q4table) <- c("Z","P(Z)")  
rownames(Q4table) <- c("","","","","","","","","","","")  
Q4table <- as.table(Q4table)  
Q4table

## Z P(Z)  
## 2 0.11  
## 3 0.20  
## 6 0.05  
## 7 0.00  
## 9 0.02  
## 12 0.2   
## 18 0.08  
## 21 0.05  
## 27 0.1   
## 42 0.1   
## 63 0.09

#This creates my table of the probability distribution Z=XY

5.c.

data<-matrix(c(0.11,0.20,0,0.05,0.02,0.05,0.20,0,0.1,0.08,0.1,0.09),nrow=3)  
row.names(data)<-c("X=2","X=3","X=7")  
colnames(data)<-c("Y=1","Y=3","Y=6","Y=9")  
  
#cov(X,Y) - Covariance matrix  
cov(data)

## Y=1 Y=3 Y=6 Y=9  
## Y=1 0.01003333 -0.00145 -0.0045 0.00045  
## Y=3 -0.00145000 0.00030 0.0015 -0.00015  
## Y=6 -0.00450000 0.00150 0.0100 -0.00100  
## Y=9 0.00045000 -0.00015 -0.0010 0.00010

#cor(X,Y) - Correlation matrix   
cor(data)

## Y=1 Y=3 Y=6 Y=9  
## Y=1 1.0000000 -0.8357661 -0.4492519 0.4492519  
## Y=3 -0.8357661 1.0000000 0.8660254 -0.8660254  
## Y=6 -0.4492519 0.8660254 1.0000000 -1.0000000  
## Y=9 0.4492519 -0.8660254 -1.0000000 1.0000000