

BIOINFORMATICS MSc PROBABILITY AND STATISTICS

SPLUS EXERCISE SHEET 2

This exercise is in two parts; the first is aimed at practising and extending programming skills in the context of extreme value distribution calculations, the second part gives an example of data summary methods.

A script relating to this exercise set can be downloaded from:

<http://stats.ma.ic.ac.uk/~das01/BioinformaticsMSc/SPLUS/sheet02.ssc>

1. ORDER STATISTICS AND APPROXIMATIONS

A result from lectures shows that the maximum and minimum order statistics derived from independent and identically distributed samples have distributions that can be identified exactly. Minima and maxima are important random variables in sequence alignment problems

If $Y_1 = X_{(1)} = \min \{X_1, \dots, X_n\}$ and $Y_n = X_{(n)} = \max \{X_1, \dots, X_n\}$, then

$$\begin{aligned} F_{Y_n}(y_n) &= \{F_X(y_n)\}^n & f_{Y_n}(y_n) &= n \{F_X(y_n)\}^{n-1} \\ F_{Y_1}(y_1) &= 1 - \{1 - F_X(y_1)\}^n & f_{Y_1}(y_1) &= n \{1 - F_X(y_1)\}^{n-1} f_X(y_1) \end{aligned}$$

We can study such random variables using **SPLUS** .

First suppose that $X_1, \dots, X_n \sim Exponential(\lambda)$ say with $\lambda = 3$, where

$$f_X(x) = \lambda e^{-\lambda x} \quad F_X(x) = 1 - e^{-\lambda x} \quad x > 0$$

We should have that

$$F_{Y_1}(y_1) = 1 - \{1 - F_X(y_1)\}^n = 1 - e^{-n\lambda y_1}$$

and

$$F_{Y_n}(y_n) = \{F_X(y_n)\}^n = \{1 - e^{-\lambda y_n}\}^n$$

The following code implements some calculations related to the distributions of maxima and minima. Find out what each of the following steps is computing:

```

>n <- 500
>lambda <- 3
>nits <- 5000
>xmin <- rep(0,nits)
>xmax <- rep(0,nits)
>for(i in 1:nits){
+ x <- rexp(n,lambda)
+ xmin[i] <- min(x)
+ xmax[i] <- max(x)
+ }
>hist(xmin)
>hist(xmax)
>length(xmin[xmin < 0.0001])/nits
>1-exp(-n*lambda*0.0001)
>length(xmin[xmin > 0.0005])/nits
>exp(-n*lambda*0.0005)
>length(xmax[xmax <= 2])/nits
>(1-exp(-lambda*2))^n
>length(xmax[xmax > 3])/nits
>1-(1-exp(-lambda*3))^n

```

EXERCISE Find a sequence of commands (or write a function) that allows the cdf and/or pdf for the two extreme order statistics to be plotted. Use something like the following construction:

```

>x <- c(0:5000)/500
>cdf.x <- function(x,lambda){1-exp(-lambda*x)}
>cdf.xmin <- function(x,lambda,n){1-(1-cdf.x(x,lambda))^n}
>cdf.xmax <- function(x,lambda,n){(cdf.x(x,lambda))^n}

```

Maximum Order statistics from *Geometric*(θ) random variables play a particular role in pattern detection and sequence alignment for biological sequences. Using the **SPLUS** package we can simulate some suitable values for $\theta = 0.25$. The Geometric distribution in **SPLUS** has mass function and cdf

$$f_X(x) = (1 - \theta)^x \theta = (1 - \phi) \phi^x \quad F_X(x) = 1 - (1 - \theta)^{x+1} = 1 - \phi^{x+1} \quad x = 0, 1, 2, \dots$$

where $\phi = 1 - \theta$. For the simulation, proceed as follows:

```

>n <- 500
>theta <- 0.25
>nits <- 5000
>xmin <- rep(0,nits)
>xmax <- rep(0,nits)
>for(i in 1:nits){
+ x <- rgeom(n,theta)
+ xmin[i] <- min(x)
+ xmax[i] <- max(x)
+ }
>hist(xmin)
>hist(xmax)

```

EXERCISES

(i) Study the distribution for different values of θ and n .

(ii) The theory of extreme order statistics says that if $X_1, \dots, X_n \sim \text{Geometric}(\theta) \equiv \text{Geometric}(1 - \phi)$ then the maximum order statistic $Y_n = X_{(n)} = \max\{X_1, \dots, X_n\}$ has cdf given by

$$F_{Y_n}(y_n) = \{F_X(y_n)\}^n = \{1 - \phi^{y_n+1}\}^n$$

and hence that

$$P[Y_n < y_n] = F_{Y_n}(y_n - 1) = \{1 - \phi^{y_n}\}^n$$

which can be approximated (for large n) by

$$P[Y_n \geq y_n] = 1 - F_{Y_n}(y_n - 1) = 1 - \{1 - \phi^{y_n}\}^n \approx 1 - \exp\{-n\phi^{y_n}\}$$

that gives an approximate method of computing a p -value in a runs test. Study the validity of this approximation for different values of ϕ and n

Extreme Order statistics from *Uniform*(0,1) random variables play a role in pattern detection; in particular we again study the distributions of the minimum and maximum order statistics because we wish to use them as test statistics in a hypothesis test of the uniformity of the sample. We can do this as follows:

```
>n <- 500
>nits <- 5000
>umin <- rep(0,nits)
>umax <- rep(0,nits)
>for(i in 1:nits){
+ u <- runif(n)
+ y <- c(0,sort(u),1)
+ udiff <- diff(y)
+ umin[i] <- min(udiff)
+ umax[i] <- max(udiff)
+ }
>hist(umin)
>hist(umax)
```

EXERCISE Study the behaviour for different values of n

3. DATA SUMMARY AND ANALYSIS

The course website contains two data files containing protein sequences relating to a comparison of 160 transmembrane proteins and 645 general proteins; the source is

<http://www.cbs.dtu.dk/~krogh/TMHMM/>

The two files are

<http://stats.ma.ic.ac.uk/~das01/BioinformaticsMSc/Data/TMPproteinlengthdata.txt>

<http://stats.ma.ic.ac.uk/~das01/BioinformaticsMSc/Data/proteinlengthdata.txt>

respectively.

Our objective is to examine any differences between the composition and general structure of transmembrane proteins and a random sample of other proteins. The SPLUS script file contains a series of commands for reading in and processing the data.

EXERCISE: Investigate the differences between the transmembrane proteins and the randomly sampled proteins in terms of

- (i) protein lengths
- (ii) protein compositions by amino acid residue.(use a loop, and the **substring** command to analyze each protein sequence in detail)

Your objective is to find any differences between the two classes of proteins on the basis of the data available.

Use commands such as

```
hist
mean
var
quantile
```

to implement the statistical comparison