MA20226: STATISTICS 2A
http://people.bath.ac.uk/masss/ma20226.html
2011/12 Practical Sheet 1

Aim: In Section 1 of this practical, we introduce the R function `curve` which is used to draw functions of x. In Section 2 we introduce the use of matrices in R and illustrate how these can be used to help analyse repeated random samples, in this case drawn from the uniform distribution. We also introduce the function `apply` to perform operations on either the rows or columns of a matrix. This section should provide you with insight into the theoretical results for the sampling distribution, mean and variance of the maximum likelihood estimator of θ when $X_1, \ldots, X_n$ are iid $U(0, \theta)$ whilst also illustrating properties of the sample mean and variance of a series of iid observations.

Logging on to R: Bring up the Start menu, click on All Programs, then click on More Applications and go to GigaTerms. You will be asked to log on again and then follow the route Start - Programs - R where you will find R2.13.0.

Quitting R: type `q()` to quit R.

A note on R: If you wish to refresh your memory of R then the brief introduction to R sheet you received when studying MA10212 (or MA10032) is available at

http://people.bath.ac.uk/masss/ma20226/introR.pdf

This guide also reminds you of the useful R functions `help`, `help.start` and `fix`. R is freely available, and may be downloaded from the R homepage

http://www.r-project.org/

Due in: This practical does NOT form part of the assessed coursework. You should hand your solutions to this practical in at the tutorial session next week, Week 4.

1 Curve sketching using `curve`

The R function `curve` draws a curve correponding to a given expression written as a function of x, or alternatively the name of a function which will be plotted. Use `help(curve)` to explore the full description of `curve` and its options. For example, to sketch $\cos x$ for $-2\pi \leq x \leq 2\pi$ we use

```
> curve(cos, -2*pi, 2*pi,main="A plot of cos x")
```

whilst

```
> curve(3*x^3-5*x+1,xlim=c(-5,5),ylab="3x^3-5x+1",main="A plot of 3x^3-5x+1")
```

plots $3x^3 - 5x + 1$ for $-5 \leq x \leq 5$. The option `add=T` will add the curve to an already existing plot. Thus,

```
> curve(cos, -2*pi, 2*pi, ylab="f(x)",main="A plot of cos x and sin x")
> curve(sin,add=T)
```

will first plot $\cos x$ for $-2\pi \leq x \leq 2\pi$ and then add $\sin x$ to the plot.

2 Sampling from the Uniform distribution and operations on a matrix

If $X_1, \ldots, X_n$ are iid $U(0, \theta)$ where θ is unknown, then the maximum likelihood estimator of θ is $M = \max\{X_1, \ldots, X_n\}$ and the sampling distribution of M is

$$f_M(m \,|\, \theta) \quad = \quad \begin{cases} n\frac{m^{n-1}}{\theta^n} & m \leq \theta, \\ 0 & \text{otherwise.} \end{cases}$$

In Lecture 4 we noted that M is a biased estimator of θ. It is straightforward to show that (see Lecture 5) that

$$E(M \,|\, \theta) \;=\; \int_0^\theta m\left(n\frac{m^{n-1}}{\theta^n}\right) dm \;=\; \left(1 - \frac{1}{n+1}\right)\theta \tag{1}$$

so that $b(M) = -\frac{1}{n+1}\theta$.

1. Show that

$$Var(M \,|\, \theta) \quad = \quad \frac{n\theta^2}{(n+1)^2(n+2)}.$$

   (NB. You may find it preferable to assume this result for now, proving it at home, and tackle the R based questions during the practical) [4]

We now consider how to utilise R to generate repeated random samples to illustrate these theoretical results. The techniques introduced can be easily adapted for sampling from any given distribution.

The function `runif` can be used to generate a sample from a uniform distribution with a specified maximum and minimum. The default is the standard uniform distribution with maximum value 1 and minimum value 0, the $U(0, 1)$. Thus, for example, to sample 5 values from a $U(0, 1)$ we have

```
> runif(5)
[1] 0.77426543 0.41015485 0.11602202 0.08005898 0.43879827
```

(of course, your data will be different!)

Suppose we want to take 50 random samples, each of size 5, from the $U(0, 1)$. One way to do this is to generate 250 random observations from the distribution and place the results in a 5×50 matrix. Then each column of the matrix may be viewed as holding one sample of size 5. We may use the `matrix` function to create such an object which we'll call `unifsamp`.

```
> unifsamp <- matrix(runif(250),nrow=5,ncol=50)
> unifsamp
           [,1]      [,2]      [,3]       [,4]      [,5]       [,6]       [,7]
[1,] 0.81838350 0.3731599 0.9055907 0.84079471 0.2732640 0.69096902 0.32643434
[2,] 0.86791855 0.1043879 0.5042760 0.47814578 0.2941667 0.08876166 0.79494633
[3,] 0.53846222 0.3837907 0.3455644 0.04515251 0.8069612 0.71371135 0.10188719
[4,] 0.30953203 0.1437689 0.3906411 0.06033681 0.2149284 0.66441605 0.07571246
[5,] 0.04111647 0.9363339 0.3725674 0.26688525 0.4563990 0.75064348 0.71648391
```

(and so on; we've truncated the data output.)

As with vectors, subscripting is used to access the components of the matrix. We provide two subscripts separated by a comma. The first subscript represents the row and the second the column. If the first subscript is left blank then you get all the rows whilst if the second is left empty you access all the columns. For example,

```
> unifsamp[,1]
[1] 0.81838350 0.86791855 0.53846222 0.30953203 0.04111647
```

produces a vector whose elements are the first column of `unifsamp`.

The function `apply` takes a matrix and then applies a function to each element in the stated dimension. For example,

```
> apply(unifsamp,2,max)
 [1] 0.8679185 0.9363339 0.9055907 0.8407947 0.8069612 0.7506435 0.7949463
 [8] 0.7443229 0.9808621 0.7018170 0.8961141 0.7951720 0.8081190 0.9132123
[15] 0.6142397 0.9400183 0.8729429 0.6014031 0.9390012 0.8201275 0.9323425
[22] 0.7503109 0.9879486 0.7240088 0.9861100 0.8777220 0.7597170 0.7577730
[29] 0.8761161 0.9996899 0.8762952 0.8819407 0.9400231 0.9089700 0.8275121
[36] 0.7888256 0.7793062 0.8392630 0.7279343 0.7903649 0.9650384 0.8867238
[43] 0.9639231 0.7722112 0.9691267 0.9693581 0.6848280 0.9148397 0.9734345
[50] 0.9114610
```

produces a vector whose elements are the maximum value of each of our 50 samples of size 5. The '2' indicates that the dimension is the columns. To perform operations on the rows of a matrix we use the dimension '1'.

2. Write a function, `randunif`, that allows you to take r random samples each of size n from a uniform distribution with maximum value maxi and minimum value mini. Your function should have four parameters (r, n, maxi, mini) and return a matrix whose ith column contains the ith sample of size n. Hand in a copy of your code. [3]

3. Now take 50 random samples of size 15 from a uniform distribution with maximum value 16 and minimum value 0. You should store the samples in a matrix, which you may like to call `unifsamp50`, for future use. Use R to answer the following questions.

   (a) Extract the 47th sample of size 15 and hand in a copy of this sample. What is the maximum of this sample? What value did you expect to see and why? [2]

   (b) Use the function `apply` to create a vector, which you may wish to call `obsmax50`, which contains the maximum value of each of the 50 samples. Thus, `obsmax50` contains 50 observations of $M = \max\{X_1, \ldots, X_{15}\}$ where the X_i are iid $U(0, 16)$. Use the R function `hist` to plot a histogram of the 50 maximums. Your histogram should have a total area of one. Using `curve` overlay your histogram with the corresponding sampling distribution of M (see equation (1)). Hand in a copy of your histogram with overlaid sampling distribution and briefly comment upon any conclusions you draw from your plot. [4]

   (c) Use the function `mean` to calculate the sample mean of the 50 maximum values held in the vector `obsmax50`. What value, and why, do you expect to see as the mean of these 50 observations? What value do you observe? [2]

(d) Use the function `var` to calculate the sample variance of the 50 maximum values held in the vector `obsmax50`. What value, and why, do you expect to see as the variance of these 50 observations? What value do you observe? [2]

[Hint: Use the help function of R to explore what `var` does. In particular, this may help you to determine whether it gives a biased or unbiased estimate of the variance of a series of iid observations.]

4. Now take 1000 random samples of size 15 from a uniform distribution with maximum value 16 and minimum value 0, storing the samples in a matrix which you may like to call `unifsamp1000`.

   (a) Repeat questions 3(b)-(d) using the 1000 samples contained in `unifsamp1000` rather than the 50 samples contained in `unifsamp50`. Comment on any differences you observe between the two samples. [6]

   (b) What is the statistical advantage of a larger sample size? [2]