STAT2011 Statistical Models

Semester 1, 2012

Computer Exercise Week 1

Welcome to the computer class for STAT2011. This week we will learn how to log onto the system and perform a basic session using the statistical computing environment R.

When you sit down to the terminal, it shows the Microsoft Windows login dialog box. We are using these machines to access the *Linux* network, specifically the server `rome`. Thus at the Windows login enter username `rome` and password `rome` to get to a rome/Linux login window. Your username on `rome` is your 8-character unikey and your 'MyUni' password will work as initial password.

Please run R in the xterm, rather than the Rtutorial application. Once the xterm is open, just type `R` and hit Enter.

We shall firstly execute a series of commands at the command line and observe the output. Then we shall collect the commands in a text file, and use this to create a report of the session in PDF format. Where commands are given below the prompt is not shown, to facilitate copying-and-pasting from the PDF version of this file.

Some tips:

- Please read this exercise through completely before starting.
- You can scroll backwards and forwards through your R command history in the xterm using the up- and down-arrows.
- Note you can copy-and-paste commands from the PDF version of this file into either the xterm or a text file.
- Do not print large objects on the screen, if you can help it. It makes it hard to find your commands later on.

The exercise starts over the page...

1. Define an object `N` equal to the integer `2012`:

   ```
   N=2012
   ```

   This `N` can be interpreted as a "population size".

2. Suppose we didn't know the value of `N`, but could take a simple random sample (i.e. sample without replacement so that each possible sample is equally likely) from $\{1, 2, \ldots, \texttt{N}\}$. How would we estimate `N`? There are (at least) two possible approaches, one using the sample mean, the other using the sample maximum.

   - As we shall see later, if we take such a sample of size n then the sample mean should be roughly equal to $(\texttt{N} + 1)/2$. Thus an estimate based on the sample mean $\bar{x}$ is $2\bar{x} - 1$.
   - An estimate based on the sample maximum $x_{\max}$ from a sample of size n is $[(n+1)/n]x_{\max} - 1$ (see Case Study 1.2.3 page 6 of the text).

3. Take a sample of size 100, and call it `s`:

   ```
   s=sample(1:N,size=100,replace=FALSE)
   ```

   The first argument here specifies the population; in R `1:N` is shorthand for `c(1,2,3,...,N)`. The last argument is not strictly necessary (the default is `FALSE`) but it is good practice to specify it anyway.

4. Obtain your two estimates:

   ```
   est.mean=(2*mean(s))-1
   ```

   Again the extra parentheses around `2*mean(s)` are optional but it is good practice to include them in such cases so long as it doesn't make it hard to read.

   ```
   est.max=(101/100)*max(s)-1
   ```

5. Display your two estimates:

```
est.mean
est.max
```

   Which is a better estimate? You can see by which is closer to 2012. However in reality you wouldn't know, so how would one assess the performance of these two procedures? One answer is to use *Monte Carlo simulation*.

6. We can simulate this experiment many many times to get an idea of how well the estimates perform. An easy way is to do a *for-loop*: copy-and-paste the lines below into the command line:

```
est.mean.vec=0
est.max.vec=0
for(i in 1:1000){
 s=sample(1:N,size=100,replace=FALSE)
 est.mean=(2*mean(s))-1
 est.max=(101/100)*max(s)-1
 est.mean.vec[i]=est.mean
 est.max.vec[i]=est.max
}
```

   This performs your little simulation 1000 times, obtaining an estimate each time and saving it in the appropriate vector.

7. Obtain means, sds and mean-squared errors of your vectors of estimates:

```
mean(est.mean.vec)
sd(est.mean.vec)
mean( (est.mean.vec-N)^2)
mean(est.max.vec)
sd(est.max.vec)
mean( (est.max.vec-N)^2)
```

8. Obtain boxplots of them, adding a line at 2012:

```
boxplot(est.mean.vec,est.max.vec,names=c("est.mean","est.max"))
abline(h=2012,lty=2)
```

   Which procedure seems better? Add a comment to this effect at the point indicated in the text file described below.

9. Now open a text editor (e.g. Nedit from the Session Manager) and create a file of the following format:

```
STAT2011 Week 1 Barack Obama
\code
...
...    # all your commands (ONLY commands, no output)
...    # from questions 1 to 7 go here
\end

\graph
...
...    # the commands from question 8 go here
...
\end
Type a comment here.
```

   Save the file as e.g. `prac1`.

10. Go to the R prompt and type

    ```
    > process(prac1)
    ```

    (replace `prac1` with your filename if it is different). If all has gone well, a PDF viewer should open with your report. Show this (on the screen) to your tutor to get marked off for the week.