STM4PSD – Workshop 8

Calculating estimators in R

It is fairly straightforward to calculate the estimators using R. Assume that you have stored a vector of samples in a variable called `my.data`. For example, you might store a small set of data like so:

```
my.data <- c(14.3, 20.2, 13.5, 17.4)
```

The appropriate functions and techniques are listed below.

- The sample mean is calculated using `mean(my.data)`.
- The sample variance is calculated using `var(my.data)`.
- The sample standard deviation is calculated using `sd(my.data)`.
- The standard error must be calculated manually, using `sd(my.data)/sqrt(length(my.data))`. Note the use of the `length` function, which returns the number of elements in a vector (giving the required value of n).
- A confidence interval must be calculated manually. You will need to do two calculations: one for the lower bound, and one for the upper bound. A shortcut method for this will be given in the solutions for this lab.

1. Use these functions to determine the confidence intervals for the data in Questions 4, 5 and 6 of Workshop 7.
2. Write a function `interval` which takes one argument, `data`, and computes an approximate 95% confidence interval for that data. It should return a vector where the first element is the lower bound of the interval and the second element is the upper bound.

Recall that, in order to calculate an approximate $\gamma \times 100\%$ confidence interval requires evaluating the quantile function for a normally distributed random variable $Z \sim N(0,1)$. The appropriate quantile $Q_Z\left(\frac{\gamma+1}{2}\right)$ can be calculated in R by typing

```
qnorm((gamma+1)/2)
```

and replacing `gamma` with the appropriate value of γ. Note that this number is called a *critical value*.

3. Use this to verify that the appropriate critical value for an approximate 95% confidence interval is 1.96.
4. Use R to verify the confidence intervals you calculated in Question 7 of Workshop 7.

Importing data from a spreadsheet

We will now look at the pressure data mentioned in the reading notes, which was considered in [1]. This data set can be found on the LMS. They were originally downloaded from the Oxford University Press website (**click here**). To load the data in R:

- In RStudio, go to the `File` menu, select `Import Dataset` and then choose `From Text (base)...`
- Locate the file `pressure1.csv`, select the file and press the `Open` button.
- There are some options here that you can change, but you should be able to leave them untouched.
- In the bottom-right you should see a **Data Frame** with two columns. Data frames are discussed below.
- Select `Import` and the data will now be stored in a data frame called `pressure1` (unless you changed the name).
- A new window will open, displaying the data. Go back to your working script file for this lab.

Data frames are an important data structure used by R for a lot of its modelling tools. Roughly speaking, a data frame is built out of a number of rows and columns. You can think of it as a representation of a spreadsheet of data.

- You can access elements of the data frame by using square brackets. Some examples are given below.

```
pressure1[4,1]           # access the element in row 4, column 1
pressure1[4, ]           # access the entirety of row 4
pressure1[, 1]           # access the entirety of column 1
pressure1[4, "Pressure"] # access the element in row 4, column "Pressure"
```

  Note that a comma must always be used; leaving the first number blank will let you obtain an entire row, and vice-versa for the columns.

- An easier way to access an entire column is by using a `$` followed by the name of the column. For example, to access the `Pressure` column, you would type

  ```
  pressure1$Pressure
  ```

  Run these commands in R to see for yourself.

5. You can plot a histogram of the data using a command like the following:

   ```
   hist(pressure1$Pressure, freq=FALSE, xlab="Pressure",
                            main="Density histogram of pressure values")
   ```

   Note the use of `freq=FALSE` to ensure it is a density histogram. From this histogram, try and guess the mean and standard deviation of the pressure data.

6. You probably noticed that the default histogram doesn't look particularly compelling. You can use the `breaks` argument to set a larger number of bins.

   Note that R uses the `breaks` argument as a suggestion only. Behind the scenes, it uses a function called `pretty` to calculate "nice" values for the bins. So you may not always see a change when you change the `breaks` argument.

   Set `breaks=20` and then try to guess the mean and standard deviation again.

7. You can calculate the estimates for the pressure data using the following commands:

   ```
   mean(pressure1$Pressure) # Sample mean
   var(pressure1$Pressure)  # Sample variance
   sd(pressure1$Pressure)   # Sample standard deviation
   ```

   Using the `curve` function with `add=TRUE`, add the density function for a normal distribution using appropriate estimates over the top of your histogram.

8. On the LMS, you will find three more data sets: `pressure2.csv`, `pressure3.csv` and `pressure4.csv`. Create histograms of these datasets can see if you can guess the sample means and sample standard deviations for each. Compare your guesses to the true estimates found using R. It can be difficult to guess a suitable mean and standard deviation from skewed histograms. How did you go?

9. Using R, create 95% confidence intervals for each of the pressure files.

References

[1] R. Pearson. *Exploring Data in Engineering, the Sciences, and Medicine*. Oxford University Press, New York, 2011.