Assignment 1

Part I – Task 1

It is best to save the data file **Table2_1.txt** to a directory on your USB Flash Drive. For the assignments we will assume that all relevant files are stored in the directory **G:\Chemometrics**. Clicking on the data file will open it up with Notepad, and we can view the contents. At this stage we should open up R and go to File->Change Directory and change the root directory to **G:/Chemometrics**.

Part II – Task 2

We can find help on the R-commands **rnorm**, **dnorm**, and **pnorm** by typing the following[1] into the command line and hitting the return key:

```
#  GET HELP ON NORMAL DISTRIBUTION
> ?rnorm
> ?dnorm
> ?pnorm
```

When we enter the command **rnorm(10)** this generates 10 random numbers that follow the standard normal distribution with mean 0 and standard deviation 1. Shown below are 10 such random numbers. It should be noted that your own set of numbers will probably differ to the below set, since the numbers are generated randomly.

```
[1]     0.1258022     0.7081296     -0.1400774
[4]      -0.8861057   -0.8202631    0.5003591
[7]     -0.5869571    -0.7492768     0.7202016     2.0982128
```

If we want to save the outputted numbers from the **rnorm** function to a text file we use the following commands:

```
#  SAVE RANDOMLY GENERATED NUMBERS TO TEXT FILE
> x=rnorm(10)
> write(x,file=”Normal.txt”)
```

The first command generates 10 random numbers that follow the standard normal distribution and stores them in an array denoted by **x**. The second command writes the data in **x** to a text file called **Normal.txt** in the root directory which should be set to **G:/Chemometrics**. If we open up this text file with Notepad we will see the contexts of **x**.

[1] Note that all functions and variables in R are case-sensitive. Also, in these assignments all R commands are preceded by the characters > or +. Comments are in green and preceded by the hash character #. Outputs are in red.

Part II – Task 3

To create the standard normal cumulative distribution function table we use the following commands:

```
#  GENERATE DISTRIBUTION TABLE
> x=seq(-3.4,3.49,by=0.01)
> x=pnorm(x)
> x=round(x,digits=4)
> matrix(x,byrow=T,ncol=10)
```

The first command generates a vector of numbers starting with -3.4, ending with 3.49 with increments of 0.01 i.e. -3.40, -3.39, -3.38,….,3.48, 3.49. These numbers are stored in the variable **x**. The second command generates the cumulative distribution function for each of the values in **x**, overwrites the variable x, and stores the **pnorm** values in **x**. For the first value -3.4, the **pnorm** is 0.0003369293 and this represents the area under the standard normal curve with mean 0 and standard deviation 1 which is below -3.4. The third command reduces the precision of the values in **x** to four decimal places i.e. the pnorm value of 0.0003369293 is now changed to 0.0003. The final command creates a matrix with 10 columns (denoted by **ncol=10**) and the values from **x** are placed in the matrix along rows (denoted by **byrow=T** where **T** means **True**). This means that the top row will contain the first 10 values of **x** from left to right, and the next row will contain the next 10 values and so on. If **byrow** had a value of false, then this would mean that we would fill the matrix by columns, starting with the furthest left column. The value from the 7th column and 45th row of Table A.1 of the Chemometrics book is retrieved by:

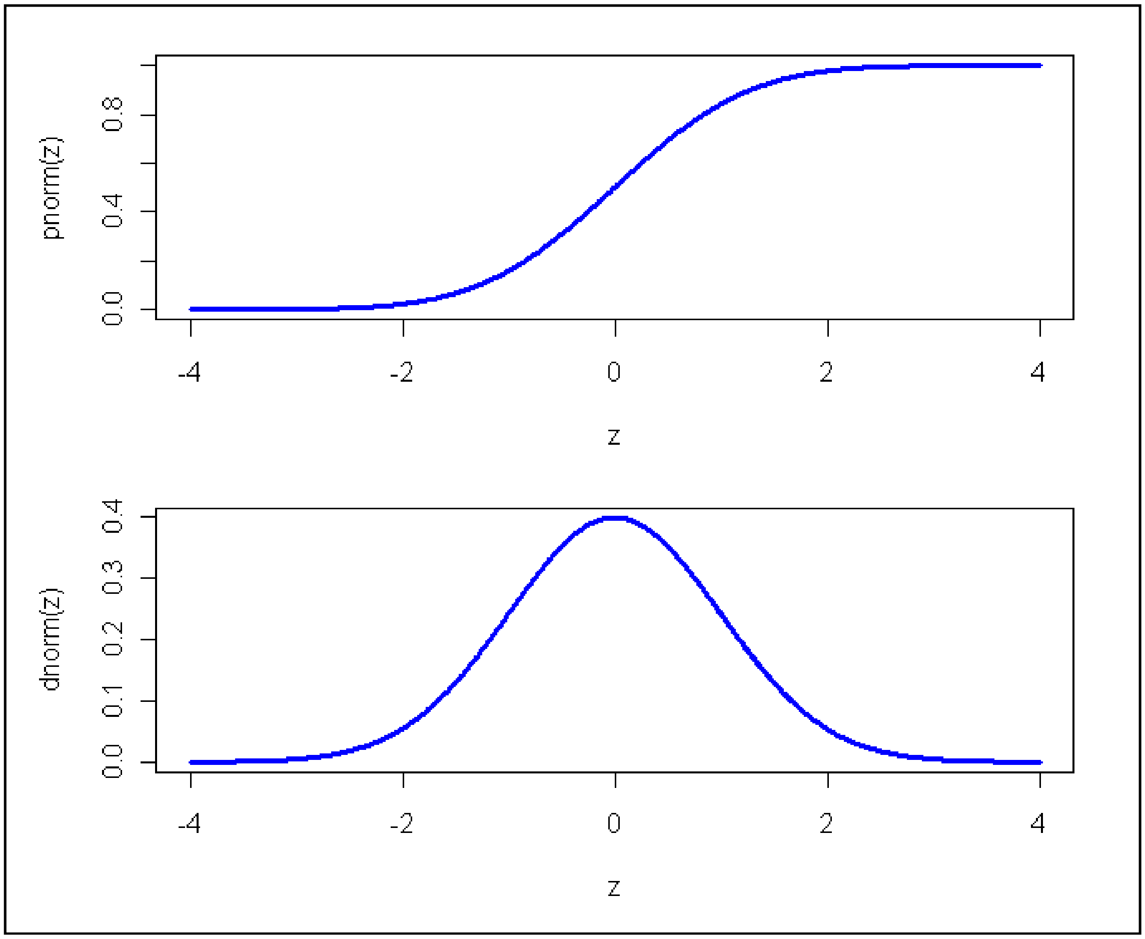
```
#  GET CUMULATIVE DISTRIBUTION VALUE FOR 1.06
> pnorm(1.06)
[1]      0.8554277
```

Part II – Task 4

To plot the cumulative distribution function and density of the standard normal model we use the following commands:

```
# PLOT DISTRIBUTION FUNCTIONS
> par(mfrow=c(2,1))
> z=seq(-4,4,by=0.01)
> plot(z,pnorm(z),type=’l’,col=”blue”,lwd=3)
> plot(z,dnorm(z),type=’l’,col=”blue”,lwd=3)
```

The first command sets the graphical parameters. The function **mfrow** specifies that we are drawing by rows, and that figures will be drawn on a 2x1 array i.e. 2 rows, 1 column. The first command generates a vector of numbers starting with -4, ending with 4 with increments of 0.01. These numbers are stored in the variable **z**. The third command graphs the cumulative distribution function of the standard normal model. The parameter **type=’l’** specifies that we want a line, and not points to be displayed on the graph. We can set the colour of the line using **col=”blue”** and finally we can set the line width to a thickness of 3 using **lwd=3**. Similarly in the final command we plot the density function of the standard normal below the first figure. Shown on the following page are the figures generated by R.



Part III – Task 5

Download and save the data sets **Table1_1.txt** and **Table2_1.txt** to the directory **G:/Chemometrics**. If we want to read the data from the first text file we can use the following command.

```
#  READ DATA FROM TEXT FILE
> Titres=read.table("Table1_1.txt", row.names=1)
```

If the root directory is not set to **G:/Chemometrics** then we can read the data as follows:

```
#  READ DATA FROM TEXT FILE
> Titres=read.table("G:/Chemometrics/Table1_1.txt",row.names=1)
```

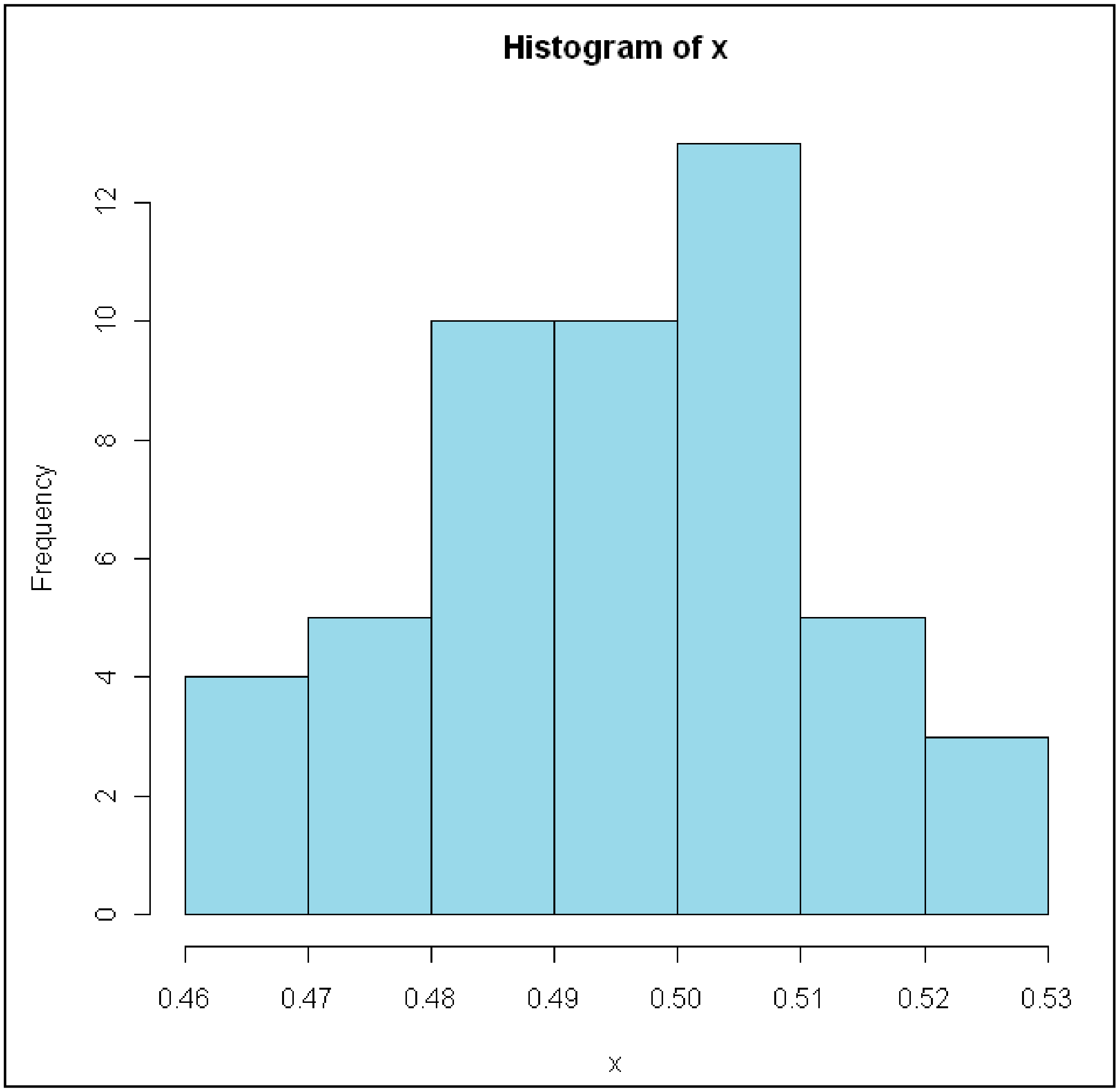
Note the forward slash character / is used for paths for R. The parameter **row.names=1** specifies that row identifiers are contained in the first column. If we did not include this parameter then the student identifiers A, B, C, D would be included with the numerical data. We can read the data from the second data set **Table2_1.txt** using the following command:

```
#  READ DATA FROM TEXT FILE, SAVE AS VECTOR
> scan("Table2_1.txt")
```

Part III – Task 6

We use the following commands to generate the histogram of the data contained in **Table2_1.txt**. The subsequent figure is also shown below. Note that the colour was added after the figure was generated by R in Microsoft Paint.

```
#  READ DATA FROM TEXT FILE, AND CREATE A HISTOGRAM
> x=scan(“Table2_1.txt”)
> hist(x)
```



Part IV– Task 4

When saving workspaces remember to include the suffix **.RData** in the name; R will not generate this file extension automatically.

Part V– Task 5

We are told in the question that the exact value is 10ml. We use the following command to check the difference between the mean titres for each student and the exact value of 10ml.

```
#  GET BIAS FOR RESULTS FOR EACH STUDENT
> rowMeans(Titres)-10
```

The outputted results are shown below in table format:

Systematic Error (Bias)			
A	B	C	D
0.10	0.01	-0.10	0.01

Next, we want to determine the standard deviations of the titres for each student, and we do this by using the following command:

```
#  GET STANDARD DEVIATION OF RESULTS FOR EACH STUDENT
> apply(Titres,1,sd)
```

The above command means that we want to apply standard deviation (denoted by **sd**) across all the rows (denoted by the value of **1**) of the matrix **Titres**

Random Error (Standard Deviation)			
A	B	C	D
0.01581139	0.17175564	0.21047565	0.03316625

We observe that students B and D have small bias while random errors are small for students A and D. We may conclude that the performance of student D is best.