Summer/Winter/Semester2 **Tutorial Solutions 2** 2015

This tutorial is an introduction to R and explores graphical summaries (stem and leaf plot, histogram and ordinate diagram) and classifying data.

Learning in Tutorials

- The role of the tutor is to help you learn more deeply, so the more work you do before the tutorial, the more you will get out of it.
- Each tutorial sheet consists of a set of practise exercises, involving both hand calculations and R. Work at your own pace in the tutorial class and then finish off all the questions at home.
- Find a study partner or small group to work with for the 3 Reports. Arrange a time to meet to work on your Reports together. You can bring your draft Reports to tutorial classes to get feedback from your tutor.
- If you finish all the tutorial questions, then work on your next Report or the Revision material.
- If you miss any of the OnlineQuizzes (due to choice or misadventure), the 'better mark principle' will automatically apply - the mark % for that task will be added to the exam. The better mark principle does not apply to the Reports.

1. Introduction to Lab

- Login to Zeno: Type in your Unikey information.
  Note: If the computer is not set up for Zeno, change it to Zeno on the desktop.
- Start the R program: Bring up the fluxbox menu by right clicking the mouse while the cursor is on the grey bakcground, and then click on the *R Command Line*
- Open up Firefox and find the MATH1005 page: maths.usyd.edu.au/MATH1005/. Now you can can access the tutorial sheets and data files.
  Note: Firefox may come up automatically when you login. Otherwise, you can find it in the fluxbox.
- Rearrange your desk top so that you can work in both R and Firefox concurrently.
  Note: Close the extra command window, which opens up automatically on login.
- Log Off: When it comes time to log off, use the fluxbox.

2. Loading data into R

Given the data 1 3 5 7 8 8

- Enter the data manually

```
> x=c(1,3,5,7,8,8)
> x
[1] 1 3 5 7 8 8
```

- Copy and paste the data from the PDF file of this page
  - At the R prompt enter `y=scan()` (the prompt changes to "`1:`").
  - Make sure your PDF viewer has its "text select" tool active, then select and copy the numbers.
  - Click next to the "`1:`" prompt, paste the numbers and hit Enter twice.

```
> y=scan()
1: 1 3 5 7 8 8
7:
Read 6 items
> y
[1] 1 3 5 7 8 8
```

- Read in the data from the internet

```
> z=scan(file=url("http://www.maths.usyd.edu.au/math1005/r/wk2q1.txt"))
Read 6 items
> z
[1] 1 3 5 7 8 8
```

3. Using Commands in R

Using your data stored in `x`, produce the following graphical summaries:

```
> table(x)
> plot(x)
> plot(table(x))
> stem(x)
> hist(x)
> boxplot(x)
```

Note that each command can be customised. Find out the options using `help()` or `?`. A list of `pch` (plotting characters) are here: http://www.statmethods.net/advgraphs/parameters.html.

Experiment with customising the commands.

```
> help(plot)
> plot(x,main="This is the main title",xlab="This is the x axis label",col="blue",pch=6)

> ?hist
> hist(x,freq=FALSE,main="Histogram",ylab="Probabilities", col="green")

> ?boxplot
> boxplot(x,horizontal=TRUE,col="red")
```

4. Graphical Summaries by hand

In an attempt to measure the 'true' heat of sublimation of platinum (in kcal/mole), Hampson and Walker (1961) recorded the following data:

136.2	136.6	135.8	135.4	134.7	135.0	134.1	143.3
147.8	148.8	134.8	135.2	134.9	146.5	141.2	135.4
134.8	135.8	135.0	133.7	134.2	134.9	134.8	134.5
134.3	135.2						

(a) Complete the following 'single' stem and leaf plot, where the break is at the decimal point. Comment on it's shape.

133	7
134	1 2 ...
⋮	
⋮	
⋮	
147	8
148	8

Note: The single stem version has 10 digits/leaves on each row/stem.

(b) Complete the following frequency table.

Interval	Frequency	Relative Frequency (3dp)	Height (3dp)
[133,134)	1	0.038	0.038
[134,135)	10		
[135,136)			
[136,137)			
[137,140)			
[140,143)			
[143,146)			
[146,149)	3	0.115	0.038
Total			

where:
Relative Frequency = Frequency/Total
Height = Relative Frequency/Interval Width

(c) Draw a histogram and describe its shape.

Solution

(a)

133	7
134	1 2 3 5 7 8 8 8 9 9
135	0 0 2 2 4 4 8 8
136	2 6
137	
138	
139	
140	
141	2
142	
143	3
144	
145	
146	5
147	8
148	8

Comment: The shape is right skewed (long right tail).
(b)

Interval	Frequency	Relative Frequency %	Height
[133,134)	1	$1/26 = 3.8$	$3.8/1=3/8$
[134,135)	10	$10/26 = 38.5$	$38.5/1=38.5$
[135,136)	8	$8/26 = 30.1$	$30.1/1=30.1$
[136,137)	2	$2/26 = 7.7$	$7.7/1=7.7$
[137,140)	0	0	$0/3=0$
[140,143)	1	$1/26 = 3.8$	$3.8/3 \approx 1.3$
[143,146)	1	$1/26 = 3.8$	$3/8/3 \approx 1.3$
[146,149)	3	$3/26 = 11.5$	$11.5/3 \approx 3.8$
Total	26	100 (rounding)	

(c)
Compare your histogram to the R output in Q5. Clearly label axes. Comment: The histogram again reveals right skewing.

5. Graphical Summaries in R

Enter the data from Question 4 into R, and produce a stem and leaf plot, frequency table and histogram.

```
> x=scan(file=url("http://www.maths.usyd.edu.au/math1005/r/wk2q4.txt"))
> stem(x)
> stem(x,scale=2)
> hist(x,breaks=c(133:137,140,143,146,149),right=F)
> hist(x,breaks=c(133:137,140,143,146,149),right=F)$counts
```

Note: `$counts` adds the counts per interval.

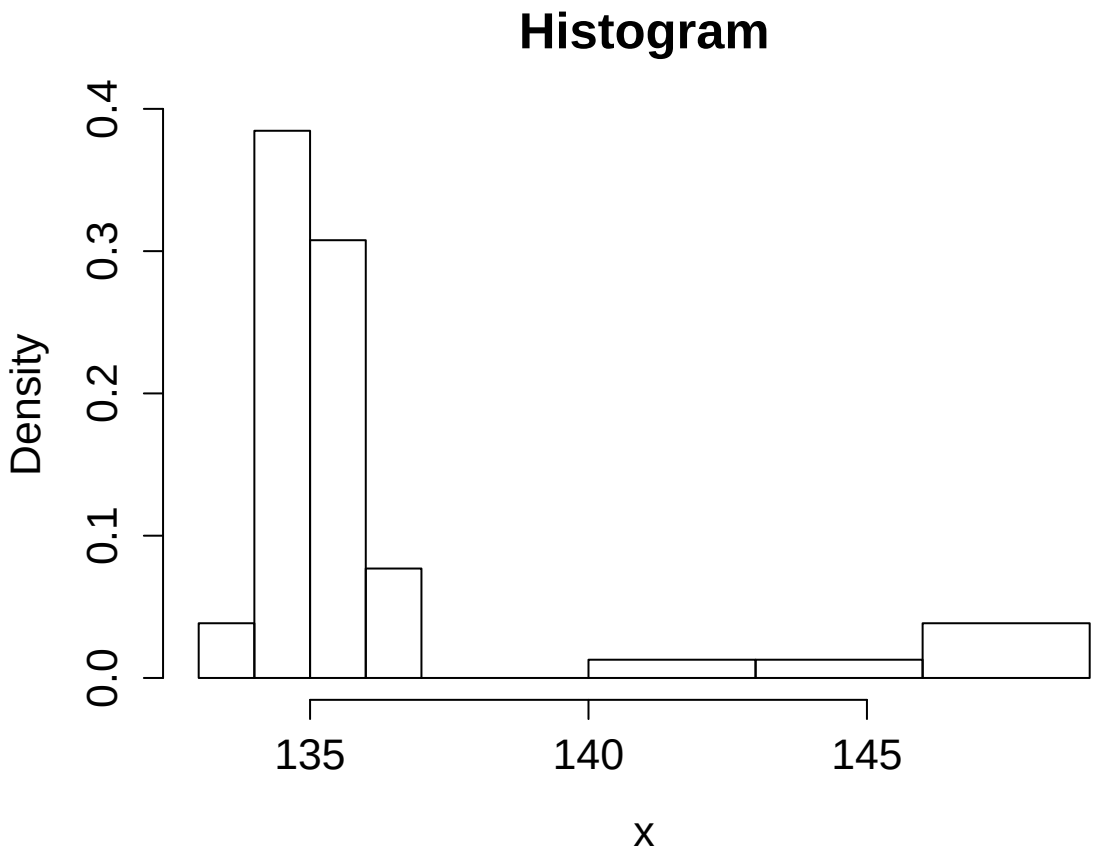
Solution

```
> x=scan()
1: 135.4 134.7 135.0 134.1 143.3 135.2 134.9 146.5 141.2 135.4 133.7 134.2 134.9 134.8
134.5
27:
Read 26 items

> stem(x)

  The decimal point is at the |

  132 | 7
  134 | 123578889900224488
  136 | 26
  138 |
  140 | 2
  142 | 3
  144 |
  146 | 58
  148 | 8

> stem(x,scale=2)

  The decimal point is at the |

  133 | 7
  134 | 1235788899
```

```
135 | 00224488
136 | 26
137 |
138 |
139 |
140 |
141 | 2
142 |
143 | 3
144 |
145 |
146 | 5
147 | 8
148 | 8
```

Note: R chooses what it considers to be an appropriate spread of the stem and leaf plot. So we use the parameter scale = to change the default layout. In this particular example, using scale = 2 produces a single stem plot.

```
>hist(x,breaks=c(133: 137,140,143,146,149),right=F,main="Histogram")
```



6. Double stem and leaf plot by hand (From the 1998 examination)

A mining company finds a body of ore and obtains 24 core samples by drilling at equally spaced intervals along the body. The samples are analysed for percentage content of a valuable mineral giving the following results:

17	18	26	18	31	31	19	17
22	13	19	17	16	14	13	10
16	14	13	23	16	20	18	30

Prepare both single and double stem-and-leaf plots. Which one is preferable and why?

Note: The double stem version has 5 digits/leaves on each row/stem.

Solution

Check your working against the following R output.

```
> x=scan()
1: 17   18   26   18   31   31   19   17   22   13   19   17   16   14   13   10  16 14 13 23
25:
Read 24 items

> stem(x)

  The decimal point is 1 digit(s) to the right of the |

  1 | 033344
  1 | 66677788899
  2 | 023
  2 | 6
  3 | 011

> stem(x,scale=0.5)

  The decimal point is 1 digit(s) to the right of the |

  1 | 03334466677788899
  2 | 0236
  3 | 011
```

Note: Here by default R chooses a double stem and leaf plot. Hence,we use scale=0.5 to get a single stem and leaf plot.

Comment: The double stem plot is preferable, because it's easier to see the shape (some right skewing). The single stem plot is a bit overcondensed.

7. Ordinate diagram by hand

The following table gives the number of ice creams sold in a coffee shop on each day in January 2002 in a Canadian city:

2	0	0	1	1	0	2	1
3	3	6	7	0	4	1	0
1	1	3	2	1	0	8	0
0	4	5	1	0	2	3	

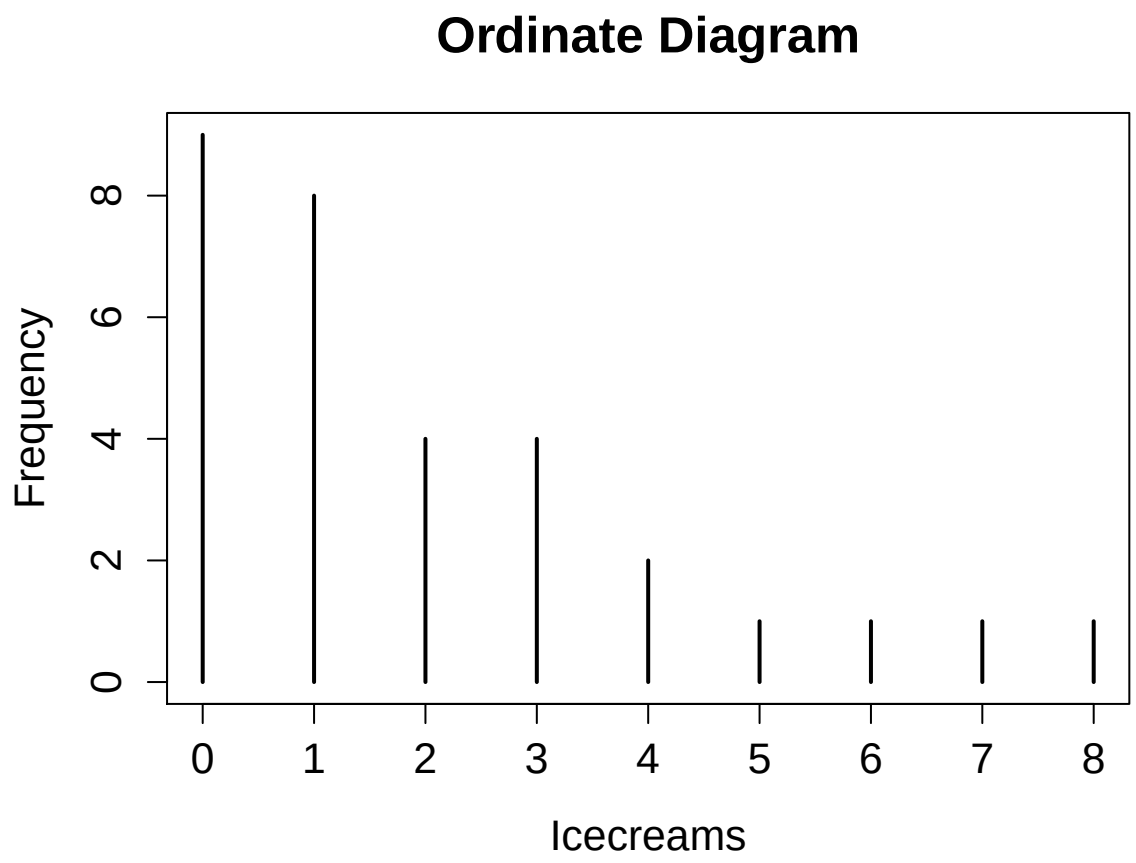
Prepare a suitable frequency distribution table for this data. Draw an ordinate diagram and comment.

Solution

Check your working against the following R output.

```
> x=scan()
1: 2   0   0   1   1   0   2   1 3   3   6   7   0   4   1   0 1   1   3   2   1   0   8   0
32:
Read 31 items

> table(x)
```

```
x
0 1 2 3 4 5 6 7 8
9 8 4 4 2 1 1 1 1

> plot(table(x),xlab="Icecreams", ylab="Frequency",main="Ordinate Diagram")
```



Comment: The data is right skewed, indicating there are many days when very few icecreams are sold. Selling icecreams in Canada is not a good idea!

8. Classifying data

Classify each of the data sets in the Appendix of the Phipps and Quine reference book.

Solution

Categorical data: None
Discrete Data: Set 1, 2, 5, 6, 10
Continuous Data: Rest