Copyright: University of Sydney

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.
- Work at your own pace during the tutorial class and then finish off all the questions at home.
- If you finish all the tutorial questions, work on your next Report or the Revision material.

1. Australian Road Fatalities

The following data is road fatalities in Australia from Jan 1989 - May 2016.

```
Road <- read.csv("http://www.maths.usyd.edu.au/u/UG/JM/MATH1005/r/StatsData/
                 AllFatalities.csv")
```

(a) What size is the data? How many variables are there? How does this data relate to the example in Topic1 lectures?

```
dim(Road)
names(Road)
```

(b) A Data Dictionary gives information about a given data set, such as collection, missing values and variables. Classify the following variables using the Data Dictionary.

http://www.maths.usyd.edu.au/u/UG/JM/StatsData.html

Variable	Numerical Continuous	Numerical Discrete	Categorical Nominal	Categorical Ordinal
State			Y	
Year		Y		
Dayweek				
CrashType				
SpeedLimit				
RoadUser				
Gender				
Age				

To check how R has classified the variables, use

```
str(Road)
```

Note: You may have rightly characterised Age as Numerical data, but here R has chosen to classify Age as Categorical Ordinal (factor).

(c) Select the Age variable and produce the following graphical summaries:

```
Age <- Road$Age
class(Age)
table(Age)
plot(table(Age))
```

(d) Now re-classify the Age variable as numerical data, and then produce the following graphical summaries:

```
AgeN<- as.numeric(levels(Age))[Age]
class(AgeN)
hist(AgeN)
boxplot(AgeN)
```

(e) Each command can be customised. Find out the options using `help()` or `?`.

Experiment with customising these commands.

```
help(hist)
hist(AgeN,freq=FALSE,main="Histogram",ylab="Probabilities", col="green")
?boxplot
boxplot(AgeN,horizontal=TRUE,col="red")
```

(f) Write a sentence summarising what the histogram and boxplot tell you about the age of road fatalities in Australia (Jan 1989 - May 2016).

(g) What is the most common and least common days for road fatalities?

```
Dayweek <- Road$Dayweek
class(Dayweek)
table(Dayweek)
plot(table(Dayweek))
```

Note to change the labels to shorter names:

```
levels(Dayweek)
levels(Dayweek) = c("F","M","S","Su","Th","T","W")
plot(table(Dayweek))
```

(h) (Extension: How can you make more sense of this graph? Reorder the factors alphabetically.)

```
DayweekOrdered <- factor(Dayweek,levels=c("M","T","W","Th","F","S","Su"))
plot(table(DayweekOrdered))
```

(i) Pick another variable and do your own investigation.

Solution

This data consists of all Australlian road fatalities from 1989-2016: 46624 fatalities with 18 variables. The data from lectures was just a recent subset of the data: Jan-March 2016.

Variable	Numerical Continuous	Numerical Discrete	Categorical Nominal	Categorical Ordinal
State			Y	
Year		Y		
Dayweek			Y	
CrashType			Y	
SpeedLimit				Y
RoadUser			Y	
Gender			Y	
Age	Y	Y		Y

The most common day is Saturday, and then least common day is Monday.

2. Efficiency of Australian Commercial Refrigerators

The following data concerns commercial refrigerators in Australia, in particular the efficiency (in kWh/24h/m).

```
Fridge <- read.csv("http://www.maths.usyd.edu.au/u/UG/JM/MATH1005/r/StatsData/
                   Refrigerators.csv")
```

(a) What size is the data? How many variables are there?

```
dim(Fridge)
names(Fridge)
```

(b) Classify the following variables using the Data Dictionary.

http://www.maths.usyd.edu.au/u/UG/JM/StatsData.html

Variable	Numerical Continuous	Numerical Discrete	Categorical Nominal	Categorical Ordinal
Brand			Y	
Country				
Sold in				
Total energy cons				
Efficiency				

(c) Comment on the country of manufacture.

```
Country <- Fridge$Country
class(Country)
plot(Country)
table(Country)
```

(d) Comment on efficiency.

```
Efficiency <- Fridge$Efficiency..kWh.24h.m..
hist(Efficiency)
boxplot(Efficiency)
```

(e) Pick another variable and do your own investigation.

Solution

The data consists of 2480 commerical refrigerators with 23 variables.

Variable	Numerical Continuous	Numerical Discrete	Categorical Nominal	Categorical Ordinal
Brand			Y	
Country			Y	
Sold in			Y	
Total energy cons		Y		
Efficiency		Y		

The most common countries are: Italy (640) and China (624), followed by New Zealand (314). Efficiency has a right skewed distribution with outliers, indicating some refrigerators have an efficiency far exceeding/worse than the rest.

3. Investigate demographics in in the USA, using the following Big Data summaries: http://datausa.io/map/

For example:
(a) What states have the highest and lowest Violent Crime rates?

(b) What states have the highest and lowest Average Salary?

(c) What states have the highest incidence of Diabetes and Obesity?

Solution

Violent Crime: Tennesee 621.3/100,000 or 0.006213% (highest); Maine 122.7 (lowest)
Average Salary: Columbia $72,810.10; Puerto Rico $25,227 USD.
Diabetes: Mississippi 12.5%; Colorado 6.8% Obesity: Mississippi 35.3%; Colorado 20.1%

Extra Questions

4. Classify Data in Everyday Life

For the following scenarios, find appropriate variables.

Scenario	Numerical Continuous	Numerical Discrete	Categorical Nominal	Categorical Ordinal
Netball Match				
KFC Drive-through				
Hospital				
Masterchef				

Solution

Scenario	Numerical Continuous	Numerical Discrete	Categorical Nominal	Categorical Ordinal
Netball Match	Height of players	Number of Goals	Position: Goals/Court	Played which Quarter
KFC Drive-through	Time	Number of Items	Gender of Staff	Rate service
Hospital	Age of building	Number of staff	State	Cleanliness standard
Masterchef	Quantity of cream	Number of contestants	Type of dish	What Star Chef

5. True heat of Platinum

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						

Source: http://nvlpubs.nist.gov/nistpubs/jres/65A/jresv65An4p289_A1b.pdf

(a) Who would be interested in ths data and why?

(b) Import the data into R.

```
Plat=scan(file=url("http://www.maths.usyd.edu.au/u/UG/JM/MATH1005/r/StatsData/
                Platinum.txt"))
```

(c) Produce graphical summaries and comment on your results.

```
stem(Plat)
stem(Plat,scale=2)
hist(Plat,breaks=c(133:137,140,143,146,149),right=F)
hist(Plat,breaks=c(133:137,140,143,146,149),right=F)$counts
```

Note:
(1) 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.

(2) `$counts` adds the counts per interval.

(d) Now complete the graphical summaries by hand. Complete the following unsorted 'single' stem and leaf plot, where the break is at the decimal point. The first 5 entries have been done already.

133	
134	7
135	8 4 . . .
136	2 6 . . .
⋮	
⋮	
⋮	
148	

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

(e) Complete the following frequency table, and draw a histogram.

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

Solution

Data could be of interest to anyone using Platinum commercially. The shape is right skewed (long right tail), indicating some very high values of 'true' heat relative to the rest of data, which could be investigated further.

Solution

Check your hand working against the following R output.

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)	

6. Mining Core Samples (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=c(17,18,26,18,31,31,19,17,22,13,19,17,16,14,13,10,16,14,13,23,16,20,18,30)
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. Icecreams in Canada

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.

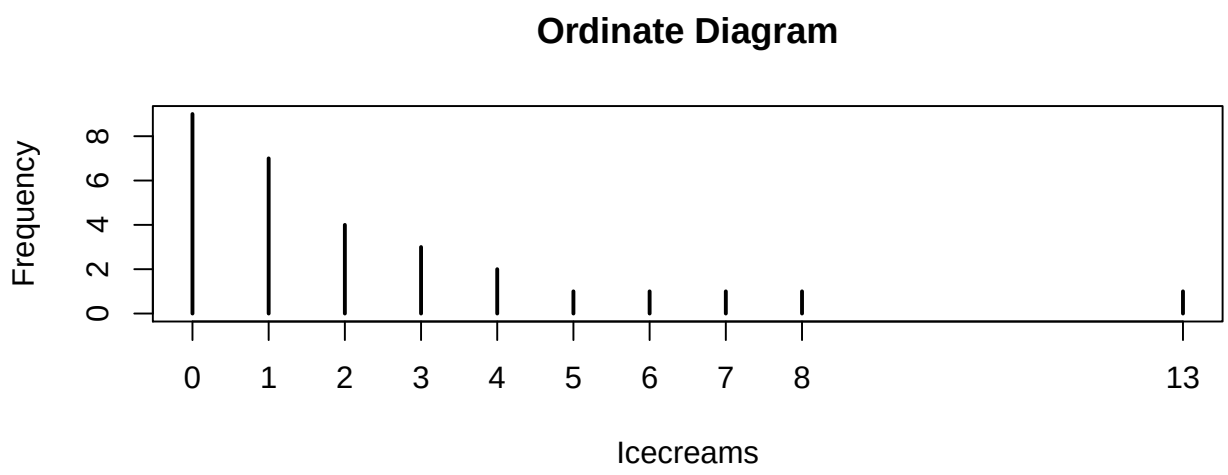
Solution

Check your working against the following R output.

```
x=c(2,0,0,1,1,0,2,13,3,6,7,0,4,1,0,1,1,3,2,1,0,8,0,0,4,5,1,0,2,3)
table(x)

## x
##  0  1  2  3  4  5  6  7  8 13
##  9  7  4  3  2  1  1  1  1  1
```

```r
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, as might be expected given climate.