Dr Isabella Gollini

Probability and Statistics

1 Descriptive Statistics, Plots and R

1.1 An Example – Lengths of Major North American Rivers

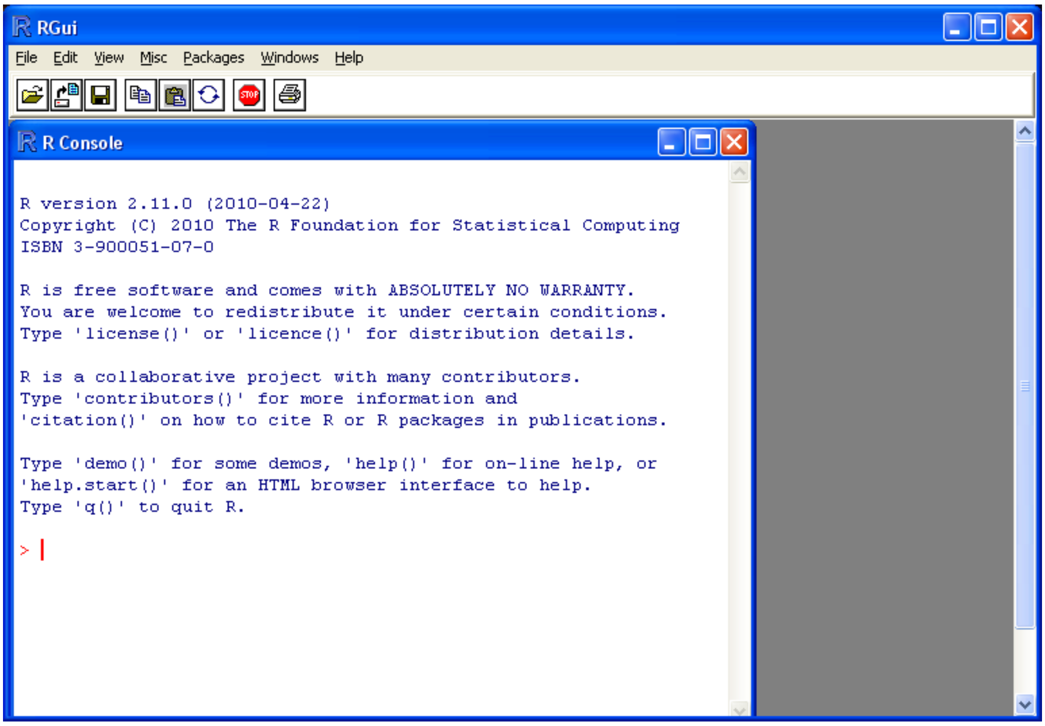
This data set gives the lengths (in miles) of 141 "major" rivers in North America, as compiled by the US Geological Survey in 1975. [1] The data set is freely available in R.

We shall carry out an initial investigation of the data using descriptive and graphical techniques, with the aid of the statistical package R.

1.2 R

The statistical software package which will be used in this module is called R. Unlike all other statistical software packages, R is free and therefore you can download it and work at home/on the bus/wherever!

R is a free software environment for statistical computing and graphics. It is open source and therefore it is constantly being updated as new 'packages' or libraries which perform different statistical techniques are added by people right across the world.



To install R on your own laptop or desktop machine go to `www.r-project.org` and click on 'CRAN'.

[1] Source: World Almanac and Book of Facts, 1975, page 406.

These notes are integrated with `R` code that is contained in the boxes, it allows you to reproduce all the examples.

```
rivers

##   [1]  735  320  325  392  524  450 1459  135  465  600  330  336
##  [13]  280  315  870  906  202  329  290 1000  600  505 1450  840
##  [25] 1243  890  350  407  286  280  525  720  390  250  327  230
##  [37]  265  850  210  630  260  230  360  730  600  306  390  420
##  [49]  291  710  340  217  281  352  259  250  470  680  570  350
##  [61]  300  560  900  625  332 2348 1171 3710 2315 2533  780  280
##  [73]  410  460  260  255  431  350  760  618  338  981 1306  500
##  [85]  696  605  250  411 1054  735  233  435  490  310  460  383
##  [97]  375 1270  545  445 1885  380  300  380  377  425  276  210
## [109]  800  420  350  360  538 1100 1205  314  237  610  360  540
## [121] 1038  424  310  300  444  301  268  620  215  652  900  525
## [133]  246  360  529  500  720  270  430  671 1770
```

In order to access the help of the dataset (and also of all the packages and functions) you have to use the symbol `?` before the object you are interested into. For example, to access the full description of the dataset use: `?rivers`

1.3 Basic descriptive statistics

A long list of data is difficult to get to grips with, so we shall carry out calculations to summarise the data in a variety of ways. Let n denote the sample size. In general, the observed values in a sample of size n may be represented by $x_1, x_2, \ldots, x_n$. This notation helps us to write down formulae for statistical calculations.

In our example the sample is represented by the lengths of the rivers:

```
x <- rivers
```

Calculate the sample size, that is the number or rivers in the dataset:

```
n <- length(x)
n

## [1] 141
```

In our example $n = 141$.

The *sample mean* $\bar{x}$ is the average of the observations,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

The sample mean is an example of a *statistic*, a number calculated from the sample data which in some way summarises the data.

Calculate the sample mean using the formula:

```
xbar <- sum(x) / n
xbar

## [1] 591.1844
```

In `R` there is the function `mean`:

```
mean(x)

## [1] 591.1844
```

Other examples of statistics are, the *sample minimum*, $\min(x_1, x_2, \ldots, x_n)$, the *sample maximum*, $\max(x_1, x_2, \ldots, x_n)$, and the *sample range* R,

$$R = \max(x_1, x_2, \ldots, x_n) - \min(x_1, x_2, \ldots, x_n).$$

Let's find the sample minimum and maximum:

```
min(x)

## [1] 135

max(x)

## [1] 3710
```

and the sample range:

```
R <- max(x) - min(x)
R

## [1] 3575
```

We can also use the function `range`, but it gives us the maximum and minimum:

```
range(x)

## [1]  135 3710
```

in order to obtain the sample range we have to make use of the function `diff` (difference between two values):

```r
diff(range(x))

## [1] 3575
```

The sample mean is the most important of the statistics described so far. It is a measure of "location" (or "centre"), in that it provides what is in some sense a typical value of the sample data. It will be important to have also a measure of "dispersion" (or "spread") of the data. The range is one such measure of dispersion, but not the most useful — it is generally too crude a measure because it is calculated only from the two extreme observations, the smallest and the largest. The most important measures of dispersion are the sample variance and the sample standard deviation.

The *sample variance* s^2 is defined by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

The *sample standard deviation* s is the square root of the sample variance. The larger the value of the sample variance the greater is the spread of the data.

The sample variance can be calculated by using the formula:

```r
s2 <- sum((x - xbar)^2) / (n - 1)
s2

## [1] 243908.4
```

or by using the function `var`:

```r
var(x)

## [1] 243908.4
```

The sample standard deviation is the square root of the sample variance:

```r
s <- sqrt(s2)
s

## [1] 493.8708
```

can also be found by using:

```r
sd(x)

## [1] 493.8708
```

The *sample median* is the value of the middle item, if the sample size is an odd number, or the average of the two middle items, if the sample size in an even number, when the data are arranged in increasing order. The sample median, like the sample mean, is a measure of location.

Sample median:

```
median(x)

## [1] 425
```

Ordered sample:

```
xo <- sort(x)
xo

##   [1]  135  202  210  210  215  217  230  230  233  237  246  250
##  [13]  250  250  255  259  260  260  265  268  270  276  280  280
##  [25]  280  281  286  290  291  300  300  300  301  306  310  310
##  [37]  314  315  320  325  327  329  330  332  336  338  340  350
##  [49]  350  350  350  352  360  360  360  360  375  377  380  380
##  [61]  383  390  390  392  407  410  411  420  420  424  425  430
##  [73]  431  435  444  445  450  460  460  465  470  490  500  500
##  [85]  505  524  525  525  529  538  540  545  560  570  600  600
##  [97]  600  605  610  618  620  625  630  652  671  680  696  710
## [109]  720  720  730  735  735  760  780  800  840  850  870  890
## [121]  900  900  906  981 1000 1038 1054 1100 1171 1205 1243 1270
## [133] 1306 1450 1459 1770 1885 2315 2348 2533 3710
```

Two further statistics, of a type similar to the median, are the *(sample) lower quartile*, Q_1, and the *(sample) upper quartile*, Q_3. Essentially Q_1 and Q_3 are given by the values of the items one quarter and three quarters of the way along when the data are arranged in increasing order, whereas the median corresponded to the value halfway along. In the present case, the median corresponds to the position

$$\frac{1}{2}(1+n) = \frac{1}{2}(1+141) = 71$$

in the ordered data.

We can find the median by using its position in the ordered sample:

```
xo[71]

## [1] 425
```

Exactly how the quartiles are calculated may differ slightly from package to package, in `R` there are 9 methods to calculate it. In general to define the quantile interpolation methods are used.

The default method in `R` to calculate the quantile q is given by:

```
h1 <- (n - 1) * .25  + 1
Q1 <- xo[floor(h1)] +
  (h1 - floor(h1)) * (xo[floor(h1) + 1] - xo[floor(h1)])
Q1

## [1] 310

h3 <- (n - 1) * .75  + 1
Q3 <- xo[floor(h3)] +
  (h3 - floor(h3)) * (xo[floor(h3) + 1] - xo[floor(h3)])
Q3

## [1] 680
```

There is a function to find the quantiles:

```
quantile(x, probs = c(0.25, 0.5, 0.75))

## 25% 50% 75%
## 310 425 680
```

Another measure of dispersion is the *(sample) interquartile range*, which is the difference between the upper quartile and the lower quartile, here

$$Q_3 - Q_1.$$

There is a function to find the interquartile range:

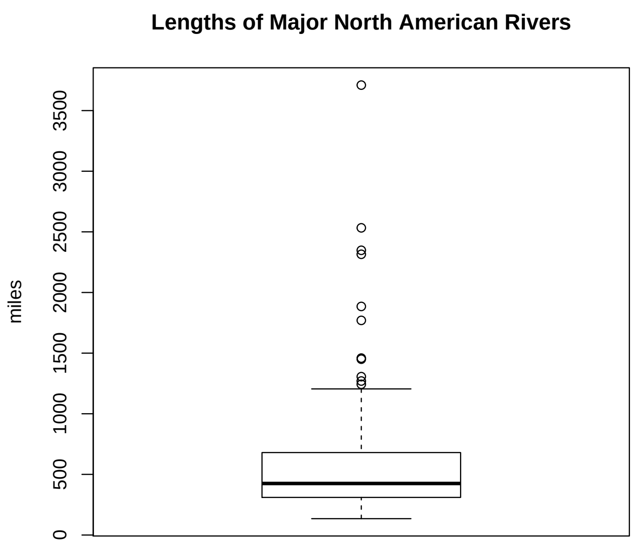
```
IQR(x)

## [1] 370
```

We have looked in detail at the calculation of some of the basic statistics on sample data, often referred to collectively as *descriptive statistics*.

The distribution of the data, including the values of the minimum, maximum, median and quartiles, may be illustrated graphically using what is known as a *boxplot* or a *box-and-whisker plot*.

```r
boxplot(x,
        ylab = "miles",
        main = "Lengths of Major North American Rivers")
```



In this plot, the lower and upper boundaries of the box are at the lower and upper quartiles, respectively. The median is shown as the horizontal line within the box. The "whiskers" extend to the minimum and the maximum, except that there is a cut-off point in that the length of either whisker is not allowed to exceed one and a half times the interquartile range. (The reason for this somewhat arbitrary limit will be discussed later in the course.) Points beyond the whiskers are regarded as "outliers" and are shown individually.

The boxplot illustrates what we might have already noticed that the main body of the sample values takes values, roughly speaking, between about 300 and 700. There are smaller values too, but a striking feature of the data is the presence of a few exceptionally large values over 2000.

1.4 Grouping of data and frequency distributions

To try to obtain a clearer picture of how the data are distributed in the sample, we calculate what is known as the *frequency distribution*. We count the number of sample members in various ranges of values of the variable being investigated. The variable in our example is the lengths of major North American rivers, measured in miles. The boundary values that we choose for our ranges of values are called the *class boundaries*. We shall take as our class boundaries 200, 400, 600 . . . , 3600, 3800. Note that the difference between successive class boundaries is the same, 200. It is usual and sensible to let the difference between successive class boundaries be the same. This difference is known as the *class interval*.

```r
cutpoint <- seq(0, 3800, by = 200)
cutpoint
```

```
##  [1]    0  200  400  600  800 1000 1200 1400 1600 1800 2000 2200
## [13] 2400 2600 2800 3000 3200 3400 3600 3800
```

The function `cut` assign each observation in `x` to a class:

```r
x_group <- cut(x, breaks = cutpoint)
```

Using the sorted sample data, we first read off the *cumulative frequencies*, the number of sample values less than each class boundary. From the partial listing of the sorted data in Table 1, for example, we see that there are no sample values less than 10, ..., there are 293 values less than 100, ..., 534 values less than 550, all 536 values are less than 580.

The *relative cumulative frequencies* are the cumulative frequencies divided by the sample size, i.e., the proportion of the sample less than each boundary value. The relative cumulative frequencies in Table 1 are expressed as percentages to one decimal place of accuracy.

Table 1: The frequency distribution of the river data

Class	Frequencies	Relative frequencies	Cumulative frequencies	Relative cum freq
(0,200]	1	0.7%	1	0.7%
(200,400]	63	44.7%	64	45.4%
(400,600]	33	23.4%	97	68.8%
(600,800]	19	13.5%	116	82.3%
(800,100]	9	6.4%	125	88.7%
(1000,1200]	4	2.8%	129	91.5%
(1200,1400]	4	2.8%	133	94.3%
(1400,1600]	2	1.4%	135	95.7%
(1600,1800]	1	0.7%	136	96.5%
(1800,2000]	1	0.7%	137	97.2%
(2000,2200]	0	0%	137	97.2%
(2200,2400]	2	1.4%	139	98.6%
(2400,2600]	1	0.7%	140	99.3%
(2600,2800]	0	0%	140	99.3%
(2800,3000]	0	0%	140	99.3%
(3000,3200]	0	0%	140	99.3%
(3200,3400]	0	0%	140	99.3%
(3400,3600]	0	0%	140	99.3%
(3600,3800]	1	0.7%	141	100%

The *class frequencies* are the counts of the number of sample values in each of the successive classes, 0-200, 200-400, 400-600, ... Thus a *class* is a range of values of the variable being investigated. The class boundaries separate successive classes, and the class interval is the length of the range of values in each class. The class frequencies are obtained by taking the differences of successive cumulative frequencies. So, for example, the frequency 19 for the class 600-800 is given by 19 = 116 - 97. *Relative frequencies* are obtained by dividing class frequencies by the sample size. They are the proportion of the sample lying in each class and have been expressed as percentages to one decimal place of accuracy.

The function **table** builds a contingency table of the number of observations belonging to each class.

```
freq <- table(x_group)
```

The function **cumsum** provides the cumulative sum, so we are able to calculate the cumulative frequencies:

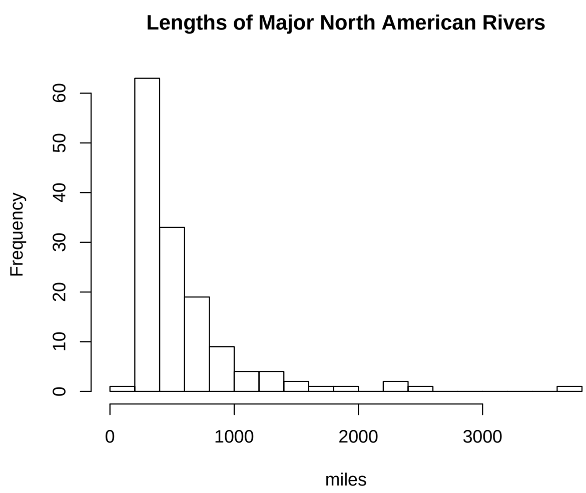
```
cum_freq <- cumsum(freq)
```

In order to calculate the relative frequencies we have to divide the frequencies for the number of observations:

```
rel_freq <- freq / n * 100
rel_cum_freq <- cum_freq / n * 100
```

The frequency distribution may be illustrated by a *histogram*, which is a plot of frequency (or relative frequency) against class, where the frequencies are represented by rectangles.

```
hist(x, breaks = cutpoint, xlab = "miles",
     main = "Lengths of Major North American Rivers")
```



- You may leave the choice of the class boundaries up to `R`, without specifying the values in `breaks`. If in `breaks` is specified as a single integer, in that case `R` uses it as the number of classes.
- In a histogram it is the area of each rectangle that should be proportional to the corresponding class frequency. If the class intervals vary from class to class then the lengths of the bases of the rectangles are proportional to the class intervals and the heights of the rectangles are not then proportional to the class frequencies. **Warning!** `R` produces histogram of the distribution of the density instead of the frequency if the classes vary. So it is safer to stick to equal class intervals.

From the histogram we obtain a visual impression of the distribution. We may imagine drawing a smooth curve that approximates the shape of the histogram. In our case, the distribution is not symmetrical. It is *positively skewed*, i.e., *skewed to the right*. If the skewness was in the opposite direction then we would say that the distribution was *negatively skewed* or *skewed to the left*.

Recall that the sample median is 425 and the sample mean is 591.18. It is a consequence of the positive skewness of the distribution that the sample mean is substantially greater than the sample median. The large sample values in the right hand tail of the distribution inflate the sample mean. For symmetric distributions the mean and median are approximately equal. If we wish to summarize the data in terms of a single number, a measure of location, that is in some sense a typical value, we might consider whether the median or the mean was the more appropriate number to use.

Consider a group of 5 employees whose salaries are 20K, 20K, 30K, 30K, 100K. Is the more typical value given by the mean, which is 40K, or by the median, which is 30K? There is no clear-cut answer, but it is important to be aware of the issue and of the fact that a single statistic gives very limited, and possibly misleading, information about the distribution.

1.5 Discrete distributions and bar charts

Consider the numbers of National Lottery jackpot winners at each draw for the period 2001-2008, where there are two draws weekly, on Wednesdays and Saturdays, 835 draws in total. The listing of the number of jackpot winners draw by draw is available in the file "LottoWinners.dat".

```r
winners <- read.table('LottoWinners.dat',
  header = FALSE,
  col.names = "winners")
```

These data take non-negative integer values only, and so are an example of *discrete data*. Continuous data are data that can take values in a continuous range, such as the length of the rivers in our earlier example, and so in principle can be recorded to any degree of accuracy, any number of decimal places. Discrete data are data that can take only a discrete set of values. Typically they represent counts and take non-negative integer values, or are restricted to a specified finite subset of the non-negative integers.

We have created a variable in `R` labelled `winners`, for the numbers of jackpot winners at each successive draw. As in the earlier example, we can generate descriptive statistics, as shown in the output below.

```
nrow(winners) # number of observations

## [1] 835

sum(is.na(winners)) # number of missing values

## [1] 0

summary(winners) # basic summary statistics

##     winners
##  Min.   : 0.000
##  1st Qu.: 1.000
##  Median : 2.000
##  Mean   : 2.143
##  3rd Qu.: 3.000
##  Max.   :15.000
```

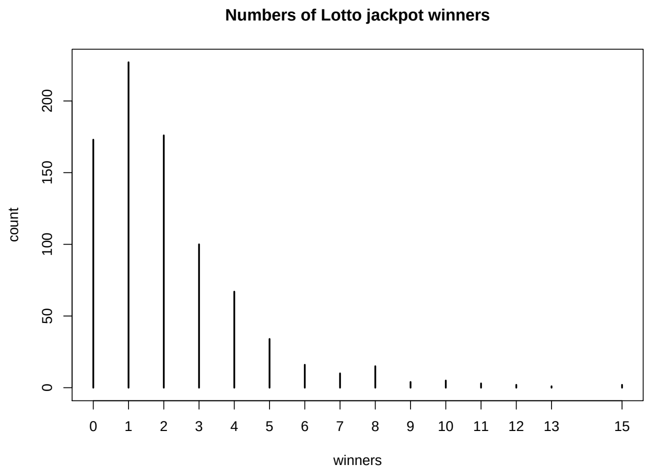
For discrete data, we can generate a *discrete frequency distribution* by using the function `table`. This distribution shows the frequencies, i.e., the counts of the numbers of occurrences of each number of jackpot winners, 0, 1, 2, ...

```
t_win <- table(winners)
t_win

## winners
##   0   1   2   3   4   5   6   7   8   9  10  11  12  13  15
## 173 227 176 100  67  34  16  10  15   4   5   3   2   1   2
```

We illustrate a discrete frequency distribution by a *bar chart*, which may be obtained by using the `plot` function. The `plot` function in `R` automatically does the plot that is appropriate for the contingency table obtained by the command `table`.

```r
plot(t_win,
     ylab = "count",
     main = "Numbers of Lotto jackpot winners")
```



The descriptive statistics show us that, over the period 2001-08, the mean number of jackpot winners per draw is 2.143, with the number of winners in any one draw ranging from 0 to 15. The lower quartile, median and upper quartile are 1, 2 and 3, respectively. The frequency distribution, gives a more comprehensive description of how the numbers of jackpot winners vary from draw to draw, and this is illustrated by the bar chart.

- The bars of the bar chart are separated from each other by spaces in between, in contrast to the rectangles of the histogram, which were adjacent to each other without any space in between. This emphasizes that the bar chart illustrates discrete data, whereas the histogram illustrates continuous data.