3
Data Summaries

In this chapter we explore various options within R for summarising data. We focus primarily on some intrinsic functions within R to obtain numerical summaries of average and spread; we also consider some of the most commonly-used graphical procedures within R for summarising data.

3.1 Numerical summaries

3.1.1 Measures of average

One of the most important and widely used measures of average is the (arithmetic) sample mean:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \, .$$

So if our data set was $\{0, 3, 2, 0\}$, then $n = 4$. Hence,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{0+3+2+0}{4} = 1.25 \, .$$

As you will see in the following examples, the mean can be distorted if our sample contains unusually large or small values relative to most of the other values in our dataset.

The sample median can overcome the sensitivity of the mean to outliers, and so is often considered a more "robust" measure of average. It is the middle observation when the data are ranked in ascending order, and so the size of the largest or smallest values is not taken into account - only the *position* of these values is considered. Denote the ranked observations as $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$. The sample median is defined as:

$$\text{Sample median} = \begin{cases} x_{(n+1)/2}, & n \text{ odd;} \\ \frac{1}{2}x_{(n/2)} + \frac{1}{2}x_{(n/2+1)}, & n \text{ even} \, . \end{cases}$$

Remember that $x_{(n+1)/2}$ is the $(n+1)/2^{\text{th}}$ ordered observation.

Although the median is more robust than the sample mean, you will see in other courses that it has less useful mathematical properties.

For our simple data set $\{0, 3, 2, 0\}$, to calculate the median we re-order

it to: $\{0,0,2,3\}$; then take the average of the middle two observations, to get 1.

The sample mode is the value which occurs with the greatest frequency. It only makes sense to calculate or use it with discrete data or qualitative data.

Warning: in R the function `mode` doesn't give you the sample mode. Use `table` instead.

3.1.2 Measures of spread

As well as knowing the location statistics of a data set, we also need to know how variable or "spread-out" our data are.

The range is easy to calculate. It is simply the largest minus the smallest[1]. In other words,

$$\text{Range } = x_{(n)} - x_{(1)} \, .$$

[1] When you get a new data set, calculating the range is useful when checking for obvious data-inputting errors.

So for our data set of $\{0,3,2,0\}$, the range is $3-0=3$. It is very useful for data checking purposes, but in general it's not very robust.

Obviously, the range can be distorted by outliers or extreme observations.

In statistics, the mean and variance are used most often. This is mainly because they have nice mathematical properties, unlike the median, say.

THE SAMPLE VARIANCE, s^2, is defined as

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \\ &= \frac{1}{(n-1)}\left\{\left(\sum_{i=1}^{n}x_i^2\right) - n\bar{x}^2\right\} \, . \end{aligned}$$

The second formula is often preferred for calculations. So for our data set $\{0,3,2,0\}$, we have

$$\sum_{i=1}^{4} x_i^2 = 0^2+3^2+2^2+0^2 = 13,$$

The divisor is $n-1$ rather than n in order to correct for the bias which occurs because we are measuring deviations from the sample mean rather than the "true" mean of the population we are sampling from - more on this in Stage 2 Statistics courses.

giving:

$$s^2 = \frac{1}{n-1}\left\{\left(\sum_{i=1}^{n}x_i^2\right) - n\bar{x}^2\right\} = \frac{1}{3}\left(13 - 4\times 1.25^2\right) = 2.25.$$

The sample standard deviation, s, is the square root of the sample variance, i.e. for our toy example $s=\sqrt{2.25}=1.5$.

The standard deviation is preferred as a summary measure as it is in the units of the original data. However, it is often easier from a theoretical perspective to work with variances.

The upper and lower quartiles are defined as follows:

- $Q1$: Lower quartile = $(n+1)/4^{\text{th}}$ smallest observation.
- $Q3$: Upper quartile = 3 $(n+1)/4^{\text{th}}$ smallest observation.

We can linearly interpolate between adjacent observations if necessary. The interquartile range is the difference between the third and first quartile, i.e. Then:

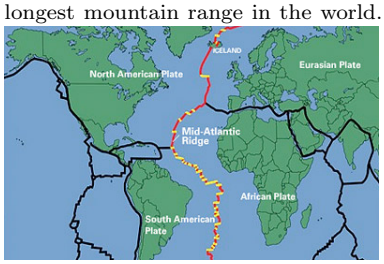
$$IQR = Q3 - Q1.$$

3.1.3 Example

The number of earth tremours recorded for five randomly chosen locations in Iceland, close to the *Mid-Atlantic ridge*[2], is recorded below:

Location	1	2	3	4	5
# tremours	7	1	14	13	20

[2] The Mid-Atlantic Ridge is a divergent tectonic plate along the floor of the Atlantic Ocean, and is part of the longest mountain range in the world.



- Calculate the mean and median number of earth tremours for this sample.

  Solution:

  The mean is:

$$\bar{x} = \frac{1}{5}(1+7+13+14+20) = 11$$

  The median is the middle value, so we get 13.

- There was a recording error for location 5, which is closest to the Mid-Atlantic Ridge; this observation *should have* been 200, and not 20. Find new values for the mean and median, and comment.

  Solution:

The mean can be distorted by unusually high or low values.

- Find the standard deviation and IQR for the earth tremours data, and comment.

  Solution:

3.2 Using R

R has many intrinsic (built-in) functions for basic statistical analyses. This includes the calculation of simple summary statistics. For example, for the movie data in section §1.3, we can easily use R to calculate averages and measures of spread. First, load the data:

```
> library(ncldata)
> data(movies)
> attach(movies)
```

To calculate averages, we use `mean` and `median`[3]:

```
> mean(Budget, na.rm = TRUE)
[1] 41245023
> median(Budget, na.rm = TRUE)
[1] 2.5e+07
```

[3] The argument `na.rm` can be invoked to remove missing values. Notice that the mean and median for movie budgets are substantially different... why?

We can also calculate measures of spread very easily[4]:

```
> range(Budget, na.rm = TRUE)
[1] 0.0e+00 1.9e+09
> var(Budget, na.rm = TRUE)
[1] 4.707623e+15
> sd(Budget, na.rm = TRUE)
[1] 68612118
```

[4] What do you think the `e` notation means? For example, `4.707623e+15` as in the output above?

To get the quartiles from R we use the `quantile` command[5], i.e.:

[5] Note the default in R gives you a slightly different quartile range, i.e. if you don't enter the `type=6` argument. As $n \to \infty$, the different quartile func-

```
> quantile(Budget, na.rm = TRUE, type = 6)
     0%     25%     50%     75%    100%
0.0e+00 9.0e+06 2.5e+07 5.0e+07 1.9e+09
> summary(Budget)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
0.000e+00 9.000e+06 2.500e+07 4.125e+07 5.000e+07 1.900e+09       998
```

Table 3.1 summarises the most useful intrinsic R commands for producing numerical summary statistics.

Command	Comment	Example
`mean`	Calculates the mean of a vector	`mean(x)`
`sd`	Calculates the standard deviation of a vector	`sd(x)`
`var`	Calculates the variance of a vector	`var(x)`
`quantile`	The vector quantiles. Make sure you use `type=6`	`quantile(x, type=6)`
`range`	Calculates the vector range	`range(x)`
`summary`	Calculates various statistics, including the quartiles. **However,** it doesn't use `type=6` quartiles.	`summary(x)`

Table 3.1: Summary of R commands in this chapter.

3.3 Graphical summaries

3.3.1 Introduction

Graphical displays of data can be very useful for showing the main features of a data set. The appropriate form of graph depends on the nature of the variables being displayed and what aspects are to be shown. However it should always be borne in mind that the object is to provide a clear and truthful representation of the data, not to distort and not to impress with unnecessary "fancy" features.

3.3.2 Qualitative data: bar charts

The most useful way to display qualitative data is usually with a bar chart. The length of each bar is proportional to the frequency of the corresponding value of the variable in the sample of data. Note that the widths of the bars should be equal to avoid giving a false impression, as should the width *between* bars[6].

Figure 3.1 shows the breakdown in MPAA ratings for the movies dataset. To create a bar chart in R we use the `table` command...

```
> table(movies$mpaa)
NC-17    PG PG-13     R
   16   526   989  3316
```

... inside the `barplot` function:

```
> barplot(table(movies$mpaa), xlab="MPAA Rating",
+         ylab="Frequency", border = "black",
+         col="mistyrose")
```

Remember to load the data first!

[6] Type `colours()` into R to get a list of colours.

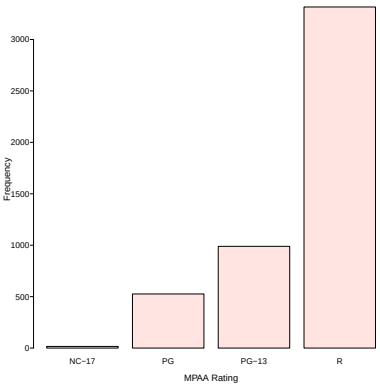


Figure 3.1: Barchart of the mpaa ratings for 4847 films.

3.3.3 Histograms

To represent the distribution of a sample of values from a continuous variable we can use a histogram. The range of values of the variable is divided into intervals, known as *classes* or *bins*, and the frequencies in bins are represented by columns. As the variable is continuous, there are no gaps between neighbouring columns, unlike a bar chart[7]. Note also that, strictly speaking, it is the *area* of the column which is proportional to the frequency, not the height. The reason for this is that columns need not be of the same width. Computer software tends to use bins of the same width - giving standard "frequency histograms" when the y-axis represents frequency[8]. However, this default can be overridden in R if you really want to do this.

[7] Unless, of course, a particular bin has zero frequency

[8] Rather than frequency *density* - important when the classes are *not* all the same width.

Rule	Formula	R command	Comment
Sturges'	$k_{ST} = \lceil \log_2 n + 1 \rceil$	Default	Tends not to be very good for $n > 30$.
Scott's	$h_{SC} = 3.49 \times s \times n^{-1/3}$	`breaks = 'Scott'`	
Freedman-Diaconis	$h_{FR} = 2 \times IQR(x) \times n^{-1/3}$	`breaks = 'FD'`	When the distribution is symmetric, this is very similar to Scott's rule.

Table 3.2: Standard bin width rules in R.

Figure 3.2 shows histograms of the film budgets. When displaying relative frequencies (or proportions; often termed "density histograms"), we can easily work out the height of each bin using:

$$\text{Height} = \frac{\text{frequency}}{n \times \text{ Bin-width}}.$$

When we display relative frequencies, the y-axis is labelled with "density" and the area under the histogram is one. Bin widths should be chosen so that you get a good idea of the distribution of the data, without being swamped by random variation.

To generate Figure 3.2 in R we use the following commands; notice the argument `freq=FALSE` to switch from frequencies to relative frequencies.

```
> hist(movies$Budget, col="grey",
+      main="Mean film budget",
+      freq=FALSE, xlab="Budget ($)")
```

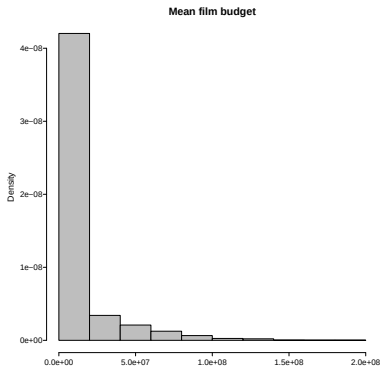


Figure 3.2: Histogram of movie budgets.

How many bins should we have?

Let:

- n: the sample size;
- k: the number of bins in the histogram;
- h: the bin-width.

Then the number of bins we use to construct a histogram is:

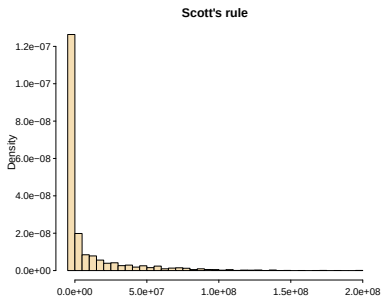
$$k = \left\lceil \frac{\max(x) - \min(x)}{h} \right\rceil \tag{3.1}$$

where $\lceil \cdot \rceil$ is the ceiling function.

Table 3.2 gives a summary of the different rules. Briefly, R uses Sturges' rule by default, which isn't always that good. Notice that Sturges' rule gives you k, but the other rules give you the bin width. Typically, it is best to go with Scotts' rule or the Freedman-Diaconis rule. You should experiment with the different rules in R to find the histogram which you think best displays your data!

To use the different rules, we use the breaks argument. For example, the following piece of code:

```
> hist(movies$Budget, col="wheat",
+      main="The FD rule",
+      freq=FALSE, xlab="Budget ($)", breaks="FD")
```



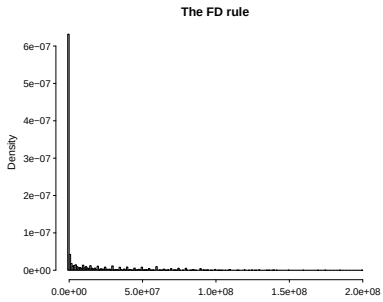


Figure 3.3: Histogram of (a) movie budgets using Scott's rule and (b) the Freedman-Diaconis rule. Compare these histograms to Figure 3.2 which uses Sturges' rule (default).

uses the Freedman-Diaconis rule - see Figure 3.3b. When we compare Figure 3.3a with Figure 3.2, we have used many more bins, which we might argue has resulted in a *better* histogram; however, it is clear that Figure 3.3b uses too *many* bins! In practical 2 you will get a chance to experiment with the different rules yourself.

3.3.4 Box and whisker plots

A box and whisker plot, often referred to simply as a boxplot, is another way to represent continuous data. This kind of plot is particularly useful for comparing two or more groups, by placing the boxplots side-by-side. Figures 3.4 and 3.5 shows boxplots of film length for different categories of film.

The central bar in the "box" is the sample *median*. The top and bottom of the box represent the upper and lower sample *quartiles*, respectively. Just as the median represents the 50% point of the data, the lower and upper quartiles represent the 25% and 75% points respectively.

The lower whisker is drawn from the lower end of the box to the smallest value that is no smaller than $1.5IQR$ below the lower quartile. Similarly, the upper whisker is drawn from the middle of the upper end of the box to the largest value that is no larger than $1.5IQR$ above the upper quantile. Points outside the whiskers are classified as outliers.

To do this in R we use the following commands:

```
> par(mfrow=c(2, 1))
> boxplot(movies$Length, ylab="Film length",
+         col="bisque")
> boxplot(movies$Length ~ movies$mpaa, ylab="Film length",
+         col="bisque")
> boxplot(movies$Length ~ movies$mpaa + movies$Romance,
+         ylab="Film length")
```

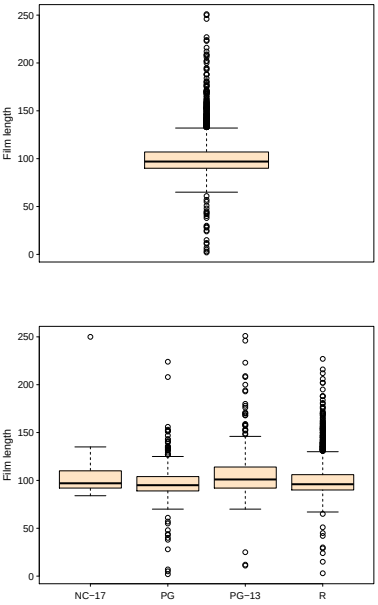


Figure 3.4: Box and whisker plots of (a) film length (b) film length split according to the mpaa rating.

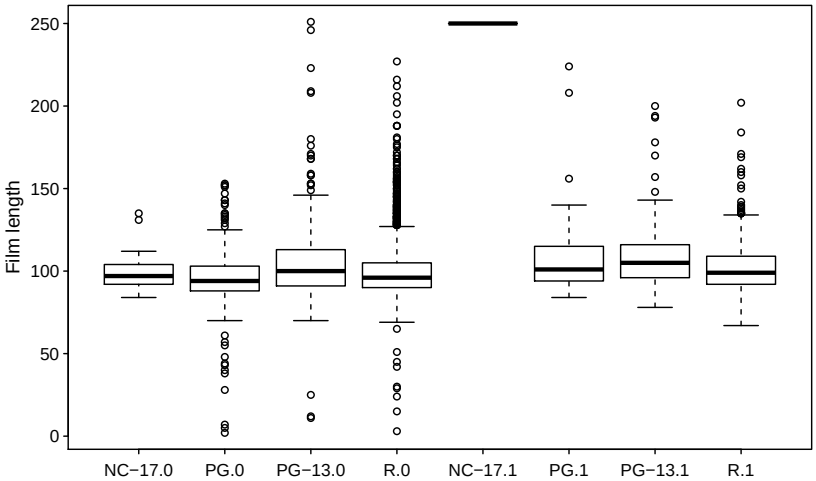


Figure 3.5: Box and whisker plots for movie length split according to MPAA rating and whether the film was a romance.

Command	Comment	Example
`table`	Contingency table	`table(x)`
`barplot`	Generate a bar chart	`barplot(table(x))`
`hist`	Histogram	`hist`
`plot`	Scatter plot. See Practical 2.	`plot(x, y)`
`points`	Add points to a plot. See Practical 2.	`points(x,y)`
`lines`	Add lines to a plot. See Practical 2.	`lines(x,y)`
`boxplot`	Box and whiskers plot	`boxplot(x)`

Table 3.3: Summary of R commands in this chapter.