2 Summary statistics

Very often we want to report a few key descriptors of a data set, rather than specifying every observation. This is done with *summary statistics*. There are two basic types of summary statistic:

- Measures of location: a quantity which is "typical" of the data or "central" in the data;
- Measures of spread: quantifies the variability in the data.

2.1 Notation

Sample size is denoted n and data are denoted $x_1, x_2, \ldots, x_n$. For example, we ask four people how many siblings they have and get: 0, 3, 2, 0. Then

$$n = 4 \quad \text{and} \quad x_1 = 0,\ x_2 = 3,\ x_3\ = 2,\ x_4 = 0\ .$$

Sums are represented by the summation symbol, a capital Sigma ($\sum$). For example,

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{4} x_i = x_1 + x_2 + x_3 + x_4 = 0 + 3 + 2 + 0 = 5$$

or

$$\sum_{i=3}^{4} x_i = x_3 + x_4 = 2 + 0 = 2\ .$$

For shorthand, sometimes we miss out the limits of the summation sign, e.g.

$$\sum x_i = \sum_{i=1}^{n} x_i\ .$$

Often we perform some operation on the data before summing, e.g. taking the square:

$$\sum_{i=1}^{n} x_i^2 = x_1^2 + x_2^2 + x_3^2 + x_4^2 = 0^2 + 3^2 + 2^2 + 0^2 = 13$$

Note in general:

$$\left(\sum_{i=1}^{n} x_i\right)^2 \neq \sum_{i=1}^{n} x_i^2\ .$$

Exercise:
Fill out the following table:

	Data	n	$\sum x_i$	$\sum x_i^2$
Data set 1	1,2,3			
Data set 2	0,0			
Data set 3	-1,0,1			

2.2 Measures of location

2.2.1 The sample mean

The most commonly used and generally most useful measure of location is the *sample mean*, denoted $\bar{x}$ (pronounced "x bar"). It is the *average observed value* and it is defined by:

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

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

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

Using R
For the leukaemia survival data, we can use R to calculate the mean:

```
url = "http://www.mas.ncl.ac.uk/~nseg4/teaching/MAS8380/survival.txt"
survival = read.table(url, header=TRUE)
mean(survival$Time) # Gives 925.1163
```

2.2.2 Sample median

The sample median is the middle observation when the data is ranked in increasing order. Denote the ranked observations $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$. (Note the brackets in the subscripts.) If there are an even number of observations the median is defined to be the sample mean of the middle two observations

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

The sample median is more robust against extreme observations than the sample mean, but has less useful mathematical properties. Example: if our data set was $\{0, 3, 2, 0\}$, then we re-order it to: $\{0, 0, 2, 3\}$ and take the average of the middle two observations to get 1.

For the leukaemia survival data, we can obtain the median using the following R command:

```
median(survival$Time) # Gives 702.
```

Other examples:

```
# Try these too:
x = c(1,2,3,4)
median(x) # Gives 2.5.
x = c(1,2,3,1000)
median(x) # Gives 2.5 too!
```

Notice that the mean and median are significantly different in this last case.

2.2.3 Sample Mode

The mode is the value which occurs with the greatest frequency. This only makes sense with discrete data, so it is not as useful as the other measures of location.

Exercise:
For the following data calculate the mean and median for each row: ✎

Data					n	$\sum x_i$	$\bar{x}$	Median
1	7	13	14	20				
7	13	14	20					
1	7	13	14	200				

2.3 Measures of spread

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

2.3.1 Range

The range is the difference between the largest and smallest observations.

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

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

```
range(survival$Time) # Gives 7 2509, i.e. the smallest and largest values.
minmax = range(survival$Time)
minmax[2] - minmax[1] # Gives 2502, the required range.
```

2.3.2 Sample variance and standard deviation

The sample variance, s^2 is defined as

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

The second formula is easier for calculations. 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 later. The sample standard deviation is often denoted σ_{n-1} on calculators.

For our data set, $\{0, 3, 2, 0\}$, we have

$$\sum x_i^2 = 0^2 + 3^2 + 2^2 + 0^2 = 13$$

so,

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

The sample standard deviation, s, is the square root of the sample variance, i.e. for our toy example $s = \sqrt{2.25} = 1.5$. It 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. A quick method for approximately calculating s is to note that the entire range of data is usually covered by $4 \times s$. So

$$s \simeq \frac{\text{Range}}{4}$$

which for our data would give 0.75 (a bit small, but not bad for a first guess).

Note:

- Adding a constant to all values doesn't change the sample variance (or sample standard deviation).
- Multiplying by a constant changes the sample variance by multiplying by the square of the constant (and the sample standard deviation by the absolute value).

Again R can calculate these quantities very easily. For the survival data:

```
var(survival$Time) # Gives 494568.
sd(survival$Time)  # Gives 703.2553.
sqrt(var(survival$Time)) # Gives 703.2553.

Time1 = survival$Time + 20 # Adds 20 to every value.
var(Time1) # Gives 494568.

Time2 = survival$Time*20 # Multiplies every value by 20.
var(Time2) # Gives 197827185 = var(survival$Time)*20^2.
```

2.3.3 Quartiles and the interquartile range

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

As for the median, we need to linearly interpolate between adjacent observations if $\frac{(n+1)}{4}$ or $\frac{3(n+1)}{4}$ are not whole numbers. For example, if there are $n = 10$ observations, $\frac{(n+1)}{4} = 2.75$ so the lower quartile is $x_{(2)} + 0.75 \times (x_{(3)} - x_{(2)})$.

The interquartile range (IQR) is the difference between the upper and lower quartiles. For our data set: $\{0, 3, 2, 0\}$ we re-order to get $\{0, 0, 2, 3\}$, so $\{\text{LQ, UQ}\} = \{0, 2.75\}$. As a rough approximation, the standard deviation is often roughly $0.75\times$ interquartile range.

To get the quartiles from R we use the `quantile` command. However, be **warned**: R can compute quartiles and quantiles in different ways from that described here! You need to include the argument `type=6`, in order to get an answer that matches up with these notes, as below:

```
> quantile(survival$Time, type=6)
0%  25%  50%  75% 100%
 7  440  702 1367 2509
> summary(survival$Time) # NB: quartiles are not computed in the same way!!!
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  7.0   442.5   702.0   925.1  1350.0  2509.0
```

Sometimes you might be asked for the "MQMQM" summary of some data: this means the minimum, LQ, median, UQ, and maximum.

2.3.4 Coefficient of variation

The coefficient of variation is defined as

$$\text{Coefficient of variation} = \frac{s}{\bar{x}}$$

This has no units and it does not change if data are rescaled.

Exercise:
For the following data calculate the Range, s, s^2 and IQR:

Data					$\sum x^2$	$n\bar{x}^2$	Range	s	s^2	IQR
1	7	13	14	20						
7	13	14	20							
1	7	13	14	200						