Statistics for Computing MA4413

Lecture 2

Numerical Summaries of Centrality and Dispersion and the Boxplot

Kevin Burke

kevin.burke@ul.ie

Numerical Summaries

We focus here on *numerical data* only. We have seen how the frequency table and corresponding histogram describe the whole distribution of data.

Often however, we would like to summarise the **main features** of the distribution without using the whole frequency table, i.e., a few numbers - *"numerical summaries"* - which provide the relevant info.

We will look at measures of:

- **Centrality** - a numeric value indicating the centre of the distribution, i.e., an "average" or "typical" value.
- **Dispersion** - a numeric value indicating the degree to which measurements *vary* about this centre, i.e., is the distribution of values tightly packed around its centre or not?

Numerical Summaries

- **Centrality**
  - mean: arithmetic average (you all know this).
  - median: middle number in the *ordered* data.
- **Dispersion**
  - range: $\max(x) - \min(x)$.
  - variance: measure of variation around the *mean*.
  - standard deviation: square root of the variance.
  - quartiles: *three* numbers - Q_1,Q_2 and Q_3 - which split the *ordered* data into four parts (note: $Q_2 =$ the median).
  - inter-quartile range: $IQR = Q_3 - Q_1$.

We also introduce the **boxplot** - a graphical method for numerical data. This could have gone into the “Visualising Numerical Data” section of Lecture 1 but we need Q_1,Q_2, Q_3 and IQR to draw it.

The Mean

The mean is just the usual arithmetic average: add all of the individual data values in the sample and divide by the number of values.

Remember that ***n*** is the symbol for the *sample size*, i.e., the number of values. The sample mean is:

$$\boxed{\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum x_i}{n}}.$$

We have introduced the *sum notation* - don't be put off by this! $\sum$ just means "the sum of" and x_i means "individual value". So $\sum x_i$ means "the sum of all values".

Remember that we have seen the symbol $\bar{x}$ before. Also, recall that it is a *statistic* which *estimates* the population mean μ *(parameter)*.

The Mean: Example

Let's say we have the annual income (in thousands) of $n = 5$ individuals living in a particular apartment block:

25	29	33	35	40

The average income is

$$\bar{x} = \frac{25 + 29 + 33 + 35 + 40}{5} = \frac{162}{5}$$
$$= 32.4$$
$$= €32,400.$$

The Mean - Skewed Data

An issue with the mean is its sensitivity to *outliers* - data values much larger / smaller than the main body of data - which lead to *skewness* (remember: data can be skewed to the right / left).

Let's now assume that the 5th individual is *much* wealthier than the others:

25	29	33	35	500

The average income is

$$\bar{x} = \frac{25 + 29 + 33 + 35 + 500}{5} = \frac{622}{5}$$
$$= 124.4$$
$$= €124,400.$$

The Mean - Skewed Data

It is clear that €124,400 is not a good representation of the centre of the income distribution. It is not a typical income for an individual living in that apartment block.

The mean gets *pulled towards* the outliers, i.e., it is pulled away from the centre in the direction of the skew.

- Data skewed to the right (caused by large values)
  $\Rightarrow$ the mean gets pulled towards the right.

- Data skewed to the left (caused by small values)
  $\Rightarrow$ the mean gets pulled towards the left.

The Median

The median, Q_2, is the value that splits the *ordered* data in half: 50% of the data lies above Q_2 and 50% of the data lies below it.
(Note: we call the median Q_2 as it is the second quartile - more on this later)

To find Q_2:

1. Put the data *in order* - smallest to largest - if it is not already.
2. Q_2 is then the value in *position* $\boxed{\dfrac{n+1}{2}}$ where n is the sample size.

Important: $\frac{n+1}{2}$ is not the value of the median, it is its *position* in the ordered data.

The Median: Example

	Position				
	1	*2*	*3*	*4*	*5*
Symmetrical:	25	29	33	35	40
Skewed to the Right:	25	29	33	35	500

The *position* of the median is:

$$\frac{n+1}{2} = \frac{5+1}{2} = \frac{6}{2} = 3,$$

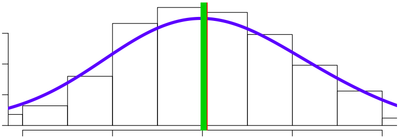
i.e., the 3rd number.

$\Rightarrow$ The *value* is: $Q_2 = 33 = €\,33,000$.

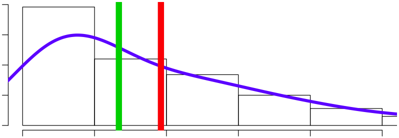
The median is unaffected by skewness - it still gives an accurate estimate of the centre.

Mean Vs Median

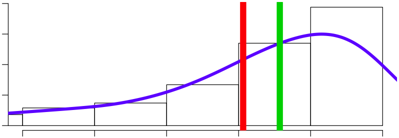
- Symmetrical data: $\bar{x} \approx Q_2$.



- Right-skewed: $\bar{x} > Q_2$.



- Left-skewed: $\bar{x} < Q_2$.



The Median: There's No Middle of 6 Numbers!

Let's say we have 6 numbers - what is the median?

Position

1	*2*	*3*	*4*	*5*	*6*
10	13	15	21	32	42

The *position* of the median is:

$$\frac{n+1}{2} = \frac{6+1}{2} = \frac{7}{2} = 3.5,$$

i.e., the median lies *between* the 3rd and 4th numbers.

$\Rightarrow$ Its *value* is simply the *average* of the numbers in position 3 and 4:

$$Q_2 = \tfrac{15+21}{2} = \tfrac{36}{2} = 18.$$

Question 1

Consider the following *sample* of numbers:

2	4	2	1	5	3	0	4	1	8

a) What is the value of n?

b) Calculate the mean and use the appropriate symbol.

c) What is the symbol for the population mean? What is its value?

d) Calculate Q_2 (hint: need to order the data first).

e) Construct a frequency table with 3 classes (let zero be the first breakpoint).

f) Draw the corresponding histogram.

Dispersion

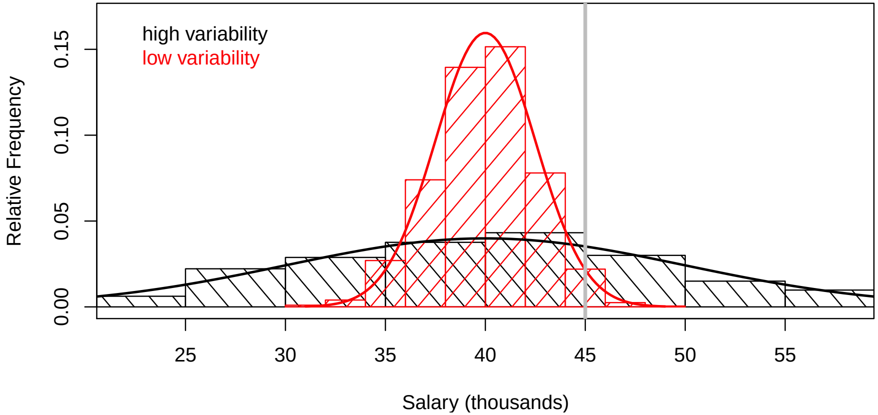
The centre of the distribution is only part of the story. We also need to know how spread out - *dispersed* - the data values are.

Consider the following:

A software engineer is offered a job with an annual salary of €45,000. The employer says that this is a very attractive salary as it is above the average for this type of job (€40,000).

Is this a good offer?... We don't know. Not without knowing how much the data *varies* about the central value.

Dispersion



If the distribution of salaries is highly variable, there are many posts available with a better salary. On the other hand, if variability is very low, we have been offered one of the highest salaries in the field.

The Range

The most basic measure of dispersion is the **range** of the data.

$$\boxed{\text{range} = \max(x) - \min(x)},$$

i.e., the largest value minus the smallest value in the set of data.

Disadvantage: It only tells us the *overall spread* of the data. But what we really want to know is how the data varies about its centre.

We mainly focus on other techniques (standard deviation and inter-quartile range).

The Variance

The **variance** is the *average squared distance from the mean* and is given by:

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}$$

$$= \frac{\sum (x_i - \bar{x})^2}{n-1}$$

In words, subtract the mean from each value, square the results and then add them all together. Finally, divide by $n - 1$.

(For technical reasons, in the case of variance, we divide by $n - 1$ rather than n)

The units of variance are *squared-units*, for example, if we were looking at income (in euro) then the variance would be in euros-squared.

The Variance

It turns out that the previous formula can be rewritten as:

(if you're good with sums, i.e., Σ, then you can show this)

$$\boxed{s^2 = \frac{\sum x_i^2 - n\,\bar{x}^2}{n-1}}$$

This version of the formula involves less computation so we will use it.

The Standard Deviation

The **standard deviation** is a *very* important quantity in statistics (as we will see later in this course).

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

$$\boxed{s = \sqrt{s^2} = \sqrt{\frac{\sum x_i^2 - n\,\bar{x}^2}{n-1}}}.$$

Since the variance is in squared-units, the standard deviation has *the same units as the data* (as a result of taking the square root).

The Standard Deviation: Example

We return to our earlier example - the incomes of 5 individuals.

						$\sum$
x_i	25	29	33	35	40	162
x_i^2	625	841	1089	1225	1600	5380

Using the above and the mean value, $\bar{x} = \frac{162}{5} = 32.4$, we then calculate the variance:

$$s^2 = \frac{\sum x_i^2 - n\,\bar{x}^2}{n-1} = \frac{5380 - 5\,(32.4^2)}{5-1} = \frac{5380 - 5\,(1049.76)}{4}$$

$$= \frac{5380 - 524.8}{4}$$

$$= \frac{131.2}{4} = 32.8 = 32,800\,€^2.$$

The Standard Deviation: Example

The standard deviation is then

$$s = \sqrt{s^2} = \sqrt{32.8} = 5.727 = 5,727\,€.$$

Note the units are euros.

Symbols

It is worth preparing ourselves for things to come:

- For the *sample standard deviation* we use the symbol s as shown.
- For the true *population standard deviation* we use σ (the Greek letter “sigma”).

Naturally we have s^2 and σ^2 for the sample variance and population variance.

Don’t forget, sample *statistics* estimate the true population *parameters*.

Important Note on Dispersion Measures

Variance and standard deviation are ***always positive numbers***.

In fact *all* measures of dispersion are positive numbers.

Consider the following four numbers: $\boxed{-10,\ -9,\ -5,\ -4}$.

Clearly the centre is negative; however, standard deviation will not be. Show that $\bar{x} = -7$ and $s = 2.94$.

Question 2

25 individuals were asked how long their laptop lasts on a full charge. The recorded times (measured in hours) are as follows:

(we saw this dataset before - lecture 1, Q5)

2.2	0.4	4.2	12.9	1.5	3.0	5.7	0.7	1.0	3.3
0.2	0.2	5.6	1.6	3.0	0.1	14.3	3.4	0.9	6.1
1.4	1.0	0.7	5.4	2.3					

a) Calculate $\bar{x}$.

b) Calculate s.

c) What is the symbol for the population standard deviation? What is our best estimate of this?

d) What is the value of μ?

e) What is the value of $\hat{p}$?

The Standard Deviation - Skewed Data

Recall that $\bar{x}$ is *not* a good measure of centrality when data is skewed (use the median, Q_2, instead).

If this is the case, we are then not interested in s either since this measures the dispersion about $\bar{x}$.

So what goes with the median? - The *inter-quartile range*.

Quartiles

There are **three quartiles** which split the (ordered!) data into four parts:

$$25\% - \mathbf{Q_1} - 25\% - \mathbf{Q_2} - 25\% - \mathbf{Q_3} - 25\%$$

The process of finding the quartiles is essentially the same as the case of finding the median (i.e., the second quartile Q_2).

1. Put the data *in order* - smallest to largest.
2. The *position* of Q_k (quartile number k) is: $\boxed{\frac{n+1}{4} \times k}$.

Q_1 is in position $\frac{n+1}{4}$, Q_2 is in position $\frac{n+1}{4} \times 2$ and Q_3 is in position $\frac{n+1}{4} \times 3$.

Inter-Quartile Range

The **inter-quartile range** is the range of the middle 50% of data.

Calculation of IQR is straightforward once we have the quartiles:

$$\boxed{IQR = Q_3 - Q_1},$$

i.e., it is simply the difference between the upper and lower quartiles.

Quartiles and IQR: Example

Consider the following sample of $n = 10$ values:

2	4	2	1	5	3	0	4	1	8

First we must *sort* the values. The reordered dataset is:

Positions:	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Values:	0	1	1	2	2	3	4	4	5	8

Quartile	**Position**	**Value**
Q_1	$\frac{10+1}{4} = \frac{11}{4} = 2.75 \Rightarrow$ between 2 & 3	$\frac{1+1}{2} = \mathbf{1}$
Q_2	$\frac{11}{4} \times 2 = 2.75 \times 2 = 5.5 \Rightarrow$ between 5 & 6	$\frac{2+3}{2} = \mathbf{2.5}$
Q_3	$\frac{11}{4} \times 3 = 2.75 \times 3 = 8.25 \Rightarrow$ between 8 & 9	$\frac{4+5}{2} = \mathbf{4.5}$

$\Rightarrow$ **IQR** $= Q_3 - Q_1 = 4.5 - 1 = \mathbf{3.5}$.

(3.5 units covers the middle 50% of data)

Question 3

We return to the laptop battery life data:

2.2	0.4	4.2	12.9	1.5	3.0	5.7	0.7	1.0	3.3
0.2	0.2	5.6	1.6	3.0	0.1	14.3	3.4	0.9	6.1
1.4	1.0	0.7	5.4	2.3					

a) What is the value of n ?

b) Find the values of the quartiles.

c) Calculate IQR.

d) Calculate $\bar{x}$ and compare this to Q_2. Is the data skewed? If so, in what direction?

Boxplot

A **boxplot** is a graph containing the following items:

1. Quartiles: Q_1, Q_2 and Q_3.
2. Mimimum/maximum values *not classed as outliers*.
3. Outliers (values much smaller/larger than the main body of data).

We know how to get quartiles. All we need to know is how to classify data as being outliers.

Outlier Detection

To find outliers we first calculate the **lower fence** and **upper fence**:

$$\boxed{\begin{aligned} LF &= Q_1 - 1.5 \times IQR \\ UF &= Q_3 + 1.5 \times IQR \end{aligned}}$$

- **Outliers** are then:
  - Values smaller than LF.
  - Values greater than UF.

Outlier Detection: Example

Let's look at the laptop battery data. In Question 3 we should have found that $Q_1 = 0.8$, $Q_2 = 2.2$, $Q_3 = 4.8$ and $IQR = 4$.

So we have

$$\begin{aligned} LF &= Q_1 - 1.5 \times IQR \\ &= 0.8 - 1.5 \times 4 = -5.2. \end{aligned}$$

$$\begin{aligned} UF &= Q_3 + 1.5 \times IQR \\ &= 4.8 + 1.5 \times 4 = \quad 10.8. \end{aligned}$$

Any value in the data less than -5.2 or greater than 10.8 is classed as an outlier.

Outlier Detection: Example

Looking at the *ordered* data:

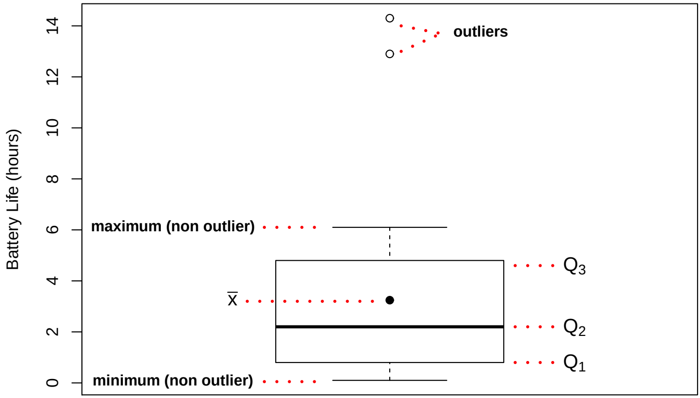
0.1	0.2	0.2	0.4	0.7	0.7	0.9	1.0	1.0	1.4
1.5	1.6	2.2	2.3	3.0	3.0	3.3	3.4	4.2	5.4
5.6	5.7	6.1	12.9	14.3					

- Values less than $LF = -5.2$: **none**.
- Values greater than $UF = 10.8$: **12.9 and 14.3**.

- Minimum of non-outliers: **0.1**.
- Maximum of non outliers: **6.1**.

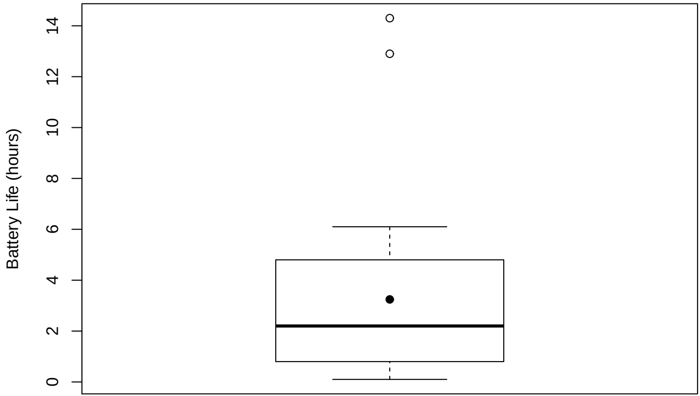
We can now draw the boxplot.

Boxplot: Example



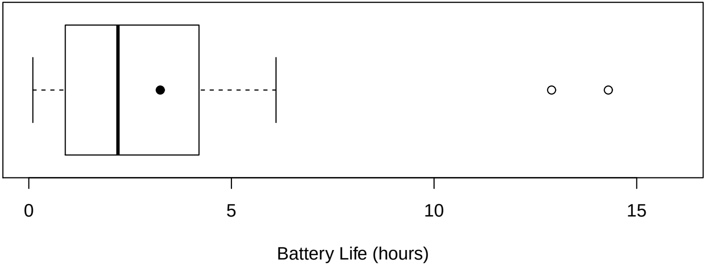
- Labelled boxplot. Note - it is also useful to include $\bar{x}$.

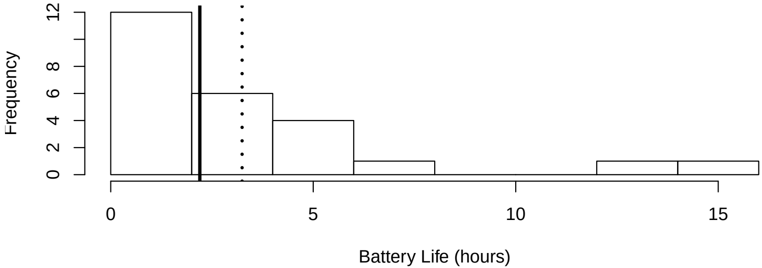
Boxplot: Example



- Boxplot without labels.

Boxplot Vs Histogram





Question 4

It turns out that the laptops can be split into two groups. The battery lives for each of the 25 laptops is shown below:

Type 1:	0.1	0.2	0.2	0.4	0.7	0.9	1.0	1.5	2.3	4.2	5.6
Type 2:	0.7	1.0	1.4	1.6	2.2	3.0	3.0	3.3	3.4	5.4	5.7
	6.1	12.9	14.3								

a) Calculate the means for the two groups: $\bar{x}_1$ and $\bar{x}_2$.

b) What are the values of n_1 and n_2?

c) Draw the boxplots for each group (side by side on the same graph) and comment.

d) Are there outliers in either group?

e) Are either of the distributions skewed? If so, in what direction?

Recap of Symbols

Firstly, the sample size is n. The other symbols are given below:

Quantity	Sample Statistic	Population Parameter
Proportion	$\hat{p}$	p
Mean	$\bar{x}$	μ
Variance	s^2	σ^2
Standard Deviation	s	σ
Quartiles	Q_1, Q_2, Q_3	—

(we did not assign symbols to population quartiles)

R Code: Centrality

The code used to calculate the mean and median for the income example is:

```
income = c(25, 29, 33, 35, 40)
mean(income)
median(income)
```

R Code: Dispersion

The code used to calculate the variance and standard deviation for the income example is:

```
income = c(25, 29, 33, 35, 40)
var(income)
sd(income)
```

Quartiles (as well as the minimum, maximum and mean) are given by the `summary` function:

```
x = c(2, 4, 2, 1, 5, 3, 0, 4, 1, 8)
summary(x)
```

R uses a slightly different method for calculating Q_1 and Q_3 to what we use in this course - so the results will be different to the lecture.

Finally IQR is found via `IQR(x)` - again different to the lecture.

R Code: Sort

Another useful function is the `sort` function which orders the data from smallest to largest.

```
laptop = c(2.2, 0.4, 4.2, 12.9, 1.5,
           3.0, 5.7, 0.7, 1.0, 3.3,
           0.2, 0.2, 5.6, 1.6, 3.0,
           0.1, 14.3, 3.4, 0.9, 6.1,
           1.4, 1.0, 0.7, 5.4, 2.3)
laptop = sort(laptop)
laptop
```

R Code: Boxplot

Using the `laptop` data from the previous slide, we can draw a boxplot (and include the mean) using

```
boxplot(laptop, xlab="Battery Life (hours)")
points(x=1,y=mean(laptop),pch=20)
```

(Remember that R uses a different formula to get Q_1 and Q_3. So the boxplot will be slightly different to the one you do by hand)

A horizontal boxplot is given by

```
boxplot(laptop, horizontal=T,
        xlab="Battery Life (hours)")
points(x=mean(laptop),y=1,pch=20)
```

R Code: Two Boxplots

Two boxplots side by side with mean values shown:

```
laptop1 = c(0.1, 0.2, 0.2, 0.4, 0.7, 0.9, 1.0,
            1.5, 2.3, 4.2, 5.6)
laptop2 = c(0.7, 1.0, 1.4, 1.6, 2.2, 3.0, 3.0,
            3.3, 3.4, 5.4, 5.7, 6.1, 12.9, 14.3)

boxplot(laptop1,laptop2,
            xlab="Battery Life (hours)")
points(x=1,y=mean(laptop1),pch=20)
points(x=2,y=mean(laptop2),pch=20)
```