MAST7866
Numerical summaries of data

Jian Zhang

University of Kent

Measures of Location

- Data can be summarised using numerical measures such as the mean, median and mode.
- Consider the following data of the number of letters in a sample of 11 words from a page of text:

3 2 10 5 9 4 2 6 3 4 3

- With a small sample like this we could sort the data which helps:

2 2 3 3 3 4 4 5 6 9 10

Mean

- The sample mean is the average of the observed numbers.
- To calculate, we add all of the observations up and divide by the total number of observations.
- Mathematically, for a sample of size n, with observations $\{x_1,x_2,\ldots,x_n\}$ this is written as:

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

- The mean is a very common measure of location and one advantage of it is that every data value is taken into account.
- However, the mean is sensitive to outliers (extreme values) in the data.

Mean

- To calculate this in R:
  - Input the data:

```
> data<-c(3,2,10,5,9,4,2,6,3,4,3)
```

  - Calculate the mean:

```
> mean(data)
[1] 4.636364
```

Median

- The median is the middle value in a data set that has been ordered from smallest to largest.
- When data have been ordered in increasing order, the median lies in the $\frac{(n+1)}{2}^{th}$ value.

- To calculate this in R:

```
> median(data)
[1] 4
```

Median

- The median is a more robust measure than the mean, since it is not affected by outliers.
- It is often a better measure than the mean if the data are highly skewed (see later!) .
- However, it can be a disadvantage that the median only uses position and does not consider the specific values of the data.

Mode

- The mode of a data set is the value that occurs most frequently.
- Note that some data sets may have more than one mode, or may have none at all.

- Note: there is no inbuilt function in R to calculate mode – see commands on the worksheet.

Measures of dispersion

- Measures of location summarise a data set with one number, but they do not tell us everything about a distribution.
- Consider the following data showing the overall percentages children from School A and School B achieved in several tests.

School A

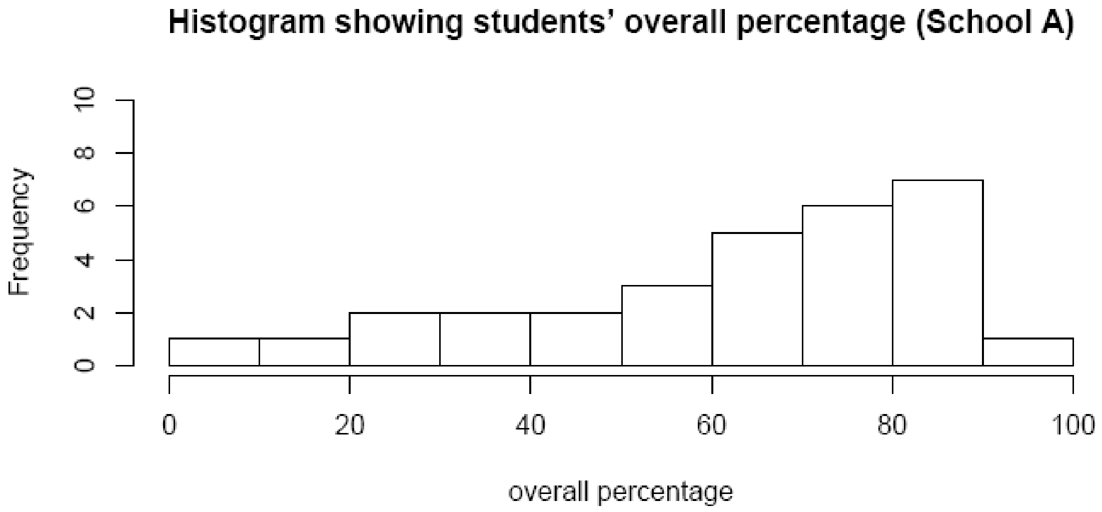
31.1	71.3	91.7	41.6	54.7
62.3	61.2	89.9	23.7	20.3
10.8	42.9	86.7	51.8	62.1
64.4	75.3	52.4	83.6	81.1
70.2	60.2	72.4	73.5	85.6
80.1	73.3	83.9	39.0	2.9

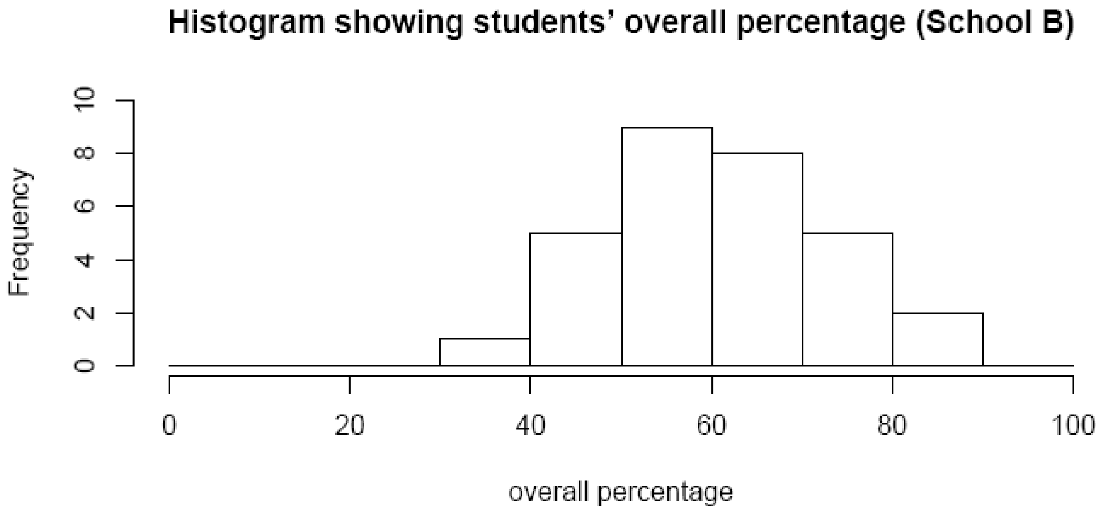
School B

60.9	62.1	31.8	75.2	51.3
66.8	44.2	78.1	44.0	70.6
58.8	46.5	57.9	65.9	53.7
60.3	79.8	42.7	80.4	59.6
63.2	54.1	58.1	71.5	62.2
53.1	47.9	60.9	57.4	81.0

The mean overall percentage is 60 for both schools, but the histograms below clearly show the distributions are different; there is a much larger variation of values in School A compared to School B.

We need measures of dispersion to provide information on the variation in a data set.



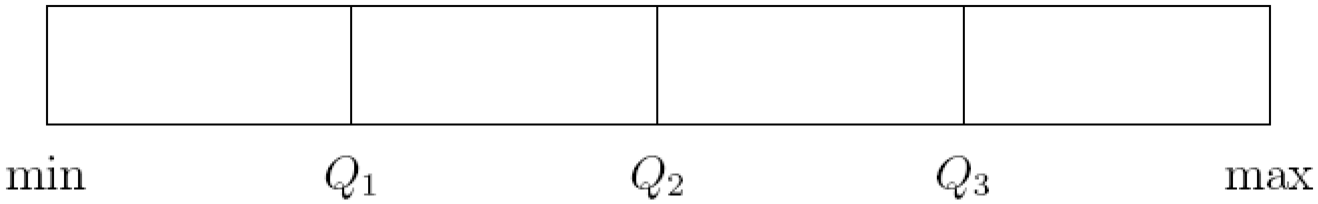


Range

- The range is the difference between the largest and smallest values in a dataset.
- However, the range is very sensitive to outliers (extreme values) in the data.

Quartiles

Before we look at interquartile range (IQR), we need to know how to compute *quartiles*, which divide an ordered set of data into quarters.



After dividing the data into quarters, 25% of the data lies below the lower quartile Q_1, 50% below the median Q_2 and 75% below the upper quartile Q_3.

The Interquartile Range (IQR)

- The quartiles of a data set can be calculated by
  - ordering the data from smallest to largest
  - find the median, Q_2 (we have already learnt how to do this)
  - Q_1 is the median of the lower half of the data (of the values below Q_2)
  - Q_3 is the median of the upper half of the data (of the values above Q_2)
- The interquartile range is the range of the middle 50% of the data,

$$IQR = Q_3 - Q_1.$$

- This is a more robust measure than the range as it is not affected by outliers.

Five Number Summary

- The five number summary of a data set consists of the minimum value, Q_1, median, Q_3 and maximum value.

- A five number summary can then be used to create a **boxplot**.

- In R:

```
> fivenum(data)
[1] 2.90 42.90 63.35 80.10 91.70
```

Boxplot

For the data on overall percentages achieved by students in School A considered earlier, the five number summary is:

min = 2.9

Q_1 = 42.9

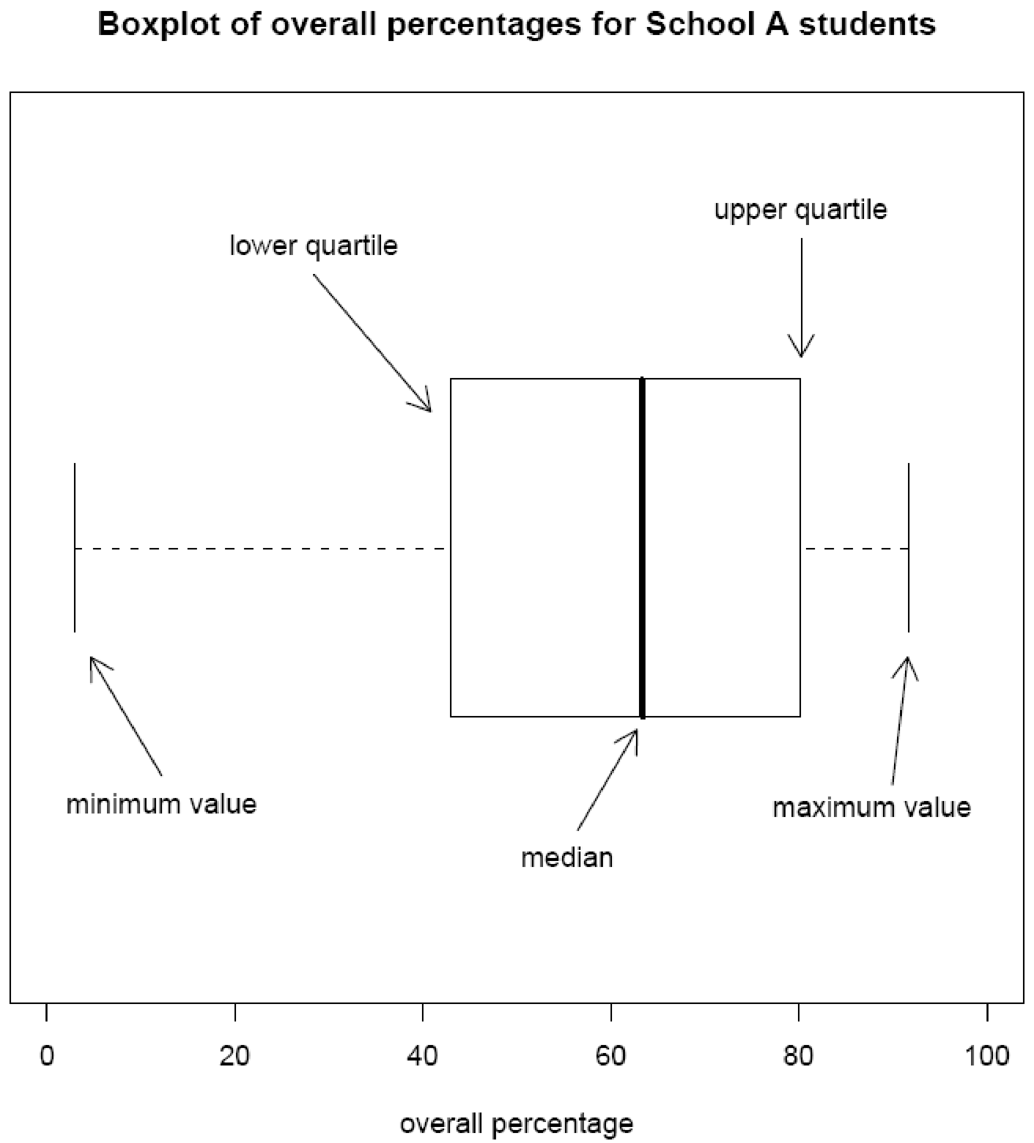
median = 63.35

Q_3 = 80.1

max = 91.7

In R:

```
> boxplot(data)
```



Sample variance and sample standard deviation

- The sample variance (s^2) and sample standard deviation (s) are the most common measures of spread, and they explain how much variation there is in the data.
- The best way to interpret the standard deviation is to think of it as the average distance of each of the observations in the data from the mean.
- If all the observations were the same, then the standard deviation would be zero.

Sample variance and sample standard deviation

The formula for the sample variance is

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

but it is often easier to use

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

Standard deviation in R

- In R:

```
> sd(data)
```