

# **SM-1402/CU-0304 Basic Statistics**

## **Chapter 1 (Summarising Data)**

**Dr. Haziq Jamil**

**Semester 2, 2019/20**

# Representation of Data

# Types of data

Numeric data can be one of two types: *discrete* or *continuous*.

## Discrete data

Discrete data can take **only exact (whole) values**, for example:

- the number of lecture rooms in UBD,
- the number of students in a class,
- the number of durians on each tree in a plantations.

Example of raw discrete data: *An orange tree produces roughly a number of fruits per tree. A survey of 30 orange trees yielded the following number of fruits per tree.*

```
[1] 31 32 34 30 32 33 31 34 32 33 35 32 32 33 32 32 33 31 32 33  
[21] 32 30 31 31 32 30 33 32 33 33
```

# Types of data

Numeric data can be one of two types: *discrete* or *continuous*.

## Continuous data

Continuous data *cannot* take exact (whole) values, but can be given only within a specified range, or measured to a specified degree of accuracy.

- the heights of students in UBD,
- the length of classrooms in UBD,
- the travel time in minutes from home to UBD.

Example of raw continuous data: *A survey of the heights of 20 children in a sports club yielded the following measurements (rounded to the nearest centimetre)*

```
[1] 133 136 120 138 133 131 127 141 127 143 130 131 125 144 128  
[16] 134 135 137 133 129
```

# Grouping data

Raw data can be grouped for a more concise presentation.

## Discrete data (oranges from trees)

Count the number of times each value occurs and summarise these in a table known as a **frequency distribution**

```
[1] 31 32 34 30 32 33 31 34 32 33 35 32 32 33 32 32 33 31 32 33  
[21] 32 30 31 31 32 30 33 32 33 33
```

Number of oranges	30	31	32	33	34	35	
Frequency	3	5	11	8	2	1	Total = 30

# Grouping data

Raw data can be grouped for a more concise presentation.

## Continuous data (children's heights)

To form a **frequency distribution** of the heights of the 20 children, group the information into **classes or intervals**

[1] 133 136 120 138 133 131 127 141 127 143 130 131 125 144 128  
[16] 134 135 137 133 129

Height (cm)	119.5-124.5	124.5-129.5	129.5-134.5	134.5-139.5	139.5-144.5	
Freq.	1	5	7	4	3	Total = 20

The values 119.5, 124.5, 129.4, ... are called the **class boundaries**. The **upper class boundary** of one interval is the lower class boundary of the next interval.

# Grouping data

Raw data can be grouped for a more concise presentation.

## Terminology

- The **mode** (for discrete data) is the value which occurs most often. This can be read off directly from the frequency distribution.

For the orange data, the mode is *32 oranges*.

- The **width of an interval** (for grouped data) is the difference between boundaries.

For the children's height frequency distribution, the interval width used was *5cm*.

# Ways of grouping data

## Example 1

*The length of 30 metal rods were measured to the nearest millimetres.*

Length (mm)	27-31	32-36	37-46	47-51
Freq.	4	11	12	3

- The interval 27-31 means  $26.5\text{mm} \leq \text{length} < 31.5\text{mm}$ , and so on.
- The class boundaries are 26.5, 31.5, 36.5, 46.5, and 51.5.
- The class widths<sup>1</sup> are 5, 5, 10, and 5.

---

[1] Class widths do not necessarily have to be the same for all classes.



# Ways of grouping data

## Example 2

*100 students took a test, and their marks are tabulated.*

Mark	30-39	40-49	50-59	60-69	70-79	80-99
Freq.	10	14	26	20	18	12

The frequency distribution can be interpreted in two ways:

- As **discrete** data.
  - The interval 30-39 represents  $30 \leq \text{mark} < 40$ .
  - The class boundaries are 30, 40, 50, 60, 70, 80, 100.
  - The class widths are 10, 10, 10, 10, 10, 20.
- As **continuous** data assuming marks are to the nearest integer.
  - The interval 30-39 represents  $29.5 \leq \text{mark} < 39.5$ .
  - The class boundaries are 29.5, 39.5, 49.5, 59.5, 69.5, 79.5, 99.5.
  - The class widths are 10, 10, 10, 10, 10, 20.

# Ways of grouping data

## Example 3

*Ages (in completed years) of applicants for a teaching post.*

Age (years)	21-24	25-28	29-32	33-40	41-52
Freq.	4	2	2	1	1

- Since ages are given in *completed years* (not to the nearest year), then '21-24' means  $21 \leq \text{age} < 25$ <sup>1</sup>.
- We can also write the categories as '21-', '25-', etc.
- The class boundaries are 21, 25, 29, 33, 41, 53.
- The class widths are 4, 4, 4, 8, 12.

---

[1] Someone who is 24 years and 11 months is not yet 25 years old, so would come under this category.

# Histograms

Consider the following table for the distribution of ages of 118 passengers on a shuttle flight from Bandar Seri Begawan, Brunei to Kota Kinabalu, Sabah.

Age (years)	0-19	20-39	40-49	50-69	70-100
Frequency	4	44	36	28	6

A **histogram** provides a visual representation of *grouped data*.

# Histograms

Age (years)	0-19	20-39	40-49	50-69	70-100
Frequency	4	44	36	28	6

# Histograms

- In a histogram, the area of each bar is proportional to the frequency that it represents, and thus

$$\text{total area} \propto \text{total frequency}$$

- The vertical axis is not labelled frequency, but *frequency density*<sup>1</sup>, where

$$\text{freq. density} = \frac{\text{frequency}}{\text{interval width}}$$

---

[1] You may also find histograms with the vertical axis labelled 'density' (c.f. density plots), or even 'frequency' (this type of histogram visualises where values are concentrated).

# Histograms

- Using the formula given, we can calculate the complete table for the histogram as follows:

Age (years)	Interval width	Frequency	Freq. density
0-19	20	4	$4 \div 20 = 0.2$
20-39	20	44	$44 \div 20 = 2.2$
40-49	10	36	$36 \div 10 = 3.6$
50-69	20	28	$28 \div 20 = 1.4$
70-99	30	6	$6 \div 30 = 0.2$

# Histograms

- Histograms may have bars of different widths, so the height of the bar must be adjusted accordingly.
- The **modal class** is the interval with the greatest frequency density, i.e. the interval represented by *the highest bar* in the histogram. For the current example, the modal class is the interval '40-49'<sup>1</sup>.

---

[1] Notice that this interval does not have the greatest frequency, but it does have the greatest frequency density.

# Frequency polygons

An alternative way of displaying grouped data is using **frequency polygons**.



# Frequency polygons

Age (years)	0-19	20-39	40-49	50-69	70-100
Frequency	4	44	36	28	6

# Frequency polygons

To construct a frequency polygon,

- Calculate the mid-interval value, where

$$\text{mid-interval value} = \frac{1}{2}(\text{LCB} + \text{UCB})$$

- Calculate the frequency densities.
- Join points together with a straight line.

# Frequency polygons

- Calculate the complete table for the frequency polygon:

Age (years)	Mid-interval value	Interval width	Frequency	Freq. density
0-19	$(0 + 20) \div 2 = 10$	20	4	$4 \div 20 = 0.2$
20-39	$(20 + 40) \div 2 = 30$	20	44	$44 \div 20 = 2.2$
40-49	$(40 + 50) \div 2 = 45$	10	36	$36 \div 10 = 3.6$
50-69	$(50 + 70) \div 2 = 60$	20	28	$28 \div 20 = 1.4$
70-99	$(70 + 100) \div 2 = 85$	30	6	$6 \div 30 = 0.2$

# Pie charts

Yet another diagram to display grouped data is a pie chart.

- A circle (pie) is divided into several **sectors** (slices of the pie), with each sector corresponding to a class interval.
- The central angle  $\theta$  of the sectors are *proportional* to the frequencies of the intervals they represents, according to the formula

$$\theta = \frac{\text{class freq.}}{\text{total freq.}} \times 360^\circ$$

- The area of a sector  $A$  of a circle with radius  $r$  is given by the formula

$$A = \frac{\text{class freq.}}{\text{total freq.}} \times \pi r^2 = \frac{\theta}{360} \times \pi r^2$$

# Pie charts

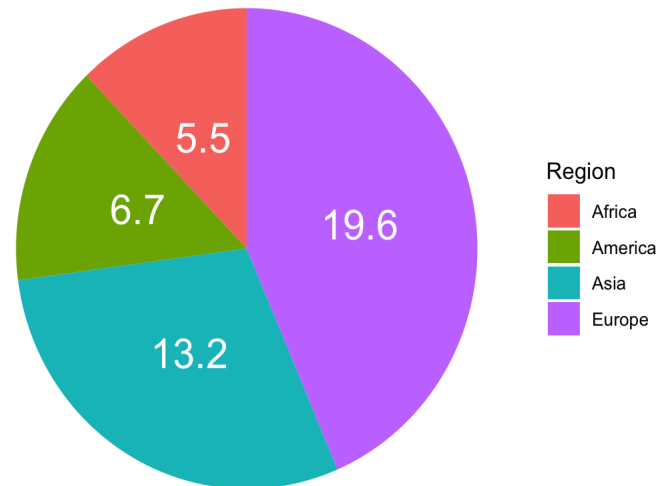
**WARNING: Pie charts should only be used for grouped data with equal class intervals**

*Annual sales (millions of dollars) of a company by region*

Region	Africa	America	Asia	Europe
Sales	5.5	6.7	13.2	19.6

*Pie chart table*

Region	Sales	Angle	Area (r=1)
Africa	5.5	$5.5/45 \times 360$ = 44	$5.5/45 \times \pi$ = 0.38
America	6.7	$6.7/45 \times 360$ = 54	$6.7/45 \times \pi$ = 0.47
Asia	13.2	$13.2/45 \times 360$ = 106	$13.2/45 \times \pi$ = 0.92
Europe	19.6	$19.6/45 \times 360$ = 156	$19.6/45 \times \pi$ = 1.37
Total	45.0	360	$\pi$



# Data Summaries

# The mean

- A 'typical' or 'average' value is useful when interpreting data. One such value is the **mean**<sup>1</sup>.
- The formula for the mean is

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

- Consider 5 numbers

0.9, 1.4, 2.8, 3.1, 5.6

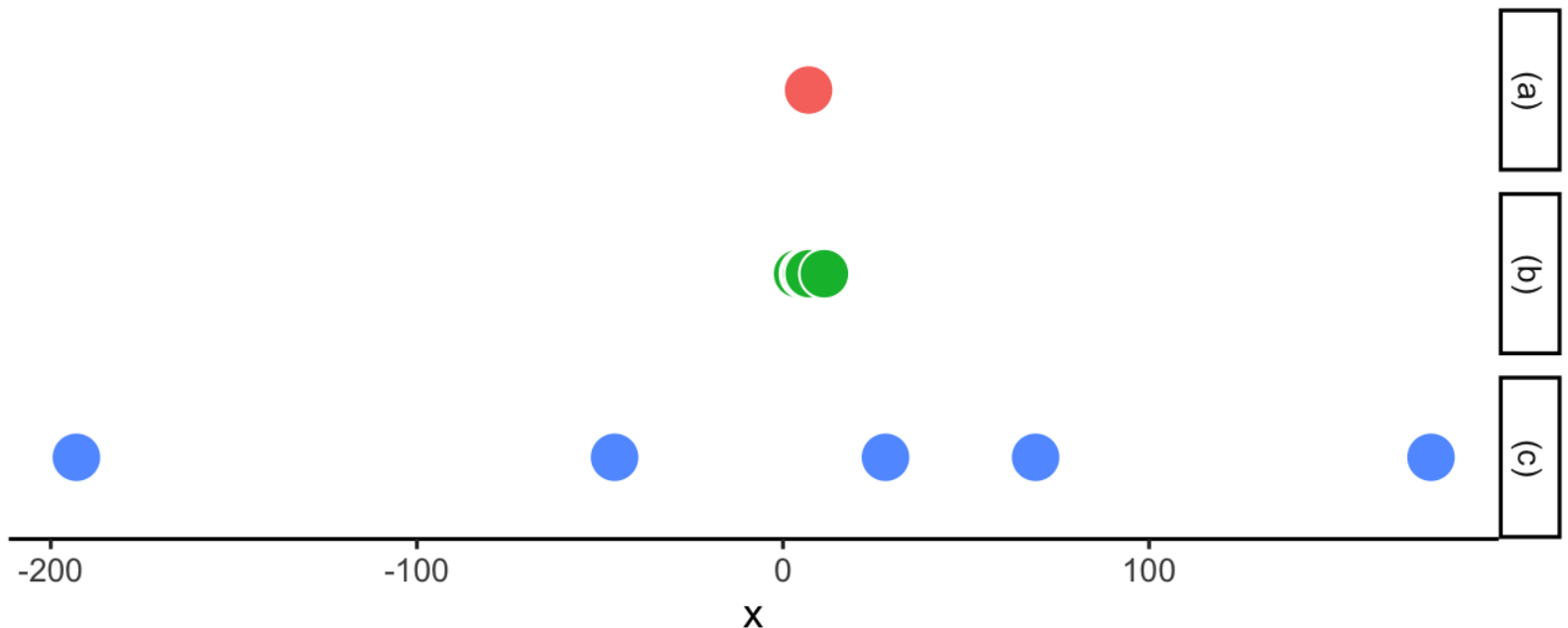
The mean is  $(0.9 + 1.4 + 2.8 + 3.1 + 5.6)/5 = 13.8/5 = 2.76$ .

[1] 'Average' does not always refer to the 'mean'. There are other measures of *central tendencies*, such as the mode, median and others.

# Variability

- Each of these sets of numbers has a mean of 7, but the *spread* of each set is different.

(a): 7, 7, 7, 7, 7  
(b): 4, 6, 6.5, 7.2, 11.3  
(c): -193, -46, 28, 69, 177





# Variability

We can measure the **variability** or spread of a distribution by calculating the *range* and the *standard deviation*.

## The range

- The range is based entirely on the extreme values of the distribution.

range = highest value – lowest value

```
(a): range(7, 7, 7, 7, 7)      = 7 - 7      = 0
(b): range(4, 6, 6.5, 7.2, 11.3) = 11.3 - 4    = 7.3
(c): -193, -46, 28, 69, 177    = 177 - (-193) = 370
```

# Variability

We can measure the **variability** or spread of a distribution by calculating the *range* and the *standard deviation*.

## The standard deviation and the variance

- The standard deviation  $s$  gives a measure of the deviations of the readings from the mean  $\bar{x}$ .
- Steps to calculate  $s$  (for  $i = 1, \dots, n$ )
  1. For each reading  $x_i$ , calculate  $x_i - \bar{x}$ , its deviation from the mean
  2. Square this deviation to give  $(x_i - \bar{x})^2$
  3. Sum all of these squared deviations to give  $\sum_{i=1}^n (x_i - \bar{x})^2$
  4. Divide this sum by  $n$  to give the **variance**
  5. Finally, take the positive square root of the variance to obtain the standard deviation  $s$

# Variability

We can measure the **variability** or spread of a distribution by calculating the *range* and the *standard deviation*.

## The standard deviation and the variance

- The formula for the variance is given by

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

- The formula for the standard deviation is given by

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

# Calculating the s.d.

Set (a)

$i$	Data $x_i$	Mean $\bar{x}$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	7	7	0	0
2	7	7	0	0
3	7	7	0	0
4	7	7	0	0
5	7	7	0	0
				SUM = 0

$$s^2 = 0/5 = 0$$

$$s = \sqrt{s^2} = \sqrt{0} = 0$$

# Calculating the s.d.

Set (b)

$i$	Data $x_i$	Mean $\bar{x}$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	4	7	-3	9
2	6	7	-1	1
3	6.5	7	-0.5	0.25
4	7.2	7	0.2	0.04
5	11.3	7	4.3	18.49
				SUM = 28.78

$$s^2 = 28.78/5 = 5.756$$

$$s = \sqrt{s^2} = \sqrt{5.756} = 2.4 \text{ (1 d.p.)}$$

# Notes on the s.d.

- The units of the standard deviation are the *same* as the units of the data.
- Standard deviations are useful when comparing sets of data; the higher the standard deviation, the greater the variability in the data.
- Alternative formula for the s.d.

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

# Cumulative frequencies

- The cumulative frequency is the **total frequency** *up to a particular item*.
- A cumulative frequency distribution can be obtained from a frequency distribution and can be illustrated
  - When the data are *discrete* and *ungrouped* (**step diagram**);
  - When the data are *continuous* or in the form of *grouped discrete distribution* (**cumulative frequency polygon/curve**).

# Step diagram

*Example: The table above shows the number of attempts needed to pass the driving test by 100 candidates at a particular test centre.*

No. of attempts	1	2	3	4	5	6
Freq. (No. of candidates)	33	42	13	6	4	2

The cumulative frequency table is created by adding the attempts cumulatively.

No. of attempts	$\leq 1$	$\leq 2$	$\leq 3$	$\leq 4$	$\leq 5$	$\leq 6$
Cumulative freq.	33	75	88	94	98	100

Notice that the last entry of the cumulative frequency table must equal to the number of observations in total.

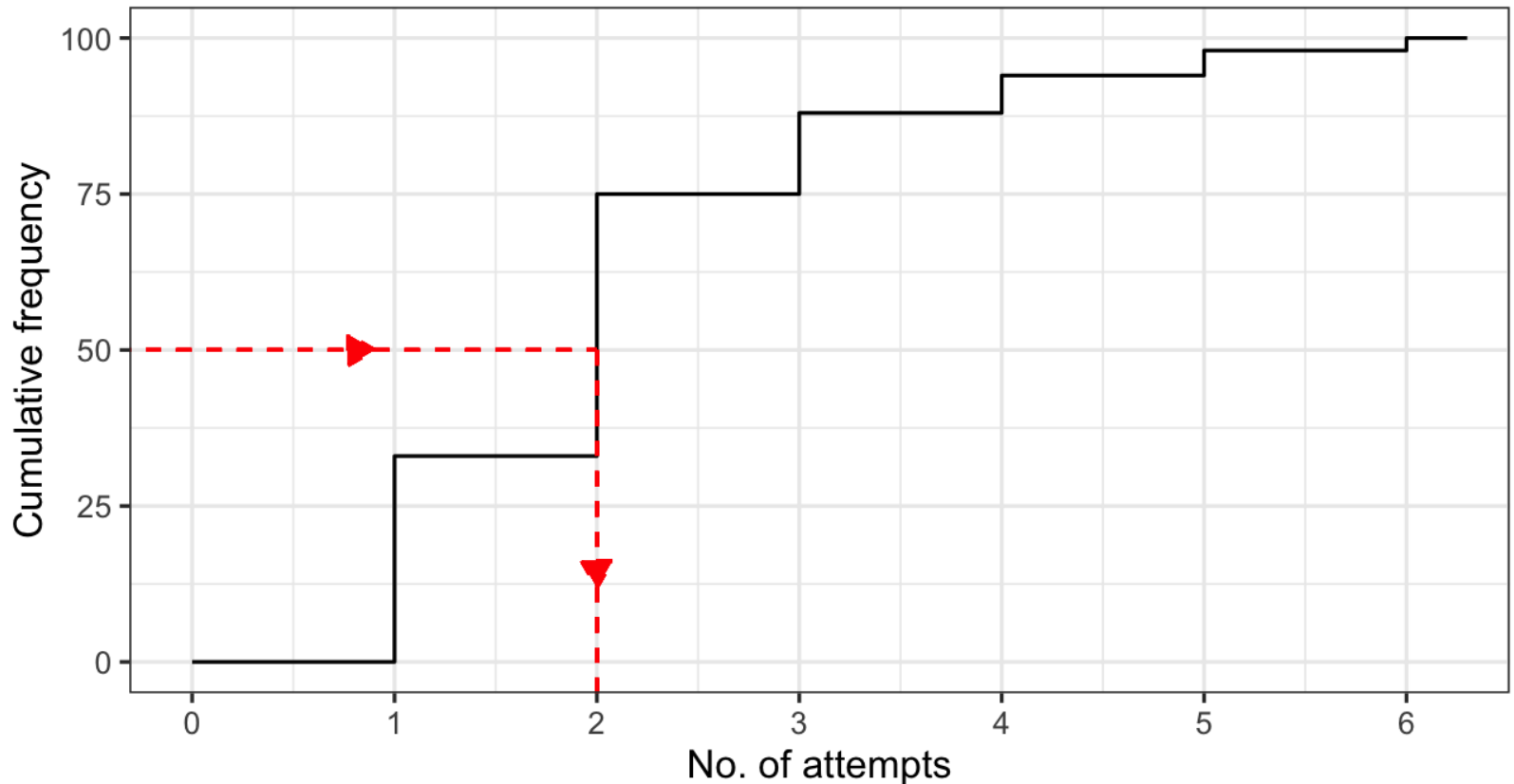


# Step diagram

No. of attempts	$\leq 1$	$\leq 2$	$\leq 3$	$\leq 4$	$\leq 5$	$\leq 6$
Cumulative freq.	33	75	88	94	98	100

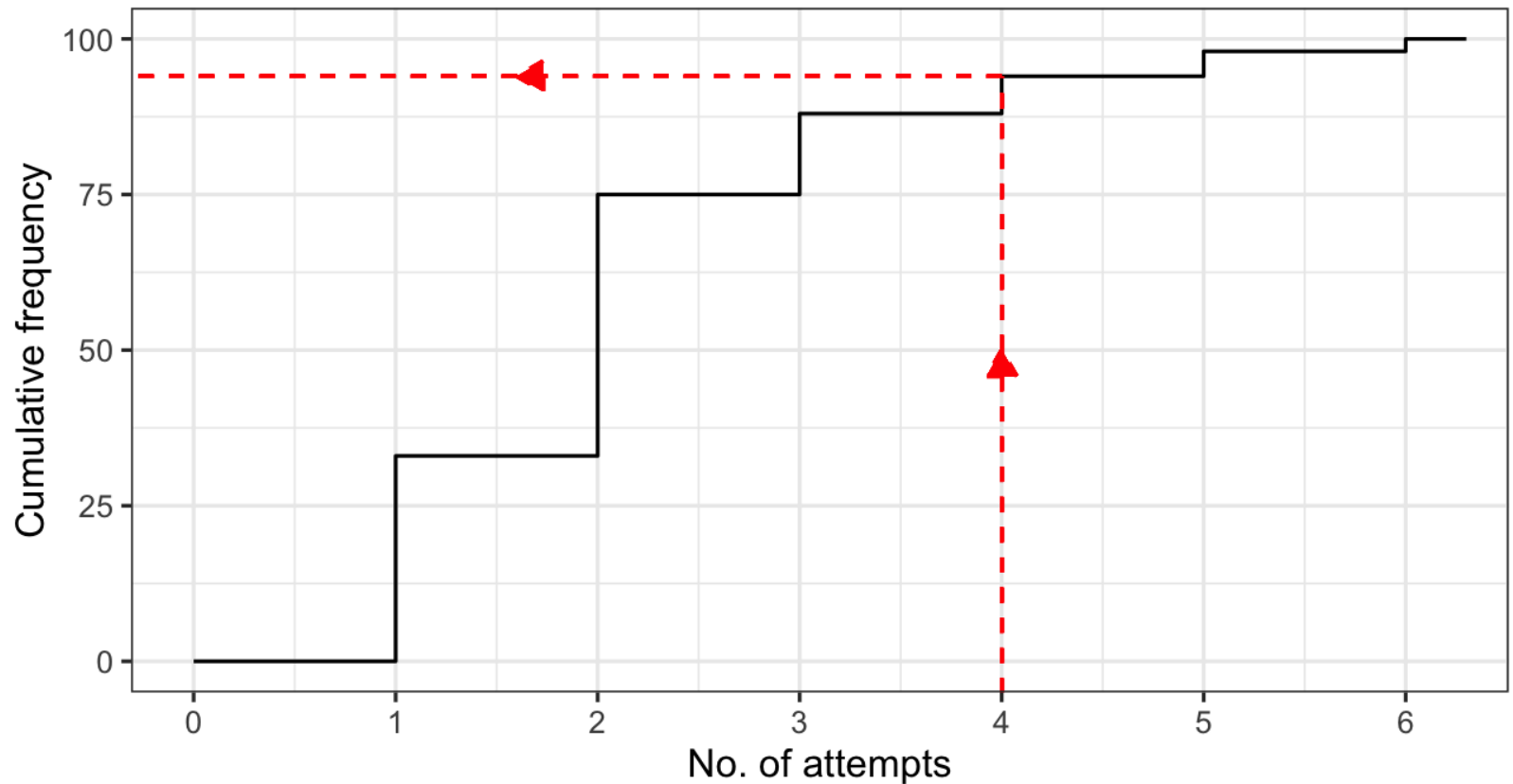
# Step diagram

Arrange candidates in ascending order of number of attempts. How many tries did the 50th candidate take to pass the test?



# Step diagram

How many candidates took up to four attempts?



# Step diagram

- It only makes sense when you read from the discrete whole values on the horizontal axis. It would be silly to consider 3.6 attempts, for example.
- In the step diagram, the *mode* is given by the value of the variable that gives the 'steepest' step.

# Cumulative frequency polygons

*Example: Six weeks after planting, the heights of 30 broad bean plants were measured and the frequency distribution formed as shown.*

Height $x$ (cm)	$3 \leq x < 6$	$6 \leq x < 9$	$9 \leq x < 12$	$12 \leq x < 15$	$15 \leq x < 18$	$18 \leq x < 21$
Frequency	1	2	11	10	5	1

- The cumulative frequency is calculated up to each UCB (6, 9, 12, 15, 18, 21).
- The lower boundary of the first class is 3.

Height $x$ (cm)	$< 3$	$< 6$	$< 9$	$< 12$	$< 15$	$< 18$	$< 21$
Cumulative freq.	0	1	3	14	24	29	30

# Cumulative frequency polygons

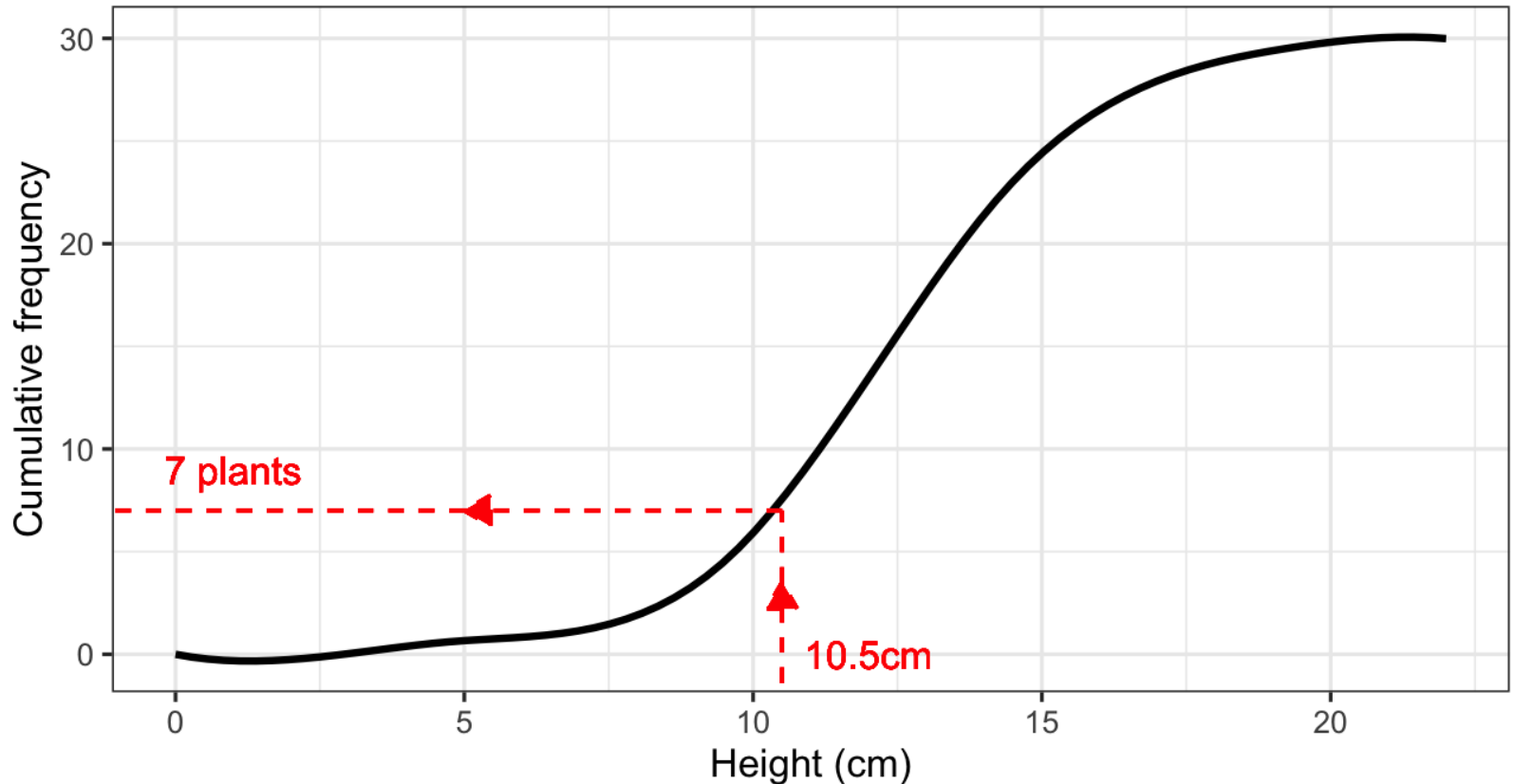
Joining all the points with a straight line gives a **cumulative frequency polygon**. This assumes that the readings are *evenly* distributed throughout the intervals.

# Cumulative frequency curve

Joining all the points with a smooth line gives a **cumulative frequency curve**. This assumes that the readings are *not evenly* distributed throughout the intervals.

# Cumulative frequency curve

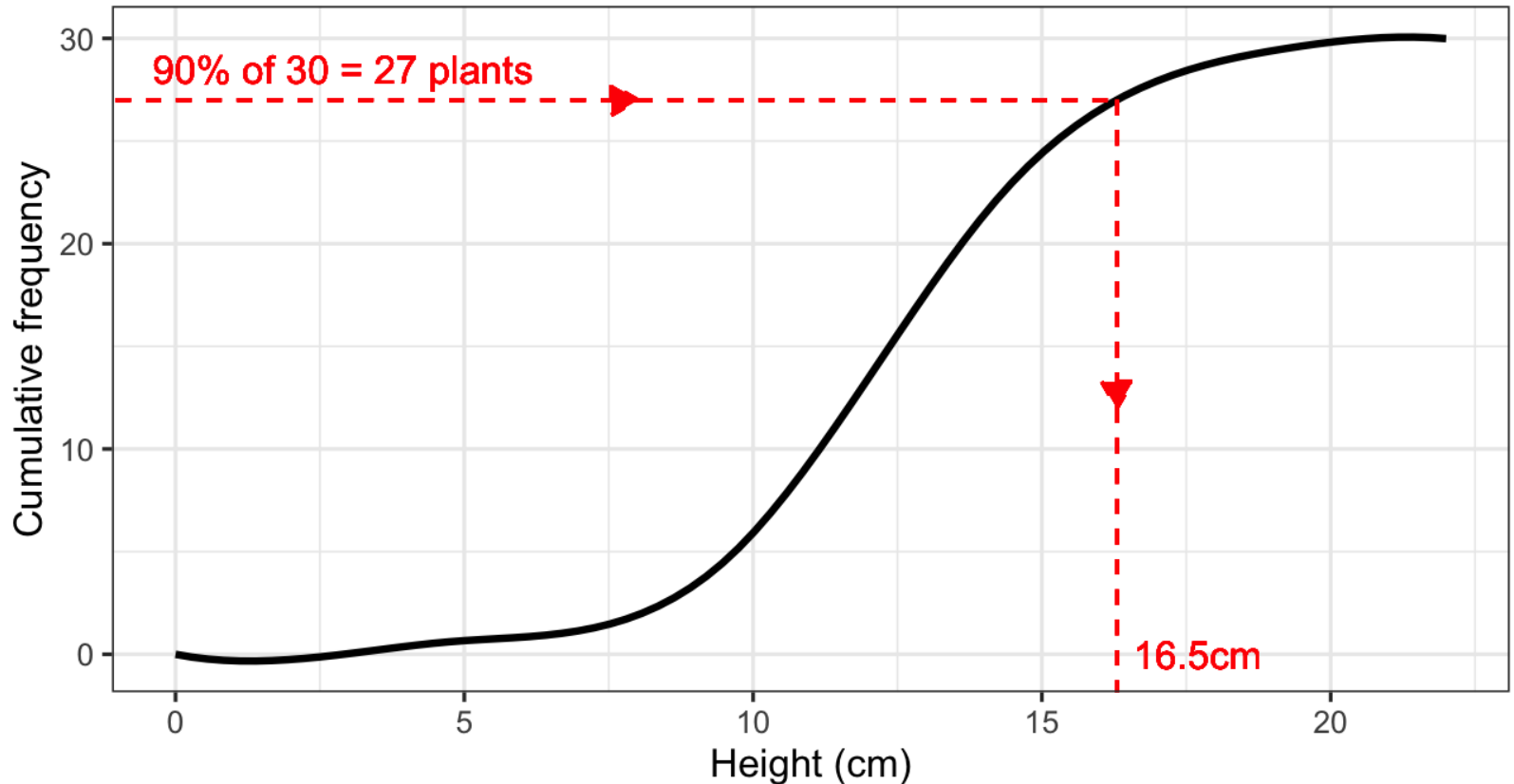
How many plants were less than 10.5cm tall?





# Cumulative frequency curve

Find  $x$  where 90% of the plants were less than  $x$  cm tall.



# Median and quartiles

For data arranged in order of size,

- The **lower quartile**  $Q_1$  is the value 25% of the way through the distribution.
- The **median**  $Q_2$  is the value 50% of the way through the distribution.
- The **upper quartile**  $Q_3$  is the value 75% of the way through the distribution.

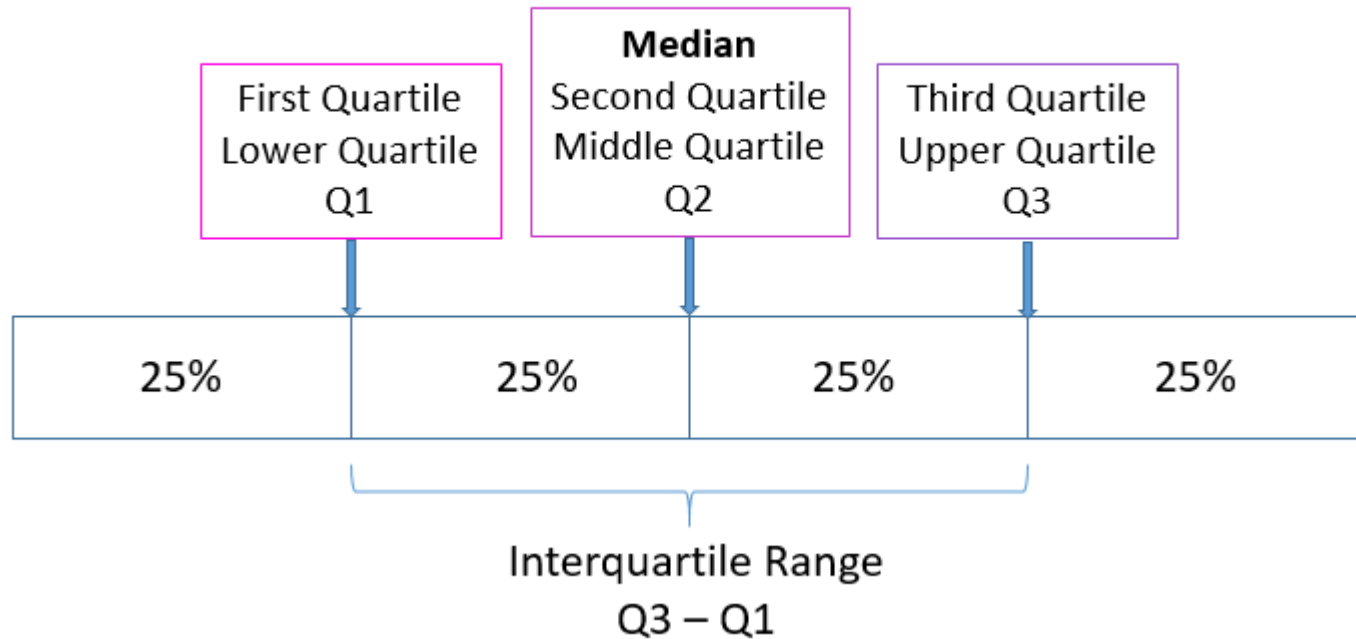
Therefore, the median, lower quartile and upper quartile together split the distribution into four equal parts.

- The **inter-quartile range**, defined as the difference between the quartiles, tells you the range of the middle 50% of the distribution.

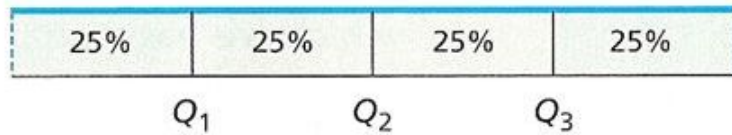
$$\text{IR} = Q_3 - Q_1$$

# Median and quartiles

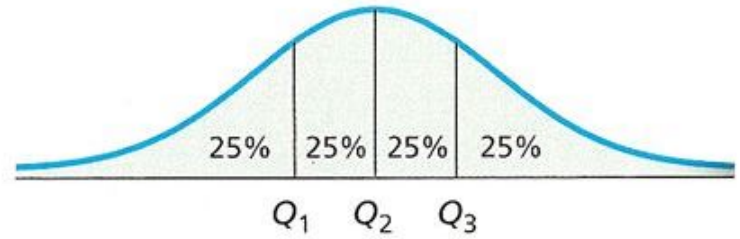
## Median and Quartiles



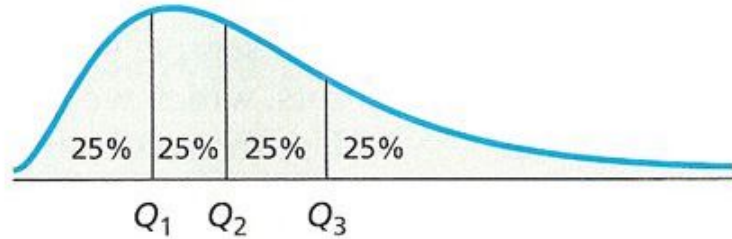
# Median and quartiles



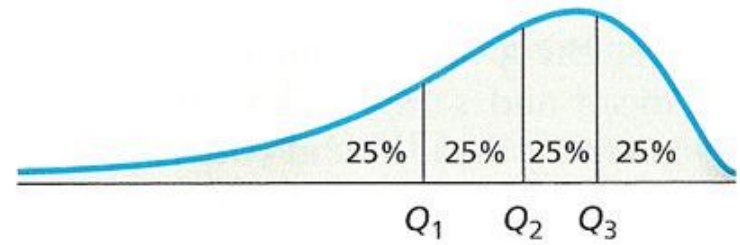
(a) Uniform



(b) Bell shaped



(c) Right skewed



(d) Left skewed

# Quartiles for ungrouped data

- For ungrouped data consisting of  $n$  observations arranged in ascending order of size, the median is the

$$Q_2 = \frac{1}{2}(n + 1)\text{th observation}$$

- If there is an odd number of observations, the median is the middle value.
- If there is an even number of observations, there are two middle values, and the median is the average of these two numbers.
- The quartiles should divide the two distributions either side of the median in half. Re-apply the formula for the median for each half to find the lower and upper quartiles.

# Quartiles for ungrouped data

*Example: Consider the set of numbers  $\{7, 7, 2, 3, 4, 2, 7, 9, 31\}$*

- Since there are 9 numbers, the median is the  $(9 + 1)/2$  th observation, i.e. the 5th observation, which is 7.

2, 2, 3, 4, [7], 7, 7, 9, 31

- The median of the left distribution  $\{2, 2, 3, 4\}$  gives the lower quartile, i.e. the middle value between 2 and 3 =  $(2 + 3) / 2 = 2.5$ .

2, [2, 3], 4, [7], 7, 7, 9, 31

- The median of the right distribution  $\{7, 7, 9, 31\}$  gives the lower quartile, i.e. the middle value between 7 and 9 =  $(7 + 9) / 2 = 8$ .

2, 2, 3, 4, [7], 7, [7, 9], 31

# Quartiles for ungrouped data

- Sometimes, the following rule is used to find the quartiles:

$$Q_1 = \frac{1}{4}(n + 1)\text{th observation}$$

$$Q_3 = \frac{3}{4}(n + 1)\text{th observation}$$

This rule agrees with the above method when  $n$  is odd, but there is a discrepancy when  $n$  is even. But it does not make a great deal of difference whichever method is used.

# Quartiles for ungrouped data

To find the median and quartiles of data in the form of an ungrouped frequency distribution, it is useful to find the cumulative frequency.

*Example: The table shows the cumulative number of attempts needed to pass the driving test by 100 candidates at a particular test centre.*

No. of attempts	$\leq 1$	$\leq 2$	$\leq 3$	$\leq 4$	$\leq 5$	$\leq 6$
Cumulative freq.	33	75	88	94	98	100

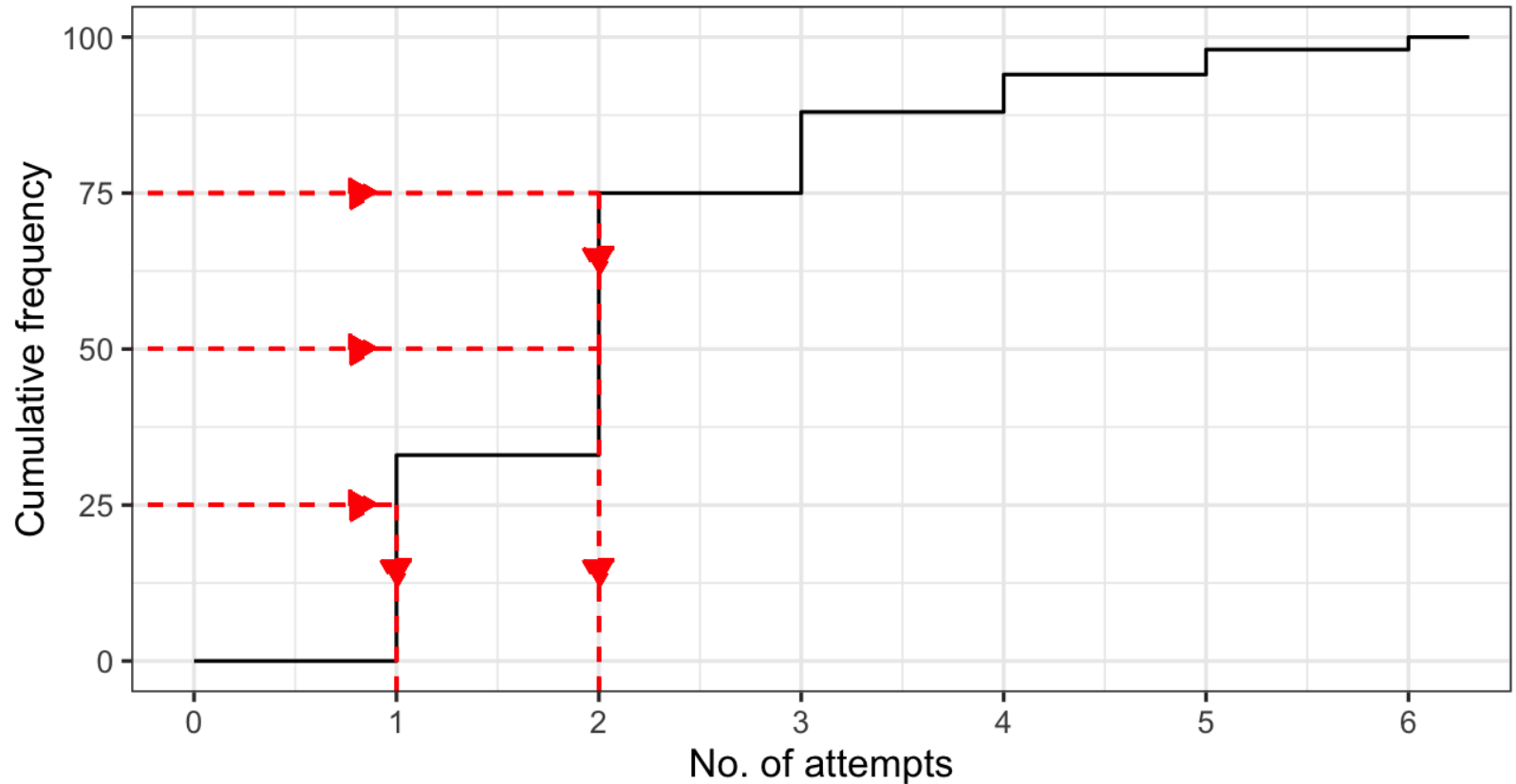
Since there are  $n = 100$  observations,

- $Q_2$  is the middle value of 50 and 51, i.e. the 50.5th observation. This lies in the class of ' $\leq 2$ ' attempts.
- $Q_1$  is the 25th observation, which lies in the class of ' $\leq 1$ ' attempts.
- $Q_3$  is the 75th observation, which lies in the class of ' $\leq 2$ ' attempts.



# Quartiles for ungrouped data

Or, we can read from the step diagram

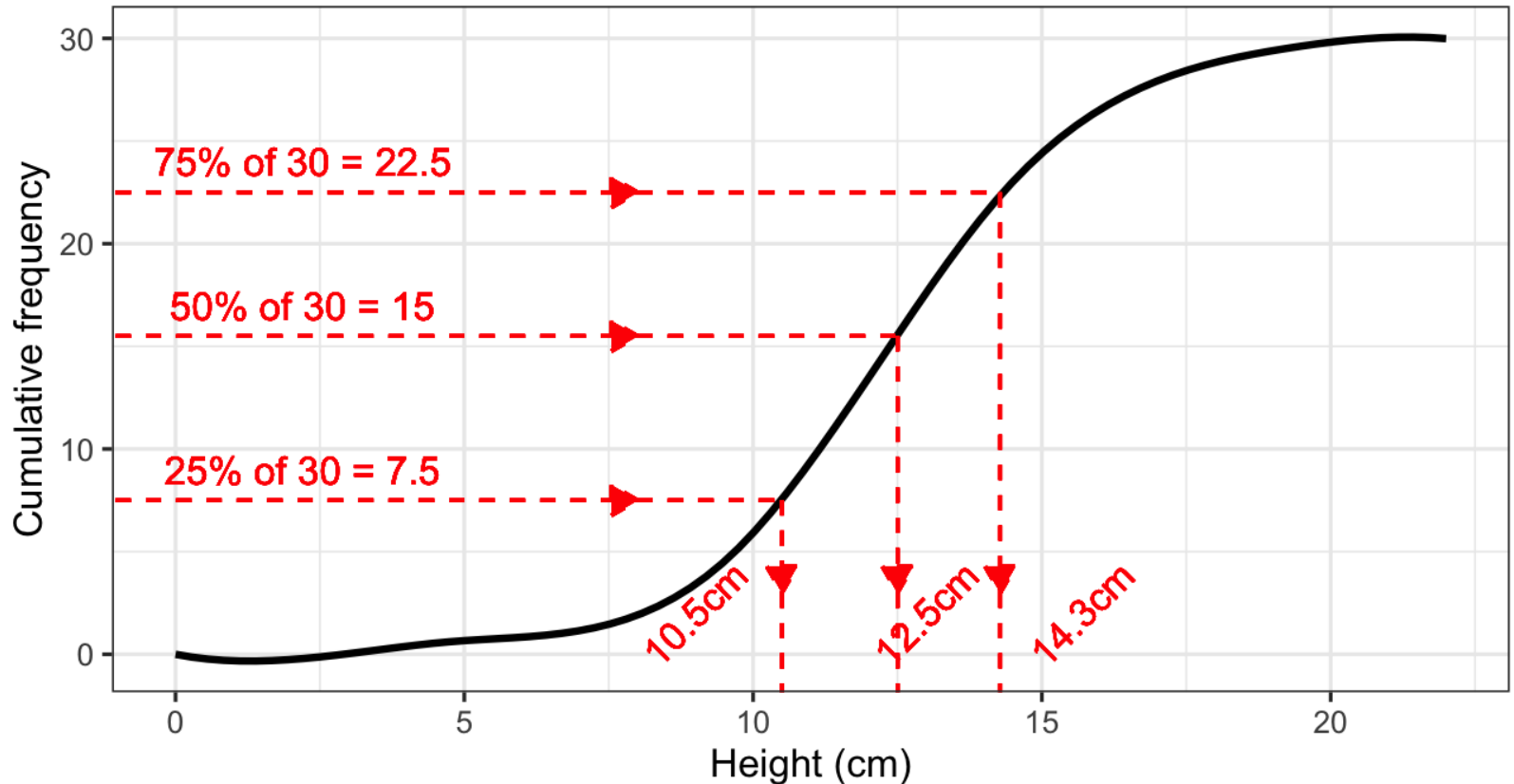


# Quartiles for grouped data

- When data have been grouped into intervals, the original information has been lost, so it is only possible to make *estimates* of the median and quartiles.
- One way of doing this is to use a cumulative frequency graph (polygon or curve).
  - $Q_1$  is the  $\frac{1}{4}$ th reading.
  - $Q_2$  is the  $\frac{1}{2}$ th reading.
  - $Q_3$  is the  $\frac{3}{4}$ th reading.

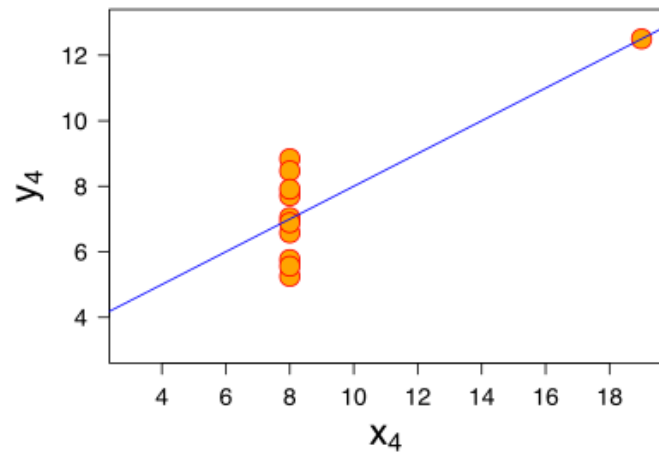
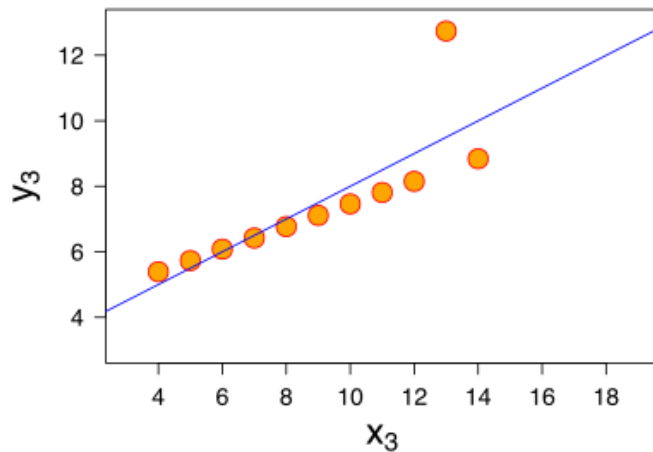
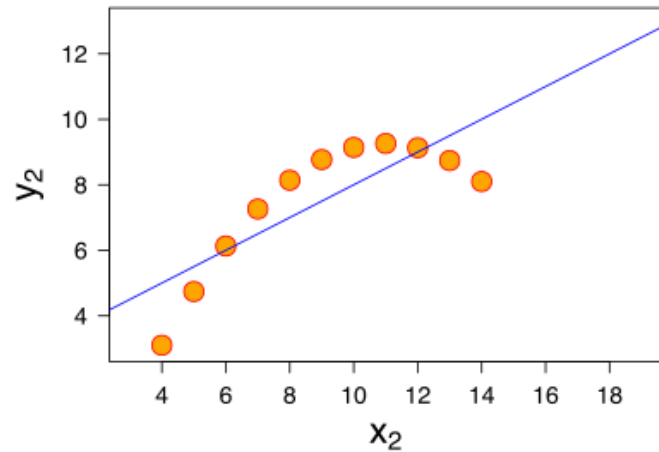
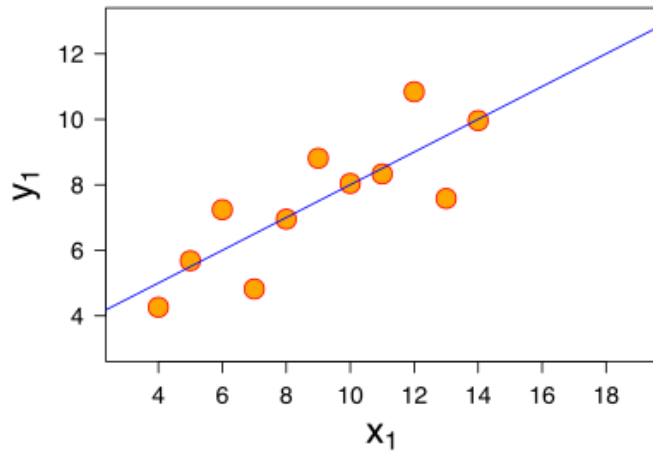
# Quartiles for grouped data

Back to the heights of 30 broad beans six weeks after planting.



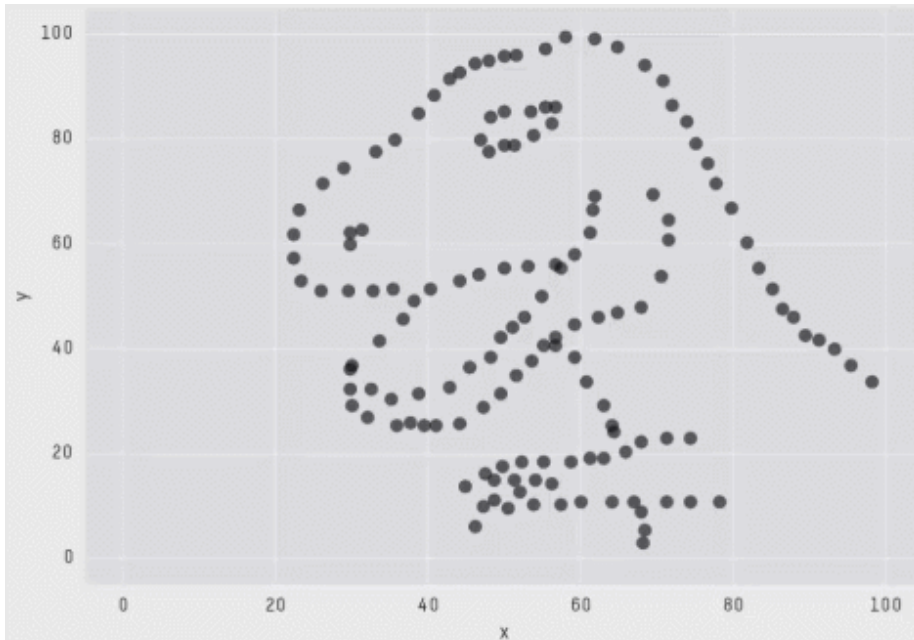
# Anscombe's Quartet

A cautionary tale in summary statistics



Mean of  $x = 9$ , Variance of  $x = 11$ . Mean of  $y = 7.5$ , Variance of  $y = 4.125$ .  
Correlation between  $x$  and  $y = 0.816$ .

# The Datasaurus Dozen



X Mean: 54.2659224  
Y Mean: 47.8313999  
X SD : 16.7649829  
Y SD : 26.9342120  
Corr. : -0.0642526

<https://www.autodeskresearch.com/publications/samestats>

**END**