

# Statistical Methods

## 3. Descriptive Statistics

Based on materials provided by Coventry University and  
Loughborough University under a National HE STEM  
Programme Practice Transfer Adopters grant



# Workshop outline

We will consider:

- ☐ Types of data
- ☐ Tables
- ☐ Charts
- ☐ Measures of middle value
- ☐ Measures of spread
- ☐ Choice of table, chart and measure

# Data types

There are three main data types in statistics:

- ❑ **Nominal – categories** that do not have a natural order, e.g. gender, eye colour, types of building
- ❑ **Ordinal – categories** which have a natural order but are not numerical, e.g. Likert response scales
- ❑ **Scale/continuous** – numerical data ordered against a constant scale, e.g. date, temperature, length, weight, frequency


You can carry out more operations (and better tests) with the scale data type

# Data types and their analysis

Type	Permitted maths operations	Permitted statistics
<b>Nominal</b>	Counting (frequency)	Percentage, mode
<b>Ordinal</b>	Counting, <, >	Percentage, median, mode, minimum, maximum, upper and lower quartiles
<b>Scale/ continuous</b>	(Interval) Counting, <, >, +, -	Percentage, median, mode, minimum, maximum, range, upper and lower quartiles, inter-quartile range, mean, standard deviation, skewness, kurtosis
	(Ratio) Counting, <, >, +, -, $\times$ , $\div$	

# Example: CensusAtSchool Questionnaire

☐ Completed by 673  
school children



## Phase 6 CensusAtSchool Questionnaire

For students aged  
11 and above

<p>1. State the first part of your postcode (eg N63 or PL23)</p> <div style="border: 1px solid black; height: 20px; width: 100%;"></div> <p>2. Are you a ?</p> <p><input type="checkbox"/> Boy    <input type="checkbox"/> Girl</p> <p>3. Please state your age in completed years.</p> <p><input type="text"/> years</p> <p>4. Complete the following measurements.</p> <p>HEIGHT.....centimetres</p> <p>FOOT LENGTH.....centimetres</p> <p>Right Elbow to Wrist.....cm</p> <p>Open Arm Span.....cm</p> <p>5. What is your favourite food type?</p> <p><input type="checkbox"/> Dairy (milk, cheese, eggs)</p> <p><input type="checkbox"/> Protein (beans, meat, fish)</p> <p><input type="checkbox"/> Carbohydrates (bread, pasta)</p> <p><input type="checkbox"/> Fruit/Veg (apples, carrots)</p> <p>6. In a normal week on how many days do you eat meat?</p> <p>0 1 2 3 4 5 6 7</p> <p>7. How many (palm of hand) portions of the following do you regularly eat per day?</p> <p><input type="checkbox"/> Fruit</p> <p><input type="checkbox"/> Vegetables</p> <p><input type="checkbox"/> Sweets</p> <p><input type="checkbox"/> Crisps</p> <p>8. In the last year have you gone on a diet, changed your eating habits or done anything to control your weight. (leave blank if you wish)</p> <p><input type="checkbox"/> Yes    <input type="checkbox"/> No</p>	<p>9. When going out on sunny days in the summer do you:</p> <p>Use Sun Cream <input type="checkbox"/> Always <input type="checkbox"/> Sometimes <input type="checkbox"/> Never</p> <p>Wear a hat for protection <input type="checkbox"/> Always <input type="checkbox"/> Sometimes <input type="checkbox"/> Never</p> <p>Wear sunglasses <input type="checkbox"/> Always <input type="checkbox"/> Sometimes <input type="checkbox"/> Never</p> <p>10. On how many days last week did you do physical activity that made you huff and puff, sweat or get tired?</p> <p>0 1 2 3 4 5 6 7</p> <p>11 a) How often do you (honestly) brush your teeth each day?</p> <p><input type="text"/></p> <p>11 b) How many fillings do you have?</p> <p><input type="text"/>    Unsure <input type="checkbox"/></p> <p>12. Which do you think is the most important environmental issue that needs to be dealt with in the next 10 years?</p> <p><input type="checkbox"/> Air Pollution</p> <p><input type="checkbox"/> Global Warming</p> <p><input type="checkbox"/> Water Pollution</p> <p><input type="checkbox"/> Flooding</p> <p><input type="checkbox"/> Energy Sources</p> <p><input type="checkbox"/> Road Congestion</p> <p><input type="checkbox"/> Landfill Sites</p> <p><input type="checkbox"/> Other - state _____</p> <p>13. Do you think that YOU personally do enough to improve the environment.</p> <p><input type="checkbox"/> Yes    <input type="checkbox"/> No    <input type="checkbox"/> Unsure</p> <p>14. Which of the following does your household recycle? (Tick all that apply)</p> <p><input type="checkbox"/> Paper    <input type="checkbox"/> Glass    <input type="checkbox"/> Tins</p> <p><input type="checkbox"/> Plastic    <input type="checkbox"/> Other    <input type="checkbox"/> Nothing</p>	<p>15. What one thing do you think would improve your local environment?</p> <p><input type="checkbox"/> Less Traffic</p> <p><input type="checkbox"/> Cycle Paths</p> <p><input type="checkbox"/> Less Litter</p> <p><input type="checkbox"/> Playgrounds</p> <p><input type="checkbox"/> More Trees</p> <p><input type="checkbox"/> More Shops</p> <p><input type="checkbox"/> More Sports Facilities</p> <p><input type="checkbox"/> Other - state _____</p> <p>16. What best describes the kind of building you live in?</p> <p><input type="checkbox"/> Detached House/ Bungalow</p> <p><input type="checkbox"/> Semi Detached</p> <p><input type="checkbox"/> Terrace</p> <p><input type="checkbox"/> Apartment/ Flat</p> <p><input type="checkbox"/> Other _____</p> <p>17. How do you usually travel to school?</p> <p><input type="checkbox"/> Walk    <input type="checkbox"/> Bus    <input type="checkbox"/> Car</p> <p><input type="checkbox"/> Cycle    <input type="checkbox"/> Rail    <input type="checkbox"/> Other</p> <p>18. If you had £1000 to give to a charity of your choice what kind of organisation would you choose?</p> <p><input type="checkbox"/> Arts</p> <p><input type="checkbox"/> Children</p> <p><input type="checkbox"/> Education/Youth development</p> <p><input type="checkbox"/> Environment</p> <p><input type="checkbox"/> Health</p> <p><input type="checkbox"/> International Aid</p> <p><input type="checkbox"/> Law/Justice</p> <p><input type="checkbox"/> Sport</p> <p><input type="checkbox"/> Wildlife/ Animals</p> <p><input type="checkbox"/> Other _____</p> <p>19. Estimate how often you contact your friends each week:</p> <p><input type="checkbox"/> Text    <input type="checkbox"/> e-mail</p> <p><input type="checkbox"/> Telephone (landline)</p> <p><input type="checkbox"/> Telephone (mobile)</p> <p>20. Estimate the 3 angles given by the online questionnaire.</p> <p><input type="text"/>    <input type="text"/>    <input type="text"/></p>
--	---	---

This resource is from the CensusAtSchool project at [www.censusatschool.ntu.ac.uk](http://www.censusatschool.ntu.ac.uk)

# Ways of describing category-based data

- ☐ Frequency tables
- ☐ Bar charts
- ☐ Pie charts
- ☐ Two-way frequency tables
- ☐ Percentage frequencies of rows and columns
- ☐ Multi series bar charts

# Frequency table for Favourite Food Type (question 5)

Also called frequency

Created by dividing the count by the total and multiplying by 100%

Favourite Food Type	Count	Percentage
Carbohydrate	272	40%
Protein	153	23%
Dairy	151	22%
Fruit and Vegetables	97	14%
<b>Total</b>	<b>673</b>	<b>100%</b>

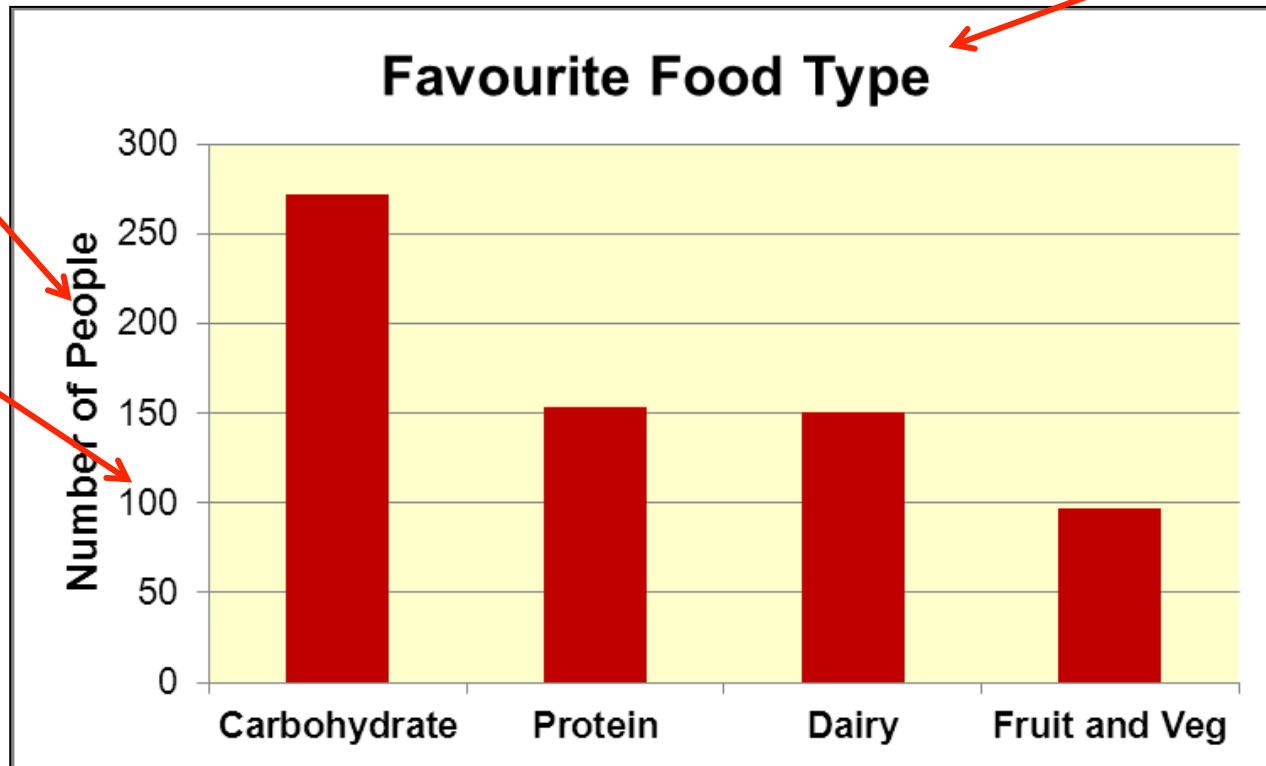
# Bar chart for Favourite Food Type

Also called a **column chart** or a **histogram**

Chart title

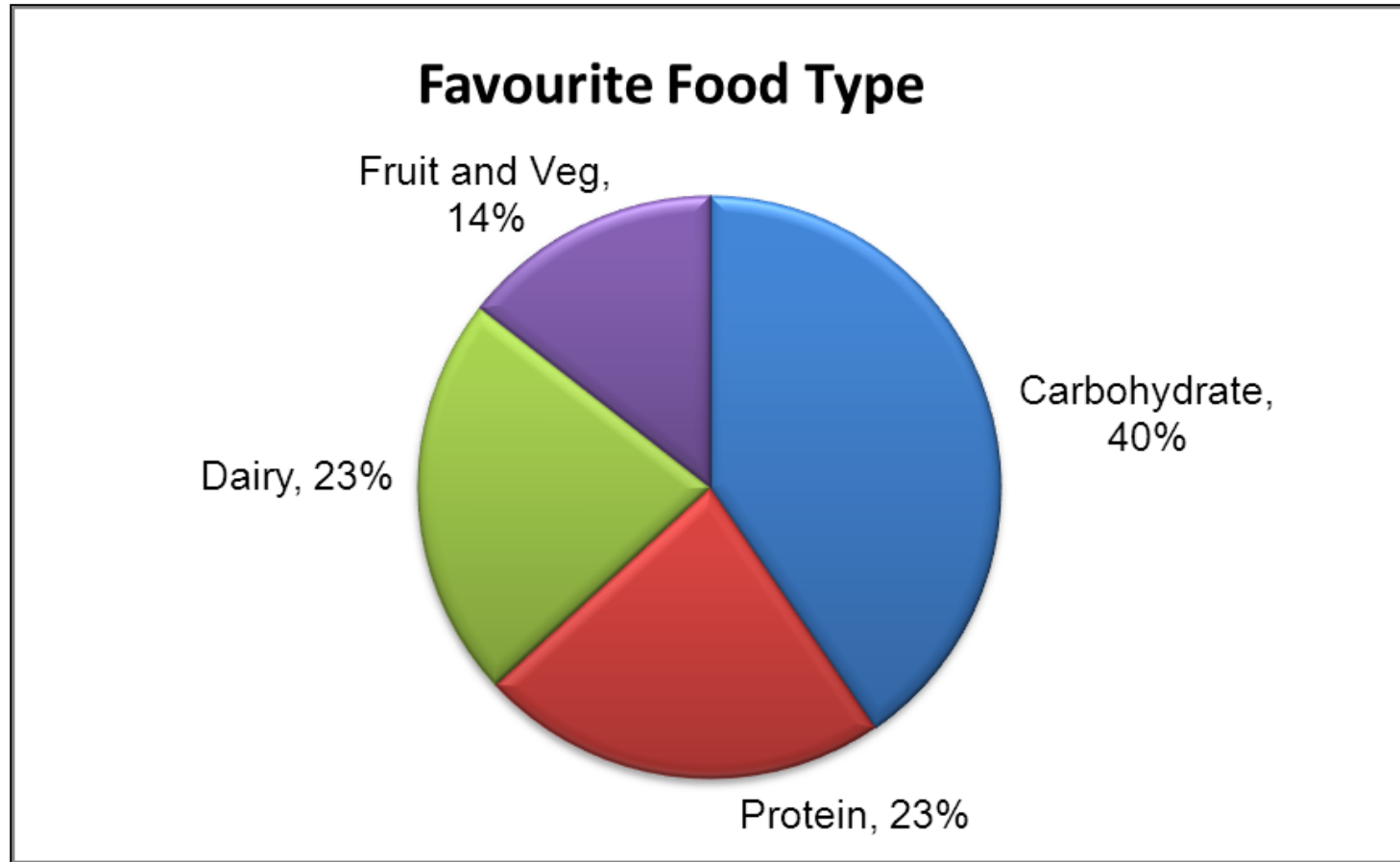
Axis title

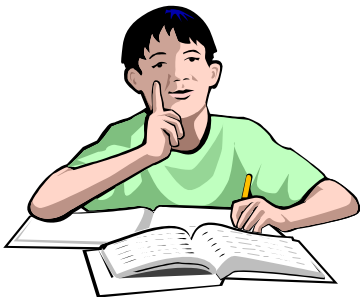
Axis labels





# Pie Chart for Favourite Food Type

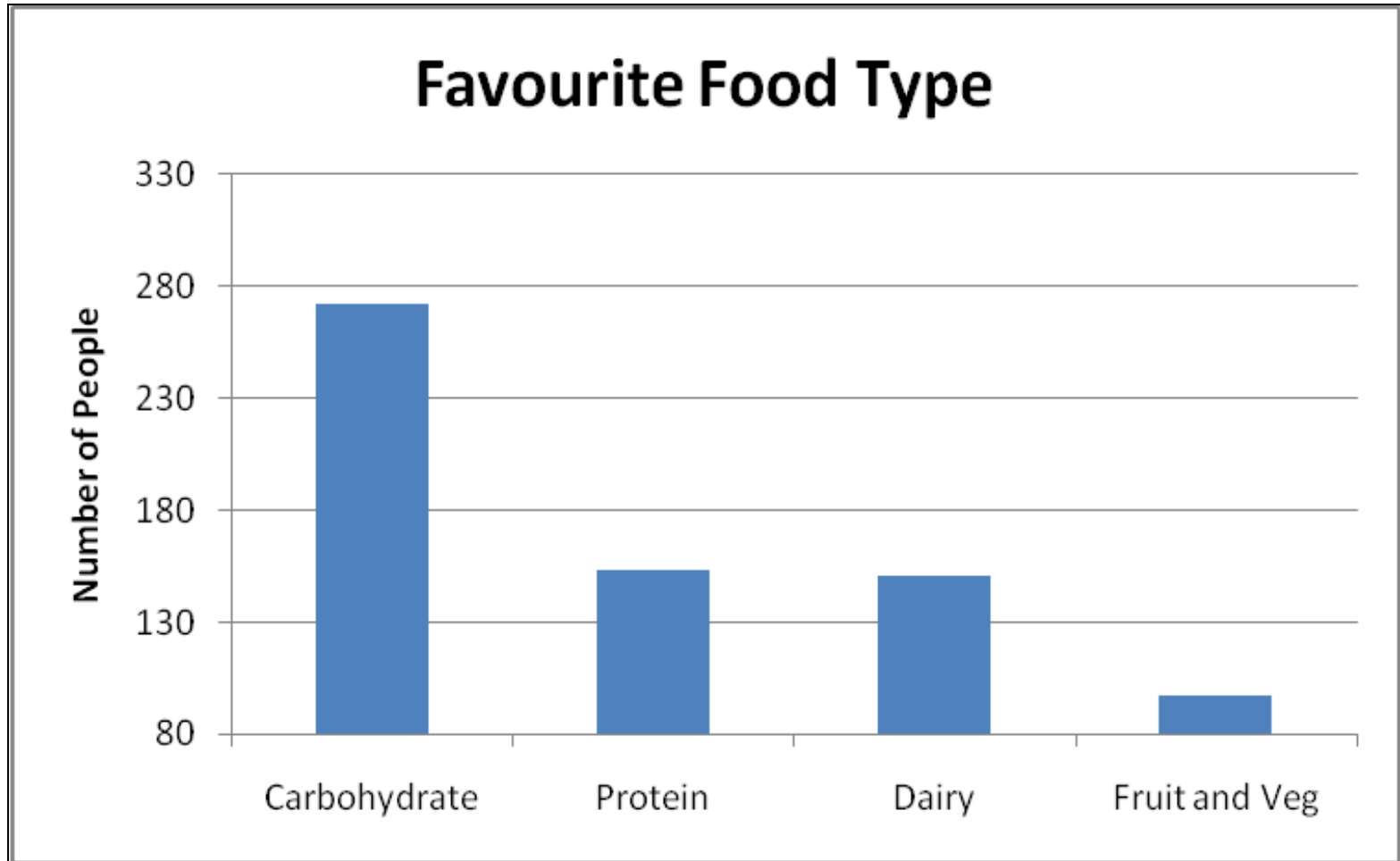




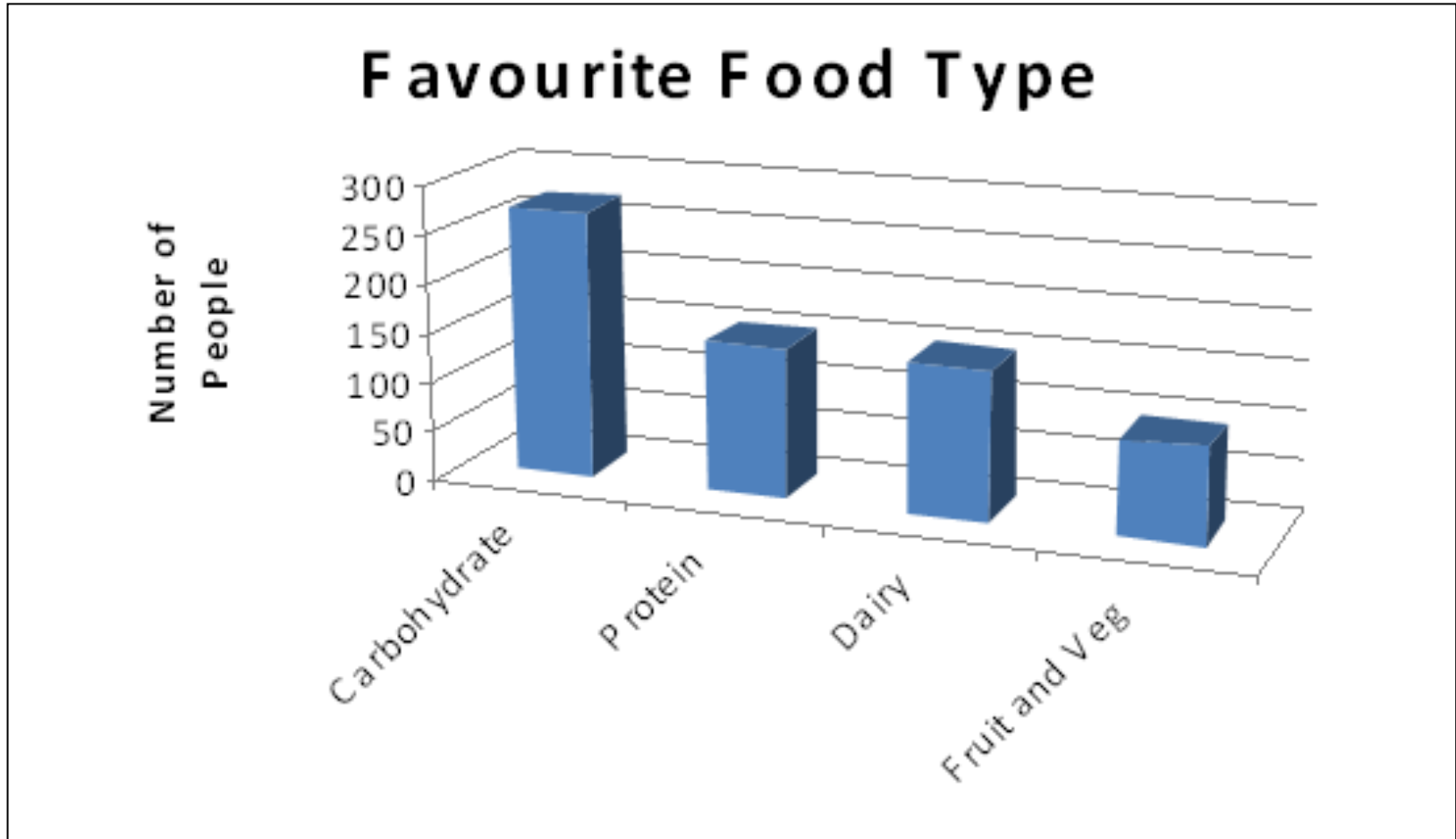
# Activity

- ☐ Consider the four charts on the handout
- ☐ Each has at least one aspect that could be improved
- ☐ Spend no more than one minute on each discussing what the **main** improvement is that could be made

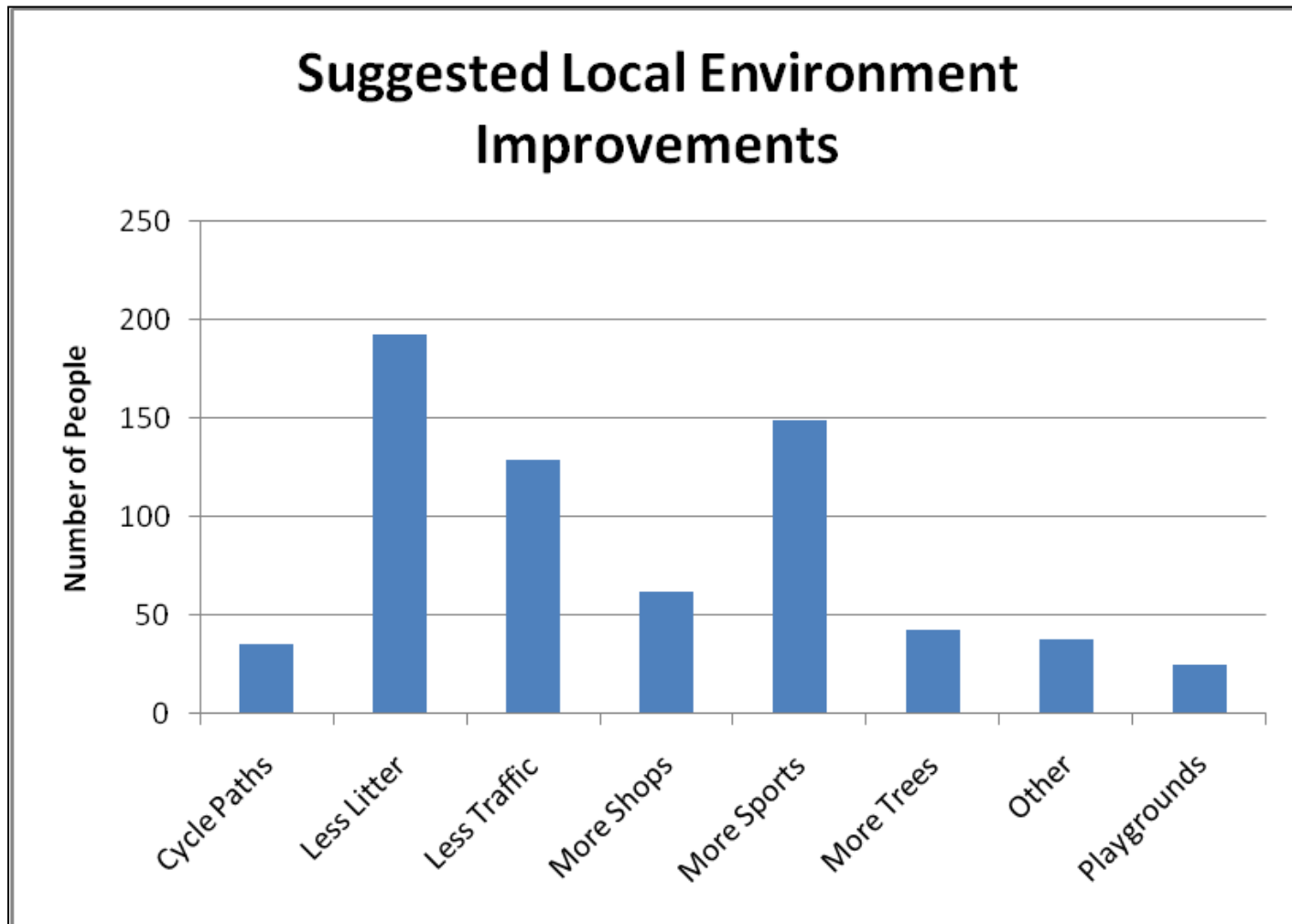
# Chart 1



# Chart 2

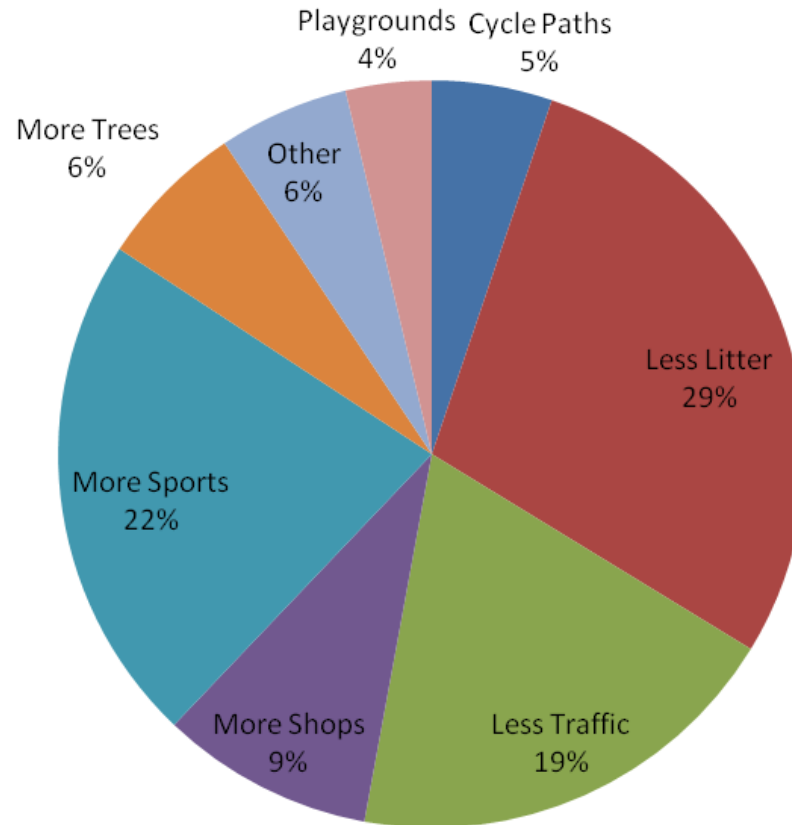


# Chart 3

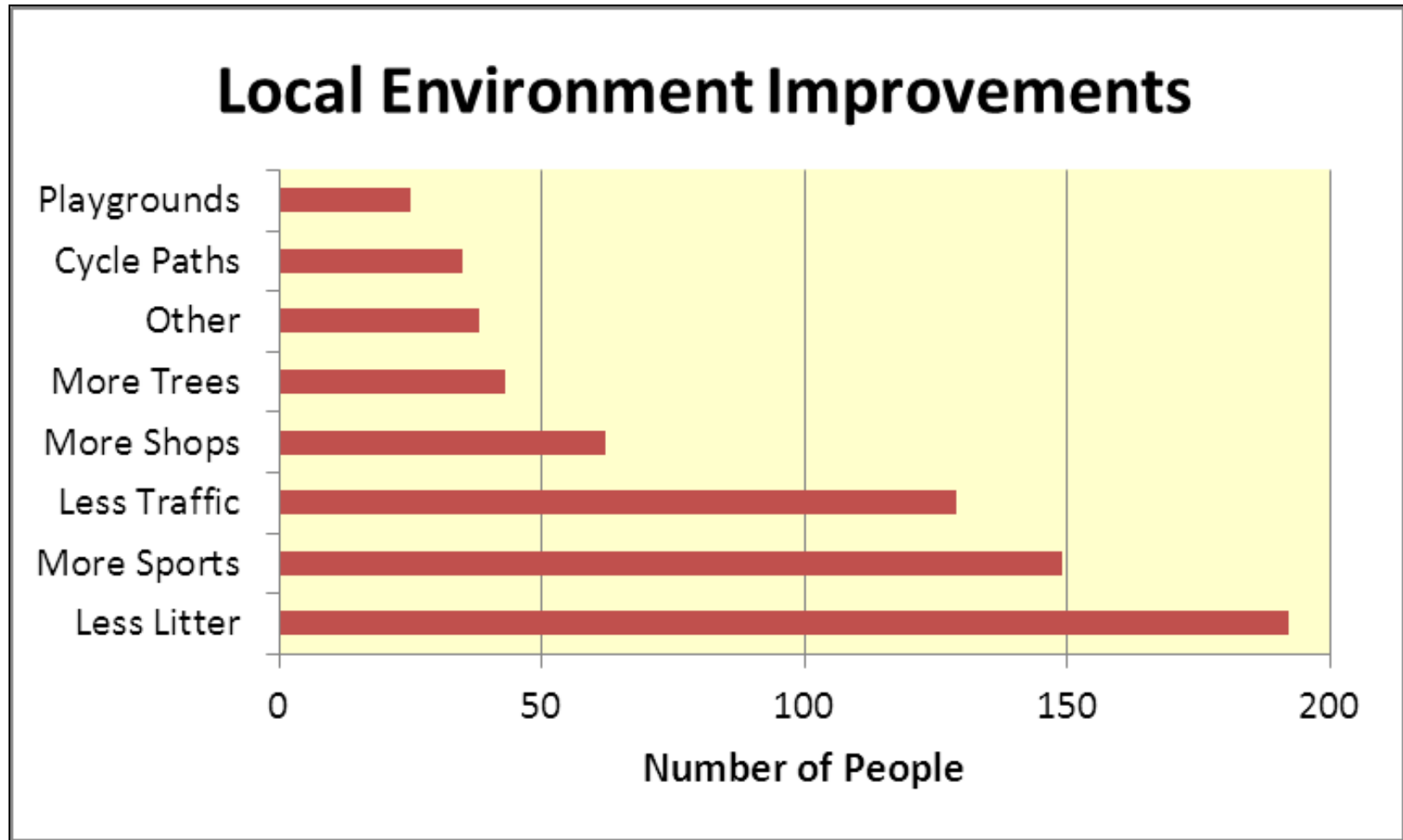


# Chart 4

## Local Environment Improvements



# Clearer bar chart for Chart 3

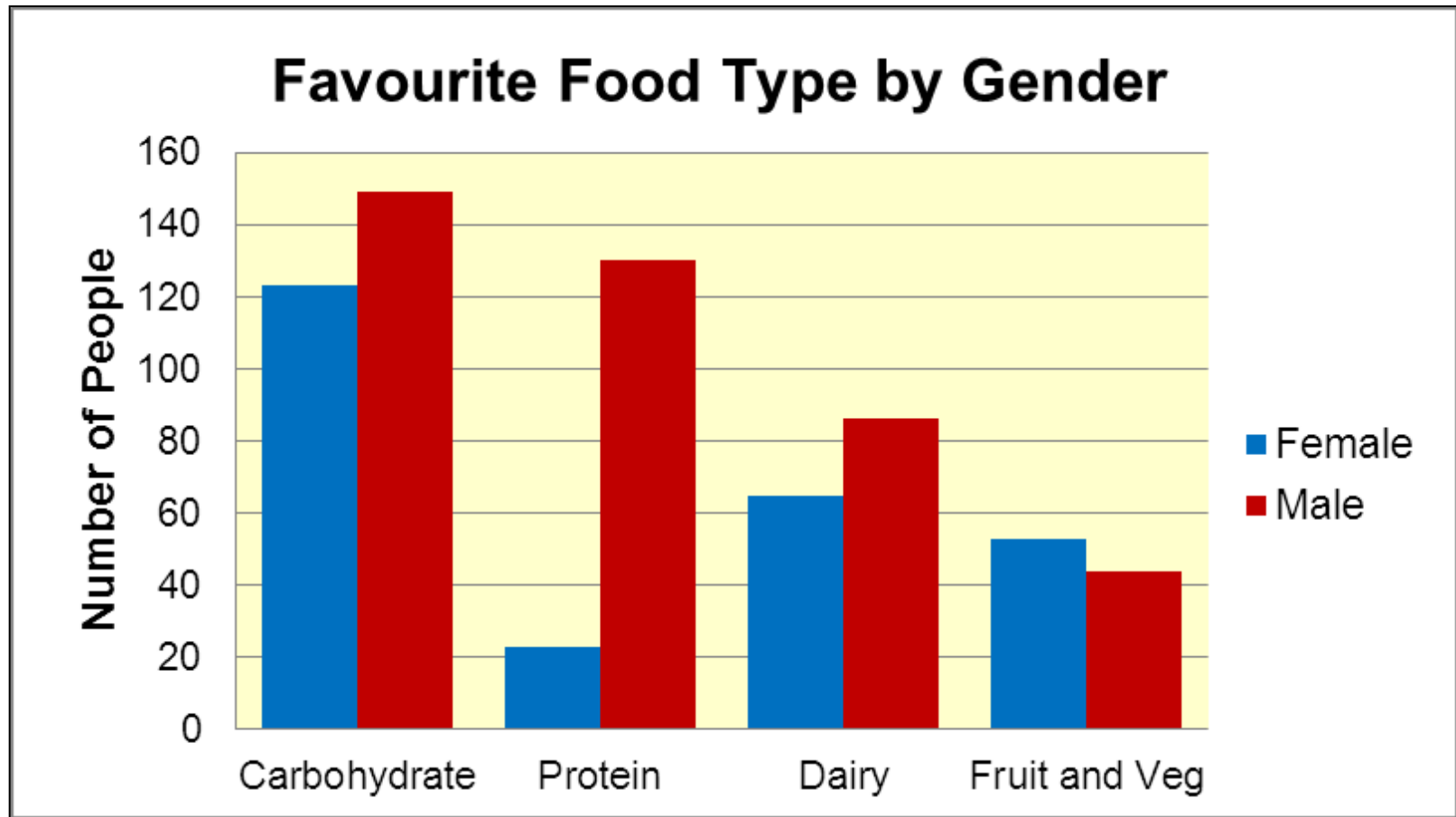


# Two-way frequency table of Favourite Food Type by Gender

Favourite Food Type	Female	Male	Total
Carbohydrate	123	149	272
Protein	23	130	153
Dairy	65	86	151
Fruit and Veg	53	44	97
<b>Total</b>	<b>264</b>	<b>409</b>	<b>673</b>



# Bar chart for Favourite Food Type by Gender



# Percentage frequency table for Favourite Food Type by Gender

Divide the count by the column total then  $\times 100\%$

Column %'s all sum to 100%

<b>Favourite Food Type</b>	<b>Female (%)</b>	<b>Male (%)</b>	<b>Total (%)</b>
Carbohydrate	47	36	<b>40</b>
Protein	9	32	<b>23</b>
Dairy	25	21	<b>22</b>
Fruit and Veg	20	11	<b>14</b>
<b>Total</b>	<b>100</b>	<b>100</b>	<b>100</b>

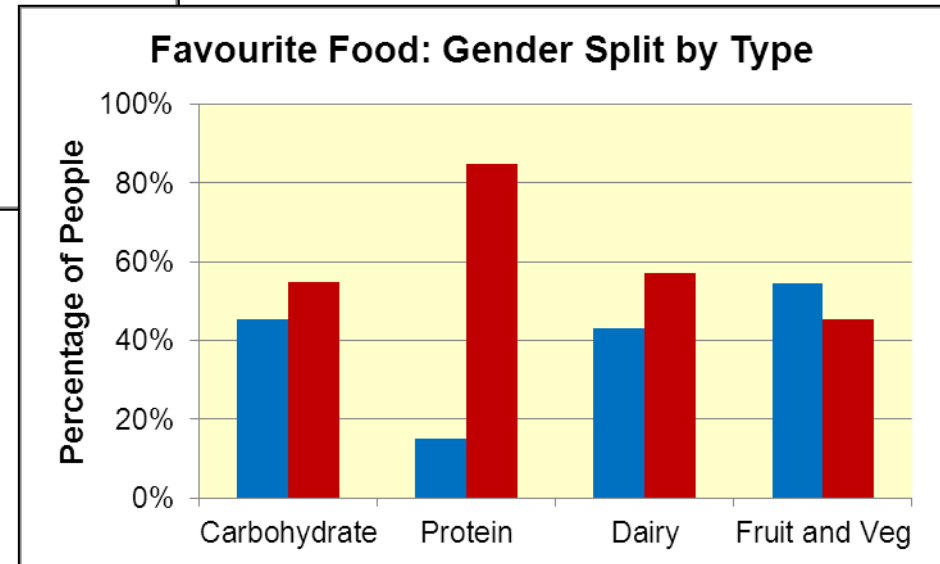
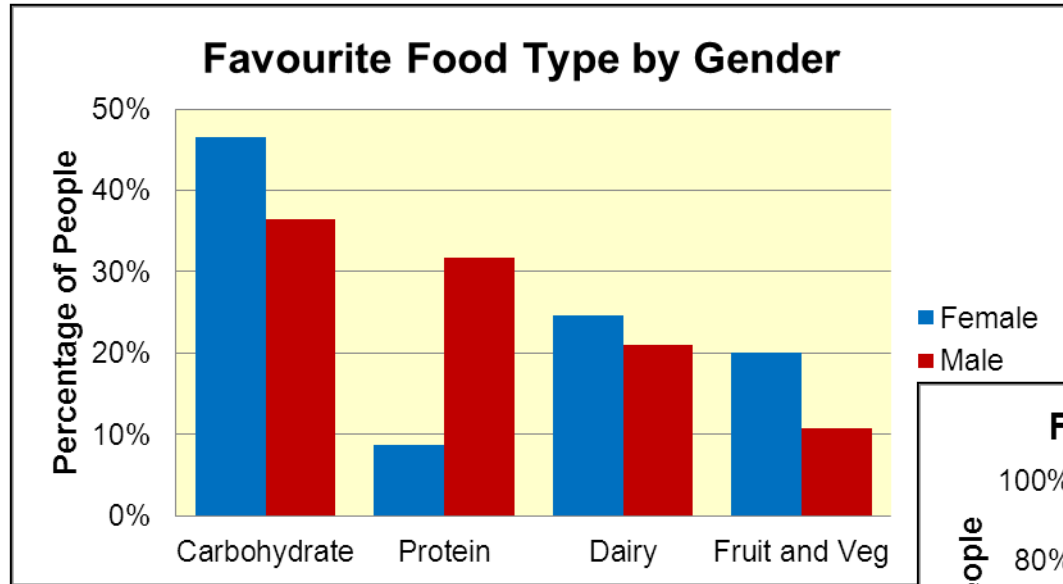
# Percentage frequency table for Favourite Food Type by Gender

Row %'s all sum to 100%

Divide  
the count  
by the  
row total  
then  
 $\times 100\%$

<b>Favourite Food Type</b>	<b>Female (%)</b>	<b>Male (%)</b>	<b>Total (%)</b>
Carbohydrate	45	55	<b>100</b>
Protein	15	85	<b>100</b>
Dairy	43	57	<b>100</b>
Fruit and Veg	55	45	<b>100</b>
<b>Total</b>	<b>40</b>	<b>60</b>	<b>100</b>

# Percentage bar charts for Favourite Food Type by Gender



# Using tables and charts for category data

- ❑ Frequency tables: OK for smaller data sets – don't have too many numbers in your write-up
- ❑ Bar charts: series of values:
  - Always use 2D
  - Put items in descending order when appropriate
- ❑ Pie charts: parts of a whole
- ❑ Multiple series bar charts: better to use percentage frequencies than frequencies
- ❑ Best option: a suitable chart with a verbal description

# Describing scale-based data

- ☐ Grouped frequency tables
- ☐ Histograms
- ☐ Measures of central tendency:
  - Mean
  - Mode
  - Median
- ☐ Measures of spread:
  - Range
  - Variance
  - Inter-quartile range
- ☐ Box plots
- ☐ Scatter plots

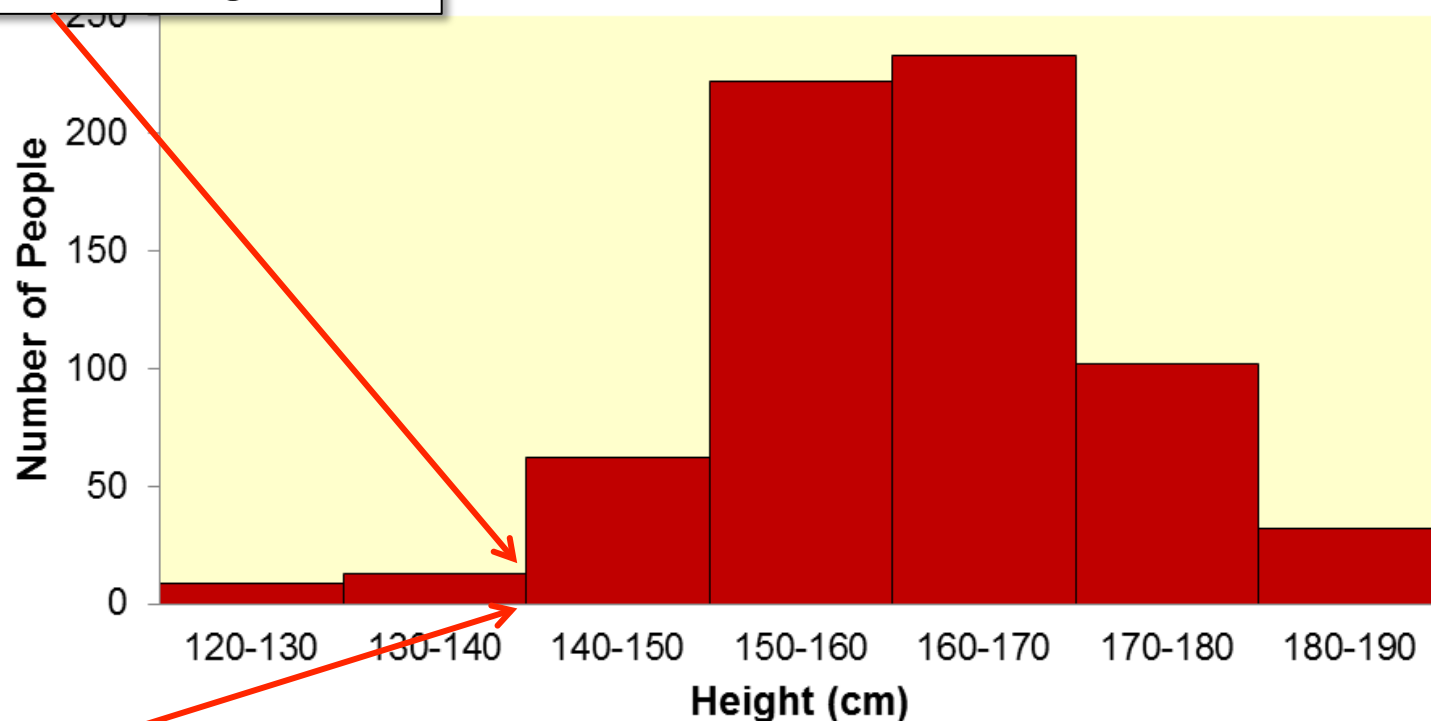
# Grouped frequency distribution for Height

Height	Frequency	Percentage
$\geq 120$ & $< 130$	9	1.3%
$\geq 130$ & $< 140$	13	1.9%
$\geq 140$ & $< 150$	62	9.2%
$\geq 150$ & $< 160$	222	33.0%
$\geq 160$ & $< 170$	233	34.6%
$\geq 170$ & $< 180$	102	15.2%
$\geq 180$ & $< 190$	32	4.8%
<b>Total</b>	<b>673</b>	<b>100.0%</b>

# Histogram of Height

Bars drawn together to emphasise they represent intervals, not categories

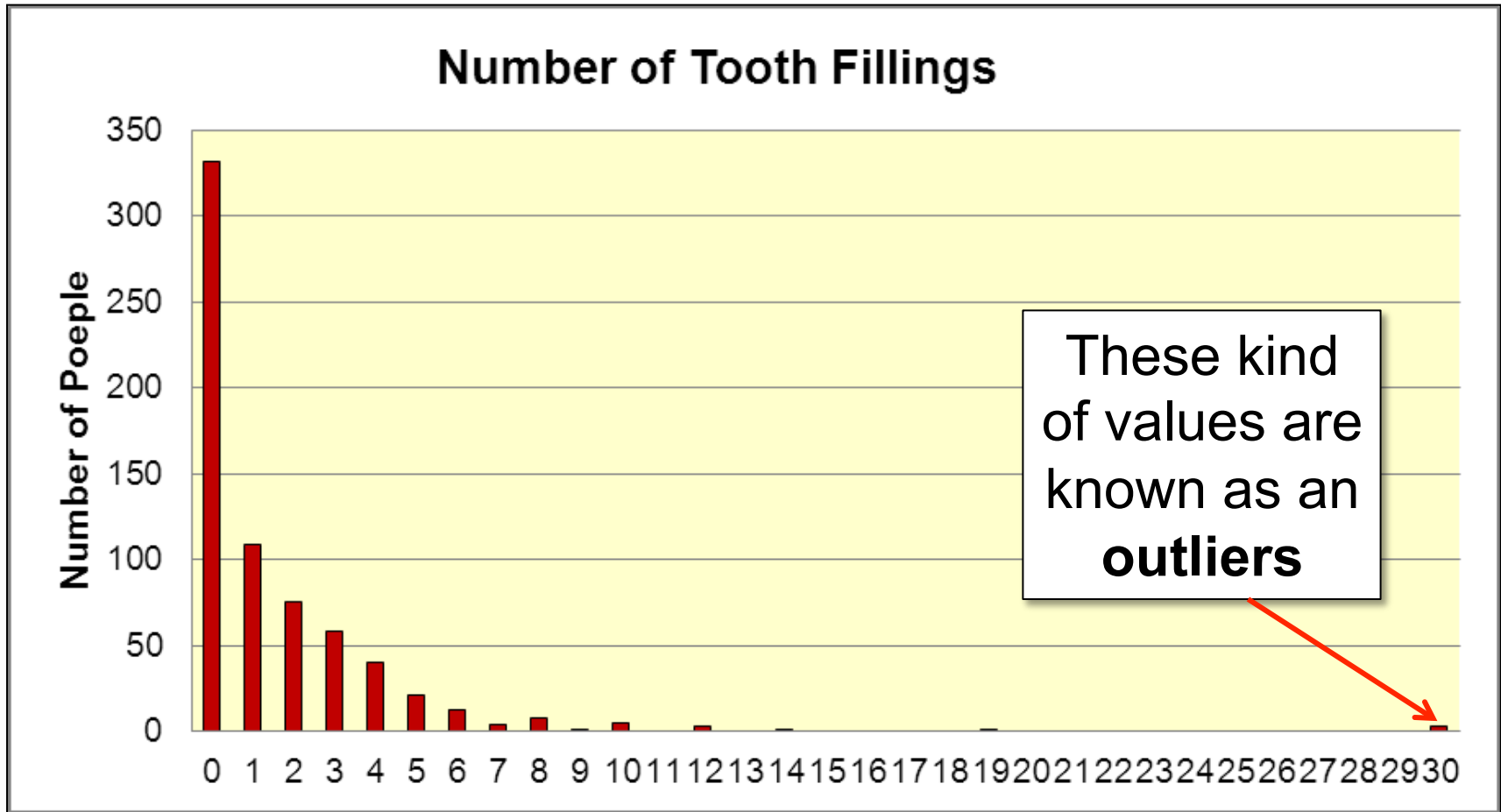
Distribution of Height



Border value is 140



# Frequency distribution of Tooth Fillings



# Important types of statistic

## 1. Measures of middle value

- ☐ **Mean (or average):** Add all the data values together and divide by the number of values
- ☐ **Mode:** The most popular data value, category or interval
- ☐ **Median:** The middle data value when the values are put in order

# Mean

The mean of a data set is a measure of its middle value.

**Example:** The number of nuclear power stations in various countries in 1989.

Country	Number	Country	Number
Canada	22	Spain	10
France	52	Sweden	12
Japan	43	UK	41
South Korea	9	USA	119
Soviet Union	73	West Germany	23

To calculate the mean, add all the data values together and divide by the number of values.

$$\bar{X} = \frac{22 + 52 + 43 + 9 + 73 + 10 + 12 + 41 + 119 + 23}{10} = \frac{384}{10} = 38.4$$

# Median

**Median:** Place the data values in ascending order:

9, 10, 12, 22, 23, 41, 43, 52, 73, 119

The median is the average of the 5<sup>th</sup> & 6<sup>th</sup> values, i.e.

$$\frac{23 + 41}{2} = \frac{64}{2} = 32$$

The data is too spread out to calculate a **mode** (as no value is repeated) but we could divide it into intervals and calculate the modal interval.

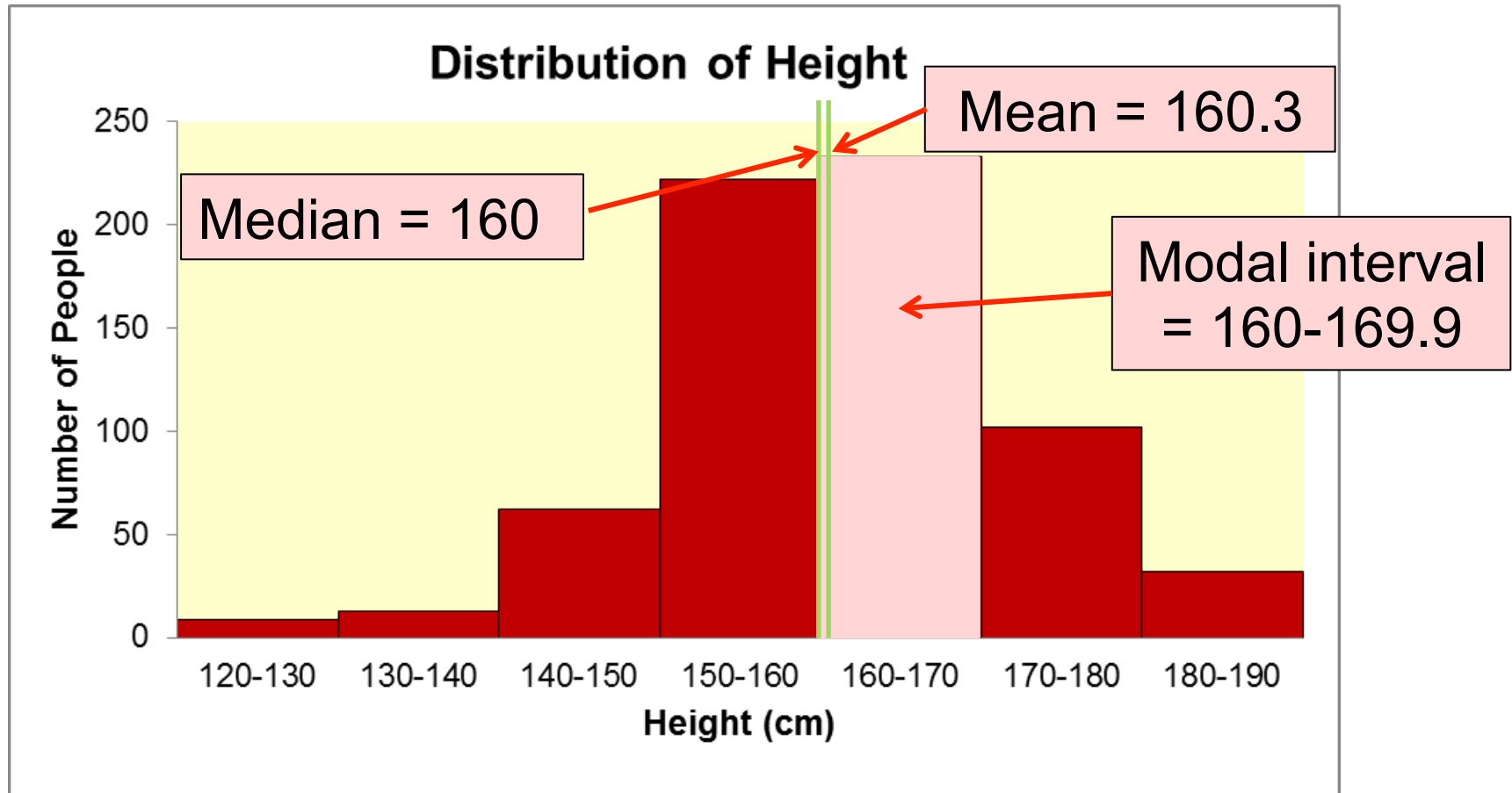
# Calculating the mode using data intervals

- ❑ First we need to decide on the size of the intervals
- ❑ As our data values could be any positive whole number and they range from 9 to 119, we could choose intervals of size 25 in order to obtain several frequencies higher than 1

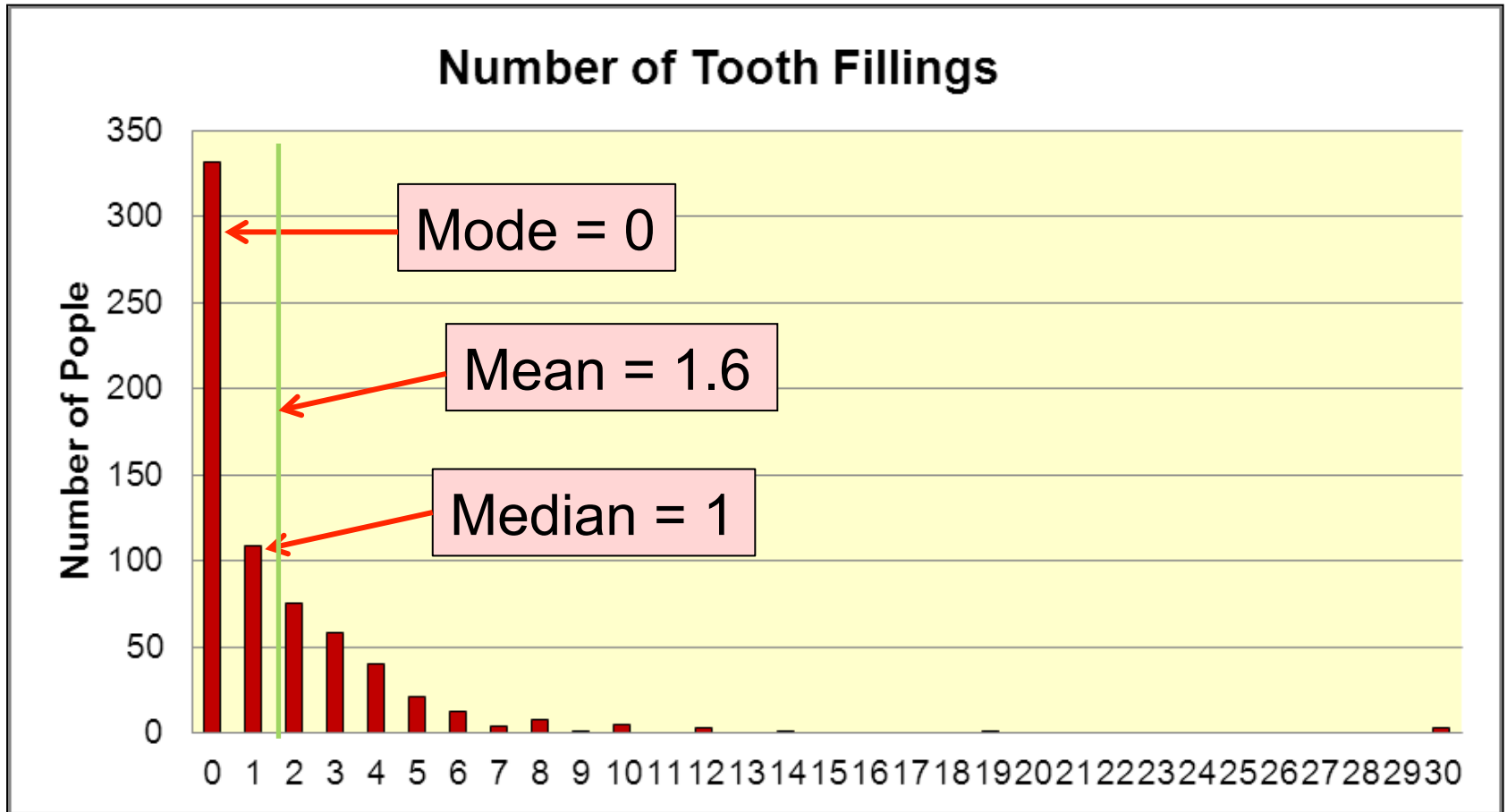
Interval	Data values within interval	Frequency (number of data values)
0 – 24	9, 10, 12, 22, 23	5
25 – 49	41, 43	2
50 – 74	52, 73	2
75 – 99		0
100 – 124	119	1

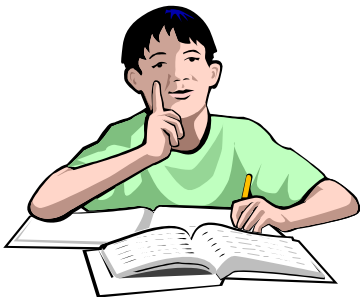
The most popular, or modal, interval is therefore 0 – 24.

# Frequency distribution of Height with measures of middle values indicated



# Frequency distribution for Tooth Fillings with measures of middle values indicated





# Activity

- ☐ The following three data sets were obtained and a histogram (or bar chart) has been produced
- ☐ In each case, discuss whether the mean, median or mode provides the best measure of the middle value



# Data set 1

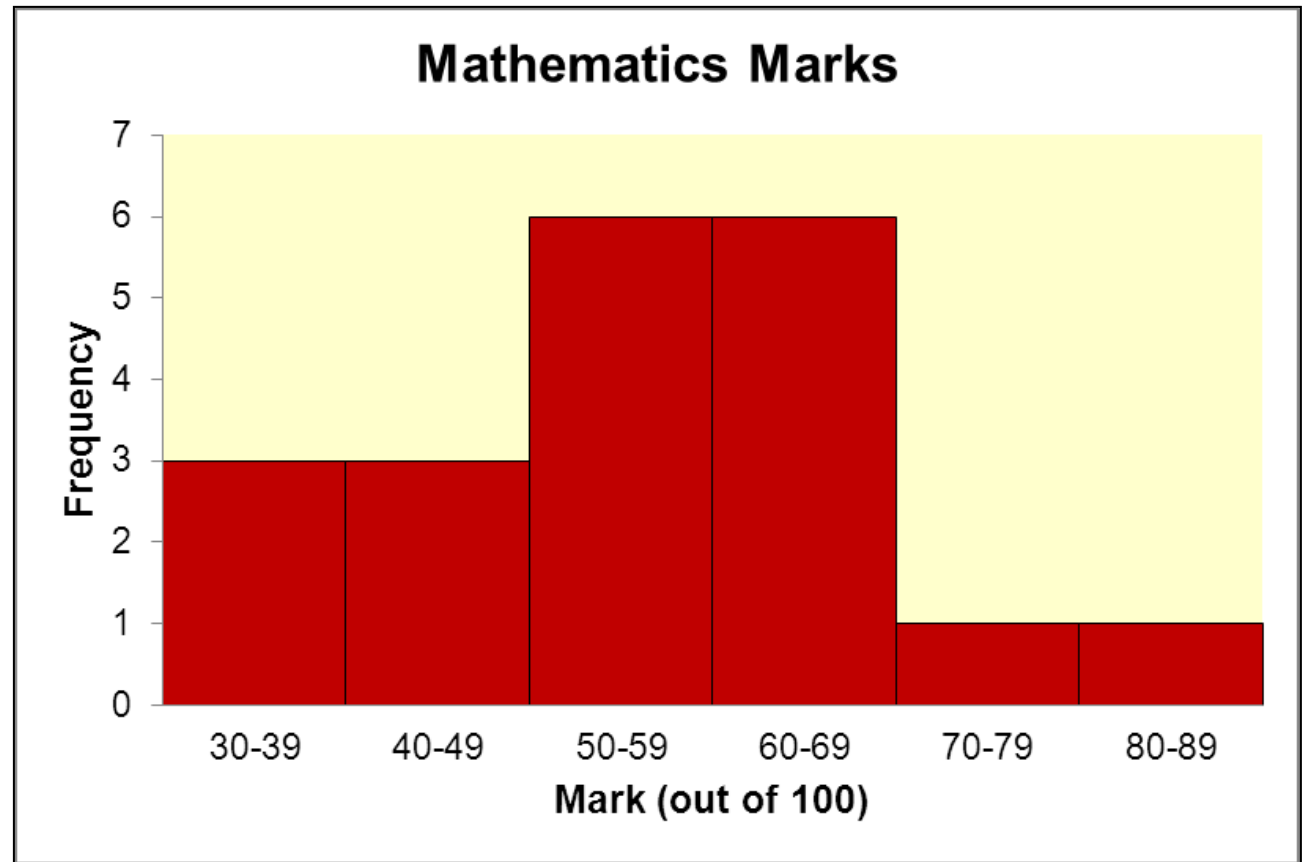
30	35	37	40	40	49	51	54	54	55
57	58	60	60	62	62	65	67	74	89

Marks (out of 100) for a mathematics exam for 20 students, sorted into intervals of size 10

Mean = 55.0

Median = 56.0

Modal interval = both 50-59 and 60-69



# Data set 2

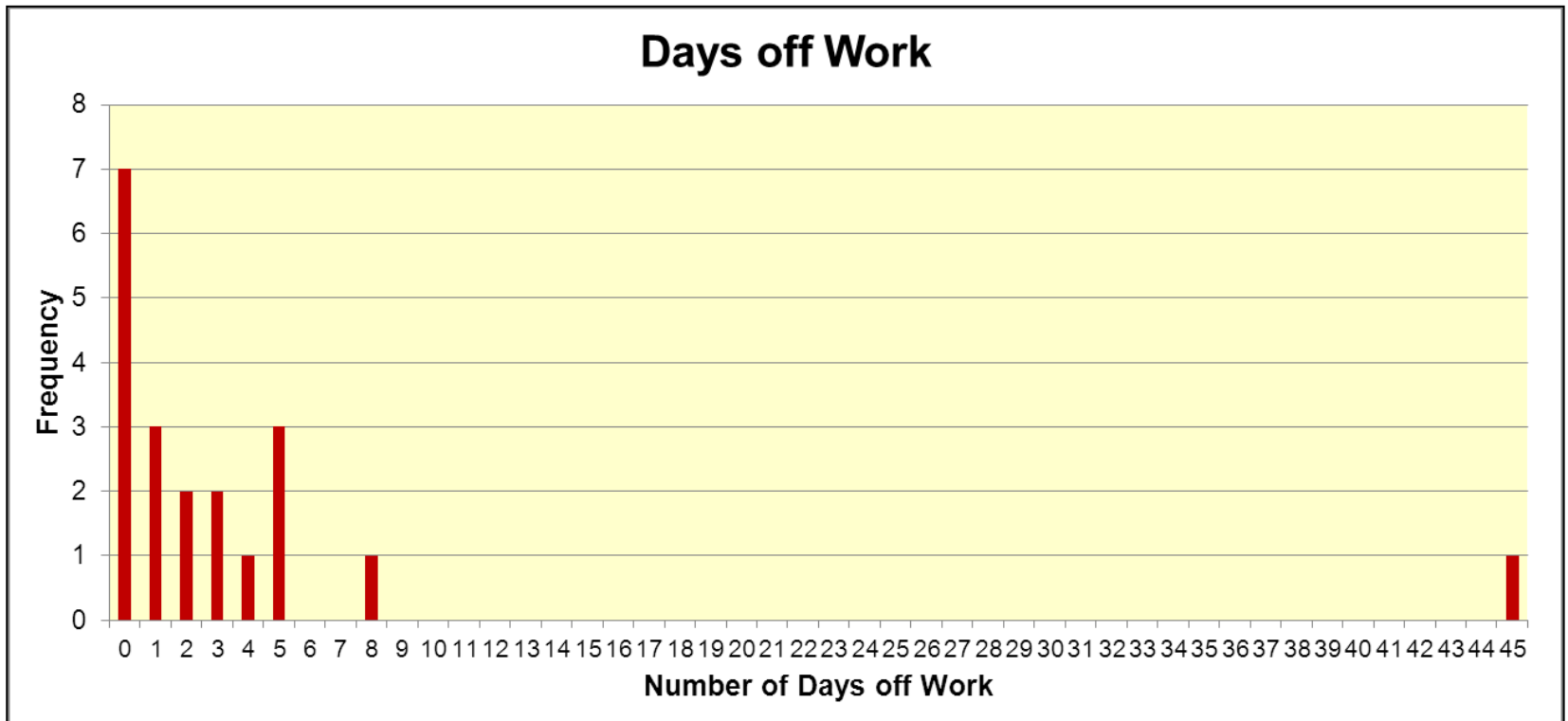
The numbers of days work missed by 20 workers in one year

Mean = 4.3

Median = 1.5

Mode = 0

0	0	0	0	0	0	0	1	1	1
2	2	3	3	4	5	5	5	8	45



# Data set 3

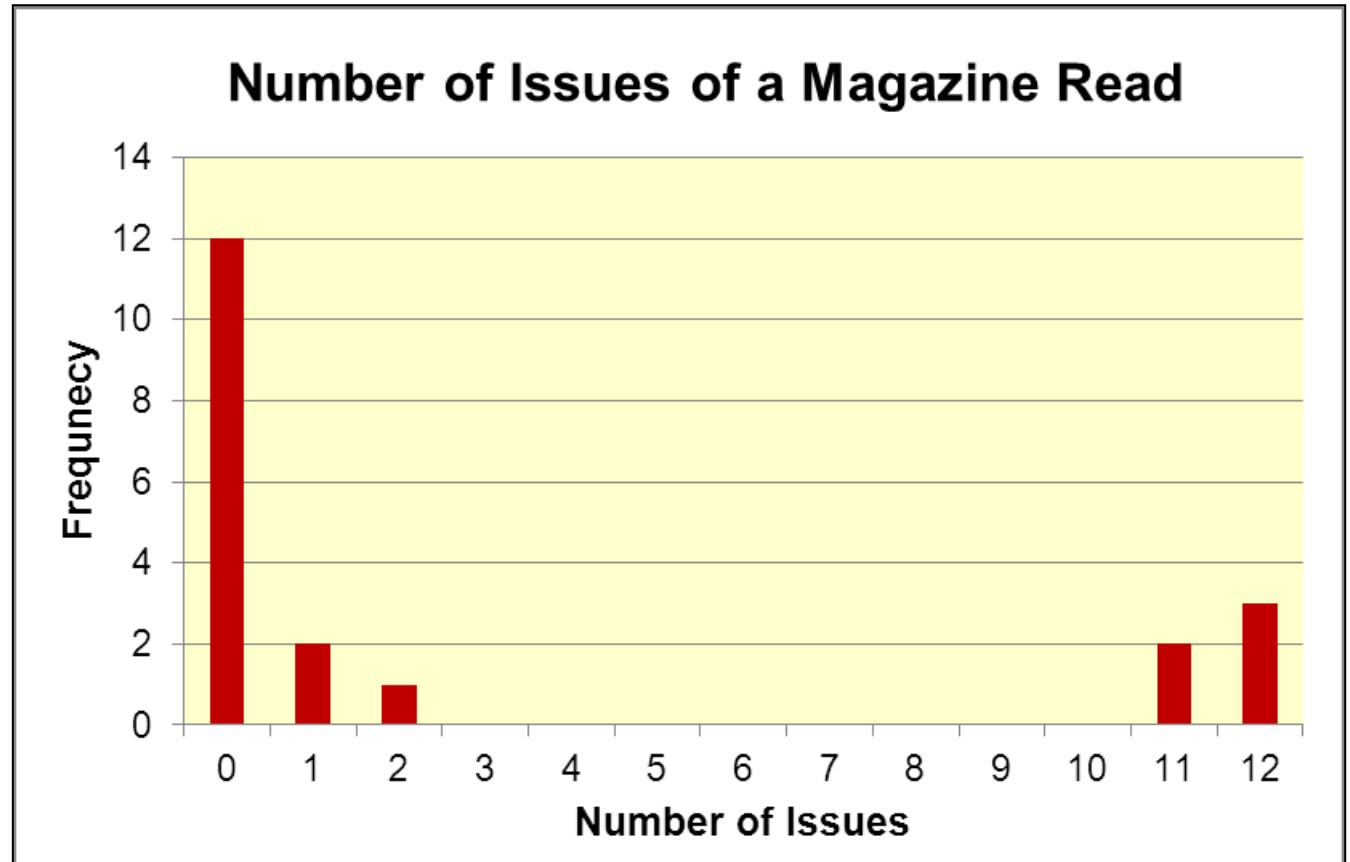
The number of issues of a particular monthly magazine read over a year by 20 people

Mean = 3.1

Median = 0

Mode = 0

0	1	11	0	0	0	2	12	0	0
12	1	0	0	0	0	12	0	11	0



# Important types of statistic

## 2. Measures of spread

- ❑ **Range:** The difference between the highest and lowest data values
- ❑ **Standard deviation:** A measure of how much the data values vary from the average value
- ❑ **Inter-quartile range:** The distance between the 25% point in the data values (**lower quartile**) and the 75% point in the data values (**upper quartile**) when they are put in ascending order
- ❑ **Example:** 22, 52, 43, 9, 73, 10, 12, 41, 119, 23

Recall the mean,  
 $\bar{X}$ , was 38.4

# Calculating the range and standard deviation

$X$	$X - \bar{X}$	$(X - \bar{X})^2$
22	-16.4	268.96
52	13.6	184.96
43	4.6	21.16
9	-29.4	864.36
73	34.6	1197.16
10	-28.4	806.56
12	-26.4	696.96
41	2.6	6.76
119	80.6	6496.36
23	-15.4	237.16
<b>Sum</b>		<b>10780.4</b>

□ **Range** =  $119 - 9 = 110$   
(highest value – lowest value)

□ **Sample variance**

$$= \sum (X - \bar{X})^2 / n - 1 = 10780.4 / 9 = 1197.8$$

□ **Sample standard deviation**

$$= \sqrt{\text{Sample Variance}} \\ = \sqrt{1197.8} = 34.6$$

□ For the **population variance** and **population standard deviation** we divide by  $n$  instead of  $n - 1$

# Calculating the inter-quartile range

Place the data values in ascending order:

9, 10, 12, 22, 23, 41, 43, 52, 73, 119

**Lower quartile** = 25% of way through data = 10 + 1/4<sup>th</sup> value

$$= 2\frac{3}{4}^{\text{th}} \text{ value} = 2^{\text{nd}} \text{ value} + \frac{3}{4} \text{ of } (3^{\text{rd}} \text{ value} - 2^{\text{nd}} \text{ value})$$

$$= 10 + \frac{3}{4} (12 - 10) = 10 + 3 \times 2/4 = 11.5$$

**Upper quartile** = 75% of way through data = 3 × 10 + 1/4<sup>th</sup> value

$$= 8\frac{1}{4}^{\text{th}} \text{ value} = 8^{\text{th}} \text{ value} + \frac{1}{4} \text{ of } (9^{\text{th}} \text{ value} - 8^{\text{th}} \text{ value})$$

$$= 52 + \frac{1}{4} (73 - 52) = 52 + 21/4 = 57.25$$

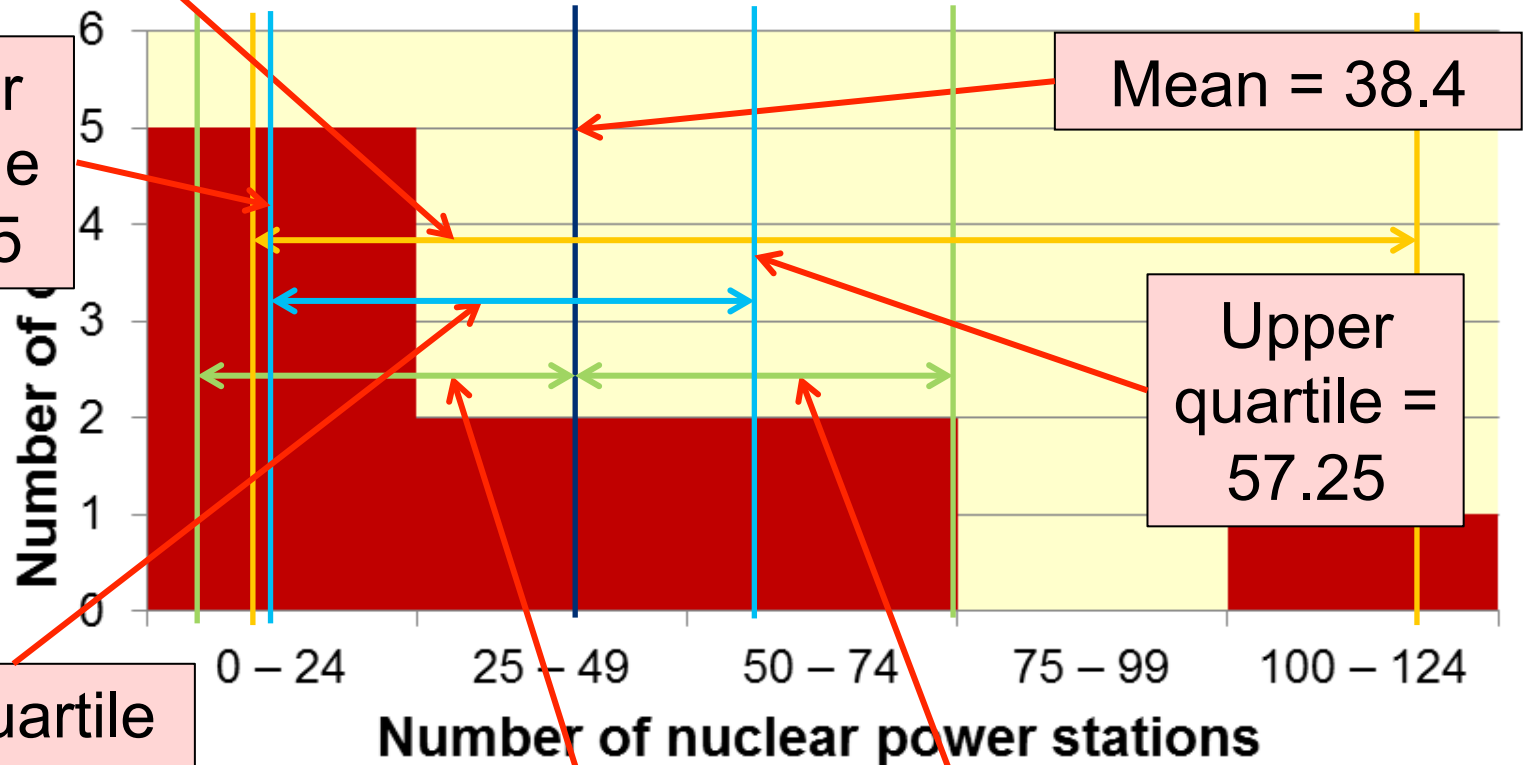
**Inter-quartile range** = difference between upper quartile and lower quartile = 57.25 - 11.5 = 45.75

Range = 110

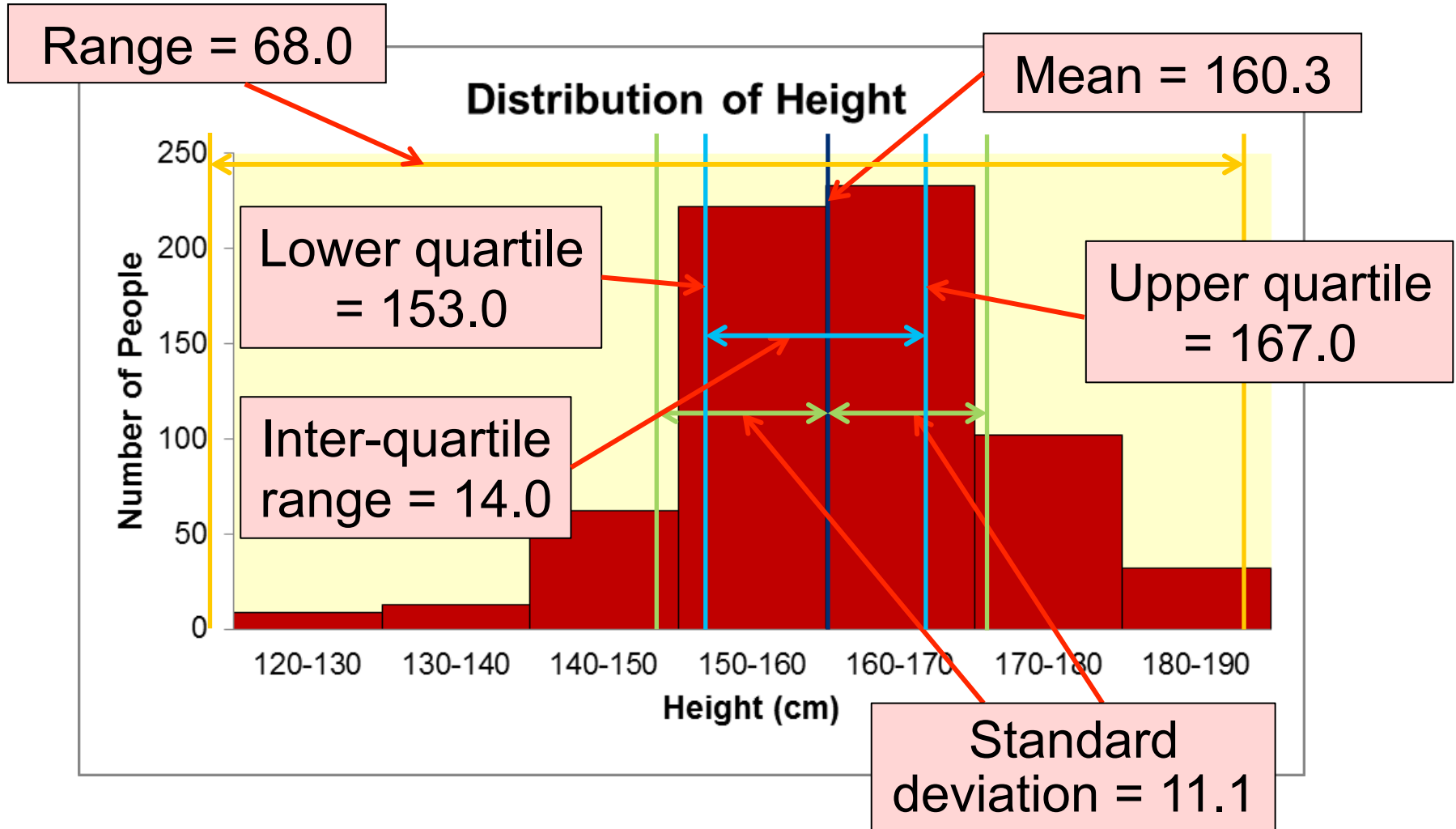
Lower  
quartile  
= 11.5

Inter-quartile  
range = 45.75

## Distribution of nuclear power stations

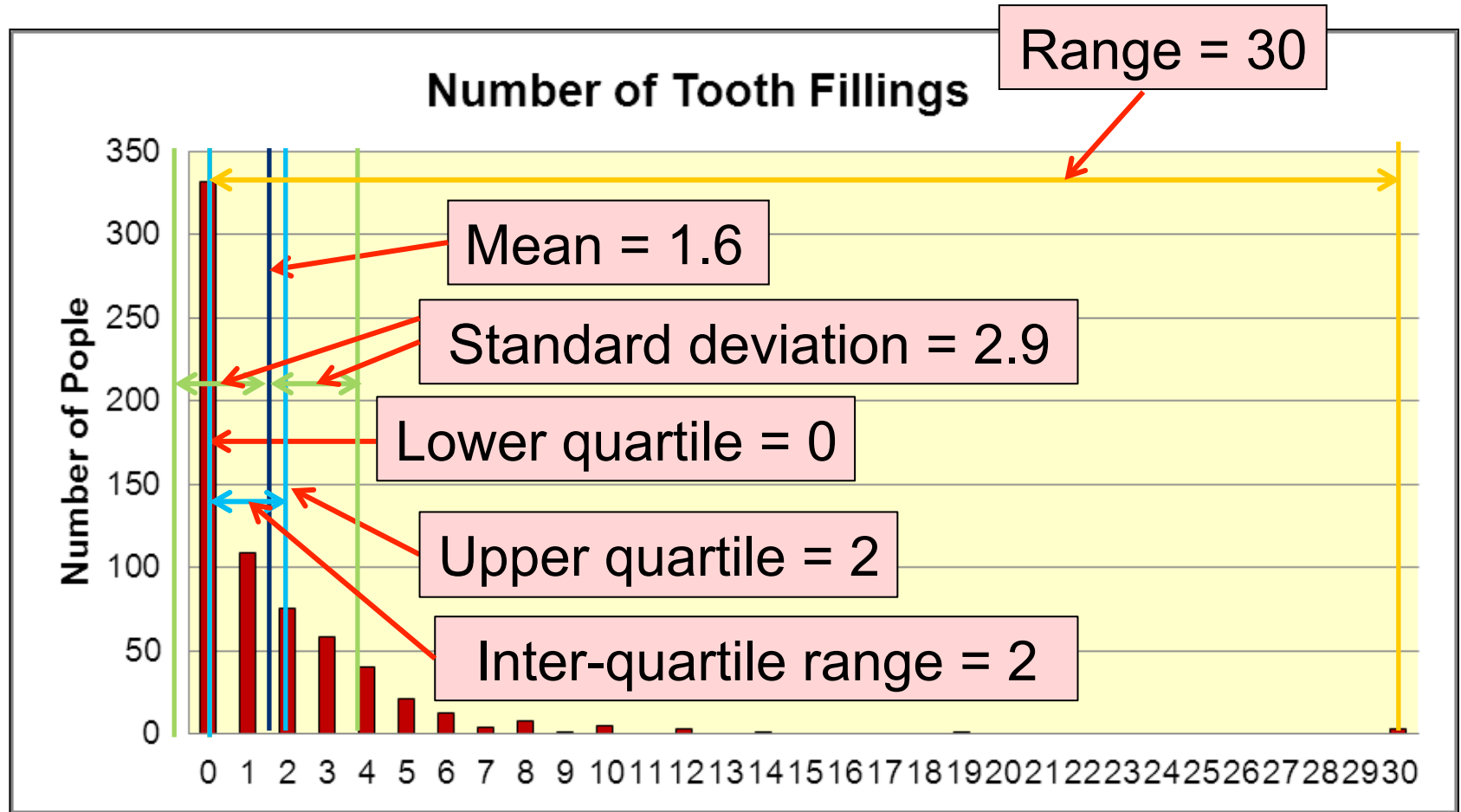


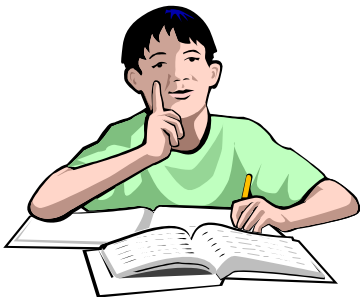
# Frequency distribution of Height with measures of spread indicated





# Frequency distribution for Tooth Fillings with measures of spread indicated





# Activity

❑ For Data Set 3:

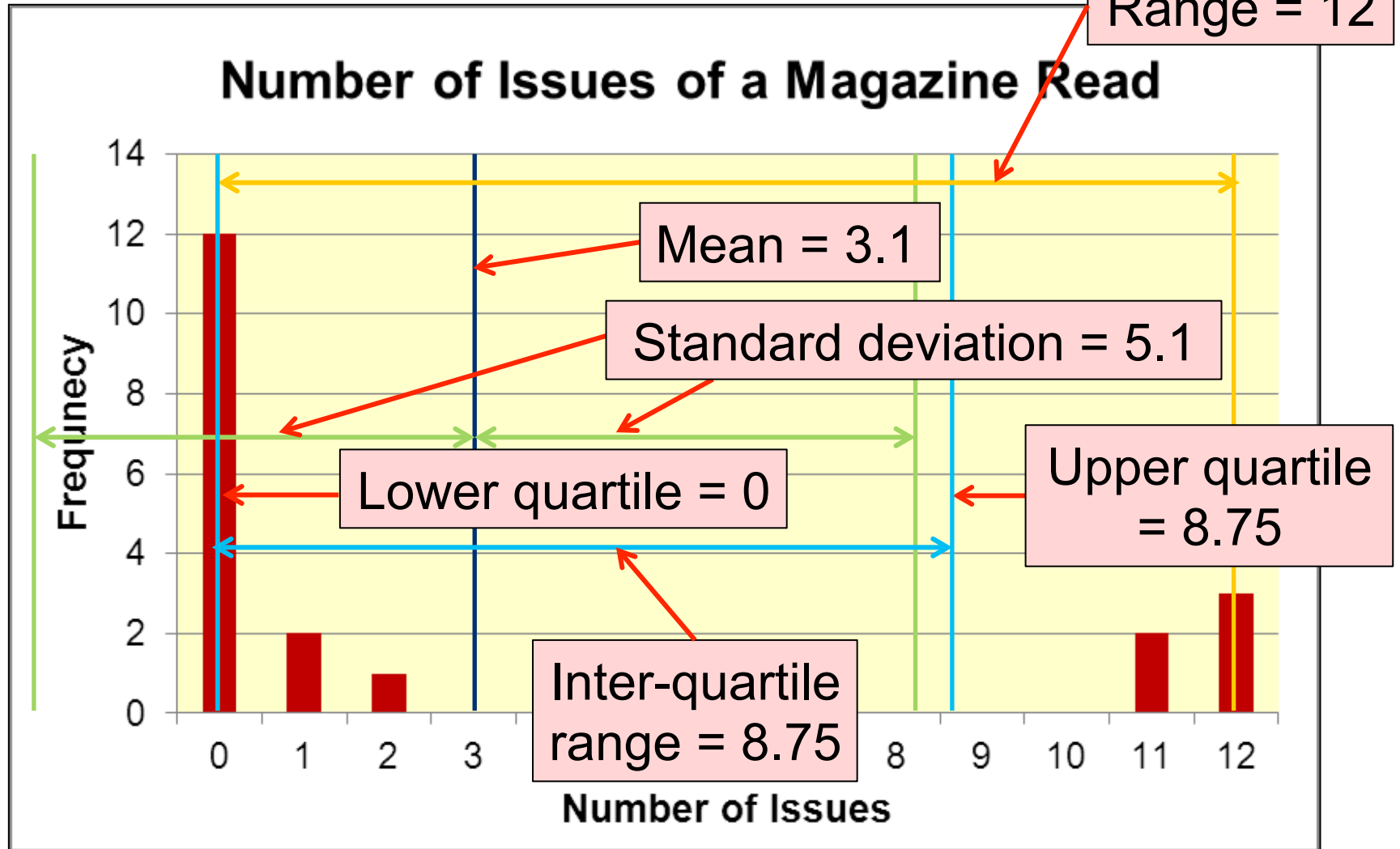
0	1	11	0	0	0	2	12	0	0
12	1	0	0	0	0	12	0	11	0

❑ Draw the range, standard deviation and inter-quartile range on the Data Set 3 chart on your handout, using:

- Lower quartile = 0
- Upper quartile = 8.75
- Mean = 3.1
- Standard deviation = 5.1

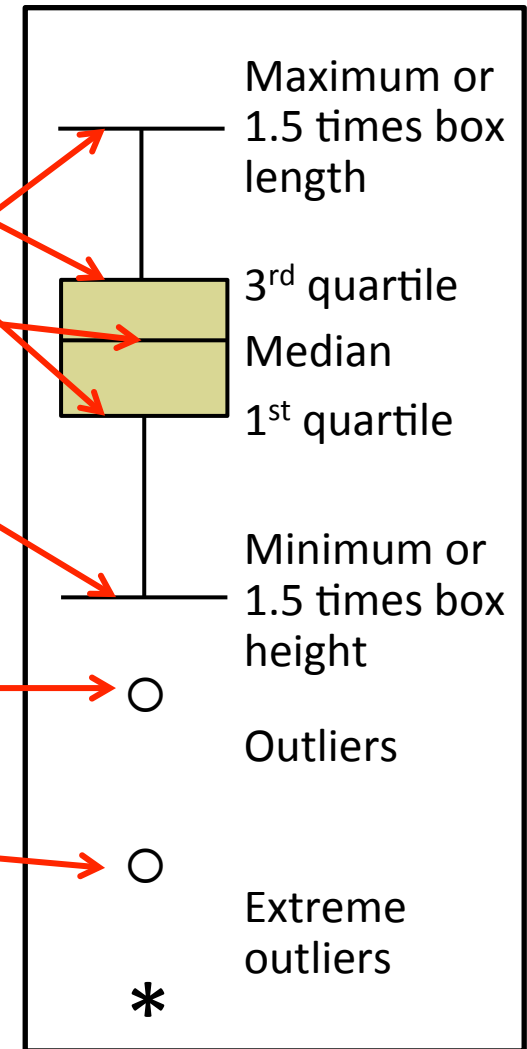
❑ Discuss in small groups which measure of spread you think is the most useful

# Solution

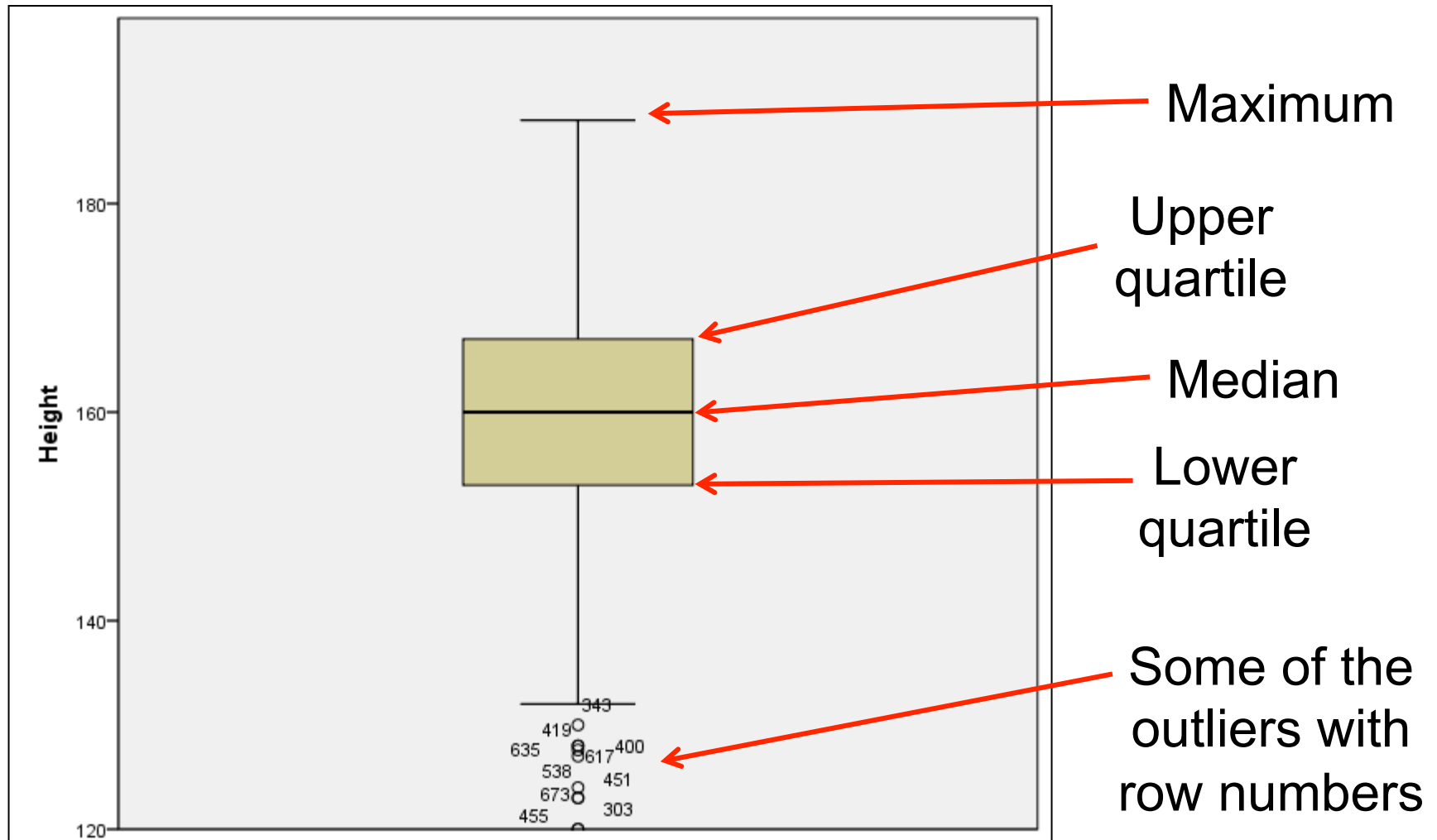


# Box and whisker diagrams (boxplots)

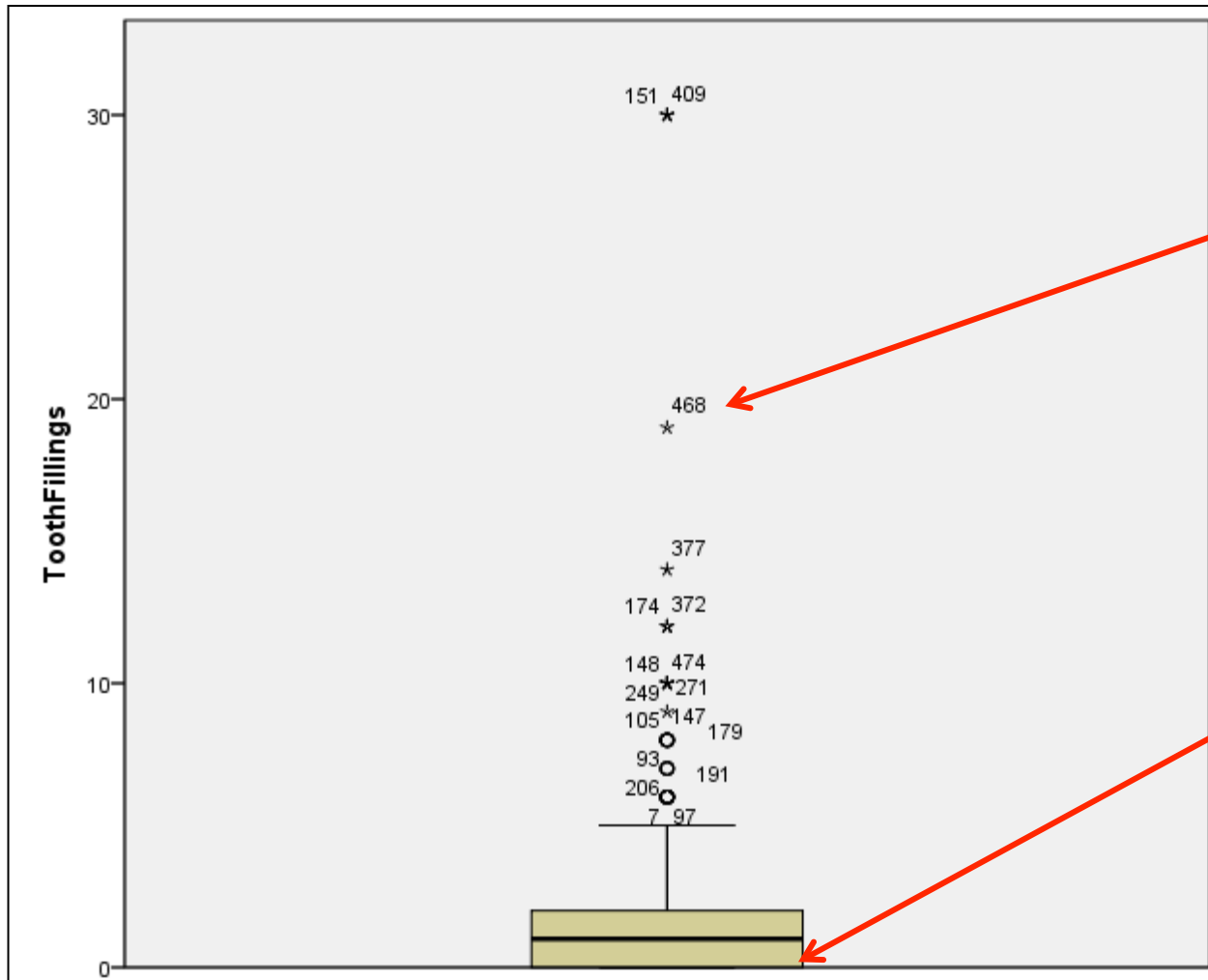
- ❑ Give a good 'picture' of scale data
- ❑ Box ends are the lower and upper quartiles
- ❑ Box middle line is the median
- ❑ Whiskers are the minimum/maximum value of the data or 1.5 times the box height from the end of the box, whichever is nearer the median
- ❑ Circles represent outliers (up to 3 times the box height from the end of the box)
- ❑ Asterisks represent extreme outliers (>3 times box height from the end of the box)



# Boxplot of Height



# Boxplot of Tooth Fillings



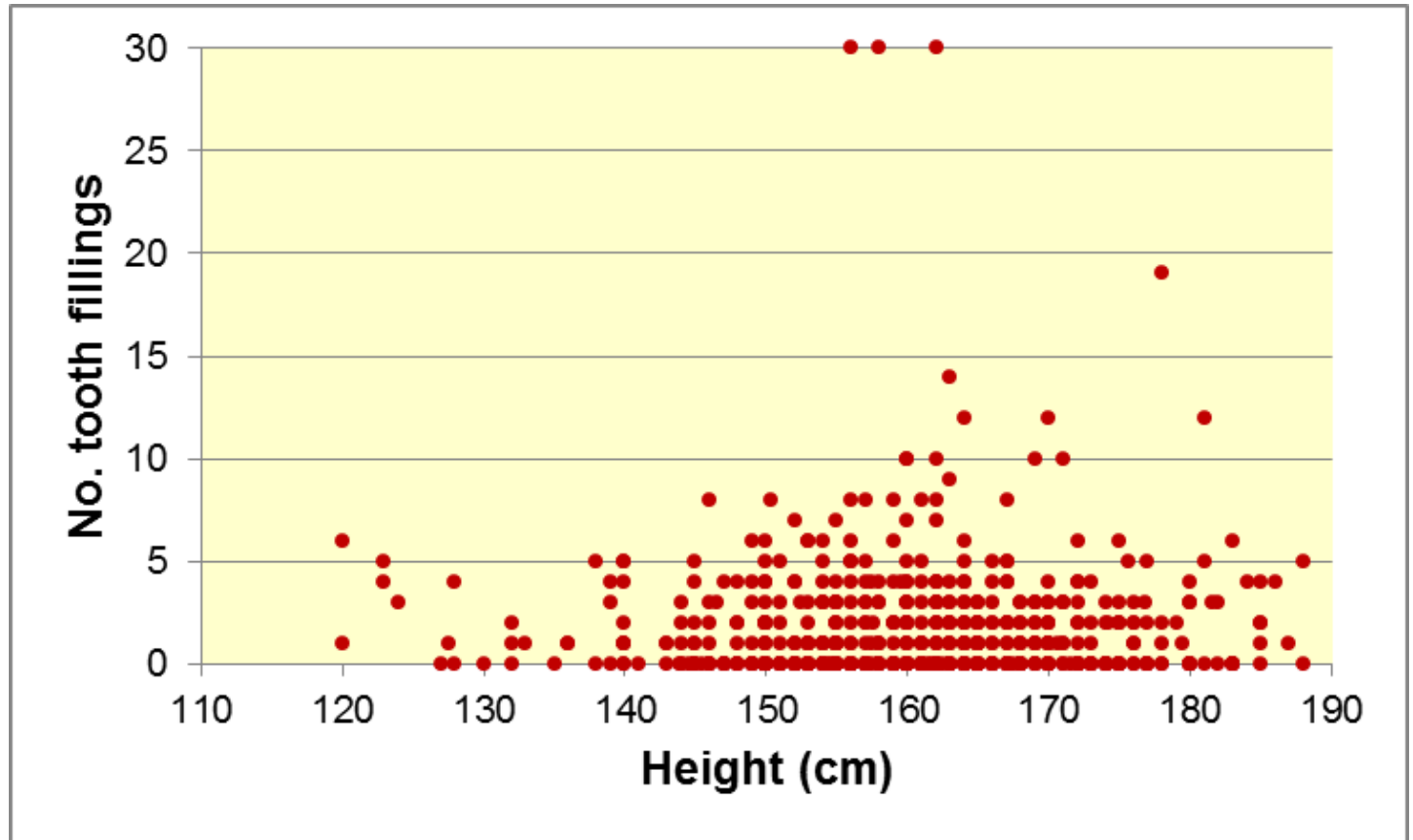
Extreme outliers with row numbers

Minimum and lower quartile have same value

# Scatter plots

- ❑ A two dimensional display of data values for one scale variable against another

Figure shows number of Tooth Fillings against Height



# Using charts and statistics for scale based data

- ❑ Bar charts with interval categories – good for summarising a scale
- ❑ Mean and standard deviation – better when data is spread around the middle
- ❑ Median and inter-quartile range – better when data spread is unusual
- ❑ Boxplots – good for summarising data
  - Can be used to compare several data groups
  - Need at least 15 data points per group
- ❑ Scatter plots – show relationships between data series



# Recap

We have considered:

- ☐ Types of data
- ☐ Tables
- ☐ Charts
- ☐ Measures of middle value
- ☐ Measures of spread
- ☐ Choices of tables, charts and measure for different data types