Chapter 2 - Statistics

Measures of Location and Spread

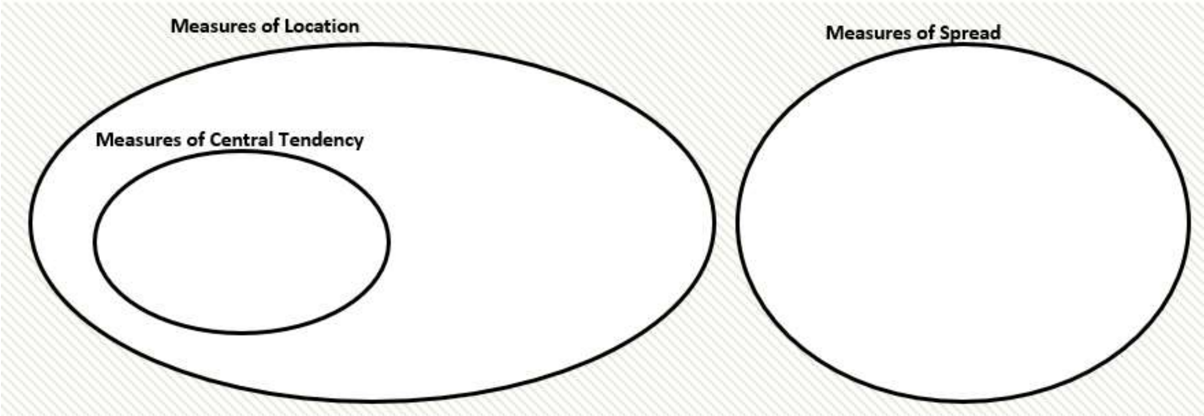
Chapter Overview

1. Measures of Central Tendency
2. Other measures of location
3. Measures of Spread
4. Variance and Standard Deviation
5. Coding

	2.3	Interpret measures of central tendency and variation, extending to standard deviation.	Data may be discrete, continuous, grouped or ungrouped. Understanding and use of coding. Measures of central tendency: mean, median, mode. Measures of variation: variance, standard deviation, range and interpercentile ranges. Use of linear interpolation to calculate percentiles from grouped data is expected.
		Be able to calculate standard deviation, including from summary statistics.	Students should be able to use the statistic x $S_{xx} = \sum(x-\overline{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n}$ Use of standard deviation = $\sqrt{\frac{S_{xx}}{n}}$ (or equivalent) is expected but the use of $S = \sqrt{\frac{S_{xx}}{n-1}}$ (as used on spreadsheets) will be accepted.

1. Measure of Central Tendency

Measures of…



Finding the mean
Using your calculator

On a Classwiz:

- Select 1-Variable.
- Enter each value above, pressing = after each entry.
- Press AC to start a statistical calculation.
- Press the OPTN button. "1-Variable Calc" will calculate all common statistics (including all on the left). Alternatively, you can construct a statistical expression yourself – in the OPTN menu press Down. "Variable" for example contains $\bar{x}$. This will insert it into your calculation; press = when done.

Diameter of coin x (cm)	2.2	2.5	2.6	2.65	2.9

Grouped Data

Height h of bear (in metres)	Frequency
$0 \leq h < 0.5$	4
$0.5 \leq h < 1.2$	20
$1.2 \leq h < 1.5$	5
$1.5 \leq h < 2.5$	11

Mini-Exercise

1.

Num children (c)	Frequency (f)
0	2
1	6
2	1
3	1

2.

IQ of L6Ms2 (q)	Frequency (f)
$80 < q \leq 90$	7
$90 \leq q < 100$	5
$100 \leq q < 120$	3
$120 \leq q < 200$	1

3.

Time t	Frequency (f)
$9.5 < t \leq 10$	32
$10 \leq t < 12$	27
$12 \leq t < 15$	47
$15 \leq t < 16$	11

Exercise 2A/2B Pages 22-23, 24-25

Combined Mean

Example

The mean maths score of 20 pupils in class A is 62.

The mean maths score of 30 pupils in class B is 75.

a) What is the overall mean of all the pupils' marks.

b) The teacher realises they mismarked one student's paper; he should have received 100 instead of 95. Explain the effect on the mean and median.

Question

Archie the Archer competes in a competition with 50 rounds. He scored an average of 35 points in the first 10 rounds and an average of 25 in the remaining rounds. What was his average score per round?

Finding the Median

You need to be able to find the median of both listed data and of grouped data.

Listed data

Items	n	Position of median	Median
1,4,7,9,10	5		
4,9,10,15	4		
2,4,5,7,8,9,11	7		
1,2,3,5,6,9,9,10,11,12	10		

Can you think of a rule to find the position of the median given n?

Grouped data

IQ of L6Ms2 (q)	Frequency (f)
$80 \leq q < 90$	7
$90 \leq q < 100$	5
$100 \leq q < 120$	3
$120 \leq q < 200$	2

Position to use for median:

Linear Interpolation

Height of tree (m)	Freq	C.F.
$0.55 \leq h < 0.6$	55	55
$0.6 \leq h < 0.65$	45	100
$0.65 \leq h < 0.7$	30	130
$0.7 \leq h < 0.75$	15	145
$0.75 \leq h < 0.8$	5	150

Formula

Examples

Weight of cat (kg)	Freq	C.F.
$1.5 \leq w < 3$	10	10
$3 \leq w < 4$	8	18
$4 \leq w < 6$	14	32

Time (s)	Freq	C.F.
$8 \leq t < 10$	4	4
$10 \leq t < 12$	3	7
$12 \leq t < 14$	13	20

Class width

Weight of cat to nearest kg	Frequency
10 – 12	7
13 – 15	2
16 – 18	9
19 – 20	4

Linear Interpolation with gaps

Example

Summarised below are the distances, to the nearest mile, travelled to work by a random sample of 120 commuters.

Distance (to the nearest mile)	Number of commuters
0 – 9	10
10 – 19	19
20 – 29	43
30 – 39	25
40 – 49	8
50 – 59	6
60 – 69	5
70 – 79	3
80 – 89	1

For this distribution,

~~(*a*) describe its shape, **(1)**~~

(*b*) use linear interpolation to estimate its median. **(2)**

Test Your Understanding

Use linear interpolation to estimate the median of the following:

1)

Age of relic (years)	**Frequency**
0-1000	24
1001-1500	29
1501-1700	12
1701-2000	35

2)

Shark length (cm)	**Frequency**
$40 \leq x < 100$	17
$100 \leq x < 300$	5
$300 \leq x < 600$	8
$600 \leq x < 1000$	10

Supplementary Exercise 1

Q1) Solomon Paper A Q5b

The number of patients attending a hospital trauma clinic each day was recorded over several months, giving the data in the table below.

Number of patients	10 - 19	20 - 29	30 - 34	35 - 39	40 - 44	45 - 49	50 - 69
Frequency	2	18	24	30	27	14	5

Use linear interpolation to estimate the median of these data.

Q2) Solomon Paper E Q4a

The ages of 300 houses in a village are recorded given the following table of results.

Age a (years)	Number of houses
$0 \leq a < 20$	36
$20 \leq a < 40$	92
$40 \leq a < 60$	74
$60 \leq a < 100$	39
$100 \leq a < 200$	14
$200 \leq a < 300$	27
$300 \leq a < 500$	18

Use linear interpolation to estimate the median.

Q3) Solomon Paper L Q7a

A cyber-café recorded how long each user stayed during one day giving the following results.

Length of stay (minutes)	Number of houses
$0 \leq l < 30$	15
$30 \leq l < 60$	31
$60 \leq l < 90$	32
$90 \leq l < 120$	23
$120 \leq l < 240$	17
$240 \leq l < 360$	2

Use linear interpolation to estimate the median of these data.

Q4) S1 May 2013 Q4

The following table summarises the times, *t* minutes to the nearest minute, recorded for a group of students to complete an exam.

Time (minutes) t	11 – 20	21 – 25	26 – 30	31 – 35	36 – 45	46 – 60
Number of students f	62	88	16	13	11	10

[You may use $\sum ft^2 = 134281.25$]

(*a*) Estimate the mean ~~and standard deviation~~ of these data. **(5)**

(*b*) Use linear interpolation to estimate the value of the median. **(2)**

Exercise 2C Pages 27-28

2. Other measures of location

Quartiles

Listed Data

Items	n	Position of LQ & UQ	LQ & UQ
1,4,7,9,10	5		
4,9,10,15	4		
2,4,5,7,8,9,11	7		
1,2,3,5,6,9,9,10,11,12	10		

Quartiles – Listed Data

Grouped Data

Items	n	Position of LQ & UQ	LQ & UQ
1,4,7,9,10	5		
4,9,10,15	4		
2,4,5,7,8,9,11	7		
1,2,3,5,6,9,9,10,11,12	10		

Quartiles – Grouped Data

Percentiles

Notation

Lower Quartile:

Median:

Upper Quartile:

57th Percentile:

3. Measures of Spread

The interquartile range and interpercentile range are examples of **measures of spread.**

Interquartile Range = Upper Quartile − Lower Quartile

Why might we favour the interquartile range over the range?

Test your understanding

Age of relic (years)	Frequency
0-1000	24
1001-1500	29
1501-1700	12
1701-2000	35

Shark length (cm)	Frequency
$40 \leq x < 100$	17
$100 \leq x < 300$	5
$300 \leq x < 600$	8
$600 \leq x < 1000$	11

Q1) S1 May 2013 Q4 (continued)

The following table summarises the times, *t* minutes to the nearest minute, recorded for a group of students to complete an exam.

Time (minutes) t	11 – 20	21 – 25	26 – 30	31 – 35	36 – 45	46 – 60
Number of students f	62	88	16	13	11	10

(*c*) Show that the estimated value of the lower quartile is 18.6 to 3 significant figures.

(1)

(*d*) Estimate the interquartile range of this distribution.

(2)

Q2) S1 June 2005 Q2

The following table summarises the distances, to the nearest km, that 134 examiners travelled to attend a meeting in London.

Distance (km)	Number of examiners
41–45	4
46–50	19
51–60	53
61–70	37
71–90	15
91–150	6

(*c*) Use interpolation to estimate the median Q_2, the lower quartile Q_1, and the upper quartile Q_3 of these data.

Q3) The ages of 300 houses in a village are recorded given the following table of results.

Age a (years)	Number of houses
$0 \leq a < 20$	36
$20 \leq a < 40$	92
$40 \leq a < 60$	74
$60 \leq a < 100$	39
$100 \leq a < 200$	14
$200 \leq a < 300$	27
$300 \leq a < 500$	18

Use linear interpolation to estimate the lower quartile, upper quartile and hence the interquartile range.

Q4)

A cyber-café recorded how long each user stayed during one day giving the following results.

Length of stay (minutes)	Number of houses
$0 \leq l < 30$	15
$30 \leq l < 60$	31
$60 \leq l < 90$	32
$90 \leq l < 120$	23
$120 \leq l < 240$	17
$240 \leq l < 360$	2

Use linear interpolation to estimate:

a) The lower quartile.

b) The upper quartile.

c) The 90th percentile.

Q5)

Distance (to the nearest mile)	Number of commuters
0 – 9	10
10 – 19	19
20 – 29	43
30 – 39	25
40 – 49	8
50 – 59	6
60 – 69	5
70 – 79	3
80 – 89	1

Find the interquartile range for the distance travelled by commuters.

4. Variance and Standard Deviation

Variance

Examples

1. 3, 11 Variance

Standard Deviation

2. 2, 3, 3, 5, 7 Variance

Standard Deviation

3. 2, 4, 6 Variance

Standard Deviation

4. 1, 2, 3, 4, 5 Variance

Standard Deviation

Variance – frequency tables

Examples

4. The following table summarises the times, t minutes to the nearest minute, recorded for a group of students to complete an exam.

Time (minutes) t	11 – 20	21 – 25	26 – 30	31 – 35	36 – 45	46 – 60
Number of students f	62	88	16	13	11	10

[You may use $\sum \text{f}t^2 = 134281.25$]

(*a*) Estimate the mean and standard deviation of these data. **(5)**

An agriculturalist is studying the yields, y kg, from tomato plants. The data from a random sample of 70 tomato plants are summarised below.

Yield (y kg)	**Frequency** (f)	**Yield midpoint** (x kg)
$0 \le y < 5$	16	2.5
$5 \le y < 10$	24	7.5
$10 \le y < 15$	14	12.5
$15 \le y < 25$	12	20
$25 \le y < 35$	4	30

(You may use $\sum \text{f}x = 755$ and $\sum \text{f}x^2 = 12\,037.5$)

(*c*) Estimate the mean and the standard deviation of the yields of the tomato plants. **(4)**

5. Coding

Rules of coding

Suppose our original variable (e.g. heights in cm) was x. Then y would represent the heights with 10cm added on to each value.

Coding	Effect on $\bar{x}$	Effect on σ
$y = x + 10$		
$y = 3x$		
$y = 2x - 5$		

You might get any **linear** coding (i.e. using $\times + \div -$). We might think that any operation on the values has the same effect on the mean. But note for example that **squaring** the values would not square the mean; we already know that $\Sigma x^2 \neq (\Sigma x)^2$ in general.

Quick-fire Questions

Old mean $\bar{x}$	Old σ_x	Coding	New mean $\bar{y}$	New σ_y
36	4	$y = x - 20$		
		$y = 2x$	72	16
35	4	$y = 3x - 20$		
		$y = \frac{x}{2}$	20	$\frac{3}{2}$
11	27	$y = \frac{x + 10}{3}$		
		$y = \frac{x - 100}{5}$	40	5

Example Exam Question

The following table summarises the times, t minutes to the nearest minute, recorded for a group of students to complete an exam.

Time (minutes) t	11 – 20	21 – 25	26 – 30	31 – 35	36 – 45	46 – 60
Number of students f	62	88	16	13	11	10

[You may use $\sum ft^2 = 134281.25$]

(*a*) Estimate the mean and standard deviation of these data. **(5)**
(*b*) Use linear interpolation to estimate the value of the median. **(2)**
(*c*) Show that the estimated value of the lower quartile is 18.6 to 3 significant figures. **(1)**
(*d*) Estimate the interquartile range of this distribution. **(2)**
(*e*) Give a reason why the mean and standard deviation are not the most appropriate summary statistics to use with these data. **(1)**

The person timing the exam made an error and each student actually took 5 minutes less than the times recorded above. The table below summarises the actual times.

Time (minutes) t	6 – 15	16 – 20	21 – 25	26 – 30	31 – 40	41 – 55
Number of students f	62	88	16	13	11	10

(*f*) Without further calculations, explain the effect this would have on each of the estimates found in parts (*a*), (*b*), (*c*) and (*d*). **(3)**