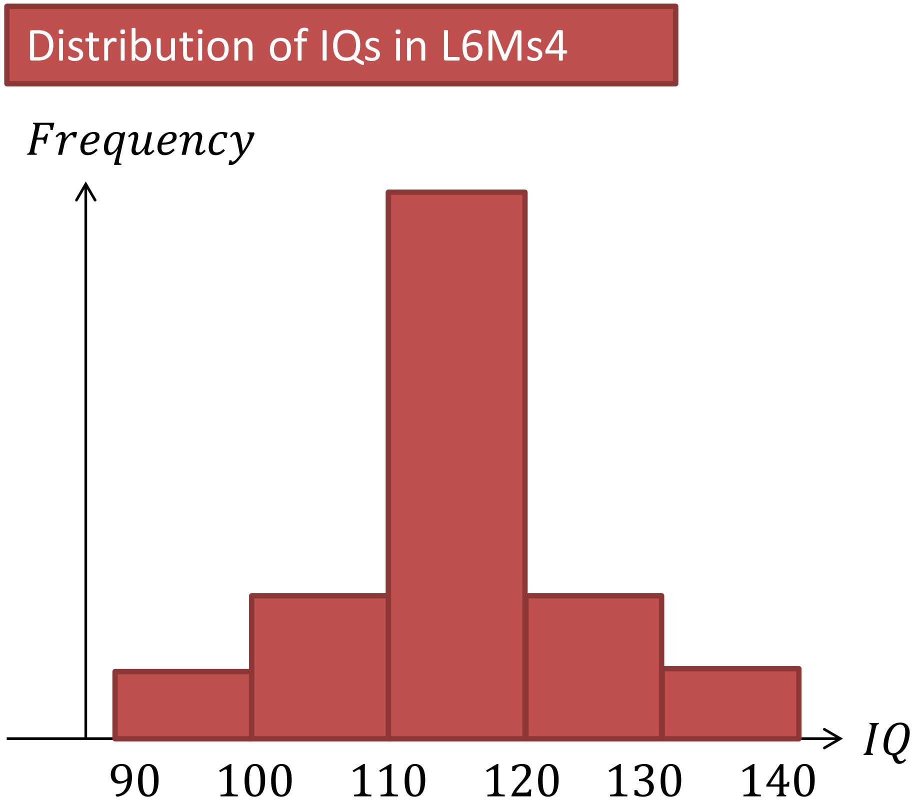
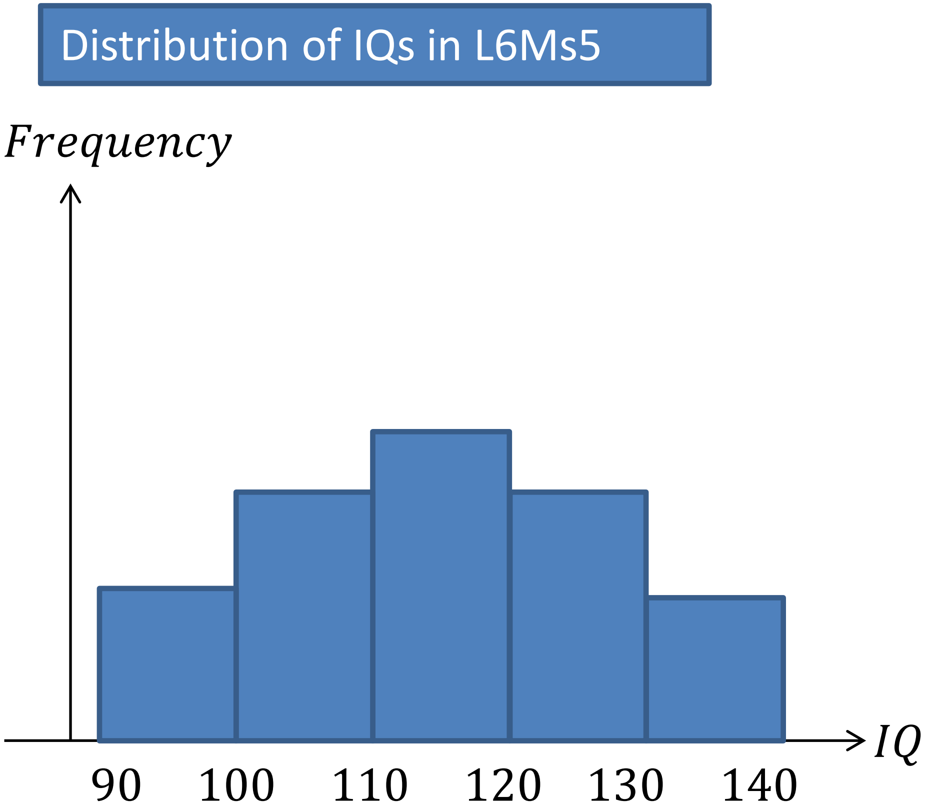
Stats1 Chapter 2: Measures of Data

Variance and Standard Deviation

What is variance?





Here are the distribution of IQs in two classes of the same size. What's the same, and what's different?

The (estimated) mean IQ is the same for the two classes, as is the (estimated) range, but the overall spread of values is greater for the second class. The interquartile range would convey this, but do we have a method of measuring spread that takes into account all the values?

Variance

Variance is a measure of spread that takes all values into account.
Variance, by definition, is the **average squared distance from the mean**.

$$\sigma^2 = \frac{\Sigma\left(x - \bar{x}\right)^2}{n}$$

Notation: While Σ is 'uppercase sigma' and means 'sum of', σ is 'lowercase sigma' (we'll see why we have the squared in a sec)

Distance from mean…

Squared distance from mean…

Average squared distance from mean…

Simpler formula for variance

But in practice you will never use this form, and it's possible to simplify the formula to the following*:

Variance

"The mean of the squares minus the square of the mean ('msmsm')"

$$\sigma^2 = \frac{\Sigma x^2}{n} - \bar{x}^2$$

*** Proof: (certainly not in syllabus!)**
Note that $\bar{x}$ is constant for a fixed variable, and that in general, $\Sigma k f(x) = k \Sigma f(x)$ for a constant k, i.e. we can factor out constants out of a summation.

$$\sigma^2 = \frac{\Sigma(x-\bar{x})^2}{n} = \frac{\Sigma(x^2 - 2x\bar{x} + \bar{x}^2)}{n}$$
$$= \frac{\Sigma x^2}{n} - \frac{\Sigma(2x\bar{x})}{n} + \frac{\Sigma \bar{x}^2}{n}$$
$$= \frac{\Sigma x^2}{n} - 2\bar{x}\left(\frac{\Sigma x}{n}\right) + \frac{\bar{x}^2}{n}\Sigma 1$$
$$= \frac{\Sigma x^2}{n} - 2\bar{x}^2 + \frac{\bar{x}^2}{n} \cdot n$$
$$= \frac{\Sigma x^2}{n} - 2\bar{x}^2 + \bar{x}^2$$
$$= \frac{\Sigma x^2}{n} - \bar{x}^2$$

Standard Deviation

$$\sigma = \sqrt{Variance}$$

The standard deviation **can 'roughly' be thought of as the <u>average distance from the mean</u>.**

Examples

3, 11

Variance

$$\sigma^2 = 65 - 7^2 = 16$$

Standard Deviation

$$\sigma = \sqrt{16} = 4$$

So note that that in the case of two items, the standard deviation **<u>is</u>** indeed the average distance of the values from the mean.

2cm 3cm 3cm 5cm 7cm

Variance

$$\sigma^2 = 19.2 - 4^2 = 3.2\text{cm}$$

Standard Deviation

$$\sigma = \sqrt{3.2} = 1.79\text{cm}$$

Practice

Find the variance and standard deviation of the following sets of data.

$$2 \quad 4 \quad 6$$

Variance = ?

Standard Deviation = ?

$$1 \quad 2 \quad 3 \quad 4 \quad 5$$

Variance = ?

Standard Deviation = ?

Practice

Find the variance and standard deviation of the following sets of data.

2 4 6

Variance = 2.67

Standard Deviation = 1.63

1 2 3 4 5

Variance = 2

Standard Deviation = 1.41

Extending to frequency/grouped frequency tables

We can just mull over our mnemonic again:

Variance: "The mean of the squares minus the square of the mean ('msmsm')"

$$Variance = \frac{\Sigma f x^2}{\Sigma f} - \bar{x}^2$$

Fro Tip: It's better to try and memorise the mnemonic than the formula itself – you'll understand what's going on better.

Fro Exam Note: In an exam, you will pretty much certainly be asked to find the standard deviation for grouped data, and not listed data.

Example

May 2013 Q4

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 $\sum ft^2 = 134281.25$]

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

We can use our STATS mode to work out the various summations needed (and "1-Variable Calc" will contain this amongst its list). Just input the table as normal. Note that, as per the discussion before, Σx^2 on a calculator actually gives you $\Sigma f x^2$ because it's **already taking the frequencies into account**.

$\bar{x} =$?

$\sigma^2 =$?

$\sigma =$?

You ABSOLUTELY must check this with the σx you can get on the calculator directly.

Example

May 2013 Q4

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 $\sum ft^2 = 134281.25$]

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

We can use our STATS mode to work out the various summations needed (and "1-Variable Calc" will contain this amongst its list). Just input the table as normal. Note that, as per the discussion before, Σx^2 on a calculator actually gives you $\Sigma f x^2$ because it's **already taking the frequencies into account**.

$$\bar{x} = \frac{4837.5}{200} = 24.1875$$

$$\sigma^2 = \frac{134281.25}{200} - 24.1875^2 = 86.37$$

$$\sigma = \sqrt{86.37} = 9.29$$

You ABSOLUTELY must check this with the σx you can get on the calculator directly.

Test Your Understanding

May 2013 (R) Q3

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)
$0 \leq y < 5$	16	2.5
$5 \leq y < 10$	24	7.5
$10 \leq y < 15$	14	12.5
$15 \leq y < 25$	12	20
$25 \leq y < 35$	4	30

(You may use $\sum fx = 755$ and $\sum fx^2 = 12\ 037.5$)

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

(c)

?

Test Your Understanding

May 2013 (R) Q3

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 $\sum fx = 755$ and $\sum fx^2 = 12\,037.5$)

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

(c)	$[\bar{x} =]\ \frac{755}{70}$ or **awrt 10.8**	B1	
	$[\sigma_x =]\ \sqrt{\frac{12037.5}{70} - \bar{x}^2} = \sqrt{55.6326...}$	M1A1ft	
	= **awrt 7.46** (Accept s = awrt 7.51)	A1	(4)

Most common exam errors

- Thinking $\Sigma f x^2$ means $(\Sigma f x)^2$. It means the sum of each value squared!
- When asked to calculate the mean followed by standard deviation, using a rounded version of the mean in calculating the standard deviation, and hence introducing rounding errors.
- Forgetting to square root the variance to get the standard deviation.

<u>ALL these mistakes can be easily spotted</u> if you check your value against "σx" in STATS mode.

Exercise 2.4

Pearson Statistics & Mechanics Year 1/AS
Pages 12-13

Homework Exercise

1 Given that for a variable x: $\Sigma x = 24$ $\Sigma x^2 = 78$ $n = 8$

Find:

a the mean **b** the variance σ^2 **c** the standard deviation σ.

2 Ten collie dogs are weighed (w kg). The summary data for the weights is:

$\Sigma w = 241$ $\Sigma w^2 = 5905$

Use this summary data to find the standard deviation of the collies' weights. **(2 marks)**

3 Eight students' heights (h cm) are measured. They are as follows:

165 170 190 180 175 185 176 184

a Work out the mean height of the students.

b Given $\Sigma h^2 = 254\,307$ work out the variance. Show all your working.

c Work out the standard deviation.

4 For a set of 10 numbers: $\Sigma x = 50$ $\Sigma x^2 = 310$

For a different set of 15 numbers: $\Sigma x = 86$ $\Sigma x^2 = 568$

Find the mean and the standard deviation of the combined set of 25 numbers.

5 Nahab asks the students in his year group how much pocket money they get per week. The results, rounded to the nearest pound, are shown in the table.

Number of £s	8	9	10	11	12
Frequency	14	8	28	15	20

a Use your calculator to work out the mean and standard deviation of the pocket money. Give units with your answer. **(3 marks)**

b How many students received an amount of pocket money more than one standard deviation above the mean? **(2 marks)**

Homework Exercise

6 In a student group, a record was kept of the number of days of absence each student had over one particular term. The results are shown in the table.

Number of days absent	0	1	2	3	4
Frequency	12	20	10	7	5

Use your calculator to work out the standard deviation of the number of days absent. **(2 marks)**

7 A certain type of machine contained a part that tended to wear out after different amounts of time. The time it took for 50 of the parts to wear out was recorded. The results are shown in the table.

Lifetime, *h* (hours)	$5 < h \leqslant 10$	$10 < h \leqslant 15$	$15 < h \leqslant 20$	$20 < h \leqslant 25$	$25 < h \leqslant 30$
Frequency	5	14	23	6	2

The manufacturer makes the following claim:

> 90% of the parts tested lasted longer than one standard deviation below the mean.

Comment on the accuracy of the manufacturer's claim, giving relevant numerical evidence.

Problem-solving

You need to calculate estimates for the mean and the standard deviation, then estimate the number of parts that lasted longer than one standard deviation below the mean.

(5 marks)

8 The daily mean windspeed, x (kn) for Leeming is recorded in June 2015. The summary data is:

$\Sigma x = 243 \qquad \Sigma x^2 = 2317$

a Use your calculator to work out the mean and the standard deviation of the daily mean windspeed in June 2015. **(2 marks)**

The highest recorded windspeed was 17 kn and the lowest recorded windspeed was 4 kn.

b Estimate the number of days in which the windspeed was greater than one standard deviation above the mean. **(2 marks)**

c State one assumption you have made in producing this estimate. **(1 mark)**

Homework Answers

1 **a** 3 **b** 0.75 **c** 0.866

2 3.11 kg

3 **a** 178 cm **b** 59.9 cm^2 **c** 7.74 cm

4 Mean 5.44, standard deviation 2.35

5 **a** Mean £10.22, standard deviation £1.35

b 18

6 1.23 days

7 Mean 16.1 hours, standard deviation 4.69 hours
One standard deviation below mean 11.41 hours.
41 parts tested (82%) lasted longer than one standard deviation below the mean. According to the manufacturers, this should be 45 parts (90%), so the claim is false.

8 **a** Mean 8.1 kn, standard deviation 3.41 kn

b 12 days

c The windspeeds are equally distributed throughout the range.