Stats1 Chapter 2: Measures of Data

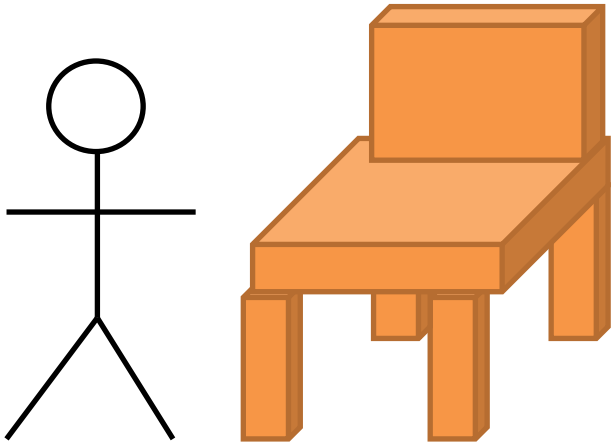
Coding

Coding

What do you reckon is the mean height of people in this room?

Now, stand on your chair, as per the instructions below.

INSTRUCTIONAL VIDEO



Is there an easy way to recalculate the mean based on your new heights? And the variance of your heights?

The mean would increase by the height of the chairs. The spread however is unaffected thus the variance would remain the same.

Starter

Suppose now after a bout of ‘stretching you to your limits’, you’re now all 3 times your original height.

What do you think happens to the **standard deviation** of your heights?

?

What do you think happens to the **variance** of your heights?

?

Extension Question: Can you prove the latter using the formula for variance?

?

Starter

Suppose now after a bout of ‘stretching you to your limits’, you’re now all 3 times your original height.

What do you think happens to the **standard deviation** of your heights?

It becomes 3 times larger (i.e. your heights are 3 times as spread out!)

What do you think happens to the **variance** of your heights?

It becomes 9 times larger. We use the scale factor of the standard deviation, squared.

Extension Question: Can you prove the latter using the formula for variance?

$$\sigma^2 = \frac{\Sigma(3x)^2}{n} - \left(\frac{\Sigma(3x)}{n}\right)^2$$

$$= \frac{\Sigma 9x^2}{n} - \left(\frac{3\Sigma x}{n}\right)^2 = \frac{9 \cdot \Sigma x^2}{n} - 9\left(\frac{\Sigma x}{n}\right)^2$$

$$= 9\left(\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2\right)$$

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.

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
$y = x + 10$	$\bar{x}$ will similarly increase by 10 (to get $\bar{y}$)	As discussed, adding (and subtracting) has no effect on standard deviation or any measure of spread.
$y = 3x$	$\bar{x}$ will get 3 times bigger.	Standard deviation will get 3 times larger.
$y = 2x - 5$	$\bar{y} = 2\bar{x} - 5$, i.e. effect on values is same effect on mean.	-5 has no effect but standard deviation will get 2 times larger.

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.

The point of coding

Cost x of diamond ring (£)

£1010	£1020	£1030	£1040	£1050

We 'code' our variable using the following:

$$y = \frac{x - 1000}{10}$$

New values y:

£1	£2	£3	£4	£5

Standard deviation of y (σ_y): $\sqrt{2}$

therefore…

Standard deviation of x (σ_x): $10\sqrt{2}$

The jist of coding: We want to find the mean/standard deviation of a variable. We transform the values, using some rule, to make them simpler. We can then more easily calculate the mean/standard deviation of the 'coded' data, and from this we can then determine what the mean/standard deviation would have been for the original uncoded data.

Quickfire 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

Quickfire Questions

Old mean $\bar{x}$	Old σ_x	Coding	New mean $\bar{y}$	New σ_y
36	4	$y = x - 20$	16	4
36	8	$y = 2x$	72	16
35	4	$y = 3x - 20$	85	12
40	3	$y = \frac{x}{2}$	20	$\frac{3}{2}$
11	27	$y = \frac{x + 10}{3}$	7	9
300	25	$y = \frac{x - 100}{5}$	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 $\Sigma ft^2 = 134281.25$]

(*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)**

Suppose we've worked all these out already.

f)

Coding is

$$t_2 = t_1 - 5$$

Mean decreases by 5.
σ unaffected.
Median decreases by 5.
Lower quartile decreases by 5.
Interquartile range unaffected.

Chapter 2 Summary

I have a list of 30 heights in the class. What item do I use for:

- Q_1?
- Q_2?
- Q_3?

?

?

?

For the following grouped frequency table, calculate:

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

a) The estimate mean: ?

b) The estimate median: ?

c) The estimate variance:
(you're given $\Sigma f h^2 = 67.8125$) ?

Chapter 2 Summary

I have a list of 30 heights in the class. What item do I use for:

- Q_1? 8^{th}
- Q_2? Between 15^{th} and 16^{th}
- Q_3? 23^{rd}

For the following grouped frequency table, calculate:

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

a) The estimate mean: $\bar{h} = \dfrac{(0.25 \times 4) + (0.85 \times 20) + \cdots}{40} = \dfrac{46.75}{40} = 1.17m \; to \; 3sf$

b) The estimate median: $0.5 + \left(\dfrac{16}{20} \times 0.7\right) = 1.06m$

c) The estimate variance:
(you're given $\Sigma f h^2 = 67.8125$)

$$\sigma^2 = \frac{67.8125}{40} - \left(\frac{46.75}{40}\right)^2 = 0.329 \; to \; 3sf$$

Chapter 2 Summary

What is the standard deviation of the following lengths: 1cm, 2cm, 3cm

?

The mean of a variable x is 11 and the variance 4.
The variable is coded using $y = \frac{x+10}{3}$. What is:

a) The mean of y? ?
b) The variance of y? ?

A variable x is coded using $y = 4x - 5$.
For this new variable y, the mean is 15 and the standard deviation 8.
What is:

a) The mean of the original data? ?
b) The standard deviation of the original data? ?

Chapter 2 Summary

What is the standard deviation of the following lengths: 1cm, 2cm, 3cm

$$\sigma^2 = \frac{14}{3} - 2^2 = \frac{2}{3} \qquad \sigma = \sqrt{\frac{2}{3}}$$

The mean of a variable x is 11 and the variance 4.
The variable is coded using $y = \frac{x+10}{3}$. What is:

a) The mean of y? $\boldsymbol{\bar{y} = 7}$
b) The variance of y? $\boldsymbol{\sigma_y^2 = \frac{4}{9}}$

A variable x is coded using $y = 4x - 5$.
For this new variable y, the mean is 15 and the standard deviation 8.
What is:

a) The mean of the original data? $\boldsymbol{\bar{x} = 5}$
b) The standard deviation of the original data? $\boldsymbol{\sigma_x = 2}$

Exercise 2.5

Pearson Statistics & Mechanics Year 1/AS
Pages 13-14

Homework Exercise

1 A set of data values, x, is shown below:

110 90 50 80 30 70 60

a Code the data using the coding $y = \frac{x}{10}$.

b Calculate the mean of the coded data values.

c Use your answer to part **b** to calculate the mean of the original data.

2 A set of data values, x, is shown below:

52 73 31 73 38 80 17 24

a Code the data using the coding $y = \frac{x - 3}{7}$.

b Calculate the mean of the coded data values.

c Use your answer to part **b** to calculate the mean of the original data.

3 The coded mean price of televisions in a shop was worked out. Using the coding $y = \frac{x - 65}{200}$ the mean price was 1.5. Find the true mean price of the televisions. **(2 marks)**

4 The coding $y = x - 40$ gives a standard deviation for y of 2.34.
Write down the standard deviation of x.

Watch out Adding or subtracting constants does not affect how spread out the data is, so you can ignore the '−40' when finding the standard deviation for x.

Homework Exercise

5 The lifetime, x, in hours, of 70 light bulbs is shown in the table.

Lifetime, x (hours)	$20 < x \leqslant 22$	$22 < x \leqslant 24$	$24 < x \leqslant 26$	$26 < x \leqslant 28$	$28 < x \leqslant 30$
Frequency	3	12	40	10	5

The data is coded using $y = \dfrac{x - 1}{20}$.

a Estimate the mean of the coded values $\bar{y}$.

b Hence find an estimate for the mean lifetime of the light bulbs, $\bar{x}$.

c Estimate the standard deviation of the lifetimes of the bulbs.

Problem-solving

Code the midpoints of each class interval. The midpoint of the $22 < x \leqslant 24$ class interval is 23, so the coded midpoint will be $\dfrac{23 - 1}{20} = 1.1$.

6 The weekly income, i, of 100 women workers was recorded.

The data was coded using $y = \dfrac{i - 90}{100}$ and the following summations were obtained:

$\Sigma y = 131, \ \Sigma y^2 = 176.84$

Estimate the standard deviation of the actual women workers' weekly income. **(2 marks)**

7 A meteorologist collected data on the annual rainfall, x mm, at six randomly selected places.

The data was coded using $s = 0.01x - 10$ and the following summations were obtained:

$\Sigma s = 16.1, \ \Sigma s^2 = 147.03$

Work out an estimate for the standard deviation of the actual annual rainfall. **(2 marks)**

Homework Exercise

8 A teacher standardises the test marks of his class by adding 12 to each one and then reducing the mark by 20%.

If the standardised marks are represented by t and the original marks by m:

a write down a formula for the coding the teacher has used. **(1 mark)**

The following summary statistics are calculated for the standardised marks:

$n = 28 \qquad \bar{t} = 52.8 \qquad S_{tt} = 7.3$

b Calculate the mean and standard deviation of the original marks gained. **(3 marks)**

9 From the large data set, the daily mean pressure, p hPa, in Hurn during June 2015 is recorded.

The data is coded using $c = \frac{p}{2} - 500$ and the following summary statistics are obtained:

$n = 30 \qquad \bar{c} = 10.15 \qquad S_{cc} = 296.4$

Find the mean and standard deviation of the daily mean pressure. **(4 marks)**

Homework Answers

1 **a** 11, 9, 5, 8, 3, 7, 6 **b** 7 **c** 70

2 **a** 7, 10, 4, 10, 5, 11, 2, 3 **b** 6.5 **c** 48.5

3 365

4 2.34

5 **a** 1.2 hours **b** 25.1 hours **c** 1.76 hours

6 22.9

7 416 mm

8 **a** $t = 0.8(m + 12)$ or $t = \frac{m + 12}{1.25}$

b Mean 54, standard deviation 0.64

9 Mean 1020 hPa, standard deviation 6.28 hPa