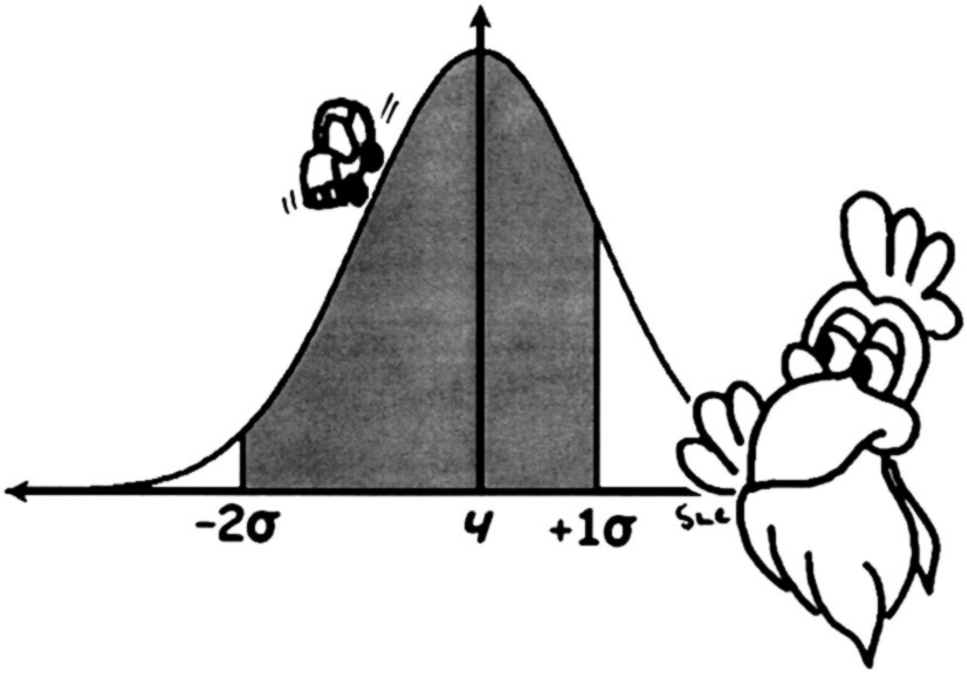
STATISTICS AND STANDARD DEVIATION



Statistics and Standard Deviation

STSD-A Objectives

- To calculate the mean and standard deviation of lists, tables and grouped data
- To determine the correlation co-efficient
- To calculate *z*-scores
- To use normal distributions to determine proportions and values
- To use Chebyshev's theorem
- To determine correlation between sets of data
- To construct scatterplots and lines of best fit
- To calculate correlation coefficient and regression equation for data sets.

STSD-B Calculating Mean

The **mean** is a measure of central tendency. It is the value usually described as the average. The mean is determined by summing all of the numbers and dividing the result by the number of values.

The mean of a population of *N* values (scores) is defined as the sum of all the scores, *x* of the population, $\sum x$, divided by the number of scores, *N*.

The **population mean** is represented by the Greek letter μ (mu) and calculated by using $\boxed{\mu = \frac{\sum x}{N}}$.

Often it is not possible to obtain data from an entire population. In such cases, a sample of the population is taken. The **mean of a sample** of *n* items drawn from the population is defined in the same way and is denoted by $\bar{x}$, pronounced ***x* bar** and calculated using $\boxed{\bar{x} = \frac{\sum x}{n}}$.

Example STSD-B1

Calculate the mean of the following student test results percentages.

92% 66% 99% 75% 69% 51% 89% 75% 54% 45% 69%

$$\bar{x} = \frac{\sum x}{n}$$

- write out formula

$$= \frac{92+66+99+75+69+51+89+75+54+45+69}{11}$$

- add together all scores
- divide by number of scores

$$= \frac{784}{11} = 71.\overline{27}$$

The mean of the student test results is 71.27 % (rounded to 2d.p.).

When calculating the mean from a frequency distribution table, it is necessary to multiply each score by its frequency and sum these values. This result is then divided by the sum of the frequencies.

The formula for the mean calculated from a frequency table is $\boxed{\bar{x} = \frac{\sum fx}{\sum f}}$

Calculations using this formula are often simplified by setting up a table as shown below.

Example STSD-B2

Calculate the mean number of pins knocked down from the frequency table.

Pins (*x*)	Frequency (*f*)	*fx*
0	2	$0 \times 2 = 0$
1	1	$1 \times 1 = 1$
2	2	$2 \times 2 = 4$
3	0	$3 \times 0 = 0$
4	2	$4 \times 2 = 8$
5	4	20
6	9	54
7	11	77
8	13	104
9	8	72
10	8	80
Total	$\sum f = 60$	$\sum fx = 420$

$$\text{mean} = \bar{x} = \frac{\sum fx}{\sum f}$$
$$= \frac{420}{60} = 7$$

The mean number of pins knocked down was 7 pins.

Note: It is rare for an exact number to result from a mean calculation.

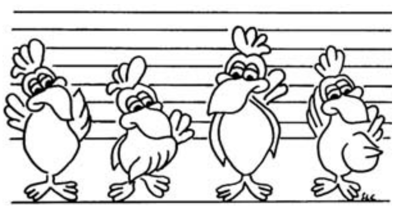
If the frequency distrubution table has grouped data, intervals, it is necessary to use the mid-value of the interval in mean calculations. The mid-value for an interval is calculated by adding the upper and lower boundaries of the interval and dividing the result by two.

$$\boxed{\text{mid value: } x = \frac{upper + lower}{2}}$$

Example STSD-B3

Calculate the mean height of students from the frequency table.

Height (*cm*)	**mid-value (*x*)**	**Frequency(*f*)**	***fx***
140 – 144.9	$\frac{140+145}{2} = 142.5$	1	142.5
145 – 149.9	147.5	1	147.5
150 – 154.9	152.5	2	305
155 – 159.9	157.5	6	945
160 – 164.9	162.5	5	812.5
165 – 169.9	167.5	2	335
170 – 174.9	172.5	1	172.5
175 – 179.9	177.5	2	355
		$\Sigma f = 20$	$\Sigma fx = 3215$



$$\text{mean } \bar{x} = \frac{\sum fx}{\sum f} = \frac{3215}{20} = 160.75cm$$

The mean height is 160.75*cm*.

Exercise STSD-B1

Calculate the mean of the following data sets.

(a) Hockey goals scored.

5, 4, 3, 2, 2, 1, 0, 0, 1, 2, 3

(b) Points scored in basketball games

Points Scored (*x*)	**Frequency *(f)***
10	1
11	0
12	4
13	1
14	3
15	1
Total	10

(c) Number of typing errors

Typing errors	***(f)***
0	6
1	8
2	5
3	1
Total	20

(d) Babies' weights

Baby Weight (kg)	**Freq *(f)***
2.80 – 2.99	2
3.00 – 3.19	1
3.20 – 3.39	3
3.40 – 3.59	2
3.60 – 3.79	5
3.80 – 3.99	2
Total	15

(e) ATM withdrawals

Withdrawals ($)	***(f)***
0 – 49	7
50 – 99	9
100 – 149	5
150 – 199	5
200 – 249	2
250 – 299	2
Total	30

STSD-C Definition of Variance and Standard Deviation

To further describe data sets, measures of spread or dispersion are used. One of the most commonly used measures is **standard deviation**. This value gives information on how the values of the data set are varying, or deviating, from the mean of the data set.

Deviations are calculated by subtracting the mean, $\overline{x}$, from each of the sample values, *x*, i.e. **deviation** $= x - \overline{x}$. As some values are less than the mean, negative deviations will result, and for values greater than the mean positive deviations will be obtained. By simply adding the values of the deviations from the mean, the positive and negative values will cancel to result in a value of zero. By squaring each of the deviations, the problem of positive and negative values is avoided.

To calculate the standard deviation, the deviations are squared. These values are summed, divided by the appropriate number of values and then finally the square root is taken of this result, to counteract the initial squaring of the deviation.

The **standard deviation of a population**, σ , of *N* data items is defined by the following formula.

$$\boxed{\sigma = \sqrt{\frac{\Sigma\left(x-\mu\right)^2}{N}}} \quad \text{where } \mu \text{ is the population mean.}$$

For a **sample** of *n* data items the **standard deviation**, *s*, is defined by,

$$\boxed{s = \sqrt{\frac{\Sigma\left(x-\overline{x}\right)^2}{n-1}}} \quad \text{where } \overline{x} \text{ is the sample mean.}$$

NOTE: When calculating the sample standard deviation we divide by (*n* – 1) not *N*. The reason for this is complex but it does give a more accurate measurement for the variance of a sample.

Standard deviation is measured in the same units as the mean.

It is usual to assume that data is from a sample, unless it is stated that a population is being used.

To assist in calculations data should be set up in a table and the following headings used:

x	$x-\mu$ ***OR*** $x-\overline{x}$	$\left(x-\mu\right)^2$ ***OR*** $\left(x-\overline{x}\right)^2$

Example STSD-C1

Determine the standard deviation of the following student test results percentages.

92% 66% 99% 75% 69% 51% 89% 75% 54% 45% 69%

x	$x-\overline{x}$	$\left(x-\overline{x}\right)^2$
92	$92-71.3=20.7$	$\left(20.7\right)^2=428.49$
66	–5.3	28.09
99	27.7	767.29
75	3.7	13.69
69	–2.3	5.29
51	–20.3	412.09
89	17.7	313.29
75	3.7	13.69
54	–17.3	299.29
45	–26.3	691.69
69	–2.3	5.29
$\Sigma x = 784$		$\Sigma\left(x-\overline{x}\right)^2 = 2978.19$

$$\overline{x} = \frac{\Sigma x}{n} = \frac{784}{11} \approx 71.3$$

$$s = \sqrt{\frac{\Sigma\left(x-\overline{x}\right)^2}{n-1}}$$
$$= \sqrt{\frac{2978.19}{11-1}}$$
$$\approx 17.26$$

The standard deviation of the test results is approximately 17.26%.

The **variance** is the average of the squared deviations when the data given represents the population. The lower case Greek letter sigma squared, σ^2, is used to represent the **population variance**.

$$\boxed{\sigma^2 = \frac{\Sigma\left(x-\mu\right)^2}{N}}$$

where μ is the population mean, and N is the population size.

The **sample variance**, which is denoted by s^2, is defined as

$$\boxed{s^2 = \frac{\Sigma\left(x-\overline{x}\right)^2}{n-1}}$$

where $\overline{x}$ is the sample mean, and n is the sample size.

As variance is measured in squared units, it is more useful to use standard deviation, the square root of variance, as a measure of dispersion.

STSD-D Calculating Standard Deviation

The previously mentioned formulae for standard deviation of a population, σ and a sample standard deviation, *s*,

$$\boxed{\sigma = \sqrt{\frac{\Sigma\left(x-\mu\right)^2}{N}} \qquad s = \sqrt{\frac{\Sigma\left(x-\overline{x}\right)^2}{n-1}}}$$

can be manipulated to obtain the following formula which are easier to use for calculations. These are commonly called computational formulae.

$$\boxed{\sigma = \sqrt{\frac{\Sigma x^2 - \frac{\left(\Sigma x\right)^2}{N}}{N}} \qquad s = \sqrt{\frac{\Sigma x^2 - \frac{\left(\Sigma x\right)^2}{n}}{n-1}}}$$

To perform calculations again it is necessary to set up a table. The table heading in this case will be:

x	***x*²**

Example STSD-D1

Determine the standard deviation of the following student test results percentages.

92% 66% 99% 75% 69% 51% 89% 75% 54% 45% 69%

x	***x*²**
92	$92^2 = 8464$
66	4356
99	9801
75	5625
69	4761
51	2601
89	7921
75	5625
54	2916
45	2025
69	4761
$\Sigma x = 784$	$\Sigma x^2 = 58856$

$$\begin{aligned} s &= \sqrt{\frac{\Sigma x^2 - \frac{\left(\Sigma x\right)^2}{n}}{n-1}} \\ &= \sqrt{\frac{58856 - \frac{784^2}{11}}{11-1}} \\ &= \sqrt{\frac{58856 - 55877.\overline{81}}{10}} \\ &\approx 17.26 \end{aligned}$$

NOTE: This is approximately the same value as calculated previously. This value will actually be more accurate as it only uses rounding in the final calculation step.

The standard deviation of the test scores is approximately 17.26%.

When data is presented in a frequency table the following computational formulae for populations standard deviation, σ , and sample standard deviation, *s*, can be used.

$$\sigma = \sqrt{\frac{\Sigma fx^2 - \frac{(\Sigma fx)^2}{\Sigma f}}{\Sigma f}} \qquad s = \sqrt{\frac{\Sigma fx^2 - \frac{(\Sigma fx)^2}{\Sigma f}}{\Sigma f - 1}}$$

If the data is presented in a grouped or interval manner, the mid-values are used as with the calculation of the mean.

The table heading for calculations will include.

x	***f***	***fx***	***x²***	***fx²***

Examples STSD-D2

Calculate the standard deviations for each of the following data sets.

(a) Number of pins knocked down in ten-pin bowling matches

Pins (x)	***f***	***fx***	***x²***	***fx²***
0	2	0	0	0
1	1	1	1	1
2	2	4	4	8
3	0	0	9	0
4	2	8	16	32
5	4	20	25	100
6	9	54	36	324
7	11	77	49	539
8	13	104	64	832
9	8	72	81	648
10	8	80	100	800
	$\Sigma f = 60$	$\Sigma fx = 420$		$\Sigma fx^2 = 3284$

$$s = \sqrt{\frac{\Sigma fx^2 - \frac{(\Sigma fx)^2}{\Sigma f}}{\Sigma f - 1}}$$

$$= \sqrt{\frac{3284 - \frac{420^2}{60}}{60 - 1}} \approx 2.41$$

The standard deviation of the number of pins knocked down is approximately 2.41 pins.

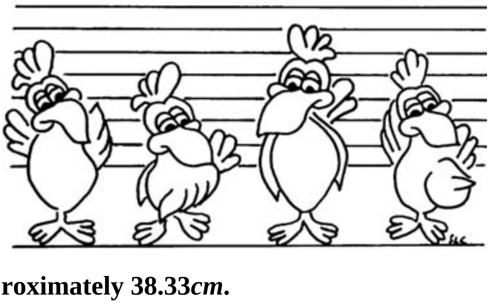
(b) Heights of students

Heights	***x***	***f***	***fx***	***x²***	***fx²***
140 – 144.9	142.5	1	142.5	20306.25	20306.25
145 – 149.9	147.5	1	147.5	21756.25	21756.25
150 – 154.9	152.5	2	305	23256.25	46512.5
155 – 159.9	157.5	6	945	24806.25	148837.5
160 – 164.9	162.5	5	812.5	26406.25	132031.25
165 – 169.9	167.5	2	335	28056.25	56112.5
170 – 174.9	172.5	1	172.5	29756.25	29756.25
175 – 179.9	177.5	2	355	31506.25	63012.5
		$\Sigma f = 20$	$\Sigma fx = 3215$		$\Sigma fx^2 = 544731.25$

$$s = \sqrt{\frac{\Sigma fx^2 - \frac{(\Sigma fx)^2}{\Sigma f}}{\Sigma f - 1}}$$

$$= \sqrt{\frac{544731.25 - \frac{3215^2}{20}}{20 - 1}}$$

$$\approx 38.33$$



The standard deviation of the heights is approximately 38.33*cm*.

Exercise STSD-D1

Calculate the standard deviations for each of the following data sets.

(a) Hockey goals scored.

5, 4, 3, 2, 2, 1, 0, 0, 1, 2, 3

(b) Points scored in basketball games.

Points Scored (x)	Frequency (f)
10	1
11	0
12	4
13	1
14	3
15	1
Total	10

(c) Number of typing errors.

Typing errors	(f)
0	6
1	8
2	5
3	1
Total	20

(d) Babies weights

Baby Weight (kg)	Freq (f)
2.80 – 2.99	2
3.00 – 3.19	1
3.20 – 3.39	3
3.40 – 3.59	2
3.60 – 3.79	5
3.80 – 3.99	2
Total	15

(e) ATM withdrawals

Withdrawals ($)	(f)
0 – 49	7
50 – 99	9
100 – 149	5
150 – 199	5
200 – 249	2
250 – 299	2
Total	30

STSD-E Co-efficient of Variation

Without an understanding of the relative size of the standard deviation compared to the original data, the standard deviation is somewhat meaningless for use with the comparison of data sets. To address this problem the **coefficient of variation** is used.

The coefficient of variation, *CV*, gives the standard deviation as a percentage of the mean of the data set.

$$CV = \frac{s}{\bar{x}} \times 100\% \qquad CV = \frac{\sigma}{\mu} \times 100\%$$

for a sample   for a population

Example STSD-E1

Calculate the coefficient of variation for the following data set.

The price, in cents, of a stock over five trading days was 52, 58, 55, 57, 59.

x	x^2
52	2704
58	3364
55	3025
57	3249
59	3481
$\Sigma x = 281$	$\Sigma x^2 = 15823$

$$\bar{x} = \frac{\sum x}{n} = \frac{281}{5} = 56.1$$

$$s = \sqrt{\frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}} = \sqrt{\frac{15823 - \frac{281^2}{5}}{5-1}} \approx 2.77$$

$$CV = \frac{s}{\bar{x}} \times 100\% = \frac{2.77}{56.2} \times 100\% \approx 4.93\%$$

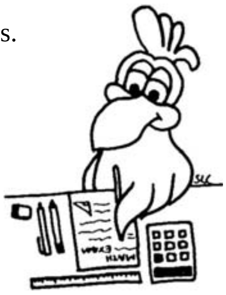
The coefficient of variation for the stock prices is 4.93%. The prices have not showed a large variation over the five days of trading.

The coefficient of variation is often used to compare the variability of two data sets. It allows comparison regardless of the units of measurement used for each set of data.

The larger the coefficient of variation, the more the data varies.

Example STSD-E2

The results of two tests are shown below. Compare the variability of these data sets.



Test 1 (out of 15 marks): $\overline{x} = 9 \quad s = 2$

Test 2 (out of 50 marks): $\overline{x} = 27 \quad s = 8$

$$CV_{test1} = \frac{s}{\overline{x}} \times 100\% = \frac{2}{9} \times 100\% \approx 22.2\% \qquad CV_{test2} = \frac{s}{\overline{x}} \times 100\% = \frac{8}{27} \times 100\% \approx 29.6\%$$

The results in the second test show a great variation than those in the first test.

Exercise STSD-E1

1. Calculate the coefficient of variation for each of the following data sets.
   (a) Stock prices: 8, 10, 9, 10, 11
   (b) Test results: 10, 5, 8, 9, 2, 12, 5, 7, 5, 8
2. Compare the variation of the following data sets.
   (a) Data set A: 35, 38, 34, 36, 38, 35, 36, 37, 36
   Data set B: 36, 20, 45, 40, 52, 46, 26, 26, 32
   (b) Boy's Heights: $\overline{x} = 141.6cm \quad s = 15.1cm$
   Girl's Heights: $\overline{x} = 143.7cm \quad s = 8.4cm$

STSD-F Normal Distribution and *z*-Scores

Another use of the standard deviation is to convert data to a standard score or **z-score**. The *z*-score indicates the number of standard deviations a raw score deviates from the mean of the data set and in which direction, i.e. is the value greater or less than the mean?

The following formula allows a raw score, *x*, from a data set to be converted to its equivalent standard value, *z*, in a new data set with a mean of zero and a standard deviation of one.

$$\boxed{z = \frac{x - \overline{x}}{s} \quad \text{sample}} \qquad \boxed{z = \frac{x - \mu}{\sigma} \quad \text{population}}$$

A *z*-score can be positive or negative:

- positive *z*-score – raw score greater than the mean
- negative *z*-score – raw score less than the mean.

Examples STSD-F1

1. Given the scores 4, 7, 8, 1, 5 determine the z-score for each raw score.

$$\bar{x} = \frac{\sum x}{n} = \frac{25}{5} = 5$$

$$s = \sqrt{\frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1}}$$
$$\approx 2.7386$$

x	x^2
4	16
7	49
8	64
1	1
5	25
$\Sigma x = 25$	$\Sigma x^2 = 155$

raw score	z-score	meaning
4	$z = \frac{4-5}{2.7386} \approx -0.37$	0.37 standard deviations below the mean
7	$z = \frac{7-5}{2.7386} \approx 0.73$	0.73 standard deviations above the mean
8	$z = \frac{8-5}{2.7386} \approx 1.1$	1.1 standard deviations above the mean
1	$z = \frac{1-5}{2.7386} \approx -1.46$	1.46 standard deviations below the mean
5	$z = \frac{5-5}{2.7386} \approx 0$	at the mean

2. Given a data set with a mean of 10 and a standard deviation of 2, determine the z-score for each of the following raw scores, x.

$x = 8$	$z = \frac{8-10}{2} = -1$	8 is 1 standard deviations below the mean.
$x = 10$	$z = \frac{10-10}{2} = 0$	10 is 0 standard deviations from the mean, it is equal to the mean.
$x = 16$	$z = \frac{16-10}{2} = 3$	16 is 3 standard deviations above the mean.

The z-scores also allow comparisons of scores from different sources with different means and/or standard deviations.

Example STSD-F2

Jenny obtained results of 48 in her English exam and 75 in her History exam. Compare her results in the different subjects considering:

English exam : class mean was 45 and the standard deviation 2.25
History exam : class mean was 78 and the standard deviation 2.4

$$z_{English} = \frac{48-45}{2.25} \approx 1.33$$

$$z_{History} = \frac{75-78}{2.4} = -1.25$$

Jenny’s English result is 1.33 standard deviations above the class mean, while her History was 1.25 standard deviations below the class mean.

It is also possible to determine a **raw score** for a given *z*-score, i.e. it is possible to find a value that is a specified number of standard deviations from a mean. The *z*-score formula is transformed to

$\boxed{x = \overline{x} + s \times z \quad \text{sample}}$ $\boxed{x = \mu + \sigma \times z \quad \text{population}}$

Examples STSD-F3

A data set has a mean of 10 and a standard deviation of 2. Find a value that is:

(i) 3 standard deviations above the mean

$z = 3$ $$\begin{aligned} x &= \overline{x} + s \times z \\ &= 10 + 2 \times 3 \\ &= 16 \end{aligned}$$ 16 is three standard deviations above the mean.

(ii) 2 standard deviations below the mean

$z = -2$ $$\begin{aligned} x &= \overline{x} + s \times z \\ &= 10 + 2 \times -2 \\ &= 6 \end{aligned}$$ 6 is two standard deviations below the mean.

Exercise STSD-F1

1. Given the scores 56, 82, 74, 69, 94 determine the *z*-score for each raw score.
2. Given a data set with a mean of 54 and a standard deviation of 3.2, determine the *z*-score for each of the following raw scores, *x*.

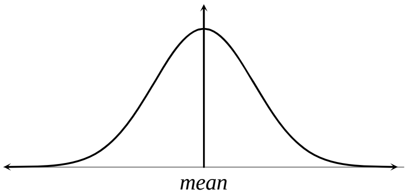
   (i) $x = 81$ (ii) $x = 57$
3. Peter obtained results of 63 in Maths and 58 in Geography. For Maths the class mean was 58 and the standard deviation 3.4, and for Geography the class mean was 55 and the standard deviation 2.3. Compared to the rest of the class did Peter do better in Maths or Geography?
4. A data set has a mean of 54 and a standard deviation of 3.2. Find a value that is:

   (i) 2 standard deviations below the mean.

   (ii) 1.5 standard deviations above the mean.

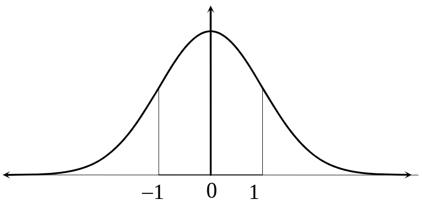
The distribution of *z*-scores is described by the **standard normal curve**.

Normal distributions are symmetric about the mean, with scores more concentrated in the centre of the distribution than in the tails. Normal distributions are often described as bell shaped.

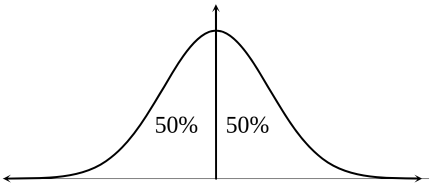


Data collected on heights, weights, reading abilities, memory and IQ scores often are approximated by normal distributions.

The **standard normal distribution** is a normal distribution with a mean of zero and a standard deviation of one.



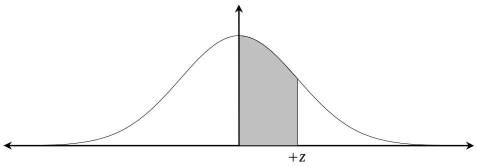
For normally distributed data 50% of the data is below the mean and 50% above the mean.

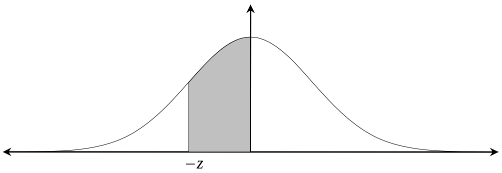


In a normal distribution, the distance between the mean and a given *z*-score corresponds to a proportion of the total area under the curve, and hence can be related to a proportion of a population. The total area under a normal distribution curve is taken as equal to 1 or 100%.

The values of the proportions, written as decimals, for various *z*-scores are provide in a statistical table, Normal Distribution Areas (see *Appendix*).

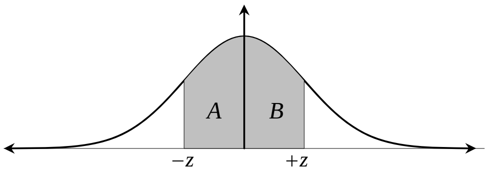
The values in the *Normal Distribution Areas table* give a proportion value for the area between the mean and the raw score greater than the mean, converted to a positive *z*-score.





As the distribution is symmetric, it is possible to calculate proportions for raw scores less than the mean. If the calculated *z*-score is negative, the corresponding positive value from the table is used.

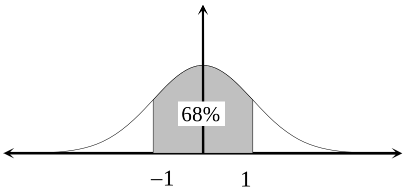
So values for *z* are the same distance from the mean whether they are a number of standard deviations more or a number of standard deviations less than the mean, and will result in the same proportion.



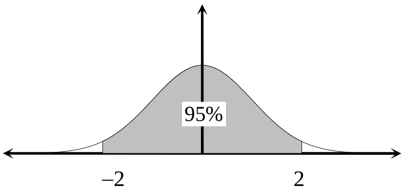
$$\boxed{\text{proportion } A = \text{proportion } B}$$

In a normal distribution approximately:

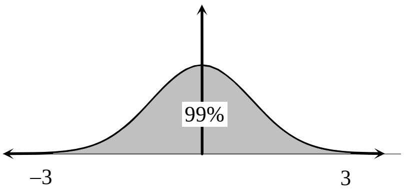
68% of values lie within ±1 s.d. of the mean



95% of values lie within ±2 s.d. of the mean



99% of values lie within ±3 s.d. of the mean



When using the *Normal Distribution Areas table*, the *z*-score is structured from the row and column headings of the table and the required proportion is found in the middle of the table at the intersection of the corresponding rows and columns.

Examples STSD-F4

Use the *Normal Distribution Areas table* to determine the proportions that correspond to the following *z*-scores.

(i) $z = 2$

- $2 = 2.00 = 2.0 + .00 \Rightarrow$ 2.0 row, .00 column

proportion = 0.4772 = 47.72%

(ii) $z = 2.1$

- $2.1 = 2.10 = 2.1 + .00 \Rightarrow$ 2.1 row, .00 column

proportion = 0.4821 = 48.21%

(iii) $z = 2.12$

- $2.12 = 2.1 + .02 \Rightarrow$ 2.1 row, .02 column

proportion = 0.4830 = 48.30%

(iv) $z = -2.2$

- use $z = 2.2$
- $2.2 = 2.20 = 2.2 + .00 \Rightarrow$ 2.2 row, .00 column

proportion = 0.4861 = 48.61%

(v) $z = -2.21$

- use $z = 2.21$
- $2.21 = 2.2 + .01 \Rightarrow$ 2.2 row, .01 column

proportion = 0.4864 = 48.64%

z	.00	.01	.02
⋮	⋮	⋮	⋮
2.0	0.4772	0.4778	0.4783
2.1	0.4821	0.4826	0.4830
2.2	0.4861	0.4864	0.4868
⋮	⋮	⋮	⋮

It is also possible to find a *z*-score for a required proportion from the *Normal Distribution Areas table*. It is necessary to find the proportion in the middle of the table and read back to the row and column headings to determine the *z*-score.

Examples STSD-F5

Use the *Normal Distribution Areas table* to determine the *z*-scores that correspond to the following proportions.

(i) $48.21\% = 0.4821$

- 2.1 row, .00 column $\Rightarrow 2.1 + .00 = 2.10 = 2.1$

The z-score would be either $z = 2.1$ or $z = -2.1$.

(ii) $48.68\% = 0.4868$

- 2.2 row, .02 column $\Rightarrow 2.2 + .02 = 2.22$

The z-score would be either $z = 2.22$ or $z = -2.22$.

z	.00	.01	.02
⋮	⋮	⋮	⋮
2.0	0.4772	0.4778	0.4783
2.1	0.4821	0.4826	0.4830
2.2	0.4861	0.4864	0.4868
⋮	⋮	⋮	⋮

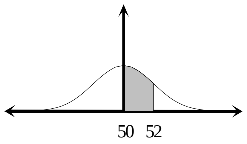
Values from the *Normal Distribution Areas table* enable the determination of proportions of the data set that lie above a raw value, below a raw value or even between two raw values within the data set.

Drawing a quick sketch of the distribution curve with the position of the mean and the raw scores(s) and shading the required area can assist with the understanding of the required calculations.

Examples STSD-F6

1. The weights of chips in packets have a mean of 50*g* and standard deviation of 3*g*.

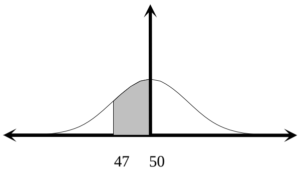
(a) Find the proportion of the packets of chips with a weight between 50*g* and 52*g*.



$z_{52} = \dfrac{52-50}{3} = 0.67 \;\Rightarrow\; 0.2486$ *(from the table)*

24.86% of the chip packets have a weight between 50*g* and 52*g*.

(b) Find the proportion of the packets of chips with a weight between 47*g* and 50*g*.

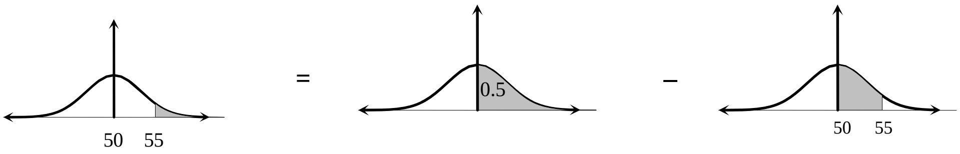


$z_{47} = \dfrac{47-50}{3} = -1$ *(from the table)*

$\Rightarrow\; 0.3413$

34.13% of the chip packets have a weight between 47*g* and 50*g*.

(c) Find the proportion of the packets of chips with a weight greater than 55*g*.



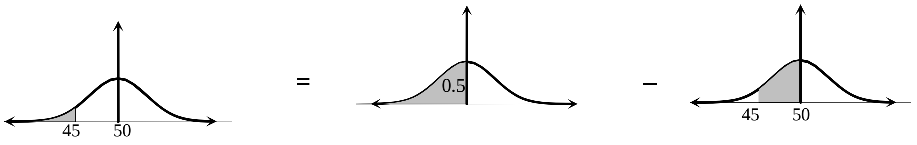
$z_{55} = \dfrac{55-50}{3} = 1.67$

$\Rightarrow 0.4525$

$$\begin{aligned}\text{Area} > 55 &= 0.5 - 0.4525\\ &= 0.0475\end{aligned}$$

4.75% of the chip packets have a weight greater than 55*g*.

(d) Find the proportion of the packets of chips with a weight less than 45*g*.



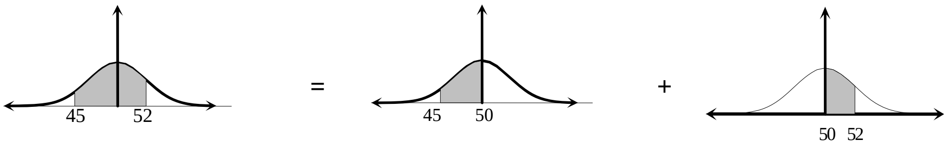
$z_{45} = \dfrac{45-50}{3} = -1.67$

$\Rightarrow 0.4525$

$$\begin{aligned}\text{Area} < 45 &= 0.5 - 0.4525\\ &= 0.0475\end{aligned}$$

4.75% of the chip packets have a weight less than 45*g*.

(e) Find the proportion of the packets of chips with a weight between 45*g* and 52*g*.



$z_{45} = \dfrac{45-50}{3} = -1.67$ $\quad z_{52} = \dfrac{52-50}{3} = 0.67$

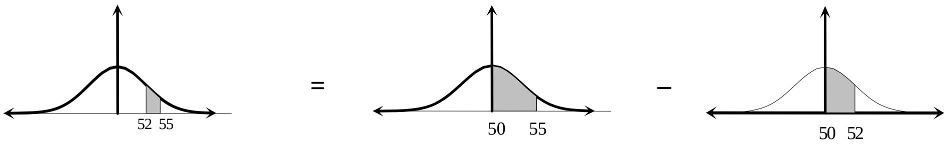
$\Rightarrow 0.4525$ $\quad\Rightarrow\; 0.2486$

$$\begin{aligned}\text{Area between 45 \& 52} &= 0.4525 + 0.2486\\ &= 0.7011\end{aligned}$$

70.11% of the chip packets have a weight between 45*g* and 52*g*.

Examples STSD-F6 continued

1. (f) Find the proportion of the packets of chips with a weight between 52*g* and 55*g*.



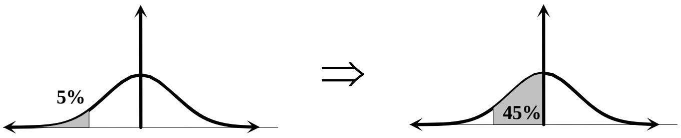
$z_{52} = \dfrac{52-50}{3} = 0.67$ $\quad z_{55} = \dfrac{55-50}{3} = 1.67$

$\Rightarrow\ 0.2486$ $\quad \Rightarrow 0.4525$

Area between 52 & 55 $= 0.4525 - 0.2486$
$= 0.2039$

20.39% of the chip packets have a weight between 52*g* and 55*g*.

2. (a) If the company selling the chips in the previous question rejects the 5% of the chip packets that are too light, at what weight should the chip packets be rejected?



$\bar{x} = 50g \ \ s = 3g$

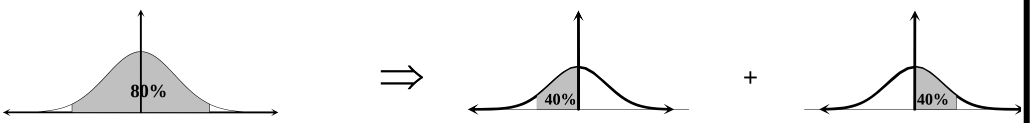
$x = \bar{x} - s \times z$
$= 50 - 3 \times 1.645$
≈ 45.065

proportion $= 0.4500$ (below mean)

$\Rightarrow z \approx -1.645$ *between* 1.64 and 1.65

Packets should be rejected if they weigh 45*g* or less.

(b) Between which weights do 80% of the chip packets fall?



$\bar{x} = 50g \ \ s = 3g$

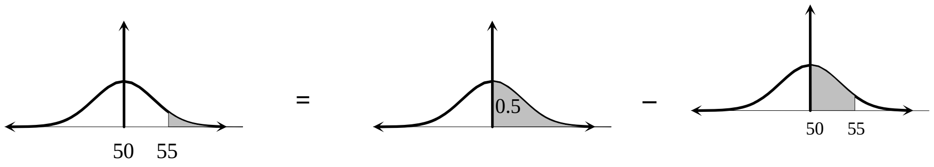
$x = \bar{x} \pm s \times z$
$= 50 \pm 3 \times 1.28$
≈ 53.84 and 46.16

proportion $= 0.4$ (above and below mean)

$\Rightarrow z \approx \pm 1.28$

80% of the packets weigh between 46.16*g* and 53.84*g*.

3. If there are 80 packets in a vending machine, how many packets would be expected to weight more than 55*g*?



$z_{55} = \dfrac{55-50}{3} = 1.67$

$\Rightarrow 0.4525$

Area > 55 $= 0.5 - 0.4525$
$= 0.0475$

Number > 55 $= 0.0475 \times 80 = 3.8$

Approximately 4 chip packets would be expected to have a weight of greater than 55*g*.

Exercise STSD-F2

1. If scores are normally distributed with a mean of 30 and a standard deviation of 5, what percentage of the scores is

(i) greater than 30?
(ii) between 28 and 30?
(iii) greater than 37?
(iv) between 28 and 34?
(v) between 26 and 28?

Exercise STSD-F2 continued

2. IQ scores have a mean of 100 and a standard deviation of 16. What proportion of the population would have an IQ of:

   (i) greater than 132?
   (ii) less than 91?
   (iii) between 80 and 120?
   (iv) between 80 and 91?
   (v) If Shane is smarter than 75% of the population, what is his IQ score?

3. The heights of boys in a school are approximately normally distributed with a mean of 140*cm* and a standard deviation of 20*cm*.

   (a) Find the probability that a boy selected at random would have a height of less than 175*cm*.

   (b) If there are 400 boys in the school how many would be expected to be taller than 175*cm*?

   (c) If 15% of the boys have a height that is less than the girls' mean height, what is the girls' mean height?

4. Charlie is a sprinter who runs 200*m* in an average time of 22.4 seconds with a standard deviation of 0.6*s*. Charlie's times are approximately normally distributed.

   (a) To win a given race a time of 21.9*s* is required. What is the probability that Charlie can win the race?

   (b) The race club that the sprinter is a member of has the record time for the 200*m* race at 20.7*s*. What is the likelihood that Charlie will be able to break the record?

   (c) A sponsor of athletics carnivals offers $100 every time a sprinter breaks 22.5*s*. If Charlie competes in 80 races over a year how much sponsorship money can he expect?

STSD-G Chebyshev's Theorem

A Russian mathematician, P.L. Chebyshev, developed a theorem that approximated the proportion of data values lying within a given number of standard deviations of the mean regardless of whether the data is normally distributed or not. **Chebyshev's theorem** states:

For any data set, at least $\left(1-\frac{1}{k^2}\right)$ of the values lie within k standard deviations either side of the mean. $(k > 1)$

So for any set of data at least:

$\left(1-\frac{1}{\mathbf{2}^2}\right)=\frac{3}{4}=75\%$ of the values lie within **2** standard deviations.

$\left(1-\frac{1}{\mathbf{3}^2}\right)=\frac{8}{9}\approx 89\%$ of the values lie with in **3** standard deviations.

This theorem allows the determination of the least percentage of values that must lie between certain bounds identified by standard deviations.

Examples STSD-G1

1. The heights of adult dogs in a town have a mean of 67.3*cm* and a standard deviation of 3.4*cm*.

(a) What can be concluded from Chebyshev's theorem about the percentage of dogs in the town that have heights between 58.8*cm* and 75.8*cm*?

$$z_{58.8} = \frac{58.8-67.3}{3.4} = -2.5 \qquad z_{75.8} = \frac{75.8-67.3}{3.4} = 2.5$$

$$\left(1-\frac{1}{2.5^2}\right)$$
$$=1-0.16$$
$$=84\%$$

At least 84% of the adult dogs would have heights between 58.8*cm* and 75.8*cm*.

2. (b) What would be the range of heights that would include at least 75% of the dogs?

$$75\% = 1-\frac{1}{k^2}$$
$$0.75 = 1-\frac{1}{k^2}$$
$$\frac{1}{k^2} = 1-0.75$$
$$\frac{1}{k^2} = 0.25$$
$$\frac{1}{0.25} = k^2$$
$$4 = k^2 \therefore k = \sqrt{4} = 2$$

Chebyshev's theorem suggests that 75% of the heights are within ± 2 standard deviations of the mean.

$$x = \mu \pm z\sigma$$
$$= 67.3 \pm 2 \times 3.4$$
$$= 74.1 \text{ and } 60.5$$

75% of the dog's heights would range between 60.5*cm* and 74.1*cm*.

Exercise STSD-G1

1. The weights of cattle have a mean of 434*kg* and standard deviation of 69*kg*. What percentage of cattle will weigh between 330.5*kg* and 537.5*kg*?
2. The age of pensioners residing in a retirement village has a mean of 74 years and standard deviation of 4.5 years. What is the age range of pensioners that contains at least 89% of the residents?
3. It was found that for a batch of softdrink bottles, the mean content was 994*ml*. If 75% of the bottles contained between 898*ml* and 1090*ml*, what was the standard deviation for the softdrink batch?
4. On a test the mean is 50 marks and standard deviation 11. At most, what percentage of the results will be less than 17 and greater than 83 marks?

STSD-H Correlation and Scatterplots

When two different data variables, quantities, are collected from the one source it is possible to determine if a relationship exists between the variables. A simple method of determining the relationship between two variables, if it exists, is by constructing a scatterplot.

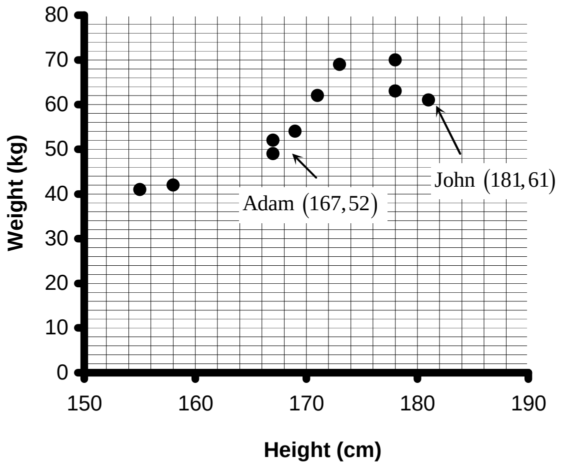
A **scatterplot** (scatter graph or scatter diagram) is a graph that is created by plotting one variable, quantity, on the horizontal axis and the other on the vertical axis. If one variable is likely to be dependent on the other, the dependent variable should be plotted on the vertical axis and the independent variable on the *x*-axis. The scales on the vertical and horizontal axes do not need to be the same or even use the same units. Also the axes do not need to commence at zero.

Examples STSD-H1

1. The heights and weights of 10 students are recorded below. Construct a scatterplot for this data.

Student	Height (*cm*)	Weight (*kg*)
Adam	167	52
Brent	178	63
Charlie	173	69
David	155	41
Eddy	171	62
Fred	167	49
Gary	158	42
Harry	169	54
Ian	178	70
John	181	61

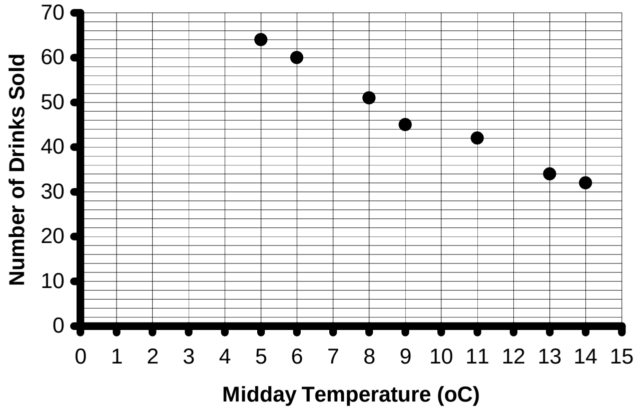
Scatterplot of Weight against Height



2. For one week the midday temperature and the number of hot drinks sold were recorded. Construct a scatterplot for this data.

	Sun	Mon	Tues	Wed	Thur	Fri	Sat
Temp (°C)	9	13	14	11	8	6	5
Number of Drinks	45	34	32	42	51	60	64

Scatterplot of Hot Drinks Sales Against Temperature



Exercise STSD-H1

Construct scatterplots for the following data sets.

1.

Student	A	B	C	D	E	F	G	H	I	J
Test 1 (/50)	33	36	15	29	16	29	44	30	44	23
Final Exam %	75	87	34	56	39	45	92	69	93	59

2.

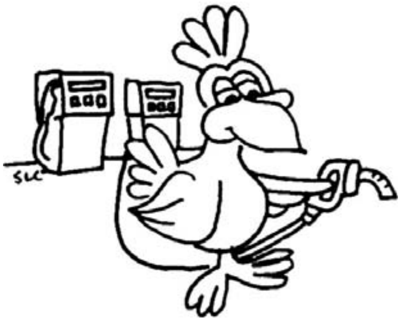
Day	Mon	Tue	Wed	Thur	Fri	Sat	Sun
Temperature (°C)	24	28	30	33	32	35	31
Softdrink Sales (cans)	17	22	27	29	30	36	29

3.

Name	Ann	Lee	May	Jan	Tom	Wes
Height (*cm*)	176	181	173	169	178	180
Shoe Size	9	9.5	8.5	8	10.5	10

4.

Speed (*km/hr*)	50	55	65	70	80	100	120	130
Fuel Economy (*km/l*)	18.9	18.6	18.1	17.3	16.7	14.7	13.2	11.2



The points in a scatterplot often tend towards approximating a line. It is possible to summarise the points of a scatterplot by drawing a line through the plot as a whole, not necessarily through the individual points. This line is called the **line of best fit**.

The line of best fit line need not pass through any of the original data points, but is used to represent the entire scatterplot. A line of best fit can be thought of as an average for the scatterplot, in a way similar to a mean is the average of a list of values.

To **sketch a line of best fit** for a scatterplot:

1. calculate the mean of the independent variable values, $\overline{x}$, and the mean of the dependent variable values, $\overline{y}$;
2. plot this mean point, $\left(\overline{x},\overline{y}\right)$ to the scatterplot;
3. sketch a through the mean point, that has a slope the follows the general trend of the points of scatterplot.

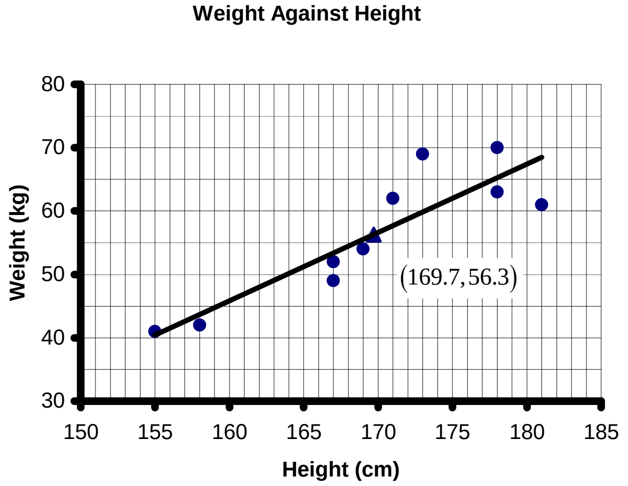
Wherever possible the number of points below the line should equal the number of points above. Outlying points need not strongly influence the line of best fit, and are often not included.

The line of best fit can be used to predict values for data associated with the scatterplot.

Examples STSD-H2

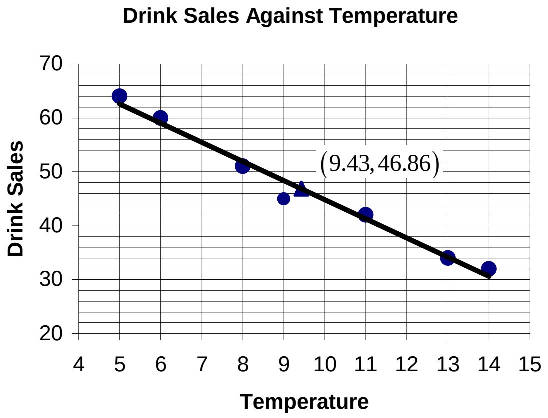
1. The heights and weights of 10 students are recorded below. Construct a scatterplot for this data and draw a line of best fit.

Student	Height (*cm*)	Weight (*kg*)
Adam	167	52
Brent	178	63
Charlie	173	69
David	155	41
Eddy	171	62
Fred	167	49
Gary	158	42
Harry	169	54
Ian	178	70
John	181	61
Means	$\bar{x} = \frac{\Sigma x}{n} = \frac{1697}{10} = 169.7$	$\bar{y} = \frac{\Sigma x}{n} = \frac{563}{10} = 56.3$



2. For one week the midday temperature and the number of hot drinks sold were recorded. Construct a scatterplot for this data.

	Sun	Mon	Tues	Wed	Thur	Fri	Sat	means
Temp (°C)	9	13	14	11	8	6	5	9.43
Number of Drinks	45	34	32	42	51	60	64	46.86

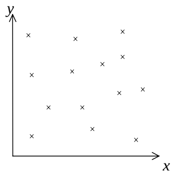
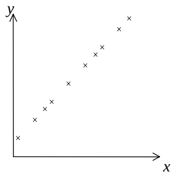
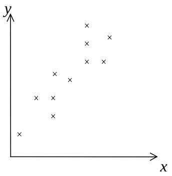
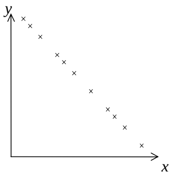
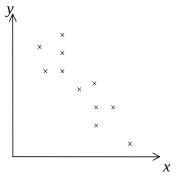


Exercise STSD-H2 Draw lines of best fit on scatterplots drawn in Exercise **STSD-H1**.

Correlation is a measure of the relationship between two measures, variables, on sets of data. Correlation can be positive or negative.

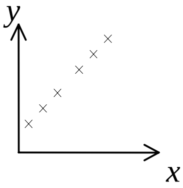
A **positive correlation** means that as one variable increases the other variable increases, eg. height of a child and age of the child. **Negative correlation** implies as one variable increases the other variable decrease, eg. value of a car and age of the car. If variables have **no correlation** there is no relationship between the variables, i.e. one measure does not affect the other.

Scatterplots enable the visual determination of whether correlation exists between variables and the type of correlation.

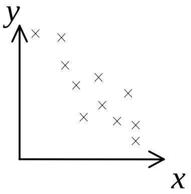
 no correlation	 perfect positive correlation	 positive correlation	 perfect negative correlation	 negative correlation

Exercise STSD-H3

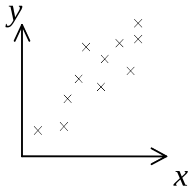
1. For each of the following scatterplots determine if the correlation is perfect positive, positive, no correlation, negative or perfect negative.



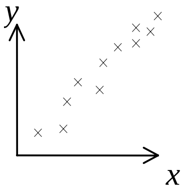
A



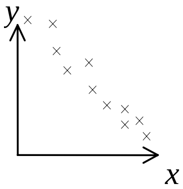
B



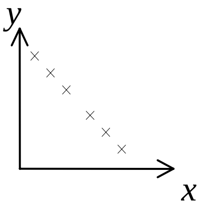
C



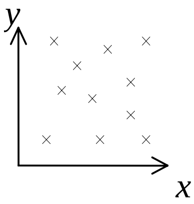
D



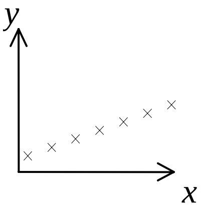
E



F



G



H

2. Select the scatterplot from those above that would best describe the relationship between the following variables.

   (i) Height at 4 years, height at 16 years.

   (ii) Age of a used car, price of a used car.

   (iii) Temperature at 6am, temperature at 3pm.

   (iv) Shoe size of mother, number of children in family.

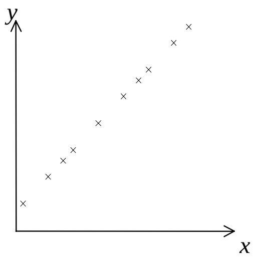
   (v) Average exam result, class size.

3. For each of the scatterplots drawn in exercise **STSD- H1** state the type of correlation between the variables.

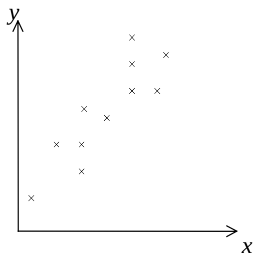
STSD-I Correlation Coefficient and Regression Equation

It is possible to quantify the correlation between variables. This is done by calculating a **correlation coefficient**. A correlation coefficient measures the strength of the linear relationship between variables.

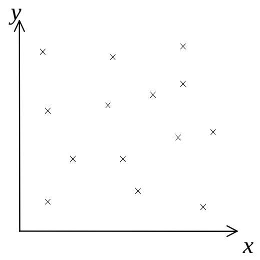
Correlation coefficients can range from –1 to +1. A value of –1 represents a perfect negative correlation and a value of +1 represents a perfect positive correlation. If a data set has a correlation coefficient of zero there is no correlation between the variables.



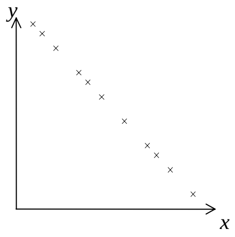
perfect positive correlation
$r = 1$



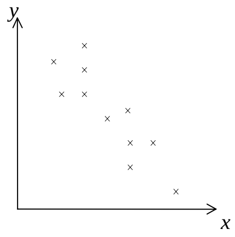
positive correlation
$r \approx 0.7$



no correlation
$r = 0$



perfect negative correlation
$r = -1$



negative correlation
$r \approx -0.7$

The most widely used type of correlation coefficient is Pearson's, r, simple linear correlation. The value of r is determined with the formula below

$$\boxed{r = \frac{\Sigma\left(X-\overline{X}\right)\left(Y-\overline{Y}\right)}{\sqrt{\Sigma\left(X-\overline{X}\right)^2\,\Sigma\left(Y-\overline{Y}\right)^2}}}$$

This formula uses the sums of deviations from the means in both the X values and Y values.

However for ease of calculation the following **calculation formula** is often used.

$$\boxed{r = \frac{n\Sigma XY - \Sigma X\,\Sigma Y}{\sqrt{\left(n\Sigma X^2 - \left(\Sigma X\right)^2\right)\left(n\Sigma Y^2 - \left(\Sigma Y\right)^2\right)}}}$$

where,	n	number of data points
	ΣX	sum of the X values
	ΣY	sum of the Y values
	ΣXY	sum of the product of each set of X and Y values
	ΣX^2	sum of X^2
	ΣY^2	sum of Y^2
	$\left(\Sigma X\right)^2$	the square of the sum of the X values
	$\left(\Sigma Y\right)^2$	the square of the sum of the Y values

To assist in calculations data can be set up in a table and the following headings used:

X	Y	X^2	Y^2	XY

Example STSD-I1

Calculate the correlation coefficient for the height weight data below.

Height , X	Weight , Y	X^2	Y^2	XY
167	52	27889	2704	8684
178	63	31684	3969	11214
173	69	29929	4761	11937
155	41	24025	1681	6355
171	62	29241	3844	10602
167	49	27889	2401	8183
158	42	24964	1764	6636
169	54	28561	2916	9126
178	70	31684	4900	12460
181	61	32761	3721	11041
$\Sigma X = 1697$	$\Sigma Y = 563$	$\Sigma X^2 = 288627$	$\Sigma Y^2 = 32661$	$\Sigma XY = 96238$

$$r = \frac{n\Sigma XY - \Sigma X \Sigma Y}{\sqrt{\left(n\Sigma X^2 - (\Sigma X)^2\right)\left(n\Sigma Y^2 - (\Sigma Y)^2\right)}}$$

$$= \frac{10\times 96238 - 1697\times 563}{\sqrt{\left(10\times 288627 - 1697^2\right)\left(10\times 32661 - 563^2\right)}} \approx 0.882998$$

There is a positive correlation between the height and weight of students.

NOTE: r^2 , sometimes referred to as the **co-efficient of determination**, represents the proportion (percentage) of the relationship between the variables that can be explained by a linear relationship. The greater the r^2 value, the greater the linear relationship between the variables.

Example STSD-I2

For the previous height/weight example

$$r \approx 0.882998$$
$$r^2 \approx 0.779685$$

$\therefore$ **approximately 78% of the relationship can be explained by a linear correlation. This is a moderately strong correlation.**

Exercise STSD-I1

Determine the correlation co-efficient for the following data sets.

1.

Temperature (°C)	9	13	14	11	8	6	5
Hot drink sales (cups)	45	34	32	42	51	60	64

2.

Speed (*km/hr*)	50	55	65	70	80	100	120	130
Fuel economy (*km/l*)	18.9	18.6	18.1	17.3	16.7	14.7	13.2	11.2

3. Data collected of students results for sitting Test 1 and the final exam (**exercise STSD H1 question 1**) Determine the correlation co-efficient for this data.
X : Test 1 results (out of 15 marks) Y : Final exam result (out of 50 marks)
$\Sigma X = 299, \Sigma Y = 649, \Sigma X^2 = 9849, \Sigma Y^2 = 46387, \ \Sigma XY = 21237, n = 10$

The relationship between two sets of data can be represented by a linear equation called a **regression equation**. The regression equation gives the variation of the dependent variable for a given change in the independent variable. It is extremely important to correctly determine which variable is dependent.

The regression equation can be used to construct the **regression line** (line of best fit) on the associated scatterplot. Because the equation is for a line, the regression equation takes on the general linear equation format, $y = mx + c$. Usually, however, for a regression equation this is written as $y = \alpha + \beta x$, where α is the y-intercept and β the slope of the line.

$$y = \alpha + \beta x \quad \text{where} \quad \beta = \frac{\Sigma XY - n\overline{X}\,\overline{Y}}{\Sigma X^2 - n\overline{X}^2}$$

$$\alpha = \overline{Y} - \beta\overline{X}$$

The slope of the line depends on whether the correlation is positive or negative.

A regression equation can be used to predict dependent variables from independent inputs within the range of the scatterplot values. It should **not** be used:

- to predict *x* given *y*
- to predict outside the bounds of given *x* values.

The stronger the correlation between the two variables the better the prediction made by the regression equation.

Again setting up a table of values, with the following headings, can assist in calculations.

X	Y	X^2	XY

Examples STSD-I3

1. Calculate the correlation coefficient for the height/weight data below.

The weight of a student should depend on the height rather than vice versa, so height is the independent, x, variable and weight the dependent, y, variable.

Height, X	Weight, Y	X^2	XY
167	52	27889	8684
178	63	31684	11214
173	69	29929	11937
155	41	24025	6355
171	62	29241	10602
167	49	27889	8183
158	42	24964	6636
169	54	28561	9126
178	70	31684	12460
181	61	32761	11041
$\Sigma X = 1697$	$\Sigma Y = 563$	$\Sigma X^2 = 288627$	$\Sigma XY = 96238$

$$\overline{X} = \frac{\Sigma X}{n} = \frac{1697}{10} = 169.7$$

$$\overline{Y} = \frac{\Sigma Y}{n} = \frac{563}{10} = 56.3$$

$$\begin{aligned}\beta &= \frac{\Sigma XY - n\overline{X}\,\overline{Y}}{\Sigma X^2 - n\overline{X}^2} = \frac{96238 - 10\times 169.7\times 56.3}{288627 - 10\times 169.7^2}\\ &= 1.0786255997...\\ &\approx 1.08\end{aligned}$$

$$\begin{aligned}\alpha &= \overline{Y} - \beta\overline{X}\\ &= 56.3 - 1.078..\times 169.7\\ &\approx -126.7\end{aligned}$$

The regression equation is $y = 1.08x - 126.7$ **OR** $w = 1.08h - 126.7$

Examples STSD-I3 continued

2. Use the regression equation to predict the following:

(i) The weight of a student who is 160*cm* tall.

$h = 160 \qquad \therefore w = 1.08h - 126.7$

$= 1.08 \times 160 - 126.7$

$= 46.1kg$

The student should weigh approximately 46.1*kg*.

(ii) The weight of a student who is 175*cm* tall.

$h = 175 \qquad \therefore w = 1.08h - 126.7$

$= 1.08 \times 175 - 126.7$

$= 62.3kg$

The student should weigh approximately 62.3*kg*.

(iii) The weight of a student who is 185*cm* tall.
Can not be predicted as input is outside range of recorded heights.

(iv) The height of a student who weighs 65*kg*.
Regression equation can not be used to predict height from weight.

Exercise STSD-I2

1. Determine the regression equation for each the following data sets. (Use sums calculated in the previous exercise and the equations to predict the requested values.)

(a)

Temperature (°C)	9	13	14	11	8	6	5
Hot drink sales (cups)	45	34	32	42	51	60	64

(i) Predict the drink sales when the temperature is 10°C.

(ii) Predict the drink sales if the temperature is 25°C.

(b)

Speed (*km/hr*)	50	55	65	70	80	100	120	130
Fuel economy (*km/l*)	18.9	18.6	18.1	17.3	16.7	14.7	13.2	11.2

(i) Predict the fuel economy for a speed of 75*km/hr*.

(ii) Predict what speed a car would be travelling if it was getting 17.5*km/l*.

2. Data collected of students results for sitting Test 1 and the final exam (**exercise STSD H1 question 1**)

X : Test 1 results (out of 15 marks) Y : Final exam result (out of 50 marks)

$\Sigma X = 299, \Sigma Y = 649, \Sigma X^2 = 9849, \Sigma Y^2 = 46387, \ \Sigma XY = 21237, n = 10$

Find the regression equation for this data.

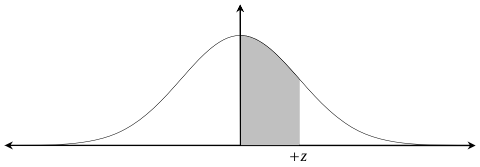
STSD-J Summary

The **mean** is a measure of central tendency. **Standard deviation** measures spread or dispersion of a data set. The **coefficient of variation**, *CV*, gives the standard deviation as a percentage of the mean of the data set. The ***z*-score** indicates how far, the number of standard deviations, a raw score deviates from the mean of the data set.

The following formulae can be used to calculate the the given statistical measures.

Statistical Measure	**Population Formula**	**Sample Formula**
Mean	$\mu = \frac{\sum x}{N}$	$\overline{x} = \frac{\sum x}{n}$
Mean from frequency table	$\mu = \frac{\sum fx}{\sum f}$	$\overline{x} = \frac{\sum fx}{\sum f}$
Standard Deviation	$\sigma = \sqrt{\frac{\Sigma x^2 - \frac{(\Sigma x)^2}{N}}{N}}$	$s = \sqrt{\frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}}$
Standard Deviation from frequency table	$\sigma = \sqrt{\frac{\Sigma fx^2 - \frac{(\Sigma fx)^2}{\Sigma f}}{\Sigma f}}$	$s = \sqrt{\frac{\Sigma fx^2 - \frac{(\Sigma fx)^2}{\Sigma f}}{\Sigma f - 1}}$
Coefficient of Variation	$CV = \frac{\sigma}{\mu} \times 100\%$	$CV = \frac{s}{\overline{x}} \times 100\%$
z-Score	$\boxed{z = \frac{x - \mu}{\sigma}}$	$\boxed{z = \frac{x - \overline{x}}{s}}$
Raw Score from z-Score	$\boxed{x = \mu + \sigma \times z}$	$\boxed{x = \overline{x} + s \times z}$

The **standard normal distribution** is a normal distribution with a mean of zero and a standard deviation of one. The distance between the mean and a given *z*-score corresponds to a proportion of the total area under the curve, and hence can be related to a proportion of a population. The total area under a normal distribution curve is taken as equal to 1 or 100%. The values in the *Normal Distribution Areas table* give a proportion value for the area between the mean and the raw score greater than the mean, converted to a positive *z*-score.



In a normal distribution approximately: 68% of values lie within ± 1 s.d. of the mean; 95% of values lie within ± 2 s.d. of the mean; and 99% of values lie within ± 3 s.d. of the mean.

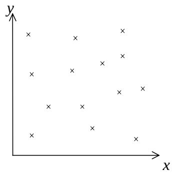
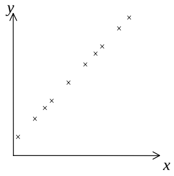
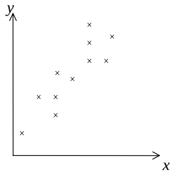
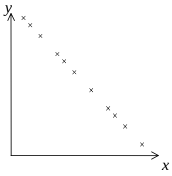
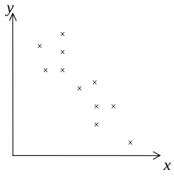
Chebyshev's theorem states: for any data set, at least $\left(1 - \frac{1}{k^2}\right)$ of the values lie within *k* standard deviations either side of the mean. $(k > 1)$.

A **scatterplot** is a graph that is created by plotting one variable, quantity, on the horizontal axis and the other on the vertical axis.

To sketch a **line of best fit** for a scatterplot:

1. calculate the mean of the independent variable values, $\overline{x}$, and the mean of the dependent variable values, $\overline{y}$;
2. plot this mean point, $\left(\overline{x}, \overline{y}\right)$ to the scatterplot;
3. sketch a line through the mean point, that has a slope that follows the general trend of the points of scatterplot.

Correlation is a measure of the relationship between two measures, variables, on sets of data. Correlation can be positive or negative.

				
no correlation	perfect positive correlation	positive correlation	perfect negative correlation	negative correlation

A correlation coefficient measures the strength of the linear relationship between variables.

The most widely used type of correlation coefficient is Pearson's, *r*, simple linear correlation. The value of *r* is determined with the **calculation formula**

$$r = \frac{n\Sigma XY - \Sigma X\,\Sigma Y}{\sqrt{\left(n\Sigma X^2 - \left(\Sigma X\right)^2\right)\left(n\Sigma Y^2 - \left(\Sigma Y\right)^2\right)}}$$

r^2 , sometimes referred to as the **co-efficient of determination**, represents the proportion (percentage) of the relationship between the variables that can be explained by a linear relationship.

The relationship between two sets of data can be represented by a linear equation called a **regression equation**.

$$y = \alpha + \beta x \qquad \text{where} \qquad \beta = \frac{\Sigma XY - n\overline{X}\,\overline{Y}}{\Sigma X^2 - n\overline{X}^2}$$

$$\alpha = \overline{Y} - \beta\overline{X}$$

A regression equation can be used to predict dependent variables from independent inputs within the range of the scatterplot values. It should **not** be used:

- to predict *x* , the independent variable, given *y*, the dependent variable.
- to predict outside the bounds of given *x* values.

The stronger the correlation between the two variables the better the prediction made by the regression equation.

STSD-K Review Exercise

1. For each of the following data sets calculate
   (i) the mean
   (ii) the standard deviation
   (iii) the coefficient of variation.

   (a) **Store Sales for a week**

   $552 $698 $547 $720 $645 $451

   (b) **Student Mark in a 5 Mark Test**

Mark	Frequency
0	1
1	2
2	4
3	10
4	8
5	5

   (c) **Daily Rainfall in millimetres**

Rainfall (*mm*)	Frequency (*days*)
0 – 4	2
5 – 9	8
10 – 14	4
15 – 19	3
20 – 24	4

2. A soft-drink filling machine uses cans with a maximum capacity of 340*ml*. The machine is set to output softdrink with a mean capacity of 330*ml*. It has been found that due to machine error the amount outputted varies with a standard deviation of 8*ml* and the amount outputted is normally distributed.
   (a) What proportion of cans will have between 330 *ml* and 340*ml* of softdrink?
   (b) What percentage of cans will have between 325*ml* and 340*ml*?
   (c) What percentage of cans will overflow?
   (d) If the smallest 5% of drinks must be rejected, what is the smallest amount which will be accepted?

3 (a) If a set of data has a mean of 76 and a standard deviation of 28.8, what is the interval that should contain at least 75% of the data?
   (b) A data set has a mean of 827 and a standard deviation of 98. At least what percentage of values should lie been 582 and 1072?
   (c) A set of data has a mean of 468. If 89% of the data values lie between 336 and 600, what is the standard deviation for the data set?

Exercise STSD-K continued

4. (a) Match each of the correlation coefficients with a scatterplot below.

(i)	$r = 0.6$	(iv)	$r = 0.9$
(ii)	$r = 0$	(v)	$r = -1$
(iii)	$r = -0.9$		

(b) Which scatterplot best approximates the correlation between each of the two variables below? (a scatterplot may be used more than once)

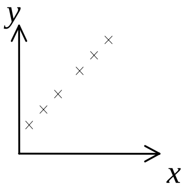
(i) days on a good diet, weight

(ii) temperature outside, temperature in a non-air conditioned car

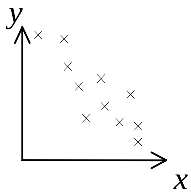
(iii) hand span, height

(iv) rainfall, level of water in river

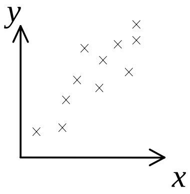
(v) length of finger nails, intelligence



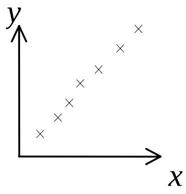
A



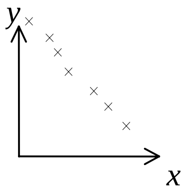
B



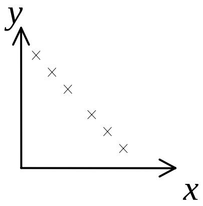
C



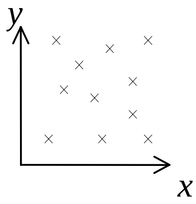
D



E



F



G

5. The average test results for a standard examination and corresponding class size were recorded for five schools. The results are summarised in the table below.

School	Class Size	Test Result
A	28	82%
B	33	50%
C	25	80%
D	14	98%
E	20	90%

Use the class size/test result data to:

(a) construct a scatterplot

(b) draw a line of best fit

(c) determine the correlation coefficient

(d) comment on the correlation

(e) determine the regression equation

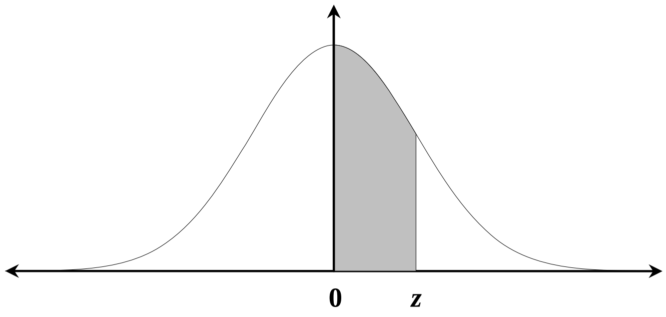
(f) use the regression equation to predict

(i) the expected result if a school had a class size of 30 students

(ii) the expected result for a class of 10 students

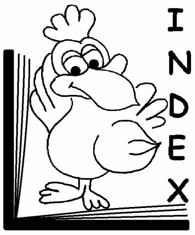
(iii) how many student there would be in a class if the test result was 75%

STSD-L Appendix – Normal Distribution Areas Table



z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

STSD-Y Index



STSD-Z Solutions

STSD –B1 Calculating Means

(a)

$$\text{mean} = \frac{\sum x}{n} = \frac{0+0+1+1+2+2+2+3+3+4+5}{11}$$
$$= \frac{23}{11} \approx 2.09$$

The mean number of hockey goals scored is approximately 2.09 goals.

(b)

Points Scored *(x)*	**Frequency *(f)***	*fx*
10	1	10
11	0	0
12	4	48
13	1	13
14	3	42
15	1	15
Total	$\Sigma f = 10$	$\Sigma fx = 128$

$$\text{mean} = \bar{x} = \frac{\sum fx}{\sum f} = \frac{128}{10} = 12.8$$

The mean number of points scored is 12.8 points.

(c)

Typing errors (*x*)	Frequency (*f*)	*fx*
0	6	0
1	8	8
2	5	10
3	1	3
Total	$\Sigma f = 20$	$\Sigma fx = 21$

$$\text{mean} = \bar{x} = \frac{\sum fx}{\sum f} = \frac{21}{20} = 1.05$$

The mean number of typing errors is 1.05 errors.

(d)

Baby Weight (kg)	**Mid-value *(x)***	**Frequency *(f)***	*fx*
2.80 – 2.99	$\frac{2.8+3}{2} = 2.9$	2	5.8
3.00 – 3.19	3.1	1	3.1
3.20 – 3.39	3.3	3	9.9
3.40 – 3.59	3.5	2	7
3.60 – 3.79	3.7	5	18.5
3.80 – 3.99	3.9	2	7.8
	Total	$\Sigma f = 15$	$\Sigma fx = 52.1$

$$\text{mean} = \bar{x} = \frac{\sum fx}{\sum f} = \frac{52.1}{15} = 3.47\bar{3} \approx 3.47$$

The mean baby weight is approximately 3.47*kg*.

(e)

Withdrawals ($)	**Mid-value *(x)***	**Frequency *(f)***	*fx*
0 – 49	$\frac{0+50}{2} = 25$	7	175
50 – 99	75	9	675
100 – 149	125	5	625
150 – 199	175	5	875
200 – 249	225	2	450
250 – 299	275	2	550
	Total	$\Sigma f = 30$	$\Sigma fx = 3350$

$$\text{mean} = \bar{x} = \frac{\sum fx}{\sum f} = \frac{3350}{30} = 111.\bar{6} \approx 111.67$$

The mean withdrawal was approximately $111.67.

STSD-D1 Calculating Standard Deviations

(a)

x	x^2
5	25
4	16
3	9
2	4
2	4
1	1
0	0
0	0
1	1
2	4
3	9
$\Sigma x = 23$	$\Sigma x^2 = 73$

$$s = \sqrt{\frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}}$$
$$= \sqrt{\frac{73 - \frac{23^2}{11}}{11-1}}$$
$$\approx 1.58$$

The standard deviation of the number of goals scored is approximately 1.58 goals.

(b)

Points (x)	*f*	*fx*	*x²*	*fx²*
10	1	10	100	100
11	0	0	121	0
12	4	48	144	576
13	1	13	169	169
14	3	42	196	588
15	1	15	225	225
	$\Sigma f = 10$	$\Sigma fx = 128$		$\Sigma fx^2 = 1658$

$$s = \sqrt{\frac{\Sigma fx^2 - \frac{(\Sigma fx)^2}{\Sigma f}}{\Sigma f - 1}} = \sqrt{\frac{1658 - \frac{128^2}{10}}{10-1}} \approx 1.48$$

The standard deviation of number of basketball points is approximately 1.48 points.

(c)

Errors (x)	*f*	*fx*	*x²*	*fx²*
0	6	0	0	0
1	8	8	1	8
2	5	10	4	20
3	1	3	9	9
	$\Sigma f = 20$	$\Sigma fx = 21$		$\Sigma fx^2 = 37$

$$s = \sqrt{\frac{\Sigma fx^2 - \frac{(\Sigma fx)^2}{\Sigma f}}{\Sigma f - 1}}$$
$$= \sqrt{\frac{37 - \frac{21^2}{20}}{20-1}} \approx 0.89$$

The standard deviation of the number of typing errors is approximately 0.89 errors.

(d)

Weights	*x*	*f*	*fx*	*x²*	*fx²*
2.80 – 2.99	2.9	2	5.8	8.41	16.82
3.00 – 3.19	3.1	1	3.1	9.61	9.61
3.20 – 3.39	3.3	3	9.9	10.89	32.67
3.40 – 3.59	3.5	2	7	12.25	24.5
3.60 – 3.79	3.7	5	18.5	13.69	68.45
3.80 – 3.99	3.9	2	7.8	15.21	30.42
		$\Sigma f = 15$	$\Sigma fx = 52.1$		$\Sigma fx^2 = 182.47$

$$s = \sqrt{\frac{\Sigma fx^2 - \frac{(\Sigma fx)^2}{\Sigma f}}{\Sigma f - 1}} = \sqrt{\frac{182.47 - \frac{52.1^2}{15}}{15-1}} \approx 0.33$$

The standard deviation of the baby weights is approximately 0.33*kg*.

STSD-D1 continued

(e)

Withdrawals	*x*	*f*	*fx*	x^2	fx^2
0 – 49	25	7	175	625	4375
50 – 99	75	9	675	5625	50625
100 – 149	125	5	625	15625	78125
150 – 199	175	5	875	30625	153125
200 – 249	225	2	450	50625	101250
250 – 299	275	2	550	75625	151250
		$\Sigma f = 30$	$\Sigma fx = 3350$		$\Sigma fx^2 = 538750$

$$s=\sqrt{\frac{\Sigma fx^2-\frac{(\Sigma fx)^2}{\Sigma f}}{\Sigma f-1}}=\sqrt{\frac{538750-\frac{3350^2}{30}}{30-1}}\approx 75.35$$

The standard deviation of the ATM withdrawals is approximately $75.35.

STSD-E1 Coefficient of Variation

1. (a) $\bar{x}=\frac{\sum x}{n}=\frac{8+10+9+10+11}{5}$
$=\frac{48}{5}=9.6$

$\sum x^2=8^2+10^2+9^2+10^2+11^2$
$=466$

$$s=\sqrt{\frac{\Sigma x^2-\frac{(\Sigma x)^2}{n}}{n-1}}=\sqrt{\frac{466-\frac{48^2}{5}}{5-1}}\approx 1.14$$

$$CV=\frac{s}{\bar{x}}\times 100\%=\frac{1.14}{9.6}\times 100\%\approx 11.9\%$$

(b) $\bar{x}=\frac{\sum x}{n}=\frac{71}{10}=7.1$

$\sum x^2=581$

$$s=\sqrt{\frac{\Sigma x^2-\frac{(\Sigma x)^2}{n}}{n-1}}=\sqrt{\frac{581-\frac{71^2}{10}}{10-1}}\approx 2.92$$

$$CV=\frac{s}{\bar{x}}\times 100\%=\frac{2.92}{7.1}\times 100\%\approx 41.1\%$$

2. (a) **Data Set A:**

$\bar{x}=\frac{\sum x}{n}=\frac{325}{9}=36.\bar{1}$

$\sum x^2=11751$

$$s=\sqrt{\frac{\Sigma x^2-\frac{(\Sigma x)^2}{n}}{n-1}}=\sqrt{\frac{11751-\frac{325^2}{9}}{9-1}}\approx 1.36$$

$$CV=\frac{s}{\bar{x}}\times 100\%=\frac{1.36}{36.1}\times 100\%\approx 3.77\%$$

Data Set B:

$\bar{x}=\frac{\sum x}{n}=\frac{323}{9}=35.\bar{8}$

$\sum x^2=12517$

$$s=\sqrt{\frac{\Sigma x^2-\frac{(\Sigma x)^2}{n}}{n-1}}=\sqrt{\frac{12517-\frac{323^2}{9}}{9-1}}\approx 10.75$$

$$CV=\frac{s}{\bar{x}}\times 100\%=\frac{10.75}{35.9}\times 100\%\approx 29.96\%$$

There is greater relative variation in Data set B.

STSD-E1 continued

2. (b) **Boys' Heights:** $\bar{x} = 141.6cm$ $s = 15.1cm$

$$\begin{aligned} CV &= \frac{s}{\bar{x}} \times 100\% \\ &= \frac{15.1}{141.6} \times 100\% \\ &\approx 10.7\% \end{aligned}$$

Girls' Heights: $\bar{x} = 143.7cm$ $s = 8.4cm$

$$\begin{aligned} CV &= \frac{s}{\bar{x}} \times 100\% \\ &= \frac{8.4}{143.7} \times 100\% \\ &\approx 5.8\% \end{aligned}$$

There is greater relative variation in boys' heights.

STSD-F1 ***z*-Scores**

1. $\bar{x} = \frac{\sum x}{n} = \frac{375}{5} = 75$

$$\begin{aligned} s &= \sqrt{\frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1}} \\ &\approx 14.2127 \end{aligned}$$

x	x^2
56	3136
82	6724
74	5476
69	4761
94	8836
$\Sigma x = 375$	$\Sigma x^2 = 28933$

raw score	***z*-score**	**meaning**
56	$z = \frac{56-75}{14.2127} \approx -1.34$	1.34 standard deviations below the mean
82	$z = \frac{82-75}{14.2127} \approx 0.49$	0.49 standard deviations above the mean
74	$z = \frac{74-75}{14.2127} \approx -0.07$	0.07 standard deviations below the mean
69	$z = \frac{69-75}{14.2127} \approx -0.42$	0.42 standard deviations below the mean
94	$z = \frac{94-75}{14.2127} \approx 1.34$	1.34 standard deviations above the mean

2. (i) $x = 81$ $z = \frac{81-54}{3.2} \approx 8.44$ 8.44 s.d. above the mean

(ii) $x = 57$ $z = \frac{57-54}{3.2} \approx 0.94$ 0.94 s.d. above the mean

3. $z_{Maths} = \frac{63-58}{3.4} \approx 1.47$ $z_{Geography} = \frac{58-55}{2.3} = 1.30$

Peter's Maths result is 1.47 standard deviations above the class mean, while his geography was 1.30 standard deviations above the class mean. So Peter did slightly better with his Maths result compared to the rest of the class.

4. (i) $z = -2$

$$\begin{aligned} x &= \bar{x} + s \times z \\ &= 54 + 3.2 \times -2 \\ &= 47.6 \end{aligned}$$

47.6 is two standard deviations below the mean.

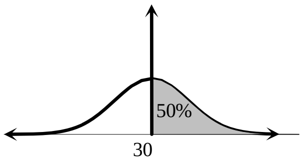
(ii) $z = 1.5$

$$\begin{aligned} x &= \bar{x} + s \times z \\ &= 54 + 3.22 \times 1.5 \\ &= 58.8 \end{aligned}$$

58.8 is 1.5 standard deviations above the mean.

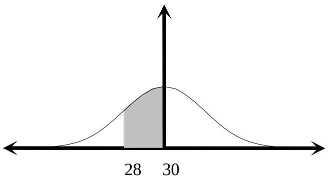
STSD-F2 Normal Distributions

1. (i)



50% of the scores are greater than 30, the mean.

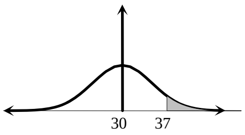
(ii)



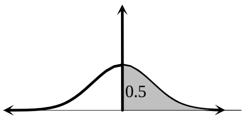
$$z_{28} = \frac{28-30}{5} = -0.4 \Rightarrow \ 0.1554$$

15.54% of the scores are between 28 and 30.

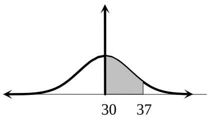
(iii)



=



–

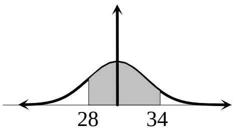


$$z_{37} = \frac{37-30}{5} = 1.4$$
$\Rightarrow 0.4192$

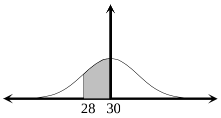
Area > 37 $= 0.5 - 0.4192$
$= 0.0808$

8.08% of the scores are greater than 37.

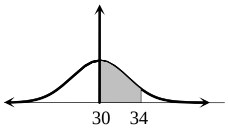
(iv)



=



+



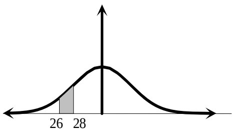
$$z_{28} = \frac{28-30}{5} = -0.4$$
$\Rightarrow 0.1554$

$$z_{34} = \frac{34-30}{5} = 0.8$$
$\Rightarrow \ 0.2881$

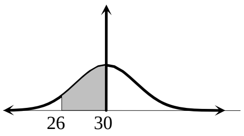
Area between 28 & 34
$= 0.1554 + 0.2881$
$= 0.4435$

44.35% of the score are between 28 and 34.

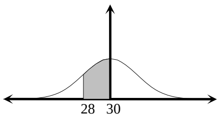
(v)



=



–



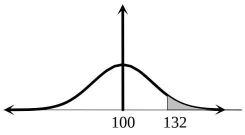
$$z_{26} = \frac{26-30}{5} = -0.8$$
$\Rightarrow \ 0.2881$

$$z_{28} = \frac{28-30}{5} = -0.4$$
$\Rightarrow 0.1554$

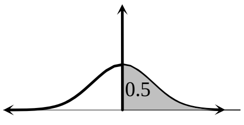
Area between 26 & 28
$= 0.2881 - 0.1554$
$= 0.1327$

13.27% of the score are between 26 and 28.

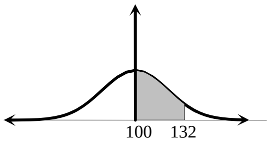
2. (i)



=



–



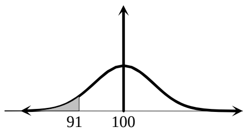
$$z_{132} = \frac{132-100}{16} = 2$$
$\Rightarrow 0.4772$

Area > 132 $= 0.5 - 0.4772$
$= 0.0228$

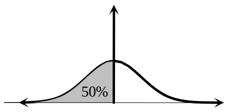
2.28% of the population have IQs greater than 132.

STSD-F2 continued

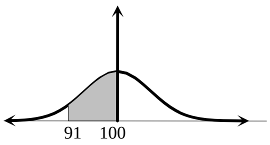
2. (ii)



=



−

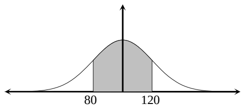


$$z_{91} = \frac{91-100}{16} = -0.5625$$
$\Rightarrow 0.2123$

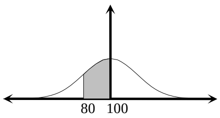
$$\text{Area} < 91 = 0.5 - 0.2123$$
$$= 0.2877$$

28.77% of the population have IQs less than 91.

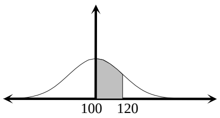
(iii)



=



+



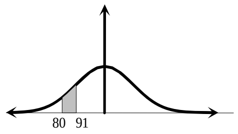
$$z_{80} = \frac{80-100}{16} = -1.25$$
$\Rightarrow 0.3944$

$$z_{120} = \frac{120-100}{16} = 1.25$$
$\Rightarrow$ 0.3944

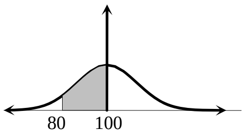
Area between 80 & 120
$= 2 \times 0.3944$
$= 0.7888$

78.88% of the population have IQs between 80 and 120.

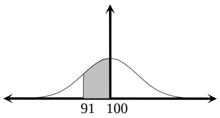
(iv)



=



−



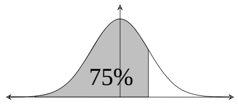
$$z_{80} = \frac{80-100}{16} = -1.25$$
$\Rightarrow 0.3944$

$$z_{91} = \frac{91-100}{16} = -0.5625$$
$\Rightarrow 0.2123$

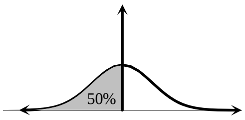
Area between 80 & 91
$= 0.3944 - 0.2123$
$= 0.1821$

18.21% of the population have IQs between 80 and 91.

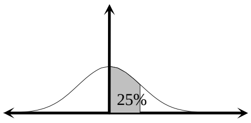
(v)



=



+



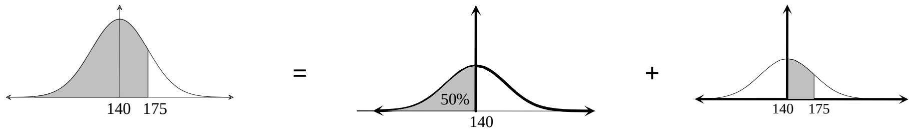
proportion $= 0.25$ (above mean)
$\Rightarrow z \approx 0.675$
between 0.67 and 0.68

$$x = \overline{x} + s \times z$$
$$= 100 + 16 \times 0.675$$
$$\approx 110.8$$

Shane would have an IQ of over 110.

STSD-F2 continued

3. (a)



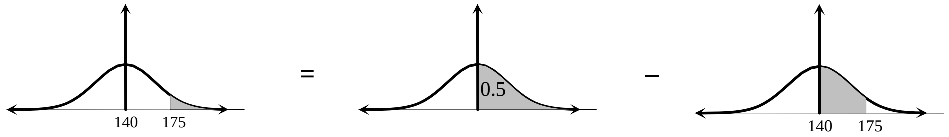
$z_{175} = \dfrac{175-140}{20} = 1.75$

$\Rightarrow 0.4599$

$\text{Area} < 175 = 0.5 + 0.4599$
$= 0.9599$

95.99% of the boys have a height of less than 175*cm*.

(b)



$z_{175} = \dfrac{175-140}{20} = 1.75$

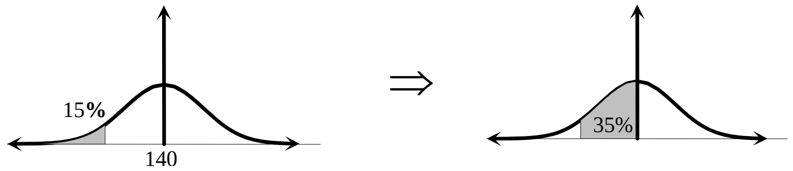
$\Rightarrow 0.4599$

$\text{Area} > 175 = 0.5 - 0.4599$
$= 0.0401$

$\textit{Boys} > 175 = 0.0401 \times 400$
≈ 16

Approximately 16 boys have a height of greater than 175*cm*.

(c)



$\text{proportion} = 0.35\ (\text{below mean})$

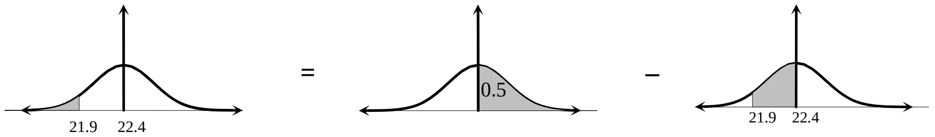
$\Rightarrow z \approx -1.035$

between 1.03 and 1.04

$x = \overline{x} - s \times z$
$= 140 - 20 \times 1.035$
≈ 119.3

The girls mean height is approximately 119.3*cm*.

4. (a)



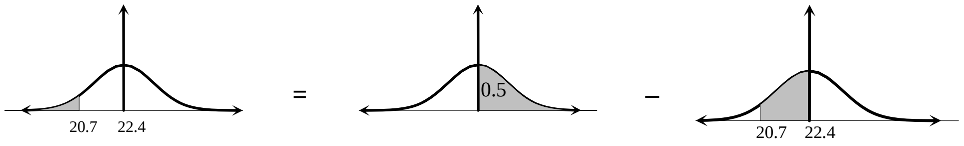
$z_{21.9} = \dfrac{21.9-22.4}{0.6} = -0.8\overline{3}$

$\Rightarrow 0.2967$

$\text{Area} < 21.9 = 0.5 - 0.2967$
$= 0.2033$

Charlie has approximately 20% probability of winning the race.

(b)



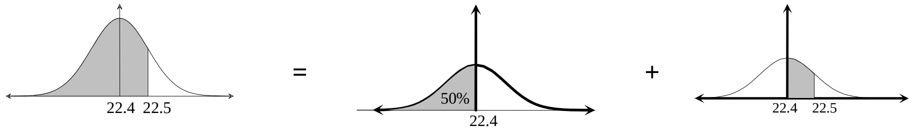
$z_{20.7} = \dfrac{20.7-22.4}{0.6} = -2.8\overline{3}$

$\Rightarrow 0.4977$

$\text{Area} < 20.7 = 0.5 - 0.4977$
$= 0.0023$

Charlie has approximately 0.23% probability of breaking the record.

(c)



$z_{22.5} = \dfrac{22.5-22.4}{0.6} \approx 0.17$

$\Rightarrow 0.0675$

$\text{Area} < 22.5 = 0.5 + 0.0675$
$= 0.5675$

$\textit{Prize} = 0.5675 \times 80 \times \$100$
$= \$4540$

Charlie expect $4540 in prize money.

STSD-G1 Chebyshev’s Theorem

1.

$$z_{330.5} = \frac{330.5 - 434}{69} = -1.5 \qquad z_{537.5} = \frac{537.5 - 434}{69} = 1.5$$

$$\begin{aligned} &\left(1 - \frac{1}{1.5^2}\right) \\ &= 1 - 0.\overline{4} \\ &= 55.\overline{5}\% \end{aligned}$$

At least 56% of the cattle would have weights between 330.5 *kg* and 537.5 *kg*.

2.

$$\begin{aligned} 89\% &= 1 - \frac{1}{k^2} \\ 0.89 &= 1 - \frac{1}{k^2} \\ \frac{1}{k^2} &= 1 - 0.89 = 0.11 \\ k^2 &= \frac{1}{0.11} = 9 \\ \therefore k &= \pm 3 \end{aligned}$$

$$\begin{aligned} &z = +3 \\ &x = \overline{x} + s \times z \\ &\quad = 74 + 4.5 \times 3 \\ &\quad = 87.5 \\ &z = -3 \\ &x = \overline{x} - s \times z \\ &\quad = 74 - 4.5 \times 3 \\ &\quad = 60.5 \end{aligned}$$

At least 89% of the pensioners would have ages between 60.5 years and 87.5 years.

3.

$$\begin{aligned} 75\% &= 1 - \frac{1}{k^2} \\ 0.75 &= 1 - \frac{1}{k^2} \\ \frac{1}{k^2} &= 1 - 0.75 = 0.25 \\ k^2 &= \frac{1}{0.25} = 4 \\ \therefore k &= \pm 2 \end{aligned}$$

$$\begin{array}{llcll} z = +2 & & & z = -2 & \\ x & = \overline{x} - s \times z & & x & = \overline{x} - s \times z \\ 1090 & = 994 + s \times 2 & \text{OR} & 898 & = 994 - s \times 2 \\ 1090 - 994 = 2s & & & 2s & = 994 - 898 \\ \frac{96}{2} & = s = 48 & & s & = \frac{96}{2} = 48 \end{array}$$

The standard deviation is 48*ml*.

4.

$$z_{17} = \frac{17 - 50}{11} = -3 \qquad z_{83} = \frac{83 - 50}{11} = 3$$

$$\begin{aligned} &\left(1 - \frac{1}{3^2}\right) \\ &= 1 - 0.\overline{1} \\ &\approx 89\% \end{aligned}$$

At least 89% of the test result would be between 17 and 83, therefore 100% – 89% = 11% of the results will be less than 17 and greater than 83 marks

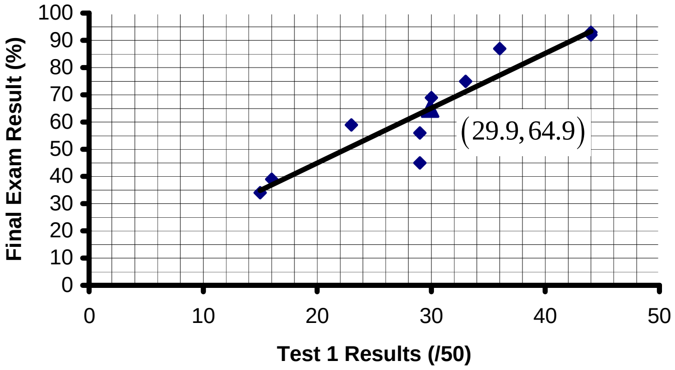
STSD-H1/H2 Scatterplots/Lines of Best Fit

1.

Student	A	B	C	D	E	F	G	H	I	J
Test 1 (/50)	33	36	15	29	16	29	44	30	44	23
Final Exam %	75	87	34	56	39	45	92	69	93	59

$mean_{test1} = \frac{299}{10} = 29.9$ $mean_{test1} = \frac{649}{10} = 64.9$

Scatterplot of Final Exam against Test 1

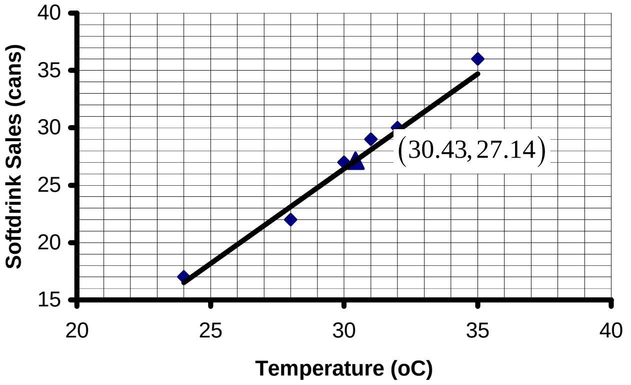


2.

Day	Mon	Tue	Wed	Thur	Fri	Sat	Sun
Temperature (°C)	24	28	30	33	32	35	31
Softdrink Sales (cans)	17	22	27	29	30	36	29

$mean_{temp} = \frac{213}{7} \approx 30.43$ $mean_{sales} = \frac{190}{7} = 27.14$

Scatterplot of Drink Sales against Temperature

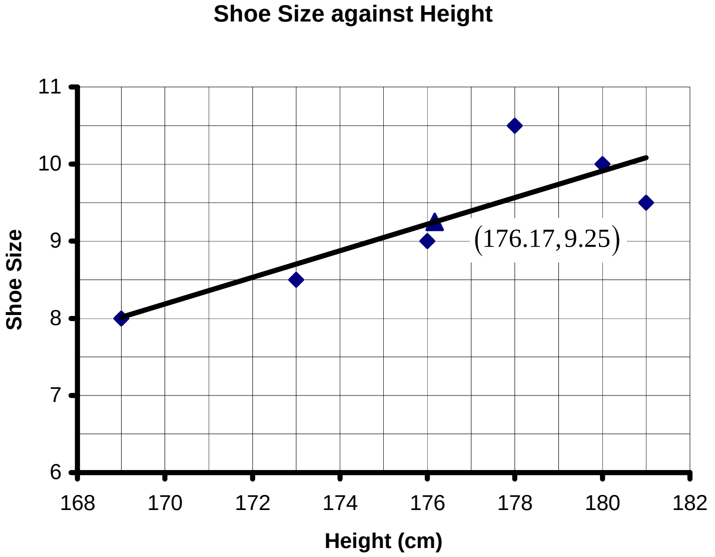


STSD-H1/H2 continued

3.

Name	Ann	Lee	May	Jan	Tom	Wes
Height (*cm*)	176	181	173	169	178	180
Shoe Size	9	9.5	8.5	8	10.5	10

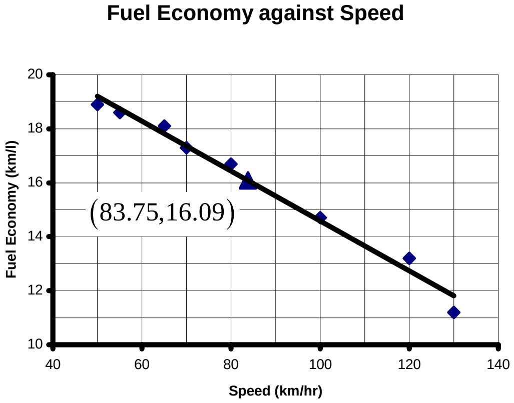
$$mean_{height} = \frac{1057}{6} \approx 176.17 \qquad mean_{shoesize} = \frac{55.5}{6} = 9.25$$



4.

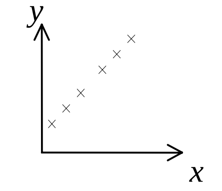
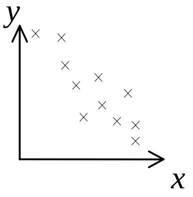
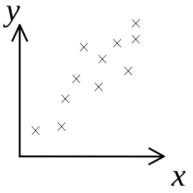
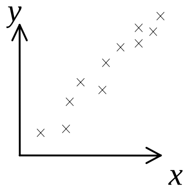
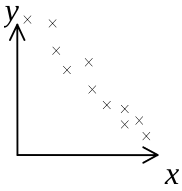
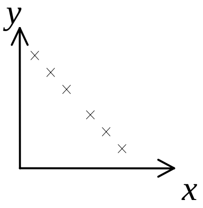
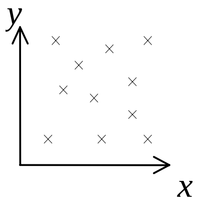
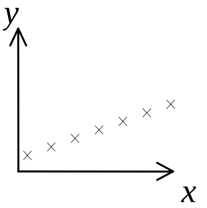
Speed (*km/hr*)	50	55	65	70	80	100	120	130
Fuel Economy (*km/l*)	18.9	18.6	18.1	17.3	16.7	14.7	13.2	11.2

$$mean_{speed} = \frac{670}{8} \approx 83.75 \qquad mean_{economy} = \frac{128.7}{8} = 16.09$$



STSD-H3 Correlation

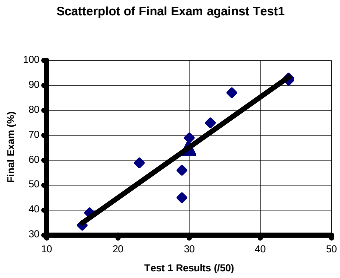
1.

 **A – perfect positive**	 **B- negative**	 **C - positive**	 **D - positive**
 **E - negative**	 **F – perfect negative**	 **G – no correlation**	 **H – perfect positive**

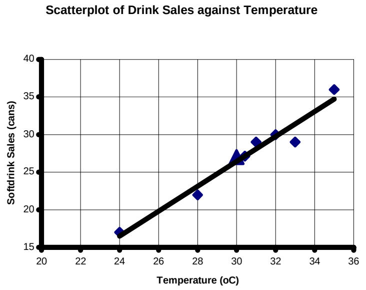
2.

(i)	Height at 4years, height at 16 years.	**C**
(ii)	Age of a used car, price of a used car.	**E**
(iii)	Temperature at 6am, temperature at 3pm.	**C**
(iv)	Shoe size of mother, number of children in family.	**G**
(v)	Average exam result, class size.	**E**

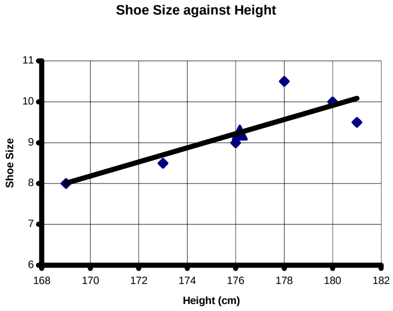
3. 1. **positive correlation** – the better a student did in Test 1 the better they did for the final examination



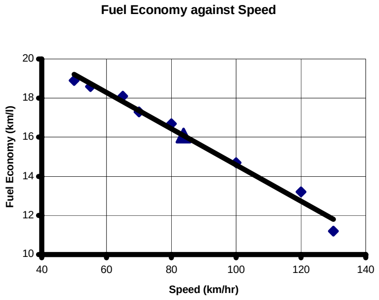
2. **positive correlation** – the hotter the temperature the more softdrinks sold.



3. **positive correlation** – the taller a person the larger their shoe size.



4. **negative correlation** – the faster a car the less economical it is.



STSD-I1 Correlation Coefficient

1.

Temperature, X	Hot Drink Sales Y	X^2	Y^2	XY
9	45	81	2025	405
13	34	169	1156	442
14	32	196	1024	448
11	42	121	1764	462
8	51	64	2601	408
6	60	36	3600	360
5	64	25	4096	320
$\Sigma X = 66$	$\Sigma Y = 328$	$\Sigma X^2 = 692$	$\Sigma Y^2 = 16266$	$\Sigma XY = 2845$

$$r = \frac{7\times 2845 - 66\times 328}{\sqrt{\left(7\times 692 - 66^2\right)\left(7\times 16266 - 328^2\right)}}$$
$$\approx -0.9901$$

There is a strong negative correlation, -0.9901, between the sales of hot drinks and the temperature. $r^2 = 0.9803$

2.

Speed, X	Economy, Y	X^2	Y^2	XY
50	18.9	2500	357.21	945
55	18.6	3025	345.96	1023
65	18.1	4225	327.61	1176.5
70	17.3	4900	299.29	1211
80	16.7	6400	278.89	1336
100	14.7	10000	216.09	1470
120	13.2	14400	174.24	1584
130	11.2	16900	125.44	1456
$\Sigma X = 670$	$\Sigma Y = 128.7$	$\Sigma X^2 = 62350$	$\Sigma Y^2 = 2124.73$	$\Sigma XY = 10201.5$

$$r = \frac{8\times 10201.5 - 670\times 128.7}{\sqrt{\left(8\times 62350 - 670^2\right)\left(8\times 2124.73 - 128.7^2\right)}}$$
$$\approx -0.99195$$

There is a strong negative correlation, –0.992, between the speed and the fuel economy. $r^2 = 0.984$

3.

$$r = \frac{10\times 21237 - 299\times 649}{\sqrt{\left(10\times 9849 - 299^2\right)\left(10\times 46387 - 649^2\right)}}$$
$$\approx 0.9302$$

There is a positive correlation, 0.93, between the result in test 1 and the result in final exam. $r^2 = 0.865$

STSD-I2 Regression Equation

1. (a)

Temperature, X	Hot Drink Sales, Y	X^2	XY
9	45	81	405
13	34	169	442
14	32	196	448
11	42	121	462
8	51	64	408
6	60	36	360
5	64	25	320
$\Sigma X = 66$	$\Sigma Y = 328$	$\Sigma X^2 = 692$	$\Sigma XY = 2845$

$$\overline{X} = \frac{\Sigma X}{n} = \frac{66}{7} \approx 9.43$$

$$\overline{Y} = \frac{\Sigma Y}{n} = \frac{328}{7} \approx 46.86$$

$$\beta = \frac{\Sigma XY - n\overline{X}\,\overline{Y}}{\Sigma X^2 - n\overline{X}^2} = \frac{2845 - 7\times 9.43\times 46.86}{692 - 7\times 9.43^2} \approx -3.57$$

$$\alpha = \overline{Y} - \beta\overline{X} = 46.86 - (-3.57)\times 9.43 \approx 80.53$$

The regression equation is $y = -3.57x + 80.53$

(i) $x = 10 \quad \Rightarrow \quad y = -3.57\times 10 + 80.53 \approx 45$

It would be expected that 45 hot drinks would be sold at 10°C.

(ii) **Can not be predicted as input is outside range of recorded temperatures.**

(b)

Speed, X	Economy, Y	X^2	XY
50	18.9	2500	945
55	18.6	3025	1023
65	18.1	4225	1176.5
70	17.3	4900	1211
80	16.7	6400	1336
100	14.7	10000	1470
120	13.2	14400	1584
130	11.2	16900	1456
$\Sigma X = 670$	$\Sigma Y = 128.7$	$\Sigma X^2 = 62350$	$\Sigma XY = 10201.5$

$$\overline{X} = \frac{\Sigma X}{n} = \frac{670}{8} = 83.75$$

$$\overline{Y} = \frac{\Sigma Y}{n} = \frac{128.7}{8} \approx 16.09$$

$$\beta = \frac{\Sigma XY - n\overline{X}\,\overline{Y}}{\Sigma X^2 - n\overline{X}^2} = \frac{10201.5 - 8\times 83.75\times 16.09}{62350 - 8\times 83.75^2} \approx -0.093$$

$$\alpha = \overline{Y} - \beta\overline{X} = 16.09 - (-0.093)\times 83.75 \approx 23.88$$

The regression equation is $y = -0.093x + 23.88$**.**

(i) $x = 75 \quad \Rightarrow \quad y = -0.093\times 75 + 23.88 \approx 16.91$

It would be expected that at 75*km/hr* the economy would be approx. 16.9*km/l*.

(ii) **Regression equation can not be used to predict speed from economy.**

2.

$$\overline{X} = \frac{\Sigma X}{n} = \frac{299}{10} = 29.9 \qquad \overline{Y} = \frac{\Sigma Y}{n} = \frac{649}{10} = 64.9$$

$$\beta = \frac{\Sigma XY - n\overline{X}\,\overline{Y}}{\Sigma X^2 - n\overline{X}^2} = \frac{21237 - 10\times 29.9\times 64.9}{9849 - 10\times 29.9^2} \approx 2.016$$

$$\alpha = \overline{Y} - \beta\overline{X} = 64.9 - (2.016)\times 29.9 \approx 4.622$$

The regression equation is $y = 2.016x + 4.622$**.**

STSD-K Review Exercise

1. (a) **Store Sales for a week**

Sales *(x)*	***x*^{2}**
552	304704
698	487204
547	299209
720	518400
645	416025
451	203401
$\Sigma x = 3613$	$\Sigma x^2 = 2228943$

$$\bar{x} = \frac{\sum x}{n} = \frac{3613}{6} \approx \$602.17$$

$$CV = \frac{s}{\bar{x}} \times 100\% = \frac{103.26}{602.17} \times 100\% \approx 17.15\%$$

$$s = \sqrt{\frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}} = \sqrt{\frac{2228943 - \frac{3613^2}{6}}{6-1}} \approx 103.26$$

(i) **the mean sales is approximately $602.17**
(ii) **the standard deviation of the sales is approximately $103.26**
(iii) **the coefficient of variation is approximately 17.15%**

(b) **Student Mark in a 5 Mark Test**

Mark (x)	Frequency, (f)	fx	x^2	fx^2
0	1	0	0	0
1	2	2	1	2
2	4	8	4	16
3	10	30	9	90
4	8	32	16	128
5	5	25	25	125
	$\Sigma f = 30$	$\Sigma fx = 97$		$\Sigma fx^2 = 361$

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{97}{30} \approx 3.23$$

$$s = \sqrt{\frac{\Sigma fx^2 - \frac{(\Sigma fx)^2}{\Sigma f}}{\Sigma f - 1}} = \sqrt{\frac{361 - \frac{97^2}{30}}{30-1}} \approx 1.28$$

$$CV = \frac{s}{\bar{x}} \times 100\% = \frac{1.28}{3.23} \times 100\% \approx 39.6\%$$

(i) **the mean mark is approximately 3.23**
(ii) **the standard deviation of the marks is approximately 1.28**
(iii) **the coefficient of variation is approximately 39.6%**

(c) **Daily Rainfall in millimetres**

Rainfall	x	f	fx	x^2	fx^2
0 – 4	2	2	4	4	8
5 – 9	7	8	56	49	392
10 – 14	12	4	48	144	576
15 – 19	17	3	51	289	867
20 – 24	22	4	88	484	1936
		$\Sigma f = 21$	$\Sigma fx = 247$		$\Sigma fx^2 = 3779$

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{247}{21} \approx 11.76$$

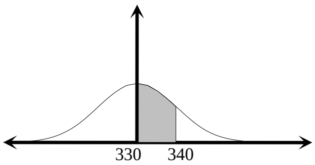
$$s = \sqrt{\frac{\Sigma fx^2 - \frac{(\Sigma fx)^2}{\Sigma f}}{\Sigma f - 1}} = \sqrt{\frac{3779 - \frac{247^2}{21}}{21-1}} \approx 6.61$$

$$CV = \frac{s}{\bar{x}} \times 100\% = \frac{6.61}{11.76} \times 100\% \approx 56.2\%$$

(i) **the mean rainfall was approximately 11.76*mm***
(ii) **the standard deviation of the rainfall was approximately 6.61*mm***
(iii) **the coefficient of variation is approximately 56.2%**

STSD-K continued

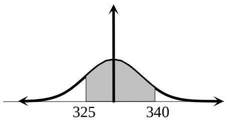
2. (a)



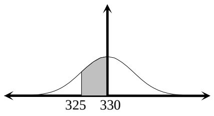
$$z_{340} = \frac{340-330}{8} = 1.25$$
$$\Rightarrow \ 0.3944$$

A proportion of 0.3944 of the cans between 330 *ml* and 340*ml* of softdrink.

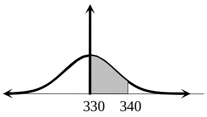
(b)



=



+



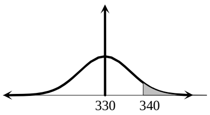
$$z_{325} = \frac{325-330}{8} = -0.625$$
$$\Rightarrow 0.2340$$
$$(\textit{between } 0.62 \ \& \ 0.63)$$

$$z_{340} = \frac{340-330}{8} = 1.25$$
$$\Rightarrow \ 0.3944$$

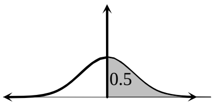
Area between 325*ml* & 340*ml*
$= 0.2340 + 0.3944$
$= 0.6284$

62.84% of the cans between 325 *ml* and 340*ml* of softdrink.

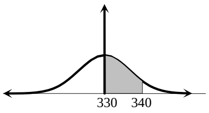
(c)



=



−

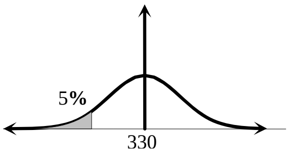


$$z_{340} = \frac{340-330}{8} = 1.25$$
$$\Rightarrow \ 0.3944$$

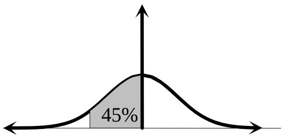
Area > 340 = 0.5 − 0.3944
= 0.1056

10.56% of the cans will overflow.

(d)



$\Rightarrow$



proportion = 0.45 (below mean)
$\Rightarrow z \approx -1.645$
between 1.64 and 1.65

$x = \overline{x} - s \times z$
$= 330 - 8 \times 1.645$
≈ 316.84

The smallest amount of softdrink that would be accepted would be 316.84*ml*.

3. (a)

$$75\% = 1 - \frac{1}{k^2}$$
$$0.75 = 1 - \frac{1}{k^2}$$
$$\frac{1}{k^2} = 1 - 0.75 = 0.25$$
$$k^2 = \frac{1}{0.25} = 4 \therefore k = \pm 2$$

$z = +2$
$x = \overline{x} - s \times z$
$= 76 + 2 \times 28.8$
$= 133.6$

$z = -2$
$x = \overline{x} - s \times z$
$= 76 - 28.8 \times 2$
$= 18.4$

At least 75% of values would lie between 18.4 and 133.6.

(b)

$$z_{582} = \frac{582-827}{98} = -2.5$$

$$z_{1072} = \frac{1072-827}{98} = 2.5$$

$$\left(1 - \frac{1}{2.5^2}\right)$$
$$= 1 - 0.16$$
$$= 84\%$$

At least 84% of the values should lie been 582 and 1072.

STSD-K continued

3. (c)

$$89\% = 1 - \frac{1}{k^2}$$

$$0.89 = 1 - \frac{1}{k^2}$$

$$\frac{1}{k^2} = 1 - 0.89 = 0.11$$

$$k^2 = \frac{1}{0.11} = 9 \therefore k = \pm 3$$

$z = +3$

$$\begin{aligned} x &= \overline{x} - s \times z \\ 600 &= 468 + s \times 3 \\ 600 - 468 &= 2s \\ \frac{132}{3} &= s = 44 \end{aligned}$$

OR

$z = -3$

$$\begin{aligned} x &= \overline{x} - s \times z \\ 336 &= 468 - s \times 3 \\ 3s &= 468 - 336 \\ s &= \frac{132}{3} = 44 \end{aligned}$$

The standard deviation is 44.

4. (a)

(i)	$r = 0.6$	**C**	(iv)	$r = 0.9$	**D**
(ii)	$r = 0$	**G**	(v)	$r = -1$	**F**
(iii)	$r = -0.9$	**E**			

(b) (i) Days on a good diet, weight **B**

(ii) temperature outside, temperature in a non-airconditioned car **D**

(iii) hand span, height **C**

(iv) rainfall, level of water in river **C**

(v) length of finger nails, intelligence **G**

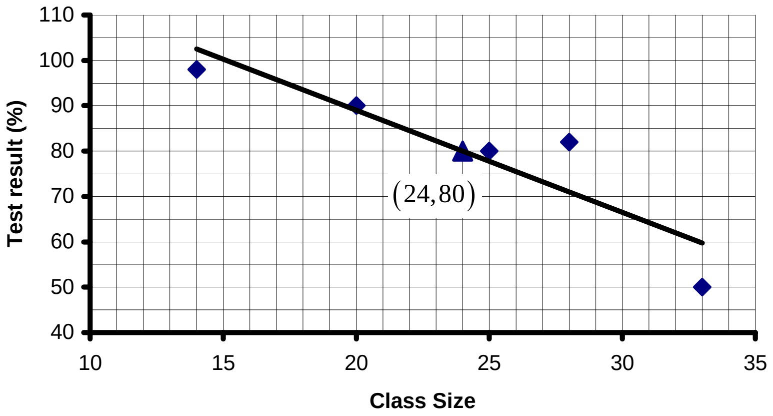
5.

School	Class Size, X	Test Result, Y	X^2	Y^2	XY
A	28	82	784	6724	2296
B	33	50	1089	2500	1650
C	25	80	625	6400	2000
D	14	98	196	9604	1372
E	20	90	400	8100	1800
	$\Sigma X = 120$	$\Sigma Y = 400$	$\Sigma X^2 = 3094$	$\Sigma Y^2 = 33328$	$\Sigma XY = 9118$

$$\overline{X} = \frac{120}{5} = 24 \qquad \overline{Y} = \frac{400}{5} = 80$$

(a)/(b)

Scatterplot of Test Result against Class size



STSD-K **continued**

5. (c) $$r = \frac{5\times 9118 - 120\times 400}{\sqrt{\left(5\times 3094 - 120^2\right)\left(5\times 33328 - 400^2\right)}}$$
$$\approx -0.9042$$

(d) **There is a negative correlation between the class size and test result.** The smaller the class the better the test result. $r^2 = 0.8175$

(e) $$\beta = \frac{\Sigma XY - n\overline{X}\,\overline{Y}}{\Sigma X^2 - n\overline{X}^2} = \frac{9118 - 5\times 24\times 80}{3094 - 5\times 24^2} \approx -2.252$$

$$\alpha = \overline{Y} - \beta\overline{X} = 80 - \left(-2.252\right)\times 24 \approx 134.05$$

The regression equation is $y = -2.252x + 134.05$

(f) (i) $x = 30$
$$y = -2.252\times 30 + 134.05 \approx 66.5$$

It would be expected that the test result would be 66.5% when the class size was 30 students.

(ii) **Can not be predicted as input is outside range of recorded class sizes.**

(iii) **Regression equation can not be used to predict class size from test result.**