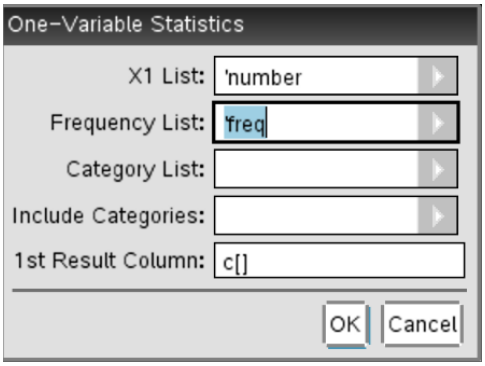
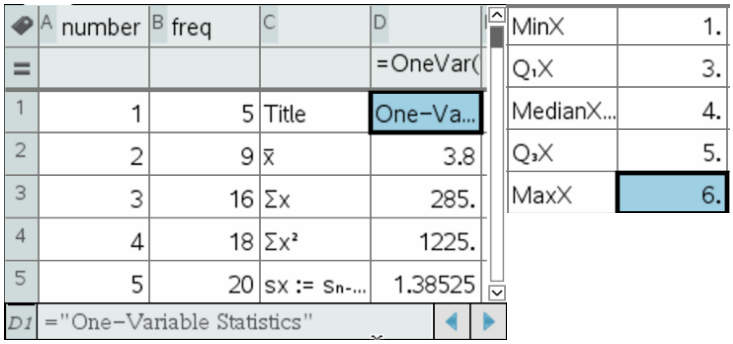
Statistics
Worked solutions for practice problems

1. Given the following frequency distribution, find
(a) the median;
(b) the mean.

Number (x)	**1**	**2**	**3**	**4**	**5**	**6**
Frequency (f)	**5**	**9**	**16**	**18**	**20**	**7**

(a) Use a calculator, one-variable statistics.





The median is 4.

(b) The mean is 3.8.
You might have to do these without a calculator. The reason that wasn't required here is that the total frequency is 75, and dividing 285 by 75 is not a good use of your time on a no-calculator question.

2. From January to September, the mean number of car accidents per month was 630. From October to December, the mean was 810 accidents per month.
What was the mean number of car accidents per month for the whole year?
Nine months at 630 each month is $9 \cdot 630 = 5670$; 3 months at 810 each month is 2430. That's a total of 8100 accidents in 12 months.
$\frac{8100}{12} = 675$ accidents per month

3. At a conference of 100 mathematicians there are 72 men and 28 women. The men have a mean height of 1.79 m and the women have a mean height of 1.62 m. Find the mean height of the 100 mathematicians.
This problem is really just like the last one.
$\frac{1.79 \cdot 72 + 1.62 \cdot 28}{100} = 1.7424$ m exactly, or 1.74 m to 3 s.f.

4. The table shows the scores of competitors in a competition.

Score	10	20	30	40	50
Number of competitors with this score	1	2	5	k	3

The mean score is 34. Find the value of k.

$$\frac{10 \cdot 1 + 20 \cdot 2 + 30 \cdot 5 + 40 \cdot k + 50 \cdot 3}{1 + 2 + 5 + k + 3} = 34$$

$$\frac{350 + 40k}{11 + k} = 34$$

$$350 + 40k = 374 + 34k$$

$$6k = 24$$

$$k = 4$$

5. The number of hours of sleep of 21 students are shown in the frequency table below.

Hours of sleep	Number of students
4	2
5	5
6	4
7	3
8	4
10	2
12	1

Find

(a) the median;

(b) the lower quartile;

(c) the interquartile range.

(a) There are 21 students. That means 10 on either side of the median and 1 in the middle. The middle term is the 11th one. The 11th student got 6 hours of sleep, so the median is 6.

(b) There are 10 students in each half, so between the fifth and sixth students will be the lower quartile. That gives 5 hours of sleep.

(c) The upper quartile would be 8, so the interquartile range is 8 – 5 = 3.

6. Three positive integers a, b, and c, where $a < b < c$, are such that their median is 11, their mean is 9 and their range is 10. Find the value of a.

Since b is known to be the middle number, $b = 11$.

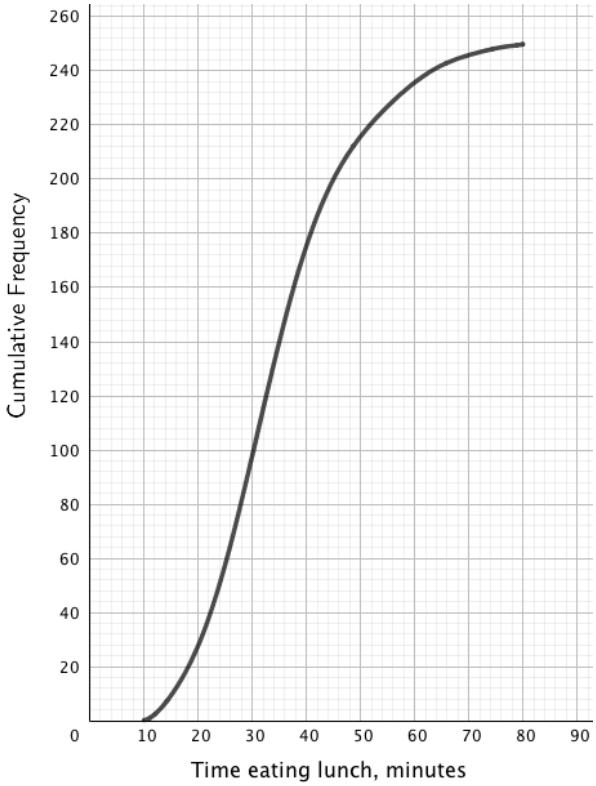
The mean gives $\frac{a + 11 + c}{3} = 9$, so $a + c = 16$.

The range gives $c - a = 10$. Adding those two equations eliminates the a and gives $2c = 26$, so $c = 13$. Then $a = 3$.

7. The cumulative frequency curve at right indicates the amount of time 250 students spend eating lunch.

(a) Estimate the number of students who spend between 20 and 40 minutes eating lunch.

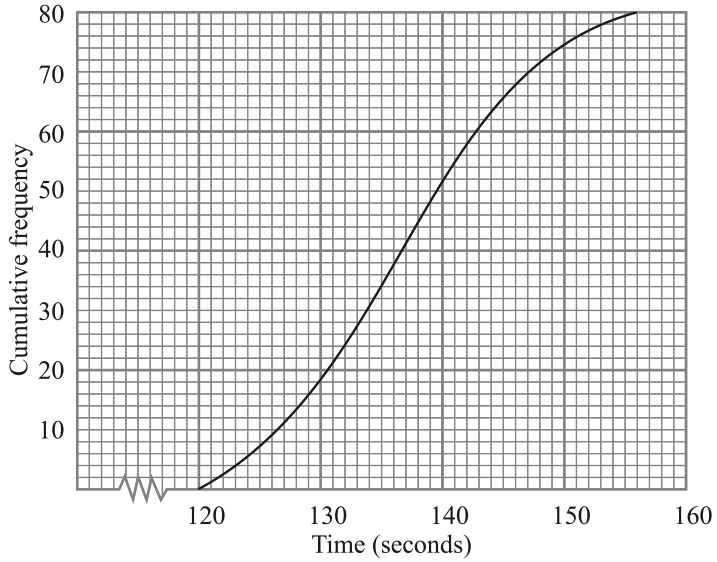
(b) If 20% of the students spend more than x minutes eating lunch, estimate the value of x.



(a) From the graph, 20 minutes corresponds to two squares above 20 on the vertical axis. On that axis, 5 squares are 20 students, so each square is 4. Therefore 20 minutes corresponds to 28 students. Similarly, 40 minutes corresponds to about 174 students. Therefore 174 – 28 = 146 students spend between 20 and 40 minutes eating lunch.

(b) 20% of 250 = $0.2 \cdot 250 = 50$
The top 50 students can be seen at a y-coordinate of 200. That goes with 45 minutes. So $x = 45$ minutes.

8. The 80 applicants for a Sports Science course were required to run 800 metres and their times were recorded. The results were used to produce the following cumulative frequency graph.



Estimate

(a) the median;

(b) the interquartile range.

(a) Halfway through the 80 applicants is 40. The 40th person corresponds to 136.5 seconds, the median time.

(b) The lower quartile, corresponding to 20 applicants, is 130.5 seconds. The upper quartile, corresponding to 60 applicants, is 142.5 s.
IQR = 142.5 – 130.5 = 12 seconds

9. One thousand candidates sit an examination. The distribution of marks is shown in the following grouped frequency table.

Marks	1–10	11–20	21–30	31–40	41–50	51–60	61–70	71–80	81–90	91–100
Number of candidates	15	50	100	170	260	220	90	45	30	20

(a) Copy and complete the following table, which presents the above data as a cumulative frequency distribution.

Marks	≤ 10	≤ 20	≤ 30	≤ 40	≤ 50	≤ 60	≤ 70	≤ 80	≤ 90	≤ 100
Number of candidates	15	65					905			

(b) Draw a cumulative frequency graph of the distribution, using a scale of 1 cm for 100 candidates on the vertical axis and 1 cm for 10 marks on the horizontal axis.

(c) Use your graph to answer parts (i)–(iii) below,

(i) Find an estimate for the median score.

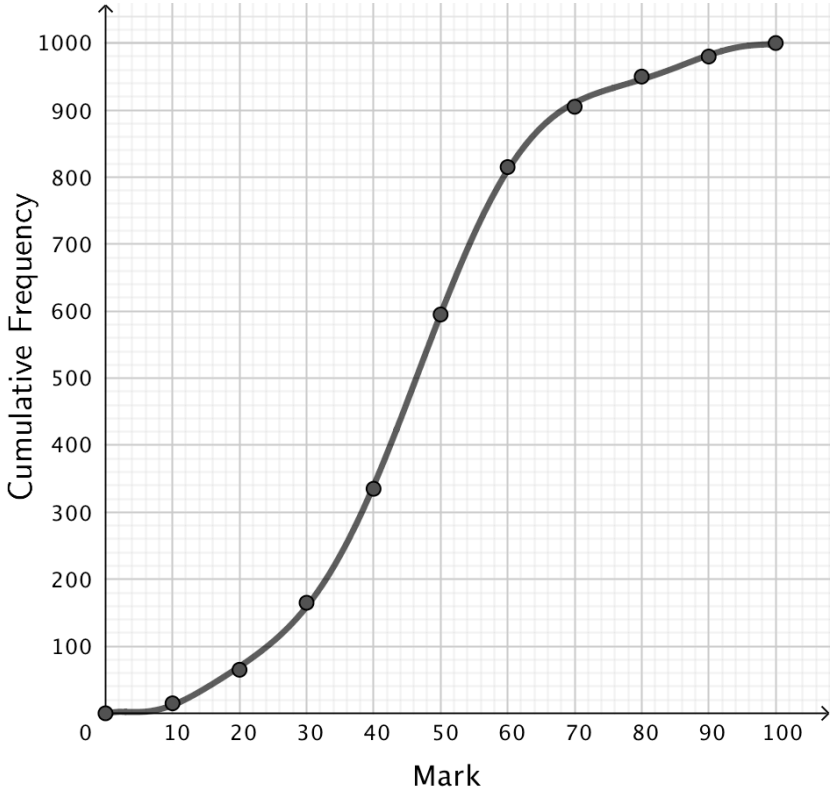
(ii) Candidates who scored less than 35 were required to retake the examination. How many candidates had to retake?

(iii) The highest-scoring 15% of candidates were awarded a distinction. Find the mark above which a distinction was awarded.

(a) The "copy" instruction is because this was from Section B, where you write in an answer booklet.

Mark	≤ 10	≤ 20	≤ 30	≤ 40	≤ 50	≤ 60	≤ 70	≤ 80	≤ 90	≤ 100
Number of candidates	15	65	**165**	**335**	**595**	**815**	905	**950**	**980**	**1000**

(b)



(c) (i) Half of 1000 candidates is 500; the 500th candidate has a score of approximately 46.

(ii) A score of 35 marks corresponds to about 240 candidates who had to retake.

(iii) The top 15% would be the top 150 people, which gives a cumulative frequency of 850. This goes with a score of about 63.

Note that your answers to part (c) would be based on your (reasonable) graph in part (b).

10. In a suburb of a large city, 100 houses were sold in a three-month period. The following cumulative frequency table shows the distribution of selling prices (in thousands of dollars).

Selling price P ($1000)	$P \leq 100$	$P \leq 200$	$P \leq 300$	$P \leq 400$	$P \leq 500$
Total number of houses	12	58	87	94	100

(a) Represent this information on a cumulative frequency curve, using a scale of 1 cm to represent $50000 on the horizontal axis and 1 cm to represent 5 houses on the vertical axis.

(b) Use your curve to find the interquartile range.

The information above is represented in the following frequency distribution.

Selling price P ($1000)	$0 < P \leq 100$	$100 < P \leq 200$	$200 < P \leq 300$	$300 < P \leq 400$	$400 < P \leq 500$
Number of houses	12	46	29	a	b

(c) Find the value of *a* and of *b*.

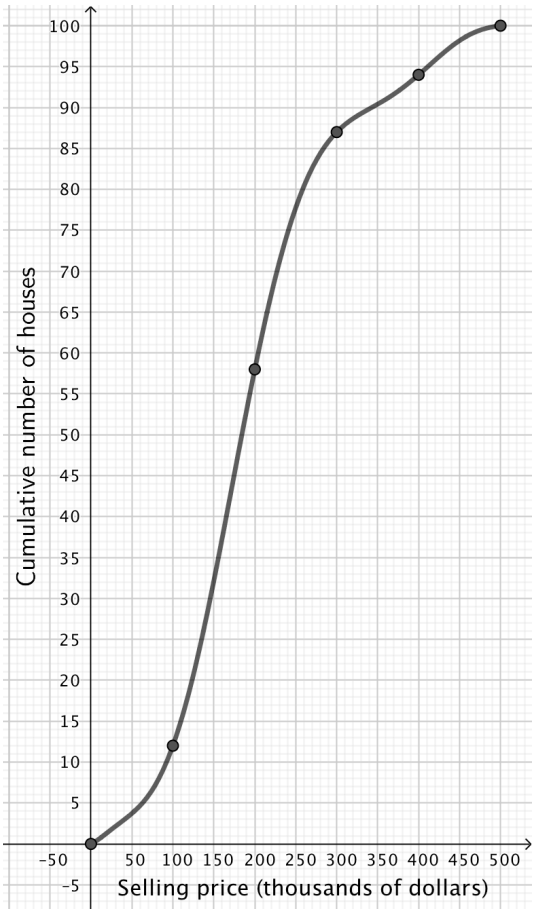
(d) Use mid-interval values to calculate an estimate for the mean selling price.

(e) Houses which sell for more than $350000 are described as De Luxe.

(i) Use your graph to estimate the number of De Luxe houses sold. Give your answer to the nearest integer.

(ii) Two De Luxe houses are selected at random. Find the probability that both have a selling price of more than $400000.

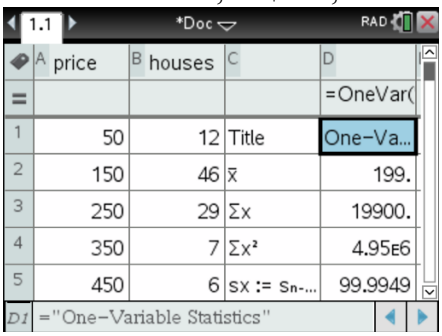
(a)



(b) $Q_1 = 135$, $Q_3 = 240$
IQR = 240 – 135 = 105, or \$105,000.

(c) $a = 94 - 87 = 7$
$b = 100 - 94 = 6$

(d) The mean is 199, or \$199,000.



(e) (i) \$350,000 corresponds with about 91 houses. There are 9 houses above this cutoff, so 9 De Luxe houses.

(ii) 6 of the 9 houses have a selling price in the \$400-500 thousand dollar range, so the probability both of them are here is $\frac{6}{9} \cdot \frac{5}{8} = \frac{30}{72} = \frac{5}{12}$. The fraction doesn't have be reduced.

11. The speeds in km h^{-1} of cars passing a point on a highway are recorded in the following table.

Speed v	Number of cars
$v \leq 60$	0
$60 < v \leq 70$	7
$70 < v \leq 80$	25
$80 < v \leq 90$	63
$90 < v \leq 100$	70
$100 < v \leq 110$	71
$110 < v \leq 120$	39
$120 < v \leq 130$	20
$130 < v \leq 140$	5
$v > 140$	0

(a) Calculate an estimate of the mean speed of the cars.

(b) The following table gives some of the cumulative frequencies for the information above.

Speed v	Cumulative frequency
$v \leq 60$	0
$v \leq 70$	7
$v \leq 80$	32
$v \leq 90$	95
$v \leq 100$	a
$v \leq 110$	236
$v \leq 120$	b
$v \leq 130$	295
$v \leq 140$	300

(i) Write down the values of a and b.

(ii) On graph paper, construct a cumulative frequency *curve* to represent this information. Use a scale of 1 cm for 10 km h^{-1} on the horizontal axis and a scale of 1 cm for 20 cars on the vertical axis.

(c) Use your graph to determine

(i) the percentage of cars travelling at a speed in excess of 105 km h^{-1};

(ii) the speed which is exceeded by 15% of the cars.

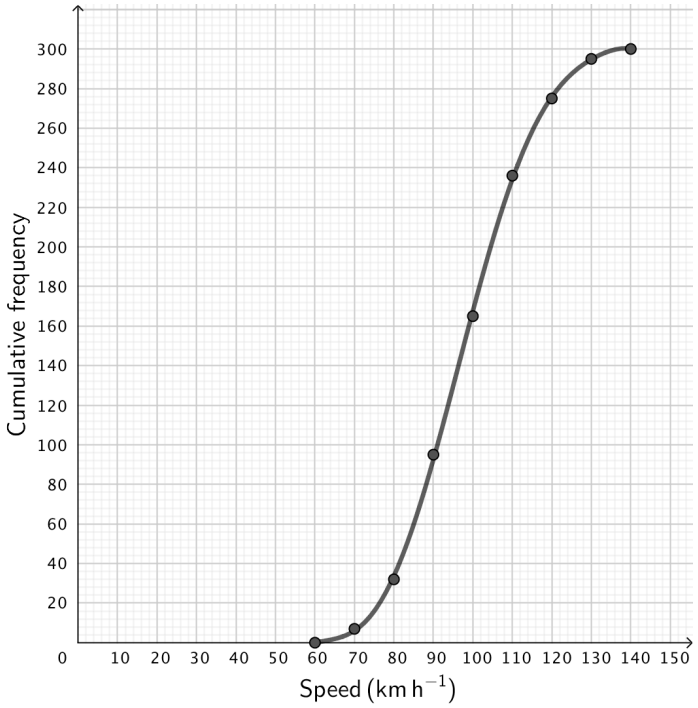
(a) Use the midpoint values of each interval. The mean is 98.2 km h^{-1}

A speed	B cars	C	D
			=OneVar(
65	7	Title	One-Va...
75	25	$\bar{x}$	98.1667
85	63	Σx	29450.
95	70	Σx^2	2.9593E6
105	71	sx := $s_{n-...}$	15.1129

(b) (i) $a = 95 + 70 = 165$

$b = 236 + 39 = 275$

(ii)



Ordinarily, I would do a broken scale on the horizontal axis to skip that empty space between 0 and 60, but that's way too much trouble in GeoGebra.

(c) (i) 105 km h^{-1} corresponds to 200 cars. That gives 100 cars above that speed out of the 300 total; this is 33.3%.

(ii) 15% of 300 is $0.15 \cdot 300 = 45$ cars. On the vertical axis, that corresponds to 255 cars, or about 114 km h^{-1}.

12. The random variable X is distributed normally with mean 30 and standard deviation 2. Find $P(27 \le X \le 34)$.

This is just a straightforward use of the calculator.

$X \sim N\left(30, 2^2\right)$

normCdf(27,34,30,2) 0.910443

$P(27 \le X \le 34) \approx 0.910$

13. In a country called Tallopia, the height of adults is normally distributed with a mean of 187.5 cm and a standard deviation of 9.5 cm.

(a) What percentage of adults in Tallopia have a height greater than 197 cm?

(b) A standard doorway in Tallopia is designed so that 99% of adults have a space of at least 17 cm over their heads when going through a doorway. Find the height of a standard doorway in Tallopia. Give your answer to the nearest cm.

(a) Let X be the height of an adult in Tallopia. Then $X \sim N(187.5, 9.5^2)$.
$P(X > 197) \approx 0.159$, or 15.9%.

normCdf(197,999999,187.5,9.5) 0.158655

(b) $P(X < k) = 0.99$ gives $k \approx 209.6$ cm, which is the height of the adult at the 99th percentile.

invNorm(0.99,187.5,9.5) 209.6

This still needs a clearance of 17 cm; $209.6 + 17 \approx 226.6$.
To the nearest centimeter, the standard doorway should be 227 cm tall.

14. It is claimed that the masses of a population of lions are normally distributed with a mean mass of 310 kg and a standard deviation of 30 kg.

(a) Calculate the probability that a lion selected at random will have a mass of 350 kg or more.

(b) The probability that the mass of a lion lies between *a* and *b* is 0.95, where *a* and *b* are symmetric about the mean. Find the value of *a* and of *b*.

(a) Let X be the mass of a lion. Then $X \sim N(310, 30^2)$.
$P(X \geq 350) \approx 0.0912$

normCdf(350,999999,310,30) 0.091211

(b) Since 1 – 0.95 = 0.05, half of this amount is to the left of a and half to the right of b.
$P(X < a) = 0.025$ gives $a \approx 251$ kg.

invNorm(0.025,310,30) 251.201

$P(X > b) = 0.025$, so $P(X < b) = 0.975$, and $b \approx 369$ kg.

invNorm(0.975,310,30) 368.799

15. The speeds of cars at a certain point on a straight road are normally distributed with mean μ and standard deviation σ. 15% of the cars travelled at speeds greater than 90 km h^{-1} and 12% of them at speeds less than 40 km h^{-1}. Find μ and σ.

Missing mean and standard deviation indicate a need for the z-score.
$X \sim N(\mu, \sigma^2)$
$P(X > 90) = 0.15$, which gives a z-score of 1.036, so $1.036 = \dfrac{90 - \mu}{\sigma}$.

invNorm(1−0.15,0,1) 1.03643

$P(X < 40) = 0.12$, which gives a z-score of –1.175, so $-1.175 = \dfrac{40 - \mu}{\sigma}$.

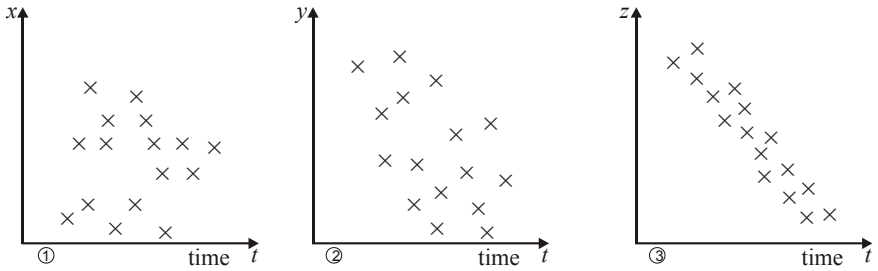
invNorm(0.12,0,1) -1.17499

Together, those give a system of equations in μ and σ. In order to use the system solver on a numeric TI-nspire, it's necessary to multiply the sigma terms off the denominator.

$$\text{linSolve}\left(\begin{cases} 1.0364333797684 \cdot s = 90 - m \\ -1.17498679146 \cdot s = 40 - m \end{cases}, \{m, s\}\right)$$
$$\{66.5663, 22.6099\}$$

Thus $\mu \approx 66.6$ km h^{-1} and $\sigma \approx 22.6$ km h^{-1}.

16. The sketches below represent scatter diagrams for the way in which variables *x*, *y* and *z* change over time, *t*, in a given chemical experiment. They are labelled ①, ② and ③.



(a) State which of the diagrams indicate that the pair of variables
(i) is not correlated;
(ii) shows strong linear correlation.

(b) A student is given a piece of paper with five numbers written on it. She is told that three of these numbers are the product moment correlation coefficients for the three pairs of variables shown above. The five numbers are

0.9, –0.85, –0.20, 0.04, 1.60

(i) For each sketch above state which of these five numbers is the most appropriate value for the correlation coefficient.
(ii) For the two remaining numbers, state why you reject them for this experiment.

(a) (i) Graph ① shows no correlation.

(ii) Graph ③ shows a strong correlation.

(b) (i) For sketch ①, we need the correlation coefficient closest to 0, which is 0.04.

Sketch ② shows a weak negative correlation, so that is –0.20.

Sketch ③ is a strong negative correlation, which would be –0.85.

(ii) A correlation of 0.9 would show an upward-sloping, tightly grouped set of points. None of the graphs looks like this. Correlation is always a number on the interval [–1, 1], so 1.60 could never be a correlation coefficient.

17. A shop keeper recorded daily sales s of ice cream along with the daily maximum temperature t °C. The results for one week are shown below.

t	29	31	34	23	19	20	27
s	104	92	112	48	56	72	66

(a) Write down the equation of the regression line for s on t.

(b) Use your equation to predict the ice cream sales on a day when the maximum temperature is 24°C. Give your answer correct to the nearest whole number.

(a) Use a calculator's regression plan for this.

A temp	B sales	C	D
			=LinRegM
31	92	RegEqn	m*x+b
34	112	m	3.56444
23	48	b	-14.6133
19	56	r²	0.682589
20	72	r	0.82619

$s = 3.56t - 14.6$

(b) $s(24) = 3.56(24) - 14.6 \approx 71$ orders of ice cream

18. The number of bottles of water sold at a railway station on each day is given in the following table.

Day	0	1	2	3	4	5	6	7	8	9	10	11	12
Temperature (T°)	21	20.7	20	19	18	17.3	17	17.3	18	19	20	20.7	21
Number of bottles sold (n)	150	141	126	125	98	101	93	99	116	121	119	134	141

(a) Write down

(i) the mean temperature;

(ii) the standard deviation of the temperatures.

(b) Write down the correlation coefficient, r, for the variables n and T.

(c) Comment on your value for r.

(d) The equation of the line of regression for n on T is $n = dT - 100$.

(i) Write down the value of d.

(ii) Estimate how many bottles of water will be sold when the temperature is 19.6°.

(e) On a day when the temperature was 36° Peter calculates that 314 bottles would be sold. Give one reason why his answer might be unreliable.

(a) Both of these come from one-variable statistics using only the temperature data.

A temp	B	C
		=OneVar(
20.7	$\bar{x}$	19.1538
20	Σx	249.
19	Σx^2	4796.56
18	$sx := s_{n-...}$	1.50699
17.3	$\sigma x := \sigma_{n...}$	1.44787

(i) mean: 19.2°

(ii) standard deviation: 1.45°

(b) Correlation coefficient comes from two variable statistics with both n and t.

A temp	B bottles	C	D
			=TwoVar(
19	121	Σy^2	192232.
20	119	$sy := s_{n-...}$	18.4182
20.7	134	$\sigma y := \sigma_{n...}$	17.6957
21	141	Σxy	30270.5
		r	0.942389

$r \approx 0.942$

(c) There is a strong positive linear correlation between temperature and number of bottles of water sold.

(d) (i) Using the linear regression command, $d \approx 11.5$.

(ii) $n = 11.5(19.6) - 100 \approx 125$ bottles of water

	=LinRegM
Title	Linear R...
RegEqn	m*x+b
m	11.5177
b	-100.301
r^2	0.888097

(e) None of the temperatures in the data collected is anywhere near 36°. Using the model for extrapolation in this way is not a good idea.

19. The following table shows the marks scored by seven students on two different mathematics tests.

Test 1 (x)	15	23	25	30	34	34	40
Test 2 (y)	20	26	27	32	35	37	35

Let L_1 be the regression line of x on y. The equation of the line L_1 can be written in the form $x = ay + b$.

(a) Find the value of a and the value of b.

Let L_2 be the regression line of y on x. The lines L_1 and L_2 pass through the same point with coordinates (p, q).

(b) Find the value of p and the value of q.

(a) Entering the data into a Lists and Spreadsheets page and finding a linear regression **using Test 2 (y) as the "X List" and Test 1 (x) as the "Y List"** gives $a \approx 1.29$ and $b \approx -10.4$.

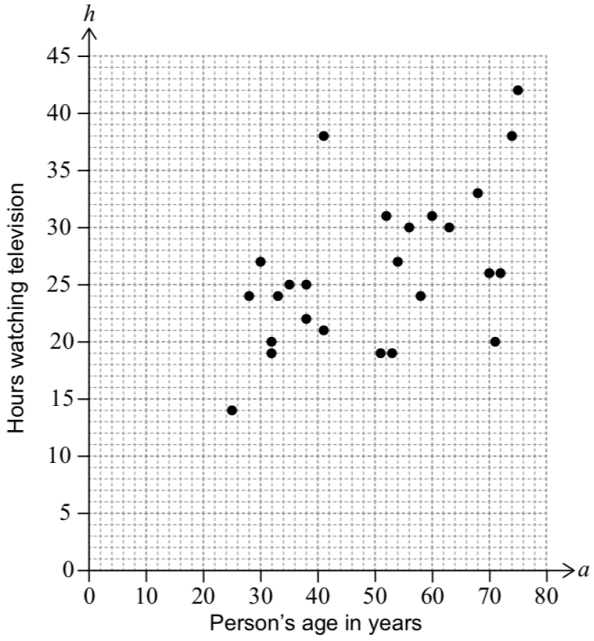
(b) The least-squares linear regression always passes through the point with coordinates at the mean of each of the variables. That means that $(\bar{x}, \bar{y})$ is on both lines.

$p = \bar{x} \approx 28.7$ and $q = \bar{y} \approx 30.3$

20. A survey was carried out to investigate the relationship between a person's age in years (a) and the number of hours they watch television per week (h). The scatter diagram represents the results of the survey.

The mean age of the people surveyed was 50.

For these results, the equation of the regression line h on a is $h = 0.22a + 15$.



(a) Find the mean number of hours that the people surveyed watch television per week.

(b) Draw the regression line on the scatter diagram.

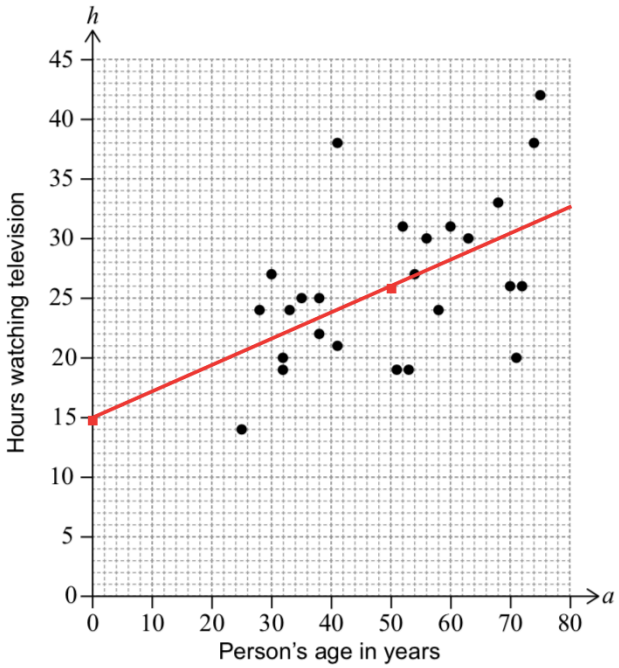
(c) By placing a tick (✓) in the correct box, determine which of the following statements is true:

The correlation between h and a is positive.	
The correlation between h and a is negative.	
There is no correlation between h and a.	

(d) Diogo is 18 years old. Give a reason why the regression line should not be used to estimate the number of hours Diogo watches television per week.

(a) The point $(\bar{x}, \bar{y})$ always lies on the least-squares regression line. Since we know that the mean x-value (called a here) is 50, the mean y-value is $\bar{h} = 0.22 \cdot 50 + 15 = 26$ hours.

(b) In general, the line should pass through (50, 26) and have about the same number of points above and below. Since we know the equation, we can also use the y-intercept of 15 to get a really accurate graph. The markscheme here requires that the graph you draw has the correct y-intercept and passes through the mean point.



(c)

The correlation between h and a is positive.	✓
The correlation between h and a is negative.	
There is no correlation between h and a.	

(d) The youngest person in the study is 22; Diogo is younger than that. Using an age of 18 would be extrapolation. (There are several ways you could express this idea to earn credit here.)