Workbook Answers
Level 1 Maths
Chance and Data

Section One

1. Definitions

a. i. The median is the midway point in a set of numbers, i.e. half the numbers are larger and half the numbers are smaller.

ii. The mean is the 'average' i.e. the result if you add up all the values in a dataset, then divide by the number of values.

iii. The lower quartile is the value that 25% (one quarter) of the numbers are smaller than, and 75% are greater than. The upper quartile is the value that 25% of the numbers are greater than.

iv. The interquartile range is the difference between the upper quartile and lower quartile: IQR = UQ – LQ.

v. Skew is the opposite of symmetry. Data is skewed if it looks like it has been 'stretched' out to one side.

vi. An outlier is a value that is much higher or lower than we would expect it to be, based on all the other values in a dataset.

vii. A seasonal trend is a pattern that repeats regularly over the course of the time series like every year, or every week.

viii. Sampling variability is the difference between samples that we get due to random chance, because each sample is slightly different from the overall population.

ix. A study is biased if there is something other than the truth and random chance affecting the answers, usually because the sample wasn't chosen at random.

2. Probability

a. $\frac{2}{5} = 40\%$

$2 \div 5 = 0.4$

$0.4 \times 100 = 40$

b. $78\% = 0.78$

$78 \div 100 = 0.78$

c. i. The probability that a randomly selected customer got strawberry ice cream with sprinkles is $\frac{14}{116} = 0.121$

There are 116 total customers, and out of those 116, 14 chose strawberry ice cream with sprinkles. When giving a probability as a decimal, it's common to round to 3 decimal places.

ii. The probability that a randomly selected customer got no toppings on their ice cream is $\frac{36}{116} = 0.31$

Looking at the 'totals' column, 36 are in the 'no toppings' row, out of 116 overall total customers.

iii. The probability that a randomly selected customer either got chocolate ice cream with sprinkles or vanilla ice cream with peanuts is: $\frac{29}{116} = 0.25$

The probability of getting chocolate ice cream with sprinkles is $\frac{22}{116}$
The probability of getting vanilla ice cream with peanuts is $\frac{7}{116}$

$\frac{22}{116} + \frac{7}{116} = \frac{29}{116}$

iv. The probability that a randomly selected customer did not get strawberry ice cream is $\frac{93}{116} = 0.802$

If a customer did not get strawberry ice cream, they must have gotten chocolate or vanilla:

The probability of getting vanilla ice cream is $\frac{44}{116}$
The probability of getting chocolate ice cream is $\frac{49}{116}$

$\frac{44}{116} + \frac{49}{116} = \frac{93}{116}$

v. The probability that a customer got chocolate ice cream, given that they chose peanuts as a topping, is $\frac{9}{49} = 0.184$.

9 customers got chocolate ice cream with peanuts.
49 customers in total got chocolate ice cream, and we know that our selected customer must be one of these 49 people.
Therefore our probability is $\frac{9}{49}$

3. Summary Statistics

a. i. Charlotte's mean score is 103.5

Add up all the numbers:
$120 + 125 + 89 + 97 + 137 + 140 + 152 + 42 + 101 + 60 + 75 = 1138$

There are 11 *numbers so divide by* 11:
$1138 \div 11 = 103.5$

ii. Charlotte's median score is 101

Write all the numbers in order:
$42\ 60\ 75\ 89\ 97\ 101\ 120\ 125\ 137\ 140\ 152$

The median is the middle number. There are 11 *numbers so pick the* 6^{th} *number*:
$42\ 60\ 75\ 89\ 97\ (101)\ 120\ 125\ 137\ 140\ 152$

iii. The lower quartile is 75 and the upper quartile is 137

To find the lower quartile, look at just the numbers below the median, and find the middle of these:
$42\ 60\ (75)\ 89\ 97$

To find the upper quartile, find the middle of the numbers above the median:
$120\ 125\ (137)\ 140\ 152$

iv. The range of her scores is 110

The range is the difference between the highest and lowest values.
$152 - 42 = 110$

v. Her mean score will go up.

205 is higher than Charlotte's previous mean score of 103.5.
Scoring better than the mean will bring her mean score up.

4. Reading Graphs

a. i. The median is approximately 9.

The median is represented by the middle bar in the box which is at roughly 9. Usually markers will accept anything within a sensible range (so if you said something like 9.5, you'd get the mark).

ii. The maximum is roughly 15.

The maximum is represented by the end of the upper whisker, which is at roughly 15.

iii. The interquartile range is about 5.

The IQR is the difference between the upper and lower quartiles, which are represented by the ends of the middle box on the graph. The box starts at around 7 and ends at around 12.
$12 - 7 = 5$

b. i. The most common age is 11 years.

Each dot represents one child, so the most common value is wherever the graph is highest. The peak is at 11 (with 5 children having this age).

ii. Yes – there is one unusually young child who is only 1 year old.

This point is visually distant from all the other points, so it is an outlier.
When you identify an outlier, you should state whether it is unusually high or unusually low.

iii. If you ignore the outlier, the graph looks pretty symmetrical.

Symmetry and skew can be a matter of personal opinion, but you should always justify your answer. You want to look at the 'big picture' and ignore if there are one or two unusual points.

c. i. The proportion of people who went to Europe for their last overseas holiday is roughly 0.15.

Look at just the 'Europe' bar and (using a ruler) see how it lines up with the y-axis. Again, any number in a sensible range will be marked correct, so saying 0.14 or 0.16 would be okay.

ii. The proportion of people who went to either North or South America for their last overseas holiday is 0.2.

d. i.

Your line of best fit may be slightly different, but should look a lot like this. Roughly half the points should be above the line and half should be below the line.

ii. The relationship between test 1 scores and test 2 scores is positive.

The graph mostly slopes upwards from left to right, so that people who got high test 1 scores mostly also got high test 2 scores. This is a positive relationship.

iii. The line is a pretty good fit for the data, except for a couple of outliers. The graph definitely seems to be linear, not curved.

When you are deciding whether a line of best fit is a good fit, you should explain whether you think a curve would be a better fit. In this case the data does not look curved.

iv.

This point is very far along the x-axis (with a test 1 score of roughly 90) but very low down the y-axis (with a test 2 score of roughly 20).

e. i. The most common first period class was Science.

The largest slice in the pie chart is the yellow slice, which contains 30% of the students. Looking at the key, this corresponds to Science.

ii. 10% of students had PE first period.

The key tells us that PE is represented by light blue. The light blue slice of the chart is labelled with 10%.

iii. 40 of students had English first period.

From the chart, 20% of students had English first period. To find 20% of 200, convert 20% to a decimal, then multiply by 200.
20 ÷ 100 = 0.20
0.20 × 200 = 40.

f. i. 25% of employees at company 2 either walk or bike to work.

Reading off the company 2 bar, the 'walking' section is labelled with 15%, and the 'Bike' section is labelled with 10%.
15% + 10% = 25%

ii. 80% of employees at company 1 do not bus to work.

Adding up all the sections for company 1 except the 'Bus' section, we get:
50% + 10% + 10% + 10% = 80%

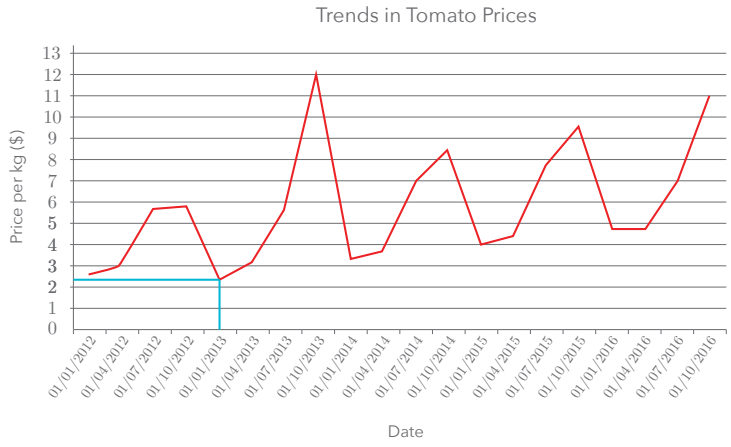
Alternatively, 20% of employees bus to work.
Therefore, 100% - 20% = 80% of employees do not bus to work.

iii. Possible differences include:

- None of the employees at company 1 answered 'other', while 5% of the employees at company 2 answered 'other'.
- Company 1 employees were much more likely to bus to work than company 2 employees, 20% vs 10%.
- Company 2 employees were more likely to take the train to work than company 1 employees, 15% vs 10%.

There are other possible differences you could state, make sure to back them up with numbers or observations from the graph.

g. i. The lowest price that tomatoes got to was just over $\$2$ per kg (around $\$2.20$) in January of 2013.



This is represented by the lowest point in the red line – using a ruler, this lines up with 1/01/2013 (January) and just above $2 per kg. $2.20 is a sensible guess.

ii. In general, tomatoes seem to be cheapest during January and April, quite a bit more expensive in July, and most expensive in October.

A similar pattern repeats each year. There are 4 points per year, so you can essentially rank them when describing the seasonal trend. This pattern makes sense, since tomatoes are in season in warmer months (like January and April) and out of season in colder months (like July and October).

iii. There is a spike in October of 2013 where tomatoes cost roughly $12/kg. This is the highest peak on the graph and is much higher than we would expect from other years.

In a time series, unusual features are points that don't follow the overall trend or seasonal trend. Make sure you state the value associated with any unusual features.

h. i. Group 1 has a higher median score.

The middle bar for group 1 is further along the graph than the middle bar for group 2, meaning group 1 has a higher median.

ii. Group 2 has a larger IQR. From the graph, we can see this by comparing the width of the middle 50% boxes, the middle 50% box for group 2 is clearly wider than for group 1.

iii. The scores for group 1 seem to be skewed to the right, as the middle 50% box is quite asymmetrical (although the whiskers are similar lengths). The scores for group 2 are more symmetrical, but still slightly skewed to the right.

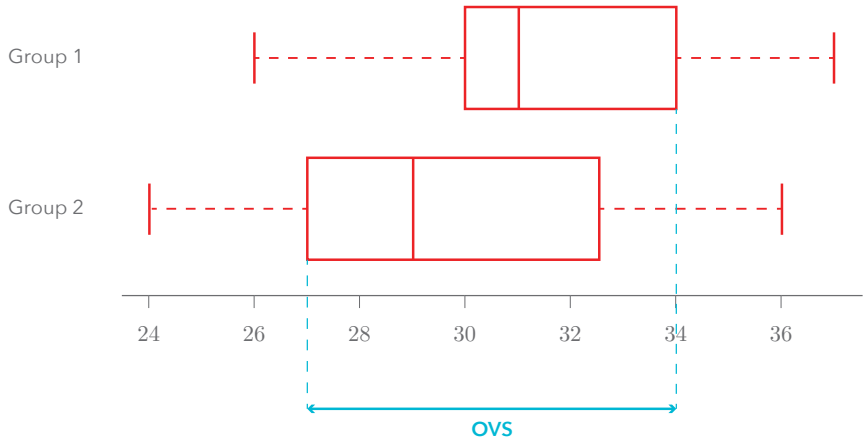
Again, symmetry and skew can depend on your personal judgement. In general, the middle 50% box should be more important to your decision than the whiskers because the whiskers could be affected by outliers.

iv. DBM = 2

The median for group 1 is 31, and the median for group 2 is 29.
The DBM is the difference between these two numbers:
31 - 29 = 2.

v. OVS = 7

The OVS is the total distance between the lowest LQ and the highest UQ, as shown in the diagram on the following page:



This goes from 27 to 34, so the OVS is:
34 – 27 = 7.

vi. The median for group 1 overlaps with the middle 50% for group 2. However, the median for group 2 is lower than the LQ for group 1, meaning that at least 50% of group 2 members scores lower than at least 75% of group 1 members. This is some evidence that group 2 tend to get lower scores than group 1.

Discussing shift and overlap involves comparing each median to the middle 50% for the other group. You should use percentages like in the second sentence of the answer above to explain what your observations mean. Shift gives some evidence of a difference between groups, although it is better to use the DBM and OVS!

Section Two

1. Probability

1. a. i. What is the probability that the first student is present?

$\frac{351}{443} = 0.7923$

ii. What is the probability that the second student is present. Remember that we have already selected one student.

$\frac{350}{442} = 0.7919$

We've rounded to 4 decimal places here just to show how this is different from the previous probability. When we selected the first student, we removed one student from the total sample – so we have 442 students left to choose from, and 350 'present' students left to choose from.

iii. What is the probability that both students are present?

$\frac{351}{443} \times \frac{350}{442} = 0.627$

Note: in past years NZQA have sometimes accepted answers where this process was not done, i.e., you would be marked right even if you said the probability of the second student being present was 351/443. However, to be safe and to build your understanding for future years, it is best to take into account the student we have already selected.

b. i. Calculate the probability of a year 11 student being present:

$\frac{132}{148} = 0.892$

There were 148 year 11 students in total, and 132 of them were present.

ii. Calculate the probability of a year 12 student being present:

$\frac{121}{176} = 0.688$

iii. Calculate the probability of a year 13 student being present:

$\frac{98}{119} = 0.824$

iv. *Write a sentence in context to answer the question.*

Year 11 students were most likely to be present, as the probability of being present was 0.892 for year 11s, while it was only 0.688 for year 12s and 0.824 for year 13s.

c. i. Calculate the probability of being marked 'justified absent' for a year 12 student:

$\frac{42}{176} = 0.239$

ii. Calculate the probability of being marked 'justified absent' for a year 11 student:

$\frac{11}{148} = 0.074$

iii. Divide the probability for year 12s by the probability for year 11s:
$\frac{0.239}{0.074} = 3.23$

iv. Write a sentence in context to answer the question using the number you just calculated:

Year 12 students were 3.23 times as likely to be marked 'justified absent', compared to year 11 students.

2. a. i. *Pick out the two percentages from the question that are relevant.*
We are looking at days where it does rain: 20%
And days that Matilda goes for a walk, if it rains: 15%

ii. *Convert these percentages into decimals.*
$20\% = 0.20$
$15\% = 0.15$

iii. *Use the probabilities to calculate the probability of rain and going for a walk.*
To find the probability of two things happening, we multiply their probabilities together:
$0.20 \times 0.15 = 0.03$
The probability that it rains and Matilda goes for a walk is 0.03.

2. Describing and Comparing Graphs

1. a. i. *Did Elana tend to spend more in year 11 or year 12? How can you see this on the graph?*

Elana tended to spend more in year 12 than in year 11. We can see this on the graph by the fact that the bars for year 12 are higher than the bars for year 11 in every month except August.

ii. *Why might this be?*

Prices might have gone up at the canteen, which would mean Elana would spend slightly more even if she bought the same things. Elana might have had more money to spend in year 12 than year 11, so she would have been able to buy more.

iii. *Which months were particularly low each year? Why might this be?*

Elana didn't spend any money at the canteen in January, November or December in either year. This is probably because she would have been on exam break or summer holidays during those months. She also spent less than usual in April, July and October both years (about half as much as in other months). This could be because school holidays would be around these times, so she wouldn't be at school for as many days.

iv. *Which months did Elana spend the most in each year? Why might this be?*

Elana spent the most in May and June each year. These are cold months, so maybe she wanted to buy warm food and spent more.

v. *Was Elana's pattern of spending overall similar between the two years?*

Overall, Elana's pattern of spending was fairly similar between year 11 and year 12, except for spending slightly more in year 12.

It's always a good idea to wrap up with a summary sentence when you're comparing graphs!

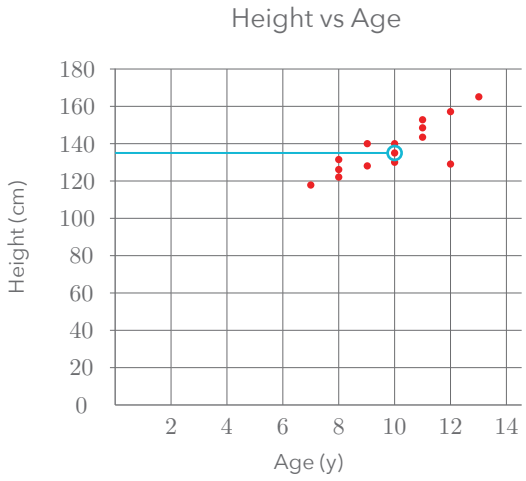
You really only need one good explanation for each point, even if it's different from the examples we've given. The important thing is to give a sensible possible reason and then explain how it would lead to your observation.

2. a. i. *Count the total number of points*

There are 15 points in total.

ii. *Find the middle point and read off the height it corresponds to*

This means we need to find the 8th point up from the bottom.



The median height of girls in this group is roughly 135cm.

Just like in previous questions, it's fine if your answer is different by a few centimetres, so long as it's between 130 and 140cm.

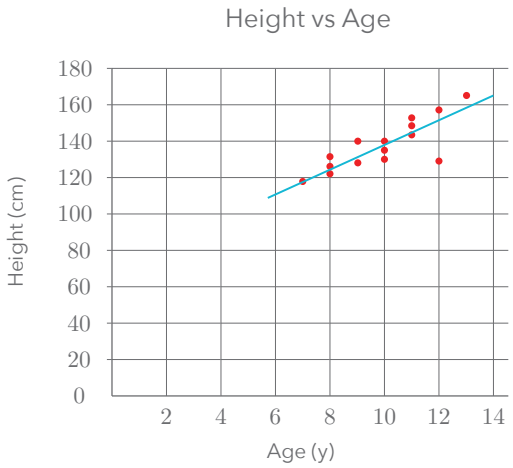
iii. Write a sentence to explain your working.

E.g., "I counted how many points there were, then I found the middle point and read its value off

the y-axis".

Although we told you what to do in this case, you should always either show or explain your working.

b. i. *Draw a line of best fit for the graph*



ii. *What does your line of best fit show about the relationship between age and height?*

The line of best fit shows that there is a positive relationship between age and height; as age increases, height tends to increase.

iii. *What shape is the relationship (linear or curved)?*

The relationship between age and height in the graph is linear.

iv. *How much scatter is there around the line of best fit? Is there the same amount of scatter across all ages?*

There is not much scatter around the line of best fit, which means that height and age are strongly related in this group. The scatter is fairly constant across the graph.

v. *Are there any outlier(s)? Where are they? Are they unusually high or low?*

There is one point that is unusually low; a person who is 12 years old and roughly 130cm tall.

vi. *How would you deal with the outlier(s)?*

This seems like a realistic value, so it doesn't seem like a mistake. It is reasonable to think that a 12 year old might be quite a bit shorter than other people their age, because people can be lots of different heights. Therefore we would keep it in the data.

When discussing an outlier, you should always state whether you think it is a mistake, or a real value, with a reason. Mistakes should usually be taken out of data, while real values should be kept in.

3. a. i. *What is the overall trend over time? (Do prices seem to be increasing, decreasing or staying roughly the same?)*

Tomato prices seem to be increasing over time, as the graph slopes slightly upwards for both supermarket and market prices. For example, in January of 2012, the supermarket price was around $2.50/kg, but in January of 2016 the supermarket price was almost $5/kg!

Notice you can write about both supermarket and market prices in the same sentence.
Giving numbers from the graph is a great way to back up your observations and raise your grade.

ii. *What is a potential reason for this long term trend?*

This could happen because food in general got more expensive over this time period, or maybe the sellers figured out they could raise their prices and people would still buy tomatoes.

You don't need to know anything about tomatoes to give reasons, just try to say something that could happen in the real world.

iii. *What is the seasonal trend? What is a potential reason for this pattern?*

For both the market and supermarket prices, tomatoes tend to be most expensive in October and cheapest in January and April each year. They are slightly cheaper in July than in October.
This is probably because tomatoes are in season in warm months, so they cost less to buy then.

iv. *How do supermarket prices and market prices compare overall? Why might this be?*

Supermarket prices and market prices tend to follow a pretty similar pattern across time, but market prices are usually lower than supermarket prices.
This is probably because prices are affected by the same factors at both the supermarket and the market, but markets can sell for cheaper. They might have less good looking tomatoes!

v. *Are there any unusually high or low points? Can you think of an explanation?*

The highest price in this time period is $12/kg at the supermarket in October of 2013. This is much higher than prices usually are at this time of year and clearly stands out on the graph. The market price also spikes up to around $9/kg at the same time, which is also unusually expensive. Maybe this was a particularly cold winter, or there was an issue with the tomato crop that year.

See page 13, "Time Series Graph", page 15 , "Outliers", pages 21-23, "How Graphs Follow Trends but are Still Cool", and page 29, "Comparing Two Sets of Data" of the Walkthrough Guide.

4. a. i. *Which dot plot looks more spread out, at a glance?*

The ballet plot seems more spread out, from both the dot plot and the middle 50% boxes of the

box plots.

ii. *Calculate the interquartile range of ages for each class.*
IQR = UQ – LQ

IQR for hip hop ages = 32 – 25 = 7
IQR for ballet ages = 27 – 23 = 4

iii. *Use the previous points to explain which class had greater variability, referring back to the IQRs and the graph.*

The hip hop class had more variability in ages. The IQR of the ages in the ballet class is only 4 years, while the IQR of the ages in the hip hop class is 7 years. The dot plot for hip hop is also much more spread out than the dot plot for ballet.

b. i. *Are there any outliers in the dot plots?*

There is an unusually old student in the ballet class (at 37 years old). There are no clear outliers in the hip hop class but there are a few points that seem distant from most of the others on the older side – with the oldest being 44.

ii. *Do outliers affect the range?*

Outliers do affect the range, as the range only takes into account the maximum and minimum values in the data.

iii. *Do outliers affect the IQR?*

Outliers do not affect the interquartile range because the top and bottom 25% of values are cut out. The IQR only measures the spread of the middle 50% of the data.

iv. *Why would this matter? Hard question, but an important one!*

We want to be able to make statements about the data in general, so it's unhelpful if one outlier affects our results a lot! Using the IQR means we can talk about spread without worrying about outliers changing our results.

c. i. *Where are the most common values?*

The plot of the ages for the ballet class has one mode at 24 years.
The plot of the ages for hip hop has three rough peaks – one just under 25, one around 26-27, and one at 32. However, this is probably just because of random chance.

ii. *Are the plots skewed or symmetrical?*

The plot of the ages for ballet is slightly skewed, with a tail to the right.
The plot of the ages for the hip hop class is more symmetrical, although there are a few quite high

values.

Describing shape is definitely an area where you could get a different answer from someone else and both be marked right! Just make sure you always back up your statement with numbers and/or reasons.

d. i. *What is the difference between the medians?*

The difference between the medians is 4 years (28 - 24 = 4).

Writing a calculation in brackets is an easy way to show your working within an answer.

ii. *What is the overall visible spread?*

The overall visible spread is 9 years (32 – 23 = 9).

iii. *Decide whether to use the $\frac{1}{3}$ rule or $\frac{1}{5}$ rule to compare the medians. (Hint: look at the sample size).*

The sample size is 60 for hip hop and 45 for ballet. This seems closer to 30 per group than 100 per group, so we should use the $\frac{1}{3}$ rule.

Always justify which rule you're using by referring to the sample size in your answer.

iv. Using the rule you chose, compare the difference between the medians to the overall visible spread.

$\frac{1}{3}$ of the overall visible spread would be 9 / 3 = 3 years. The difference between the medians is 4 years, which is more than this.

v. *Based on this, can we conclude that there is a difference in ages between the two dance classes?*

Since the difference between the medians is more than $\frac{1}{3}$ of the overall visible spread, we can say there is a difference in ages between the hip hop class and the dance class. People who go to the ballet class tend to be younger.

vi. *If you got stuck on the steps above, you can also use shift and overlap to support your answer: does the median for one group fall outside the middle 50% for the other group?*

The median for the ballet class falls below the LQ for the hip hop class, meaning at least 50% of the ballet sample were younger than at least 75% of the hip hop sample.

The median for the hip hop class is also higher than the UQ for the ballet class, meaning that at least 50% of the hip hop sample were older than at least 75% of the ballet class.

This gives us evidence that there probably is a difference between the two classes, with the ballet class tending to be younger.

The DBM/OVS answer will generally be at an Excellence level, while the shift/overlap answer is more

of a Merit technique.

e. i. *What impact does sample size have on the dot plot?*

Since each dot in the dot plot represents one person, a larger sample size will mean more dots in total. This means all the peaks etc will look higher, which makes it hard to compare across different sized groups.

ii. *What impact does the sample size have on the box plots?*

The box plots are not affected by sample size, since they just summarise the minimum, maximum, LQ, UQ, and median.

iii. *Should we be concerned about using the plots overall to make conclusions?*

Overall we can still use the box plots to make conclusions, and just not pay as much attention to the dot plots. Both sample sizes are big enough to be reliable (both more than 30) so our results will still be valid.

f. i. *Is 10 a big enough sample size to be confident about results? Why or why not?*

10 people overall probably isn't a big enough sample size to get good results, because the results will be very affected by sampling variability (random chance). We want each sample to be ideally at least 30.

As a rule of thumb, comparing sample sizes to 30 is a quick way to decide whether results are likely to be valid. Talking about sampling variability is an extra boost to your answer.

ii. *What is the difference between the way Petra selected her sample, and the way the dance school selected their sample?*

The dance school selected a random sample. Petra asked her close friends, who might be different from a random sample (they are probably similar to her!).

iii. *Are there any factors that might be different between the two surveys?*

Petra's friends might not be representative of all people who go to adult dance classes, especially since they all probably know each other.
Petra's friends might also go to a different dance school, which could attract people from a different group.

5. a. Once again, your answers might be different, but you should try to have at least one advantage and one disadvantage for each plot type.

i. *Advantage(s) of pie charts:*

Pie charts make it clear that everything adds up to 100% since each student could only be in one

group.

ii. *Disadvantage(s) of pie charts:*

It's inconvenient to need two different graphs to compare the groups.
Small slices like the 'languages' slice are hard to read and compare.
It's also hard to tell at a glance (before checking the percentages) how the slices compare.

iii. *Advantage(s) of proportional bar chart:*

It's much easier to compare two groups when they're on the same graph. It's also easy to see which bars are higher than others at a glance.

iv. *Disadvantage(s) of proportional bar chart:*

This kind of chart would start getting hard to read if we had more than two groups.
It's also less obvious that each group adds up to 100%, compared to using a pie chart.

b. The most sensible other option would be a stacked bar chart.

Like a pie chart, these show all the sections adding up to 100%, but it's easier to visually judge size on a bar compared to a circle.

6. a. The proportional bar chart isn't affected by the different sample sizes, because it shows everything as a proportion of the overall group. The frequency bar chart is affected by sample sizes, because the heights of the bars are given by the number of people in each category. There are more year 10 students than year 11 students, which makes the bars for year 10 look higher.

b. If we want to compare the two groups, we should use the proportional bar chart. The sample size usually shouldn't affect our comparisons, so it's better to use a graph that isn't affected by sample size.

3. Making and Assessing Predictions

1. a. i. *What proportion of the year 11s in the table were marked as 'justified absent'?*

$$\frac{11}{148} = 0.074$$

Be really careful when doing these calculations – we are only looking at year 11s, so the total is 148, not 443.

ii. *Multiply this proportion by the total number of students the prediction is about (31):*

$$0.074 \times 31 = 2.294$$

iii. *Round to the nearest whole number and put into context:*

We would predict about 2 justified absences in this year 11 class.

b. i. *Is there any other information that would be helpful?*

It would be good to know why those students in the table were marked as justified absent (e.g. sickness). It could also be useful to know how many students are usually justified absent, not just on one day.

ii. *There is a high number of justified absences for year 12 students in the table - why could this be?*

It could be that there was a field trip or class event on that day for year 12s, which would mean a lot of students would be marked as justified absent. If there was a class trip for year 11s the next day, this would affect the absences as well.

iii. *Decide whether the same pattern would be seen on Friday as on Thursday:*

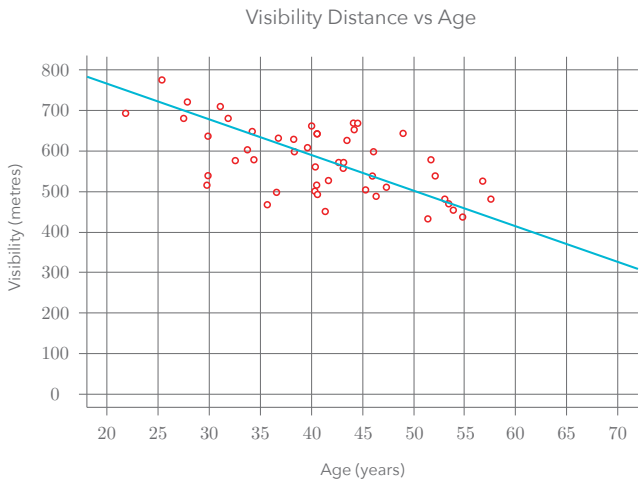
The pattern of absences on Fridays could be different to Thursdays, especially if e.g., class trips are more likely to be at the end of the week.

You could also argue that Fridays would be similar to Thursdays! You're probably getting by now that these answers are all about justifying yourself.

iv. *Sum up briefly whether the prediction is likely to be useful overall:*

Overall, there are too many other factors that could affect the number of justified absences, so this prediction is probably not very useful.

2. a. i. *Draw a line of best fit on the graph.*



ii. *Find the gradient of the line (using rise/run).*

Rise: $400 - 700 = -300$

Run: $62 - 25 = 37$

Gradient $= \frac{-300}{37} = -8.1$

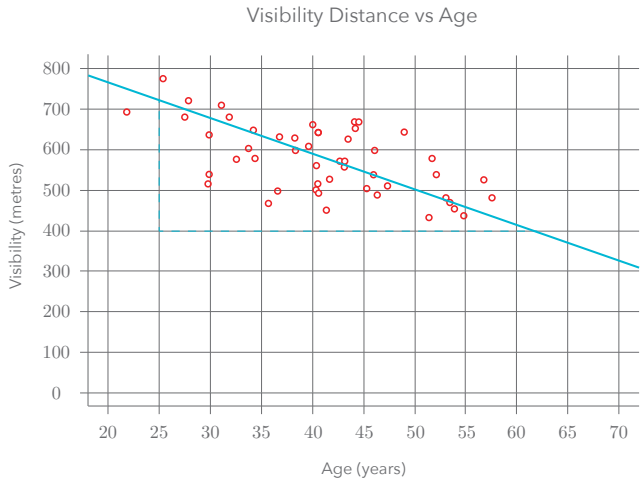
It doesn't matter what points you use, but it's best to use points that are fairly far apart and are easy to estimate the position of. Show your calculations clearly, as there is often a lot of variety in the values you could get and your marker needs to know where your gradient value came from.

iii. *Put the gradient into context to answer the question.*

On average, visibility distance decreases by about 8 metres per year of age.

Make sure to include units in your context sentence! We've rounded here because the line of best fit is just a 'best guess'.

b. i. *Label the point on your line of best fit where the age is 70.*



ii. *State what the visibility distance is at this point:*

Roughly 330 metres. *Your answer will depend on your line of best fit.*

iii. *Explain how you got your predicted visibility distance:*

E.g: I found the point on my line of best fit that lined up with 70 along the x-axis, then I used a ruler to find what visibility distance this corresponded to.

c. i. *Is the relationship from the graph clearly visible?*

The negative relationship between visibility distance and age is quite clear from the graph.

ii. *How strong is this relationship? What does this mean in relation to the graph?*

This is a moderate relationship (not a very strong one). There is a reasonable amount of scatter around the line of best fit.

iii. *How would you use the linear relationship to predict visibility distance?*

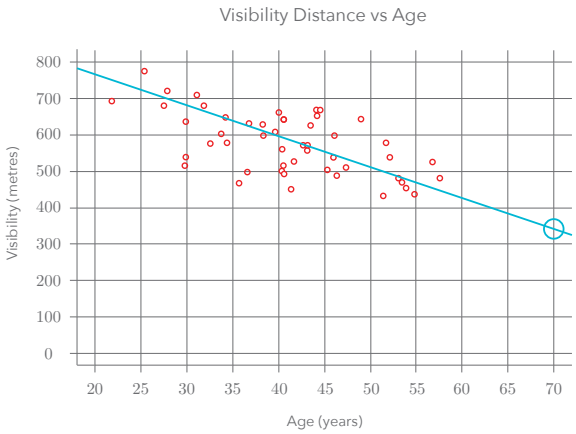
To predict visibility distance using this relationship, we would use the line of best fit to find the expected visibility distance for each age.

iv. *Are there any other factors than age that you would want to take into account to make a prediction?*

Visibility distance is probably affected by other factors. For example, if participants wore glasses or contacts, or if they had some kind of eye impairment.

v. *Briefly sum up the usefulness of age for predicting visibility distance.*

Overall, age seems to be useful for predicting visibility distance, but because of the scatter and other factors that could contribute, we should not be overly confident about predictions.



Section Three
Practice Exam

Question One

a. i. $\frac{14}{30} = 0.467$

Achieved for correct answer.

ii. $\frac{2}{14} = 0.143$

Merit for correct answer.

iii. Probably not. It's true that in the sample, the majority did not get 5+ a day (16/30 did not), but:

- 30 people is a pretty small sample to make a statement about all New Zealanders: there would still be quite a lot of sampling variability. Maybe if the doctor asked a different sample of 30, she would get a different result.
- The doctor only got people to fill in the diary for one day, and the number of servings of fruits and vegetables they eat each day may vary.
- Patients at the doctor's office might be less healthy overall than people who didn't go to the doctor (since they probably went because they were sick!)
- Patients at this particular doctor's office might not be representative of all New Zealanders because they will all live in similar areas.
- The results depend on people filling in a food diary, which they might not have done accurately.

Achieved for saying it is a fair claim with reference to the graph, or for saying it is not a fair claim with a valid reason.
Merit for saying it is not a fair claim with two valid reasons.
Excellence for saying it is not a fair claim with three valid reasons.

Your reasons may be different from the examples given.

b. i. There is a negative, linear relationship between the number of servings of fruits and veges, and resting heartrate. As the number of servings increases, the resting heartrate decreases.

Achieved for correctly stating relationship.

ii. Eli's predicted heartrate would be about 50bpm (anywhere between 40 and 60). To find this, I drew a line of best fit on the graph, found the point where 'number of servings' was 9 and used a ruler to see what heartrate this corresponded with.

Either: This prediction seems valid, because...

- The relationship between servings and heartrate is strong, most of the points are close to the line of best fit, so it is valid to use this line to predict resting heart rate.
- The relationship looks linear, so it is valid to use the line of best fit to approximate resting heart rate by using the number of servings of fruits and veggies consumed.

Or: This prediction is not very valid, because:

- There are no points at 9 servings, so we can't be sure that there is still a strong relationship at this point.
- Too many other factors could affect Eli's heartrate, like how much exercise he does, how old he is, whether he has a heart problem, etc.
- This is quite a bit lower than all the other points, including the one at 11 servings, so it doesn't seem realistic.

Achieved for correct prediction with working explained.
Merit for correct prediction, and one valid reason explaining whether it is valid or not.
Excellence for correct prediction, and two valid reasons explaining whether it is valid or not with at least one reason referring to the graph.

iii. Number of servings of fruits and vegetables is more strongly related to resting heart rate, than to fasting blood sugar.

- There is less scatter on the resting heart rate graph than the fasting blood sugar graph.
- There is more variability between points on the fasting blood sugar graph.

Achieved for correctly identifying resting heart rate.
Merit for using scatter/variation between points to justify your answer.

Question Two

a. i. There is more variety in the amount of water consumed by teenagers than adults.

- The interquartile range of the amount of water for teenagers is 450mL (1950 – 1500) which is quite a bit larger than the interquartile range for adults, which is 300mL (2000-1700).
- The range of the amount of water for teenagers is 5100, which is also a lot more than the range for adults, which is 800mL (1800-1000).
- The dots on the 'Teenagers' dot plot are more spread out than the dots on the 'Adult' dot plot.

Achieved for stating teenagers, with at least one reference to the graph.

ii.
- The median amount of water consumed by adults is 1800mL, which is 200mL higher than the median for teenagers.
- Both graphs have one peak. The most common amount of water consumed by teenagers is about 1700mL, while the most common amount for adults is around 1900mL.

- The graph for adults looks pretty symmetrical. The graph for teenagers is more skewed, with a few very high values.
- There is a definite outlier in the 'teenagers' graph, at 6000mL, while there are not high outliers in the graph for adults.
- There are some slightly low points in both the teenagers and adults group, at around 1000mL, that are visibly lower than the rest.

Achieved for clearly describing one feature.
Merit for clearly describing two features, including one direct comparison between the graphs.
Excellence for clearly describing three features, including two direct comparisons between the graphs.

iii. The median is found by identifying the middle number in the data. The mean is found by adding up all the values, then dividing by the number of values. The median is not affected by extremely high or low values, because it is just the middle value. Mostly the teenagers seem to drink less water, so the median is lower than for adults.

However, there are also some very high values in the teenagers sample. One person said they drank 6000mL, and one said they drank just under 4000mL. These very high values bring the mean up, so the mean for teenagers ends up higher than for adults.

Achieved for explaining how mean and median are found, or mentioning outliers.
Merit for stating that outliers affect the median, but do not affect the mean.
Excellence for writing a clear explanation, with reference to the graph.

b. i. 8-10 hours.

Achieved for correct answer.

ii. Reading off the graph, roughly 12 teenagers got less than 6 hours, 21 got 6-8 hours, 18 got 8-10 hours, and 7 got more than 10 hours.

In total, this gives 12 + 21 + 18 + 7 = 58 teenagers in this sample.
So, the proportion who got 8-10 hours sleep is $\frac{18}{59} = 0.31$.

Achieved for correct answer (any value between 0.25 and 0.35).
Merit for correct answer with clear working.

iii. The most common amount of hours of sleep was 6-8 for both teenagers and adults.
The second most common amount was less than 6 hours for adults, and 8-10 hours for teenagers.
Teenagers were more likely to get 8-10 hours than adults, and much more likely to get more than 10 hours.

Achieved for one valid comparison.
Merit for two valid comparisons.

Question Three

a. i. There is a similar variety between adults and teeangers drinking coffee. However, it is more common for a teenager to not drink coffee (0.143) versus an adult (0.085).

Teeangers: $\frac{5}{35} = 0.143$

Adults (22+): $\frac{22}{259} = 0.085$

Achieved for correct answer.

ii. The proportion of people who never drank coffee in age group was:

Younger than 18:	$\frac{5}{35} = 0.143$
18-21:	$\frac{32}{208} = 0.154$
21-25:	$\frac{17}{192} = 0.089$
25-30:	$\frac{4}{50} = 0.08$
Older than 30:	$\frac{1}{17} = 0.059$

The likelihood of never drinking coffee was highest for the 18-21 year old age group.

Achieved for correct answer.
Merit for correct answer with clear working.

iii. This information could be presented in:

- A pie chart for each group.
- A stacked bar graph.
- A clustered bar graph.

Reasons for which graph you would prefer to use (and which you would prefer not to) could include:

- I would prefer to use a pie chart because it makes it obvious that all the different groups (of coffee consumption) add up to 100% for each group.
- I would prefer not to use a pie chart because I would need one chart for each age group.
- I would prefer not to use a pie chart because it is hard to compare the values.
- I would prefer to use a stacked bar graph because it makes it obvious that all the different groups (of coffee consumption) add up to 100% for each group.
- I would prefer to use a stacked bar graph because it is easy to compare groups at a glance.
- I would prefer to use a clustered bar graph because it allows all the groups to be shown on the same graph.
- I would prefer not to use a clustered bar graph because there would be a lot of bars, which could be hard to read.

You could also have other good reasons to use or not use your chosen graph.

Achieved for stating two valid graph types, or one graph type with a reason to use it.
Merit for stating two valid graph types, with one justification of your preference.
Excellence for two valid graph types, with at least two justifications of your preference.

iv) In the sample, $\frac{1}{17}$ students older than 30 drank coffee 'Never', and $\frac{1}{17}$ drank coffee 'occasionally'. In total, $\frac{1}{17} + \frac{1}{17} = \frac{2}{17}$ drank coffee either 'Never' or 'Occasionally'.

$\frac{2}{17} = 0.118$
$0.118 \times 782 = 92.276$

We would expect 89 or 90 students aged over 30 to drink coffee either 'never' or 'occasionally', at this university.

Achieved for getting to $\frac{2}{17}$ or 0.118.
Merit for getting the correct answer, even if not rounded.
Excellence for getting the correct answer, with working shown, and rounded to a whole number.

b.
- Overall, the caffeine level in the student's bloodstream is slightly higher on Friday until about 5pm, and is slightly higher on the Thursday after then.
- On both days, the overall trend is for the student's caffeine level to rise between 7am and 5pm, and then fall again between 5pm and 12am.
- There is a sort of seasonal pattern where her caffeine level seems to spike every 3 or 4 hours (probably when she drinks a coffee). This happens at 8am, 12pm, and 4pm on both Thursday and Friday.
- On both days, the lowest caffeine level is around 25mg, at 7am. The highest level is around 250mg. This level is reached at 12pm and 4pm on the Thursday, and at 4pm on the Friday.
- There is a small spike at 11pm on the Thursday, but not on the Friday.

Achieved for one valid point.
Merit for two valid points.
Excellence for three valid points, including the first point or something very similar.

For each question, give yourself:

N0	N1	N2	A3	A4	M5	M6	E7	E8
No correct working	Some correct working with no correct answers	One achieved point	Two achieved points	Three achieved points	Two merit points	Three merit points	One excellence point	Two excellence points

Add up your scores on the three questions to get your overall mark:

0 - 8 = N
9 - 13 = A
14 - 18 = M
19 - 24 = E

