SmartPrep Math Syllabus: Student Pack

Author: Arne du Toit

August 2020

1 Workshop 5

Abstract

Focus on Questions 1 and 2 of exam paper 2.
Question 1 & 2: Statistics

1.1 Statistics

1.1.1 Terminology

Median:

The median value refers to the middle value of an organised (arranged data set). The arrangement of the data should either me in ascending or descending order. For an odd amount of data points you divide the number by two and round up to the next integer to identify the position of the median. Consider the following organised data set consisting of 11 data points (x_i-values),

$$2, 2, 4, 5, 5, 5, 7, 8, 8, 11, 15.$$

Since $n = 11$ is an odd number of data points we divide by two and round up to the nearest integer.

$$\begin{aligned} \frac{n}{2} &= \frac{11}{2} \\ &= 5.5 \\ &\implies 6. \end{aligned}$$

This the sixth position (x_6) holds the value of the median. In this case the median is $x_6 = 5$. If the total number of data points are even then you still divide the value by two, but in order to determine the median you apply the following method:

$$\begin{aligned} \frac{n}{2} &= a \\ \frac{n}{2}+1 &= b \\ median &= \frac{x_a + x_b}{2}. \end{aligned}$$

Consider the following 14 data points:

$$4, 7, 7, 9, 14, 14, 15, 18, 22, 24, 25, 25, 25, 32$$

$$\begin{aligned} \frac{14}{2} &= 7 \\ \frac{14}{2}+1 &= 8 \\ median &= \frac{x_7 + x_8}{2} \\ median &= \frac{15+18}{2} \\ median &= 16,5. \end{aligned}$$

Thus the median for this set of values is the addition of data points x_7 and x_8 divided by two such that the median equals 16,5.

Mean:

The mean value is considered the average of a data set. If all of the values in a data set had to be represented by a single value then the mean would be that value. In order to calculate the mean you sum all values and divide that number by the total number of values in your data set.

$$\overline{x} = \frac{\Sigma_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + x_3 + ... + x_{n-1} + x_n}{n}.$$

Thus in the numerator you take into account the actual value assigned to x_i while the denominator creates a weighted factor that depends on how many values are assigned a certain value. This is a similar concept as the one we saw in probability theory (W4) where the likelihood of an event happening is the number of ways that event can take place divided by the total number of outcomes.

Variance:

The variance represents how 'spread out' the data is. Spread out refers to how far each value is from the mean value. Thus in order to calculate the variance you first need to calculate the mean of the data set. The variance is calculated from taking each data point and subtracting the the mean value from it, this gives us the difference between the two values. Next we square this value, to convert it into a positive value. Since you summed a calculation consisting of each data point it makes sense that you would once again divide by the total number of data points, to once again get that averaged effect. When variance is denoted by σ^2 then the total equation looks as follow:

$$\sigma^2 = \frac{\Sigma_{i=1}^{n}(x_i - \overline{x})^2}{n}$$

Let's look at two data sets and discuss their variance:

$$A = 4, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 8 \qquad B = 1, 2, 3, 4, 5, 6, 6, 7, 8, 9, 10, 11$$

Both these data sets have a mean of 6 but they consist of very different data points. All values of set A are very closely distributed around the mean value, with the biggest difference being two. On the other hand, set B has a much wider spread with the largest difference between the mean and a data point being 5. What do we expect the corresponding variance values to be? I'll leave it to each student to calculate that the respective variances are:

$$A : \sigma^2 = 1,17 \qquad B : \sigma^2 = 9,17.$$

We see that set A, with the data distribution closely around the mean value, has a smaller variance than that of set B, who's data points have a wider distribution. This illustrates how interpretations concerning the data set can be made if you are given the variance.

Standard deviation:

The standard deviation is closely related to the variance since it is equal to the square root of the variance $\sigma = \sqrt{\sigma^2}$. Such that,

$$\sigma = \sqrt{\frac{\Sigma_{i=1}^{n}(x_i - \overline{x})^2}{n}}.$$

Why is standard deviation important? If you do not know the exact values in the data set, the standard deviation is used as a statistical indication of where you can expect a certain amount of data points to be. For example, 68% of a data sets values are found within ± 1 standard deviation from the mean value and 95% of values are distributed between ± 2 standard deviations from the mean.

Range:

When we discussed functions we saw the range of the function consisted of all the y-values for which the function is defined. This is similar in the way that the range indicates the interval within all data points lie. This can be obtained by subtracting the minimum value from the maximum value, however this does not mean that every value in the interval must appear in the data set. This can also be thought of as the length of a box and whisker diagram, but more on that a bit later.

Inter-quartile range (IQR):

The inter-quatile range contains the middle 50% of the data. The quartiles are calculated similarly to the median since they are the middle values between the minimum and the median and the median and the maximum. Lets look at an arranged set of data consisting of an odd amount of values.

$$1, 2, 5, 6, 7, 9, 12, 15, 18, 19, 27.$$

Once again we have $n = 11$ data points, thus our median will be at position $\frac{n}{2} = \frac{11}{2} = 5.5 \implies 6$, such that our median is $x_6 = 9$. This splits our data points into two sets of data consisting of the the lower and upper half respectively. The quartiles are the middle values of these two sets. Thus,

$$s_1 : 1, 2, \mathbf{5}, 6, 7 \qquad\qquad s_2 : 12, 15, \mathbf{18}, 19, 27.$$

$Q_1 = 5$ and $Q_3 = 18$. Why is it quartile 1 and 3, what happened to quartile 2? Technically $Q_2 =$ the median. These quartiles can be used to indicate the dispersion of the data in a different way to the variance and standard deviation. The inter-quartile range consist of the 2^{nd} and 3^{rd} quarters of the data set and can be determined by subtracting the values of Q_1 from Q_3. Thus for this data set the $IQR = Q_3 - Q_1 = 18 - 5 = 13$.

How does this process change when the data set consist of an even amount off data points? Consider the data set:

$$3, 5, 7, 8, 9, 11, 15, 16, 20, 21.$$

We now have $n = 10$ data points. Thus our median can be calculated as,

$$\begin{aligned}
\frac{10}{2} &= 5\\
\frac{10}{2} + 1 &= 6\\
median &= \frac{x_5 + x_6}{2}\\
median &= \frac{9 + 11}{2}\\
median &= 10.
\end{aligned}$$

Position wise the middle of this data set is between x_5 and x_6, thus this will be the position where the split between the data will occur. Such that,

$$s_1 : 3, 5, \mathbf{7}, 8, 9 \qquad\qquad s_2 : 11, 15, \mathbf{16}, 20, 21.$$

The identification of the quartiles remain trivial as it is just the initial split between the data that occurs differently. We have that $Q_1 = 7$ and $Q_3 = 16$ while the $IQR = 16 - 7 = 9$.

As mentioned above the values associated with IQR can be used to give us an indication of the dispersion of the data set, and this can all be represented by a box and whisker diagram (BWD). Lets look at the last mentioned data set's box and whisker diagram in Figure 1.

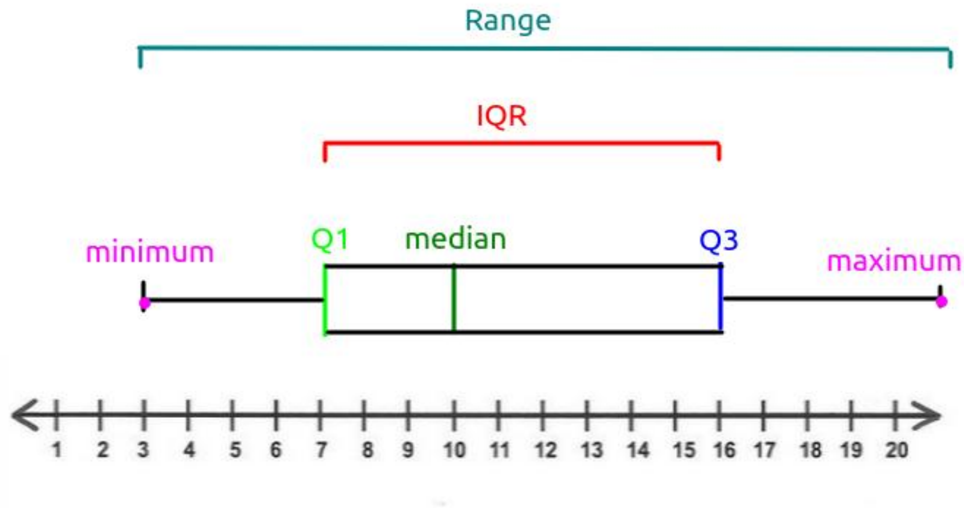


Figure 1: Box and Whiskers diagram (BWD)

We should be able to interpret a box and whisker diagram (BWD) in order to draw conclusions about the data set especially in the cases where we have not seen the entire data set. These conclusions get drawn due to the symmetry or lack of symmetry of the diagrams. We see that the diagram in Figure 1 is not symmetrical, since the distance between $Q1$ and the median is much smaller than the distance between the median and Q_3. This box and whisker diagram indicates a distribution that is skewed right. Lets take a look at the three possible distributions and discuss some characteristics (generally) associated to them.

Skewed right

- The mean value is greater than that of the median.
- The median is closer to Q_1 than it is to Q_3.
- Histogram: There is a steep incline on the left side of the distribution.
- Histogram: There is a steady decline on the right side of the distribution.

Symmetric

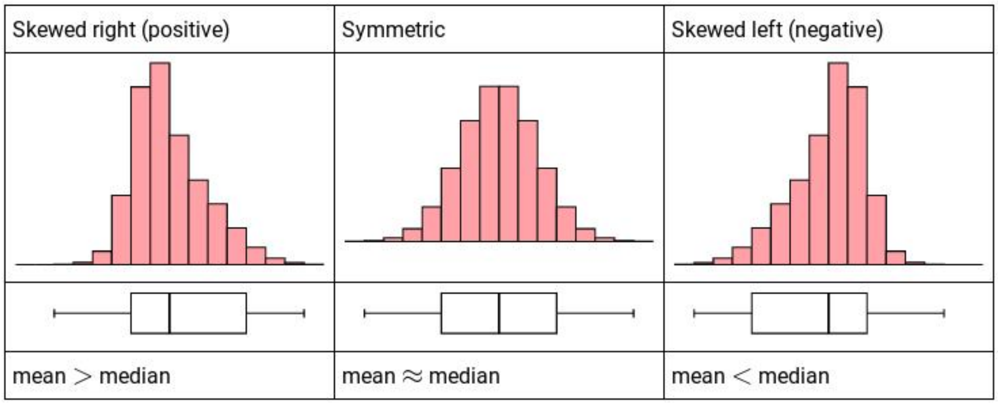


Figure 2: Different Box and Whiskers diagram [1]

- The mean value is very close to or sometimes equal to the median.
- The median is an equal distance from Q_1 and Q_3.
- Histogram: The distribution is very even on each side and resembles a graph that is reflected around the $y-$axis.

Skewed left

- The mean value is less than that of the median.
- The median is closer to Q_3 than it is to Q_1.
- Histogram: There is a steady incline on the left side of the distribution.
- Histogram: There is a steep decline on the right side of the distribution.

1.1.2 Scatter-plots

A scatter plot is a two-dimensional data visualization that uses dots to represent the values associated with a given data set. Since scatter plots are two dimensional they represent bivariate (depending on two variables) data. Figure 3 show a scatter plot where each dot represents a person, its placement with respect to the $x-$axis indicates the person height while the placement with respect to the $y-$axis indicates the persons weight.

Thus a scatter plot can be used to examine the relationship between two data sets (or the lack of a relationship). Data can follow a linear, quadratic or exponential trend, figure 4 illustrate the cases respectively.

Depending on the specific trend the data follows, one can estimate where a certain value might feature that is not already represented in the data set. We will look at these estimations regarding linear trends by means of 'Lines of Best Fit'. Drawing this line yourself will not represent the data perfectly, since each persons 'guess' to where the line should lie will be

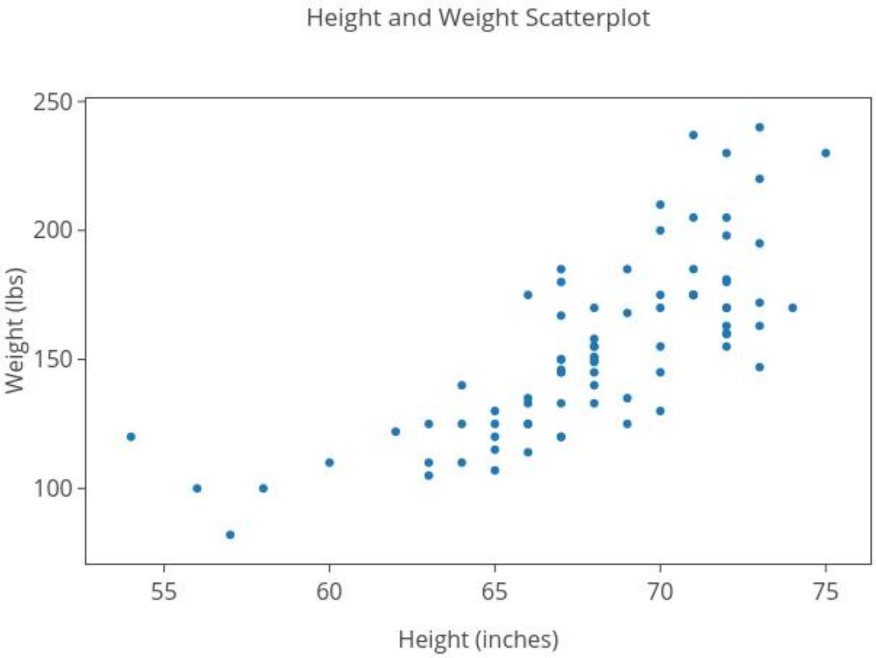
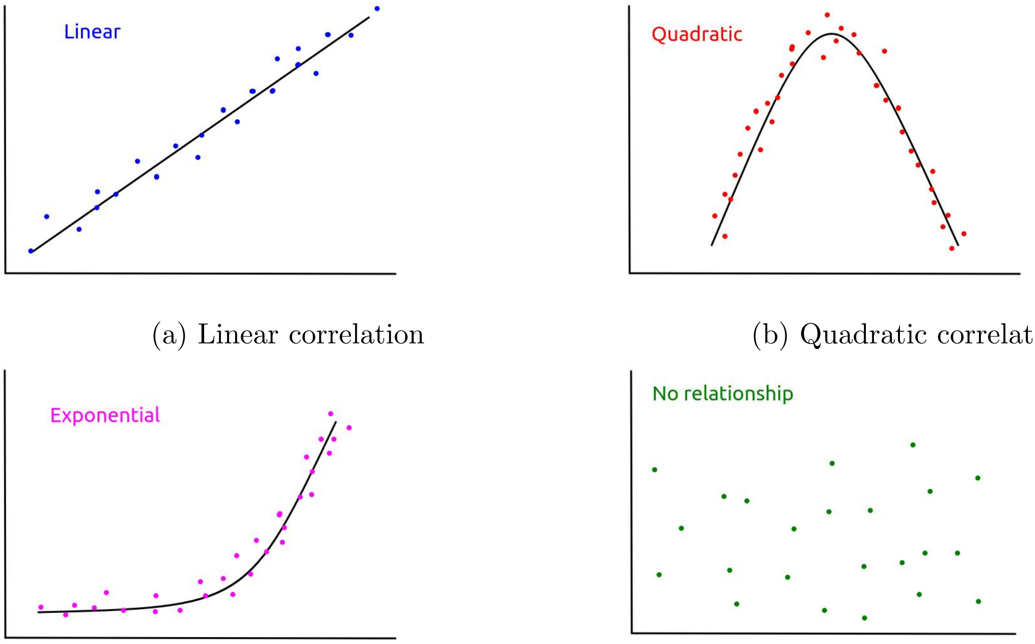


Figure 3: Bivariate data represented by a scatter plot [2]



(a) Linear correlation

(b) Quadratic correlation

(c) Exponential correlation

(d) No correlation

Figure 4: Different relationships

slightly different, but it is the first step in using an aid to estimate future values following the trend. These lines are drawn with the rough guideline of passing through some data points whilst aiming to have the same amount of data points above the line as underneath the line. Outliers need not be taken into consideration when determining these estimates. Outliers don't give accurate representation of the data since they usually appear when something completely out of the ordinary took place.

Example of when an outlier may occur: John owns a small beverage tuck shop in a small quiet town. He decides to invest a certain percentage of his profits, once a month, into marketing to see if his amount of sales will increase. In the first 3 months he starts seeing a slow but steady increase, but during month four a convoy of tour busses passed through the town and decided to make an unplanned pit stop in the town. This created an opportunity for passengers to step out of the bus and buy a beverage. This spiked Johns sales by an exponential factor for month four. The rest of the year the business returns back to normal and the advertising adds to a slow but steady increase in sales.

This represents a situation where an outlier in data occurred that does not add to the information portrayed by the data, as the advertising John had done had no effect on the passengers in the tour bus who accidentally ended up in the town. It is due to reasons similar to this specific situation that you can discard outliers to determine line estimates.

The more precise version of a line of best fit is called a regression line calculated by the least squares method. The equation to this line is the same equation we know for a straight line in a slightly different form, $y = a + bx$ where a represents the $y-$intercept and b represents the gradient of the slope. You are encouraged to solve for these variables using a calculator. Lets look at an example of such a calculator program [3]:(note: they assumed an outlier at $x = 6$ that's why it is left out)

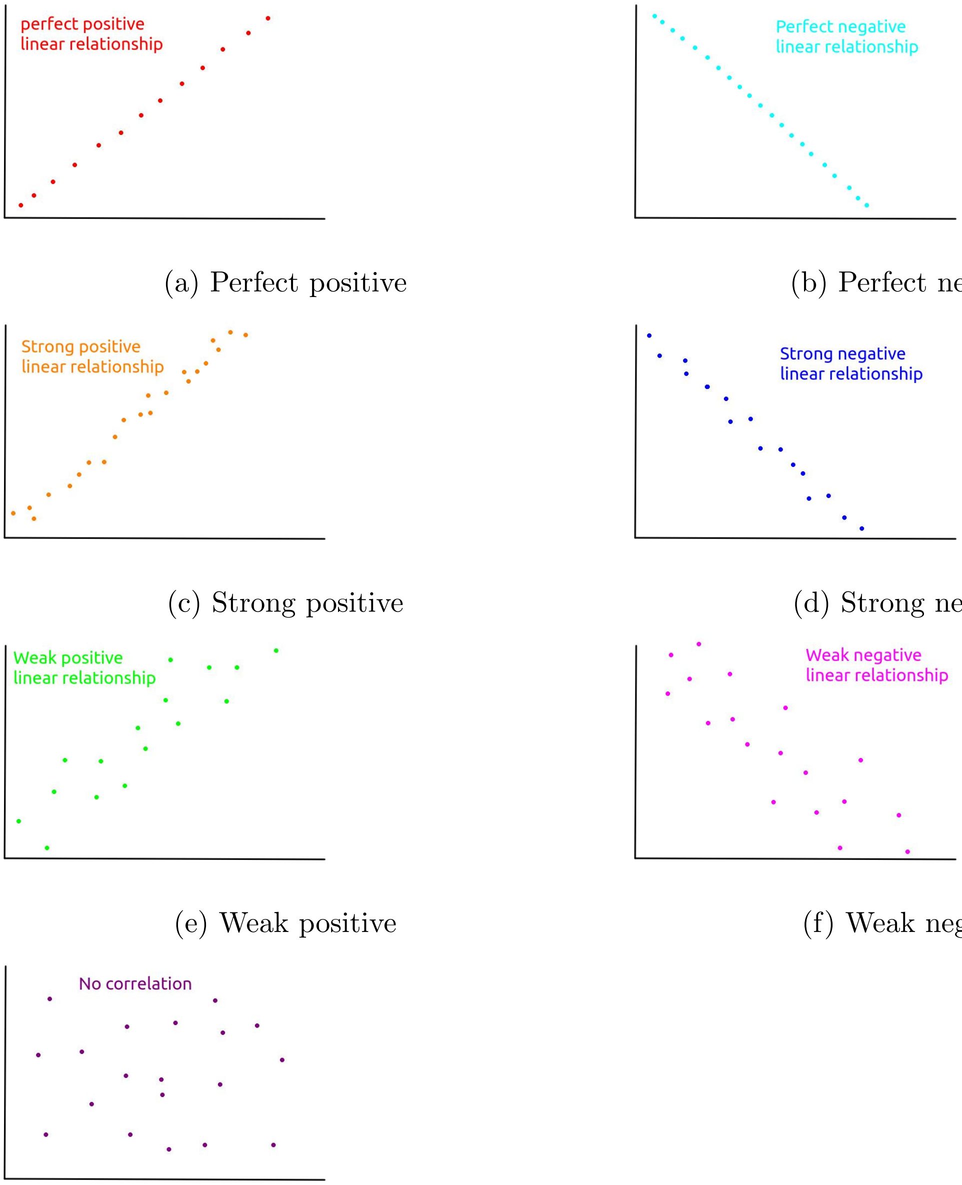
```
MODE 2
PRESS 2: A+BX
ENTER DATA POINTS:
COLUMN(X)
1= 2= 3= 4= 5= 7= 8= 9= 10= 11= 12=
COLUMN(Y)
10 22 20 38 46 48 62 61 74 88 86
THEN PRESS AC
PRESS SHIFT 1
PRESS 5:REG
PRESS 1: A = (to get the value of a which is) 5,315923567
PRESS SHIFT 1
PRESS 5:REG
PRESS 2: B = (to get the value of b which is) 6,896178344
PRESS AC
THEN PRESS MODE 1 TO GET BACK TO NORMAL MODE
```

Thus the equation of the regression line is $y = 6,896178344x + 5,315923567$. Something important to keep in mind or check is that the line of regression should pass through the mean points $(\overline{x}, \overline{y})$.

The last aspect that we must discuss is the strength of a data sets linear relationship. If the dots in a data set very closely resemble a straight line then the linear relationship is strong, likewise when the resemblance can not correlate to a straight line strongly then it is a weak linear relationship. If there is no resemblance to a straight line what so ever, then there is simply no correlation. The strength is indicated by a value r which can be displayed along with the line coefficients ($a\&b$) on your calculator. The value r is to be interpreted in the following way:

- $r = 1$: Perfect positive linear association.
- $r = 0$: No positive/negative correlation whatsoever.
- $r > 0$ and very close to 1: Strong positive linear association.
- $r > 0$ and fairly close to 1: Moderately positive linear association.
- $r > 0$ and $\leq 0,5$: Weak positive linear association.
- $r < 0$ and very close to -1: Strong negative linear association.
- $r < 0$ and fairly close to -1: Moderately negative linear association.
- $r < 0$ and $\geq -0,5$: Weak negative linear association.

A weak a strong linear relationship can further be categorised as being either positive or negative, depending of the gradient of the data points. Lets look at a visual summary of all these possibilities.



(a) Perfect positive

(b) Perfect negative

(c) Strong positive

(d) Strong negative

(e) Weak positive

(f) Weak negative

(g) No correlation

Figure 5: Strength of relationships

1.2 Example questions from past papers

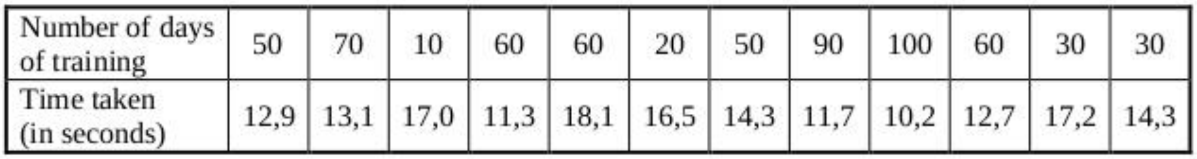
Lets consider some examples from past papers. Note the following marking criteria that will be followed when assessing your papers.

- If a candidate supplies two answer for a single question, then only the first attempt will be marked. (Thus if you truly wish to change your answer, completely cross out your first attempt and try again).

- If a candidate canceled his/her first attempt to answer a question and they do not supply a second attempt, then the first attempt should get marked. (This is not recommended to any students, it merely supplies comfort if you run out of time.)
- Candidates may not assume any values or answers to solve problems, it is unacceptable. (If you are unsure of any values in your exam I advise you to raise your hand and ask for assistance. Furthermore, have confidence in the values you supply as answers, be able to stand by your answer).

`EXAMPLE 1`
[From 2014 Exemplar exam Q1][11 marks]

Twelve athletes trained to run the 100m sprint event at the local athletics club trails. Some of them took their training more serious than others. The following table and scatter plot shows the number of days that n athlete trained and the time taken to run the event. The time taken, in seconds, is rounded to one decimal place.



Number of days of training	50	70	10	60	60	20	50	90	100	60	30	30
Time taken (in seconds)	12,9	13,1	17,0	11,3	18,1	16,5	14,3	11,7	10,2	12,7	17,2	14,3

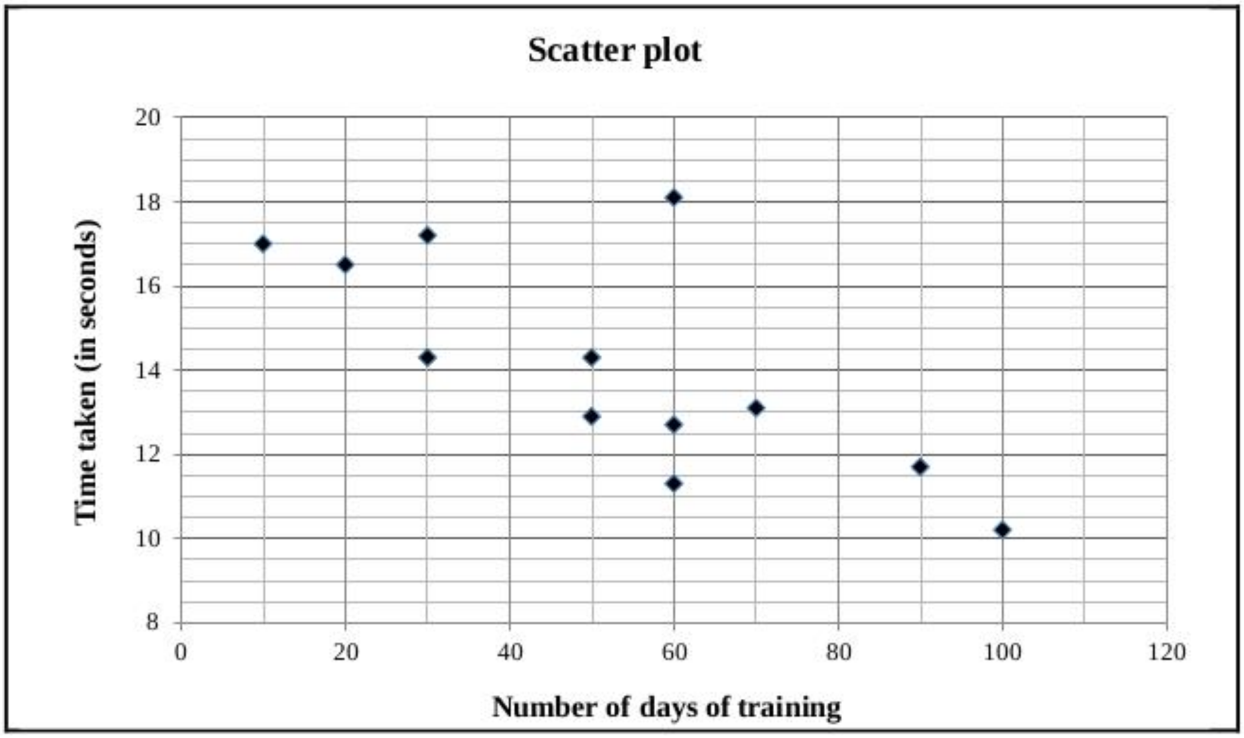


Figure 6: Question 1, paper 2, 2014 Exemplar exam

1. Discuss the trend of the data collected.

2. Identify any outlier(s) in the data.
3. Calculate the equation of the least squares regression line.
4. Predict the time taken to run the 100m sprint for an athlete training for 45 days.
5. Calculate the correlation coefficient.
6. Comment on the strength of the relationship between the variables.

EXAMPLE 2
[From 2014 Exemplar exam Q2][10 marks]

The table below shows the amount of time (in hours) that learners aged between 14 and 18 spent watching television during 3 weeks of the holiday.

Time (hours)	**Cumulative frequency**
$0 \leq t < 20$	25
$20 \leq t < 40$	69
$40 \leq t < 60$	129
$60 \leq t < 80$	157
$80 \leq t < 100$	166
$100 \leq t < 120$	172

1. Draw an ogive (cumulative frequency curve) on DIAGRAM SHEET 1 to represent the above data.
2. Write down the modal class of the data.
3. Use the ogive (cumulative frequency curve) to estimate the number of learners who watched television more than 80% of the time.
4. Estimate the mean time (in hours) that learners spent watching television during 3 weeks of the holiday.

EXAMPLE 3
[From past paper 2, February 2014, Q1][12 marks]

The tuck shop at Great Future High School sells cans of soft drinks. The Environmental Club at the school decided to have a can-collection project for the three weeks to make learners aware of the effect of litter on the environment.

The data below shows the number of cans collected on each school day of the three-week project.

	Day 1	**Day 2**	**Day 3**	**Day 4**	**Day 5**
Week 1	58	83	85	89	94
Week 2	97	98	100	105	109
Week 3	112	113	114	120	145

1. Calculate the mean number of cans collected over the three-week period.
2. Calculate the standard deviation.
3. Determine the lower and upper quartiles of the data.
4. Draw a box and whiskers diagram to represent the data (a scaled line diagram will be supplied in finals but is not necessary to practise on.)
5. On how many days did the number of cans collected lie outside ONE standard deviation of the mean?

```
EXAMPLE 4
```

[From past paper 2, February 2014 Q2][10 marks]

The histogram below shows the time, in minutes, spent by customers while shopping at Excellent Supermarket.

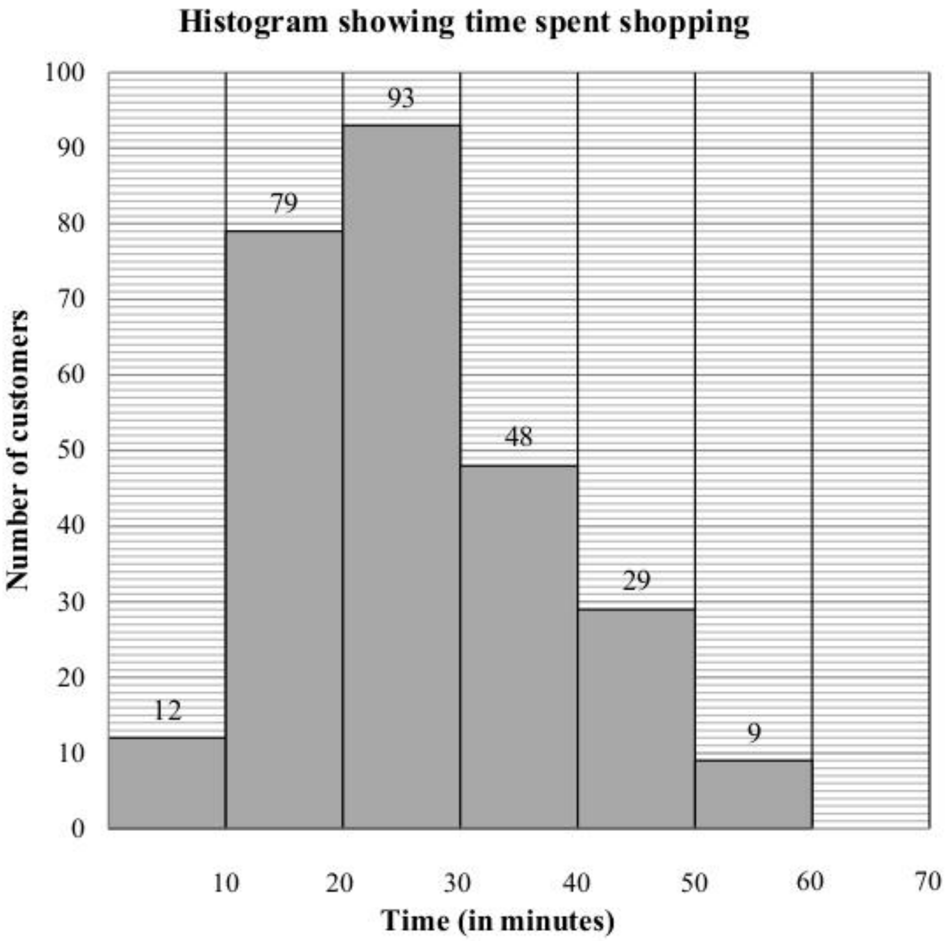


Figure 7: Question 2, paper 2, February 2014

1. Complete the frequency column and cumulative frequency column below:

2. Use the grid to draw the ogive of the above data (DIAGRAM SHEET 2).
3. Use the ogive to estimate the median time that customers spent at this supermarket.
4. Comment on the skewness of the data.

EXAMPLE 5
[From past paper 2, February 2014, Q3][6 marks]

The scatter plot below shows the age and the time for each of the first ten swimmers of a swimming club to complete n open water swimming event. The time taken is rounded to the nearest half-minute.

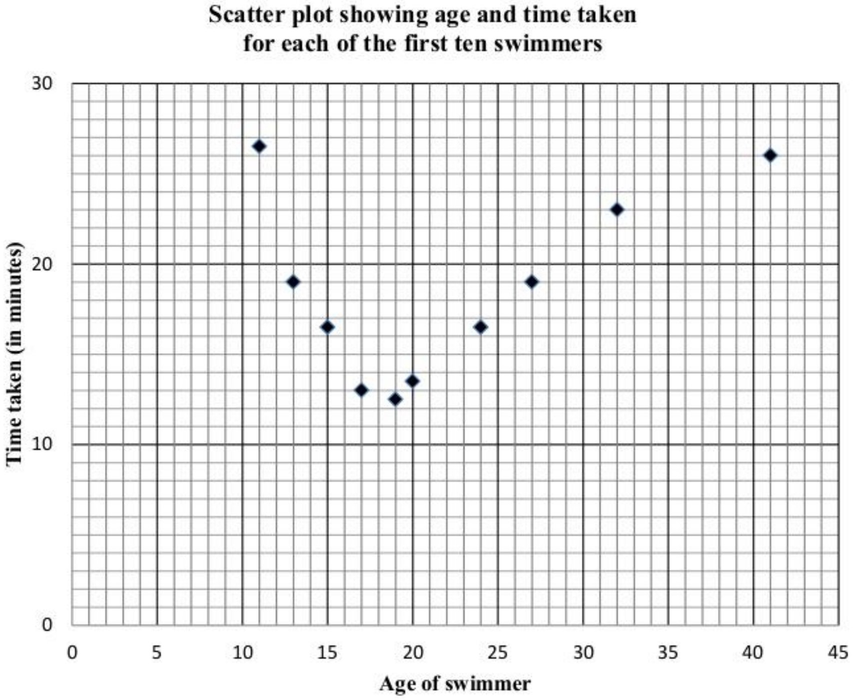


Figure 8: Question 3, paper 2, February 2014

1. Write down the coordinates of an outlier in the scatter plot.
2. Which of the following functions will best fit the data: linear, quadratic or exponential?
3. Give an explanation for the trend observed in this set of data.
4. If the two worst (longest) times are disregarded from the set of data, how will this affect the following:
   (a) The standard deviation of the original set of data.
   (b) The mean of the original set of data.

References

[1] Revision — Statistics — Siyavula. *Siyavula.com*. `https://www.siyavula.com/read/maths/grade-12/statistics/09-statistics-01`.

[2] Height and Weight Scatterplot — scatter chart made by Oli_stanford — plotly. *Plot.ly*. `https://plot.ly/ OLI_Stanford/198.embed`

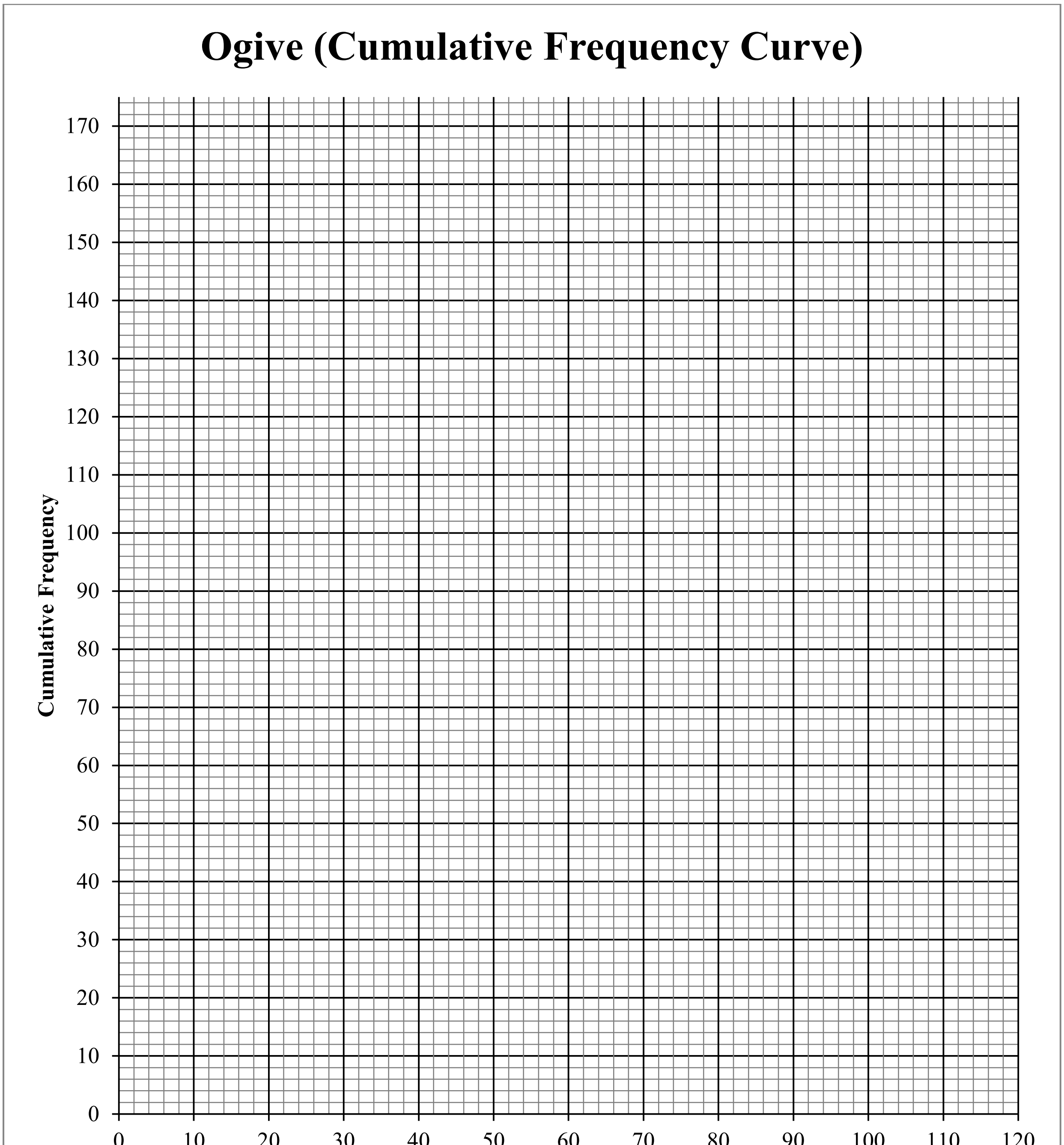
[3] Mind Action Series Mathematics 12 Textbook. *M.D. Phillips, J. Basson, C. Botha.*`ISBN 13: PRINT: 9781869214821`

NAME:

GRADE/CLASS:

DIAGRAM SHEET 1

QUESTION 2.1

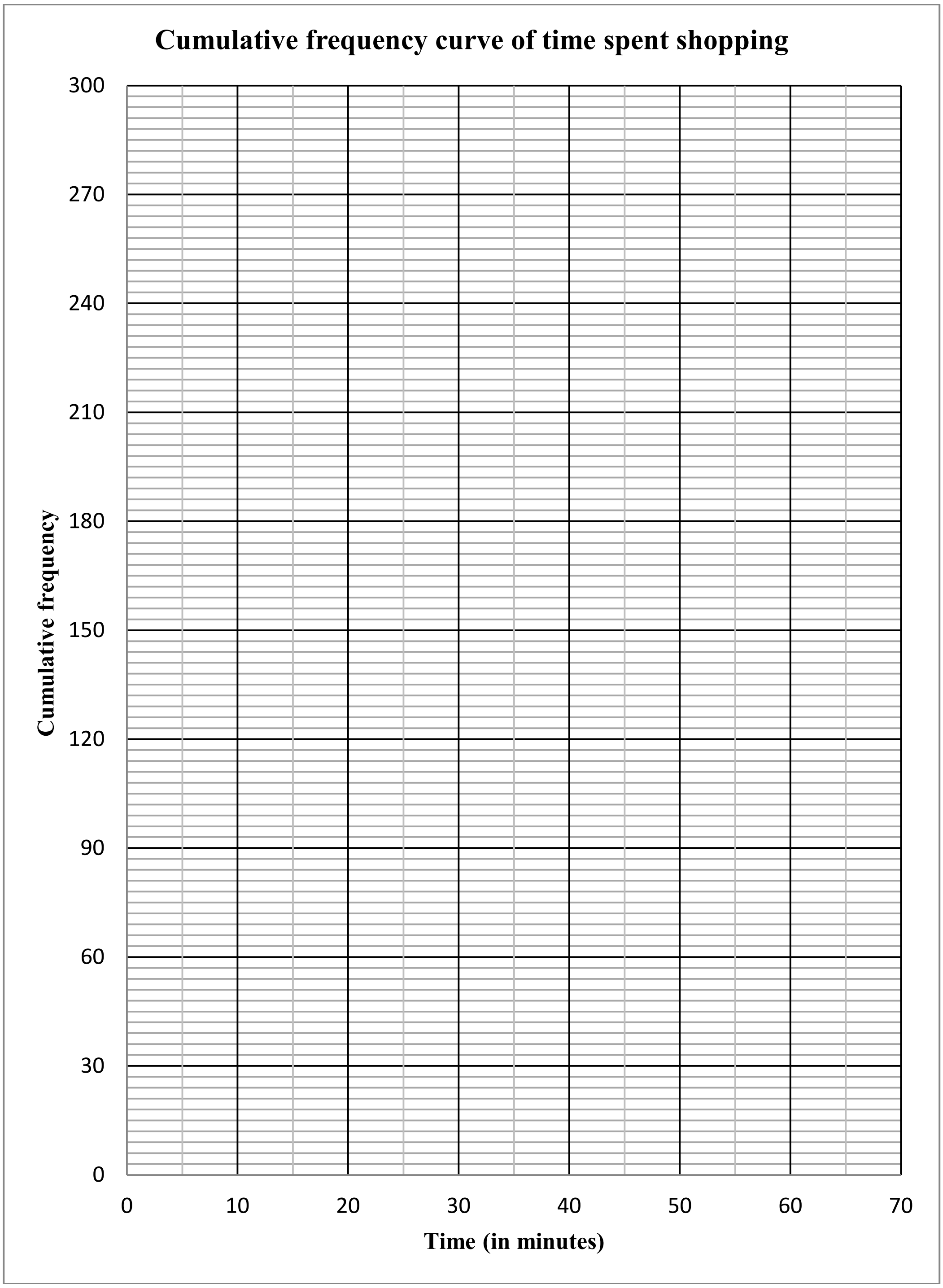


CENTRE NUMBER: □□□□□□□□

EXAMINATION NUMBER: □□□□□□□□□□□□□

DIAGRAM SHEET 2

QUESTION 2.2


Cumulative frequency curve of time spent shopping