A Guide to Advanced Statistics

Learners should continually be encouraged to critically engage with the data in terms of the data collection, analyses and interpretation. This higher level of engagement is only possible once learners have understood the basic concepts central to working with statistics. Standard deviation, mean and grouping data and constructing graphs are emphasized in the videos and tasks that accompany the series. Encourage learners to understand the calculations they are doing, rather than just applying formulae.

The concept of data being skewed or symmetrical is first revised in this series by looking at standard deviation, mean and normal distribution. Although learners have come across these concepts in Grade 11, learners tend to need a lot of revision and practice in order to further master the calculation and application of these when working in the context of grouped data and more than one set of data.

Bivariate data is also covered as well as trendlines, regression and correlation coefficients. The terms interpolation and extrapolation are also defined and explained. The lessons are all approached within real-life contexts, which are an integral part of data handling.

This is also a good section to encourage learners to study and revise using a concept map or a mind map as a learning tool for summarizing the various concepts covered. This helps learners to see how the different topics in statistics link to each other and how this in turn forms a bigger picture that helps one to analyse and interpret the data.

Video Summaries

Some videos have a 'PAUSE' moment, at which point the teacher or learner can choose to pause the video and try to answer the question posed or calculate the answer to the problem under discussion. Once the video starts again, the answer to the question or the right answer to the calculation is given.

Mindset suggests a number of ways to use the video lessons. These include:

- Watch or show a lesson as an introduction to a lesson
- Watch of show a lesson after a lesson, as a summary or as a way of adding in some interesting real-life applications or practical aspects
- Design a worksheet or set of questions about one video lesson. Then ask learners to watch a video related to the lesson and to complete the worksheet or questions, either in groups or individually
- Worksheets and questions based on video lessons can be used as short assessments or exercises
- Ask learners to watch a particular video lesson for homework (in the school library or on the website, depending on how the material is available) as preparation for the next days lesson; if desired, learners can be given specific questions to answer in preparation for the next day's lesson

1. **Working with Mean and Standard Deviation**

   In this video Zinzi helps Justice with how to use data to improve his restaurant business. They cover how to select a sample and calculate the mean and standard deviation of a data set and what this means.

2. **Normal Distribution**

   This video deals with another example of calculating the mean and standard deviation of grouped data. The lesson also deals with the graphical representation of the data in the histogram and then introduces the normal curve.

3. **The Bell Curve**

   The statistical measures of mean, mode and standard deviation are discussed in context of the normal distribution as well as the skewing of the distribution. Its appearance in industry is then presented.

4. **Working with Bivariate Data**

   This video deals with bivariate data. Scatter plots are drawn and linear relationships between variables identified. The lesson then deals with calculating an equation to best describe the relationship, this concept being regression.

5. **Working with Correlation Coefficients**

   This video studies the notion of correlation, this being the relationship between two variables, i.e. the nature of the relationship and its strength. The lesson also shows how to calculate the linear regression line/line of best fit, with the use of a calculator.

6. **How to Misuse Statistics**

   This video discusses how bias in data arises and how it can be prevented. The lesson also deals with the misrepresentation of data through the use of incorrectly drawn graphs.

Resource Material

Resource materials are a list of links available to teachers and learners to enhance their experience of the subject matter. They are not necessarily CAPS aligned and need to be used with discretion.

1. Working with Mean and Standard Deviation	http://www.mathsisfun.com/data/standard-deviation-formulas.html	A website explaining and summarizing mean and standard deviation with worked examples.
	http://video.about.com/statistics/How-to-Calculate-a-Standard-Deviation.htm	A video explaining standard deviation and mean.
2. Normal Distribution	http://www.bozemanscience.com/standard-deviation/	A video explaining normal distribution and standard deviation.
3. The Bell Curve	http://www.mathsisfun.com/data/standard-normal-distribution.html	A webpage explaining normal distribution, bell curve and standard deviation. Easy to follow and worked examples.
4. Working with Bivariate Data	http://www.sophia.org/tutorials/bivariate-data-two-variables--2	A tutorial explaining bivariate data with examples.
	http://www.mathsisfun.com/data/univariate-bivariate.html	Examples and explanation of univariate and bivariate data.
	http://www.youtube.com/watch?v=jzw4ktrwaN8	Video explaining univariate and bivariate data.
5. Working with Correlation Coefficients	http://www.mathsisfun.com/data/correlation.html	Explanation of correlation coefficients and examples
	http://mathbits.com/MathBits/TISection/Statistics2/correlation.htm	Explanation of regression and correlation coefficient.
6. How to Misuse Statistics	http://www.dummies.com/how-to/content/how-to-identify-statistical-bias.html	Explains how to identify statistical bias.
	http://websites.wnc.edu/~downs/Math120/120s13-2.pdf	Explains misuse of statistics.
	http://www.youtube.com/watch?v=VgBOcjxgzKg	A video explaining how statistics can be misleading

Task

Question 1

This table presents data from Justice's restaurant. Customers were asked to evaluate the quality of the salad dishes on his menu on a scale from 1 to 10, with 10 being the highest score in terms of quality. One hundred customers took part in this part of the survey.

Score	1	2	3	4	5	6	7	8	9	10
Frequency	0	2	4	7	29	47	8	1	2	0

1.1 Calculate the mean for this set of data.
1.2 What does this mean tell you?
1.3 Determine the standard deviation for the set of data.

Question 2

Four waiters were rated by customers on their service. The results were tabulated in this table.

Service score	**Waiter 1**	**Waiter 2**	**Waiter 3**	**Waiter 4**
[0 ; 1)	5	2	0	0
[1 ; 2)	7	3	0	0
[2 ; 3)	22	6	0	0
[3 ; 4)	11	17	8	1
[4 ; 5)	4	9	24	6
[5 ; 6)	0	7	17	11
[6 ; 7)	0	2	5	16
[7 ; 8)	0	0	3	10
[8 ; 9)	0	0	0	4
[9 ; 10)	0	0	0	0

2.1 Using these results, draw the histograms for each waiter.
2.2 State whether each graph is symmetrical or skewed.
2.3 Use the data to work out a) means and b) standard deviations for each waiter.
2.4 Use the information you have to suggest which of the waiter(s) need to improve their service.

Question 3

Justice considers moving his restaurant to a busy shopping centre. The management have told him that the rent differs depending on where he wants his restaurant. Justice wants to know if the number of people visiting a restaurant is related to the rent. So he has collected data from sixteen existing restaurants for one month. This table summarises these results.

Restaurant	**Rent (Rands)**	**No. of people**
1	4 000	420
2	6 000	610
3	5 000	490
4	4 000	405
5	5 500	555
6	4 500	480

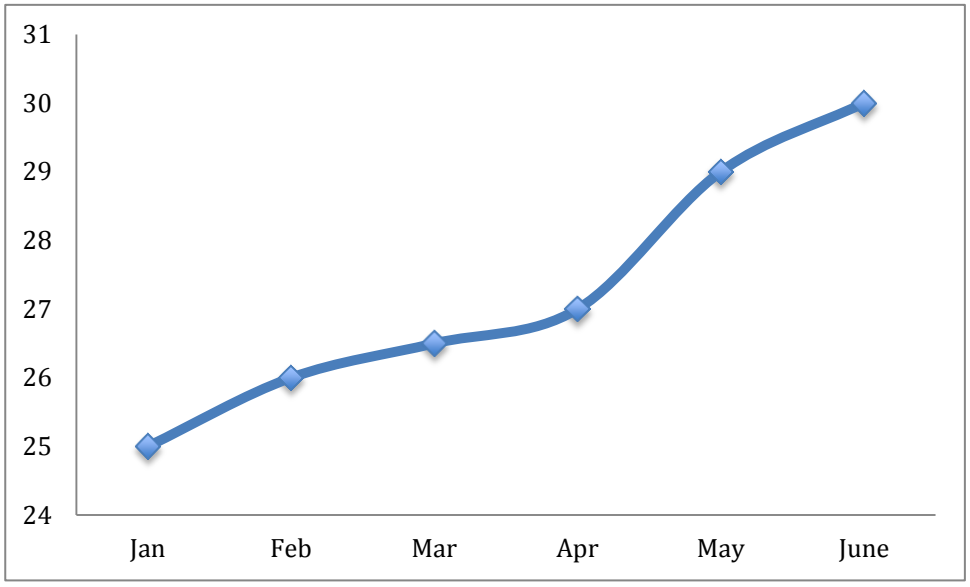
7	6 000	620
8	6 000	575
9	4 500	450
10	5 500	545
11	3 500	340
12	5 500	550
13	5 000	505
14	4 500	430
15	4 000	390
16	5 000	505

3.1 Draw a scatter plot to represent the data. Decide what values will be shown on the x-axis and what will be shown on the y-axis.
3.2 Fit a curve to the data by inspection.
3.3 Find the best-fit curve mathematically using the "least squares" formula.
3.4 Now use your calculator to find the regression line (linear curve) for this data.
3.5 Find the correlation coefficient for the regression line.
3.6 Write a short explanation for Justice about the relationship between the rent and the popularity of the restaurant.
3.7 Based on this data, estimate the number of customers that would visit a restaurant in one month where the rent is a) R5 250 per month and b) R7 000 per month. State whether you are using interpolation or extrapolation to do each estimation.

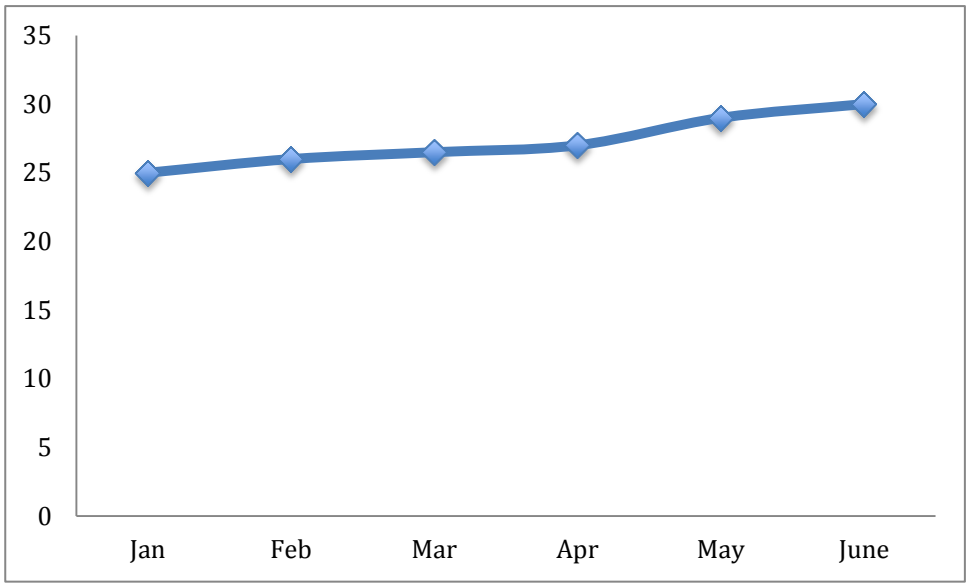
Question 4

The following two graphs represent the sale of kilograms of cement at a company during the first six months of a year. The two graphs represent the same data.

Graph A



Graph B



4.1 If the graphs represent the same data, explain why they look different?

4.2 Give a reason why you think a company may want to misrepresent data in this way.

4.3 This table provides the value used to draw the line graphs above. Use this data to construct two different bar graphs that represent the data differently.

Month	Kilograms sold
January	25
February	26
March	26,5
April	27
May	29
June	30

Task Answers

Question 1

Score	1	2	3	4	5	6	7	8	9	10
Frequency	0	2	4	7	29	47	8	1	2	0

1.1

$$\bar{x} = \frac{(2\times 2)+(4\times 3)+(7\times 4)+(29\times 5)+(47\times 6)+(8\times 7)+(1\times 8)+(2\times 9)}{100}$$

$$= \frac{4+12+28+145+282+56+8+18}{100}$$

$$= \frac{553}{100}$$

$$= 5,53$$

1.2 The mean tells us that the average rating that this sample of 100 customers gave for the salads at Justice's restaurant is 5,53. This mean indicates that the quality of the salads is "average" in that they are not too bad but not too good. So Justice can use this information to gather information on specific salads or to improve the overall quality of all the salads at his restaurant.

1.3

Score (*x*)	**Frequency (*f*)**	***fx***	***x*²**	***fx*²**
1	0	0	1	0
2	2	4	4	8
3	4	12	9	36
4	7	28	16	112
5	29	145	25	725
6	47	282	36	1692
7	8	56	49	392
8	1	8	64	64
9	2	18	81	162
10	0	0	100	0

$$\text{Variance}(s^2) = \frac{\sum fx^2}{n} - \bar{x}^2$$

$$= \frac{3191}{100} - (5,53)^2$$

$$= 31,91 - 30,5809$$

$$= 1,3291$$

$$= 1,3$$

$$\text{Standard deviation} = \sqrt{1,3291} = 1,152..$$

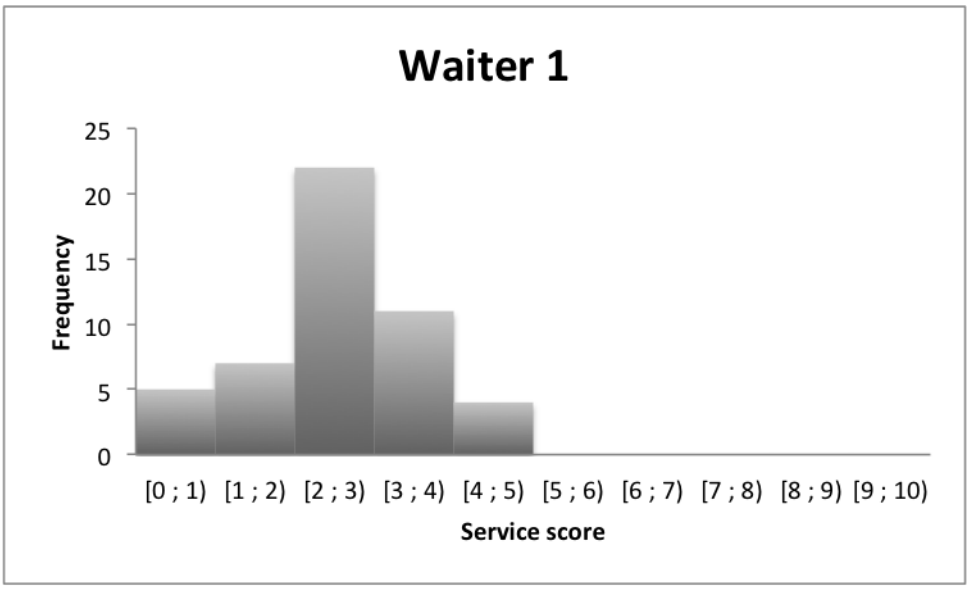
$$= 1,15$$

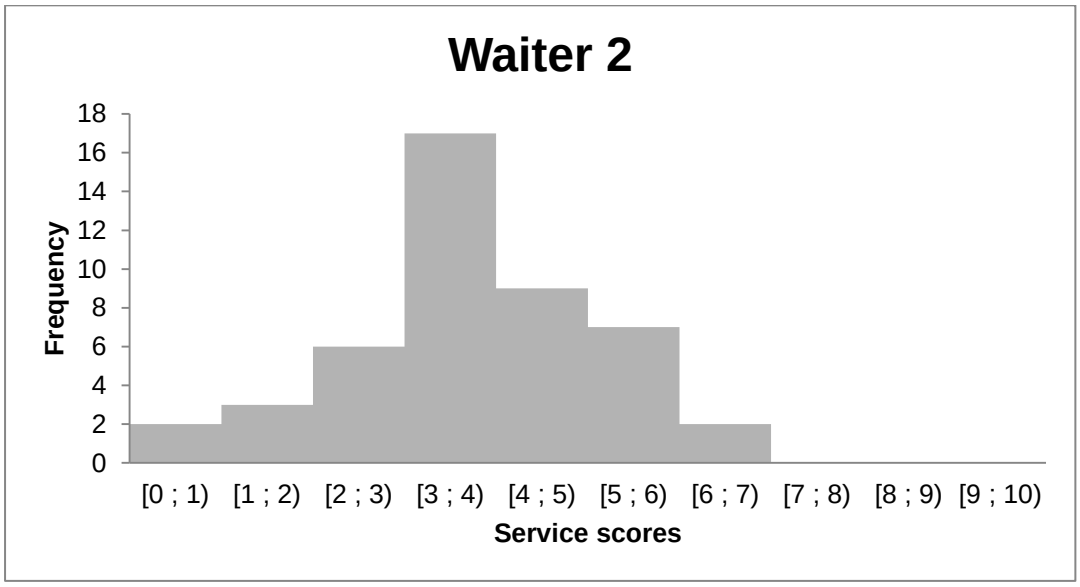
Question 2

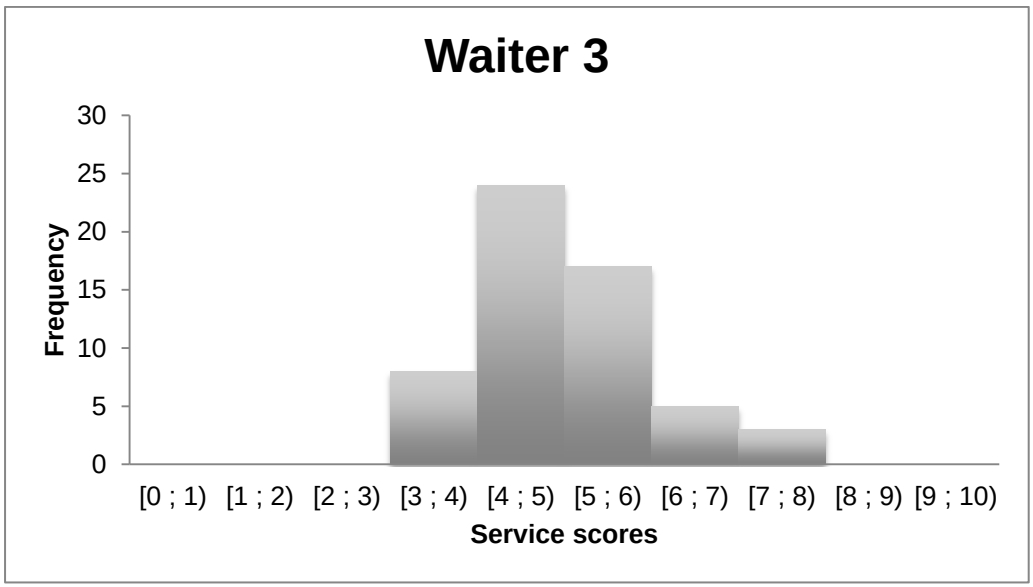
Here is the data from the survey conducted in Justice's restaurant on the level of service.

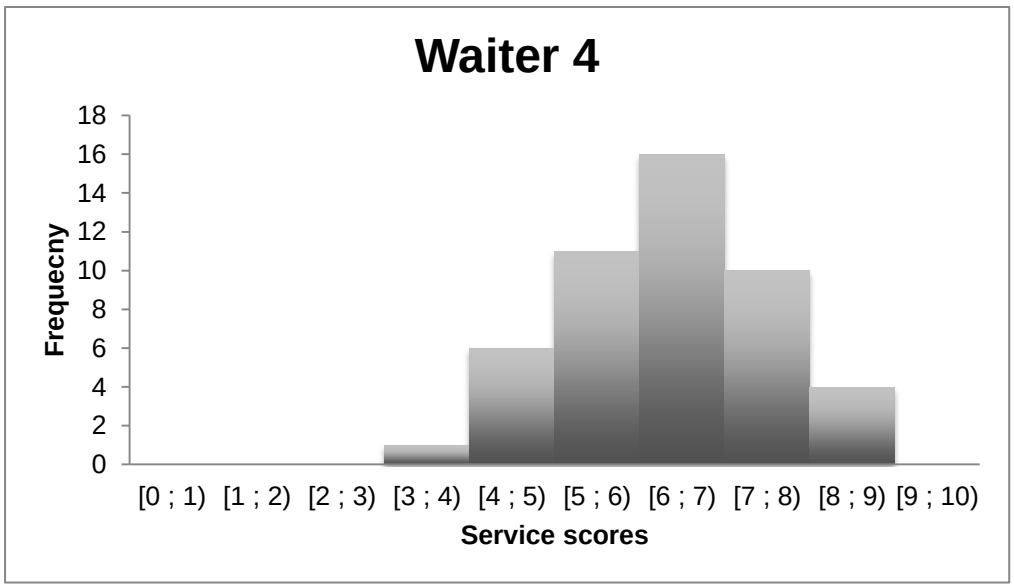
Service score	Waiter 1	Waiter 2	Waiter 3	Waiter 4
[0 ; 1)	5	2	0	0
[1 ; 2)	7	3	0	0
[2 ; 3)	22	6	0	0
[3 ; 4)	11	17	8	1
[4 ; 5)	4	9	24	6
[5 ; 6)	0	7	17	11
[6 ; 7)	0	2	5	16
[7 ; 8)	0	0	3	10
[8 ; 9)	0	0	0	4
[9 ; 10)	0	0	0	0

2.1









2.2 The histograms of Waiters 1, 2 and 4 are skewed, while the graph of Waiter 3 is more symmetrical.

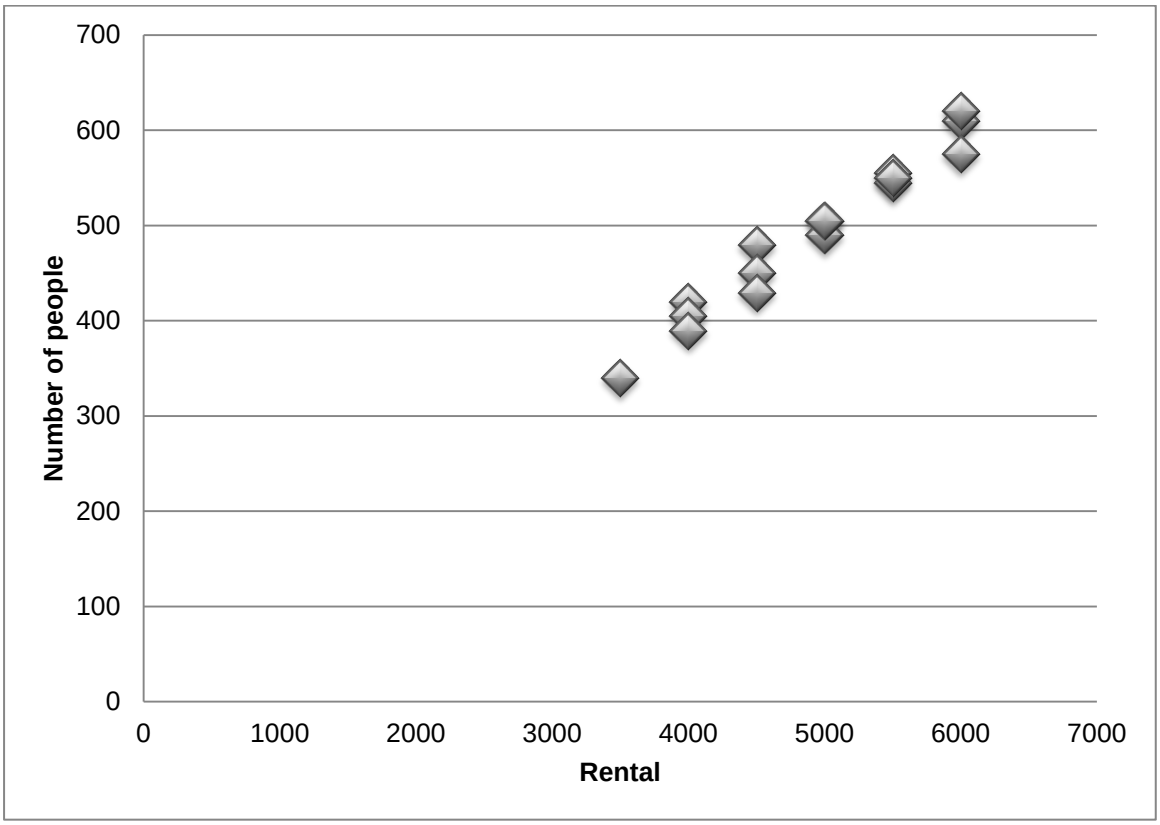
2.3 a) Waiter 1: $\bar{x} = 2,04$
Waiter 2: $\bar{x} = 3,24$
Waiter 3: $\bar{x} = 4,49$
Waiter 4: $\bar{x} = 5,83$

b) Waiter 1: $s = 15,26$
Waiter 2: $s = 17,89$
Waiter 3: $s = 34,36$
Waiter 4: $s = 33,03$

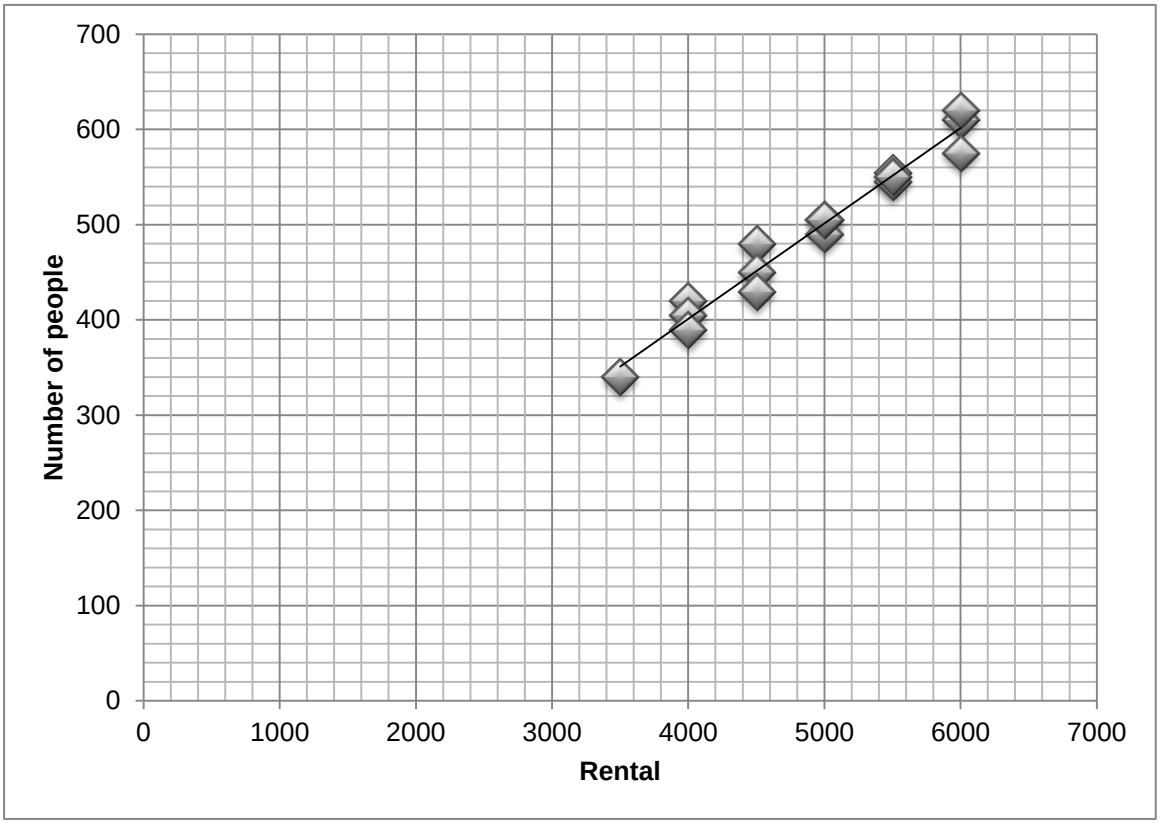
2.4 Waiters 1 and 2 need the most training according to the customer survey. The standard deviation on their scores is also the lowest indicating that in their survey the opinion of the customers vary the least from the mean. Waiters 3 and 4 have higher scores, although they would also benefit from more training. The higher standard deviation scores on their data though also indicate that the customers' opinion varied more on their service. Overall, all the service evaluated on this survey can benefit from improvement.

Question 3

3.1



3.2



3.3 .

Restaurant (*n*)	Rent (Rands) (*x*)	No. of people (*y*)	*xy*	x^2
1	4000	420	1680000	16000000
2	6000	610	3660000	36000000
3	5000	490	2450000	25000000
4	4000	405	1620000	16000000
5	5500	555	3052500	30250000
6	4500	480	2160000	20250000
7	6000	620	3720000	36000000
8	6000	575	3450000	36000000
9	4500	450	2025000	20250000
10	5500	545	2997500	30250000
11	3500	340	1190000	12250000
12	5500	550	3025000	30250000
13	5000	505	2525000	25000000
14	4500	430	1935000	20250000
15	4000	390	1560000	16000000
16	5000	505	2525000	25000000

The least squares formula:

$$A = \overline{y} - B\overline{x}$$

$$B = \frac{n\sum xy - \sum x \sum y}{n\sum (x)^2 - \left(\sum x\right)^2}$$

$$= \frac{16(39575000) - (78500)(7870)}{16(394750000) - (78500)^2}$$

$$= \frac{633200000 - 617795000}{6316000000 - 6162250000}$$

$$= \frac{15405000}{153750000}$$

$$= 0,10019...$$

$$= 0,1002$$

$$A = \overline{y} - B\overline{x}$$

$$= \frac{7870}{16} - (0,1002)\left(\frac{78500}{16}\right)$$

$$= 491,875 - 491,60625$$

$$= 0,26875$$

So $A = 0,27$

Therefore the best-fit equation is:

$y = 0,27 + 0,1x$

3.4 y = 0,29 + 0,10x

3.5 r = 0,98

3.6 As the correlation coefficient (r) is close to a value of 1, this indicates that there is a high correlation between the rent and the popularity of the restaurant.

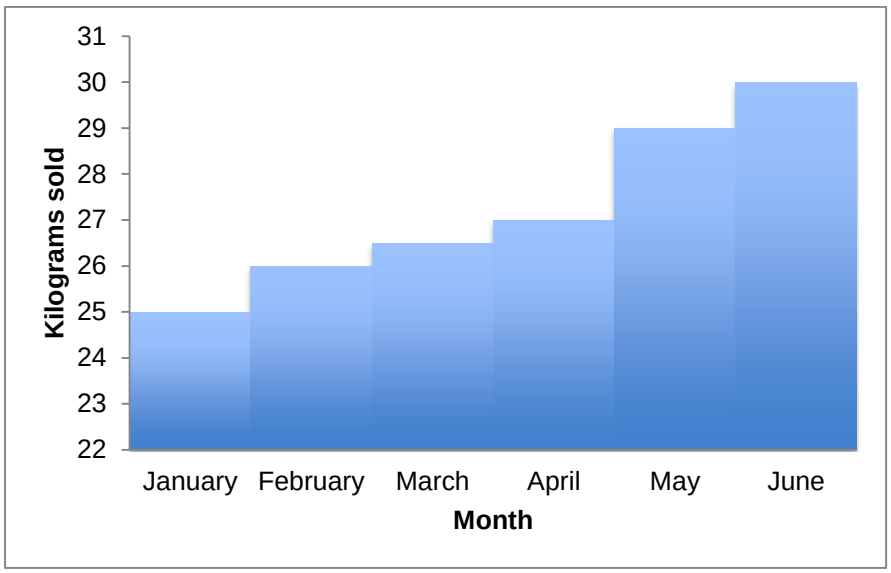
3.7 a) Interpolation; approximately 520 people
b) Extrapolation; approximately 700 people

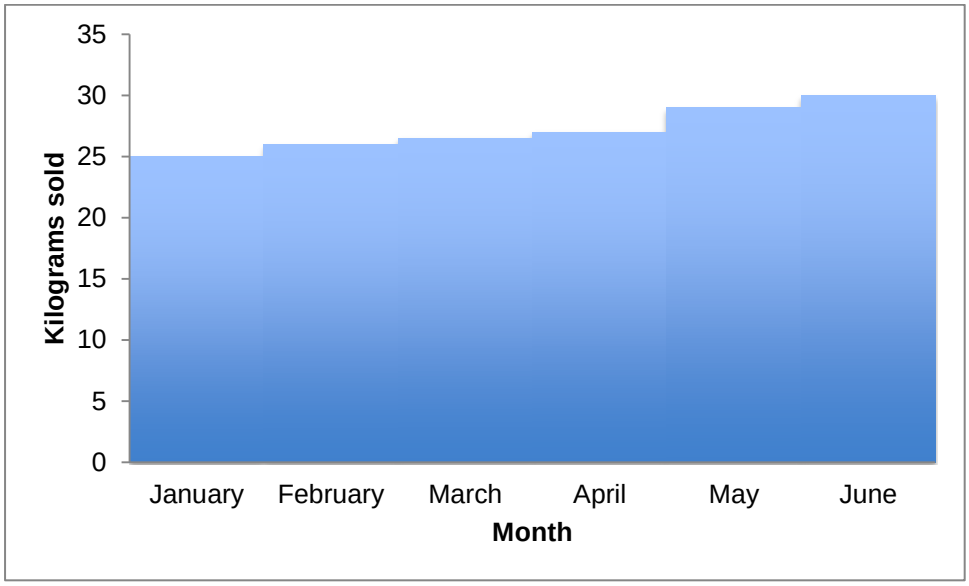
Question 4

4.1 The y-axis on the two graphs is different. Graph A the y-axis starts at 24 and the unit is 1. Graph B starts at 0 and the unit used is 5. Graph A therefore looks like a much greater increase in kilograms sold than the actual reality which is better represented by Graph B.

4.2 Perhaps they want to make the increase each month seem more drastic/higher/greater than it actually is.

4.3


31
30
29
28
27
26
25
24
23
22
Kilograms sold
January
February
March
April
May
June
Month


35
30
25
20
15
10
5
0
Kilograms sold
January
February
March
April
May
June
Month

Acknowledgements

Mindset Learn Executive Head	Dylan Busa
Content Manager Classroom Resources	Jenny Lamont
Content Coordinator Classroom Resources	Helen Robertson
Content Administrator	Agness Munthali
Content Developer	Hannah Barnes
Content Reviewer	Mlamuli Jiyane

Produced for Mindset Learn by Traffic

Facilities Coordinator	Cezanne Scheepers
Facilities Manager	Belinda Renney
Director	Alriette Gibbs
Editor	Belinda Renney
Presenter	JT Medupe
Studio Crew	Abram Tjale
Graphics	Wayne Sanderson