Solutions Manual

to accompany

Australasian Business Statistics
4th Edition

Updated by
Paul Burke & Taha Chaiechi

WILEY

Chapter 3
Descriptive summary measures

SOLUTIONS TO PROBLEMS IN CHAPTER 3

3.1 Determine the mode, median and mean of the following data.

mode =**16**, median = **16**, mean = **12.8**

3.2 The owner of a new Indian restaurant is wondering how its prices compare with others in the local area. Use the following price information on a sample of vegetarian Rogan Josh main dishes to write a short report to the restaurant owner about the central tendency of the data.

mode = $**14.50**, median = $**13.20**, mean = **$12.96**
Restaurants in the local area charge an average of $12.96 for a vegetarian Rogan Josh main dish. This means that there are vegetarian dishes on the menu that may cost less than $12.96 and there are dishes that cost more. The most frequently occurring price is $14.50. But since the mean is strongly affected by extremely low or high values, we are also interested in Median price as an alternative measure. The median price in this restaurant is the middle price in the list of sorted dish prices and it basically means 50% of the cafes charge $13.20 or less and 50% charge $13.20 or more.
This information on central tendency shows that in this case the median is higher than the mean with couple of dishes at higher end of price range ($14.50). The information conveys that the majority of vegetarian dishes with prices less than $12.96 are more frequent in the menu than pricier dishes.

3.3 A fitness consultant working with a leading rugby league team measured the height of a sample of 100 male players. The output is broken down by position in terms of summary measures associated with a sample of 50 players who predominantly play in the forwards and a sample of 50 players who predominantly play in the backs. Explain in plain language what the figures imply.

Male rugby league players within the team who play in the backs have an average height of 177.8cm (this means there are player taller than 177.8 cm and there are players who are shorter); forwards in the same team have an average height of 180.5cm. Comparing the two teams it is obvious that payers in forwards in average are taller than players in the backs. The most frequently occurring height (mode of height) is 172cm among those players who are backs and 179cm among those who play in the forwards. The median height for the backs is 172 cm, meaning 50% of the backs have a height of 175cm or less and 50% have a height of 175cm or more. Comparison between the mean and the median figures of the players height in the backs indicates that height is positively skewed this is because the mean (177.8) is greater than the median (172). 50% of the forwards have a height of 181cm or less and 50% have a height of 181cm or more. Comparison between the mean and the median figures of the player's height in the forwards also indicates that height is positively skewed.

3.4 A University has collected data to summarise the method of transportation that students use to predominantly travel to the main campus. Students were asked to respond to one category only. In cases where multiple modes were used, students were asked to indicate the method that represents the majority of their travel time.
Data were collected for each case using the above coding scheme. Of the mean, median and mode, which summary measure(s) are appropriate to describe the data? Justify your answer.

As the listed methods of transportation used by students are nominal data, the only appropriate measure of the central tendency is the mode. It shows the most frequently occurring method of transport for travel to the main campus. Other measures of the central tendency are in this case meaningless. Remember that Median and Mean can only be used if the data is quantitative. The codes are arbitrarily assigned to the categories for easy data collection and any ordering of codes or arithmetic operations performed on them is meaningless.

3.5 Compute the 20th percentile, the 60th percentile, Q_1, Q_2 and Q_3 for the following data.

Rearranging the data into ascending order:

146, 204, 280, 298, 320, 356, 445, 450, …
470, 786, 800, 820, 849, 918, 957, 964.

20th percentile is a value such that 20% of the data are equal or below the value and no more than 80% are above the value. Similarly 60th percentile is a value that 60% of the data are equal or below the value and no more than 40% are above the value.

Note n=16.

For the 20th percentile:

$$i = \frac{20}{100}(16) = 3.2$$

Since i is not a whole number, the *P*th percentile value is found by rounding i up to the next integer and reporting the value at this location. This is equivalent to the value at the location given by the whole number part of i + 1:

P_{20} is located at the 3 + 1 = 4th term

P_{20} = **298**

For the 60th percentile:

$$i = \frac{60}{100}(16) = 9.6$$

P_{60} is located at the 9 + 1 = 10th term

P_{60} = **786**

$Q_1 = P_{25}$

$$i = \frac{25}{100}(16) = 4$$

Since i is a whole number, $Q_1 = P_{25}$ is the average of the values located at the i=4th and (i+1)th =4+1=5th observations.

$Q_1 = P_{25}$ = (298+320)/2 = **309**

$Q_2 = P_{50}$= Median

The median is located at:

$$i = \frac{50}{100}(16) = 8$$

Since i is a whole number, $Q_2 = P_{50}$ is the average of the values located at the i=8th and (i+1)th =8+1=9th observations.

$Q_2 = P_{50}$ = (450+470)/2 = **460**

$Q_3 = P_{75}$

$$i = \frac{75}{100}(16) = 12$$

Since i is a whole number, $Q_3 = P_{75}$ is the average of the values located at the i=12th and (i+1)th =12+1=13th observations.

$Q_3 = P_{75}$ = (820+849)/2 = **834.5**

3.6 Compute P_{35}, P_{65}, P_{90}, Q_1, Q_2 and Q_3 for the following data.

Rearranging the data in ascending order:

10,11,11,12,13,16,18,18,21,21,22,23,23,27,29,29,30,30,30,41,42,42,46.

n = 23

For the 35th percentile:

$$i = \frac{35}{100}(23) = 8.05$$

P_{35} is located at the 8 + 1 = 9th term
P_{35} = **21**

For the 65th percentile:

$$i = \frac{65}{100}(23) = 14.95$$

P_{65} is located at the 14 + 1 = 15th term
P_{65} = **29**

For the 90th percentile:

$$i = \frac{90}{100}(23) = 20.7$$

P_{90} is located at the 20 + 1 = 21st term
P_{90} = **42**

$Q_1 = P_{25}$

$$i = \frac{25}{100}(23) = 5.75$$

Q_1 is located at the 5+1 = 6th term
Q_1 = **16**

Q_2 = Median
The median is located at the:

$$i = \frac{50}{100}(23) = 11.5$$

Q_2 is located at the 11+1 = 12th term
Q_2 = **23**

$Q_3 = P_{75}$

$$i = \frac{75}{100}(23) = 17.25$$

Q_3 is located at the 17+1=18th term
Q_3 = **30**

3.7 A hairdresser franchisor is concerned about the time taken by staff to complete a standard haircut for male customers at one of its newly opened stores. The franchisor decides to visit the store and record the time taken for such a category of haircut on 30 random occasions. A benchmark of 20 minutes has been set as a reasonable objective based on the franchisors' experience at their other stores. Interpret the following output to help the franchisor understand the time taken to provide male customers with a standard haircut in relation to this benchmark.

The benchmark of 20 minutes has been reached on at least 75% of the occasions sampled. A quarter of haircuts sampled took 10.5 minutes or less and 10% of haircuts took 8.2 minutes or less. Further, half of the haircuts took 12.5 minutes or less and 75% of haircuts took 19.7 minutes or less. Finally, the 90th percentile indicates that 90% of haircuts took 32 minutes or less, whilst 10% took 32 minutes or more.

The output seems satisfactory as it shows that at least three quarters of haircuts (minimum of 22 out of 30) met the benchmark of 20 minutes time-wise. Q2 shows that 50% of the number of haircuts took 12.5 minutes or less, Q3 shows that 75% of the number of haircuts took 19.7 minutes or less. Only one quarter of haircuts exceeded the expected time and from that 25% (around 7 haircuts) only 3 haircuts took longer than 32 minutes.

3.8 A data set contains the following seven values.
6 2 4 7 8 3 5
a. Calculate the range.
b. Calculate the population variance.
c. Calculate the population standard deviation.
d. Calculate the interquartile range.
e. Calculate the z-score for each value.
f. Calculate the coefficient of variation.

x	x -μ	(x-μ)^2
6	6-5 = 1	1
2	-3	9
4	-1	1
7	2	4
8	3	9
3	-2	4
5	0	0
Σx = 35	Σ│x-μ│ = 12	Σ(x -μ)^2 = 28

$$\mu = \frac{\Sigma x}{N} = \frac{35}{7} = 5$$

(a) Range = 8 - 2 = **6**

(b) $\sigma^2 = \dfrac{\Sigma(x-\mu)^2}{N} = \dfrac{28}{7} = \mathbf{4}$

(c) σ = $\sqrt{\frac{\Sigma(x-\mu)^2}{N}} = \sqrt{4}$ = **2**

(d) 2, 3, 4, 5, 6, 7, 8

$Q_1 = P_{25}$

i = $\frac{25}{100}(7)$ = 1.75

Q_1 is located at the 1 + 1 = 2th term, Q1 = 3

$Q_3 = P_{75}$:

i = $\frac{75}{100}(7)$ = 5.25

Q_3 is located at the 5 + 1 = 6th term, Q3 = 7

IQR = $Q_3 - Q_1$ = 7 - 3 = **4**

(e) z = $\frac{6-5}{2}$ = **0.5**

z = $\frac{2-5}{2}$ = **-1.5**

z = $\frac{4-5}{2}$ = **-0.5**

z = $\frac{7-5}{2}$ = **1**

z = $\frac{8-5}{2}$ = **1.5**

z = $\frac{3-5}{2}$ = **-1**

z = $\frac{5-5}{2}$ = **0**

(f) $CV = \frac{\sigma}{\mu} \times 100 = \frac{2}{5} \times 100 = \mathbf{40\%}$

3.9 A data set contains the following eight values.
4 3 0 5 2 9 4 5
a. Calculate the range.
b. Calculate the sample variance.
c. Calculate the sample standard deviation.
d. Calculate the interquartile range.
e. Calculate the coefficient of variation.

x	$(x-\bar{x})$	$(x-\bar{x})^2$
4	0	0
3	-1	1
0	-4	16
5	1	1
2	-2	4
9	5	25
4	0	0
5	1	1
Σx = 32		$\Sigma(x-\bar{x})^2$ = 48

$$\bar{x} = \frac{\Sigma x}{n} = \frac{32}{8} = 4$$

(a) Range = 9 - 0 = **9**

(b) $s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1} = \frac{48}{7} = \mathbf{6.857}$

(c) $s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} = \sqrt{6.857} = \mathbf{2.619}$

(d) Numbers in order: 0, 2, 3, 4, 4, 5, 5, 9

$Q_1 = P_{25}$

$i = \frac{25}{100}(8) = 2$

Q_1 is located at the average of the 2nd and 3rd terms,

$Q_1 = \frac{(2+3)}{2} = \mathbf{2.5}$

$Q_3 = P_{75}$

$$i = \frac{75}{100}(8) = 6$$

Q_3 is located at the average of the 6th and 7th terms

$Q_3 = (5 + 5)/2 = \mathbf{5}$

$IQR = Q_3 - Q_1 = 5 - 2.5 = \mathbf{2.5}$

(e) $CV = \frac{s}{\bar{x}} \times 100 = \frac{2.619}{4} \times 100 \approx \mathbf{65.5\%}$

3.10 According to Chebyshev's theorem, at least what proportion of the data is within $\mu \pm k\sigma$ for each of the following values of k?

a. 2
b. 2.5
c. 1.6
d. 3.2

(a) $1 - \frac{1}{2^2} = 1 - \frac{1}{4} = \frac{3}{4} = .75$

(b) $1 - \frac{1}{2.5^2} = 1 - \frac{1}{6.25} = .84$

(c) $1 - \frac{1}{1.6^2} = 1 - \frac{1}{2.56} = .609$

(d) $1 - \frac{1}{3.2^2} = 1 - \frac{1}{10.24} = .902$

3.11 A car fleet manager working for a local council is thinking of gradually replacing the current fleet of vehicles used by the council with vehicles that use LPG (gas) rather than unleaded petrol. The concern is not so much the average price of petrol but rather the variability in price that occurs, as this becomes problematic for budgeting and managing reimbursements to employees. The fleet manager will upgrade the fleet to LPG-powered vehicles so long as the variability in the LPG price is lower than that of unleaded petrol. The fleet manager uses a website that collates data on fuel prices in the council region to produce the following summary statistics based on a random sample of prices drawn from the past year.
Interpret the output and comment on the variability observed for the price of each fuel. If you make this comparison using the standard deviation for each fuel type, what conclusion can you reach? Suppose the fleet manager tells you

that they prefer to compare variability relative to the size of each fuel's mean price. Make a recommendation to the fleet manager about which vehicles should be used in the upgrade.

In the local council region unleaded petrol costs an average of 150.43 cents per litre with a standard deviation of 6.19 cents per litre, and LPG costs an average of 79.82 cents per litre with a standard deviation of 5.44 cents per litre. When comparing the standard deviations only, the price of unleaded petrol has a greater variation than LPG. Therefore, upgrading the fleet to LPG powered vehicles should be recommended.

Unleaded petrol: $CV = \frac{s}{\bar{x}} \times 100 = \frac{6.19}{150.43} \times 100 = 4.11\%$

LPG: $CV = \frac{s}{\bar{x}} \times 100 = \frac{5.44}{79.82} \times 100 = 6.82\%$

However, if the fleet manager wants to make a decision based on the comparison of the variability of the fuel price relative to the mean price, they should be recommended not to replace the current fleet with LPG-powered vehicles, because the coefficient of variation for the unleaded petrol is lower than the coefficient of variation for the LPG.

3.12 A wine industry association reported in its magazine that a particular wine was being marketed by online wine distributors with an average market price of $125 and standard deviation of $12, with the distribution of prices being approximately bell shaped. One boutique wine distributor is concerned by this report as they are charging $50 per bottle for this particular wine. Between what two price points would approximately 68% of prices fall? Between what two numbers would 95% of the prices fall? Between what two values would 99.7% of the prices fall? Write a short report informing the distributor whether the current price being charged is comparable to others.

Approximately 68% of the online distributors of this particular wine charge between $113 and $137 per bottle, approximately 95% charge between $101 and $149 and approximately 99.7% charge between $89 and $161. Nearly all of the online distributors of this particular wine charge more than this boutique wine distributor, who charges $50 per bottle.

3.13 An employment agency is concerned that some of its clients for whom it has found part-time work are not receiving enough hours of employment. It examines a sample of clients and asks them to report how many hours they had worked in the last month. The agency notes that the data are not normally distributed.

If the mean hours worked is 38 and the standard deviation is 6 hours, what proportion of values would fall between 26 hours and 50 hours? What proportion of values would fall between 14 hours and 62 hours? Between what two values would 89% of the values fall? Explain your findings in simple terms to the employment agency's management team.

$\bar{x} = 38, s = 6$

Since the distribution is not normal, Chebyshev's theorem applies.

Both observations, 26 and 50, are two standard deviations from the mean.

$$1 - \frac{1}{k^2} = 1 - \frac{1}{2^2} = \mathbf{0.75}$$

At least 75% of the clients of this employment agency, for whom the agency has found part-time work, had worked in the last month between 26 and 50 hours.

Both observations, 14 and 62, are four standard deviations from the mean.

$$1 - \frac{1}{k^2} = 1 - \frac{1}{4^2} = \mathbf{0.9375}$$

At least 93.75% of the clients of this employment agency, for whom the agency has found part-time work, had worked in the last month for between 14 and 62 hours.

$$1 - \frac{1}{k^2} = 0.89 \Rightarrow k = \pm 3.015$$

At least 89% of the values are within $38 \pm (3.015)(6)$.

At least 89% of the clients of this employment agency, for whom it has found part-time work, had worked in the last month for between 19.91 and 56.09 hours.

3.14 Baycoast City Real Estate Agents has been approached by a government organisation to examine rental affordability in the Baycoast area. The organisation seeks information on the weekly rental price, particularly information about the average weekly rental for a house and the range of rental amounts that defines where 95% of rentals will fall. Using the Baycoast data set, calculate the necessary statistics and write a short report to the government organisation in answer to their question (assuming the data are bell shaped).

Prospective tenants in the Baycoast area can expect to rent a house for an average of $604 per week. The standard deviation of the weekly rent is $226. 95% of the houses cost between $152 and $1056 per week to rent.

3.15 The size in square metres of properties sold by a major real estate firm in Australia in the last year was analysed using descriptive summary measures. The mean size of a one-bedroom residential apartment sold by this firm was 60

square metres, the median was 55 square metres, and the standard deviation was12 square metres. Compute the value of the Pearsonian coefficient of skewness and interpret the result.

Remember that this coefficient compares the mean and median in regards with the magnitude of the standard deviation. Note that, if the distribution is symmetrical, the mean and median are the same value and hence the coefficient of skewness is equal. (Page 76)

$$S_k = \frac{3(\mu - M_d)}{\sigma} = \frac{3(60-55)}{12} = \mathbf{1.25}$$

Positive values indicate the positively skewed data and negative coefficient values indicate negatively skewed data. The coefficient here is 1.25 indicates that the distribution is positively skewed. That is, the distribution is skewed to the right.

3.16 A survey of drivers asked respondents to list the age in years of the vehicle that they predominantly drive. The following data represent a sample of 18 responses provided. Use the data to construct a box and whisker plot. List the median, *Q_1*, *Q_3*, the endpoints of the inner fences and the endpoints of the outer fences. Are any outliers present in the data?

$n = 18$ $Q_1 = P_{25}$:

$$i = \frac{25}{100}(18) = 4.5$$

Q_1 = 4+1 = 5th term = **3**

$Q_3 = P_{75}$:

$$i = \frac{75}{100}(18) = 13.5$$

Q_3 = 14th term = **9**

$$\text{Median: } \frac{(n+1)^{th}}{2} = \frac{(18+1)^{th}}{2} = \frac{19}{2}^{th} = 9.5^{\text{th}} \text{ term}$$

Median = average of 9th and 10th term = **5**

IQR = Q_3 - Q_1 = –9 – 3 = **6**

Inner Fences: Q_1 - 1.5 IQR = 3 - 1.5 (6) = **-6**

Q_3 + 1.5 IQR = 9 + 1.5 (6) = **18**

Outer Fences: Q_1 - 3.0 IQR = 3 - 3.0 (6) = **-15**

Q_3 + 3.0 IQR = 9 + 3.0 (6) = **27**

Remember that as a distance of 1.5 IQR outward from the lower and upper quartiles is what is referred to as inner fences. Outer fences are constructed based on a distance that is 3.0 IQR from the lower and upper quartiles. In this example, we can see that there are no outliers identified as all the reported numbers are inside these boundaries. There are no extreme outliers nor mild outliers.

3.17 An online retailer that sells board games and puzzles has produced summary measures, shown in the right-hand column, describing the cost charged to consumers for shipping. Write a short description of the data incorporating a discussion of symmetry and skewness to inform the retail owners whether the data related to shipping charges are bell shaped and how this is reflected in the measures of central tendency, particularly the median and mean.

The distribution of the shipping charges is positively skewed (with a coefficient of skewness of 1.856). This implies a greater number of observations occur in the right tail of the distribution relative to a normal distribution. This has an effect on the relative positions of the mean and the median (mean is greater than the median). The mean is located towards the tail of the distribution, drawn towards the upper end of the distribution and its value is $10.53, while the median is only $9.80. The distribution is also leptokurtic (with the coefficient of kurtosis being positive and equal to 3.002). This indicates that the shipping charges have a higher frequency of values nearer to the mean relative to a normal distribution, a relatively peaked distribution.

3.18 A manufacturer of solar power systems is doing some comparative testing of two differently designed 1.9 kWh (10 module) systems. On the basis of a sample of 52 observations, both systems, on average, produce 8.18 kWh per day. The company is seeking to further develop the system that appears to produce a fairly stable amount of output above and below this measure of central tendency. System B was rejected as it displayed too much variation in output relative to system A and did so with noticeable negative skewness. Based on the data and the plots below, do you agree with this decision?

A correct decision has been made, since system B displayed greater variation in the power output than system A. The IQR of system B is more than double the IQR of system A and the range of the power output of system B is double the range of the power output of system A. Therefore system B should be rejected as it is less consistent than system A.

3.19 Determine the value of the coefficient of correlation, *r*, for the following data.

$\Sigma x = 80$ $\quad\Sigma x^2 = 1{,}148$ $\quad\Sigma y = 69$
$\Sigma y^2 = 815$ $\quad\Sigma xy = 624$ $\quad n = 7$

$$\mathrm{r} \;=\; \frac{\sum xy - \dfrac{\sum x \sum y}{n}}{\sqrt{\left[\sum x^2 - \dfrac{(\sum x)^2}{n}\right]\left[\sum y^2 - \dfrac{(\sum y)^2}{n}\right]}} \;=$$

$$\mathrm{r} \;=\; \frac{624 - \dfrac{(80)(69)}{7}}{\sqrt{\left[1{,}148 - \dfrac{(80)^2}{7}\right]\left[815 - \dfrac{(69)^2}{7}\right]}} \;=\; \frac{-164.571}{\sqrt{(233.714)(134.857)}} \;=$$

$$\mathrm{r} \;=\; \frac{-164.571}{177.533} \;=\; \mathbf{-0.927}$$

3.20 Determine the value of *r* for the following data.

$\Sigma x = 1{,}087$ $\quad\Sigma x^2 = 322{,}345$ $\quad\Sigma y = 2{,}032$
$\Sigma y^2 = 878{,}686$ $\quad\Sigma xy = 507{,}509$ $\quad n = 5$

$$\mathrm{r} \;=\; \frac{\sum xy - \dfrac{\sum x \sum y}{n}}{\sqrt{\left[\sum x^2 - \dfrac{(\sum x)^2}{n}\right]\left[\sum y^2 - \dfrac{(\sum y)^2}{n}\right]}} \;=$$

$$\mathrm{r} \;=\; \frac{507{,}509 - \dfrac{(1{,}087)(2{,}032)}{5}}{\sqrt{\left[322{,}345 - \dfrac{(1{,}087)^2}{5}\right]\left[878{,}686 - \dfrac{(2{,}032)^2}{5}\right]}} \;=$$

$$r = \frac{65{,}752.2}{\sqrt{(86{,}031.2)(52{,}881.2)}} = \frac{65{,}752.2}{67{,}449.5} = \mathbf{.975}$$

3.21 The following data are the selling prices of houses (in $000) and the land size (in square metres) for an outer suburb of Sydney. Use the data to compute a correlation coefficient, r, to determine the correlation between house price and the size of the land. Interpret your results.

$\Sigma x = 6384.0$ $\Sigma x^2 = 3826478.0$

$\Sigma y = 7149.0$ $\Sigma y^2 = 4834005.0$

$\Sigma xy = 4248280.0$ $n = 11$

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{n}\right]\left[\sum y^2 - \frac{(\sum y)^2}{n}\right]}} =$$

$$r = \frac{4248280.0 - \frac{(6384.0)(7149.0)}{11}}{\sqrt{\left[3826478.0 - \frac{(6384.0)^2}{11}\right]\left[4834005.0 - \frac{(7149.0)^2}{11}\right]}} = \mathbf{.657}$$

3.22 The chief financial officer (CFO) for the producer of a well-known fashion line has recently come under fire for suggesting that retailers should hire younger people as frontline employees to improve sales. A human resource manager working for a retailer that stocks this fashion line is able to collate data gathered by an independent survey company for a sample of employees on a number of variables. These include the weekly sales the employee achieves, their age, years employed by the company and an average rating of friendliness on a scale of 1 to 10 (10 = most friendly). The human resource manager generates output in Excel that shows the correlation between these variables (see overleaf).

Use the information to write a response to the CFO stating whether the analysis supports their assertion. You may like to comment on associations between other variables that appear in this output, such as whether people who have worked with the retailer for a longer period are friendlier or generate better sales.

According to the data gathered by an independent survey company, there is no evident relationship between the age and the weekly sales an employee achieves

(the correlation coefficient between these two variables is -0.024). Therefore, this analysis does not support the idea of hiring young people in order to improve sales. However the correlation coefficient between friendliness and sales is 0.866 which represents a strong positive correlation. This indicates that friendlier sales employees generate greater weekly sales.
There is also a moderate positive correlation of 0.441 between the number of years employed by the company and weekly sales, which indicates that sales people who have worked with the retailer for a longer period achieve better weekly sales.

3.23 The CEO of Combaro Ltd is interested in seeing whether employees who are paid more have a greater level of job satisfaction and take fewer days off. Using the data set provided, examine whether this is the case.

	DysAbsnt	*JobSat*	*WkSalry*
DysAbsnt	1		
JobSat	0.187035	1	
WkSalry	0.148087	0.102221	1

As the coefficient of correlation between weekly salary and job satisfaction is 0.102, there is very weak evidence that the employees who are paid more have a greater level of job satisfaction. However there is no evidence to support the idea that the employees who are paid more take fewer days off, as the correlation coefficient between these two variables would have to be negative in order to support this idea, while in fact it is positive 0.148.

3.24 An investor in real estate wonders whether there is any evidence to support her belief that older houses tend to sell for less, while newer houses can command higher selling prices. Using the Baycoast City Real Estate Agents data set of recent house sales, examine whether there is an association between the selling price of a house and the age of the house in years. Write a short interpretation of this result for the real estate investor.

	Price($'000)	*Age*
Price($'000)	1	
Age	-0.363	1

The correlation coefficient between the age of the house and its selling price is -0.363, which supports the investor's belief that older houses tend to sell for less, while the newer houses can command higher prices.

3.25 The Australian Census of Population and Housing asks every household to report information on each person living in the house. Suppose that, for a sample of 30 households, the number of persons living in each was reported as follows. Compute the mean, median, mode, range, lower and upper quartiles,

and interquartile range for these data and interpret them in a brief plain-language report.

1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 4, 5, 6, 8

Mean: $\bar{x} = \frac{\Sigma x}{n} = \frac{76}{30} = \mathbf{2.53}$

Mode = **2** (There are eleven 2's)

Median: There are n = 30 terms.

The median is located at $\frac{n+1}{2}^{th} = \frac{30+1}{2} = \frac{31}{2}$ = 15.5th position.

Median is the average of the 15th and 16th value.

However, since these are both 2, the median is **2**.

Range = 8 - 1 = **7**

$Q_1 = P_{25}$:

$$i = \frac{25}{100}(30) = 7.5$$

Q_1 is the 8th term = **1**

$Q_3 = P_{75}$:

$$i = \frac{75}{100}(30) = 22.5$$

Q_3 is the 23rd term = **3**

IQR = $Q_3 - Q_1$ = 3 - 1 = **2**

3.26 The Australian Census of Population and Housing asks for each resident's age. Suppose that a sample of 40 households taken from the census data showed the age of the first person recorded on the census form as follows. Compute P_{10}, P_{80}, Q_1, Q_3, the interquartile range and the range of these data.

P_{10}:

$$i = \frac{10}{100}(40) = 4$$

P_{10} = 4.5th term = **23**

P_{80}:

$$i = \frac{80}{100}(40) = 32$$

P_{80} = 32.5th term = **49.5**

$Q_1 = P_{25}$:

$$i = \frac{25}{100}(40) = 10$$

P_{25} = 10.5th term = **27.5**

$Q_3 = P_{75}$:

$$i = \frac{75}{100}(40) = 30$$

P_{75} = 30.5th term = **47.5**

IQR = $Q_3 - Q_1$ = 47.5 - 27.5 = **20**

Range = 81 - 19 = **62**

3.27 Determine the Pearson product–moment correlation coefficient for the following data.

$\Sigma x = 24$	$\Sigma x^2 = 94$
$\Sigma y = 430$	$\Sigma y^2 = 31068$
$\Sigma xy = 1316$	$n = 7$

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{n}\right]\left[\sum y^2 - \frac{(\sum y)^2}{n}\right]}} =$$

$$r = = \frac{1316 - \frac{(24)(430)}{7}}{\sqrt{\left[94 - \frac{(24)^2}{7}\right]\left[31068 - \frac{(430)^2}{7}\right]}}$$

$$r = = \frac{-158.286}{\sqrt{(11.71429)(4653.714)}} \quad \frac{-158.286}{233.4843} = \mathbf{-.678}$$

Remember that this correlation coefficient ranges from -1 to + 1representing the direction and relative strength of the linear relationship between the variables. Value of +1 denotes a perfect (linear) positive relationship and value of -1 denotes a perfect (linear) negative relationship and value of 0 means no linear relationship is present between the two variables. In this case r= -0.678 which indicates there is an imperfect negative relationship between the variables, indicating as one variable gets larger the other gets smaller.

3.28 Financial analysts like to use the standard deviation as a measure of risk for a stock. The greater the deviation in a stock price over time, the more risky it is to invest in the stock. However, the average prices of some stocks are considerably higher than the average prices of others, allowing for the potential of a greater standard deviation of price. For example, a standard deviation of $5.00 on a $10.00 stock is considerably different from a $5.00 standard deviation on a $40.00 stock. In this situation, a coefficient of variation might provide insight into risk. Suppose stock X costs an average of $13.21 per share and has shown a standard deviation of $5.28 for the past 30 days. Suppose stock Y costs an average of $2.52 per share and has shown a standard deviation of $0.50 for the past 30 days. Use the coefficient of variation to determine the variability for each stock. Based on the coefficient of variation, which is the riskier stock?

$$CV_X = \frac{\sigma_x}{\mu_x}(100\%) = \frac{5.28}{13.21}(100\%) = \mathbf{39.97\%}$$

$$CV_Y = \frac{\sigma_y}{\mu_y}(100\%) = \frac{0.5}{2.52}(100\%) = \mathbf{19.84\%}$$

Many investors use this coefficient to determine the risk involved with the expected return from investment, the lower the ratio the less expected hence lower risk. Based on the coefficient of variation in this example, stock X has a greater relative variability. From this perspective, stock X is riskier.

3.29 An NRMA report stated that the average age of a car in Australia is 10.5 years. Suppose the distribution of ages of cars on Australian roads is approximately bell shaped. If the standard deviation is 2.4 years, between what two values would 95% of the car ages fall?

$\mu = 10$

From the Empirical Rule: 99.7% of the values lie in $\mu \underline{+} 3\sigma = 10 \underline{+} 3\sigma$

$6\sigma = 20 - 1 = 19$

$\boldsymbol{\sigma = 3.17}$

suppose that $\mu = 10.5$, $\sigma = 2.4$:

95% lie within $\mu \underline{+} 2\sigma = 10.5 \underline{+} 2(2.4) = 10.5 \underline{+} 4.8$
Between 5.7 and 15.3

3.30 According to a *Human Resources* report, a worker in the IT industry spends on average 419 minutes (or 6.98 hours) a day on the job. Suppose the standard deviation of time spent on the job is 27 minutes. a. If the distribution of time spent on the job is approximately bell shaped, between what two times would 68% of the data fall? 95%? 99.7%? b. If the shape of the distribution of times is unknown, approximately what percentage of the times would be between 359 and 479 minutes? c. Suppose a worker spent 400 minutes on the job. What would that worker's z-score be and what would it tell the researcher?

$\mu = 419$, $\sigma = 27$

(a) 68%: $\mu \underline{+} 1\sigma$ $419 \underline{+} 27$ **392 to 446**

95%: $\mu \underline{+} 2\sigma$ $419 \underline{+} 2(27)$ **365 to 473**

99.7%: $\mu \underline{+} 3\sigma$ $419 \underline{+} 3(27)$ **338 to 500**

(b) Use Chebyshev's:

The distance from 359 to 479 is 120

$\mu = 419$ The distance from the mean to the limit is 60.

k = (distance from the mean)/σ = 60/27 = 2.22

Proportion = $1 - 1/k^2 = 1 - 1/(2.22)^2 = .797$ = **79.7%**

(c) $x = 400.\ z = \dfrac{400-419}{27} = \mathbf{-0.704}$.

This worker is in the lower half of workers but within one standard deviation of the mean.

3.31 According to the Australian Taxation Office, the average taxable income in an affluent suburb of Sydney is $94 720. Suppose the median taxable income in this area is $90 050 and the mode is $89 200. Is the distribution in this area skewed? If so, how? Which of these measures of central tendency would you use to describe these data? Why?

Mean	$94,720
Median	$90,050
Mode	$89,200

Since these three measures are not equal, the distribution is skewed. The distribution is skewed to the right. Often, the median is preferred in reporting income data because it yields information about the middle of the data while ignoring extremes.

3.32 A hire car company is interested in summary statistics that are useful in describing travel times between the central business district and the domestic terminal at Sydney Airport. They locate a report that indicates that the average total time for travel by car is 14 minutes. The shape of the distribution of travel times is unknown, but in addition it is reported that 35% of travel times are between 10.5 and 17.5 minutes. Use Chebyshev's theorem to determine the value of the standard deviation associated with travel times.

35% of observations within a given range implies under Chebyshev's theorem:
$1 - 1/k^2 = .35$.

Solving for k, $k = 1.240347$

The distance from $\mu = 14$ to upper value of $x = 17.5$ is $17.5 – 14 = 3.5$

$1.240347\sigma = 3.5$

$\sigma = 3.5/1.240347 = \mathbf{2.822}$

3.33 The *Monthly Banking Statistics* published by the Australian Prudential Regulation Authority provides selected information on the banking business of individual banks within the domestic market. It contains high-level breakdowns of the domestic assets and liabilities of each bank as well as more detail on loans and advances to and deposits by different sectors of the economy. Total resident assets refer to all assets on the banks' domestic books that are due from residents. The following Excel descriptive statistics output lists the variable total domestic assets of banks in Australia ($ million). Study the output and describe in your own words what you can learn about the domestic assets of banks in Australia.

The mean value for total domestic assets of banks in Australia is $37,020 million and the median value is given by $6,478 million. The standard deviation of the distribution is $83,148 million. The distribution of total domestic assets of banks in Australia is positively skewed as a result of extreme values.

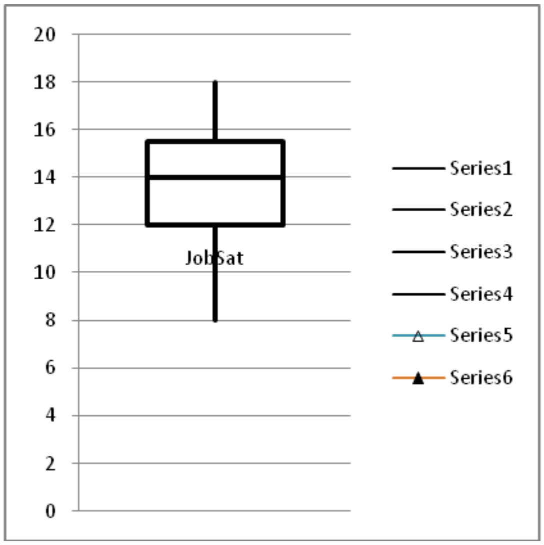
3.34 The CEO of Combaro would like to know whether employees are satisfied in their positions. In particular, the CEO would like to know about the central tendency of the data and whether the data are skewed in some way. For instance, the CEO suspects that there may be many employees who are quite satisfied, but average satisfaction levels are being distorted by a few individuals who are extremely unhappy. Use a boxplot to investigate whether this is the case. Write a report to the CEO on your findings, including supporting numerical measures such as skewness.

JobSat	
Mean	13.666667
Standard Error	0.368692
Median	14
Mode	17
Standard Deviation	2.5543732
Sample Variance	6.5248227
Kurtosis	-0.441206
Skewness	-0.321773
Range	10
Minimum	8
Maximum	18
Sum	656
Count	48

<u>Boxplot Output</u> JobSat

First Quartile	12.0000
Median	14.0000
Third Quartile	15.5000
Interquartile Range	3.5000

Moderate Outliers (△)	0
Extreme Outliers (▲)	0



Overall, the employees of Combaro are satisfied with their jobs. Their job satisfaction scores range from 8 to 18 on a scale from 1 to 20, with a mean of 13.67 and a median of 14. The box-plot and the coefficient of skewness of -0.322 indicate that the distribution is slightly negatively skewed.

The boundaries of the inner fence are:

$$Q_1 - 1.5 \times IQR = 12 - 1.5 \times 3.5 = 6.75$$

$$Q_3 + 1.5 \times IQR = 15.5 + 1.5 \times 3.5 = 20.75$$

All of the satisfaction scores are within these boundaries, therefore none of the employees seem to be extremely unhappy.

3.35 **The following scatter plot examines the potential association between years of education and weekly salary. The data come from the Combaro data set. Write short descriptions about what you would expect to see in the graph and what you actually see. In examining both descriptions, estimate the correlation coefficient in each case. Using the Combaro data set, calculate the correlation coefficient. What does your analysis suggest about the years of education an employee has completed and the amount they are paid?**

One would like to hope that individuals with more years of education are being paid higher salaries, with a correlation coefficient of say about 0.5. However, the graph indicates that there is only a very weak positive correlation between the years of education and the weekly salary. The estimated correlation coefficient is say 0.15.

When using Excel to calculate the correlation coefficient between these variables we obtain a value of 0.07, which indicates that there is practically no linear relationship between these two variables.

	EducYrs	*WkSalry*
EducYrs	1	
WkSalry	0.069935	1

The actual correlation coefficient is only 0.070, indicating that the weekly salary is not related to one's years of education.

3.36 A recent article in the *Baycoast City Times* states that the days of large backyards have disappeared, and that larger properties are in short supply. In turn, the journalist writes that the Baycoast region provides the perfect place for those wishing to find houses on larger lots and property developers to have a greater number of options for subdivision. This article was based on a report by Baycoast City Real Estate Agents that the mean lot size of recent properties sold was 1175 square metres. The journalist interpreted this to imply that 50% of properties are equal to or bigger than 1175 square metres. The journalist also claimed that this is made even more exciting because larger lot sizes always mean higher house prices.

a. Using the BCREA data set and KaddStat, construct a box and whisker plot to verify how representative this mean lot size is and verify the journalist's claims about lot sizes in the Baycoast region. Write a short report to explain how to interpret different summary measures of central tendency and how the journalist may have misinterpreted the original summary measure.

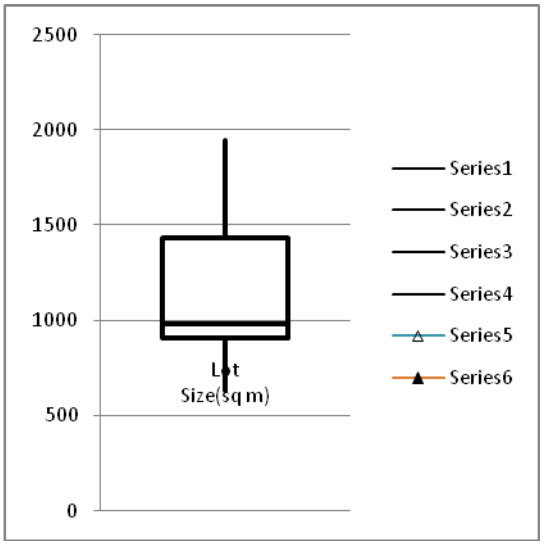
b. Use an appropriate measure of linear association to determine whether house prices and lot sizes are related. Interpret this measure to determine whether it supports the journalist's claims.

c. Which other factors in the Baycoast data set could also be associated with higher house prices? Pick one these and see whether the sample data in the Baycoast data set support your suspicions.

a)

<u>Boxplot Output</u>	Lot Size(sq m)
First Quartile	911.0000
Median	980.0000
Third Quartile	1436.0000

Interquartile Range	525.0000
Moderate Outliers (Δ)	0
Extreme Outliers (▲)	0



The journalist is correct by stating that the mean lot size in the Baycoast region is 1175 square metres. However, the distribution is positively skewed and therefore the mean is not a representative measure of the central tendency. The median value of $980 should be used instead, which implies that 50% of the properties in the Baycoast region are equal to or greater than 980 square metres.

b)

	Price($'000)	*Lot Size(sq m)*
Price($'000)	1	
Lot Size(sq m)	0.41083	1

The correlation coefficient of 0.411 indicates that there is a moderate positive correlation between the house price and the lot size. Houses with larger backyards sell at higher prices.

c)

	Price($'000)	*Area (sq m)*
Price($'000)	1	
Area (sq m)	0.567769	1

The correlation coefficient between the house area and house price is 0.568, which indicates that larger houses are more expensive than smaller houses.

3.37 A sales report describes the current price of various digital cameras on the market with 16-megapixel resolution and 10× optical zoom using z-scores. Prices are described as following a normal distribution with an average price of $219 and a standard deviation of $13. The z-score associated with one particular model is 1.5. What is its sales price? How much more expensive is this camera compared to another camera with a z-score of −2.5?

$\mu = \$219, \quad \sigma = \$13$

Given $z = \frac{x-\mu}{\sigma}$, we can solve for x: $x = \mu + z\sigma$

A camera with a z-score of 1.5 and assuming a normal distribution is appropriate implies a price of: x = 219+(1.5)(13) = **$238.50**

A camera with a z-score of -2.5, has a price of x = 219+(-2.5)(13) = $186.50

The first camera is 238.50 – 186.50 = **$52** more expensive than the second camera.