Segment: Descriptive Statistics

Topic: Numerical Representation of Data

Table of Contents

Introduction

Measures of central location, variability, and shape help to describe a data set. Measures of location reveal centrality of the data set, while variability measures reveal deviance of the data set from the centre. Measures of shape enable us to compare the shape of a distribution relative to common distributions found in data. We also introduce an alternative graphical method of presenting a distribution, known as a boxplot. In addition, it is sometimes useful to measure the linear association between two variables.

Correlation is a useful measure of this association between the two variables. In this topic, we will be investigating measures of central location, spread, and shape. In addition, we will also be exploring a technique to summarise the strength of the relationship between two variables.

Learning Objectives

At the end of this segment, you will be able to:

- distinguish between the three alternatives for measuring central location
- identify the various methods used to measure deviations
- compare the methods used to measure the shape
- use Excel to generate summaries of relationships
- identify the measures used to summarise linear relationships between two variables
- evaluate the importance of using boxplots for investigating the distribution of variables.

1. Measures of Central Location

The central location of a numerical data set provides useful information.

Firstly, the central location is a representative value of the data set.

Secondly, it can serve as a basis for making comparisons – if set A is centred at 10 and set B at 20, it makes sense to say that the numbers in set A are mostly smaller than those in set B.

There are three alternatives for the measure of central location:

1. Mean

2. Median

3. Mode

Read below to explore the differences between these three measures of central location using the sales data set {6, 9, 10, 12, 13, 14, 14, 15, 16, 16, 16, 17, 17, 18, 18, 19, 20, 21, 22, 24}.

Three Measures of Central Location

1. Mean

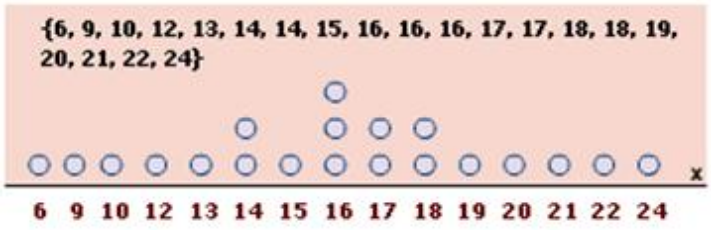
Mean = Average

The mean is the **average** of all the numbers in the set. It is calculated by taking the sum of all the numbers and dividing the sum by the number of numbers. It is also the most commonly used measure. In the given data set, the Mean is

(6 + 9 + 10 + 12 + 13 + 14 + 14 + 15 + 16 + 16 + 16 + 17 + 17 + 18 + 18 + 19 + 20 + 21 + 22 + 24)/20

= 317/20

= 15.875



2. Median

Median = value below which half the data points lie

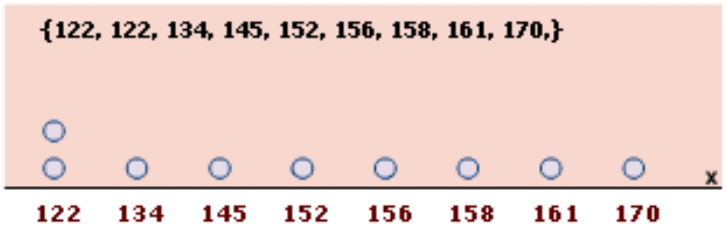
The median is defined such that half the numbers in the set are greater and the other half are smaller than it.

If the numbers in a data set are **sorted** in increasing order, then **"the number in the middle position"** is clearly the median.

If there are an odd number of numbers, say, nine, then the middle is the fifth position and the number in the fifth position is the median.

For odd number data set {122, 122, 134, 145, 152, 156, 158, 161, 170} with 9 numbers

median = 152



But if there are an even number of numbers, say, ten, then there is no middle position.

Rather, the fifth and the sixth positions together form the middle. In this case, the average of the two numbers in positions five and six is declared the median.

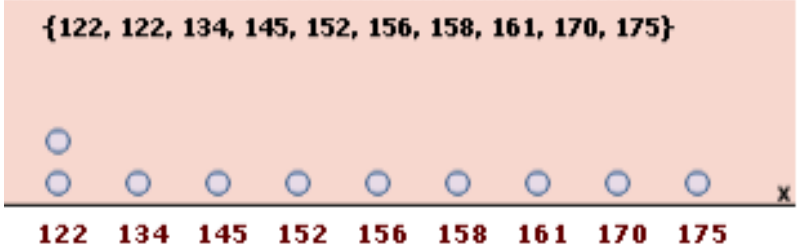
For the data set {122, 122, 134, 145, 152, 156, 158, 161, 170, 175}

with ten numbers, the median is the average of 152 and 156, which is 154.

For **even** number data set n {122, 122, 134, 145, 152, 156, 158, 161, 170, 175} with 10 numbers

median = (152 + 156)/2

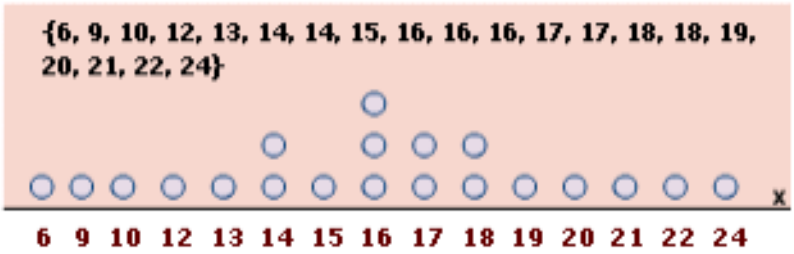
median = 154



For the above sales data set,

Median = (16 + 16)/2

Median = 16

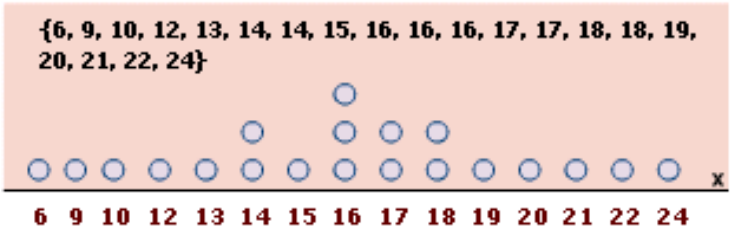


3. Mode

Mode = most frequently occurring value

The mode is defined as the most frequently occurring value in the set.

In the given data set, the Mode is also 16. because it has the largest frequency of 3.

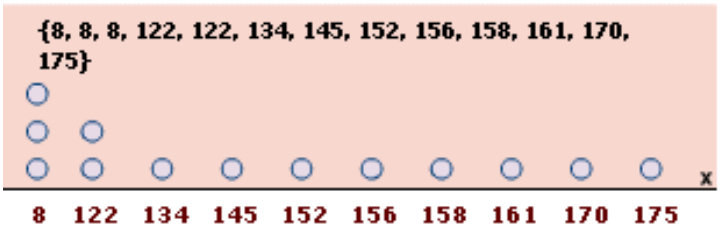


The use of the mode as the centre or a representative value of a data set is better suited for large data sets, say, with at least a hundred numbers.

To see why, consider the data set

{8, 8, 8, 122, 122, 134, 145, 152, 156, 158, 161, 170, 175}

Its mode is 8, but 8 is hardly a representative value or the centre of the set.



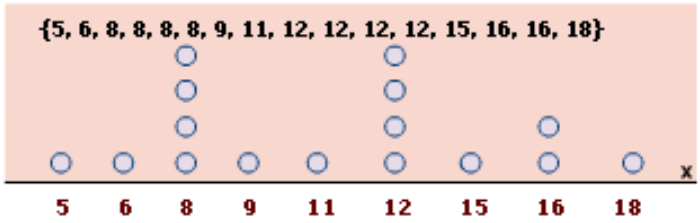
Such anomalies are unlikely in a large data set.

Sometimes there can be two (or more) modes for a data set.

Consider the set {5, 6, 8, 8, 8, 8, 9, 11, 12, 12, 12, 12, 15, 16, 16,18}

It has two modes, namely, 8 and 12.

This can also pose problems. Once again, this is unlikely to occur in a large data set.



1.1 Comparing the Mean, Median, and Mode

The following table presents a comparison of the three measures.

Table 1: Comparing Mean, Median, and Mode

	Advantages	Disadvantages
Mean	• Possesses some useful mathematical properties • Related to the sum of all the numbers • Suitable for both small data sets (that do not have extreme values) and large data sets	• Affected by extreme values in small data sets that can make it unsuitable for representation • Needs some calculation involving adding and dividing
Median	• Very suitable for small data sets if there are extreme values • No calculation beyond sorting and counting needed	• Needs sorting
Mode	• No calculation needed • Suitable for design applications	• May not be unique • Needs frequency counts

The mean has an additional desirable mathematical property:

The deviations of all the data points from the mean will add to zero. For example, the mean of the data set {5, 7, 8, 10, 10} is 8. The deviations from the mean are {-3, -1, 0, +2, +2} and these deviations add to zero.

Using mean

In algebraic calculations that involve products of deviations, many terms become zero and vanish by virtue of this property. As a result, we get simpler results than if we used median or mode. For this reason, the mean is used in almost all the techniques we will see in later topics.

Exercise:

Below is an exercise to practice what you have learnt on mean, median, and mode.

Here is a data set of monthly income for a set of ten employees:
1,800; 1,980; 2,000; 2,400; 2,750; 3,100; 3,200; 4,320; 6,750; 12,000.

Question 1: Based on these figures, calculate the mean, median, and mode.
Question 2: Based on these figures, which is the better measure of the central tendency of data?

1. Mean
2. Median
3. Mode

1.2 Percentiles and Quartiles

Besides the measure of central location, one may also use a percentile or a quartile of a data set in order to describe it.

The n^{th} percentile of a data set is that value below which n% of the data lie. The most common percentiles used are quartiles.

- The first quartile is the 25th percentile where 25% of the observations fall below this point.
- The second quartile is the same as the 50th percentile or the median.
- The third quartile is the 75th percentile with 75% of the observations falling below this point.

Percentiles and quartiles are popular ways to describe the position of a data point within a data set.

2. Measures of Variability

There is a popular joke that goes: a statistician is one who would say your body temperature is normal if your head is in the oven and your feet are in the fridge. Of course, that is only a joke. Statisticians do pay attention to deviations from the measure of central location, because they are as important as, or at times even more important than, the average.

Common measures used to quantify deviations are:

- Inter-quartile range

This is the difference between the first and the third quartiles. It is a measure of the spread of the data:

Inter-quartile range = Third quartile – First quartile

- Mean absolute deviation

The mean absolute deviation of a data set is the average of the absolute values of the deviations of all the data points from their mean. The formula is given below:

$$\text{MAD} = \frac{\sum_{i=1}^{n} |X_i - \text{mean}|}{n}$$

- Variance

The variance of a set of observations is the average squared deviation of the data points from their mean. Sample variance is denoted by s^2 and population variance is denoted by σ^2 and the formulae for both are given below:

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \text{mean})^2}{n-1} \qquad \sigma^2 = \frac{\sum_{i=1}^{n} (X_i - \text{mean})^2}{n}$$

- Standard deviation

The standard deviation of a set of observations is the (positive) square root of the variance of the set.

Standard deviation versus Variance

As seen from the formulas found under the solutions referred to above, be aware that the variance and standard deviation, the last two terms, are calculated slightly differently when the data is of a sample and of a population. Unlike the other measures of variability, the variance is a squared value, thus if our data has the dimensions of kilograms then our variance will be in kilograms^2 (kilograms squared). Since the standard deviation is the square root of the variance, it will have the dimensions of kilograms.

The advantage of working with the standard deviation is that it has the same units as the original data.

We need to keep in mind that two standard deviations cannot be added to produce a meaningful quantity. On the other hand, two variances can be added, in certain cases, to produce a combined variance. (We can see the details of this only when we learn about random variables later.) As a result, you will see that in some instances, people mention the standard deviation of a data set and in other instances, they mention the variance. You should be wary of which one is being mentioned and use it accordingly.

The obvious disadvantage with the variance is that it is in squared units (eg, kg^2) and therefore, cannot be compared with quantities in the original unit (eg, kg).

Examples of variability

Two important applications of measures of variability are worth mentioning here.

Read below to find out more about these applications.

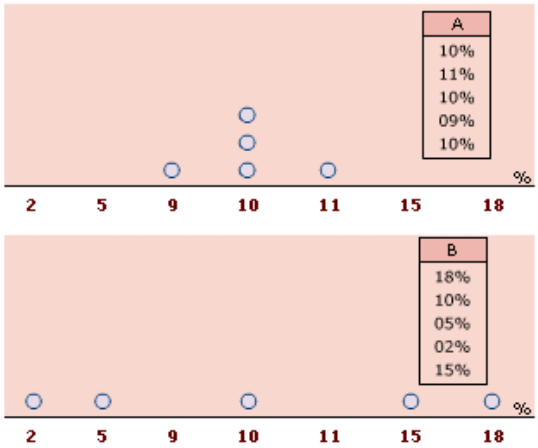
Example 1

Suppose there is some uncertainty in the rates of return from two investments A and B. Looking at the past five years' data we see that the returns from A and B are

A: 10%, 11%, 10%, 9%, 10%

B: 18%, 10%, 5%, 2%, 15%

On average, both A and B returned 10%. However, B's returns have deviated more than A's. We would therefore, declare that B is more risky. To quantify how risky an investment's returns are, it is customary to use the variance or the standard deviation of its past returns. The smaller the variance, the smaller the risk.



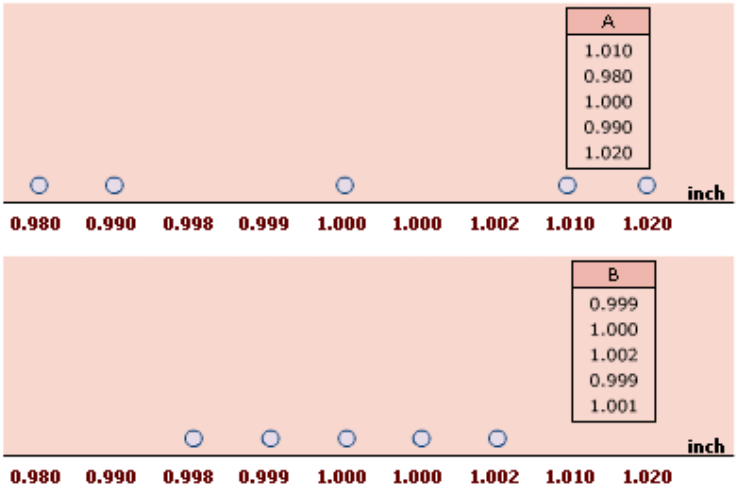
Example 2

Suppose two automatic machines A and B produce pins whose diameters need to be exactly one inch. Suppose a random sample of five pins is taken from A and another sample from B. Let the accurate measurements of the diameters of these samples be:

A: 1.010, 0.980, 1.000, 0.990, 1.020

B: 0.999, 1.000, 1.002, 0.998, 1.001

The average diameter of the pins from A and B are both 1.000. But A's pins deviate from 1.000 to a larger extent than B's. Hence B has better quality. Thus, a measure of variability can be useful for measuring the quality of machined parts. The smaller the variability the better the quality.



3. Measures of Shape

It is useful to have a measure to describe the shape of a distribution. One method to do this is to compare the distribution to a symmetrical distribution, since for a symmetrical distribution that is unimodal, the three measures of central location coincide. If the distribution is not symmetrical, the distribution is said to be skewed.

Skewness measures the deviation from symmetry, and hence zero skewness means perfect symmetry. Refer to the two figures below:

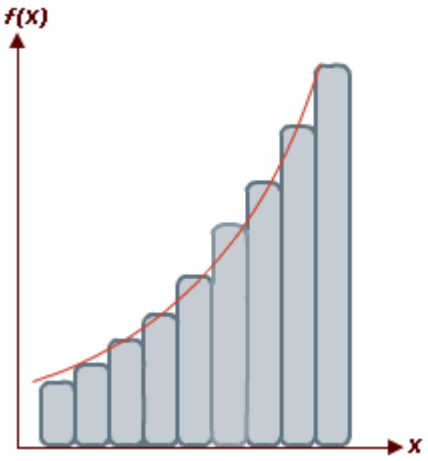


Fig. 1: Negatively Skewed

Negative skewness indicates there are some large deviations to the left.

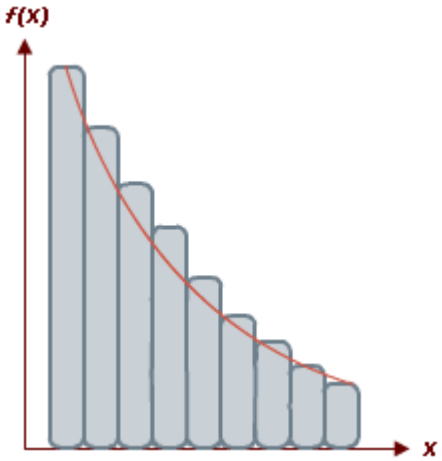


Fig. 2: Positively Skewed

Positive skewness indicates there are some large deviations to the right.

One method of investigating the skewness of a distribution is to compare the mean and median of the data set. If the mean is greater than the median, the data is said to be right skewed (or positively) skewed. Likewise, if the mean is less than the median, the data is said to be left skewed (or negatively) skewed.

4. Using Excel to Generate Summary Measures

Fortunately, there is very little need to remember the formulas for calculating the most common summary measures as Microsoft Excel provides tools for calculating these.

As an example, consider data collected as part of an investigation on new home prices for a particular building company in a major city in Australia. The investigators collected data on the price of the new home, the floor area, and several other relevant variables.

Suppose that we were interested in summarising the information on the size of each home. To do this we can make use of the Descriptive Statistics function in Excel (on the Excel menu bar, click Data/ Data Analysis/ Descriptive Statistics). In the descriptive statistics window, you will need to select the relevant 'Input Range'.

The Descriptive Statistics window is shown in the figure below.

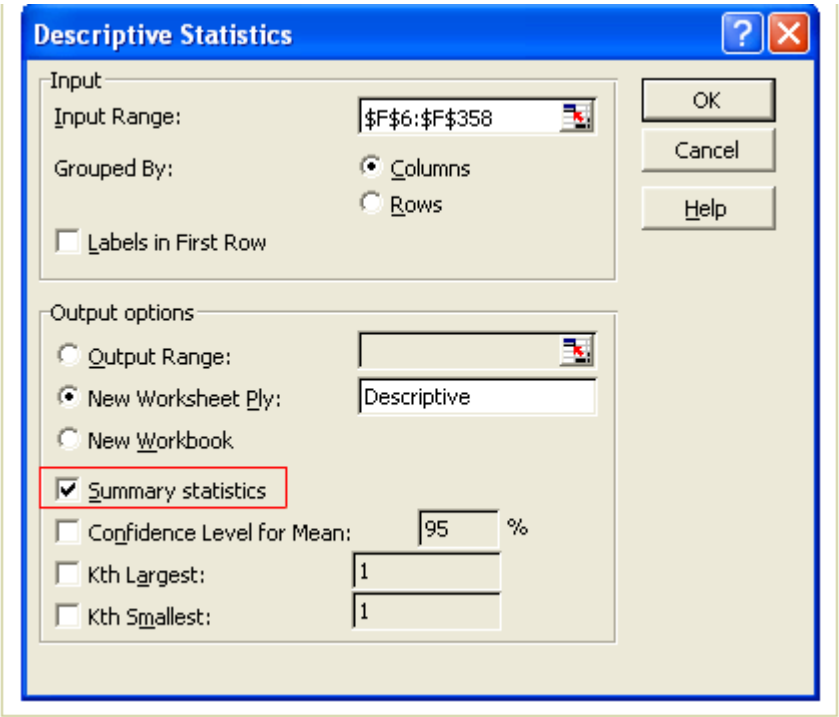


Fig. 3: Descriptive Statistics Window

If we select our data range and check the box for the 'Summary statistics', the following output as shown in the table below appears.

Table 2: Output of Summary Measures

Floor area (metres squared)	
Mean	194.58
Standard error	3.16
Median	193.36
Mode	242.98
Standard deviation	59.35
Sample variance	3,522.29
Kurtosis	0.24
Skewness	0.35
Range	331.55
Minimum	69.39
Maximum	400.94
Sum	68,685.55
Count	353

As can be seen in the table, the output includes most of the key summary measures we would be interested in determining.

- The average floor area of the new homes (mean) is 194.58 m^2
- The standard deviation is 59.35 m^2
- The median floor area is 193.36 m^2, which, if compared to the mean, suggests that the data may be very slightly right, or positively, skewed.

5. Measures of Association

We have been introduced to scatter plots to investigate the relationship between two variables. It is also useful to summarise the linear relationship between two variables. Two measures useful in summarising this relationship are:

- Covariance
- Correlation

Each of these measures describes the strength and direction of the linear relationship between two quantitative variables.

Covariance

The covariance is essentially the average of the products of the deviations from the mean of each of the variables. If the two variables tend to vary in the same direction, the covariance will tend to be positive. Likewise, if the two variables tend to vary in opposite directions, the covariance will tend to be negative.

As an example, we could consider the selling prices of homes in a particular city. We might expect the covariance between the selling price and the distance from the city centre to be negative. Likewise, we would expect that the price of a house to increase with the size of the house. In this case, we would expect the covariance between the price and the size of the house to be positive.

One limitation of using the covariance as a measure of association is that it is affected by the units in which the two variables are measured. For example, suppose you have been told that the covariance of two variables is 250. The sign is positive which suggests that the two variables

tend to vary in the same direction, but the magnitude gives us little information as to the strength of this relationship. To overcome this problem, we can derive another measure of association to describe the strength of the relationship.

Correlation

This measure is the coefficient of correlation and is obtained by dividing the covariance by the product of the standard deviation of both variables. The resulting measure is a unitless quantity that is unaffected by the units of measurement of each of the two variables.

The coefficient of correlation is *always* between -1 and +1. The closer it is to either of these extremes, the closer the points in a scatter plot are to some straight line, either in a positive or negative direction. Likewise, a coefficient of correlation close to 0 indicates that there may be no linear relationship between the two variables.

As an example, consider the new home price data collected in the previous example. Suppose that the investigators were interested in the relationship between the price of a new home and its floor area. The figure below shows this data represented graphically as a scatter plot.

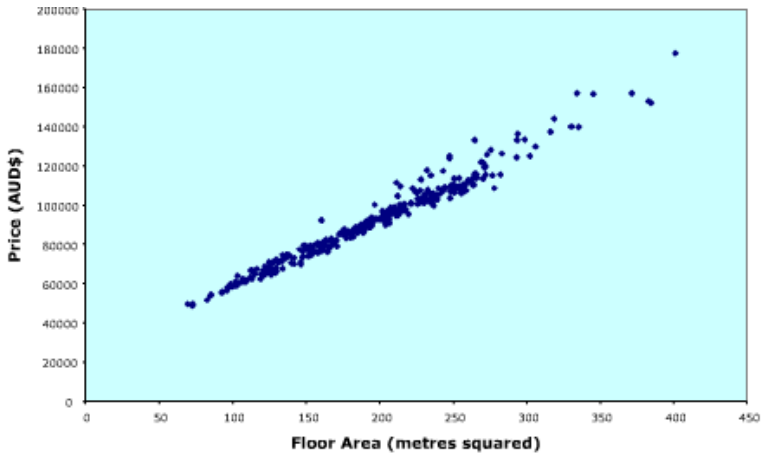


Fig. 5: New Home Price Versus Floor Area

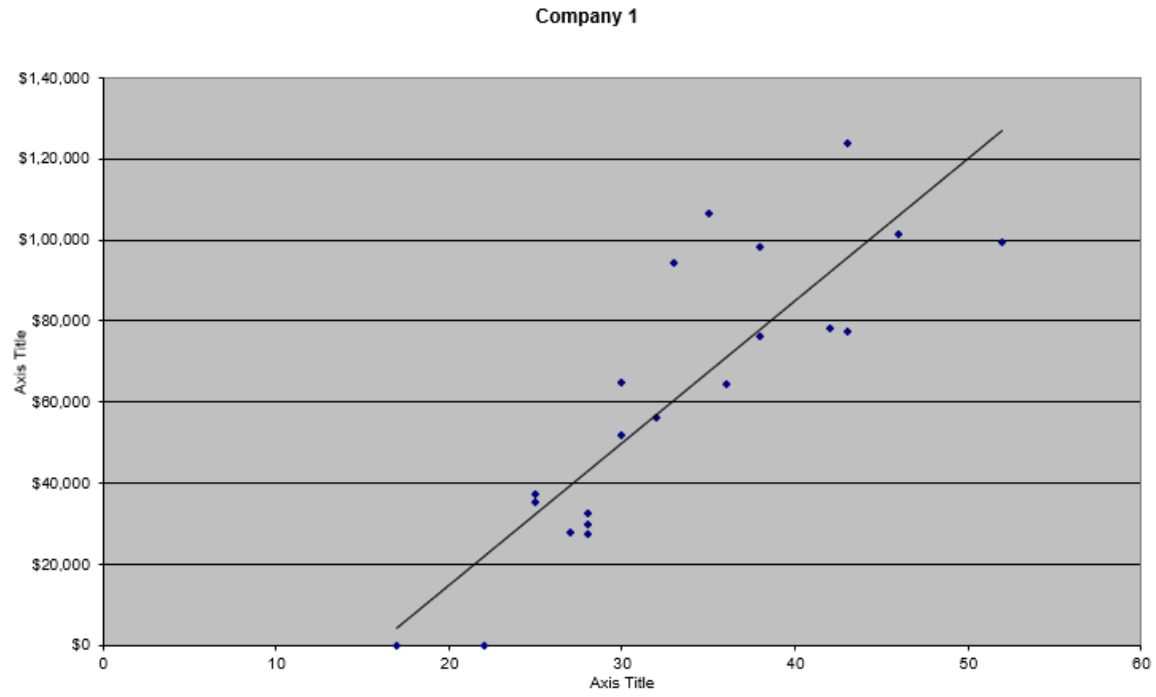
The covariance between these two variables is 1234011. This value can be derived by using the covariance data analysis option in Excel (Tools/Data Analysis/Covariance). This figure is positive which suggests that there is a positive relationship between these two variables, however the magnitude of this result is difficult to interpret unless we know the units. The coefficient of correlation for this relationship is 0.9842. As we discussed previously, this indicates that there is a positive relationship between these two variables and as the value is close to 1 it also suggests

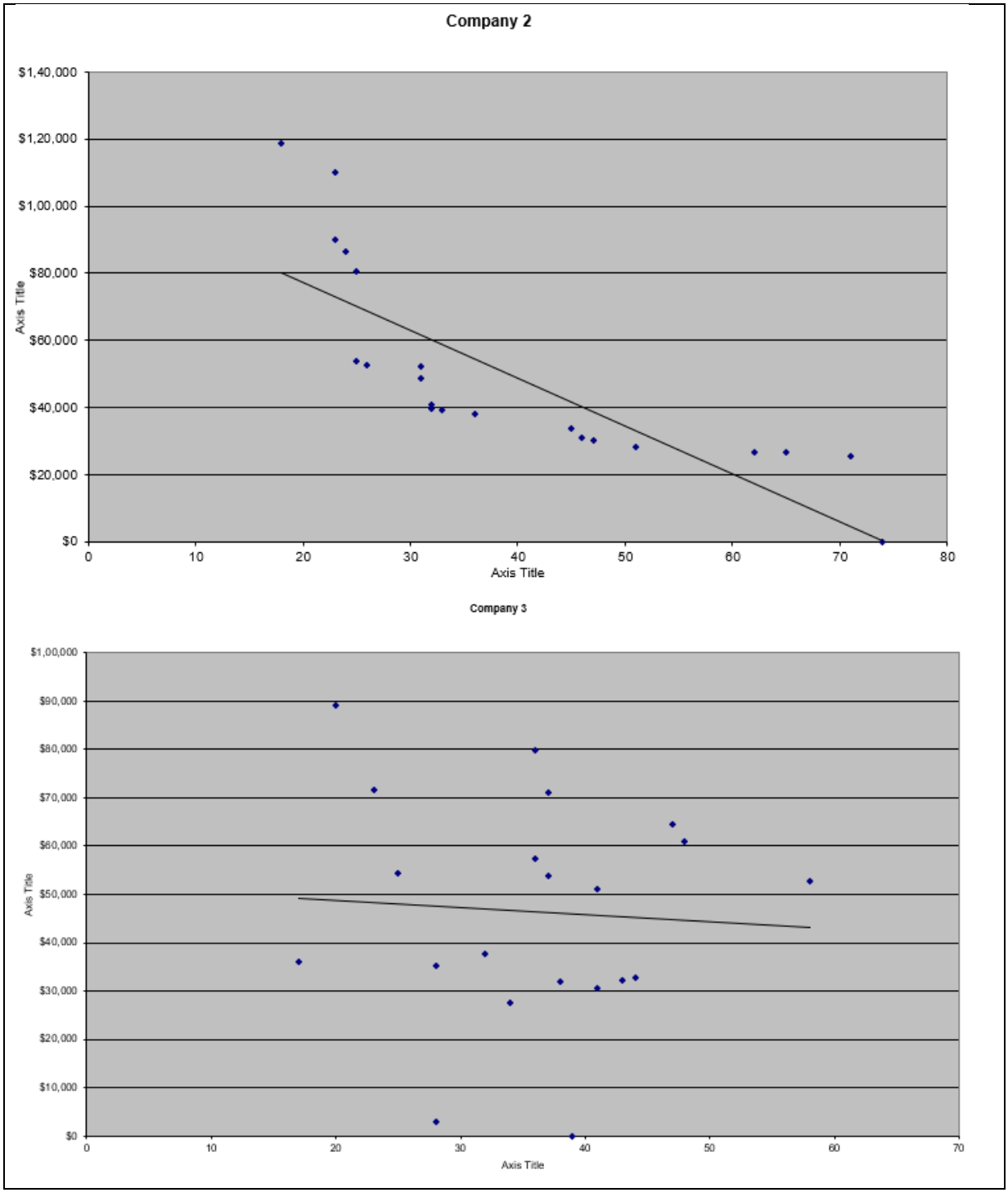
that there is a very strong linear component in the relationship between these two variables (as can be seen from the above figure).

Exercise

Below is an exercise to practise your knowledge of correlation.

Data on employee age and salaries for three different companies, Company 1, Company 2, Company 3 is shown using three scatterplots, which represent the relationship between the employee age and their respective salaries.





Company 1		Company2		Company 3	
Age	Annual_Income	Age	Annual_Income	Age	Annual_Income
38	$98,440	18	$1,18,854	39	$0
25	$37,227	23	$1,10,038	28	$2,938
38	$76,048	23	$89,865	34	$27,461
28	$27,610	24	$86,303	41	$30,450
22	$0	25	$80,443	38	$31,873
52	$99,429	25	$53,989	43	$32,112
43	$1,23,828	26	$52,569	44	$32,641
28	$32,611	31	$52,049	28	$35,205
17	$0	31	$48,774	17	$36,158
25	$35,230	32	$40,703	32	$37,709
32	$56,173	32	$39,595	41	$51,000
36	$64,241	33	$39,242	58	$52,604
27	$27,856	36	$38,224	37	$53,763
28	$29,720	45	$33,569	25	$54,487
30	$65,000	46	$31,170	36	$57,451
35	$1,06,468	47	$30,289	48	$61,019
46	$1,01,318	51	$28,189	47	$64,533
43	$77,398	62	$26,605	37	$71,152
42	$78,206	65	$26,486	23	$71,668
30	$51,728	71	$25,300	36	$79,921
33	$94,402	74	$0	20	$89,052

Question 1: Refer to the diagram "Company 1". Which of the following is the correlation coefficient for company 1?

1. 0.343
2. 0.858
3. 0.994
4. 0.996

Question 2: Refer to the diagram "Company 2". Which of the following is the correlation coefficient for company 2?

1. -0.023
2. -0.026
3. -0.799
4. 0.434

Question 3: Refer to the diagram "Company 3. Which of the following is the correlation coefficient for company 3?

1. -0.062

2. -0.532
3. 0.858
4. 0.860

6. Describing Data with Boxplots

Another useful graphical method for summarising data is by using boxplots. Boxplots make use of quartiles and can be a very useful method of graphically representing the distribution of a single variable or for comparing the distributions of two or more variables.

As an example, suppose that a pizza store has four delivery drivers, and would like to investigate the times taken to deliver pizzas by each of these drivers. The manager of the store might take a sample of the deliveries by each of these drivers and then produce histograms of the delivery times for each driver.

An alternative is to produce a boxplot as shown in the figure below.

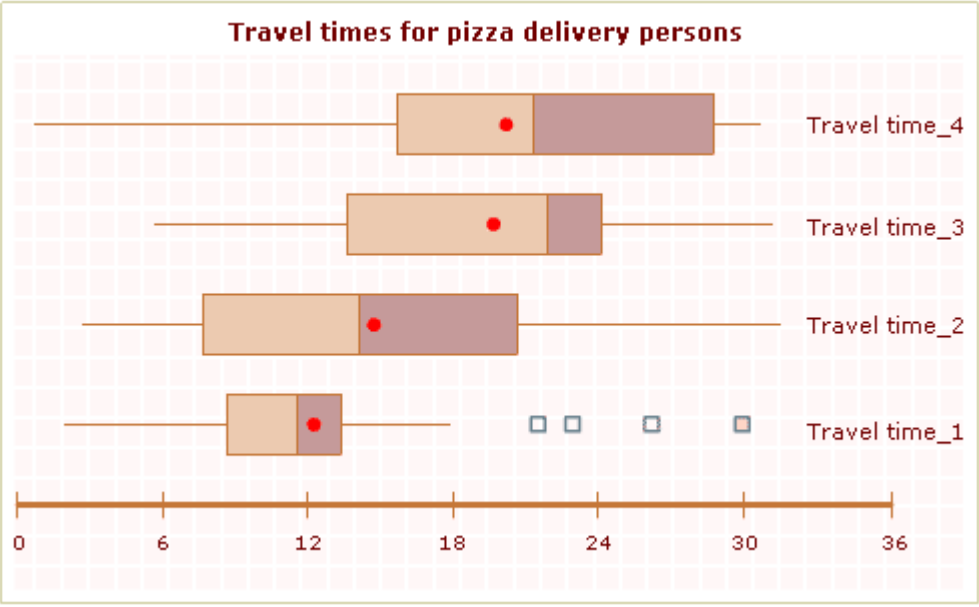


Fig. 6: Travel Times for Pizza Delivery Persons

The boxplots represent the delivery times for the four pizza drivers. The variable along the bottom of the graph is time and each of the four boxplots are stacked on top of each other allowing easy comparisons.

If we consider any one of the boxplots in the figure, we can see a box in the centre of each plot. The left and right side of the box are the first and third quartiles respectively. Therefore, the length of each box represents the interquartile range (the height of each box has no significance). The vertical line in each box represents the location for the median, while the point inside the box represents the mean.

Horizontal lines are also drawn from each side of the box. They extend to the most extreme observations on each side. Typically, these lines on each side of the box are drawn out no further than 1.5 interquartile ranges from each side. Points further out than this are considered outliers and are represented by dots (as on the right-hand side in the lower box the above figure).

In this case, it can be seen that the delivery times for driver four (the top boxplot) are, on average, longer. There may be several reasons for this and we would need to collect more information before making a decision. One obvious reason could be that driver four drives slower. Another reason is that driver four delivers pizzas to areas that are further from the store.

7. Summary

Here is a quick recap of what we have learnt so far:

- To describe a large data set, it is convenient to describe a measure of its central tendency and describe how the data points deviate from this central tendency.
- Measures of central location (mean, median, and mode) help to describe the centre of the data set.
- Measures of variability – inter-quartile range, mean absolute deviation, variance, and standard deviation – help to describe deviations from the centre.
- Skewness and similar measures can be used to describe the shape of the data sets.
- Tools such as Excel can be used to generate summaries of relationships
- Covariance and the correlation coefficient can be used to provide a measure of the strength of the linear relationship between two variables.
- Boxplots are a useful method of graphically investigating the distribution of one or more variables.

8. Glossary

Boxplot	A plot that describes the distribution of a data set, with a box in the middle and a whisker on each side. The box denotes where the middle half of the data lies, and the whiskers show the extent of the first and last quarters. Outliers are separated out to limit the length of the whiskers. The median is marked inside the box.
Percentile	The nth percentile of a data set is that value below which n% of the data points lie.
Quartile	The first quartile of a data set is that value below which a quarter of the data points lie. It is the same as the 25th percentile. The third quartile is that value below which three quarters of the data lie. It is the same as the 75th percentile.
Inter-quartile range	The inter-quartile range of a data set is the difference between its first and third quartiles.

9. Answers

Exercise: Measures of Central Location

Question 1: Correct answer is shown in the table below:

The mean is	4,030
The median is	2,925
The mode is	Nil

Mean of the average of the ten numbers in the data set, i.e.

```
(1,800 + 1,980 + 2,000 + 2,400 + 2,750 + 3,100 + 3,200 + 4,320 +
6,750 + 12,000) / 10 = 4,030
```

Median is defined such that half the numbers in the set are greater than it, and half are less than it. In this case there are 10 numbers, so the fifth and sixth numbers are added and average to get the medina, i.e.,

```
(2,750 + 3,100) / 2 = 2,925
```

Mode is the most frequently occurring variable, which, in this case, is absent from the data set- this is an instance of a case where is no unique mode.

Question 2: The correct answer is option 2, Median

The data supplied contains one outlier (12000). In this case it gives a mean which is considerably higher than the typical value in the dataset. In such cases the median might be a better measure.

Exercise: Measures of Association

Question 1: The correct answer is option 2, 0.858

The relationship seems to be quite a strong positive linear relationship and we would expect the correlation coefficient to be positive and reasonably close to one.

Question 2: The correct answer is option 3, -0.799

Although not quite a linear relationship, the relationship seems to be a moderately strong negative relationship and we would expect the correlation coefficient to be negative and reasonably large.

Question 3: The correct answer is option 1, -0.062

The linear relationship seems to be very weak and we would expect the correlation coefficient to be close to zero.