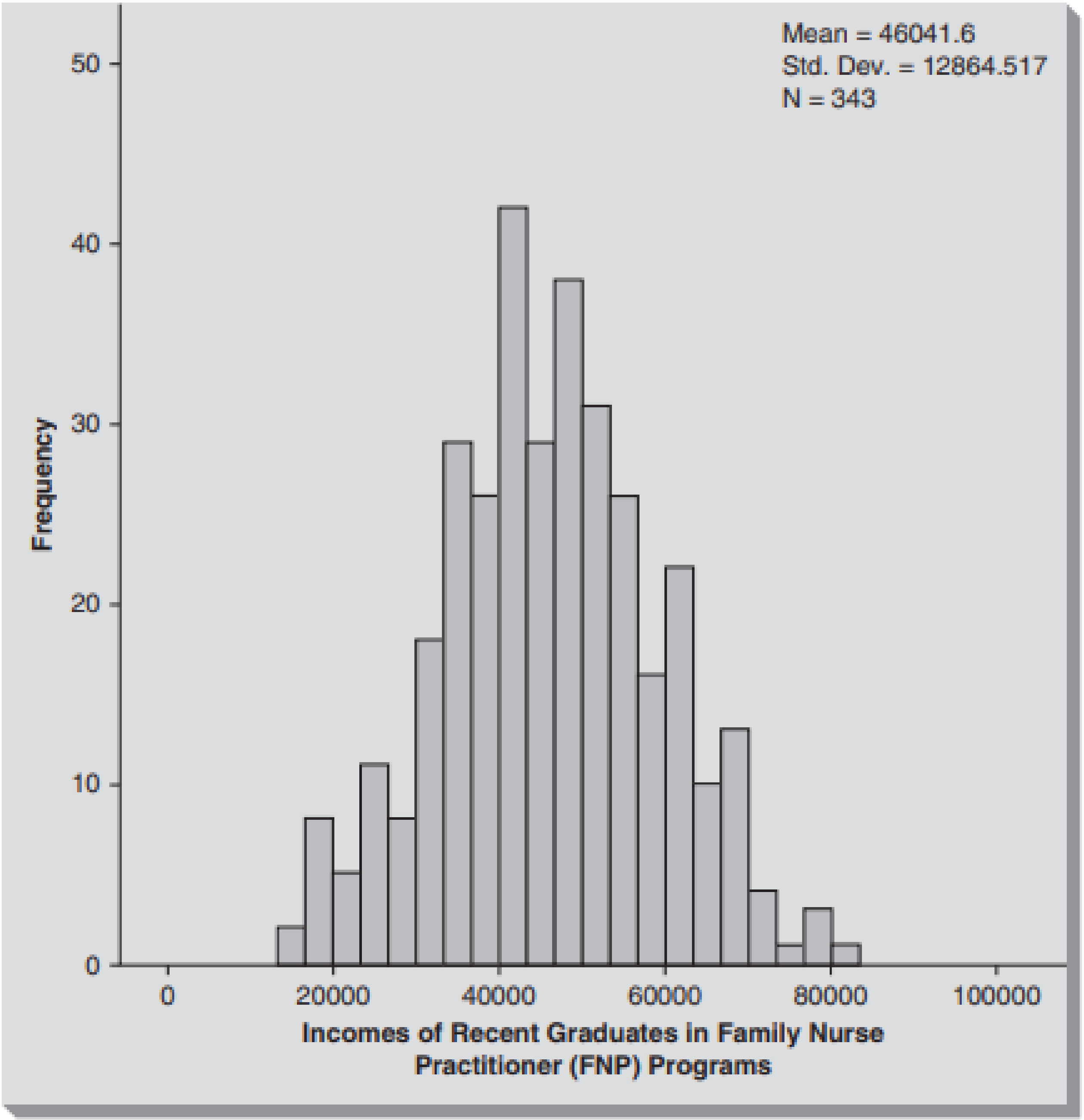
Descriptive Statistics

When the data are measured at the interval or ratio level, it is important to present the distribution of data in terms of **central tendency** (i.e., the average case) and **variability** (i.e., the range and spread of the data from the center). For example, **Figure 6-1** shows a histogram of incomes of recent graduates in Family Nurse Practitioner (FNP) programs. Questions we might ask about graduates are, "What would be the typical or average measurement value if one person was selected at random from this group?" and "How far from the average are data values spread?" These are difficult questions to answer with visual displays such as graphs, charts, and tables. We need numerical measures of central tendency and variability so that we can understand the distribution of the data on an objective basis.

These numeric measurements of central tendency and variability are examples of **descriptive statistics** and they help us to explain the data more accurately and in greater detail than graphical display. However, it is always good to begin with graphical displays of the data to visually inspect the distribution; you should then confirm what was seen in the graphical displays with numeric descriptive statistics.


Mean = 46041.6
Std. Dev. = 12864.517
N = 343
Frequency
50
40
30
20
10
0
0
20000
40000
60000
80000
100000
Incomes of Recent Graduates in Family Nurse Practitioner (FNP) Programs

- Descriptive statistics involves describing, summarizing and organizing the data so it can be easily understood.
- **Graphical displays** are often used along with the quantitative measures to enable clarity of communication.

- **Qualitative data**-

the variable which yield non numerical data.

  - E.g.- education, marital status, eye colour
  - **Frequency**- number of observations falling into particular class/ category of the qualitative variable.
  - **Frequency distribution**- table listing all classes & their frequencies.
  - Graphical representation- **Pie chart, Bar graph**.

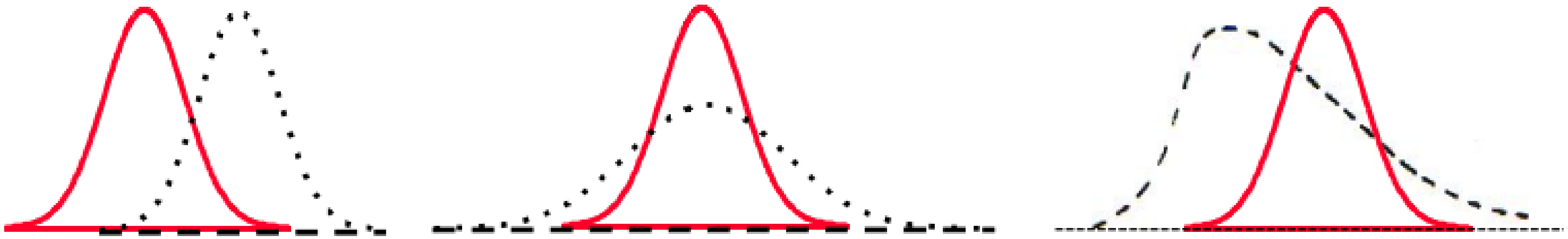
- **Quantitative data**-
  - Can be presented by a frequency distribution.
  - If the discrete variable has a lot of different values, or if the data is a continuous variable then data can be grouped into classes/ categories.

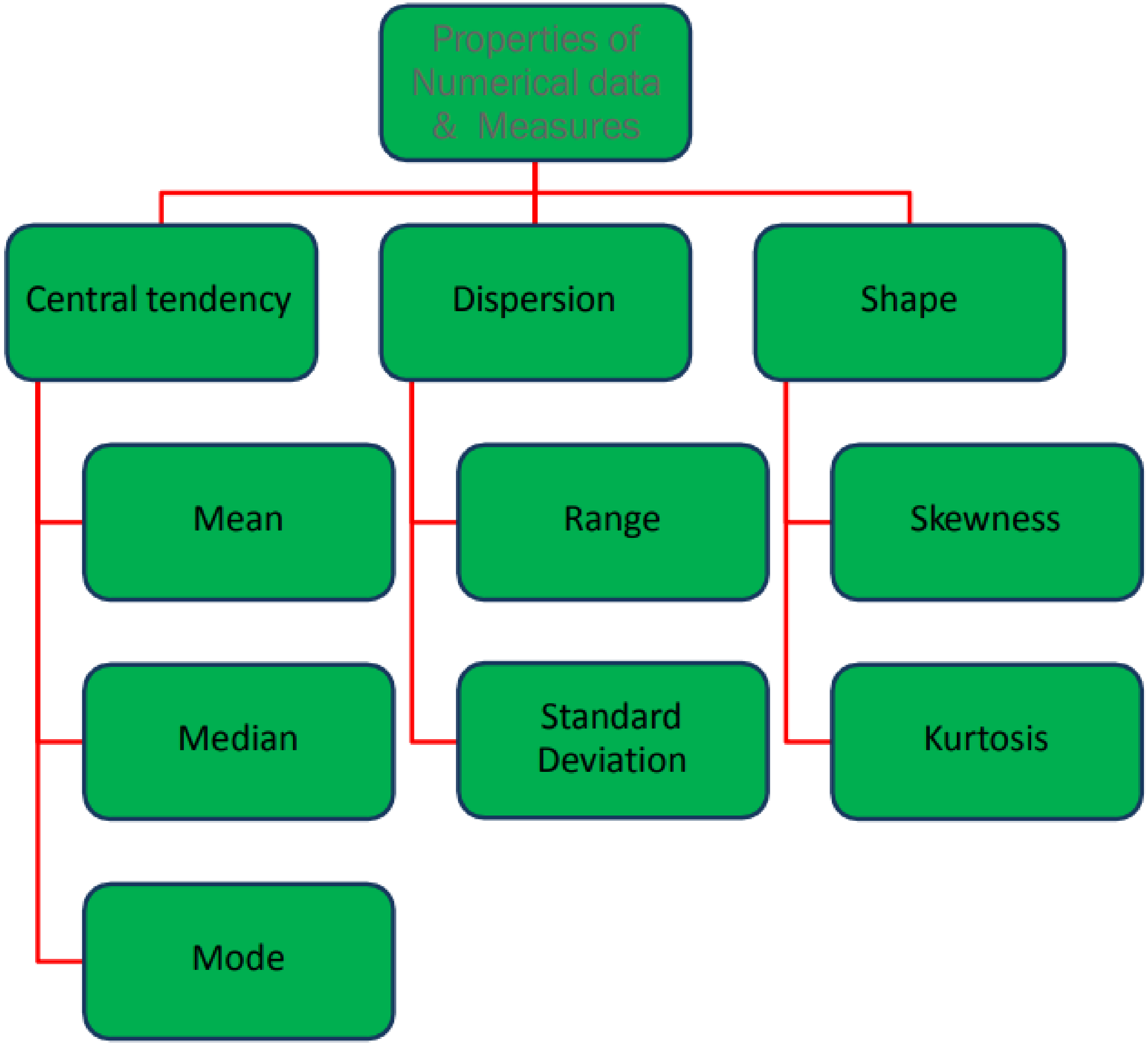
  - **Class interval / BINS**- covers the range between maximum & minimum values.
  - **Class limits**- end points of class interval.
  - **Class frequency**- number of observations in the data that belong to each class interval.

  - Usually presented as a **Histogram** or a **Bar graph**.

The following measures are used to describe a data set:

- Measures of position (also referred to as central tendency or location measures).
- Measures of spread (also referred to as variability or dispersion measures).
- Measures of shape.





Properties of Numerical data & Measures

Central tendency

Mean

Median

Mode

Dispersion

Range

Standard Deviation

Shape

Skewness

Kurtosis

Measures of Position:

- ❑ Position Statistics measure the data central tendency.
- ❑ Central tendency refers to where the data is centered.
- ❑ You may have calculated an average of some kind.
- ❑ Despite the common use of average, there are different statistics by which we can describe the average of a data set:
  - Mean.
  - Median.
  - Mode.

Measures of center

- **Central tendency**- In any distribution, majority of the observations ***pile up, or cluster around*** in a particular region.

- **Mean**- sum of observed values in a data divided by the number of observations

- **Median**- observation in the data set that divides the data set into half.

- **Mode**- value of the data set which occurs with greatest frequency

- Mean & Median can be applied only to Quantitative data

- Mode can be used either to Qualitative or Quantitative data.

- **Outlier**- observation that falls far from the rest of the data. Mean gets highly influenced by the outlier.

The mean – add up all the numbers and divide by how many numbers there are.

The median – is the middle number. It is found by putting the numbers in order
and taking the actual middle number if there is one,
or the average of the two middle numbers if not.

The mode – is the most commonly occurring number.

Let's illustrate these by calculating the mean, median and mode for the following data.

Weight of luggage presented by airline passengers at the check-in (measured to the nearest kg).

$$18 \quad 23 \quad 20 \quad 21 \quad 24 \quad 23 \quad 20 \quad 20 \quad 15 \quad 19 \quad 24$$

$$\text{Mean} = \frac{18+23+20+21+24+23+20+20+15+19+24}{11} = 20.64.$$

Median = 20.

$$15 \quad 18 \quad 19 \quad 20 \quad 20 \quad \underset{\substack{\uparrow \\ \text{middle value}}}{20} \quad 21 \quad 23 \quad 23 \quad 24 \quad 24$$

Mode = 20. The number 20 occurs here 3 times.

Here the mean, median and mode are all appropriate measures of central tendency.

The following data give the lifetime of 30 incandescent light bulbs (rounded to the nearest hour) of a particular type.

872	931	1146	1079	915	879	863	1112	979	1120
1150	987	958	1149	1057	1082	1053	1048	1118	1088
868	996	1102	1130	1002	990	1052	1116	1119	1028

Construct a frequency, relative frequency, and cumulative relative frequency table.

Solution

Note that there are $n = 30$ observations and that the largest observation is 1150 and the smallest one is 865 with a range of 285. We will choose six classes each with a length of 50.

Class	*Frequency*	*Relative frequency*	*Cumulative relative frequency*
	f_i	$\frac{f_i}{\sum f_i}$	$\sum_{k=1}^{i} \frac{f_k}{n}$
50–900	*4*	*4/30*	*4/30*
900–950	*2*	*2/30*	*6/30*
950–1000	*5*	*5/30*	*11/30*
1000–1050	*3*	*3/30*	*14/30*
1050–1100	*6*	*6/30*	*20/30*
1100–1150	*10*	*10/30*	*30/30*

The following data refer to a certain type of chemical impurity measured in parts per million in 25 drinking-water samples randomly collected from different areas of a county.

11	19	24	30	12	20	25	29	15	21
24	31	16	23	25	26	32	17	22	26
35	18	24	18	27					

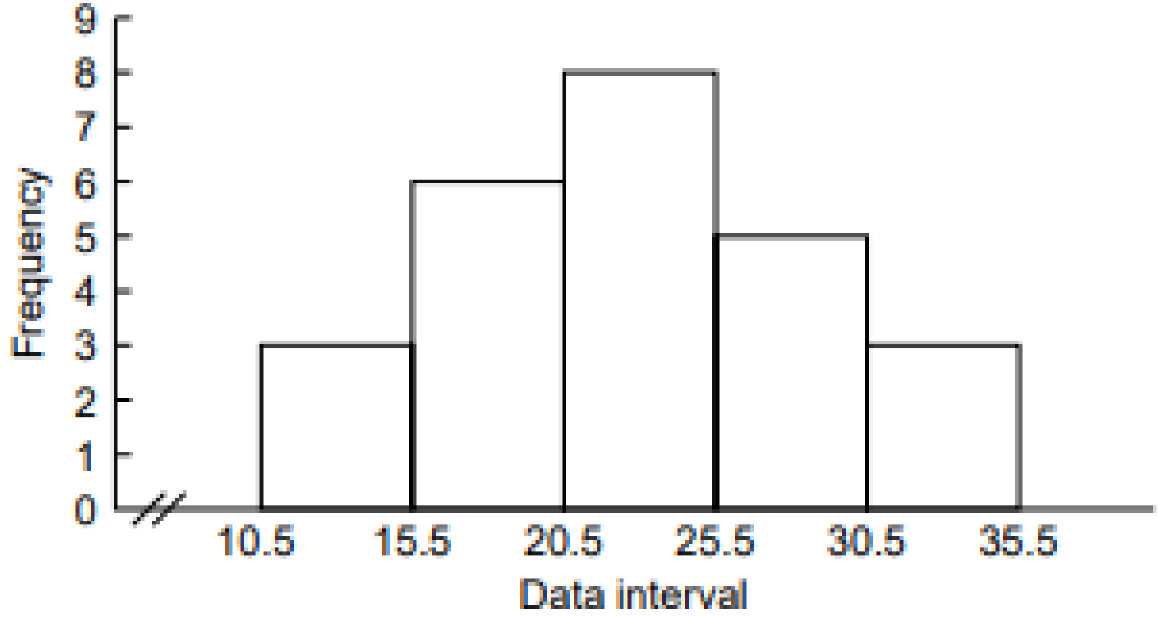
(a) Make a frequency table displaying class intervals, frequencies, relative frequencies, and percentages.

(b) Construct a frequency histogram.

Solution

(a) *We will use five classes. The maximum and minimum values in the data set are 35 and 11. Hence the class width is* $(35 - 11)/5 = 4.8 \simeq 5$. *Hence, we shall take the class width to be 5. The lower boundary of the first class interval will be chosen to be 10.5. With five classes, each of width 5, the upper boundary of the fifth class becomes 35.5. We can now construct the frequency table for the data.*

Class	*Class interval*	f_i = *frequency*	*Relative frequency*	*Percentage*
1	*10.5 – 15.5*	*3*	*3/25 = 0.12*	*12*
2	*15.5 – 20.5*	*6*	*6/25 = 0.24*	*24*
3	*20.5 – 25.5*	*8*	*8/25 = 0.32*	*32*
4	*25.5 – 30.5*	*5*	*5/25 = 0.20*	*20*
5	*30.5 – 35.5*	*3*	*3/25 = 0.12*	*12*



■ **FIGURE 1.4** Frequency histogram of impurity data.

2 Measures of Dispersion

The mean is the value usually used to indicate the centre of a distribution. If we are dealing with quantity variables our description of the data will not be complete without a measure of the extent to which the observed values are spread out from the average.

2.1 The Range

One very simple measure of dispersion is the range. Lets consider the two distributions given in Figures 3 and 4. They represent the marks of a group of thirty students on two tests.

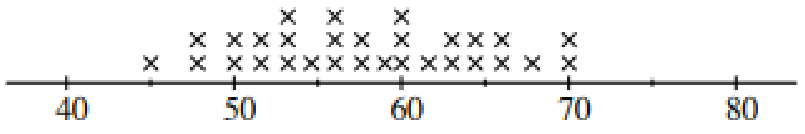


Figure 3: Marks on test A.

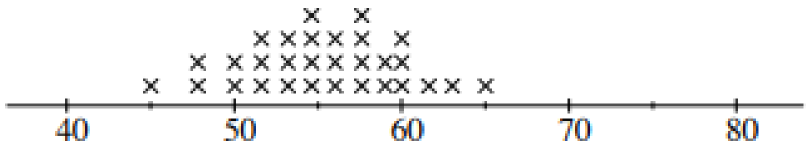


Figure 4: Marks on test B.

Here it is clear that the marks on test A are more spread out than the marks on test B, and we need a measure of dispersion that will accurately indicate this.

On test A, the range of marks is $70 - 45 = 25$.

On test B, the range of marks is $65 - 45 = 20$.

Here the range gives us an accurate picture of the dispersion of the two distributions.

However, as a measure of dispersion the range is severely limited. Since it depends only on two observations, the lowest and the highest, we will get a misleading idea of dispersion if these values are outliers. This is illustrated very well if the students' marks are distributed as in Figures 5 and 6.

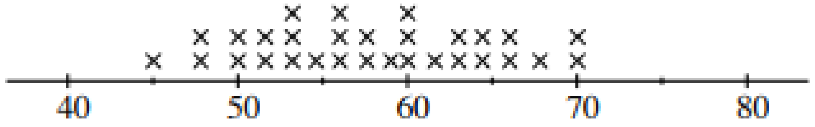


Figure 5: Marks on test A.

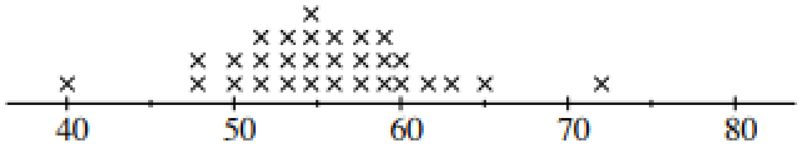


Figure 6: Marks on test B.

On test A, the range is still $70 - 45 = 25$.

On test B, the range is now $72 - 40 = 32$, but apart from the outliers, the distribution of marks on test B is clearly less spread out than that of A.

We want a measure of dispersion that will accurately give a measure of the variability of the observations. We will concentrate now on the measure of dispersion most commonly used, the standard deviation.

2.2 Standard Deviation

Suppose we have a set of data where there is no variability in the observed values. Each observation would have the same value, say 3, 3, 3, 3 and the mean would be that same value, 3. Each observation would not be different or *deviate* from the mean.

Now suppose we have a set of observations where there is variability. The observed values would deviate from the mean by varying amounts.

The standard deviation is a kind of average of these deviations from the mean.

This is best explained by considering the following example.

Take, for example, the following grades of 6 students:

$$56 \qquad 48 \qquad 63 \qquad 60 \qquad 51 \qquad 52.$$

Mean = 55.

To find how much our observed values deviate from the mean, we subtract the mean from each.

Observed values	56	48	63	60	51	52
Deviations from Mean	$+1$	-7	$+8$	$+5$	-4	-3

We cannot, at this stage, simply take the average of the deviations as their sum is zero.

$$(+1) + (-7) + (+8) + (+5) + (-4) + (-3) = 0$$

We get around this difficulty by taking the square of the deviations. This gets rid of the minus signs. (Remember $(-7) \times (-7) = 49$.)

Deviations	+1	−7	+8	+5	−4	−3
Squared deviations	1	49	64	25	16	9

We can now take the mean of these squared deviations. This is called the variance.

$$\text{Variance} = \frac{1 + 49 + 64 + 25 + 16 + 9}{6} = 27.33.$$

The variance is a very useful measure of dispersion for statistical inference, but for our purposes it has a major disadvantage. Because we squared the deviations, we now have a quantity in square units. So to get the measure of dispersion back into the same units as the observed values, we define standard deviation as the square root of the variance.

$$\text{Standard Deviation} = \sqrt{\text{Variance}} = \sqrt{27.33} = 5.228.$$

The standard deviation may be thought of as the 'give or take' number. That is, on average, the student's grade will be 55, give or take 5 marks. The standard deviation is a very good measure of dispersion and is the one to use when the mean is used as the measure of central tendency.

Example: Calculate the mean and standard deviation of the following set of data.

Birthweight of ten babies (in kilograms)

2.977 3.155 3.920 3.412 4.236 2.593 3.270 3.813 4.042 3.387

Solution:

Birthweight	Deviations from Mean	Squared Deviations
in kilograms	score − mean	(score − mean)^2
2.977	−0.5035	0.2535
3.155	−0.3255	0.1060
3.920	0.4395	0.1932
3.412	−0.0685	0.0047
4.236	0.7555	0.5708
2.593	−0.8875	0.7877
3.270	−0.2105	0.0443
3.813	0.3325	0.1106
4.042	0.5615	0.3153
3.387	−0.0935	0.0087
Sum = 34.805	Sum = 0	Sum = 2.3948

$$\text{Mean} = \frac{\text{sum of observations}}{\text{number of observations}} = \frac{34.805}{10} = 3.4805 = \mu.$$

$$\text{Variance} = \frac{\text{sum of squared deviations}}{\text{number of observations}} = \frac{2.3948}{10} = 0.2395 = \sigma^2.$$

$$\text{Standard Deviation} = \sqrt{\text{Variance}} = \sqrt{0.2395} = 0.4894 = \sigma.$$

2.3 The Interquartile Range

The interquartile range is another useful measure of dispersion or spread. It is used when the median is used as the measure of central tendency. It gives the range in which the middle 50% of the distribution lies. In order to describe this in detail, we first need to discuss what we mean by quartiles.

2.3.1 Quartiles

Suppose we start with a large set of data, say the heights of all adult males in Sydney. We can represent these data in a graph, which if smoothed out a bit, may look like Figure 9.

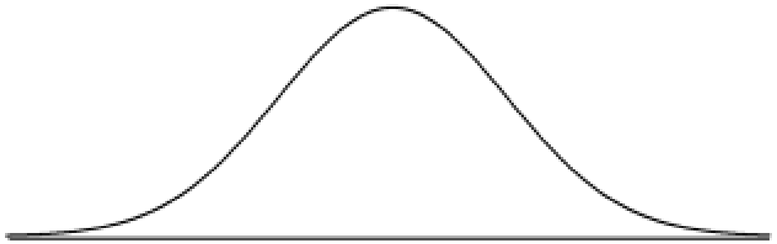


Figure 9: Graph representing heights of adult males.

As the name 'quartile' suggests, we want to divide the data into four equal parts. In the above example, we want to divide the area under our curve into four equal areas.

The second quartile or median

It is easy to see how to divide the area in Figure 9 into two equal parts, since the graph is symmetric. The point which gives us 50% of the area to the left of it and 50% to the right of it is called the second quartile or median. This is illustrated in Figure 10.

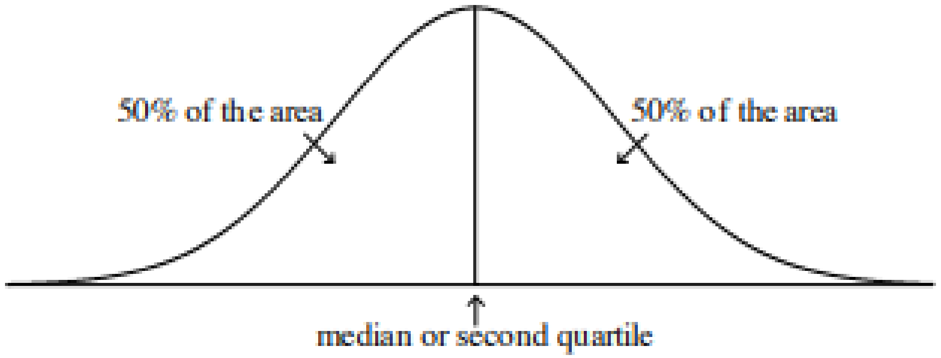


Figure 10: Graph showing the median or second quartile.

This exactly corresponds to our previous idea of median as the middle value.

The first quartile

The first quartile is the point which gives us 25% of the area to the left of it and 75% to the right of it. This means that 25% of the observations are less than or equal to the first quartile and 75% of the observations greater than or equal to the first quartile. The first quartile is also called the 25th percentile. This is illustrated in Figure 11.

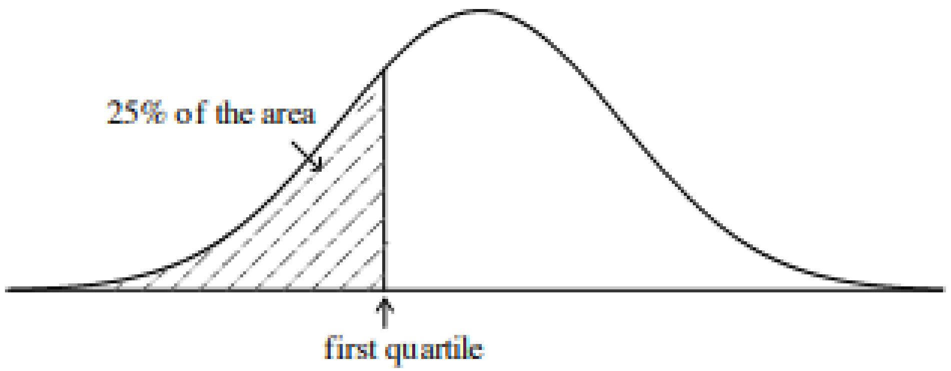


Figure 11: Graph showing the first quartile.

The third quartile

The third quartile is the point which gives us 75% of the area to the left of it and 25% of the area to the right of it. This means that 75% of the observations are less than or equal to the third quartile and 25% of the observation are greater than or equal to the third quartile. The third quartile is also called the 75th percentile. This is illustrated in Figure 12.

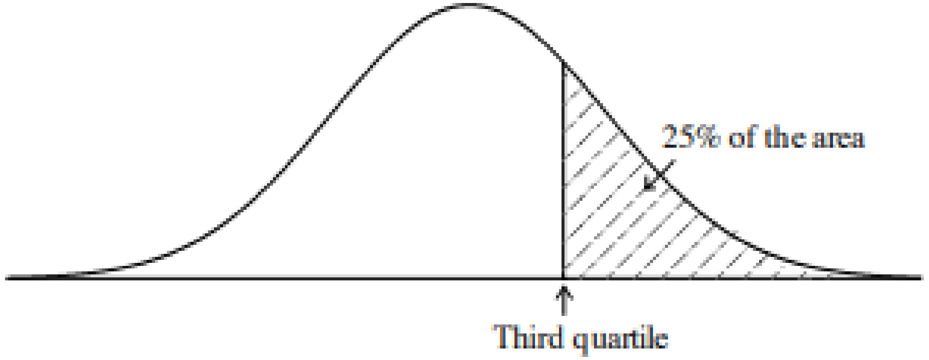


Figure 12: Graph showing the third quartile.

Summary

The first (Q_1), second (Q_2) and third (Q_3) quartiles divide the distribution into four equal parts. This is illustrated in Figure 13.

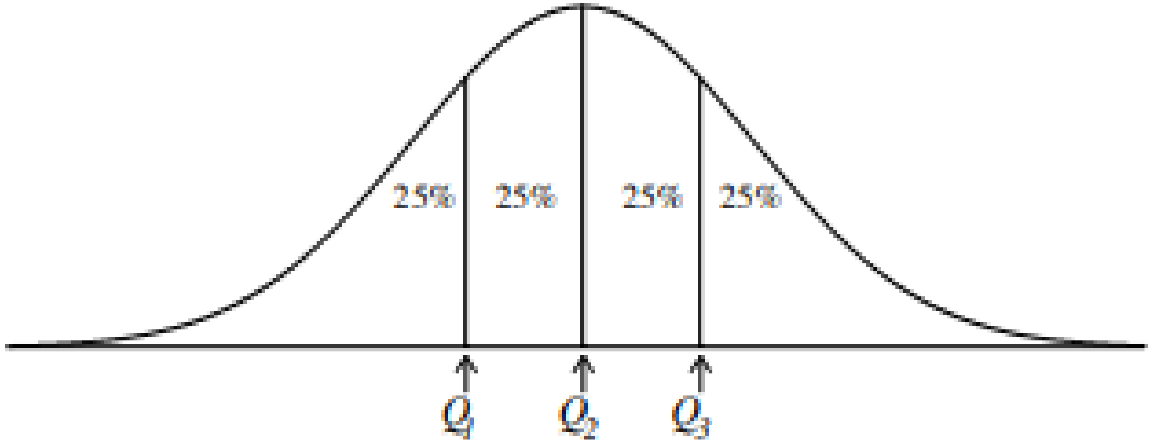


Figure 13: Graph showing all quartiles.

2.3.2 Quartiles for small data sets

Suppose now we have a small data set of twelve observations which we can write in ascending order as follows. (A data set, where the number of observations is a multiple of four, has been chosen to avoid some technical difficulties.)

$$15 \quad 18 \quad 19 \quad 20 \quad 20 \quad 20 \quad 21 \quad 23 \quad 23 \quad 24 \quad 24 \quad 25$$

In this case, we want to divide the data into four equal sets, so that there are 25% of the observations in each.

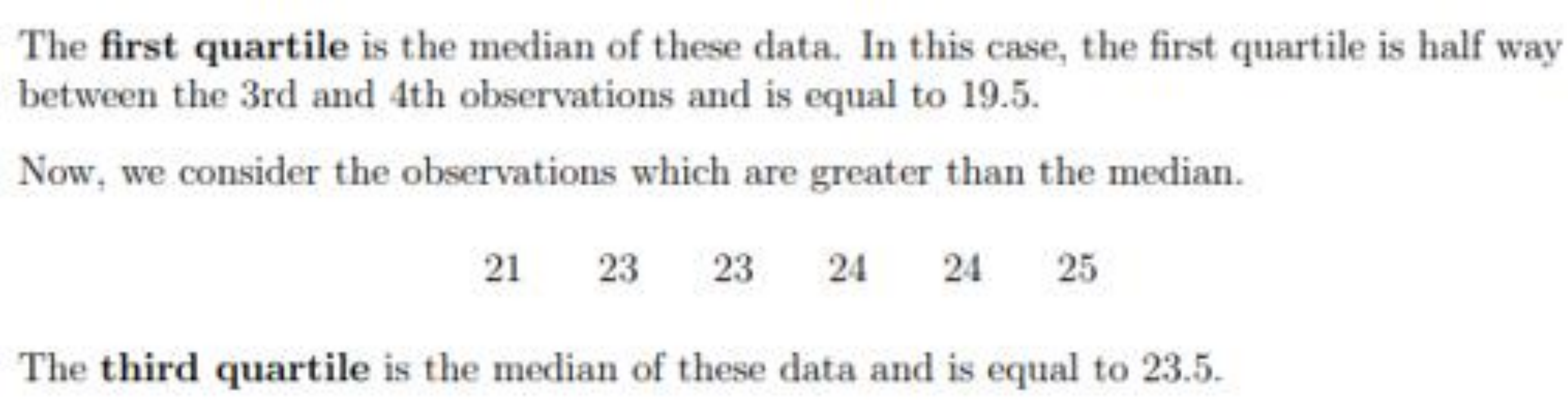
First, we find the median just as we did in Chapter 1.

$$15 \quad 18 \quad 19 \quad 20 \quad 20 \quad 20 \underset{\text{median}}{\uparrow} 21 \quad 23 \quad 23 \quad 24 \quad 24 \quad 25$$

The **median** is 20.5 (half way between the 6th and 7th observations), and divides the data into two equal sets with exactly 50% of the observations in each: the 1st to the 6th observations in the first set and the 7th to 12th observations in the other.

To find the first quartile we consider the observations less than the median.

$$15 \quad 18 \quad 19 \quad 20 \quad 20 \quad 20$$



The **first quartile** is the median of these data. In this case, the first quartile is half way between the 3rd and 4th observations and is equal to 19.5.

Now, we consider the observations which are greater than the median.

$$21 \quad 23 \quad 23 \quad 24 \quad 24 \quad 25$$

The **third quartile** is the median of these data and is equal to 23.5.

So, for our small data set of 12 observations, the quartiles divide the set into four subsets.

$$\begin{array}{ccccccccccccccc} 15 & 18 & 19 & \underset{Q_1}{\uparrow} & 20 & 20 & 20 & \underset{Q_2}{\uparrow} & 21 & 23 & 23 & \underset{Q_3}{\uparrow} & 24 & 24 & 25 \end{array}$$

We will now use the quartiles to define a measure of spread called interquartile range.

2.3.3 The interquartile range

The interquartile range quantifies the difference between the third and first quartiles. If we were to remove the median (Q_2) from Figure 13 we would have a graph like that in Figure 14.

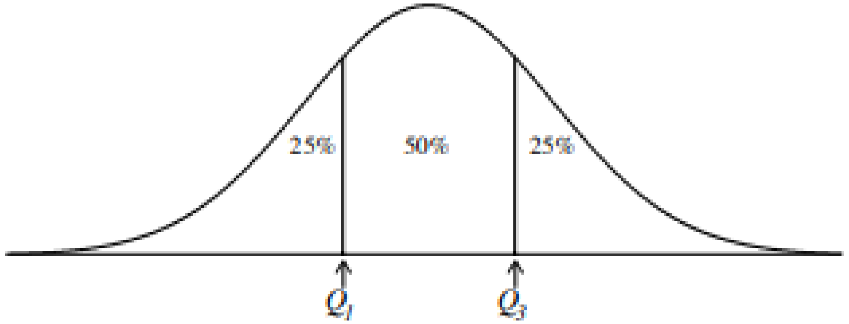


Figure 14: Graph showing the first and third quartiles.

From Figure 14, we see that 50% of the area is between the first and third quartiles. This means that 50% of the observations lie between the first and third quartiles.

We define the interquartile range as:

$$\text{The interquartile range} = \text{Third quartile} - \text{First quartile}.$$

For our small data set, the first quartile was 19.5 and our third quartile was 23.5. So, the interquartile range is $23.5 - 19.5 = 4$.

We will use the interquartile range later to draw a box-plot. For now we are interested in it as a measure of spread.

The interquartile range is particularly useful to describe data sets where there are a few extreme values. Unlike the range, and to a lesser extent the standard deviation, it is not sensitive to extreme values as it relies on the spread of the middle 50% of the distribution. So, if there are data sets which have extreme values, it can be more appropriate to use the median to describe central tendency and the interquartile range to describe the spread.

In the following exercises, data sets, where the number of observations is a multiple of four, have been given. Of course, the quartiles can be found for all other sized data sets, but we will restrict ourselves to these simple cases to avoid any technical difficulties. It is not necessary that you know how to calculate quartiles for all cases, but it is important that you understand the concept.

Formulae for Mean and Standard Deviation of a Population

The formula for the mean (average) of N observations is given by:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

where x_1 is the value of the first observation, x_2 is the value of the second observation, etc.

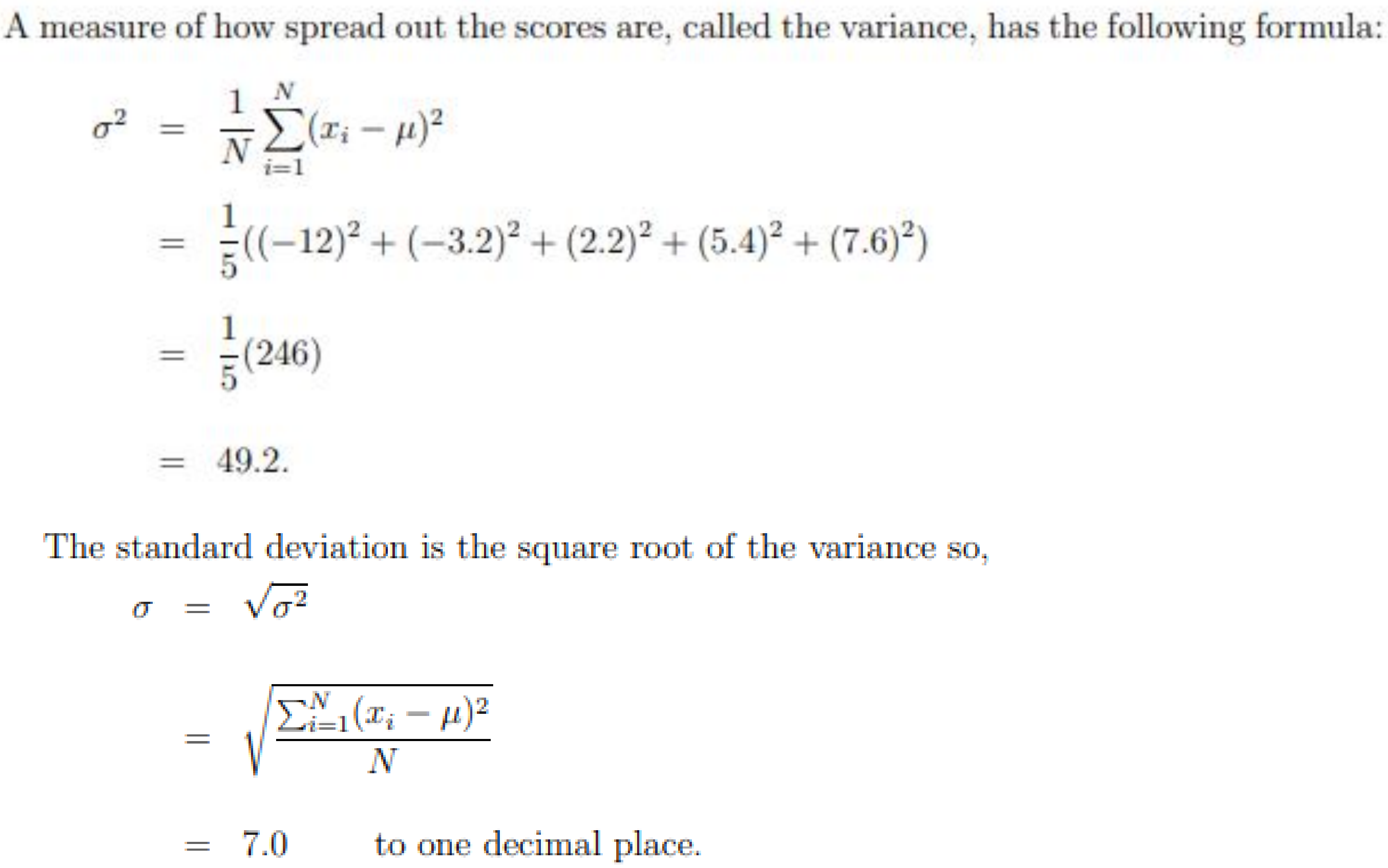
Example: The weights of five children in a family are:

$$x_1 = 3.5\text{kg} \qquad x_2 = 12.3\text{kg} \qquad x_3 = 17.7\text{kg} \qquad x_4 = 20.9\text{kg} \qquad x_5 = 23.1\text{kg}.$$

Find the mean and standard deviation of the weights of these children.

Solution:

$$\begin{aligned}
\mu = \frac{1}{N}\sum_{i=1}^{N} x_i &= \frac{1}{N}(x_1 + x_2 + x_3 + x_4 + x_5) \\
&= \frac{1}{5}(3.5 + 12.3 + 17.7 + 20.9 + 23.1) \\
&= \frac{1}{5}(77.5) \\
&= 15.5.
\end{aligned}$$



A measure of how spread out the scores are, called the variance, has the following formula:

$$\begin{aligned}\sigma^2 &= \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2\\ &= \frac{1}{5}((-12)^2+(-3.2)^2+(2.2)^2+(5.4)^2+(7.6)^2)\\ &= \frac{1}{5}(246)\\ &= 49.2.\end{aligned}$$

The standard deviation is the square root of the variance so,

$$\begin{aligned}\sigma &= \sqrt{\sigma^2}\\ &= \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}}\\ &= 7.0 \qquad \text{to one decimal place.}\end{aligned}$$

Estimates of the Mean and Variance

We have, so far, concerned ourselves with the mean, variance, and standard deviation of a population. These have been written using the Greek letters μ, σ^2, and σ respectively.

However, in statistics we are mainly concerned with analysing data from a sample taken from a population, in order to make inferences about that population. Our data sets are usually random samples drawn from the population.

When we have a random sample of size n, we use the sample information to estimate the population mean and population variance in the following way.

The mean of a sample of size n is written as $\overline{x}$ (read x bar).

To find the sample mean we add up all the sample scores and divide by the number of sample scores. This can be written using sigma notation as:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The sample mean is used to estimate the population mean. If we took many samples of size n from the population, calculated their sample means, and then averaged them, we would get a value very close to the population mean. We say that the sample mean is an unbiased estimator of the population mean.

An estimate of the population variance of a sample of size n is given by s^2 where

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Notice that we are dividing by $n - 1$ instead of n as we did to find the population variance. We need to do this because the value obtained if we divide by n, tends to underestimate the population variance. Calculated in this way, s^2 is an unbiased estimator of population variance. In fact, s^2 can be described as the *estimated population variance*. (It is sometimes called the 'sample variance' but this is strictly speaking not accurate.)