Descriptive Statistics

Contents

1 Statistics

1.1 Introduction – Descriptive Statistics

For the moment, we will take it to mean ‘the summary and description of large, possibly very large, sets of numbers.’

Our intention will be to describe the pattern in a set of numbers that may not be immediately obvious to us.

This is more correctly known as Descriptive Statistics.

We will also be referring to a set or list of number or values as a ‘dataset’.

Later on in your course, you will move on to the study of Statistics in the more specific (and powerful) sense of carrying out tests and analysis on lists of numbers being generated by an engineering, scientific or business process.

1.2 Indexed lists and Summation Notation

It is necessary to introduce some notation which will make it easier for us to denote a large lists, and to indicate that the numbers in the list are, for example, to be added.

Consider a list of 40 figures, for example, the ages of a group of students.

Let us use one symbol to denote the age of a student, say x.

The ages of each student can then be denoted by the letter x, with a subscript assigned to set them apart.

Thus the 40 ages are referred to as

$$x_1, x_2, x_3, \ldots, x_{40}.$$

Each of the symbols x_1, x_2, x_3, up to x_{40}, is a separate variable representing a particular student's age.

All the numbers in this list are ages of students; the same symbol represents them all, they are all to be dealt with together as part of one list.

We can now talk about the list of 40 figures as

$$x_1, x_2, \ldots, x_{40},$$

or they can be written as the figures

x_i, where $i = 1$ to 40.

In the second case, the variable i is called the index, or an index variable.

It is also known informally as a counter variable.

1.2.1 Summation

The most important extension of this notation is to allow a way of indicating that the numbers in a list are being added.

For our list

$$x_1, x_2, \ldots, x_{40},$$

we will indicate that these 40 numbers are to be added by writing the following expression:

$$\sum_{i=1}^{40} x_i \ .$$

The symbol Σ should be thought of as an *instruction* to add the numbers in the list. The symbol Σ, called sigma, is the Greek equivalent of capital S.

In this notation, the expression '$i = 1$' is written below the Σ and '40' is written above it means the summation goes from x_1 to x_{40}.

Thus the result of the calculation is the total of the values in the list.

1.2.2 Shorthand in summation

If it is understood that all the numbers in a given list are to be added, the sum can be written as:

$$\sum_i x_i \ .$$

The general variable *x* and the index variable *i* are shown in the expression.

To take this one step further, if it is clear that there is only one index variable present, the index variable could even be dropped from the summation symbol, to leave

$$\Sigma\, x_i.$$

Since there is only one index variable, it is the one listing the numbers to be added.

1.2.3 Example – Student Exam Results

Let x_i, for $i = 1$ to 20, be the percentage marks of 20 students in an assessment:

85, 65, 40, 55, 64, 75, 80, 66, 57, 86,

47, 94, 81, 72, 83, 51, 63, 77, 36, 68.

Find the value of

$$\sum_i x_i \, .$$

Solution:

This expression is telling us to add up the list of 20 numbers.

Carrying out this instruction gives 1,345, so we can write:

$$\sum_i x_i = 1{,}345 .$$

Since it was understood that we are to add all the numbers in the list, this result can be written more simply as

$$\Sigma\, x_i = 1{,}345.$$

1.2.4 More Arithmetical Operations

Consider the quantity

$$\sum_i x_i^{\,2} \,.$$

Since a power takes precedence over addition in the rules of arithmetic, with no brackets, this means the values of x are being squared before they are added.

Thus each number is squared, and the results added up. Therefore this instruction means ‘the sum of the squares of the numbers in the list.’

In the case of the numbers listed above for the marks in the assessment, the result of this calculation is 95,295. Thus

$$\sum_i x_i^{\,2} = 95{,}295\,.$$

An important aspect of this instruction Σx_i^2 is that there were no brackets; the squared therefore referred to each individual number in the list to be squared.

Now the same summation is presented, but now there are brackets around the sum:

$$\left(\sum_i x_i \right)^2 .$$

Following the normal rules of precedence in arithmetical operations, this means the numbers have been added, and only then is the result squared.

For example, in the case of the assessment marks,

$$\left(\sum_i x_i \right)^2 = 1{,}345^2 = 1{,}809{,}025 .$$

This is ‘the square of the sum.’

Generally the first number, the sum of squares, is smaller than this number, the square of the sum. For a list of n numbers x_i, the difference between the two quantities

$$n\sum_i x_i^2 \text{ and } \left(\sum_i x_i\right)^2.$$

is a vitally important quantity. As we will see later it is a measure of how widely distributed the numbers are.

1.3 Descriptive Statistics

Given a large set of figures, for example a list of sales figures or ages, simply looking at the numbers themselves can tell very little about any trends or patterns in the data.

The first job of Descriptive Statistics is to describe or summarise such a collection of figures, in order to give some idea of what they mean.

1.3.1 The Mean

Consider a list of n figures x_1, x_2,…, x_n. The first number we might quote to describe this data list is given by:

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

This is the total of the values, divided by the number of values on the list. It is the widely understood concept of the average.

In mathematics and statistics, this is more formally known as the *mean*.

The symbol for the mean of an indexed list of values x_i, is the symbol x with a bar over the symbol. This is referred to as ‘x bar’.

1.3.2 The Standard Deviation

This one value is however limited. It does not give us any information about how widely spread the numbers in a given list are; they could all be close to the average or there could be a very widely distributed set of numbers.

A second number is needed to quantify this.

The most important such measure is the *standard deviation*, usually denoted *s*, which is defined by the following equation:

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

This number takes the difference between each value of *x* and the average, squares this difference, adds them, and then divides by *n*-1.

Squaring the differences ensures they build up rather than cancel, and so, crucially, we know that *s* is always a positive number.

The top line of this equation can be changed, with a bit of algebra, to give the alternative equation for *s*:

$$s^2 = \frac{\sum_{i=1}^{n} x_i^{\,2} - n\left(\bar{x}\right)^2}{n-1}.$$

Take a note that the expression for the top line means that it is found by calculating the sum of squares first, then subtract the product $n(\bar{x})^2$.

This version of the equation requires fewer steps in the calculation. The first equation for *s* should be taken as being used for the definition of the idea of the standard deviation; the second for calculation of *s*.

1.3.3 Example of Two Standard Deviations

Recall the list of the percentage marks of students in an assessment mentioned above:

85, 65, 40, 55, 64, 75, 80, 66, 57, 86,

47, 94, 81, 72, 83, 51, 63, 77, 36, 68.

We can calculate the average and standard deviation for these figures, and do so quickly, using the second, simpler form of the equation for *s*.

The first step is finding the sum of the numbers:

$$\sum_i x_i = 1{,}345$$

Then the calculation for the mean is:

$$\bar{x} = \frac{\sum_i x_i}{n} = \frac{1{,}345}{20} = 67.25.$$

In the equation for the standard deviation, we will need the sum of the squares, in other words:

$$\sum_i x_i^{\,2} .$$

So square each number in the list, then add, to get (check this yourself):

$$\sum_i x_i^{\,2} = 95{,}295 .$$

Then the standard deviation equation gives:

$$s^2 = \frac{\sum_i x_i^{\,2} - n\left(\bar{x}\right)^2}{n-1} = \frac{95,295 - 20 \times 67.25^2}{19}.$$

The details of the top-line calculation are:

$$67.25^2 = 4522.5625,$$

and so then multiplying by 20 gives 90,451.25. Then

$$s^2 = \frac{95,295 - 90,451.25}{19}.$$

Taking the difference and dividing in the 19 gives

$$s^2 = \frac{4843.75}{19} = 254.934.$$

Taking the square root to get s gives $s = 15.97$.

This number is quite high, indicating that the percentage marks were quite widely distributed. The standard deviation reflects this variety in the numbers.

Now for an example of the opposite case, consider the following list of ages in a first year college class. There are:

- 7 students aged 17,
- 20 aged 18,
- 11 aged 19, and
- 2 are aged 20.

Adding the numbers gives:

$$7x17 + 20x18 + 11x19 + 2x20 = 139 + 360 + 189 + 40 = 728.$$

Thus the average age is:

$$728/40 = 18.2.$$

To get the standard deviation, the sum of the squares is needed. It is:

$$7x17^2 + 20x18^2 + 11x19^2 + 2x20^2 =$$

$$= 7x289 + 20x324 + \ldots + 2x400 = 13{,}274.$$

The standard deviation is then given by putting this value into the equation:

$$\frac{13{,}274 - 40 \times 18.2^2}{39} = \frac{13{,}274 - 13{,}249.60}{39} = \frac{24.4}{39}.$$

so $s^2 = 0.63$, and the standard deviation $s = 0.79$.

It stands out immediately that this second figure is much smaller than the first.

This reflects the fact that the group of marks were very widely distributed, whereas the list of ages were bunched very close together in a group of 18s, 19s and so on.

1.3.4 Another example of two Standard Deviations

Here is an example where two lists of numbers have the same mean, but their standard deviations distinguish between them.

List A: 12, 23, 45, 47, 62, 78, 34, 16, 73, 23, 67, 11, 12, 14.

List B: 48, 34, 51, 32, 41, 35, 36, 37, 38, 32, 39, 28, 40, 27.

Let us use:

a_i, i = 1 to 14 to refer to the first list, and:

Use b_i, i = 1 to 14, for the second.

Then a quick calculation (check this yourself) gives:

$$\Sigma\, a_i = 517, \quad \Sigma\, b_i = 518.$$

The two averages are then $\bar{a}$ = 36.93, and $\bar{b}$ = 37.

Now find the sum of squares for each list and from that, the value of s:

$$\Sigma\, a_i^2 = 27{,}055,$$

so the top line is

$$27{,}055 - 14(36.92)^2 = 7{,}962.73.$$

Then

$$s_A^2 = 7962.73/13 = 612.53,$$

so $s_A = 24.75$.

For list B, the calculation is similar:

$$\Sigma\, b_i^2 = 19{,}758,$$

so the top line is

$$19{,}758 - 14(37)^2 = 592.$$

Then

$$s_B^2 = 592/13 = 45.54,$$

so $s_B = 6.75$.

1.4 Grouped Data

We will now look at alternative ways in which large or very large data sets can be presented or gathered, and how we can calculate estimates for our concepts of the mean and standard deviation in this case.

We will illustrate these ideas with an example first, and then present equations based on the same ideas.

Consider the following case of 449 people attending a film.

Age classes	*Frequencies*
10 – 20	51
20 – 30	120
30 – 40	150
40 – 50	75
50 – 60	43
60 – 70	10

This type of information is known as grouped data.

The numbers in each group are called *frequencies*, and such grouped data is also known as a *frequency distribution*.

1.4.1 An Important Convention

One very important point with this example of group data is illustrated by asking: in which group is a person of age exactly 30 years old counted?

The convention we will adopt for our work is that the group ’20 – 30’ includes all ages of 20 or above, up to *but not including* 30.

Mathematically it is the set:

{ The set of all numbers x such that $x \geq 20$ and $x < 30$ }

This means that the number 30 goes into the 30 – 40 age group.

1.4.2 Estimating the Mean and Standard Deviation

The question now arises as to how we can calculate a value for the mean or the standard deviation for a frequency distribution.

Because the original values are not available, the mean and standard deviation must be estimated.

The first calculation is to find the total number of values, *n*. This is simply the sum of the frequencies.

Mathematically, let f_i, for $i = 1$ to 6, be the frequencies; then

$$n = \sum f_i = 449.$$

This value n is the total number of people in the cinema.

To estimate the sum of all the ages, work as if each person in the age group 10 to 20 is age 15, each person in 20 to 30 is 25, and so on.

In other words, we will act as though each of the people in a group has the mid-point value as an age.

So we will multiply each midpoint by the number of people in that group, which is the frequency.

We are now working from this table:

Age classes	*Frequencies*	*Midpoints*
10 – 20	51	15
20 – 30	120	25
30 – 40	150	35
40 – 50	75	45
50 – 60	43	55
60 – 70	10	65

The sum of each midpoint times the corresponding frequency is:

$51x15 + 120x25 + 150x35 + 75x45 + 43x55 + 10x65 = 15{,}405.$

With this estimate of the sum, we can now estimate the mean by dividing it by the total *n*:

$$15{,}405/449 = 34.31.$$

We now need an approximation of the sum of the squares.

Now each mid-point is now squared and then multiplied by the corresponding the frequency:

$$51x15^2 + 120x25^2 + 150x35^2 + 75x45^2 + 43x55^2 + 10x65^2 = 594{,}425.$$

With this estimate for the sum-of-squares, the standard deviation can now be calculated.

Substitution in the equation for the standard deviation gives:

$$\frac{594{,}425 - 449 \times (34.31)^2}{448}$$

So s^2 is

$$s^2 = 65873/448 = 147, \text{ so } s = 12.12.$$

1.4.3 Summary – Parameters for Grouped Data

Let the numbers f_1, f_2, f_3,… , be the frequencies, and let the numbers m_1, m_2, m_3,… be the midpoints of the groups.

The total number of values n is naturally found by adding the frequencies:

$$n = \sum_i f_i .$$

The *frequency average* is then

$$\bar{x} = \frac{\sum_i m_i f_i}{n}.$$

In a similar way, the *frequency standard deviation* can be calculated as

$$s^2 = \frac{\sum_i m_i^2 f_i - n\left(\bar{x}\right)^2}{n-1}.$$

1.4.4 Mean and Standard Deviation Calculations

The following are the lifetimes of machines produced by two companies labelled A and B.

Find the average and standard deviation for each data set.

Lifetimes (months)	*A*	*B*
0 to 5	2	58
5 to 10	8	25
10 to 15	19	20
15 to 20	59	7
20 to 25	26	6
25 to 30	7	4
30 to 35	3	2

For the first data set, the total of the frequencies is $n = 124$.

The sum of the midpoints times the frequencies is:

$$2.5x2 + 7.5x8 + \ldots + 32.5x3 = 2{,}210.$$

Dividing this by 124, gives the estimate for the mean of 17.82.

The estimate of the standard deviation goes as follows.

The ‘sum of the squares’ is:

$$\sum_i m_i^{\,2} f_i = 2.5^2 \text{ x2} + 7.5^2 \text{ x8} + \ldots + 32.5^2 \text{ x3} = 43{,}125.$$

Putting this in the equation for *s*:

$$s^2 = \frac{\sum_i m_i^{\,2} f_i - n\left(\bar{x}\right)^2}{n-1} = \frac{43{,}125 - 124 \times 17.82^2}{123}.$$

This works out to be $s^2 = 3{,}748.5/123 = 30.476$, so $s = 5.52$.

For the second data set, the sum of the frequencies gives

$$n = 122.$$

The first sum, of the midpoints times the frequencies, is

$$2.5\text{x}58 + 7.5\text{x}25 + \ldots + 32.5\text{x}2 = 1{,}015.$$

This giving an estimate for the mean of 1,015/122 = 8.32.

The 'sum of the squares' is:

$$\sum_i m_i^{\,2} f_i = 2.5^2 \text{ x}58 + 7.5^2 \text{ x}25 + \ldots + 32.5^2 \text{ x}2 = 15212.5.$$

Putting this in the equation for s:

$$s^2 = \frac{\sum_i m_i^2 f_i - n\left(\bar{x}\right)^2}{n-1} = \frac{15212.5 - 122 \times 8.32^2}{121}.$$

This works out to be $s^2 = 55.93$, so $s = 7.48$.

1.5 Medians and Quartiles

There are other numbers that help us describe the distribution of a large dataset called the median and the quartiles.

The essential difference between these numbers and the mean and standard deviation is that the median and quartiles are concerned with the order of the numbers in a given dataset.

1.5.1 Definition of the Median

For a given data set, when the numbers have been arranged in order, the median is that value which is mid-point in the data. Half of the numbers are greater than median, half are less.

This is straightforward in the case of an odd number of values, as in the following case:

1, 4, 5, 6, 10, 11, 12.

These numbers are arranged in order, so it can be seen that 3 numbers are above 6, and 3 below. Thus 6 is the median value.

The situation is slightly different if there are an even number of values, for example

4, 5, 6, 9, 12, 23, 25, 30.

Here the 9 is the 4th number, the 12 the 5th, out of 8 numbers.

So the midway point of the data lies between these two values.

The value quoted for the median in this case is the midpoint of 9 and 12:

$$(9+12)/2 = 10.5.$$

1.5.2 The Quartiles

Once a list of numbers has been divided into two groups by the median, the quartiles take this a step further.

The first quartile divides the lower half, the third quartile divides the upper group, in exactly the same way as the median did for the original list.

1.5.3 The Median and Quartiles for a List of Numbers

Find the median and quartiles for the following list of numbers:

7, 8, 12, 6, 5, 3, 9.

The first step is to arrange them in order:

3, 5, 6, 7, 8, 9, 12.

There are 7 numbers, so the median is the 4^{th} : 7.

To get the first quartile, the lower half of the list of numbers are:

3, 5, 6, 7.

The question arises whenever there are an odd number of values as to whether the 7 should be included in the lower list or the higher list.

The convention adopted is that it will be included in both.

The quartile is then between 5 and 6, so it is (5+6)/2 = 5.5.

For the third quartile, the upper half are:

7, 8, 9, 12.

The quartile is between 8 and 9, so it is (8+9)/2 = 8.5.

1.5.4 Another Example of Median and Quartiles

Find the median and quartiles of the following set of heights:

1.7, 1.8, 2.0, 1.6, 1.5, 1.8, 1.9, 1.7.

The first step: putting these numbers in order gives:

1.5, 1.6, 1.7, 1.7, 1.8, 1.8, 1.9, 2.0.

The calculations are:

- The median will be between 1.7 and 1.8, and so is 1.75.
- The first quartile will be between 1.6 and 1.7; 1.65.
- Similarly the third is between 1.8 and 1.9, and so is 1.85.

The diagram shows the results for this example and represents how the median and quartiles divide up the list of numbers into ‘quarters’:

1.5, 1.6,	1.7, 1.7,	1.8, 1.8,	1.9, 2.0
1.65	1.75	1.85	

1.5.5 Comparing the Average with the Median

The value of the mean compared to the median tells us something about the distribution of the data. Consider the data set from the previous example:

1.5, 1.6, 1.7, 1.7, 1.8, 1.8, 1.9, 2.0.

The mean is 1.75, and we saw the median was 1.75; the two numbers are identical.

Now consider what happens if the two last numbers are significantly increased:

1.5, 1.6, 1.7, 1.7, 1.8, 1.8, 3.9, 4.0.

If we recalculate the mean, it is now 2.25.

But note that the median has not changed.

The larger numbers at the higher end of the list (and this is where the order becomes important) have brought up the mean but not the median.

Therefore if the median is less than the mean, we can conclude that a small number of high values are dragging up the mean.

A similar example could be constructed by changing the lower numbers in the original list.

This illustrates the idea of a list of numbers being ‘skewed’, that is, having a ‘tail’ towards the higher or lower end of the values.

The conclusions are phrased as follows:

- If the mean is close to the median, then there are as many figures above average as below. The list is not skewed.
- If the mean is closer to the 1st Quartile then a few of the lower figures are ‘dragging’ the average down. This means the list is skewed to the left, the lower values.

- If it is closer to the 3rd Quartile then a few of the higher numbers are pulling the average up. This means the list is skewed to the right, the higher values.

These ideas will become very important in grouped data.

1.6 The Median for Grouped Data

In the case where grouped data is available or preferred, the median and quartiles for grouped data must be estimated in a similar way to the mean and standard deviation.

Recall the frequency distribution of ages of a group attending a film:

Age classes	*Frequencies*
10 – 20	51
20 – 30	120
30 – 40	150
40 – 50	75
50 – 60	43
60 – 70	10

An estimate could be found, if we see what its place is within one of the groups, and from this, estimate how far it is above that groups' lower bound.

To see which group the median is in, we need not just the number in each group, the frequencies, but also the number in a given group plus all those in previous groups. The results are shown here:

Age classes	*Frequencies*	*Cumulative Frequencies*
10 – 20	51	51
20 – 30	120	171
30 – 40	150	321
40 – 50	75	396
50 – 60	43	439
60 – 70	10	449

These are called the *cumulative frequencies.*

To estimate the median, there are 449 numbers, so the median would normally be the 225^{th} number.

From the table above of cumulative frequencies, we see that age group 30 – 40 has the 172^{nd} to the 321^{st} numbers, so the median is in this group.

The next step is to find the median's place within this age group.

It is the 225^{th} number, there are 171 in the previous 2 groups, so the place of the median is the

$$225 - 171 = 54^{th} \text{ number.}$$

We have found that the median is the 54^{th} number out of the 150 numbers in group 30 – 40.

To estimate its value, the best we can do is assume that the numbers *within* this group are evenly spread out.

Since the group goes from 30 to 40, that is, 10 years, we can say the value of the median is

$$\frac{54}{150} \text{ times } 10 \text{ years}$$

into the group, in other words, above the lower bound. In summary it is

$$\frac{54 \times 10}{150} + 30 \text{ years.}$$

Alternatively we had that the median is the 54^{th} number on the 30 – 40 age group.

Again, if the numbers *within* the group were evenly spread out, the mean separation between them would be 10yrs/150.

The median is the 54^{th} number in the group, so the estimate is

$\frac{10}{150}$ years times 54, above the lower bound.

Thus is the required estimate for the median is:

$$\frac{10 \times 54}{150} + 30 \text{ years.}$$

Either way, we arrive at the same estimate. The calculation gives:

$$30\text{yrs} + 3.6\text{yrs} = 33.6\text{yrs.}$$

Repeating this analysis for the general case will give us an equation for this estimate for any set of grouped data.

1.6.1 Summary – The Median for Grouped Data

For a grouped data set, let n be the total number of values in the data.

The median is the midpoint of this data, so it is value will be the number in place $(n + 1)/2$, to stay consistent with our definition of the median for even or odd lists of numbers.

Having decided which group the median is in, using the cumulative frequencies, the Median is given by the equation

$$\text{Median} = \left(\frac{n+1}{2} - cf \right) \frac{w}{f} + L$$

where the symbols are:

- cf is the cumulative frequency of the class before that of the median value,
- f is the frequency of the class containing the median value,

- L is the lower bound of this class,
- w is the width of each class.

To see why this equation gives a useful estimate of the mean, consider a data set of n figures for which a frequency distribution is available. Decide, using the cumulative frequencies, which class it falls in.

- Let w be the width of the classes (this is usually constant), let L be the lower boundary of this class and let f be the frequency.
- The values of the data in this class run from L to $L + w$.
- If cf is the cumulative frequency of the class before, and there are f data values in this class, then the numbers in this class are the $cf + 1^{st}$ to the $cf+f^{th}$ places.
- The median is the $(n + 1)/2 – cf$ number in this class.

Now assume that the data points are evenly distributed within this class. They are each separated by a step of

$$w/f.$$

Multiplying this increment by $(n + 1)/2 - cf$, the place number of the median, tells us how much above the lower bound the median value is.

The estimate is then the place number, times the increment, added to the lower bound. The resulting equation for the estimate *MD* is then

$$MD = L + \left(\frac{n+1}{2} - cf \right) \frac{w}{f} .$$

1.6.2 The Quartiles for Grouped data.

To calculate the quartiles, return to the cumulative frequencies and decide which class the $(n + 1)/4$ and $3(n + 1)/4$ data points occur in. The first and the third quartile are then given by the equations:

$$\text{Q1} = \left(\frac{n+1}{4} - cf \right) \frac{w}{f} + L \text{ and } \text{Q3} = \left(\frac{3(n+1)}{4} - cf \right) \frac{w}{f} + L$$

Where the symbols have the same meaning for the group the quartile is in.

1.6.3 An Example of Medians and Quartiles

The following table shows the lifetimes of components produced by a company, sorted into groups.

The cumulative frequencies have also been calculated.

Find the median and quartiles for this data set.

Classes	*Frequency*	*Cumulative Frequency*
0 to 5	9	9
5 to 10	15	24
10 to 15	20	44
15 to 20	30	74
20 to 25	20	94
25 to 30	6	100

In this example, there are 100 numbers, so the median would be between the 50^{th} and 51^{st} numbers. We will use 50.5. The first three groups contain 44 numbers between them, so the 50^{th} is in the 4^{th} group, 15 to 20. The numbers *f*, *cf*, *w*, and *L* must now be identified for this group.

- The frequency of the class containing the median value is 30, so: $f = 30$.

- The cumulative frequency of the previous class is 44 (remember, the 50^{th} number comes after the 44^{th}, which told us which group to look at) , so: $cf = 44$.
- The lower bound of this class is 15, so: $L = 15$.
- The width of the classes is 5: $w = 5$.

The Median is given by the equation:

$$\text{Median} = \left(\frac{n+1}{2} - cf \right) \frac{w}{f} + L$$

Putting in the values gives:

$$\text{Median} = \left(\frac{101}{2} - 44\right)\frac{5}{30} + 15 \; = (50.5 - 44)\frac{5}{30} + 15$$

$$= \frac{6.5 \times 5}{30} + 15 = 1.1 + 15 = 16.$$

The median is then 16.1.

For the first quartile, $(n + 1)/4$ is 25.25, so Q_1 would normally be just after the 25^{th} number.

This means it is in the group 10 to 15.

- Again, the value of w is 5.
- The frequency is $f = 20$.
- The value of the lower bound is $L = 10$.
- The previous cumulative frequency is $cf = 24$.

The first quartile is given by the equation:

$$\mathrm{Q1} = \left(\frac{n+1}{4} - cf \right) \frac{w}{f} + L \, .$$

This becomes:

$$\mathrm{Q1} = (25.25 - 24)\frac{5}{20} + 10 = 1.25 \times \frac{5}{20} + 10 = 0.31 + 10 = 10.31$$

The first quartile is 10.31.

For the third quartile, $3(n+1)/4$ is 75.75, so Q3 would be between the 75^{th} and 76^{th} numbers. This means it is in the group 20 to 25.

- As before, $w = 5$.
- The frequency is $f = 20$.
- The value of the lower bound is $L = 20$.
- The previous cumulative frequency is $cf = 74$.

The third quartile is given by the equation:

$$\mathrm{Q3} = \left(\frac{3(n+1)}{4} - cf \right) \frac{w}{f} + L \, .$$

This becomes:

$$\mathrm{Q3} = (75.75 - 74) \frac{5}{20} + 20$$

$$= 1.75 \times \frac{5}{20} + 20 = 0.44 + 20 = 20.44$$

The third quartile is 20.44.

1.7 Graphs

The aim of statistics is to summarise large sets of numbers, so they can be quickly understood. So far we have used particular numbers or parameters, such as the mean, median and standard deviation, to do this. The next step is to illustrate the information visually, with graphs. We will start with the frequency distribution of the cinema audience:

Age classes	*Frequencies*	*Cumulative Frequencies*
10 – 20	51	51
20 – 30	120	171
30 – 40	150	321
40 – 50	75	396
50 – 60	43	439
60 – 70	10	449

We will look at three different ways of representing this information.

Our intention here is not to set out definitive rules for constructing all possible graphs, but rather to adopt conventions or agreements on how the information is shown, in particular, conventions which are consistent with or reflect the logic behind the information.

1.7.1 The Frequency polygon

The first way of showing this data in graphical form is by plotting the frequencies of the classes.

The age groups are marked out on a horizontal axis, by indicating the bounds (or limits). The frequencies will be used for the vertical axis.

However, each frequency is defined for a group, rather than one point.

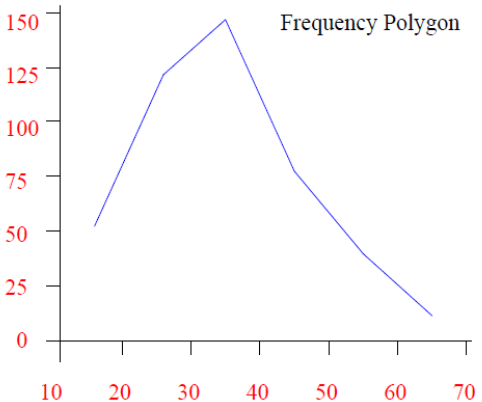
There are two approaches to how the graph is then to be drawn.

The first option is to take a representative value for each class, and plot the frequency above this value. The obvious representative value is the midpoint.

These points are then joined to form a graph composed of straight lines. The graph to be drawn is now a plot of the midpoints and the frequencies. The points used are then:

(15, 51), (25,120), (35,150), … (65,10).

These points are marked in on the graph and linked up with straight lines. The result is a **Frequency polygon**, shown in the diagram.


Frequency Polygon
150
125
100
75
50
25
0
10
20
30
40
50
60
70

The conventions for the frequency polygon are:

1. Draw a horizontal axis with notches or ticks indicating the boundaries of the groups.
2. Draw a vertical axis intersecting the horizontal axis at the first lower bound.
3. Scale the vertical axis appropriately for the values in your frequency distribution.
4. For each group, plot a point representing the frequency over the mid-point of that group.

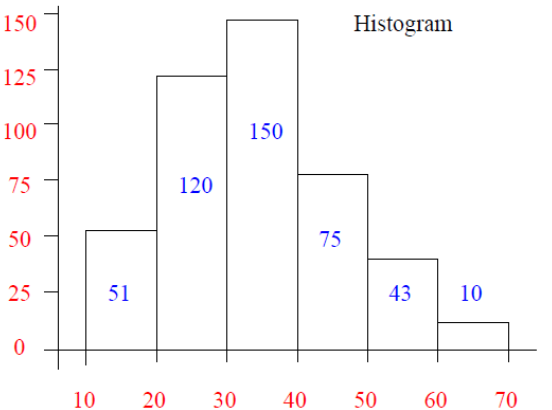
5. Join these points with straight lines.

1.7.2 The Histogram

The second approach to representing the frequencies is to draw the axes as before, and draw a **box** over each class, to the height of the appropriate frequency. This type of graph is called a **histogram.** The conventions for the histogram are:

1. Draw a horizontal axis with notches or ticks for the boundaries of the groups.
2. Draw a vertical axis intersecting the horizontal axis a small space to the left of the first lower bound.

3. Scale the vertical axis appropriately for the values in your frequency distribution.
4. For each group, draw a box to the height of the frequency over the group, the width going from the lower to the upper bound of the group.
5. The boxes for the group should have no spaces between them.


Histogram
150
125
100
75
50
25
0
51
120
150
75
43
10
10
20
30
40
50
60
70

1.7.3 Cumulative Frequency Polygon

Having drawn a graph of the frequencies, the next option is to illustrate the cumulative frequencies. Consider the following:

- If 51 is the cumulative frequency for the group 10 to 20, then there are 51 people below the age of 20.
- If 171 is the cumulative frequency for the group 20 to 30 then there are 171 people below the age of 30.

In other words, the cumulative frequency for each age class represents the number of ages below the upper bound of that class. This suggests plotting the cumulative frequencies of each group over the upper bound of the group.

To be consistent with this idea, the graph should be started at 0, plotted over the lower bound of the first group.

The points being plotted are then:

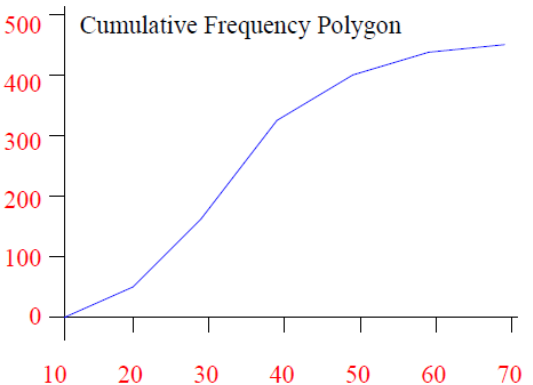
(10,0), (20,51), (30,171), (40,321), (50,396), (60,439), (70,449).

The conventions for the cumulative frequency polygon are:

- Draw a horizontal axis with notches or ticks showing the boundaries of the groups.
- Draw a vertical axis intersecting the horizontal axis at the first lower bound.

- Scale the vertical axis appropriately for the values in the cumulative frequency distribution.
- The first point of the graph is the lower bound of the first group, plotted with 0 on the vertical axis.
- For each group, plot a point representing the cumulative frequency over the upper bound of that group.
- Join these points with straight lines.

Here is the cumulative frequency polygon.


Cumulative Frequency Polygon
500
400
300
200
100
0
10
20
30
40
50
60
70

1.7.4 Comparing the Mean with the Median – Revisited

Having graphed the frequency distribution for the Cinema Audience in three different ways, we can see how the data looks and see how this relates to what a comparison of the mean and median will tell us. The values of these parameters were:

- Mean: 34.31
- Median: 33.6

Therefore the mean is slightly higher than the median.

We interpret this to mean that there are a small number of high values dragging up the mean. The distribution is skewed.

Looking at the histogram or the frequency polygon, you can see that the distribution would be more symmetric were it not for the last age group, 60 for 70.

Consider the distribution without this last age group:

Age classes	*Frequencies*	*Cumulative Frequencies*
10 – 20	51	51
20 – 30	120	171
30 – 40	150	321
40 – 50	75	396
50 – 60	43	439

If we redo all our work, we now find the following:

- Mean: 33.61
- Median: 33.27

This would indicate that there is a smaller bias, the distribution is almost symmetric, but now leaning leftward, towards the smaller numbers.