Descriptive Statistics

Damian Cox

February 6, 2019

Contents

1 Introduction

We will first set out the aims and utility of the concepts we will be studying here, then we will explore some of the fundamental ideas of descriptive statistics and accompanying definitions.

1.1 Summarising Data

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. In the first instance, this will be done by drawing up a list of parameters calculated from a given set of data; the mean, standard deviation, median and percentiles.

1.1.1 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:

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

This is the sum 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.

Take careful note that the symbol for the mean is the variable x with a bar over the symbol. This is referred to as 'x bar'.

The mean gives us some information about our list of values; it can be regarded as an indicator of the middle or centre of the set of numbers. It is formally called a measure of central tendency. This one value is limited, however. 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 mean or there could be a very widely distributed set of numbers. A second number is needed to quantify this.

1.1.2 The Standard Deviation

The first measure of dispersion we will study is the variance, defined in the following equation:

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

This number takes the difference between each value of x and the mean, squares this difference and then divides their sum by $n-1$. Squaring the differences ensures they build up rather than cancel and so, crucially, we know that this is always a positive number.

While this number does measure how widely dispersed the data x_i is around the mean, it has one disadvantage in that it is not in the same units as the original data values. Whatever units x_i is in, the variance is in the square of that unit. We therefore introduce the square root of the variance, known as the *standard deviation*, usually denoted s. It is defined by the equation

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

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

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}. \tag{3}$$

This version of the equation requires fewer steps in the calculation. The expression for the top line means that it is found by calculating the sum of squares first, then subtract the product $n\bar{x}^2$. We therefore know that the sum of the squares is always higher than the term $n\bar{x}^2$, since the result of this subtraction must be positive.

The first equation for s should be taken as being used for the definition of the idea of the standard deviation; the second is usually used for calculation of s, if computing power is limited.

1.1.3 An Example of a Calculation for s

Here is a list of the percentage marks of students in an assessment:

85, 65, 40, 55, 64, 75, 80, 66, 57, 86, 47, 94, 81, 72, 83, 51, 63, 77, 36, 68.

We will calculate the mean 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:

$$\sum x_i = 1,345.$$

Then the mean is

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

In the equation for the standard deviation, we will firstly calculate the sum of the squares:

$$\sum x_i^2 = 95,295.$$

Then the standard deviation equation gives:

$$s^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1} = \frac{95,295 - 20 \times 67.25^2}{19}.$$

The details of the top-line calculation are: $67.25^2 = 4522.5625$, so multiplying by 20 gives 90,451.25. Then the top line is 95,295 - 90,451.25 = 4,843.75, so that

$$s^2 = \frac{4,843.75}{19} = 254.93.$$

Taking the square root means that $s = 15.97$.

This number is quite high, indicating that the percentage marks were quite widely distributed around the mean mark of 67.25. Looking through the list, there were several marks in each 'decade' from the 30's, 40's and on up to the 80's. The standard deviation reflects this variety in the numbers.

1.1.4 Comparing Two Standard Deviations

Here is an example where two lists of numbers have the same mean, but their standard deviations distinguish between them. These calculations go the same way as the previous example.

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, for $i = 1$ to 14, to refer to the first list, and use b_i, for $i = 1$ to 14, for the second. A quick calculation (check this yourself) gives:

$$\sum a_i = 517, \ \sum b_i = 518.$$

The two means are then

$$\bar{a} = 36.93, \ \bar{b} = 37.$$

Thus the means of the two lists are very close. Now let us calculate the two standard deviations.

The sum of squares for each is

$$\sum a_i^2 = 27,055 \Rightarrow s_A^2 = \frac{27,055 - 14 \times 36.92^2}{13} = \frac{7962.73}{13} = 612.53.$$

Therefore the standard deviation for list A is $s_A = 24.75$. Now turn to B:

$$\sum b_i^2 = 27,055 \Rightarrow s_B^2 = \frac{119,758 - 14 \times 37^2}{13} = \frac{592}{13} = 45.54.$$

Therefore the standard deviation for list B is $s_B = 6.75$.

The standard deviation for list A is considerably smaller than for B, reflecting the narrower spread of numbers.

1.2 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 for 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. Their ages are known, but grouped by how many are in each age decade, the twenties, thirties and so on. This information is presented in the following table: This

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

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

1.2.1 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? We need an exact definition of the meaning of the group '20 – 30', as an example. 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:

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

This means that the number 30 goes into the 30 – 40 age group, the number 40 then goes into the 40 – 50 age group and so on.

1.2.2 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 cannot be calculated in the same way as before; they can only be estimated. The way to proceed is

as follows.

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

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

Here the summation always includes all of the values of f_i so no limits 1 to 6 are needed in the 'sigma' notation. 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. These numbers are then summed to give an estimate of the total of all the ages. We are now working from this table: The sum of each midpoint times the corresponding frequency is:

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

$$51 \times 15 + 120 \times 25 + 150 \times 35 + 75 \times 45 + 43 \times 55 + 10 \times 65 = 15,405.$$

This is the most useful estimate of the sum of the ages we can get from the information we have.

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

$$\frac{15,405}{449} = 34.31.$$

In order to produce an estimate of the standard deviation, we will need an approximation of the sum of the squares. This will be done in a similar way to the sum itself; each mid-point is now squared and then multiplied by the corresponding frequency:

$$51 \times 15^2 + 120 \times 25^2 + 150 \times 35^2 + 75 \times 45^2 + 43 \times 55^2 + 10 \times 65^2 = 594,425.$$

We now use this estimate for the sum-of-squares in the equation for the standard deviation:

$$s^2 = \frac{594,425 - 449x34.31^2}{448} = 65873/448 = 147.0$$

So $s = 12.12$.

1.2.3 Summary - Parameters for Grouped Data

The equations we have been using for the mean and standard deviation can now be rewritten for the case of grouped data, based on the ideas we have just outlined in the cinema audience case. 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:

$$n = \sum f_i.$$

The *frequency mean* is then

$$\bar{x} = \frac{\sum m_i f_i}{n}. \tag{4}$$

The *frequency standard deviation* is given by the two equations

$$s^2 = \frac{\sum_{i=1}^{n} (m_i - \bar{x})^2 f_i}{n-1} = \frac{\sum_{i=1}^{n} m_i^2 f_i - n\bar{x}^2}{n-1}. \tag{5}$$

This equation is the same as the previous one for s, except that the approximation for the sum of the squares found from the midpoints and frequencies is used instead of the actual values.

It can often be very useful to use these equations for the grouped version of a large set, as they are quicker to calculate. A very large data set can be grouped in the same way as the cinema audience we saw above. This work is a counting process, and cuts down on the large numbers of calculations needed to work with the raw data itself. Once this has been done, it is then be quicker to use the estimates for groups rather than the original equations.

1.2.4 An Example of Grouped Data Calculations

The following are the lifetimes of machines produced by two companies labelled A and B. Find the mean and standard deviation for each data set.

Lifetimes	Set A	Set B
100 to 105	2	58
105 to 110	8	25
110 to 115	19	20
115 to 120	59	7
120 to 125	26	6
125 to 130	7	4
130 to 135	3	2

For the first data set, the total of the frequencies is $n = 124$. The sum of the midpoints times the frequencies is:

$$102.5 \times 2 + 107.5 \times 8 + ... + 132.5 \times 3 = 14,610.$$

Dividing this by 124 gives the estimate for the mean of 117.82.

The estimate of the standard deviation goes as follows. The estimate of the sum of the squares is:

$$102.5^2 \times 2 + 107.5^2 \times 8 + \ldots + 132.5^2 \times 3 = 43,125.$$

Putting this in the equation for s:

$$s^2 = \frac{\sum m_i^2 f_i - n\bar{x}^2}{n-1} = \frac{\sum m_i^2 f_i - 124 \times 117.82^2}{123}$$

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 sum of the midpoints times the frequencies is

$$102.5 \times 58 + 107.5 \times 25 + \ldots + 132.5 \times 2 = 13,115.$$

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

The estimate of the standard deviation goes as follows. The 'sum of the squares' is:

$$102.5^2 \times 58 + 107.5^2 \times 25 + \ldots + 132.5^2 \times 2 =$$

Putting this in the equation for s:

$$s^2 = \frac{\sum m_i^2 f_i - n\bar{x}^2}{n-1} = \frac{\sum m_i^2 f_i - 124 \times 117.82^2}{123}$$

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

1.3 The Median and Percentiles

There are other parameters aside from the mean that help us describe the midpoint and distribution of a large dataset. They are called called the median and the quartiles, more generally, the percentiles. They will be useful in their own right and also when they are compared to the mean. The essential difference between these numbers and the mean and standard deviation is that the median and percentiles are concerned with the order of the numbers in a given dataset.

1.3.1 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. Consider the following list of 7 numbers, arranged in increasing order:

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

In this list, 3 numbers are above 6 and 3 are 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. We may therefore say that the midway point of the data lies between these two values. The value quoted for the median in this case will be the midpoint of 9 and 12:

$$\frac{9+12}{2} = 10.5.$$

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. This means that when the numbers in a data set are arranged in order, one quarter of the numbers are below the first quartile and three quarters are above it.

We will find the median and quartiles for the following list of 7 numbers:

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

The first step is to arrange them in order:

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

The median is the 4th number: 7. To get the first quartile, the lower half of the list of numbers are:

$$3, 5, 6, 7.$$

The question arises now whether the 7 should be included in the lower list or the higher list. This problem will arise whenever there are an odd number of values in the list. Logically it should be in both or neither - the convention adopted is that it will be included in both. The quartile is then between 5 and 6, so it is

$$\frac{5+6}{2} = 5.5.$$

For the third quartile, the upper half are:

$$7, 8, 9, 12.$$

The quartile is between 8 and 9, so it is

$$\frac{8+9}{2} = 8.5.$$

1.3.2 Example of Medians and Quartiles

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

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

Arranging these numbers in order first:

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

The calculations are:

1. The median will be between 1.7 and 1.8, i.e. 1.75.

2. The first quartile will be between 1.6 and 1.7, and so is 1.65.

3. Similarly the third is between 1.8 and 1.9, and so is 1.85.

The following 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 - 1.65 - 1.7, 1.7 - 1.75 - 1.8, 1.8 - 1.85 - 1.9, 2.0

1.3.3 Comparing the Mean with the Median

The value of the mean compared to the median can tell us something about the distribution of a dataset. This is in effect a comparison of the values of the numbers with their order. 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. The median, by contrast, 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 listed here:

1. If the mean is close to the median, then there are as many figures above the mean as there are below. The list is not skewed.

2. If the mean is below the median then there are a small number of lower figures which are 'dragging' the mean down. This means the list has a 'tail' of smaller values to the left but is skewed to the right, the higher values.

3. If the mean is above the median then there are a small number of higher figures which are 'dragging' the mean up. This means the list has a 'tail' of higher values to the right but is skewed to the left, the lower values.

These ideas will become very important for grouped data.

1.4 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. We will produce estimates for a particular frequency distribution and then apply the same logic to the general case to produce a set of equations giving estimates for the median and the other percentiles.

For the case of grouped data, we will introduce notation for the median and refer to it from now on by the symbol M.

1.4.1 The Cinema Audience

Recall the frequency distribution of ages of a group attending a film: If the numbers in the list, going from 10 years to 70 years, were equally spaced out, a good estimate of the median would be

$$10\text{years} + \frac{60\text{years}}{2} = 40\text{years}.$$

A better estimate could be found if we carry out a similar procedure for the group the median is in. Therefore we see what its place is within one of the groups and from this, estimate how far it is above that groups' lower bound.

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

To estimate the median, there are 449 numbers, so the median would normally be the 225th number. We must now determine which group this number is in. We can see it falls in the 30 to 40 years age group, by checking the 'total so far' of each group and seeing if we have gone past the 225th number.

To do this formally, we will need the number in a given group plus all those in previous groups. These are called the cumulative frequencies. This means that a cumulative frequency for a given group is the total number of data values that are less than the upper bound of that group. The cumulative frequencies for the cinema audience data are shown in the following table: From the table above of cumulative frequencies, we see that age group 30 to

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

40 years has the 172nd to the 321st 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 225th number, there are 171 in the previous two groups, so the place of the median is the 225 – 171 = 54th number within the group.

We have found that the median is the 54th 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

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

above the lower bound of the group, so that the estimate for the median is

$$30 + \frac{54}{150} \times 10\,\text{years} = 30 + \frac{54}{15} = 33.6\,\text{years}.$$

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

1.4.2 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

$$\frac{n+1}{2},$$

to stay consistent with our definition of the median for even or odd lists of numbers. Knowing this, first use the cumulative frequencies to decide which group the median is in.

The symbols in this equation have the following meanings:

- c 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.

Then the Median is given by the equation

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

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 and repeat the argument made above for the cinema audience. Decide, using the cumulative frequencies, which class the median falls in. Then the symbols as listed above, the f values in this group run from L to $L+w$. They are the values in places $c+1$ to $c+f$ in our list of values. The median is the value in place $\frac{n+1}{2} - c$ in the list, out of the f in the group of width w. Then assume the numbers are evenly spread out, so that there is a step of $\frac{w}{f}$ between each number. So multiply this fraction times the place in the list to determine how far into the group the median is:

$$\left(\frac{n+1}{2} - c\right)\frac{w}{f}$$

This number is then added to the lower bound to get the estimate for the median:

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

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. Using the same symbols as before with the corresponding meanings, the first and the third quartile are then given by the equations:

$$Q_1 = L + \left(\frac{n+1}{4} - c\right)\frac{w}{f},\ Q_3 = L + \left(\frac{3(n+1)}{4} - c\right)\frac{w}{f}.$$

At this point, note that the median is the second quartiles; this means that we could denote it by Q_2 to show the link to the quartiles.

1.4.3 Example - Component Lifetimes

The following table shows the lifetimes of components produced by a company, sorted into groups. The cumulative frequencies have also been calculated. We

Age classes	Frequencies	Cumulative Frequencies
100 – 105	9	9
105 – 110	15	24
110 – 115	20	44
115 – 120	30	74
120 – 125	20	94
125 – 130	6	100

will calculate the median and quartiles for this data set.

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

- The width of the classes is $w = 5$.
- The lower bound of this class is $L = 115$.
- The frequency of the class containing the median value is $f = 30$.
- The cumulative frequency of the previous class is $c = 44$.

The median is then

$$M = L + \left(\frac{n+1}{2} - c\right)\frac{w}{f} = 115 + (50.5 - 44)\,\frac{5}{30} = 115 + \frac{6.5}{30} = 116.1.$$

For the first quartile, $\frac{n+1}{4} = 25.25$, so Q_1 would normally be just after the 25th number. This means it is in the group 110 to 115. Now identify the values needed for our equation for the first quartile:

- The width of the classes is $w = 5$.
- The lower bound of this class is $L = 110$.
- The frequency of the class containing the median value is $f = 20$.
- The cumulative frequency of the previous class is $c = 24$.

The first quartile is then

$$Q_1 = L + \left(\frac{n+1}{4} - c\right)\frac{w}{f} = 115 + (25.25 - 24)\,\frac{5}{20} = 110 + \frac{1.25}{4} = 110.31.$$

For the third quartile, $\frac{3}{4}(n+1) = 75.75$, so Q_3 would normally be after the 75th number. This means it is in the group 120 to 125. Now identify the values needed for our equation for the third quartile:

- The width of the classes is $w = 5$.
- The lower bound of this class is $L = 120$.
- The frequency of the class containing the median value is $f = 20$.
- The cumulative frequency of the previous class is $c = 74$.

The third quartile is then

$$Q_3 = L + \left(\frac{3(n+1)}{4} - c\right)\frac{w}{f} = 120 + (75.75 - 74)\,\frac{5}{20} = 110 + \frac{1.75}{4} = 120.44.$$

2 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 provide examples of three types of graphs we intend to use, with the frequency distribution of

the cinema audience. Our intention here however 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.

The cinema audience distribution is shown again in the table here:

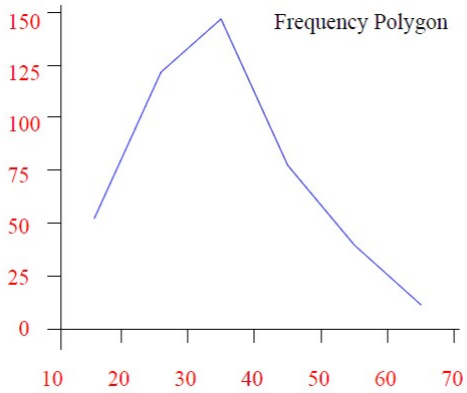
Age classes	Frequencies	Midpoints	Cumulative f.
10 – 20	51	15	51
20 – 30	120	25	171
30 – 40	150	35	321
40 – 50	75	45	396
50 – 60	43	55	439
60 – 70	10	65	449

2.1 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. For example, 51 is the frequency for the age group 10 to 20. 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 candidate for such a 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 figure 2.1. The conventions for



the frequency polygon are summarised here:

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.

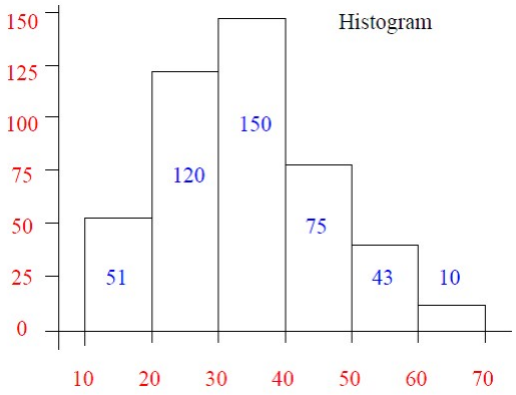
2.2 The Histogram

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 boxes should cover the group, with no space between them. The conventions for the histogram 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, draw a box to the height of the frequency over the group, the width of the box going from the lower to the upper bound of the group.
5. The boxes for the group should have no spaces between them.

The result is a Histogram, shown in figure 2.3.



2.3 Cumulative Frequency Polygon

Having drawn a graph of the frequencies, the next option is to illustrate the cumulative frequencies. We will again choose the values for the horizontal and vertical axes based on the logic behind the idea of the cumulative frequencies. Consider the following ideas:

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

2. If 171 is the cumulative frequency for the group 20 to 30 then there are 171 people below the age of 30.

This can be summarised by saying that the cumulative frequency for each age class represents the number of ages below the upper bound of that class, which 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

1. Draw a horizontal axis with notches or ticks showing the boundaries of the groups.

2. Draw a vertical axis intersecting the horizontal axis at the first lower bound.

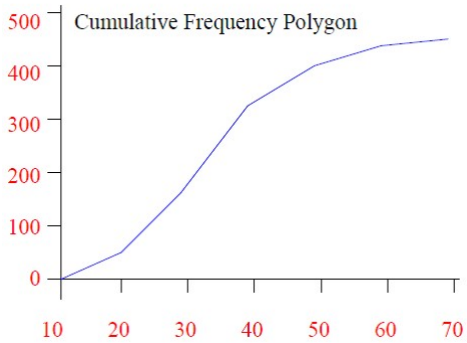
3. Scale the vertical axis appropriately for the values in the cumulative frequency distribution.

4. The first point of the graph is the lower bound of the first group, plotted with 0 on the vertical axis.

5. For each group, plot a point representing the cumulative frequency over the upper bound of that group.

6. Join these points with straight lines.

Here is the cumulative frequency polygon. This is sometimes known as an 'Ogive'.



2.3.1 Comparing the Mean with the Median - Revisited

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

1. Mean: 34.31

2. 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, the distribution would be more symmetric were it not for the last age group,

60 for 70 and the small number in that group. Consider the distribution with this last age group put into the 50 to 60 group: If we redo all our work, we

Age classes	Frequencies	Midpoints	Cumulative f.
10 – 20	51	15	51
20 – 30	120	25	171
30 – 40	150	35	321
40 – 50	75	45	396
50 – 60	53	55	449

now find the following:

1. Mean: 33.61
2. Median: 33.27

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

3 Correlation and Regression

The subject of correlation and regression are extremely important to statistics and the sciences in general. As the word 'correlation' itself suggests, the core of this subject is the idea of investigating how two quantities are linked. For example, if the numbers Y increase as X increases, this may be coincidence or it may indeed be because there is an underlying connection between the two variables. A related question arises – can we predict values of y from x, based on the data we have? We will introduce the idea here as a simple geometric or calculation-of-a-parameter idea.

3.1 Correlation

To investigate how two quantities might be related, the starting point would clearly be to obtain a list of values of the two variables x and y, listed as:

$$X_1 \dots X_n \text{ and } Y_1 \dots Y_n.$$

We will now define a number that goes some way to measuring how closely linked these values are. This is called the correlation coefficient and is usually denoted r.

3.1.1 Definition - The Correlation Coefficient

For the two lists of numbers given, the value of r is given by the equation

$$r = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2}\sqrt{\sum_i (Y_i - \bar{Y})^2}}$$

This equation may be viewed as showing r as a ratio of standard deviations, except the top line 'crosses' the variation between variables x and y. The value of r is a function of the data points X, Y, and so is a parameter for a dataset in the same way as the mean or standard deviation. It can be proved from this definition that it must always be between +1 and –1, no matter what the values of the quantities X, Y.

Just like the equation for the standard deviation itself, some algebra on this equation gives a version where the calculations can be done in fewer steps:

$$r = \frac{\sum_i X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\sum_i X_i^2 - n\bar{X}^2}\sqrt{\sum_i Y_i^2 - n\bar{Y}^2}}$$

3.1.2 A Correlation Example

The following are two lists of numbers; they are the corresponding values of two quantities x and y written in order.

x: 1.3, 1.9, 2.4, 3.1, 3.8, 4.5, 5.1

y: 1.9, 7.1, 7.3, 7.2, 11.9, 15.1, 12.8.

We will show that there is a positive correlation between the variables x and y. We will carryout the calculation using the second, quicker, version of the equation for r:

$$r = \frac{\sum_i X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\sum_i X_i^2 - n\bar{X}^2}\sqrt{\sum_i Y_i^2 - n\bar{Y}^2}}$$

The sums in this calculation involve sums, means and sums of squares, for both variables, exactly the same as those done for the mean and standard deviation. The one extra calculation to be done is the so-called 'cross product' on the top line. It is

$$\sum_i X_i Y_i = 1.3 \times 1.9 + 1.9 \times 7.1 + 2.4 \times 7.3 + \ldots + 5.1 \times 12.8 = 234.25.$$

Then:

$$r = \frac{234.25 - 7 \times 3.157 \times 9.043}{\sqrt{81.37 - 7 \times 3.157^2}\sqrt{692.61 - 7 \times 9.043^2}} = \frac{34.403}{\sqrt{11.597}\sqrt{120.197}} = 0.92.$$

At this point in our study of these concepts, we will take it that a value close to +1 indicates a strong, positive correlation between the two variables. A value close to –1 would indicate a strong negative correlation and a value close to 0 means no correlation. The exact meaning of 'close to' will be refined or defined when we study probability or random variables.

3.2 An Alternative Equation for r

An alternative version of the equation for correlation coefficient r is given by re-writing it in terms of means and standard deviations, quantities we are already familiar with. For the same list of values of two variables x and y, listed as:

$$X_1 \ldots X_n \text{ and } Y_1 \ldots Y_n.$$

The equation for r can be written, using the standard deviation s_X for the X values and s_Y for the Y values, as:

$$r = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)s_X s_Y} = \frac{\sum_i X_i Y_i - n\bar{X}\bar{Y}}{(n-1)s_X s_Y}.$$

It can make the calculations much quicker if some of these figures are already available.

3.2.1 A Correlation Between Age and Smoking

In an attempt to establish a link between age and smoking habits, a group of ten smokers were asked for their age and the number of cigarettes they smoked. Let A_i be the age recorded and X_i the number of cigarettes for each smoker $i = 1$ to 10. The following information is also available, with the usual symbols used for the means and standard deviations of quantities A and X:

$$\bar{A} = 41.8, \bar{X} = 8.9, s_A = 9.37, s_X = 3.96 \text{ and } \sum_i A_i X_i = 3,932.$$

Calculate the correlation coefficient r for this data to determine if the two variables are linked.

The first step is to calculate r:

$$r = \frac{\sum_i X_i Y_i - n\bar{X}\bar{Y}}{(n-1)s_X s_Y} = \frac{3,932 - 10 \times 41.8 \times 8.9}{9 \times 9.37 \times 3.96}$$

This works out to be:

$$r = \frac{211.8}{333.9468} = 0.634.$$

We will see later that a figure like this will be treated as a test statistic, where a more robust interpretation of the correlation value can be done. In the meantime, we can just observe that the value of r indicates that the number of cigarettes increases with age; a negative correlation.

3.3 Regression

We will now introduce a very useful technique, linked to correlation, which will allow us to construct a simple equation to model data where one variable appears to depend on another, in a situation where no underlying equations or knowledge of physics tells us how to proceed. This is what is known as an 'empirical approach.'

3.3.1 Two Correlated Variables

Consider the following set of values of two variables x and y, where each value of the variable x has a corresponding value of the variable y:

x: 1.7, 1.8, 1.9, 2.0, 2.1, 2.2, 2.3, 2.4, 2.5, 2.6.

y: 2.78, 2.13, 2.39, 2.12, 1.82, 1.5, 1.61, 1.24, 0.69, 0.82.

We have already seen how we can use correlation to establish how strongly linked the two variables are. The equation for r is:

$$r = \frac{\sum_i X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\sum_i X_i^2 - n\bar{X}^2}\sqrt{\sum_i Y_i^2 - n\bar{Y}^2}}.$$

The various sums required for this equation are calculated first:

$$\bar{X} = 2.15,\ \bar{Y} = 1.71,\ \sum_i X_i^2 = 47.05,\ \sum_i Y_i^2 = 33.31,\ \sum_i X_i Y_i = 35.0.$$

The calculation for r is then:

$$r = \frac{35.0 - 10 \times 2.15 \times 1.71}{\sqrt{47.05 - 10 \times 2.15^2}\sqrt{33.31 - 10 \times 1.71^2}} =$$

$$= \frac{-1.766}{\sqrt{0.825}\sqrt{4.0714}} = -0.964.$$

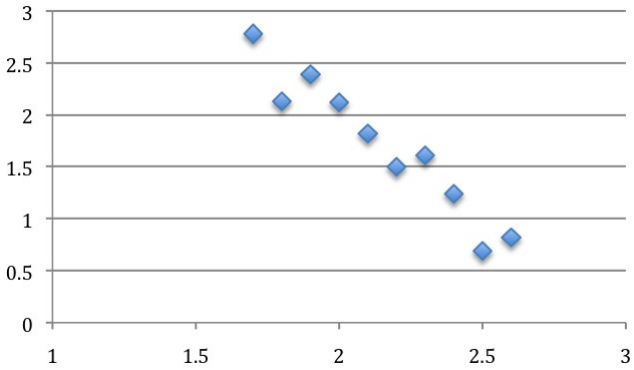
This value of $r = -0.964$ indicates strong correlation.

3.3.2 Scatterplot

The reason for this strong correlation can be seen when the two corresponding values of variables x and y are treated as points in the X-Y plane, in other words, the data is written as the list of points

$$(X_1, Y_1), (X_2, Y_2), \ldots (X_10, Y_10).$$

This is known as a scatter plot, shown here: It can be seen from the scatterplot



that the points are almost in a straight line, heading downward, reflecting the correlation coefficient being very close to -1. If it has been shown that the variables x and y are linked in some way, to the extent that a correlation test can do this, the next logical step would be to come up with a simple equation to predict the value of the y variable from the x.

3.3.3 Linear Regression

The statistical technique known as regression gives a way of linking the variable y with x, so that values of y can be predicted from x, within the range of the original data. To predict values of y from values of variable x not already listed in the original data, an equation is needed giving y as a function of x.

The simplest equation to use in this way would be a linear relation

$$y = \alpha + \beta x.$$

The values of α and β are chosen to minimise the error in predicting the values we already have of y from their corresponding values x. This idea is called regression and because we are trying to fit a line, it is referred to as linear regression.

Fitting a line to this data is done in the following way. Capital letters will be used to denote actual values of the variables, with a subscript for their place in the list so that corresponding values of x and y can be paired. So the data is two lists of n numbers:

$$X_1 \dots X_n \text{ and } Y_1 \dots Y_n.$$

So for each actual value of variable x, call it X_i, the actual value of y we have is Y_i and the predicted value from the linear equation will be $\alpha + \beta X_i$. The error is then

$$e_i = Y_i - (\alpha + \beta X_i).$$

The best way to see how the errors are building up as we go through the n values given is to square each error and then add them. Therefore we try to minimise the sum of errors

$$\sum_i e_i^2.$$

Because we are trying to find values of α and β to minimise the squared error, this method is also called least squares regression.

The following equations give values for the two constants with this constraint. Because The equation for b is

$$b = \frac{\sum_i X_i Y_i - n\bar{X}\bar{Y}}{\sum_i X_i^2 - n\bar{X}^2}$$

and the equation for a is

$$a = \bar{Y} - b\bar{X}.$$

An important geometric interpretation of the second equation is that the 'mean point' $(\bar{X}, \bar{Y})$ is on the line we are producing, in other words

$$\bar{Y} = a + b\bar{X}.$$

3.3.4 Applying The Regression Equations

All the quantities required for the calculations for a and b have already been calculated when we found our value of r. In fact, all we need is the top line from the r calculation and the value under the first square root below the line: these values can now be substituted directly into the equation for coefficient b:

$$b = \frac{-1.766}{0.825}$$

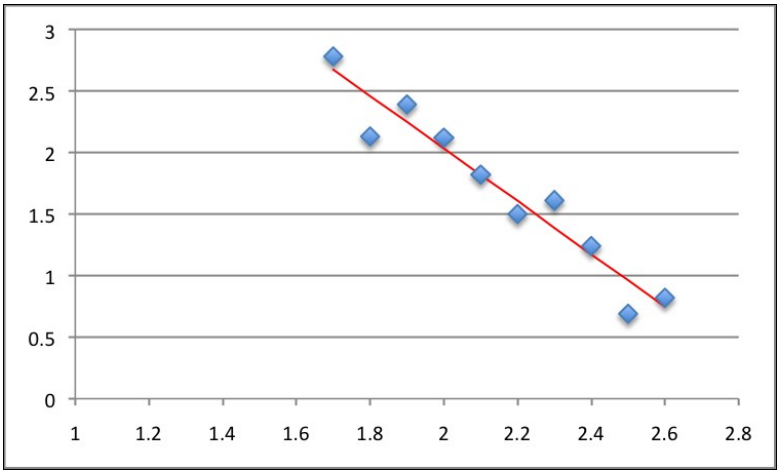
The result for a is then

$$a = 1.71 - (-2.141) \times 2.15 = 6.312.$$

The equation linking the variables y and x is then

$$y = 6.312 - 2.141x.$$

If we draw this line in the same X-Y plane as the data, in the scatterplot above, we can see how close it comes to meeting all the points from the original data. In the scatterplot, the points are in blue, the line is in red. For this data set, the line is clearly very close to the original data points; it provides an excellent predictor of the value of y from the value of x. This is expected given that the value of r was very close to -1. Naturally such a good fit will not always be the case.


3
2.5
2
1.5
1
0.5
0
1
1.2
1.4
1.6
1.8
2
2.2
2.4
2.6
2.8