Chapter 3

Numerical summaries for data

3.1 Introduction

So far we have only considered graphical methods for presenting data. These are always useful starting points. As we shall see, however, for many purposes we might also require *numerical* methods for summarising data: perhaps one or two numbers can summarize the key information about location and variability in the data. Before we introduce some ways of summarising data numerically, let us first think about some notation.

3.2 Mathematical notation

Before we can talk more about numerical techniques we first need to define some basic notation. This will allow us to generalise all situations with a simple shorthand.

Very often in statistics we replace actual numbers with letters in order to be able to write general formulae. We generally use a single letter to represent sample data and use subscripts to distinguish individual observations in the sample. Amongst the most common letters to use is x, although y and z are frequently used as well. For example, suppose we ask a random sample of three people how many mobile phone calls they made yesterday. We might get the following data: 1, 5, 7. If we take another sample we will most likely get different data, say 2, 0, 3. Using algebra we can represent the general case as x_1, x_2, x_3:

1st sample	1	5	7
2nd sample	2	0	3
typical sample	x_1	x_2	x_3

This can be generalised further by referring to the data *as a whole* as x and the ith observation in the sample as x_i. Hence, in the first sample above, the second observation is $x_2 = 5$ whilst in the second sample it is $x_2 = 0$. The letters i and j are most commonly used as the index numbers for the subscripts.

The total number of observations in a sample is usually referred to by the letter n. Hence in our simple example above $n = 3$.

The next important piece of notation to introduce is the symbol $\sum$. This is the upper case of the Greek letter "sigma". It is used to represent the phrase "sum the values". This symbol is used as follows:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n.$$

This notation is used to represent the sum of all the values in our data (from the first $i = 1$ to the last $i = n$), and is often abbreviated to $\sum x$ when we sum over all the data in our sample.

Two other mathematical basics need to be introduced. First, the use of powers is important in many statistical formulae. We all know that, for example, the square of three means raising 3 to the power 2, i.e. $3^2 = 3 \times 3 = 9$. This can be generalised to x^k, which means multiplying x by itself k times.

The other important idea is the use of brackets. Brackets are used to impose an ordering on the way operations are carried out. The operation inside the bracket is carried out before the one outside. Consider the following three cases:

$$\begin{aligned} 3 + 4^2 &= 19 \\ 3^2 + 4^2 &= 25 \\ (3 + 4)^2 &= 49. \end{aligned}$$

In the first case, we simply square 4 and then add this to 3. In the second case, we square both numbers and then add them together, while in the third case, because of the brackets, we add the numbers together and then square the result. Each one of these seemingly similar formulae gives a very different result. If we consider the last two formulae in general terms we could represent the second as $\sum x^2$, that is, we raise all the xs to the power 2 and then add them together. The third equation can be represented as $(\sum x)^2$, that is, all the xs are summed together and then this sum raised to the power 2. This is an important distinction which we will use later.

3.3 Measures of Location

These are also referred to as measures of *centrality* or, more commonly, **averages**. In general terms, they tell us the value of a "typical" observation. There are three measures which are commonly used: the *mean*, the *median* and the *mode*. We will consider these in turn.

3.3.1 The Arithmetic Mean

The arithmetic mean is perhaps the most commonly used measure of location. We often refer to it as the average or just the mean. The arithmetic mean is calculated by simply adding all our data together and dividing by the number of data we have. So if our data were 10, 12, and 14, then our mean would be

$$\frac{10 + 12 + 14}{3} = \frac{36}{3} = 12.$$

We denote the mean of our sample, or sample mean, using the notation $\bar{x}$ ("x bar"). In general, the mean is calculated using the formula

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

or equivalently as

$$\bar{x} = \frac{\sum x}{n}.$$

Example
Suppose we ask 7 Stage 2 Business Management students how many units of alcohol they drank last week and get: 16, 52, 0, 6, 10, 0, 21. The sample mean alcohol consumption of these $n = 7$ students is

For small data sets this is easy to calculate by hand, though this is simplified by using the statistics mode on a calculator.

Sometimes we might not have the raw data; instead, the data might be available in the form of a table. It is still possible to calculate the mean from such data. Let us first consider the case where we have some ungrouped discrete data. Previously we have seen the data:

Date	Cars Sold	Date	Cars Sold
1st July	9	8th July	10
2nd July	8	9th July	5
3rd July	6	10th July	8
4th July	7	11th July	4
5th July	7	12th July	6
6th July	10	13th July	8
7th July	11	14th July	9

The mean number of cars sold per day is

$$\bar{x} = \frac{9 + 8 + \cdots + 8 + 9}{14} = \frac{108}{14} = 7.71.$$

These data can be presented as the frequency table

Cars Sold $(x_{(j)})$	**Frequency** (f_j)	$x_{(j)} \times f_j$
4	1	4
5	1	5
6	2	12
7	2	14
8	3	24
9	2	18
10	2	20
11	1	11
Total	$n = 14$	108

The sample mean can be calculated from these data as

$$\bar{x} = \frac{(4 \times 1) + (5 \times 1) + (6 \times 2) + \cdots + (11 \times 1)}{14} = \frac{4 + 5 + 12 + \cdots + 11}{14} = \frac{108}{14} = 7.71.$$

We can express this calculation of the sample mean from discrete tabulated data as

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{k} x_{(j)} \times f_j.$$

Here the different values of X which occur in the data are $x_{(1)}, x_{(2)}, \ldots, x_{(k)}$. In the example $x_{(1)} = 4,\ x_{(2)} = 5, \cdots, x_{(k)} = 11$ and $k = 8$.
If we only have *grouped* frequency data, it is still possible to *approximate* the value of the sample mean. Consider the following (ordered) data:

8.4	8.7	9.0	9.0	9.2	9.3	9.3	9.5	9.6	9.6
9.6	9.7	9.7	9.9	10.3	10.4	10.5	10.7	10.8	11.4

The sample mean of these data is 9.73. Grouping these data into a frequency table gives

Class Interval	**mid–point** (m_j)	**Frequency** (f_j)	$f_j \times m_j$
$8.0 \leq x < 8.5$	8.25	1	8.25
$8.5 \leq x < 9.0$	8.75	1	8.75
$9.0 \leq x < 9.5$	9.25	5	46.25
$9.5 \leq x < 10.0$	9.75	7	68.25
$10.0 \leq x < 10.5$	10.25	2	20.50
$10.5 \leq x < 11.0$	10.75	3	32.25
$11.0 \leq x < 11.5$	11.25	1	11.25
Total		$n = 20$	195.50

When the raw data are not available, we don't know where each observation lies in each interval. The best we can do is to assume that all the values in each interval lie at the central value of the interval, that is, at its mid–point. Therefore, the (approximate) sample mean is calculated using the frequencies (f_j) and the mid–points (m_j) as

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{k} f_j \times m_j.$$

For the grouped data above, we obtain

$$\bar{x} = \frac{1}{20}\{(1 \times 8.25) + (1 \times 8.75) + \cdots + (3 \times 10.75) + (1 \times 11.25)\} = \frac{195.5}{20} = 9.775.$$

This value is fairly close to the correct sample mean and is a reasonable approximation given the partial information we have in the table.

For large samples with narrow intervals, this approximate value will be very close to the correct sample mean (calculated using the raw data).

3.3.2 The Median

The median is occasionally used instead of the mean, particularly when the data have an asymmetric profile (as indicated by a histogram – think back to last week) or there are outlying or unusual observations. The median is the middle value of the observations when they are listed in ascending order. It is straightforward to determine the median for small data sets, particularly via a stem and leaf plot.

The median is that value that has half the observations above it and half below. For example, ordering the student alcohol data gives $\{0, 0, 6, 10, 16, 21, 52\}$. Clearly the middle value is 10, so the median is 10 units per week.

Suppose we also asked four Stage 2 Marketing and Management students how many units of alcohol they drank last week, and got $\{21, 0, 12, 14\}$. Ordering the data gives $\{0, 12, 14, 21\}$ and there are now two middle values in the sample, 12 and 14. If there are two middle values we take the average of these two numbers as the median, so in this case the median is $(12 + 14)/2 = 13$ units per week.

In general, the median is calculated as the

$$\left(\frac{n+1}{2}\right)^{th} \text{ smallest observation in the sample.}$$

For example, with the original alcohol data there were $n = 7$ observations and so the median was the

$$\frac{n+1}{2} = \frac{7+1}{2} = \frac{8}{2} = 4^{th} \text{ smallest observation,}$$

which is what we observed previously; for these data the median is 10 units per week.

For the second alcohol dataset we had $n = 4$ and so the median was the

$$\frac{n+1}{2} = \frac{4+1}{2} = \frac{5}{2} = 2.5^{th} \text{ smallest observation,}$$

which just means that it is half-way between the 2nd and 3rd smallest observations. Again, this is what we found; the median is 13 units per week.

It is possible to estimate the median value from an ogive as it is half way through the ordered data and hence is at the 50% level of the cumulative frequency. The accuracy of this estimate will depend on the accuracy of the drawn ogive.

3.3.3 The Mode

This is the final measure of location we will look at. It is the value of the random variable which occurs with the highest frequency. It is usually found by inspection. For discrete data this is easy. The mode is simply the most common value. So, on a bar chart, it would be the category with the highest bar. For example, consider the following data: 2, 2, 2, 3, 3, 4, 5. Quite obviously the mode is 2 as it occurs most often. We often talk about modes in terms of categorical data. Recalling the mode of transport example from Chapter 1 (page 7), the mode was "Car", as it was the most popular mode of transport to university.

It is possible to refer to modal classes with grouped frequency data. This is simply the class with the greatest frequency of observations. For example, the model class of

Class	Frequency
$10 \leq x < 20$	10
$20 \leq x < 30$	15
$30 \leq x < 40$	30

is obviously $30 \leq x < 40$. It is not possible to put a single value on the mode with such continuous data. However, the modal class might tell you much about the data. Modal classes are also obvious from histograms, being the highest peaked bar. Of course, if we change the class boundaries, the position of the modal class may change.

So when should you use one measure of location and not the others?

Consider the student alcohol dataset $\{0, 0, 6, 10, 16, 21, 52\}$ which has a mean of 15 units per week and a median of 10 units per week. Note that we could change 52 to 152 and the median is still 10 units per week, but the mean is now 29.3 units per week.

We prefer to use the median if the distribution of the data is **asymmetric** (or **skewed**) or if there are outliers present since the mean can be distorted by extreme values. We say that the median is more *robust* than the mean to such values.

If the distribution is roughly symmetric and there are no outliers, then the mean and median will be similar. There are reasons why, in this situation, we would probably use the mean instead of the median, and these will be covered later in the course.

3.4 Measures of Spread

A measure of location is insufficient in itself to summarise data as it only describes the value of a typical outcome and not how much variation there is in the data. For example, consider the following two samples

Sample 1	6	22	38	mean $= 22$	median $= 22$
Sample 2	21	22	23	mean $= 22$	median $= 22$

Both samples have the same measures of location but they are clearly very different samples! The first set of data ranges considerably from the mean or median value while the second stays very close. Neither the mean nor the median fully represents the data. As well as knowing the location statistics of a data set, we also need to know how variable or 'spread-out' our data are.

There are three basic measures of spread which we will consider: the *range*, the *inter–quartile range* and the *sample variance*.

3.4.1 The Range

This is the simplest measure of spread. It is simply the difference between the largest and smallest observations. In our simple example above the range for the first set of numbers is $38 - 6 = 32$ and for the second set it is $23 - 21 = 2$. These clearly describe very different data sets. The first set has a much wider range than the second.

There are two problems with the range as a measure of spread. When calculating the range you are looking at the two most extreme points in the data, and hence the value of the range can be unduly influenced by one particularly large or small value, known as an *outlier*. The second problem is that the range is only really suitable for comparing (roughly) equally sized samples as it is more likely that large samples contain the extreme values of a population.

3.4.2 The Inter–Quartile Range

The inter–quartile range describes the range of the middle half of the data and so is less prone to the influence of the extreme values.

To calculate the inter–quartile range (IQR) we simply divide the ordered data into four quarters. The three values that split the data into these quarters are called the *quartiles*. The first quartile (*lower quartile*, $Q1$) has 25% of the data below it; the second quartile (*median*, $Q2$) has 50% of the data below it; and the third quartile (*upper quartile*, $Q3$) has 75% of the data below it. We already know how to find the median, and the other quartiles are calculated as follows:

$$Q1 = \frac{(n+1)}{4}\text{th smallest observation}$$
$$Q3 = \frac{3(n+1)}{4}\text{th smallest observation.}$$

Just as with the median, these quartiles might not correspond to actual observations. For example, in a dataset with $n = 20$ values, the lower quartile is the $(20 + 1)/4 = 5\,\frac{1}{4}$th smallest observation, that is, a quarter of the way between the 5th and 6th smallest observations. This calculation is essentially the same process we used when calculating the median. Consider the data:

8.4	8.7	9.0	9.0	9.2	9.3	9.3	9.5	9.6	9.6
9.6	9.7	9.7	9.9	10.3	10.4	10.5	10.7	10.8	11.4

Here the 5th and 6th smallest observations are 9.2 and 9.3 respectively. Therefore, the lower quartile is

$$Q1 = 9.2 + \frac{1}{4}(9.3 - 9.2) = 9.2 + 0.025 = 9.225.$$

Similarly the upper quartile is the $3 \times (20 + 1)/4 = 15\,\frac{3}{4}$ smallest observation, that is, three quarters of the way between the 15th and 16th smallest observations which are 10.3 and 10.4, respectively; so

The **inter–quartile range** is simply the difference between the upper and lower quartiles, that is

$$IQR = Q3 - Q1$$

which for these data is

The inter-quartile range can also be *estimated* from the ogives in a similar manner to the median. Simply draw the ogive and then read off the values for 75% and 25% and calculate the difference between them. This is especially useful if you only have grouped data. Again the accuracy depends on the quality of your graph.

The inter–quartile range is useful as it allows us to make comparisons between the ranges of two data sets, without the problems caused by outliers or uneven sample sizes.

3.4.3 The Sample Variance and Standard Deviation

The **sample variance** is the standard measure of spread used in statistics. It is usually denoted by s^2 and is simply the "average" of the squared deviations of the observations from the sample mean. That is, we use the formula

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 .$$

We can simplify this to

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

This formula is easier for calculations. The divisor is $n-1$ rather than n in order to correct for the bias which occurs because we are measuring deviations from the sample mean rather than the "true" mean of the population we are sampling from.

Note that the notation x_i^2 represents the squared value of the observation x_i. That is, $x_i^2 = (x_i)^2 = x_i \times x_i$.

The **sample standard deviation**, denoted s, is the positive square root of the sample variance. This quantity is often used in preference to the sample variance as it has the same units as the original data and so is perhaps easier to understand.

Consider again the data on the number of units of alcohol consumed by a sample of 7 students last week. The data were: 16, 52, 0, 6, 10, 0, 21. We have already calculated the sample mean as $\bar{x} = 15$. Now

$$\sum x^2 =$$

$$n(\bar{x})^2 =$$

and so the sample variance is

$$s^2 =$$

and the sample standard deviation is

$$s = \sqrt{s^2} =$$

If this appears complicated, don't worry, as most scientific calculators will give the sample standard deviation when in stats mode. Note that on a scientific calculator the correct sample standard deviation is given by the s or $x\sigma_{n-1}$ button on the calculator and **not** the σ or $x\sigma_n$ buttons.

Note also that a different calculation is needed when the data are given in the form of a grouped frequency table with frequencies (f_i) in intervals with mid–points (m_i). First the sample mean $\bar{x}$ is approximated (as described earlier) and then the sample variance is approximated as

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

3.5 Box plots

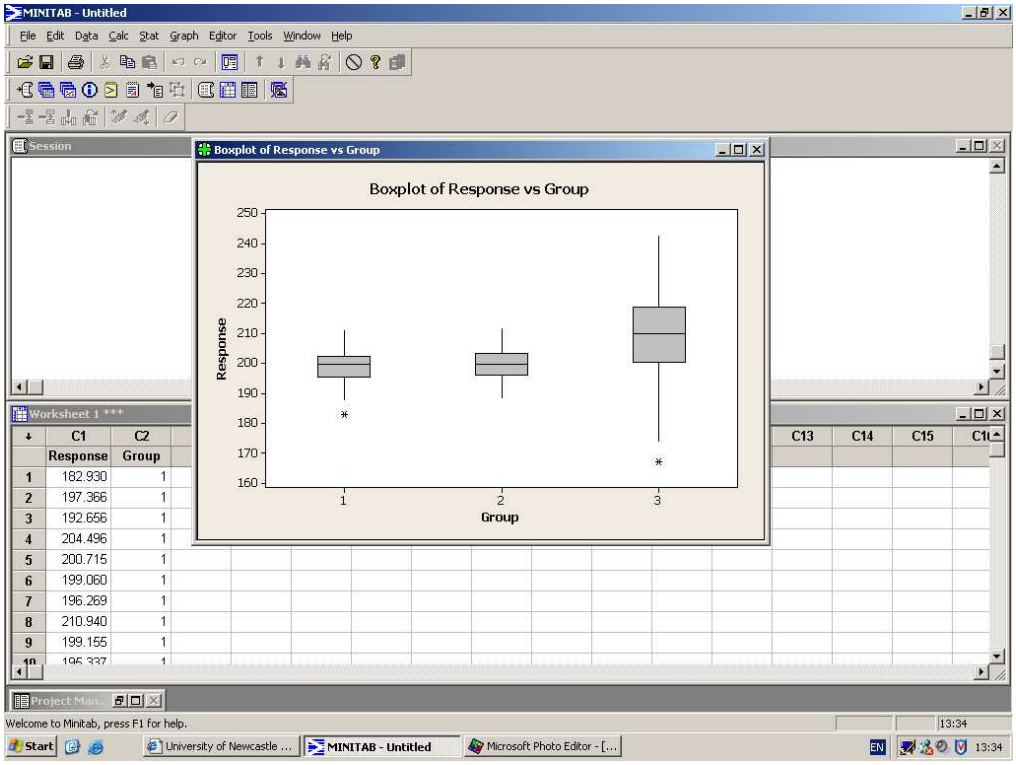
Box plots (or "box and whisker" plots) are another graphical method for displaying data and are *particularly useful for highlighting differences between groups*, for example, different spending patterns between males and females or comparing pricing within designated market segments. These plots use some of the key summary statistics we have looked at earlier, the quartiles and also the maximum and minimum observations.
The plot is constructed as follows. After laying out an x–axis for the full range of the data, a rectangle is drawn with ends at the the upper and lower quartiles. The rectangle is split into two at the median. This is the "box". Finally, lines are drawn from the box to the minimum and maximum values – these are the "whiskers".

Suppose that, from our data, we obtain the following summary statistics:

Minimum	Lower Quartile (Q1)	Median (Q2)	Upper Quartile (Q3)	Maximum
10	40	43	45	50

In the space below, construct the associated box plot.

Displaying group structure is one of the main uses of box plots. Shown below is a plot produced by `Minitab`.



It clearly shows that although there is overlap between the three sets of data, the first and second datasets contain roughly similar responses and that these are quite different from those in the third set. Note that the asterisks (*) at the ends of the whiskers is the way `Minitab` highlights outlying values.

3.6 Exercises

1. Recall the data from Exercise 1 in Chapter 2 on the weight (in kg) of 50 sacks of potatoes leaving a farm shop. The *ordered* data are presented below.

8.1	8.2	8.5	8.7	8.8
8.9	9.2	9.3	9.3	9.4
9.5	9.5	9.6	9.6	9.6
9.7	9.7	9.9	9.9	10.0
10.0	10.0	10.0	10.0	10.1
10.2	10.2	10.2	10.3	10.3
10.4	10.4	10.4	10.5	10.6
10.6	10.6	10.6	10.6	10.7
10.8	10.9	11.0	11.2	11.3
11.3	11.3	11.5	11.6	12.8

(a) Calculate the mean of the data.

(b) Calculate the median of the data.

(c) Calculate the range of the data.

(d) Calculate the inter–quartile range.

(e) Calculate the sample standard deviation.

(f) Draw a box plot for these data and comment on it.

(g) Put the data in a grouped frequency table.

(h) Find the modal class.

2* Chloe collected the following data on the weight, in grams, of "large" chocolate chip cookies produced by Millie's Cookie Company.

27.1 22.4 26.5 23.4 25.6 26.3 51.3 24.9 26.0 25.4

To summarise, Chloe was going to calculate the mean and standard deviation for this sample. However, her friend Mark warned her that the mean and standard deviation might be inappropriate measures of location and spread for these data.

(a) Do you agree with Mark? If so, why?

(b) Mark suggested the *geometric mean* as an alternative to the standard sample mean. Find out what the *geometric mean* is, and calculate this for the data collected from Millie's.

(c) Do you think Mark was right to suggest the geometric mean as an alternative measure of average? Explain.

(d) Calculate measures of location and spread that you feel are more suitable.

*** Prize question** – the "best" solution submitted before 5pm on Friday 17th October 2014 wins a prize! Solutions to me via email (`daniel.henderson@ncl.ac.uk`) or in person.