# Computational Mathematics/Information Technology

Dr Oliver Kerr

2009–10

# Introduction

Many results from statistics can be presented, accidentally or deliberately, in misleading ways.

# Introduction

Many results from statistics can be presented, accidentally or deliberately, in misleading ways.

- Most people have more than the average number of legs.

# Introduction

Many results from statistics can be presented, accidentally or deliberately, in misleading ways.

- Most people have more than the average number of legs.

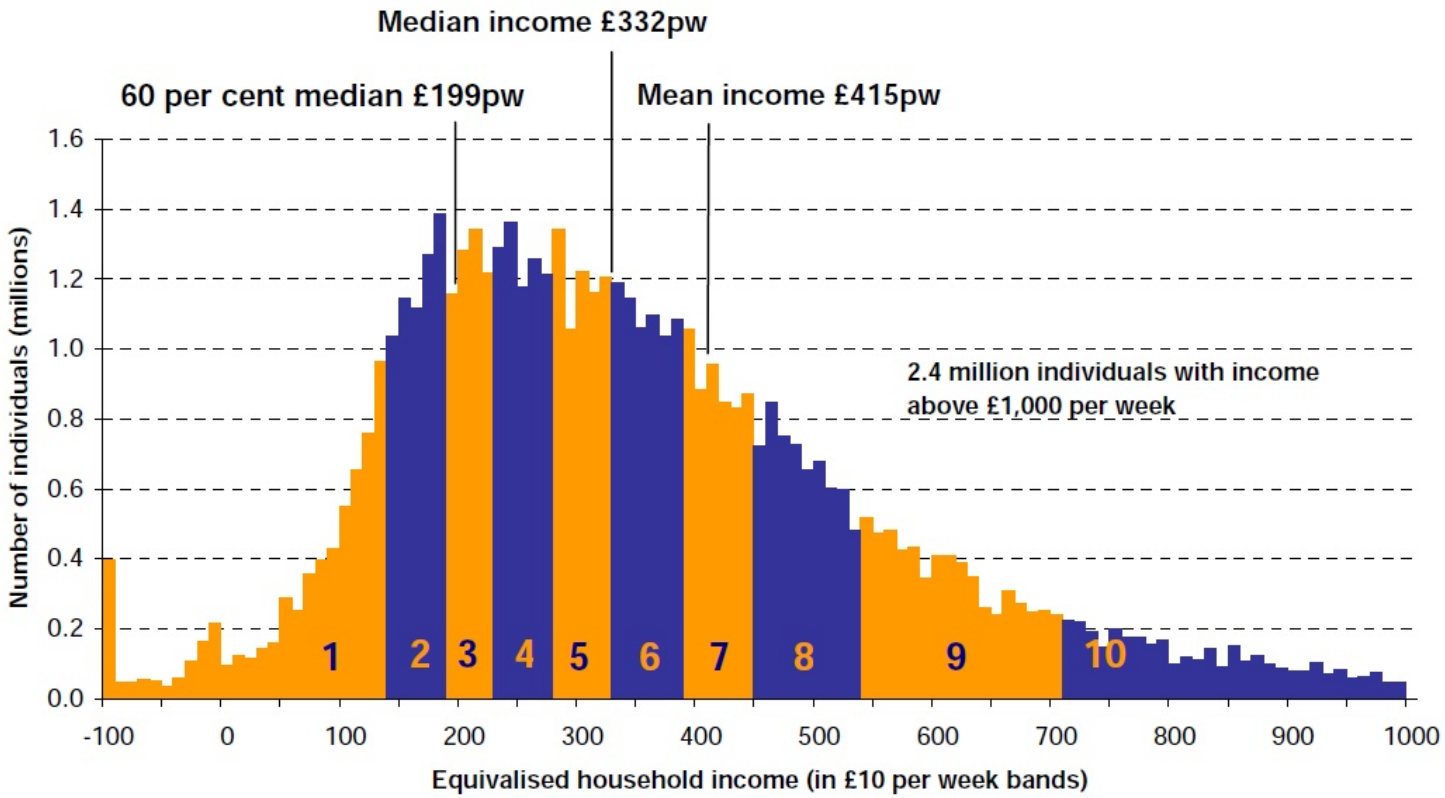
Nearly everyone has 2, some have either one or none, and none have 3 or more So the average is less than 2.

Jake the Peg (Rolf Harris)

- Nearly two thirds of households earn less than average.

- Nearly two thirds of households earn less than average.

**Figure 2.1 (AHC): Income distribution for the total population, 2007/08**




from Households Below Average Income, National Statistics

It could be worse!

If you measure something with a probability density function of

$$p(x) = \frac{2}{\pi(1+x^2)}, \qquad 0 \leq x < \infty$$

then everything is below average!

Just having one or two numbers to describe a sample can be misleading.

You need a collection of statistics that can be used, and ways of displaying your data and analysis that helps you and others understand what is going on.

We will look at Minitab, a purpose built piece of software for displaying and analysing statistical data. But first we will go over some of the figures it will produce. (Many of which you will recognise.)

# Mean and Standard Deviation

Two basic numbers can be used to describe a sample:

- What is a typical number from your data set?
- What is the spread of the data?

We have seen the first of these may not have a "best" answer, nor may the second. For the moment we will look at the **mean** and **standard deviation**:

Given a set of data $\{x_1,\ x_2 \ldots x_N\}$ the following definitions are made:

**Mean**

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

**Standard Deviation**

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

Sometimes instead of the standard deviation

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

we use the **variance**

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

Sometimes instead of the standard deviation

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

we use the **variance**

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

Which can be written as

$$s^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - 2\frac{\overline{x}\sum_{i=1}^{N} x_i}{N} + \frac{\sum_{i=1}^{N} \overline{x}^2}{N}$$

$$= \frac{\sum_{i=1}^{N} x_i^2}{N} - 2\overline{x}^2 + \overline{x}^2 = \overline{x^2} - \overline{x}^2$$

These definitions are directly applicable to given sets of data, however most of statistics is concerned with obtaining information about a large set of data (*the population*) by considering only a few of the items from the population (*the sample*).

For example we could estimate the average height of first year students at City University by finding the average height of a subgroup — the students present today. Similarly we could estimate the standard deviation of the population by considering the standard deviation of our sample.

But you can have problems if you have small samples:

If you have a large population of numbers equally divided between 1 and $-1$, the mean is 0 and the variance is 1.

But you can have problems if you have small samples:

If you have a large population of numbers equally divided between 1 and $-1$, the mean is 0 and the variance is 1.

Take a sample of 2 numbers, one after another. There are 4 equally likely outcomes

- 1, 1. Average 1, variance 0
- 1, $-1$. Average 0, variance 1
- $-1$, 1. Average 0, variance 1
- $-1$, $-1$. Average $-1$, variance 0

The mean of the averages is correct, but the mean of the variances is half of the right answer.

Small samples can cause problems.

Consider the following:

**Population**

$X_1, X_2, \ldots, X_M$

mean $= \mu$ std $= \sigma$

**Sample**

$x_1, x_2, \ldots, x_N$

mean $= \overline{x}$ std $= s$

The mean of the sample $\overline{x}$ is usually used to estimate the mean of the population $\mu$.

When estimating the standard deviation of a population from a sample the following adjusted formula is used:

$$\sigma \approx \hat{s} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}}$$

Replacing $N$ by $N-1$ gets the expected value of the variance right.
**Note**: By default Minitab always calculates $\hat{s}$ for the standard deviation.

If $N$ is large the use of $N-1$ instead of $N$ makes little difference. If your sample is small then you should always be very careful of conclusions!

"If your experiment needs statistics, you ought to have done a better experiment." — Ernest Rutherford, the first person to split the atom.

If $N$ is large the use of $N-1$ instead of $N$ makes little difference. If your sample is small then you should always be very careful of conclusions!

"If your experiment needs statistics, you ought to have done a better experiment." — Ernest Rutherford, the first person to split the atom.

"Anyone who expects a source of power from the transformation of the atom is talking moonshine" — Ernest Rutherford.

Experts aren't always right!

# Coefficient of Variation

Just as we saw with linear regression and standard errors, saying a sample had a standard deviation of, say, 10 tells us little by itself. If we were measuring the money in our pockets it would be plausible, but if we were looking at annual incomes it would be very unlikely!

Similarly if measure our heights in metres we will get a different answer compared to the answer if we measure our heights in centimetres.

To overcome the dependence on the units used and to still keep the idea of the amount of spread the following coefficient is introduced:

The **coefficient of variation** $v$ is defined as:

$$v = \frac{\sigma}{\mu} \approx \frac{\hat{s}}{\overline{x}}$$

The ratio of the amount of spread or deviation about the mean to the size of the mean.

# Quartiles and Median

Instead of the mean another quantity that can be used to describe a data set is the **median**.

If the data is arranged into ascending order then the data item halfway along the list is the median.

This will only work exactly for an odd number of items. For a list with an even number we would take the average of the middle two.

Clearly the median divides the set of data into two halves in the sense that there are an equal number of data items on either side of it. In a similar way the ordered data set may be split into four quarters in order to give an idea of the spread of the values about the median.

Taken in order the points that do this are called, the first, second and third **quartiles** and denoted $Q_1$, $Q_2$ and $Q_3$ respectively.

The second quartile is the median. However the positions of the first and third quartiles need a bit more care (particularly for small sample sizes).

**Note**: The spread of the quartiles plays a similar rôle to the standard deviation.

One of the usual way to calculate the quartiles (and the way Minitab does it) is as follows:

For the ordered data set $\{x_1, x_2, \ldots x_N\}$. We define the **quartiles** as:

$$Q_1 = x_{\frac{N+1}{4}} \quad Q_2 = x_{\frac{N+1}{2}} \quad \text{and} \quad Q_3 = x_{\frac{3(N+1)}{4}}$$

and use linear interpolation to calculate these values when the indices are not whole numbers.

**Example**: Given the data set $\{1, 3, 5, 9, 11, 11, 12, 15\}$ calculate $Q_1$, $Q_2$ and $Q_3$.

In this example $N = 8$ thus:

$$Q_1 = x_{\frac{8+1}{4}} = x_{2.25} \quad Q_2 = x_{\frac{8+1}{2}} = x_{4.5} \quad \text{and} \quad Q_3 = x_{\frac{3(8+1)}{4}} = x_{6.75}$$

Hence

- $Q_1 = x_{2.25}$ is a quarter of the way between $x_2 = 3$ and $x_3 = 5$, i.e., $Q_1 = 3.5$.
- $Q_2 = x_{4.5}$ is half way between $x_4 = 9$ and $x_5 = 11$, i.e., $Q_2 = 10$.
- $Q_3 = x_{6.75}$ is three quarters of the way between $x_6 = 11$ and $x_7 = 12$, i.e., $Q_3 = 11.75$.

**Alternative definition**: As before we define $Q_2$ to be the median of the data set. We now define $Q_1$ and $Q_3$ to be the medians of all the data points less than $Q_2$ and of all the data points greater than $Q_2$ respectively.

In the previous example with the data set $\{1, 3, 5, 9, 11, 11, 12, 15\}$ we have $Q_2 = 10$.

$Q_1$ is the median of $\{1, 3, 5, 9\}$, giving $Q_1 = 4$.

$Q_3$ is the median of $\{11, 11, 12, 15\}$, giving $Q_3 = 11.5$.

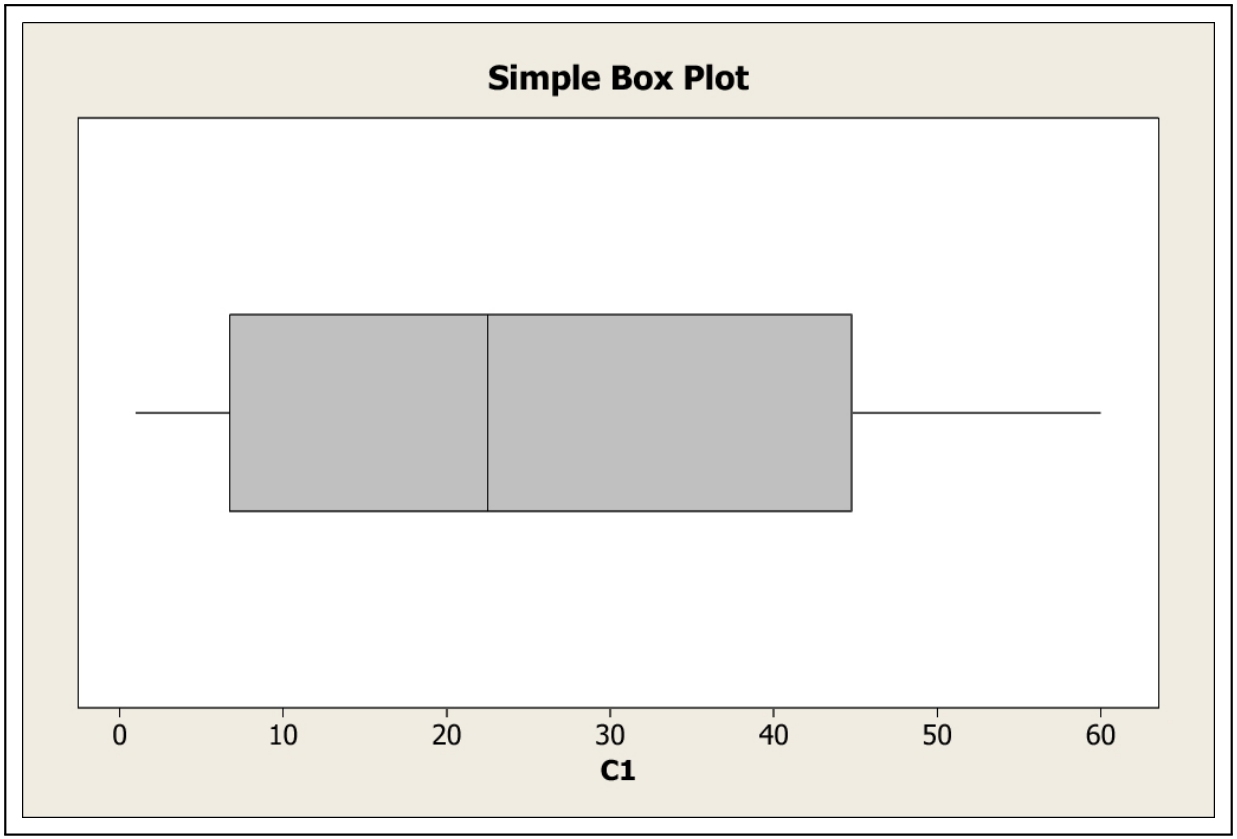
Other definitions exist! For large data sets it usually makes little difference.

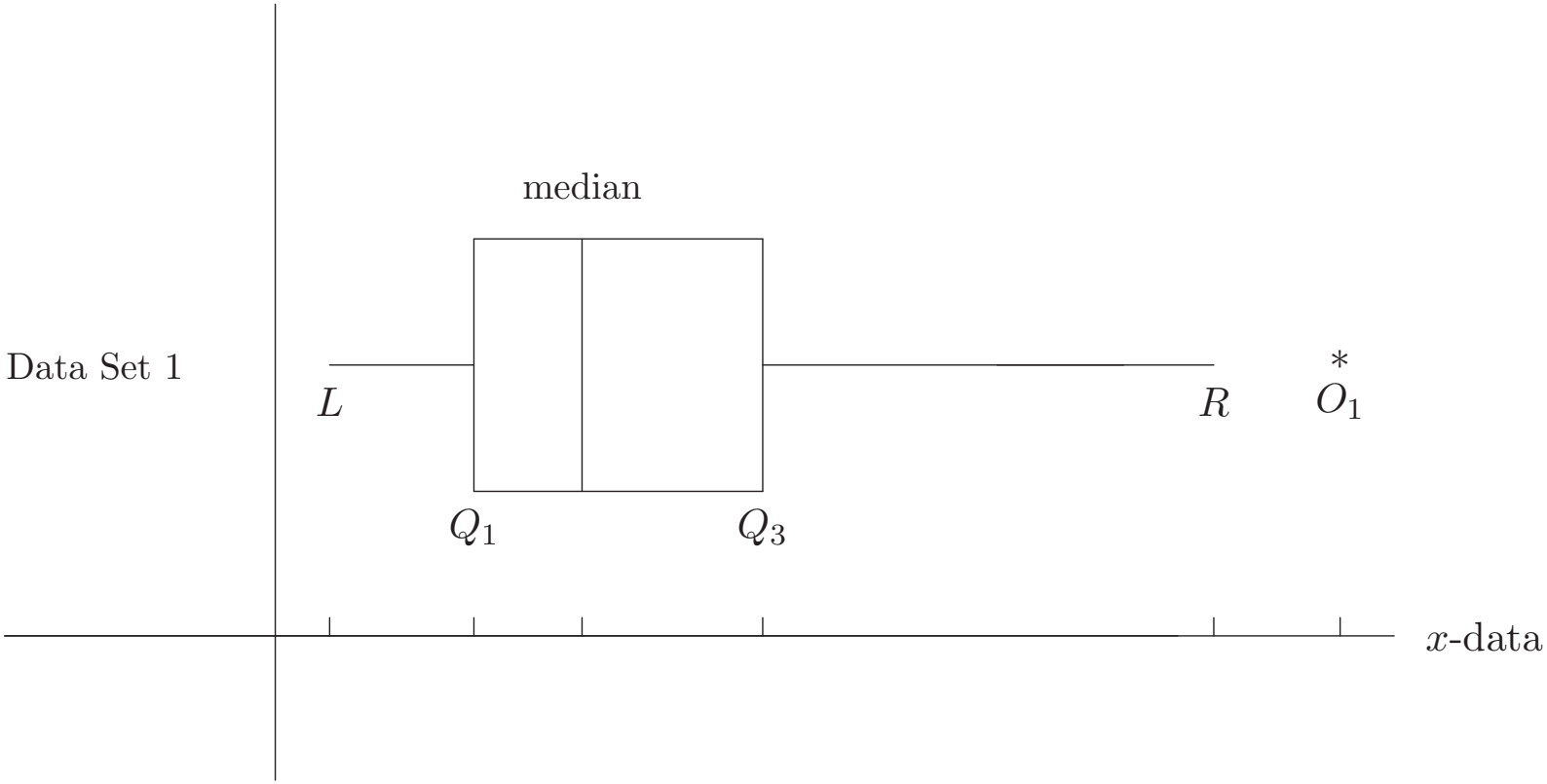
# Graphical Data Representations

We will look at the two basic ways to represent data: the **box plot** and the **histogram**.

# Box Plot

A box plot produced by Minitab looks like






$L =$ "minimum", $R =$ "maximum", $O_1 =$ outlier

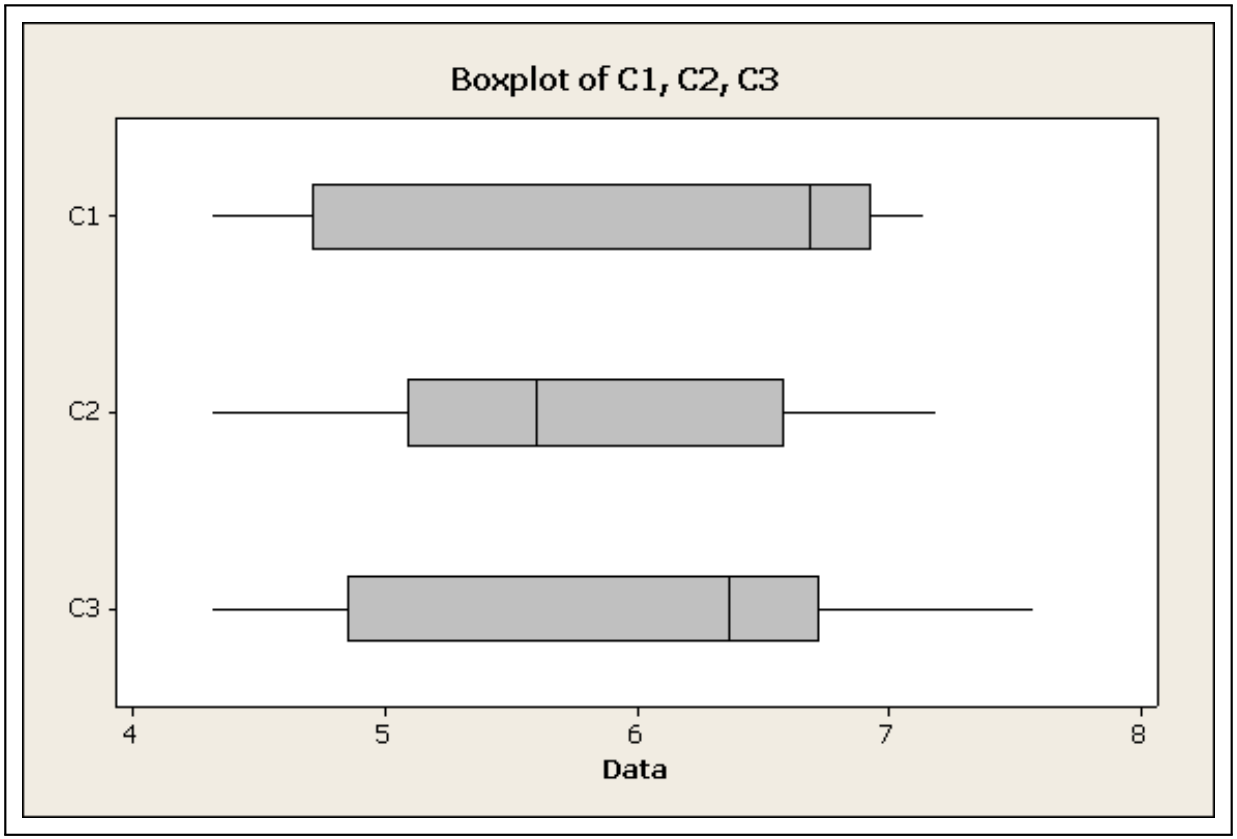
The "minimum" and "maximum" are given by

$$L = \text{Max}\begin{cases} \text{Min data item} \\ Q_1 - 1.5(Q_3 - Q_1) \end{cases} \qquad R = \text{Min}\begin{cases} \text{Max data item} \\ Q_3 + 1.5(Q_3 - Q_1) \end{cases}$$

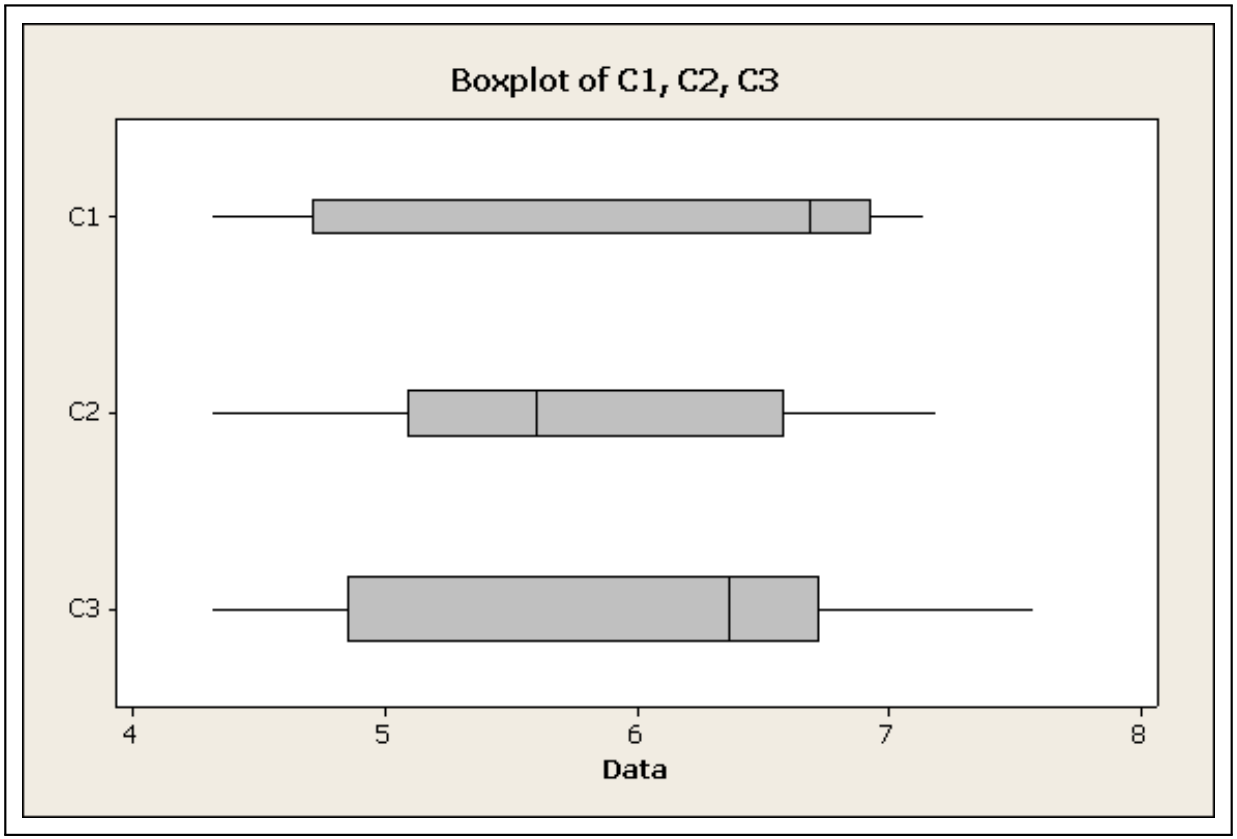
This stops the whisker being determined by an occasional exceptional point.

Care has to be taken with plots, otherwise results can be misleading.

Consider these three samples from the same population:



This one




gives an indication of the relative sizes of the samples (5, 10, 20).

# Histograms

A Histogram is a type of bar chart that indicates the frequency of occurrence of the data items across its range.

We split the interval into partitions (buckets) then we count how of the sample data fall in each bucket and plot a bar chart where the height of the bars corresponds to the the number of data points in that sub-interval.

The resulting graph is called a **histogram**.

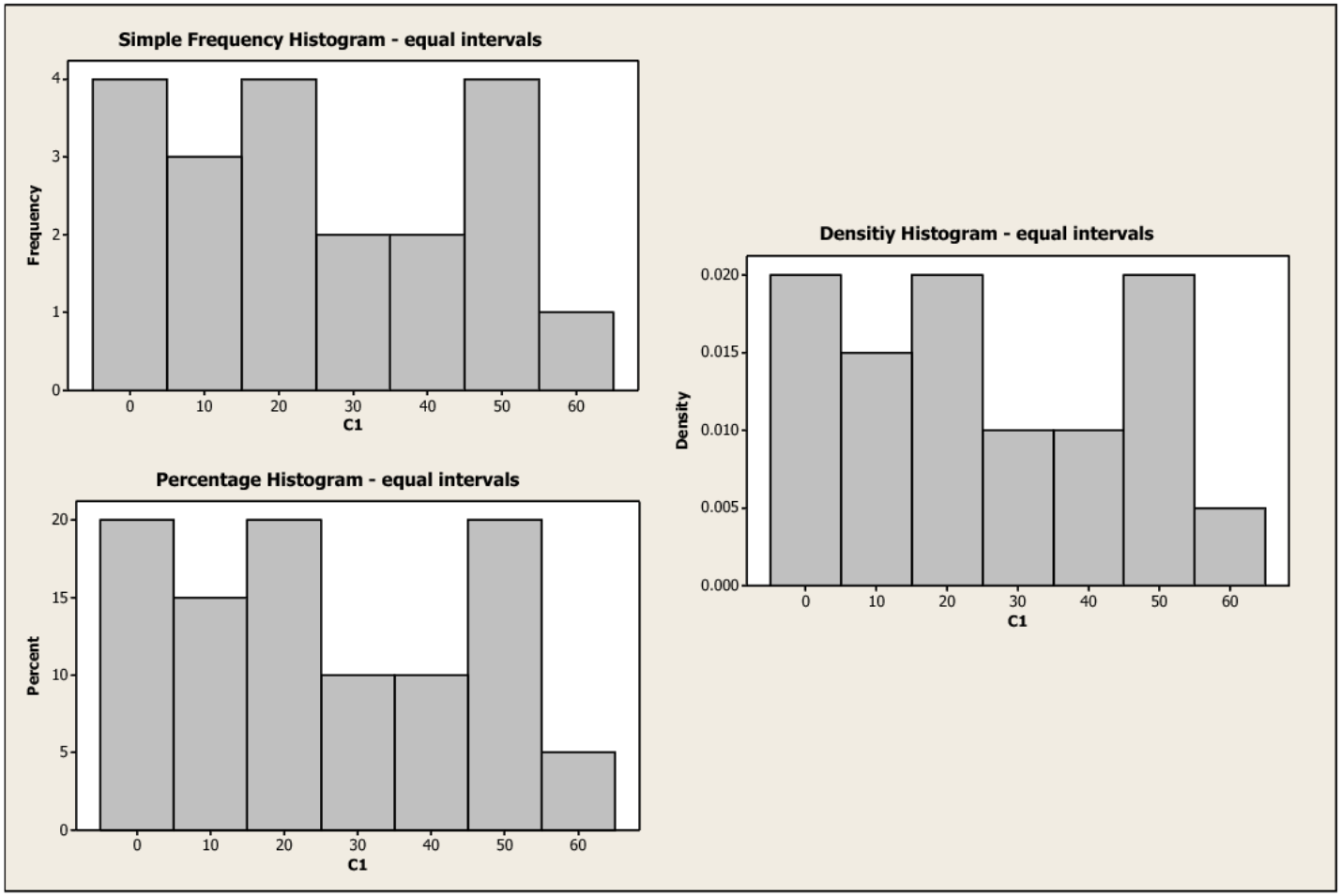
We can plot histograms of the following data set:

| $x_i =$ | 1 | 2 | 2 | 3 | 6 | 9 | 10 | 20 | 21 | 22 |
|---|---|---|---|---|---|---|---|---|---|---|
| | 23 | 27 | 30 | 35 | 44 | 45 | 47 | 50 | 52 | 60 |

Minitab provides us with three types of histogram:

- **frequency histogram**: the height of the rectangle is proportional to the frequency.
- **percentage histogram**: the height of the rectangle is proportional to the frequency expressed as a percentage.
- **density histogram**: the area of the rectangle is proportional to the frequency. (The sum of the areas of all the rectangles equals unity.)

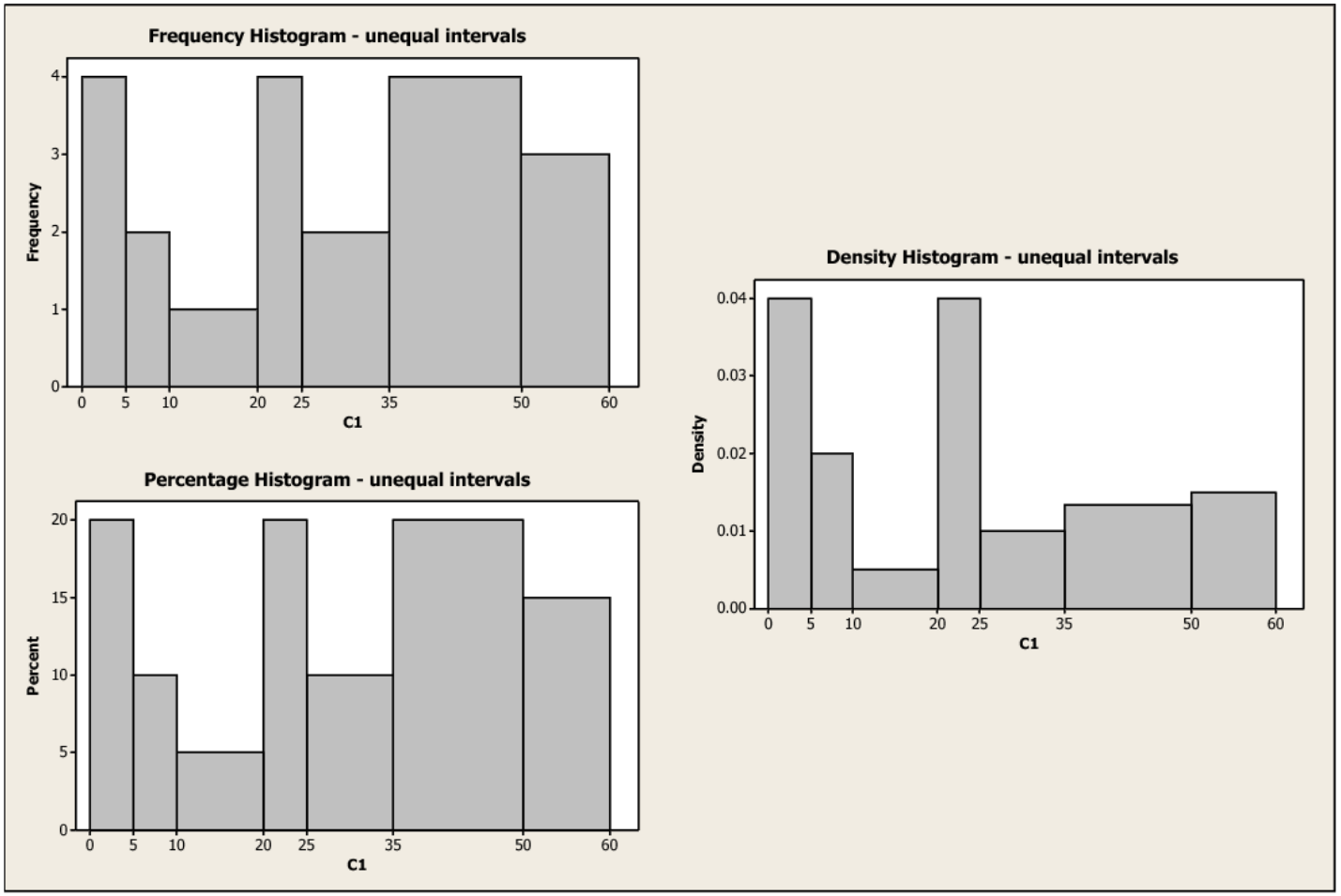
Minitab produces the following versions of these histograms:



With the interval sizes equal the only difference is in the scale on the left.

We can also have intervals of different sizes.

The same sample, different intervals:



When the intervals are not of equal length the frequency and percentage histograms are different from the density histogram it is the area.

In frequency and percentage histograms greater emphasis seems to be given to larger buckets.

In the density histogram the areas the rectangle are adjusted to reflect the frequency. If you have a very big sample and small buckets then the density histogram will approach the probability density function, even if the buckets are of unequal sizes.

Graphical representations of data can be very helpful in giving a feel of what is going on.

Care always has to be taken to ensure that your data sets are presented in ways that are clear and not misleading.