Chapter 1: Statistical Variation and the Normal Curve

1.0 Introduction

Perhaps the most striking feature of real data from almost any discipline is the variation they exhibit. If a nurse measures physical dimensions (heights, weights..) or blood parameters (Serum albumin or cholesterol..) for a sample of patients, they will (nearly) all be different. If a teacher or psychologist administers a psychological test to a sample of children, the test scores will vary greatly. If a manufacturing engineer measures physical dimensions (e.g., outer diameter or length) of components manufactured to the same specifications, they will not all be the same – in fact, they will exhibit precisely the same type of variation as the psychological test scores. Similarly, environmental scientists will find different acidity levels (pH values) in samples from different lakes or streams, or from one site sampled at different times. They will find that birds differ in their dimensions in much the same way that humans do and if they measure birds’ eggs, they will find very much the same kind of variation as did the engineer, teacher or nurse referred to above.

This variation is the reason why statistical methods and ideas are required across a wide range of disciplines. Statistics is the science of describing and analysing chance variation, and statistical methods help us distinguish between systematic effects and chance variation.

In this chapter, we will be concerned with describing such variation – at first with pictures in Section 1.1. We will then focus, in Section 1.2, on the Normal curve that can be adapted to describe the variation encountered in many practical situations, and which is fundamental to the simple statistical methods that form the basis of much of this course. We will discuss, in Section 1.3, how to decide if these curves fit our data. Section 1.4 considers the implications of combining chance quantities (for example, the implications for the overall length of a project of combining several consecutive activities, each subject to chance variation). In practice, we are very often concerned with the behaviour of averages; thus, a medical study might be carried out to decide if the use of a new drug reduces, on average, the blood pressure of patients suffering from high blood pressure. In Section 1.5, we will examine how the behaviour of averages differs from that of ‘raw’ scores – this carries implications for answering questions such as that related to blood pressure changes, referred to above. In Section 1.6 we will consider some aspects of measurement systems – if the same object is measured many times (even by the same person using the same equipment) the same result is not found. Due to chance measurement error, a range of values will be obtained, and these values often exhibit characteristics very similar to those of the industrial or psychological data discussed above. Finally, Section 1.7 illustrates some applications of a simple statistical tool (the control chart) which is related to the material introduced in the earlier sections of the chapter.
1.1 Statistical Variation

We begin with a series of examples to illustrate statistical variation in a number of disciplines. Figure 1.1.1 is based on a dataset (Grades) that comes with the Minitab software [1] that will be used extensively in this course manual. Investigators in a northeastern university in the USA collected data to investigate the relationship between SAT scores (Scholastic Aptitude Test, often used as a college admission or course placement criterion) and GPA grades (grade point average – university performance score). Figure 1.1.1 gives the Verbal SAT scores of 200 students. This picture is typical of what we might expect from psychological/educational test results – the students cluster around the central value (about 600, here) and the numbers of students who get scores far from the centre is quite small – the numbers decrease in a more or less symmetrical manner in each direction.

![Image of SAT Verbal scores for 200 university students](image)

Figure 1.1.1: SAT Verbal scores for 200 university students

Figure 1.1.2 represents capping torques (in lbs-inches) of 78 container caps; the capping torque is a measure of the energy expended in breaking the seal of the screw-capped containers. The data come from a process capability study of a newly installed filling line in a company that manufactures chemical reagents. It is interesting to note that the physical measurements display similar characteristics to the psychological scores – they vary in a roughly symmetrical way around a central value (about 10.8 for the torques).

Figure 1.1.3 represents petal lengths of a sample 50 irises from the variety Setosa. The data were originally published in the Bulletin of the American Iris Society [2], but their frequent citation in the statistical literature is due to their use in the pioneering paper on multivariate analysis by R.A. Fisher [3]. This botanical/environmental science example exhibits the same symmetry characteristics noted above for the psychological and industrial data.
Figure 1.1.2: Capping torques (lb-in) for 78 screw-capped containers

Figure 1.1.3: Petal lengths (mm) for 50 irises (Setosa)

Figure 1.1.4: Heights (cm) of 1794 pregnant women
In Figure 1.1.4 we return to humans, but this time the values represent physical measurements rather than psychological scores: they are heights (cm) of pregnant women. The data were downloaded from the website of Prof. Martin Bland (a well-known medical statistician, see for example his book [4]). Because the sample size is so large (1794 values) the histogram is much smoother and the symmetry is much better defined, than for the smaller sample sizes of the first three examples. In Section 1.3 we will see that all four datasets are consistent with underlying Normal or Gaussian distributions - smooth symmetrical bell-shaped curves (see Figure 1.1.6) that will be discussed in some detail in Section 1.2. Consequently, any departures from symmetry in our histograms are considered due to the variability necessarily associated with relatively small sample sizes. Note though, that both the women’s heights and petal lengths data have many repeated values, due to imprecise measurement systems. This would not be expected in truly Normal data, but is unlikely to matter for practical purposes.

While this symmetrical shape is very common, it is by no means universal, as illustrated by our next example. Figure 1.1.5. represents urinary β-hydroxycorticosterone (β-OHCS) values in nmol/24h “obtained on a random sample of 100 obese adult females drawn from a fully specified target population”; the data come from the book on laboratory medicine by Strike [5]. The steroid data are strongly skewed (to the right); the characteristic described by the graph is very different from those of our first four examples. Most of the steroid measurements are less than 900 but individual values range up to around 1400. While it is now possible to deal directly with skewed data, traditionally, analysts sought transformations which would allow the data to be analysed using methods based on the Normal distribution; this is still a common practice. Figure 1.1.6 shows a histogram (with a superimposed Normal curve) of the transformed data after the original measurements were subjected to transformation by taking logarithms (to base 10). As we shall see again in Section 1.3 the transformation has been successful – the transformed data are well modelled by a Normal curve. Where transformed data follow a Normal curve, the original skewed distribution is said to be ‘log-Normal’.

![Figure 1.1.5: Urinary β-hydroxycorticosterone (nmol/24h) values for 100 obese females](image-url)
There are of course very many different types of shape that may be encountered for histograms of real data (more will be presented during the lectures). However, Figures 1.1.1-1.1.6 were selected for presentation, both to illustrate the wide applicability of the Normal curve (in terms of discipline type) and the possibility of transforming data so that they follow this curve, which is, perhaps, the easiest of all statistical distributions to use – hence its popularity. It is also of fundamental importance because averages of data sampled from skewed distributions will, in many cases, follow a Normal curve (see Section 1.5) – this means that methods developed for Normal data can be applied to such averages, even though the underlying data are skewed.

1.2 The Normal curve

If the more or less symmetrical histograms of Sections 1.1 were based on several thousand observations, the bars of the histograms could be made very narrow and the overall shapes might then be expected to approximate the smooth curve shown below as Figure 1.2.1.
This curve is symmetrical about the centre, the mean, usually denoted $\mu$ - the Greek letter corresponding to $m$, and its width is measured by the standard deviation, denoted $\sigma$ - the Greek letter corresponding to $s$. The standard deviation is formally defined but can usefully be thought of as one sixth of the width of the curve, for most practical purposes. In principle, the curve stretches from minus infinity to plus infinity but in practice (as can be seen from Figure 1.2.1) the area in the tails becomes infinitesimally small very quickly.

**Interpreting the Normal Curve**

Figure 1.2.2 shows three Normal curves with mean $\mu$ and standard deviation $\sigma$. The curves are drawn in such a way that the total area enclosed by each curve is one. Approximately 68% of the area lies within 1 standard deviation of the mean; the corresponding values for two and three standard deviations are 95% and 99.7%, respectively. These areas will be the same for all Normal curves irrespective of the values of the means and standard deviations.

![Figure 1.2.2: Areas within one, two and three standard deviations of the mean](image)

The implications of this figure are that if the variation between measurements on individuals/objects (e.g. blood pressure, SAT scores, torques, etc.) can be described by a Normal curve, then 68% of all values will lie within one standard deviation of the mean, 95% within two, and 99.7% within three standard deviations of the mean. Similarly, if we have a measurement error distribution that follows a Normal curve, then in 32% of cases the chance error will result in a measured value being more than one standard deviation above or below the mean (the ‘true value’ if the measurement system is unbiased). In only 5% of cases will the observation be further than two standard deviations from the mean.
Suppose that the Normal curve shown in Figure 1.2.3 represents the masses of filled containers sampled from a filling line. The curve is interpreted as an idealised histogram for which the total area is 1.0. The area between any two points (say $x_1$ and $x_2$) represents the relative frequency of container masses between $x_1$ and $x_2$. If this area is 0.2 it means that 20% of all containers filled by this line have masses between $x_1$ and $x_2$ units. The assumption is that, provided the filling process remains stable, 20% of all future fillings will result in containers with masses between $x_1$ and $x_2$ units. Since relative frequencies are generally of interest, e.g. what fraction of values is less than or greater than a given value, or what fraction lies between two given values, then, where a Normal curve describes the variability in question, it can be used to calculate the required relative frequency.

The use of probability as a way of expressing uncertainty has become common in everyday speech. Our concept of probability is intimately bound up with the idea of relative frequency and people move easily between the two ways of describing variability. Thus, without any formal definition of probability, it seems quite natural to say that if a filled container is selected at random, the probability is 0.2 that its mass will lie between $x_1$ and $x_2$. Both of these ways of describing variability, viz., relative frequency and probability, are equivalent, and they will be used interchangeably hereafter.
Obtaining Areas under Normal Curves

![Standard Normal Curve and Arbitrary Normal Curve](image)

Figure 1.2.4: A standard Normal curve (z) and an arbitrary Normal curve (x)

All Normal curves have essentially the same shape once allowance is made for different standard deviations. This allows for easy calculation of areas, i.e., relative frequencies or probabilities. If, for any arbitrary curve with mean $\mu$ and standard deviation $\sigma$, the area between two values, say $x_1$ and $x_2$, is required, all that is necessary is to find the corresponding values $z_1$ and $z_2$ on the standard Normal curve, determine the area between $z_1$ and $z_2$ and this will also be the area between $x_1$ and $x_2$ (as shown schematically in Figure 1.2.4). A standard Normal curve has mean zero and standard deviation one.

The relationship between the $x$ and $z$ values is simple:

$$z = \frac{x - \mu}{\sigma}$$

The $z$ value is the answer to the question “by how many standard deviations does $x$ differ from the mean $\mu$?”

Once the $z$ value is calculated, the area to its left (i.e. the area from minus infinity up to this value of $z$) can be read from a standard Normal table (Table ST-1). Thus, a $z$ value of 2 has an area to its left of 0.9772, while a $z$ value of 1.5 has an area to its left of 0.9332. Therefore, the area between $z=1.5$ and $z=2$ on the standard Normal curve is 0.9772 - 0.9332 or 0.0440. Similarly, the area between two points which are 1.5 and 2 standard deviations from the mean of any Normal curve will also be 0.0440.
Example 1

Figure 1.1.4 showed a distribution of heights for a population of women. The mean of the 1794 values was 162.4 and the standard deviation was 6.284. If we take the mean and standard deviation as those of the corresponding population, we can address questions such as those below.

(i) What fraction of the population has a height greater than 170 cm?

(ii) What fraction of the population has a height less than 145 cm?

(iii) What bounds should we use if we want to enclose the central 95% of the population within them?

(i) **What fraction of the population has a height greater than 170 cm?**

To calculate the required proportion of the population, the area to the right of 170 on the height curve must be found. To do this, the equivalent value on the standard Normal curve is calculated; the area to the right of this can be found easily and this gives the required proportion, as shown in Figure 1.2.5.

![Figure 1.2.5: The standard Normal curve (z) and the women's heights curve (x)](image)

The translation from the height (x) to the standard Normal scale (z) is given by:

\[ z = \frac{x - \mu}{\sigma} \]

\[ z = \frac{170 - 162.4}{6.284} = 1.21 \]
Thus, a height of 170 corresponds to the point 1.21 on the standard Normal curve. The area to the left of 1.21 under the standard Normal curve is 0.8869, so the area to the right is 0.1131, since the total area is 1. Similarly, the area to the right of 170 is also 0.1131. Approximately 11% of the women will have a height greater than 170 cm.

We can write out our calculations as follows:

\[
P(x > 170) = P\left(Z > \frac{170 - 162.4}{6.284}\right) = P(Z > 1.21) = 1 - P(Z < 1.21)\]

\[= 1 - 0.8869 = 0.1131\]

We can read \(P(x > 170)\) as “the proportion of \(x\) values greater than 170” or, equivalently, as “the probability that a single \(x\) value, selected at random, will be greater than 170”.

(ii) **What fraction of the population has a height less than 145 cm?**

The value of \(Z\) that corresponds to 145 is given by:

\[
z = \frac{x - \mu}{\sigma}
\]

\[z = \frac{145 - 162.4}{6.284} = -2.77\]

Statistical Table-1 (ST-1) only gives areas for positive values of \(Z\). However, we can use the symmetry of the curve to find areas corresponding to negative values. Thus, \(P(z < 2.77)\) is the area below 2.77; \(1 - P(z < 2.77)\) is the area above +2.77. By symmetry this is also the area below −2.77, as shown in Figure 1.2.6

![Figure 1.2.6: The symmetry property of the Normal curve (area is 0.0028 in each tail)]
Accordingly, we have:

\[ P(z < -2.77) = 1 - P(z < 2.77) = 1 - 0.9972 = 0.0028 = P(X < 145). \]

Approximately 0.3% will have heights less than 145 cm.

(iii) What bounds \((A, B)\) should we use if we want to enclose the central 95% of the population within them?

To find the values that enclose the central 95% of the standard Normal curve (i.e., an area of 0.95) we proceed as follows. Enclosing the central 0.95 means we leave 0.025 in each tail. The upper bound, therefore, has an area of 0.975 below it (as shown in the leftmost curve in Figure 1.2.7). If we inspect the body of Table ST-1, we find that 0.975 corresponds to a \(z\) value of 1.96. By symmetry, the lower bound is \(-1.96\) (the central curve in Figure 1.2.7). Thus, a standard Normal curve has 95% of its area between \(z=-1.96\) and \(z=+1.96\). From the relation:

\[ z = \frac{x - \mu}{\sigma} \]

We get:

\[ 1.96 = \frac{B - \mu}{\sigma} \Rightarrow B = \mu + 1.96\sigma \]

and:

\[ -1.96 = \frac{A - \mu}{\sigma} \Rightarrow A = \mu - 1.96\sigma. \]

It is clear, then, that the values that enclose the central 95% of the population for an arbitrary Normal curve \((x)\) with mean \(\mu\) and standard deviation \(\sigma\) are: \(\mu-1.96\sigma\) and \(\mu+1.96\sigma\).
In the present case this gives us: $162.4 \pm 1.96(6.284)$, i.e., $162.4 \pm 12.3$. While this gives $(150.1, 174.7)$, for practical purposes the bounds would probably be described as 150 to 175 cm.

**Example 2**

Suppose the distribution of capping torques for a certain product container is Normal with mean $\mu=10.6$ lb-in and standard deviation $\sigma=1.4$ lb-in (these were the values obtained from the data on which Figure 1.1.2 is based). To protect against leaks, a minimum value for the torque must be achieved in the manufacturing process. If the lower specification limit on torque is set at 7 lb-in, what percentage of capping torques will be outside the lower specification limit?

To calculate the required percentage, the area to the left of 7 on the capping process curve must be found. To do this, the equivalent value on the standard Normal curve is calculated; the area to the left of this can be found easily and this gives the required proportion (as illustrated schematically in Figure 1.2.8).

![Figure 1.2.8: The standard Normal curve and the capping process curve](image)

The translation from the torque scale ($x$) to the standard Normal scale ($z$) is given by:

$$ z = \frac{x - \mu}{\sigma} $$

$$ z = \frac{7 - 10.6}{1.4} = -2.57 $$

Thus, a torque of 7 on the capping process curve corresponds to the point $-2.57$ on the standard Normal curve. The area to the left of 2.57 under the standard Normal
The normal curve is 0.995 (or 0.9949\(^1\) if you prefer!), so the area to the right is 0.005, since the total area is 1. By symmetry, the area to the left of \(-2.57\) is also 0.005 and, therefore, five caps per thousand will be expected to be below the capping torque lower specification limit in future production, provided the process remains stable. A simple statistical tool (a ‘control chart’) may be used to monitor the stability of manufacturing processes. Section 1.7 provides an introduction to and gives several examples of such charts.

**Exercises**

1.2.1. Weight uniformity regulations usually refer to weights of tablets lying within a certain range of the label claim. Suppose the weights of individual tablets are Normally distributed with mean 100 mg (label claim) and standard deviation 7 mg.

(a) If a tablet is selected at random, what is the probability that its weight is:
   (i) less than 85 mg?
   (ii) more than 115 mg?
   (iii) either less than 85 mg or more than 115 mg?

(b) (i) What percentage of tablets lie between 90 mg and 110 mg?
    (ii) What percentage are heavier than 105 mg?
    (iii) What percentage are lighter than 98 mg?

(c) (i) What weight value \(K\) could be quoted such that 99% of all tablets are heavier than \(K\)?
    (ii) What weight values \(A\) and \(B\) can be quoted such that 99% of all tablets lie between \(A\) and \(B\)?

1.2.2 Figures 1.1.5 and 1.1.6 are based on data cited by Strike [5]. The data are urinary \(\beta\)-hydroxycorticosteroid (\(\beta\)-OHCS) values in nmol/24h, obtained on a random sample of 100 obese adult females drawn from a fully specified target population. According to Strike it is of interest to define a 95% reference range for the \(\beta\)-OHCS measurements (a reference interval encloses the central 95% of the population of measured values) that can be used to assess the clinical status of future subjects drawn from the same population. The distribution of \(\beta\)-OHCS values is quite skewed; accordingly, in order to construct the reference interval we first transform the data to see if they can be modelled by a Normal distribution in a transformed scale. Figure 1.1.6 suggests that this is a reasonable assumption (this will be discussed further in Section 1.3). In the log\(_{10}\) scale the mean is 2.7423 and the standard deviation is 0.1604.

(i) Calculate the bounds \((a, b)\) for the 95% reference interval in the log scale.
(ii) Use your results to obtain the corresponding reference interval in the original measurement scale: the bounds will be \((10^a, 10^b)\).
(iii) What value, \(U\), should be quoted such that only 1% of women from this population would be expected to have a \(\beta\)-OHCS measurement which exceeds \(U\)?

---

\(^1\) Try re-calculating using standard deviations of 1.38 or 1.41 instead of 1.40 and note the effect on the resulting probabilities. How often (never?) do engineers know standard deviations to this level of precision?
Calculating the standard deviation

Our calculations using the Normal curve assumed the mean and standard deviation were already known. In practice, of course, they have to be calculated from observations on the system under study. Suppose that a sample of n objects has been selected and measured (e.g., SAT scores, people’s heights etc.) and the results are $x_1, x_2, \ldots x_n$. The average of these, $\bar{x}$, gives an estimate of the corresponding population value, $\mu$. The standard deviation, $s$, of the set of results gives an estimate of the corresponding population standard deviation, $\sigma$. This is calculated as:

\[
s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}
\]

i.e. the mean is subtracted from each value, the deviation is squared and these squared deviations are summed. To get the average squared deviation, the sum is divided by $n-1$, which is called the ‘degrees of freedom’\(^2\). The square root of the average squared deviation is the standard deviation. Note that it has the same units as the original measurements.

It is important always to distinguish between system parameters, such as the population standard deviation, $\sigma$, and estimates of these parameters, calculated from the sample data. These latter quantities are called ‘sample statistics’ (any number calculated from the data is a ‘statistic’) and are themselves subject to chance variation. System parameters, such as $\mu$ and $\sigma$, are considered fixed, though unknown, quantities. In practice, if the sample size $n$ is very large, the result calculated using the definition given above will often be labelled $\sigma$, since a very large set of measurements will give the ‘true’ standard deviation. If the sample size is small, the result is labelled $s$ (or sometimes $\hat{\sigma}$); this makes clear to the user that the calculated value is itself subject to measurement/sampling error, i.e. if the study were repeated a different value would be found for the sample standard deviation.

The assumption that the parameters of the Normal distributions are known without error is reasonable for the height data of Example 1 (where $n=1794$), but clearly the sample size for the capping torque data ($n=78$) is such that we cannot expect that the ‘true’ process mean or standard deviation is exactly equal to the sample value. Methods have been developed (tolerance intervals) to allow for the uncertainty involved in having only sample data when carrying out calculations like those of

\(^2\) $(n-1)$ is used instead of the sample size $n$ for a mathematical reason; the deviations (before they are squared) sum to zero, so once the sum of $(n-1)$ is found, the last deviation is determined by the constraint, thus, only $(n-1)$ deviations are free to vary at random – hence, ‘degrees of freedom’. Obviously, unless $n$ is quite small, effectively the same answer is obtained if we get the average by dividing by $n$. 

14
Examples 1 and 2, though these are rarely discussed in elementary textbooks. Hahn and Meeker [6] discuss such intervals in an engineering context.

1.3 Assessing Normality

The statistical methods that will be discussed in later chapters assume that the data come from a Normal distribution. This assumption can be investigated using a simple graphical technique, which is based on the type of calculation discussed in the previous section. Here, for simplicity, the logic of the method will be illustrated using a sample of 10 values, but such a sample size would be much too small in practice, as it would contain little information on the underlying shape, particularly on the shapes of the tails of the curve.

The relation between any arbitrary Normal value $x$ and the corresponding standard Normal value $z$ is given by:

$$z = \frac{x - \mu}{\sigma}$$

which can be re-written as:

$$x = \mu + \sigma z$$

If $x$ is plotted on the vertical axis and $z$ on the horizontal axis, this is the equation of a straight line with intercept $\mu$ and slope $\sigma$, as shown in Figure 1.3.1. Accordingly, if we make a correspondence between the $x$ values in which we are interested and $z$ values from the standard Normal table and obtain a straight line, at least approximately, when they are plotted against each other, we have evidence to support the Normality assumption.

![Figure 1.3.1: The theoretical relationship between $x$ and $z$ when $x$ is Normal](image)
The correspondence is established by finding the fraction of the x values equal to or less than each value in the dataset, and then finding the z value that has the same fraction of the standard Normal curve below it. In Table 1.3.1, the ten x values are listed in column 1, their ranks, i, (smallest 1 to largest n=10) in column 2, the fraction equal to or less than x in column 3 (i/n). Column 4 gives the z value with the area given by column 3 below it; for example, in the first row, 90% of the sample values are equal to or less than x=12.6 and column 4 indicates that 90% of the area under a standard Normal curve is below 1.282 (Table ST-1 gives z values to two decimal places, but software such as Minitab or Excel will give more, if required).

Using this method for establishing the correspondence between x and z will always result in the largest x value corresponding to z=\( \infty \) (which presents a challenge when plotting!). To avoid the infinity, some small perturbation of the calculation of the fraction i/n is required, such that the result is never i/n=1. Simple examples of possible perturbations would be (i–0.5)/n or i/(n+0.5). Minitab has several options; (i–3/8)/(n+1/4), which is based on theoretical considerations, is used in Table 1.3.1. The areas which result from this are shown in column 5 and the corresponding z values (Normal scores) in column 6. The choice of perturbation is not important: all that is required is to draw a graph to determine if the values form an approximately straight line. We do not expect a perfectly straight line even for truly Normal data, due to the effect of sampling variability. Thus, 12.6 is the ninth largest value in our dataset, but any value between 13.1 and 12.2 would also have been the ninth largest value. Consequently, 12.6 is not uniquely related to 0.841.

Figure 1.3.2 shows the Normal plot (or Normal probability plot, as it is also called) for our dataset – a ‘least squares’ line is superimposed to guide the eye (we will discuss such lines later). The data are close to the line (hardly surprising for an illustrative example!).

<table>
<thead>
<tr>
<th>x</th>
<th>Ranks</th>
<th>Fraction ≤ x</th>
<th>( \frac{i-3/8}{n+1/4} )</th>
<th>N-scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.6</td>
<td>9</td>
<td>0.9</td>
<td>1.282</td>
<td>0.841</td>
</tr>
<tr>
<td>13.1</td>
<td>10</td>
<td>1.0</td>
<td>( \infty )</td>
<td>0.939</td>
</tr>
<tr>
<td>9.8</td>
<td>2</td>
<td>0.2</td>
<td>-0.842</td>
<td>0.159</td>
</tr>
<tr>
<td>11.7</td>
<td>7</td>
<td>0.7</td>
<td>0.524</td>
<td>0.646</td>
</tr>
<tr>
<td>10.6</td>
<td>4</td>
<td>0.4</td>
<td>-0.253</td>
<td>0.354</td>
</tr>
<tr>
<td>11.2</td>
<td>6</td>
<td>0.6</td>
<td>0.253</td>
<td>0.549</td>
</tr>
<tr>
<td>12.2</td>
<td>8</td>
<td>0.8</td>
<td>0.842</td>
<td>0.744</td>
</tr>
<tr>
<td>11.1</td>
<td>5</td>
<td>0.5</td>
<td>0.000</td>
<td>0.451</td>
</tr>
<tr>
<td>8.7</td>
<td>1</td>
<td>0.1</td>
<td>-1.282</td>
<td>0.061</td>
</tr>
<tr>
<td>10.5</td>
<td>3</td>
<td>0.3</td>
<td>-0.524</td>
<td>0.256</td>
</tr>
</tbody>
</table>

Table 1.3.1 Sample data for calculating Normal Scores
Departures from Normality

The schematic diagrams in Figure 1.3.3 indicate some patterns which may be expected when the underlying distribution does not conform to the Normal curve. Plots similar to Figure 1.3.3(a) will be encountered commonly, as outliers are often seen in real datasets. Figure 1.3.3(c) might, for example, arise where the data represent blood measurements on patients - many blood parameters have skewed distributions. Figures 1.3.3(b) and (d) are less common but (b) may be seen where the data are generated by two similar but different sources, e.g. if historical product yields were being studied and raw materials from two different suppliers had been used in production. The term ‘heavy tails’ refers to distributions which have more area in the tails than does the Normal distribution (the t-distribution is an example). Heavy tails mean that data values are observed further from the mean, in each direction, more frequently than would be expected from a Normal distribution.

Figure 1.3.3: Idealised patterns of departure from a straight line
Real Data

In Section 1.1 we examined histograms for a number of datasets; here, we draw Normal plots for several of those examples. Figure 1.3.4 is the Normal plot for the SAT scores of Figure 1.1.1, while Figure 1.3.5 is the Normal plot for the capping torques of Figure 1.1.2. In both cases the lines formed by the plotted points are essentially straight. The plots also contain statistical significance tests (the Anderson-Darling test, hence AD) which test for Normality. We will discuss statistical tests in the next chapter. Here, we use a rule of thumb which says that a p-value (p-values vary between 0 and 1) greater than 0.05 supports Normality, while one less than 0.05 rejects it. For both Normal plots the p-values are large and there is no reason to question the assumption of Normality.

Figure 1.3.4: Normal plot for SAT Verbal scores for 200 university students

Figure 1.3.5: Normal plot for Capping torques for 78 screw-capped containers

Figure 1.3.6 refers to the Urinary $\beta$-OHCS measurements, which we first encountered in Figure 1.1.5. The graph is far from straight and the AD p-value is very much less than 0.05; the graph displays the characteristic curvature of a right-
skewed distribution. Figure 1.3.7 is the Normal plot of the log-transformed data and it is quite consistent with an underlying Normal distribution.

Figure 1.3.6: Normal plot for Urinary β-OHCS 100 obese females

Figure 1.3.7: Normal plot for log-transformed β-OHCS measurements

Figure 1.3.8 is a Normal plot of the women’s heights measurements of Figure 1.1.4. Although it is very close to a straight line, it has a curiously clumped, almost stair-like, appearance. The p-value is very small indeed. The reason for this is that the measurements are rounded to integer values (centimetres) and, because the dataset is so large, there are many repeated values. A Normal distribution is continuous (the sampled values can have many decimal places) and few, if any, repeated values would be expected to occur; the AD test detects this unusual occurrence. To illustrate the effect of rounding on the plots, 1794 random numbers were generated from a Normal distribution with the same mean and standard deviation as the real data (the mean and SD of the sample generated are slightly different from those specified, due to sampling variability).
Figure 1.3.8: Normal plot for women’s heights measurements

Figure 1.3.9: Normal plot for simulated women’s heights measurements

Figure 1.3.9 shows a Normal plot for the data generated; as expected from random numbers, the sample values give a very good Normal plot. The data were then rounded to whole numbers (centimetres) and Figure 1.3.10 was drawn – its characteristics are virtually identical to those of the real data, showing that the cause of the odd line is the rounding; note the change in the p-value in moving from Figure 1.3.9 to Figure 1.3.10. Figure 1.3.11 shows a Normal plot for the Iris data of Figure 1.1.3 – the effect of the measurements being rounded to only one decimal place, resulting in multiple petals having the same length, is quite striking.
The assumption of data Normality underlies the most commonly used statistical significance tests and the corresponding confidence intervals. We will use Normal plots extensively in verifying this assumption in later chapters.

1.4 Combining Random Quantities

Often the quantities in which we are interested are themselves the result of combining other quantities that are subject to chance variation. For example, a journey is composed of two parts: X is the time it takes to travel from A to B, and Y is the time to travel from B to C. We are interested in S, the sum of the two travel times.

If we can assume that X is distributed with mean $\mu_X=100$ and standard deviation $\sigma_X=10$ (units are minutes), while Y is distributed with mean $\mu_Y=60$ and standard deviation $\sigma_Y=5$, what can we say about the total travel time, $S=X+Y$?
If we can assume that X and Y are independent (i.e., if X happens, for example, to be unusually long, this fact carries no implications for Y being either unusually long or short), then the following rules hold.

**Addition Rules**

\[ \mu_S = \mu_X + \mu_Y \]
\[ \sigma^2_S = \sigma^2_X + \sigma^2_Y \]

The underlying mathematics shows that both the means and the variances (the squares of the standard deviations) are additive. This means that the standard deviation of the overall travel time, S, is:

\[ \sigma_S = \sqrt{\sigma^2_X + \sigma^2_Y} \]

We might also ask, how much longer than the trip from B to C, the trip from A to B is likely to take (i.e., what can we say about D=X−Y). To answer such questions we need subtraction rules.

**Subtraction Rules**

If D = X − Y, then:

\[ \mu_D = \mu_X - \mu_Y \]
\[ \sigma^2_D = \sigma^2_X + \sigma^2_Y \]

Note that the variances add, even though we are subtracting. This gives:

\[ \sigma_D = \sqrt{\sigma^2_X + \sigma^2_Y} \]

The independence assumption is required for the rules related to the standard deviation; it is not required for combining means.

**Normality Assumptions**

The rules given above hold irrespective of the distributions that describe the chance variation for the two journeys, X and Y. However, if we can assume that X and Y are Normally distributed, then any linear combination of X and Y (sums or differences, possibly multiplied by constants) will also be Normally distributed. In such cases, the calculations required to answer our questions are simple; they are essentially the same as for Examples 1 and 2, once we have combined the component parts into the overall measured quantity, and then found its mean and standard deviation using the above rules.
Example 3

Given our assumptions above about the means and standard deviations of the two journey times, \(X\) and \(Y\), what is the probability that the overall journey time, \(S\),

(i) will take more than three hours (\(S > 180\) minutes)?
(ii) will take less than 2.5 hours (\(S < 150\) minutes)?

(i) From our addition rules we obtain:

\[
\mu_S = \mu_X + \mu_Y = 100 + 60 = 160
\]
\[
\sigma^2_S = \sigma^2_X + \sigma^2_Y = 10^2 + 5^2 = 125
\]
\[
\sigma_S = \sqrt{\sigma^2_X + \sigma^2_Y} = \sqrt{10^2 + 5^2} = \sqrt{125} = 11.18
\]

Overall travel times are, therefore, Normally distributed with a mean of 160 and a standard deviation of 11.18, as shown schematically in Figure 1.4.1.

The probability that \(S\) is more than 180 minutes is given by:

\[
P(S > 3\text{hours}) = P(S > 180) = P(Z > \frac{180 - \mu_S}{\sigma_S}) = P(Z > \frac{180 - 160}{11.18}) = P(Z > 1.79)
\]
\[
= 1 - P(Z < 1.79) = 1 - 0.9633 = 0.0367
\]

The probability that the overall travel time, \(S\), exceeds three hours (180 minutes) is approximately 0.04, as illustrated schematically in Figure 1.4.1.
Example 4

Suppose that $X$ (as given in the introduction, i.e. mean $\mu_X=100$ and standard deviation $\sigma_X=10$ (units are minutes)) is the time it takes a part-time student to travel to college each evening and she makes ten such trips in a term.

(i) If $T$ is the total travel time to college for the term, what are the mean and standard deviation for $T$?

(ii) What is the probability $T$ exceeds 18 hours (1080 minutes)?

Since only one trip type is involved, we do not need the $X$ sub-script. We have:

$$\mu_X = \mu = 100$$
$$\sigma_X = \sigma = 10$$

The total travel time is:

$$T = \sum_{i=1}^{10} X_i$$
where $X_i$ is the journey time for trip $i$ and $i = 1, 2, 3 \ldots 10$. All trips have the same characteristics, so $\mu_i=100$ and $\sigma_i=10$.

(i) $\mu_T = \mu_1 + \mu_2 + \mu_3 \ldots \mu_{10} = 10(100) = 1000$

$\sigma^2_T = \sigma^2_1 + \sigma^2_2 + \sigma^2_3 \ldots \sigma^2_{10} = 10(10^2) = 1000$

$\sigma_T = 31.62$

Total travel time is Normally distributed with a mean of 1000 and a standard deviation of 31.62, where the units are minutes.

(ii)

![Figure 1.4.3: The standard Normal curve and the total travel time curve](image)

$P(T > 1080) = P(Z > \frac{1080 - 1000}{31.62}) = P(Z > 2.53)$

$= 1 - P(Z < 2.53) = 1 - 0.9943 = 0.0057$

The chances that her total travel time to college for the term might exceed 18 hours (1080 minutes) is only about 6 in a thousand, as illustrated schematically in Figure 1.4.3. Note that the travel characteristics for travelling back from college could be quite different (different times might mean different traffic levels), so this would need examination if we were interested in her total amount of college travel.

**Exercises**

1.4.1. A final year psychology student approaches a member of staff seeking information on doing a PhD. In particular, she is concerned at how long it will take. The staff member suggests that there are essentially four sequential activities involved: **Literature Review** (reading and understanding enough of the literature in a particular area to be able to say where there are gaps that might be investigated), **Problem Formulation** (stating precisely what will be investigated, how this will be done, what measurement instruments will be required, what kinds of conclusions might be expected), **Data Collection** (developing the necessary research tools, acquiring subjects, carrying out the data collection, carrying out the statistical analysis), and **Writing up** the results (drawing conclusions, relating them to the prior
literature and producing the final thesis). She points out that the required times are variable, but that her experience suggests that they can be modelled by independent Normal distributions with the parameters shown in Table 1.4.1, where the time units are weeks.

<table>
<thead>
<tr>
<th>Activity</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literature review (L)</td>
<td>30</td>
<td>8</td>
</tr>
<tr>
<td>Problem formulation (P)</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>Data collection (D)</td>
<td>120</td>
<td>12</td>
</tr>
<tr>
<td>Write up (W)</td>
<td>16</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 1.4.1: Estimates for the means and SDs of the four activities (weeks)

Assuming that this is a reasonably accurate description of the situation (in practice, of course, the activities will overlap), what is the probability that the doctorate will take (i) less than 3 years (156 weeks)?; (ii) more than 3.5 years (182 weeks)?; (iii) more than 4 years (208 weeks)?

1.5 The effects of averaging

It is rarely the case that decisions are based on single measured values – usually, several values are averaged and the average forms the basis of decision making. We need, therefore, to consider the effect of averaging on our data. We will discuss this in the context of measuring the percentage dissolution of tablets after 24 hours in a dissolution apparatus. The ideas are general though; precisely the same statements could be made about examination results, prices of second-hand cars or the pH measurements made on samples of lake water.

Figure 1.5.1 illustrates the averaging effect. Figure 1.5.1(a) is an idealised histogram of individual tablet measurements; it represents what we might expect to get by drawing a histogram for many thousands of measurements of the percentage dissolution of individual tablets. Because of the large number of observations, the bars of the histogram could be made very narrow and a smooth curve would be expected to fit the histogram very closely. In practice, of course, we would rarely have such a large quantity of data available to us, but that does not prevent us from visualising what would happen under idealised conditions.

The data on which Figure 1.5.1(b) is based are no longer single tablet dissolution values. Suppose we randomly group the many thousands of tablet measurements into sets of four, calculate the means for each set, and draw a histogram of the means. Four values selected at random from the distribution described by (a) will, when averaged, tend to give a mean result close to the centre of the distribution. In order for the mean to be very large (say greater than 92) all four of the randomly selected tablet measurements would have to be very large, an unlikely occurrence. Figure 1.5.1(b) represents an idealised version of this second histogram. Of course we would never undertake such an exercise in the laboratory, though we might in

---

3 If the probability that one tablet was greater than 92 was 0.05, the probability that all four would be greater than 92 would be 0.00000625!
the classroom. The implications of the exercise are, however, very important. Figure 1.5.1 shows that the properties of means are very different from those of single measurements in one key aspect: they are much less variable. This difference is important, since virtually all data-driven decision making relies on means rather than on single measurements. The narrower curve is called ‘the sampling distribution of the mean’, as it describes the variation of means, based on repeated sampling, and its standard deviation is called ‘the standard error of the mean’ or ‘standard error’ for short.

Figure 1.5.1 Idealised histograms of the distributions of percentage dissolution of individual tablets (a) and of the averages of 4 tablets (b).

Mathematical theory shows that, if the standard deviation of the distribution of individual values is \( \sigma \), then the standard error of the distribution of sample means is \( \sigma / \sqrt{n} \), where \( n \) is the number of independent values on which each mean is based. Suppose the individual values vary about a long-run average value of 90 with a standard deviation of 0.5. It follows that averages of 4 values will also vary about a long-run mean of 90, but their standard error will be 0.25. Recall that approximately 95% of the area under any Normal curve lies within two standard deviations (or standard errors) of the mean. This implies that while 95% of individual tablet measurements would be between 89 and 91, 95% of means of 4 tablet measurements would be between 89.5 and 90.5. Thus, the effect of the averaging process is that a sample mean result is likely to be closer to the long-run or batch average, than a randomly selected single tablet measurement would be. This is why we base decisions on means, in practice.

The histograms/curves of Figure 1.5.1 are Normal (probably a reasonable assumption for tablet dissolution). However, what has been described does not depend on the assumption of an underlying Normal distribution. Even if the curve for individual tablets had been skewed, the mean of the sampling distribution would
be the same as that of the individual tablets and the standard error would still be \( \sigma/\sqrt{n} \). What is interesting, though, is that, as the number of measurements on which each mean is based increases, the shape of the sampling distribution approaches that of a Normal curve, even when the underlying distribution shape is non-Normal.

We will illustrate this interesting and important result using some data simulated from a skewed distribution, specifically from a chi-square distribution with 7 degrees of freedom (we will encounter the chi-square distribution later when we discuss significance tests; here, it is used simply as a typically skewed distribution curve).

Figure 1.5.2 shows the theoretical curve, while Figure 1.5.3 shows a histogram of 200 values randomly generated from this distribution.

![Figure 1.5.2: A chi-square curve with 7 degrees of freedom](image)

![Figure 1.5.3: Histogram of 200 values from a chi-square curve with 7 df](image)

Figure 1.5.4 shows the corresponding Normal plot which has the characteristic curved shape, we saw previously for skewed distributions.
Thirty values were randomly generated from this distribution and the mean was obtained. This was repeated 200 times, resulting in a sample of 200 means. A histogram of these means is shown in Figure 1.5.5, while Figure 1.5.6 represents a Normal plot of the set of 200 means.
The effect of averaging has been particularly successful here in producing means which closely follow a Normal distribution (not all samples will be quite so good). The mathematical theory says that as the sample size tends to infinity, the means become Normal; it is clear, though, that even with a sample size of 30 in this case, the Normal approximation is very good.

In pathological cases (e.g., extremely skewed distributions) the sampling distribution curve will not be Normal for the moderate sample sizes that are likely to be used in practice. In most cases, however, where the underlying curve is not too far from a Normal curve, the sampling distribution is reasonably close to a Normal curve, and this allows us use statistical methods based on the assumption of Normality. Apart from its common occurrence in practice (as seen in several examples in Section 1.1) this property of sampling distributions is the reason why the Normal curve is so important in statistics.

1.6 Measurement Systems

The same symmetrical characteristics as seen for the data of Section 1.1 are displayed in Figure 1.6.1, but it is different in one important respect – the picture is based on 119 repeated measurements of the same pharmaceutical material, whereas all our earlier examples were based on single measurements on different people or objects. The material measured was a control material used in monitoring the measurement process in a quality control analytical laboratory in a pharmaceutical manufacturing company [See Mullins [7] Chapters 1, 2]. The methods discussed in Section 1.3 show that a Normal curve is a good model for these data, also.
In using the Normal curve in a measurement context we have a different conceptual model for our data than that for the previously cited examples. Where the numbers referred to different physical entities (whether these were people, industrial components or flowers) the curves refer to variation between those units – some entities give values much higher or lower than others, but most cluster in the centre. For measurement systems we do not have a physical distribution; we have an hypothetical distribution of the many thousands of measurements that might be made on the single physical entity that is to be measured. In practice, of course, the vast majority of these measurements will never be made, so our questions about the ‘system’ of measurements are somewhat abstract in nature, but we are so accustomed to dealing with measurements that this abstraction often goes un-discussed. Two aspects of measurement quality are illustrated by Figure 1.6.2, namely bias and precision.

---

4 Alanine aminotransferase (ALT) is an enzyme found in most human tissue.
Suppose this picture represents smoothed histograms, based on thousands of measurements made on the same object/material by two different measurement systems A, B. It is clear that System A gives lower values, on average, i.e., the long-run mean for A is smaller than that for B, $\mu_A < \mu_B$. System A is said to be biased downwards relative to System B. Which system gives the ‘right’ results, on average, (i.e., is unbiased) we cannot know unless there is a measurement ‘gold standard’.

The spread of measurements for System B is greater than that for A – A is said to have better precision. The width of each curve is measured by the standard deviation, $\sigma$; the shapes of the curves in Figure 1.6.2 mean that $\sigma_A < \sigma_B$. There is no necessary relationship between the two parameters. There is no reason why $\mu_A$ could not be bigger than $\mu_B$, while $\sigma_A$ was less than $\sigma_B$.

Figure 1.6.3 [7, 8] shows serum ALT concentrations for 129 adults. It demonstrates that there is no necessary relationship between the shape of the curve describing variation in the physical system under study and the shape of the curve that describes variation in the measurement system used to measure the physical entities. The ALT distribution (Figure 1.6.3(a)) is right-skewed, that is, it has a long tail on the right-hand side, which means that some individuals have much higher ALT levels than the majority of the population.
variation in ALT measurements. It suggests that a Normal distribution might well describe the analytical variability. If, therefore, we were interested in studying the analytical procedure – for, example, to quantify the precision, which is obviously poor – we might work directly in the arithmetic scale (i.e., the scale in which the measurements were made. On the other hand, if we wished to make statements about the distribution of ALT concentrations in the population from which the subjects derive (by obtaining a reference interval, for example), then we might choose to transform the data before proceeding with our analysis (in order to get a Normal curve which is easier to analyse).

Figures 1.6.1 and 1.6.3 underline the importance of the Normal curve for the analysis of measurement systems, as the graphs of Section 1.1 showed its importance for describing chance variation in different physical populations.

**Exercises**

1.6.1.  
(i) An unbiased analytical system is used to measure an analyte whose true value for the parameter measured is 100 units. If the standard deviation of repeated measurements is 0.80 units, what fraction of measurements will be further than one unit from the true value? Note that this fraction will also be the probability that a single measurement will produce a test result which is either less than 99 or greater than 101, when the material is measured. What fraction of measurements will be further than 1.5 units from the true value? Give the answers to two decimal places.

(ii) Suppose now that the analytical protocol requires making either 2 or 3 measurements and reporting the mean of these as the measured value. Recalculate the probabilities for an error of one unit and compare them to those obtained for a single measurement.

1.7 Case Study: Statistical Process Control Charts

In this chapter we have discussed various aspects of the Normal curve as a model for statistical variation. This final section illustrates the use of a graphical tool for monitoring the stability of a system. This tool, the control chart, is based on the properties of the Normal curve and is an example of a powerful practical tool based on simple underlying ideas.

Statistical process control (SPC) charts were introduced for monitoring production systems but, as we shall see, they have much wider applicability. Conceptually they are very simple: a small sample of product is taken regularly from the process and some product parameter is measured. The values of a summary statistic (such as the sample mean) are plotted in time order on a chart; if the chart displays other than random variation around the expected result it suggests that something has changed in the process. To help decide if this has happened control limits are plotted on the chart: the responses are expected to remain inside these limits. Rules are decided upon which will define non-random behaviour.
The most commonly used charts were developed by Shewhart [9] in Bell Laboratories in the nineteen twenties. He described variation as being due either to ‘chance causes’ or ‘assignable causes’. Chance causes are the myriad of small influences that are responsible for the characteristic Normal shape, which we saw in a variety of settings in Section 1.1. Shewhart’s crucial insight was that stability must be understood in the context of this inevitable chance variation, not in the commonplace notion of ‘unchange’. Assignable causes are larger: they tend to change the mean or the standard deviation of a process distribution. If a process is stable such that it exhibits only random variation around a given reference value and the size of that variation (as measured by the standard deviation) remains constant, Shewhart described it as being ‘in statistical control’ or simply ‘in control’. The objective in using control charts is to achieve and maintain this state of statistical control. Control charts were developed to control engineering production processes, but, as Charts 2, 3 and 4 will show, they can be applied just as effectively to measurement systems. Any apparent differences between the two situations disappear when we think of a measurement system as a production process whose output is measurements. Chart 5 illustrates the use of a control chart in a retail sales setting.

For an extensive discussion of control charts in a manufacturing environment see Montgomery [10]. For an introduction to the use of control charts for monitoring measurement systems see Mullins [7]. Stuart [11] gives a less technical introduction to control charts in industry than does Montgomery; he also discusses their application in a retail business context (see Chart 5, which is ‘borrowed’ from Stuart!).

**Chart 1: Monitoring Production Systems**

Figure 1.7.1 shows a control chart for a critical dimension (mm) of moulded plastic medical device components, each of which contained a metal insert purchased from an outside supplier.

Twenty five samples of size n=5 were taken from the stream of parts manufactured over a period of several weeks, with a view to setting up control charts to monitor the process. The five sample results were averaged and it is the average that is plotted in the chart. For now, the centre line (CL) and control limits (upper and lower control limits, UCL and LCL) will be taken as given; later, the rationale for the limits will be discussed.
In using control charts different companies/users apply different rules to identify instability. For example, the SPC manual of the Automotive Industry Action Group [12] (AIAG: a task force set up by the Ford, Chrysler and General Motors companies to unify their SPC procedures) suggests that action is required if any of the following occur:

- any point outside of the control limits;
- a run of seven points all above or all below the central line;
- a run of seven points in a row that are consistently increasing (equal to or greater than the preceding points) or consistently decreasing;
- any other obviously non-random pattern.

Runs of nine points, rather than seven, are also commonly recommended. The same basic principle underlies all the rules: a system that is in statistical control should exhibit purely random behaviour - these rules correspond to improbable events on such an assumption. Accordingly, violation of one of the rules suggests that a problem has developed in the process and that action is required. The rationale for the rules is discussed below.

The manufacturing process to which Figure 1.7.1 refers appears, by these rules, to be in statistical control. No points are outside the control limits and there are no long runs of points upwards or downwards or at either side of the central line.

**Chart 2: Monitoring a HPLC Assay**

To monitor the stability of a measurement system, a control sample may be measured repeatedly over time, together with the test samples that are of primary interest. The control sample results are then plotted on a control chart. The centre
line for this chart is the accepted correct value for the control material (typically the mean of recent measurements on the material), and newly measured values for the control material are expected to vary at random around the centre line. Sometimes the centre line might be the certified value for a bought-in or a ‘house’ reference material.

Figure 1.7.2 shows a control chart for a HPLC potency assay of a pharmaceutical product [7]. The data displayed in the chart were collected over a period of several months. At each time point two replicate measurements were made on a control material, which was just a quantity of material from one batch of the production material routinely manufactured and then measured in the laboratory. These results were averaged and it is the average that is plotted in the chart.

The average level of the measurement system to which Figure 1.7.2 refers appears, by the AIAG rules, to be in statistical control. No points are outside the control limits and there are no long runs of points upwards or downwards or on either side of the central line. Accordingly, we can feel confident that the analytical system is stable and producing trustworthy results.

**Chart 3: Monitoring both Production and Analytical Systems**

Figure 1.7.3 illustrates the benefits of having control charts in use for both production results and analytical control data, when product reviews are carried out.

Figure 1.7.3(a) shows potencies for a series of batches of a drug with a clear downwards shift in results. Such a shift suggests problems in the production system, but would often lead to production management questioning the quality of the analytical results. It has to be someone else’s fault!
As shown in Figure 1.7.3(b), a stable control chart for the analytical system for the same period is the best possible defence of the stability of the analytical results. If, on the other hand, there had been a downward shift in the analytical system it would have been reflected in Figure 1.7.3(b). Recognising such a shift would save much time that would otherwise be wasted searching for a non-existent production problem. Thus, the availability of control charts for both the production and analytical systems allows data-driven decision-making and can reduce inter-departmental friction when problems arise at the interface\(^5\).

---

**Chart 4: An Unstable Analytical System**

Figure 1.7.4, represents 49 measurements on control samples made up to contain 50 parts per billion (ppb) of iron in pure water [7]. The sample was analysed by an Inductively Coupled Plasma (ICP) Spectrometer over a period of months.

\[^5\text{In fact, the problem here could be due to changes in an in-coming raw material; so a third control chart might well be worthwhile!}\]
Figure 1.7.4: A control chart for the iron (Fe) content of a water sample (ppb)

The diagram shows an analytical system which is clearly out of control; in fact, three gross outliers were removed before this chart was drawn. The data were collected retrospectively from the laboratory records; the laboratory did not use control charts routinely. They are presented here because they exhibit the classic features of an out-of-control system: several points outside the control limits, a run of points above the centre line and a run of points downwards. There is an obvious need to stabilise this analytical system.

There are two ways in which control charts are used, viz., for assessing retrospectively the performance of a system (as in Figure 1.7.4) and for maintaining the stability of a system, which is their routine use once stability has been established. Where a chart is being used to monitor a system we would not expect to see a pattern such as is exhibited between observations 17 and 32 of Figure 1.7.4, where 16 points are all either on or above the centre line. Use of a control chart should lead to the upward shift suggested here being corrected before such a long sequence of out-of-control points could develop.

Note that if the chart were being used for on-going control of the measurement process, the centre line would be set at the reference value of 50 ppb. Here the interest was in assessing the historical performance of the system, and so the data were allowed determine the centre line of the chart.
The Theory Underlying the Control Limits

The centre line (CL) for the chart should be the mean value around which the measurements are expected to vary at random. In a manufacturing context, this could be the target value for the process parameter being measured, where there is a target, but in most cases it is the mean of the most recent observations considered to be 'in-control'. Similarly, in a measurement context, the centre line will either be the mean of recent measurements on the control material, or the accepted 'true value' for an in-house standard or a certified reference material.

The control limits are usually placed three standard deviations above and below the centre line, Figure 1.7.5(b); the standard deviation of individual measurements is used if the points plotted are individual measurements, as in Chart 4, and the standard error of the mean, if the plotted points are means of several measured values, as in Chart 1. This choice may be based on the assumption that the frequency distribution of chance causes will follow a Normal curve, or it may be regarded simply as a sensible rule of thumb, without such an assumption. As we have seen a distribution curve can be thought of as an idealised histogram: the area under the curve between any two values on the horizontal axis gives the relative frequency with which observations occur between these two values. Thus, as shown in Figure 1.7.5(a), 99.74% of the area under any Normal curve lies within three standard deviations (3σ) of the long-run mean (μ) and so, while the system remains in control, 99.7% of all plotted points would be expected to fall within the control limits.

Where only single values are measured (as frequently happens in process industries, for example) the control limits are:

Upper Control Limit (UCL) = CL + 3σ

Lower Control Limit (LCL) = CL − 3σ

where σ is the standard deviation of individual values.

If, as is more commonly the case in manufacturing of components, the point to be plotted is the average of a number of values, say n, then an adjustment has to be made to take account of the fact that means, rather than individual values, are being plotted. If individual values give a frequency distribution, as shown in Figure 1.7.5(a), with mean μ and standard deviation σ then a very large number of averages, each based on n values from the same process, would give a similar frequency distribution centred on μ but the width would be defined by σ/√n (the 'standard error'). The distribution would be tighter and more of the values would be close to μ. This simply quantifies the extent to which the averaging process reduces variability, as discussed in Section 1.5: see Figure 1.5.1, in particular.
To obtain the correct control limits when plotting averages, we simply replace $\sigma$ by $\sigma/\sqrt{n}$ in the expressions given above and obtain:

\[
\text{Upper Control Limit (UCL)} = CL + 3\sigma/\sqrt{n} \\
\text{Lower Control Limit (LCL)} = CL - 3\sigma/\sqrt{n}
\]

where $\sigma$ is (as before) the standard deviation of individual values. The chart based on single measurements is sometimes called an 'Individuals or X-chart' while the chart based on means is called an 'X-bar chart'; the two charts are, however, essentially the same. Note that in a laboratory measurement context, where $n$ is typically 2 or 3, the simplest and safest approach is to average the replicates and treat the resulting means as individual values (see Mullins [7], Chapter 2).

The limits described above are known as 'three sigma limits', for obvious reasons; they are also called 'action limits'. They were first proposed by Shewhart in the
nineteen twenties and are almost universally used with this type of control chart. Their principal justification is that if the process is in control then the probability of a point going outside the control limits is very small: about three in a thousand. Accordingly, they give very few false alarm signals. If a point does go outside the control limits it seems much more likely that a problem has arisen (e.g., the process average has shifted upwards or downwards; see Figure 1.7.5(c)) than that the process is stable, as in Figure 1.7.5(b), but the chance causes have combined to give a highly improbable result. This is the rationale for the first AIAG rule, quoted above.

The basis for the second AIAG rule (a run of seven points all above or all below the central line) is that if the system is in control the probability that any one value is above or below the central line is $1/2$. Accordingly, the probability that seven in a row will be at one side of the central line is $(1/2)^7 = 1/128$; again, such an occurrence would suggest that the process has shifted upwards or downwards. The third rule (a run of seven points in a row that are consistently increasing (equal to or greater than the preceding points) or consistently decreasing) has a similar rationale: if successive values are varying at random about the centre line, we would not expect long runs in any one direction.

The last catch-all rule (any other obviously non-random pattern) is one to be careful of: the human eye is adept at finding patterns, even in random data. The advantage of having clear-cut rules that do not allow for subjective judgement is that the same decisions will be made, irrespective of who is using the chart. Having said this, if there really is ‘obviously non-random’ behaviour in the chart (e.g., cycling of results between day and night shift) it would be foolish to ignore it.

The assumption that the data follow a Normal frequency distribution is more critical for charts based on individual values than for those based on averages. Averages, as we saw in Section 1.5, tend to follow the Normal distribution unless the distribution of the values on which they are based is highly skewed. In principle, this tendency holds when the averages are based on very large numbers of observations, but in practice samples of even four or five will often be well behaved in this regard. If there is any doubt concerning the distribution of the measurements a Normal probability plot (see Section 1.3) may be used to check the assumption.

Chart 5: A retail sales control chart

Stuart [11] considered the problem of monitoring cash variances in the till of a retail outlet – a sports club bar. At the end of each day’s trading, the cash in the till is counted. The till is then set to compute the total cash entered through the keys, which is printed on the till roll. The cash variance is the difference between cash and till roll totals (note that ‘variance’ as used here is an accounting term, which is not the same as the statistical use of the same word, i.e., the square of the standard deviation). Small errors in giving change or in entering cash amounts on the till keys
are inevitable in a busy bar. Ideally, the cash variance will be zero. In the presence of inevitable chance variation, the best that can be hoped for is that the cash variances will centre on zero. In other words, the cash variance process is ideally expected to be a pure chance process with a process mean of zero.

A control chart covering Sunday cash variances for a year is shown in Figure 1.7.6. The variances are recorded in pounds (the data pre-date the Euro). The centre line, is placed at zero, the desired process mean. The control limits are placed at ±£12; the limits were established using data from the previous year.

As may be seen from the chart, this process was well behaved for the most part, apart from three points outside the control limits. For the first point, week 12, value £14.83, the large positive cash variance indicates either too much cash having been put in the till or not enough entered through the till keys. When the bar manager queried this, one of the bar assistants admitted a vague recollection of possibly having given change out of £5 instead of out of £10 to a particular customer. The customer involved, when asked, indicated that she had suspected an error, but had not been sure. The £5 was reimbursed.

In the second case, week 34, value –£22.08, an assistant bar manager who was on duty that night paid a casual bar assistant £20 wages out of the till but did not follow the standard procedure of crediting the till with the relevant amount.

The last case, week 51, value -£18.63, was unexplained at first. However, other out of control negative cash variances on other nights of the week were observed, and these continued into the following year. Suspicion rested on a recently employed casual bar assistant. An investigation revealed that these cash variances occurred.
only when that bar assistant was on duty and such a coincidence applied to no other member of staff. The services of that bar assistant were ‘no longer required’ and the cash variances returned to normal.

1.8 Conclusion

This chapter has introduced the concept of statistical variation – the fact that objects or people selected from the same population or process give different numerical results when measured on the same characteristic. It also showed that, where the same object/person was measured repeatedly, different numerical results were obtained, due to chance measurement error. In many cases, the variation between numerical results can be described, at least approximately, by a Normal curve.

This curve will often describe the variation between means of repeated samples, also, even where the raw data on which the means are based come from distributions that are not themselves Normal. This fact is of great importance for what follows: the most commonly used statistical procedures are based on means; therefore, statistical methods based on Normal distributions will very often be applicable in their analysis. For this reason, the properties of the Normal curve and its use in some simple situations (e.g., developing reference intervals in medicine, or use of control charts in industrial manufacturing or in monitoring measurement systems) were discussed in the chapter.
References

Outline Solutions

1.2.1.

(a) A weight of 85 mg corresponds to a Z value of -2.14 as shown below – the standard Normal table gives an area of 0.0162 to the left of -2.14, so the area to the left of 85 is also 0.0162.

\[ z = \frac{x-\mu}{\sigma} = \frac{85-100}{7} = -2.14 \]

\[ P(X < 85) = P(Z < -2.14) = P(Z > 2.14) = 1 - P(Z < 2.14) = 1 - 0.9838 = 0.0162 \]

By symmetry, the area above 115 is also 0.0162

Approximately, 1.6% of tablets will have weights either greater than 115 mg or less than 85 mg, which means that about 3.2% of tablets are outside these limits.

(b) The calculations for part (b) are carried out in exactly the same way as those for (a).
(c) (i) If we inspect the standard Normal table to find the value that has an area of 0.99 below it, we find \( Z = 2.33 \); therefore, \(-2.33\) will have an area of 0.01 below it.

\[
P(Z < -2.33) = 0.01
\]

\[
Z = -2.33 = \frac{K - 100}{7}
\]

This means that \( K = 100 - 2.33(7) = 100 - 16.31 = 83.69 \)

99% of tablets would be expected to weigh more than 83.69 mg.

(ii)

\[
P(Z < 2.575) = 0.995 \Rightarrow P(Z > 2.575) = 0.005
\]

and, by symmetry, \( P(Z < -2.575) = 0.005 \)

\[
Z = 2.575 = \frac{B - 100}{7}
\]

This means that \( B = 100 + 2.575(7) = 100 + 18.025 = 118.025 \) mg

By symmetry \( A = 100 - 2.575(7) = 100 - 18.025 = 81.975 \) mg
1.2.2. In the log_{10} scale we have a Normal distribution with mean 2.7423 and standard deviation 0.1604.

(i) Calculate the bounds (a, b) for the 95% reference interval in the log scale.

The reference bounds are: μ−1.96σ and μ+1.96σ which gives:

2.7423 ± 1.96(0.1604) or 2.4279 to 3.0567.

(ii) Use your results to obtain the corresponding reference interval in the original measurement scale: the bounds will be (10^a, 10^b).

10^{2.4279}=267.9 and 10^{3.0567}=1139.5, so our reference interval in the original measurement scale is 267.9 to 1139.5 (i.e., 268 to 1140 in practice). Note that although the mean 2.7423 is the centre of the interval in the log scale, when we back transform [10^{2.7423}=552.5], the corresponding value in the arithmetic scale is not centred between the two bounds – the right-hand bound is much further above 552.5 than the lower bound is below it. This reflects the skewed shape of the original histogram. In fact, 552.5 is an estimate of the median (the steroid value below which and above which 50% of the women lie). For a symmetrical distribution the mean and the median coincide, but for right-skewed distributions the mean will be higher; thus, the mean of the 100 original steroid values is 590.8 – the long tail pulls the mean upwards. Income distributions, for example, are typically quite skewed, and the relatively small number of very high income individuals increases the average income – the median value is, therefore, a better one-number summary of a ‘typical’ member of the population.

Note also that the calculations above will be quite sensitive to the number of decimal places used.

(iii) What value, U, should be quoted such that only 1% of women from this population would be expected to have a β-OHCS measurement which exceeds U?
The critical value on the standard Normal curve which has 0.99 to the left and 0.01 to the right is 2.33. From the relation:

\[ z = \frac{x - \mu}{\sigma} \]

it is clear that the corresponding value for a Normal curve with mean \( \mu = 2.7423 \) and standard deviation \( \sigma = 0.1604 \) is \( \mu + 2.33\sigma = 3.1160 \). Only 1% of the Log\(_{10}(\beta\text{-OHCS}) \) values would be expected to exceed Log\(_{10}(U) = 3.1160 \). The corresponding value in the original measurement scale is \( U = 10^{3.1160} = 1306.2 \).

1.4.1. We begin by finding the parameters for the doctorate completion time, \( T \)

\[ \mu_T = \mu_L + \mu_P + \mu_D + \mu_W = 30 + 10 + 120 + 16 = 176 \]
\[ \sigma^2_T = \sigma^2_L + \sigma^2_P + \sigma^2_D + \sigma^2_W = 8^2 + 3^2 + 12^2 + 3^2 = 226 \]
\[ \sigma_T = \sqrt{\sigma^2_L + \sigma^2_P + \sigma^2_D + \sigma^2_W} = 15.03 \]

Since all four components of the total time are assumed to be Normally distributed, the total time, \( T \), will also have a Normal distribution.
(i) \[ P(T < 3 \text{ years}) = P(T < 156 \text{ weeks}) \]

\[ P(T < 156) = P(Z < \frac{T - \mu_T}{\sigma_T}) = P(Z < \frac{156 - 176}{15.03}) = P(T < -1.33) = P(T > 1.33) \]

\[ = 1 - P(T < 1.33) = 0.09 \]

The probability that the doctorate will take less than 3 years (156 weeks) is 0.09.

(ii) \[ P(T > 3.5 \text{ years}) = P(T > 182 \text{ weeks}) \]

\[ P(T > 182) = P(Z > \frac{T - \mu_T}{\sigma_T}) = P(Z > \frac{182 - 176}{15.03}) = P(T > 0.3992) \]

\[ = 1 - P(T < 0.40) = 1 - 0.655 = 0.345 \]

The chances of the postgraduate work taking more than three and a half years (182 weeks) are about 0.35.
Similarly, the probability that the work will take more than 4 years (208 weeks) is about 0.02, as shown below.

\[ P(T > 208) = P(Z > \frac{T - \mu_T}{\sigma_T}) = P(Z > \frac{208 - 176}{15.03}) = P(T > 2.13) \]
\[ = 1 - P(T < 2.13) = 1 - 0.9834 = 0.0166 \]

This example illustrates the analysis of project planning networks. Of course, the current set of tasks is the simplest possible ‘network’ – a linear sequence of activities. Real projects (consider for example the construction of an electricity generating station) may involve thousands of activities, some of which may proceed simultaneously (‘in parallel’), while others must be fully completed before others may begin (e.g., foundations must be in place before walls can be built – these activities are ‘in series’).

(i) A measurement of 101 corresponds to a Z value of 1.25 as shown below – the standard Normal table gives an area of 0.8944 to the left of 1.25, so the area to the right of 1.25 is 0.1056. The probability that a single measurement will exceed 101 is, accordingly, 0.1056.

\[ z = \frac{x - \mu}{\sigma} = \frac{101 - 100}{0.8} = 1.25 \]
\[ P(X > 101) = P(Z > 1.25) = 1 - P(Z < 1.25) = 1 - 0.8944 = 0.1056 \]

By symmetry, the probability that a single measurement will be less than 99 is also 0.1056. Hence, the probability that a single measurement will be more than one unit from the true value is 2(0.1056) or 0.21.

The corresponding calculation for a divergence of 1.5 units shows a probability of 0.06.

(ii) We saw in Section 1.5 that means are less variable than single measurements and, specifically, that the variability of means is given by the standard error (the standard deviation for single values divided by the square root of the number of values on which the mean is based). To allow for the averaging in our calculations, therefore, we simple replace the standard deviation by the standard error.

The probability that a mean of two measurements will exceed 101 is given by:
\[
z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{101 - 100}{0.8 / \sqrt{2}} = 1.77
\]

Where \(\bar{x}\) now represents the mean of two measurements.

\[\Pr(\bar{x} > 101) = \Pr(Z > 1.77) = 1 - \Pr(Z < 1.77) = 1 - 0.9616 = 0.0384\]

By symmetry, the probability that the mean of two measurements will be less than 99 is also 0.0384. Hence, the probability that the mean of two measurements will be more than one unit from the true value is \(2(0.0384)\) or 0.08.

The value of making even two measurements instead of one is clear – the probability of an error of one unit has been reduced from 0.21 to 0.08.

The corresponding probability for the mean of three measurements is 0.03.