4

Statistical models

All models are wrong, but some are useful. (Box, 1999, p. 23)

Overview

Control charts are versatile statistical tools with a role to play in all four of the measure, analyse, improve and control phases of Six Sigma projects. In order to develop control charts from run charts, some understanding of statistical models for both discrete and continuous random variables is required, in particular of the normal or Gaussian statistical model. The normal distribution is also fundamental in understanding of the concept of sigma quality level referred to earlier in Section 1.1. Brief reference will also be made to the multivariate normal distribution. An understanding of statistical models in turn necessitates some fundamental knowledge of probability.

Finally, knowledge of the statistical properties of sums of independent random variables yields important results concerning means and proportions – results that are vital for the assessment of whether or not changes made during the improve phase of a Six Sigma project have been effective, for an appreciation of the way in which the various sources of errors in measurement processes contribute to the overall measurement error, and for an understanding of the construction of control charts.

Unlike other chapters, this one includes a number of exercises at various points, some of which do not require the use of Minitab. They are included to help the reader understand the topics of probability and statistical models as they are developed. Solutions to these exercises are provided on the book's website.

4.1 Fundamentals of probability

4.1.1 Concept and notation

Probability theory developed in the seventeenth century due, to some extent, to the dialogue between gamblers and mathematicians concerning the odds for games of chance involving

Six Sigma Quality Improvement with Minitab, Second Edition. G. Robin Henderson.

Table 4.1 Relative frequencies of sound tees.

Number tested, n	Number sound, f	Relative frequency, f/n
10	9	0.900 000
100	85	0.850 000
1 000	780	0.780 000
10 000	7 952	0.795 200
100 000	80 042	0.800 420
1 000 000	799 631	0.799 631

dice and cards. Further impetus came from the development of astronomy in the eighteenth and nineteenth centuries and the work of mathematicians Gauss and Legendre on problems such as the development of models for the orbits of comets. Subsequently the normal or Gaussian probability distribution provided Walter Shewhart with the foundation for the control chart in 1924 and was central to the development of inferential statistics by William Gosset and Ronald Fisher around the same time.

Consider an injection-moulding process for the manufacture of plastic golf tees that is behaving in a stable, predictable manner and where we record successive tees as either sound or defective. Suppose Table 4.1 summarizes findings as the recording of data progresses.

A plot of the relative frequency of sound tees against the logarithm to base 10 of the number of tees tested is shown in Figure 4.1. (The sequence of numbers tested is 10, 100, 1000, etc., which may be written as 10^1, 10^2, 10^3, etc. The sequence of indices 1, 2, 3, etc. is the sequence of logarithms to base 10 of the sequence of numbers tested in the first column.)

For small numbers tested, the relative frequency is unstable but, as the number tested becomes large, the relative frequency stabilizes at around 0.8. For this test there are two outcomes; the tee is either sound (S) or defective (D). The set of possible *outcomes* $\{D, S\}$ is called the sample space for the testing of a tee. The value 0.8 can be assigned as the *probability*

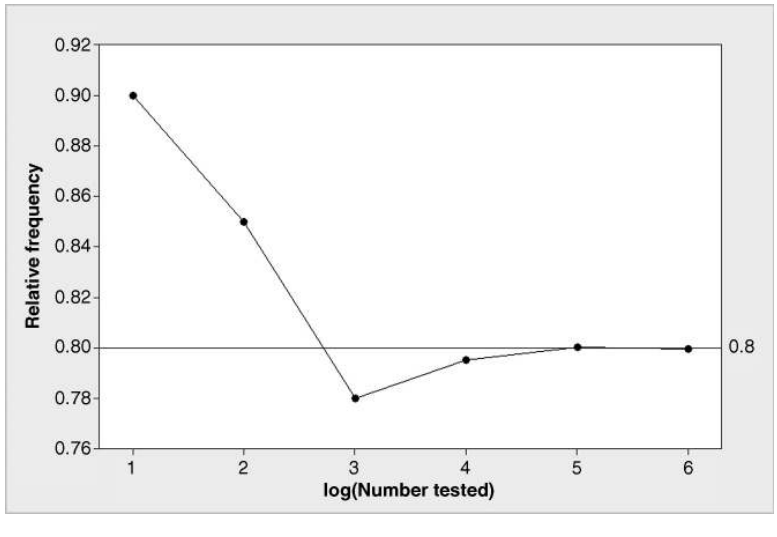

Figure 4.1 Relative frequency of sound tees.

of the outcome that a tee is sound. One can write in shorthand $P(S) = 0.8$. A little reflection should convince the reader that the relative frequency of defective tees would stabilize at 0.2. Therefore, $P(D) = 0.2$. The total of the probabilities assigned to the outcomes in a sample space is 1.

If a conventional cubic die is rolled then the sample space is $\{1, 2, 3, 4, 5, 6\}$, where the integers represent the number showing on the uppermost face of the die when it comes to rest. It is natural to assign probability 1/6 to each of the six outcomes since one would expect the relative frequency of each outcome to stabilize at the same value and the total of the six probabilities is 1. An *event* is defined as a subset of the sample space. For example, the subset $A = \{2, 4, 6\}$ of the sample space above is the event of rolling an even number. The probability of an event is the sum of the probabilities of the constituent outcomes. (Note that an event may consist of a single outcome.) Thus $P(A) = 1/6 + 1/6 + 1/6 = 1/2$. This means that in a long sequence of rolls of a die one would expect the relative frequency of an even result to stabilize at 0.5. Probability is a measure on a scale of 0 to 1 inclusive and may be considered as 'long-term' relative frequency of occurrence. If the probability of an event is 0 then it is impossible for the event to occur. If the probability of an event is 1 then the event is certain to occur.

Let us denote by $P(A \cup B)$ the probability of *either* event A *or* event B (*or* both) occurring is denoted $P(A \cup B)$, and by $P(A \cap B)$ the probability of *both* event A *and* event B occurring. (The 'cup' $\cup$ and 'cap' $\cap$ symbols are the union and intersection symbols used in the mathematics of sets. An informal way to remember that the $\cap$ symbol corresponds to *both* ... *and* ... is to think of fish $\cap$ chips!)

If for the rolling of a die the event B is defined as the score being a multiple of 3, then $B = \{3, 6\}$ so that $A \cup B = \{2, 3, 4, 6\}$ and $A \cap B = \{6\}$. The *complement* of event A, denoted by $\bar{A}$ and referred to as the event 'not-A', is the *non-occurrence* of A. Here $\bar{A} = \{1, 3, 5\}$. The event $B \mid A$ ('B given A') is the event that B occurs, knowing that A has occurred. Thus, for the example, it is the event that the score is a multiple of 3, given the information that it is an even number. Knowing that A has occurred, we are dealing with the reduced sample space consisting of the outcomes 2, 4 and 6, so we assign revised probabilities of one-third to each of these. Hence, $P(B|A) = 1/3$ since, of the three outcomes in the reduced sample space, only one is a multiple of 3. You should verify that $P(A|B) = 1/2$.

Exercise 4.1 Let C be the event that the score is a prime number, so that $C = \{2, 3, 5\}$. Write down $P(C), P(A|C), P(B|C), P(C|A)$ and $P(A \cap C)$.

4.1.2 Rules for probabilities

There are various rules for combining probabilities. Three fundamental ones will be considered.

Rule 1. $P(\bar{E}) = 1 - P(E)$.

For the example of the die above, $P(B) = 1/3$, so Rule 1 yields $P(\bar{B}) = 1 - P(B) = 2/3$. This means that the probability of a result that is not a multiple of three is $2/3$. This can be taken as an indication that out of every, say, 300 rolls of a die, one could expect two thirds of the rolls, i.e. 200 rolls, to yield results that are not multiples of three.

If the event $E_1 \cap E_2 = \{\ \}$, the empty set, then the events E_1 and E_2 are said to be *mutually exclusive*. In this case $P(E_1 \cap E_2) = 0$, i.e. when E_1 and E_2 are mutually exclusive it is impossible that both occur simultaneously.

Students in an introductory statistics course participated in a simple experiment. Each student recorded his or her height, weight, gender, smoking preference, usual activity level, and resting pulse. Then they all flipped coins, and those whose coins came up heads ran in place for one minute. Then the entire class recorded their pulses once more.

Column	Name	Count	Description
C1	Pulse1	92	First pulse rate
C2	Pulse2	92	Second pulse rate
C3	Ran	92	1 = ran in place 2 = did not run in place
C4	Smokes	92	1 = smokes regularly 2 = does not smoke regularly
C5	Sex	92	1 = male 2 = female
C6	Height	92	Height in inches
C7	Weight	92	Weight in pounds
C8	Activity	92	Usual level of physical activity: 1 = slight 2 = moderate 3 = a lot

Panel 4.1 Description of the Pulse.MTW data set.

Rule 2. If the events E_1 and E_2 are mutually exclusive then

$$P(E_1 \cup E_2) = P(E_1) + P(E_2).$$

Thus when dealing with an *either ... or* situation for ***mutually exclusive*** events one must *add* probabilities.

In order to introduce the third rule reference will be made again to the Pulse.MTW data set encountered in the previous chapter. The description of the data set displayed in Panel 4.1 may be accessed using **Help > Help**, clicking on the **Search** tab, entering Pulse in the **Type in the word(s) to search for:** window, clicking the **List Topics** button and then double-clicking on PULSE.MTW in the list of topics.

With the worksheet PULSE.MTW open, use of **Stat > Tables > Tally Individual Values...**, with Ran entered under **Variables:** and both **Counts** and **Percent** checked, yields the Session window output displayed in Panel 4.2. This indicates that 35 of the 92 students in the class ran. As the decision whether or not to run was meant to be based on the outcome of the toss of a coin, one would expect the relative frequency of heads to have been approximately 0.50 while in fact it was 0.38, to two decimal places. (Did some students cheat through not running in place in spite of obtaining a head on their coins? This type of question will be addressed formally in Chapter 7.)

Tally for Discrete Variables: Ran

Ran	Count	Percent
1	35	38.04
2	57	61.96
N=	92	

Panel 4.2 Tally of variable Ran.

Use of **Stat > Tables > Cross Tabulation and Chi-Square...** with **Categorical variables:** specified as **For rows:** Sex and **For cols:** Smokes, with **Counts** and **Row percents** checked, yields the table in Panel 4.3. The value 1 for the variable Sex indicates a male and the value 1 for the variable Smokes indicates a regular smoker. Thus, to the nearest whole per cent, 30% of the students smoked regularly, while 23% of the female students smoked regularly and 35% of the male students smoked regularly. Is smoking gender-dependent? If $P(E_2|E_1) = P(E_2)$ then we say that the events E_1 and E_2 are *independent*. Let S denote the event that a student is a regular smoker and let F denote the event that a student is female. If we regard the 92 students as a sample from a population of students then we have the estimates $P(S) = 30\% = 0.30$ and $P(S|F) = 23\% = 0.23$. The fact that these two estimates differ suggests that smoking might be gender-dependent in the student population. Formal assessment of dependence will be considered in Chapter 10.

Rule 3. If the events A and B are **independent** then

$$P(A \cap B) = P(A) \times P(B).$$

Thus when dealing with a *both ... and* situation for independent events one must *multiply* probabilities.

Consider blood groups in the UK population. The mutually exclusive groups are O, A, B and AB with respective probabilities 0.46, 0.42, 0.09 and 0.03 (British Broadcasting Corporation, 2004). Thus, for example, the probability that a randomly selected member of the population has blood of group either A or B is $0.42 + 0.09 = 0.51$. The rhesus factor is

Tabulated statistics: Sex, Smokes

Rows: Sex Columns: Smokes

	1	2	All
1	20	37	57
	35.09	64.91	100.00
2	8	27	35
	22.86	77.14	100.00
All	28	64	92
	30.43	69.57	100.00

Cell Contents: Count
% of Row

Panel 4.3 Cross-tabulation of Sex and Smokes.

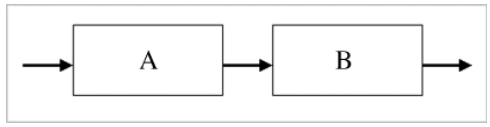

Figure 4.2 System with two subsystems.

present in the blood of 85% of the UK population and absent from the blood of the remainder of the population. Those with the factor present are said to be rhesus positive, while those with the factor absent are said to be rhesus negative. The presence or absence of the rhesus factor is independent of blood group. The author is 'O rhesus negative'. We have $P(\text{Rhesus positive}) = 0.85$, so by Rule 1,

$$P(\text{Rhesus negative}) = 1 - 0.85 = 0.15.$$

By Rule 3,

$$P(\text{Both O and rhesus negative}) = P(\text{O}) \times P(\text{Rhesus negative}) = 0.46 \times 0.15 = 0.069.$$

Thus just less than 7% of the UK population are O rhesus negative.

The multiplication rule for the combination of probabilities of independent events can be applied to systems made up of a series of subsystems or processes that function independently of each other. Figure 4.2 depicts a system comprising two subsystems, A and B. Let the probability that each of the subsystems functions correctly be 0.9. Assuming independence, the multiplication rule yields probability $0.9 \times 0.9 = 0.81$ that the overall system functions correctly. This probability of 0.81 is often referred to as the reliability of the system. With three such subsystems the reliability would be $0.9 \times 0.9 \times 0.9 = 0.9^3 = 0.729$.

Reference was made in Chapter 1 to the concept of sigma quality level (or sigma) for a process. Consider a system consisting of 10 independent subsystems where each of the subsystems is created by a process operating with a sigma quality level of 3. As can be seen from Appendix 1, a sigma quality level of 3 corresponds to 66 811 nonconformities per million, which equates to a probability of $1 - \frac{66\,811}{1\,000\,000} = 0.933\,189$ that each subsystem functions correctly. The reliability of the system would be $0.933\,189^{10} = 0.5008$, to four decimal places. Thus only 50% of the systems would function correctly. (This calculation may be performed in Minitab using **Calc** > **Calculator** with **Expression:** 0.933 189**10 and **Store result in variable:** Answer. The result may then be read from the column named Answer created in the current worksheet.)

Exercise 4.2 A system consists of 1000 subsystems each of which is produced by a Six Sigma process, i.e. a process with sigma quality level 6. Verify that the relative frequency of system failure is 3 in 1000.

A key contributor to the development of Six Sigma at Motorola was senior quality engineer Bill Smith. One product he worked on had a much higher failure rate than predicted, despite great care having been taken during its design.

> Smith came to realize that it was the accumulation of a lot of little defects made during the manufacturing process – not inherent design flaws – that caused the

> high rate of early-life failures. Eliminating the source of those defects was therefore the only way the company could deliver higher quality to its customers.
>
> (Reynard, 2007, p. 23)

Thus appreciation of the relevance of probability calculations such as that in Exercise 4.2 to the need for low failure rates amongst the large number of individual components in complex systems, such as a cellular telephone, is key.

4.2 Probability distributions for counts and measurements

4.2.1 Binomial distribution

Consider a situation where there is constant probability p that an item produced by a process is nonconforming. Let D denote a nonconforming item and S denote a conforming item. Thus we have $P(D) = p$ and $P(S) = 1 - p$ (by Rule 1), which will be denoted by q. Suppose that samples of $n = 2$ items are selected at random from the process output. The sample space consists of the four sequences SS, SD, DS, DD where, for example, the sequence DS represents the outcome that the first item selected was nonconforming and the second was conforming. Rule 3 gives the probabilities of the above four outcomes as qq or q^2, qp, pq and pp or p^2, respectively.

The number of nonconforming items in samples of $n = 2$ items is a count or, more formally, a discrete random variable. It is conventional to use an upper-case letter to denote a random variable and the corresponding lower-case letter to denote a specific value that random variable may take. Let the number of nonconforming items in samples of two items be denoted by X. Thus specific values of X will be denoted by x. Table 4.2 demonstrates the calculation of the probability that a random sample of two items includes precisely x nonconforming items for $x = 0$, 1 and 2. Thus $P(X = 0) = q^2$, $P(X = 1) = 2qp$ (note use of Rule 2) and $P(X = 2) = p^2$. The probability function for a discrete random variable is typically denoted by $f(x)$ and is defined as:

$$P(X = x) = f(x).$$

The probability function in Table 4.3 is for the specific case of samples of $n = 2$ and for probability $p = 0.2$, i.e. where 20% of items are nonconforming.

Exercise 4.3 Tabulate the probability functions for the cases where $n = 2$ and $p = 0.5$ and where $n = 2$ and $p = 0.7$.

Note that in all cases the sum of the probabilities is 1. This must always be the case for a probability function for a discrete random variable. In general the fact that the probabilities

Table 4.2 Derivation of probabilities.

No. of nonconforming items, x, in sample of two	Outcomes yielding x nonconforming items	Probability that sample includes x nonconforming items
0	SS	$qq = q^2$
1	Either SD or DS	$qp + pq = 2qp$
2	DD	$pp = p^2$

Table 4.3 Probability function.

x	$f(x)$
0	0.64
1	0.32
2	0.04

sum to 1 may be demonstrated as follows, use being made of the fact that since $q = 1 - p$ then $q + p = 1$:

$$q^2 + 2qp + p^2 = (q + p)^2 = 1^2 = 1.$$

Since the two-term expression $q + p$ is referred to as a binomial expression in mathematics, this type of discrete probability distribution is referred to as a *binomial distribution*. In order to specify the distribution, the number of items tested or number of trials, n, and the constant probability, p, that an individual item is nonconforming are required. The numbers n and p are referred to as the two *parameters* of the distribution. The short-hand $B(n, p)$ is used for the binomial distribution with parameters n and p and the short-hand $X \sim B(n, p)$ is used to indicate that the random variable X has the specified binomial distribution.

Minitab provides a facility for the calculation of binomial and other widely used probability functions. Having set up and named columns C1 and C2 as indicated in Figure 4.3, **Calc** > **Probability Distributions** > **Binomial...**, with the **Probability** option selected, may be used to reproduce the probabilities in Table 4.3. **Number of trials:** was specified as 2 and **Event probability:** as 0.2. **Input column:** specifies the column, named x, containing the values of the random variable for which probabilities are to be calculated and **Optional storage:** 'f(x)' indicates where the probabilities are to be stored. (Recall that the column named f(x) may be selected for storage by highlighting and left clicking or by typing the column name enclosed in single quotes.) The probability function is displayed as a bar chart in Figure 4.4.

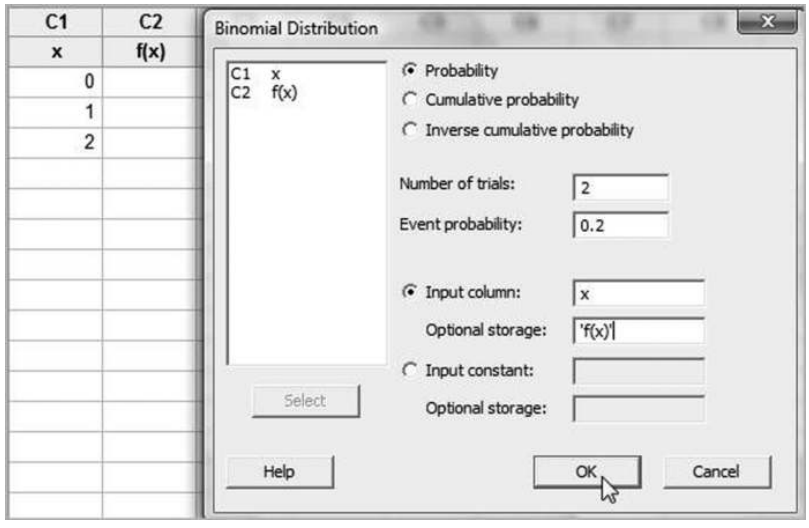

Figure 4.3 Obtaining binomial probabilities.

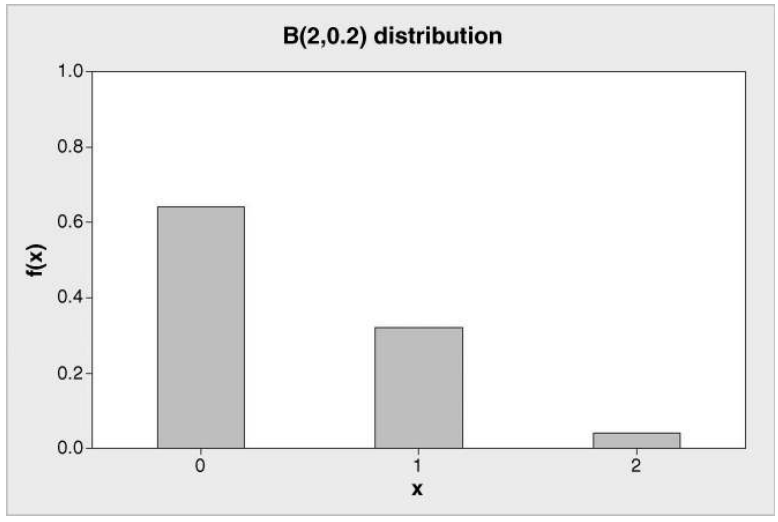

Figure 4.4 Display of $B(2, 0.2)$ probability function.

When the parameter p is less than 0.5 the distribution is positively skewed as in Figure 4.4. When $p = 0.5$ the distribution is symmetrical and when p is greater than 0.5 the distribution is negatively skewed. The display in Figure 4.4 was created using **Graph** > **Bar Chart…** – see the dialog in Figure 4.5. **Bars represent:** Values from a table was selected initially followed by **Simple** under **One column of values**. The column of probabilities f(x) was selected under **Graph variables:** and **Categorical variable:** x for the horizontal axis. **Labels…** was used to create the title B(2, 0.2) distribution. The vertical scale was altered to have maximum value 1, since a probability cannot exceed 1. (The alteration was made by double-clicking on the vertical scale, selecting the **Scale** tab and, under **Scale Range**, unchecking **Auto** for **Maximum** and entering the value 1 in the appropriate window.)

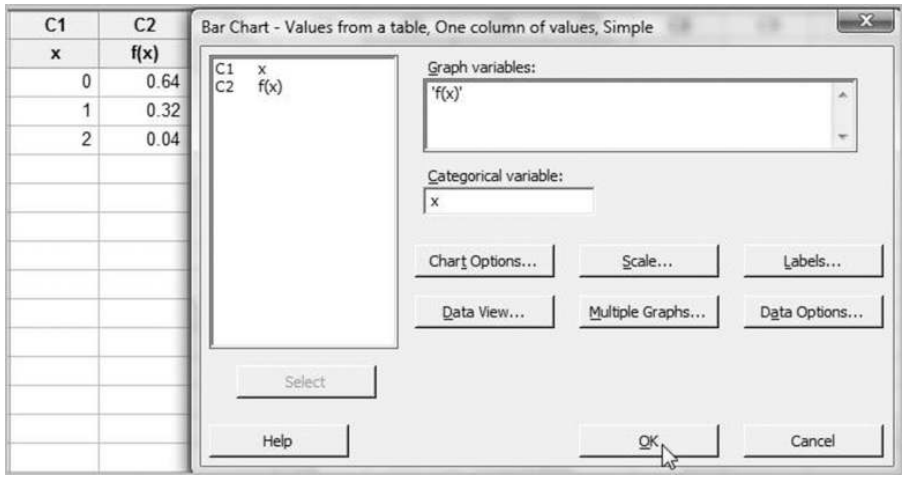

Figure 4.5 Creating display of $B(2, 0.2)$ distribution.

Exercise 4.4 Use Minitab to tabulate and display the probability functions of the $B(2, 0.5)$ and $B(2, 0.7)$ distributions. Compare the values of the probability functions with your answers to Exercise 4.3.

In addition to the probability function, $f(x)$, for a discrete random variable, X, the **cumulative probability function** $F(x)$ is important. This is given by

$$F(x) = P(X \leq x).$$

For the $B(2, 0.2)$ distribution tabulated Table 4.2 we have:

$$\begin{aligned}
F(0) &= P(X \leq 0) = P(X = 0) = f(0) = 0.64, \\
F(1) &= P(X \leq 1) = P(\text{either } X = 0 \text{ or } X = 1) \\
&= P(X = 0) + P(X = 1) = f(0) + f(1) = 0.64 + 0.32 = 0.96, \\
F(2) &= P(X \leq 2) = P(\text{either } X \leq 1 \text{ or } X = 2) \\
&= P(X \leq 1) + P(X = 2) = F(1) + f(2) = 0.96 + 0.04 = 1.
\end{aligned}$$

Both the probability function and the cumulative probability function are tabulated in Table 4.4.

You should check the distribution function, $F(x)$, using Minitab by selecting the **Cumulative probability** option in the dialog box displayed in Figure 4.3. The software does not distinguish between f(x) and F(x) as column names so one could give the name F.(x) to column C3 and then use it for storage of the results. In plain English, $f(1)$ gives the probability that a sample includes *precisely one* nonconforming item whereas $F(1)$ gives the probability that a sample includes *one or fewer* nonconforming items. In general the probability function, $f(x)$, gives the probability that a sample includes *precisely* x nonconforming items, while the cumulative probability function, $F(x)$, gives the probability that a sample includes x *or fewer* nonconforming items.

Exercise 4.5 From the tables created in answering Exercise 4.3 tabulate the cumulative probability function $F(x)$ for the $B(2, 0.5)$ and $B(2, 0.7)$ distributions. Check your answers using Minitab.

Suppose that the process referred to above continues to produce items with constant probability of 0.2 that an item is nonconforming and that it is decided to monitor the process by taking samples of $n = 25$ items at regular intervals. Grass-roots calculation of the probability function in this case would be very tedious indeed. Mathematical formulae are available but will not be introduced here. If required, the probability function and the cumulative probability function can be obtained using Minitab. One would use **Calc** > **Probability Distributions** > **Binomial...** as in Figure 4.3 with the values $0, 1, 2, \ldots, 24, 25$ set up in column C1 and the **Number of trials** specified as 25 in order to tabulate the

Table 4.4 Functions for $B(2, 0.2)$ distribution.

x	$f(x)$	$F(x)$
0	0.64	0.64
1	0.32	0.96
2	0.04	1.00

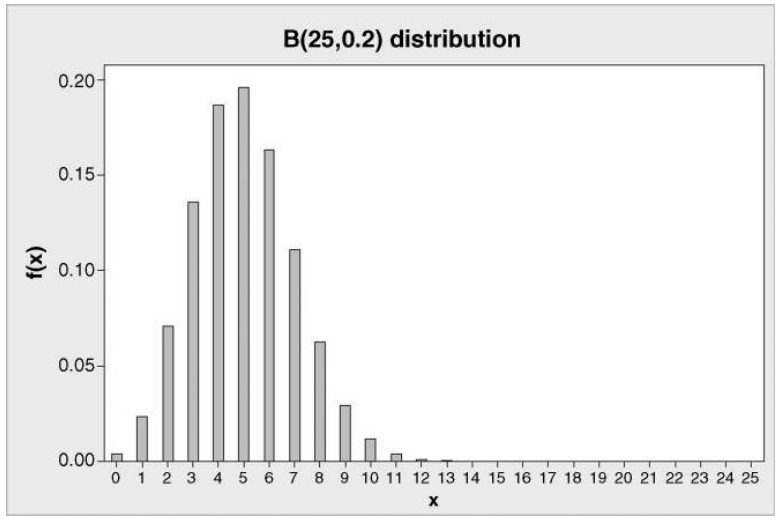

Figure 4.6 Display of $B(25, 0.2)$ distribution.

probability function. This function is displayed in Figure 4.6, with the minimum set to 0 on the vertical scale.

Note that it would be rare to obtain a sample including more than 10 nonconforming items. The most likely number of nonconforming items is 5, and this value is known as the *mode* of the distribution.

What might a run chart of a typical series of counts of numbers of nonconforming items in samples of 25 look like? Minitab enables random data from a wide variety of distributions to be generated. Having assigned No. N-C as the name of column C1, one can use **Calc** > **Random Data** > **Binomial...** as indicated in Figure 4.7. By specifying **Number of**

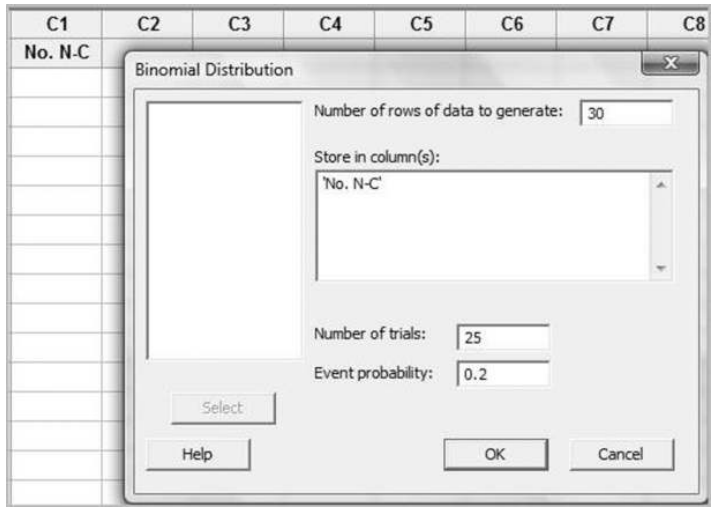

Figure 4.7 Generating data from the $B(25,0.2)$ distribution.

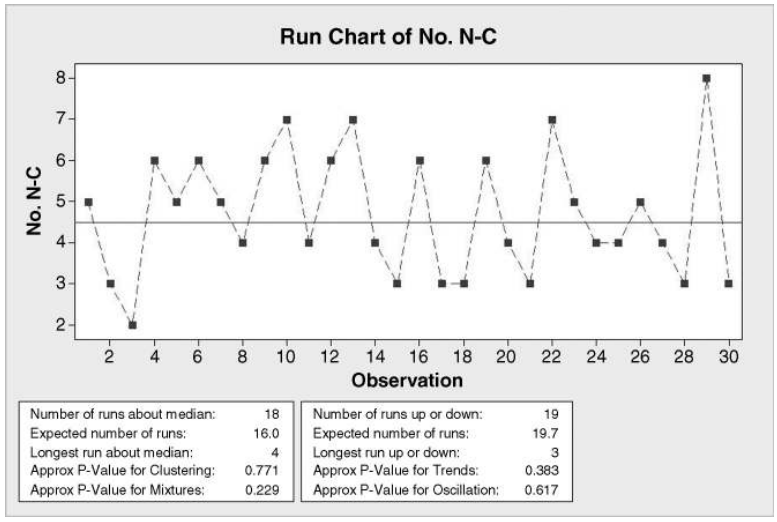

Figure 4.8 Run chart of the data.

rows to generate: 30 we are indicating that we wish to simulate the selection of 30 samples of $n = 25$ items from a population of items in which there is probability $p = 0.2$ that an item is nonconforming. No. N-C is selected for **Store in column(s):** and the distribution parameters $n = 25$ and $p = 0.2$ are specified as before.

A run chart of a data set generated by the author in this manner is displayed in Figure 4.8. Since data have been simulated here for a stable, predictable process with constant probability 0.2 of an item being nonconforming it should be no surprise that none of the *P*-values are less than 0.05 and that therefore there is no evidence of any variation other than common cause variation. The mean and standard deviation of the 30 counts were 4.700 and 1.535, respectively.

Table 4.5 gives the means and standard deviations for longer and longer series of samples of simulated counts for the same stable, predictable process. It is possible to imagine the sampling being continued *ad infinitum*. The resulting conceptual infinity of counts is known as the *population*. The long-term mean as sampling continues is denoted by μ and is called the *expected value* or *population mean* of the random variable X, the number of nonconforming items in samples of 25. The long-term mean of $(X - \mu)^2$ is denoted by σ^2 and is called the

Table 4.5 Means and standard deviations for simulated series of samples.

No. of samples	Mean	Standard deviation
30	4.7000	1.5350
100	5.0500	1.8880
1 000	5.0640	2.0100
10 000	4.9968	2.0067
100 000	4.9996	1.9997

variance of the population; σ is the *standard deviation* of the population. In standard statistical notation:

$$\begin{aligned}\text{Population mean } \mu &= \text{Expected value of } X = E[X];\\ \text{Population variance } \sigma^2 &= \text{Expected value of } (X-\mu)^2 = E[(X-\mu)^2].\end{aligned}$$

Statistical theory provides important results for the $B(n, p)$ distribution:

$$\mu = np, \quad \sigma^2 = npq, \quad \sigma = \sqrt{npq}.$$

Substitution of $n = 25$, $p = 0.2$ and $q = 1 - p = 0.8$ into the above formulae yields 5 and 2 for the population mean and standard deviation, respectively. This is in accord with the sequences of sample means and standard deviations obtained from the simulations and displayed in Table 4.5. The column of means appears to be 'homing in' on the expected value 5 and the column of standard deviations on the expected value 2. The binomial distribution provides the basis for two widely used control charts to be introduced in Chapter 5.

Exercise 4.6 A public utility company believes that currently 10% of billings of private customers are posted out after the scheduled date. Suppose that samples of 250 of the accounts posted each working day are checked for delay in posting. Denote by X the number of accounts in the sample which are posted after the scheduled date.

(i) State the distribution of X and its parameters.

(ii) Calculate the mean and standard deviation of X.

(iii) Simulate data for 50 days, display it in a run chart, obtain the mean and standard deviation of the 50 daily counts and compare with your answers to (ii).

(iv) Repeat (iii) for 500 days and for 5000 days.

(v) Obtain the probability that a sample contains fewer than 15 delayed accounts.

4.2.2 Poisson distribution

Scrutiny of the menu under **Calc > Probability Distributions** reveals a total of 24 distributions available to provide statistical models! Distributions that model count data – in other words, that model discrete random variables – are the binomial, hypergeometric, discrete, integer and Poisson. The hypergeometric distribution has important applications in acceptance sampling and is referred to later in this chapter. From the point of view of quality improvement the two most important discrete distributions are the binomial and Poisson.

The Poisson distribution is named in honour of the French mathematician Siméon Denis Poisson and may be used to model situations where a discrete random variable X may take any of the integer values 0, 1, 2, 3, ..., with no upper limit on the range of possible values. The Poisson distribution has a single parameter, usually denoted by the Greek letter λ (lambda). Both the mean and variance of the distribution are equal to λ. Thus the standard deviation is $\sqrt{\lambda}$.

Table 4.6 Frequency of V2 bomb hits.

No. hits	0	1	2	3	4	5	6	7	≥8
No. squares	229	211	93	35	7	0	0	1	0

The Poisson distribution provides an important model for counts of random events in both time and space. During the Second World War the city of London was attacked by German V2 flying bombs. In order to assess the ability of the Germans to aim at specific targets, British scientists divided the city into a set of 576 squares, each of side 0.5 km. The number of V2 bomb hits per square was counted, yielding the following frequency table displayed in Table 4.6 (data reproduced by permission of the Institute and Faculty of Actuaries from Clarke, 1946, p. 481).

In order to fit the Poisson model to this data set an estimate of the model parameter is required. The Poisson parameter λ is equal to the mean, so the mean number of hits per square will provide an estimate of this parameter. A convenient way to input the data into Minitab is to first name column C1 in a new worksheet No. Hits, make the Session window active, select the **Editor** menu and check **Enable Commands**. The software responds by presenting the prompt MTB > in the Session window. The user is now able to 'drive' using session commands as well as menu commands. The SET command now be used to input the data as indicated in Panel 4.4. Note, for example, that the notation 93(2) indicates that 93 values of 2 for the number of hits were obtained. The single value of 7 for the number of hits could have been indicated by 1(7) but the solitary 7 is sufficient.

```
MTB > set c1
DATA> 229(0) 211(1) 93(2) 35(3) 7(4) 7
DATA> end
MTB > Tally 'No. Hits';
SUBC>   Counts.

Tally for Discrete Variables: No. Hits

 No.
Hits  Count
   0    229
   1    211
   2     93
   3     35
   4      7
   7      1
  N=    576

MTB > Describe 'No. Hits';
SUBC>   Mean;
SUBC>   Count.

Descriptive Statistics: No. Hits

          Total
Variable  Count    Mean
No. Hits    576  0.9323
```

Panel 4.4 Input and analysis of V2 bomb data.

Table 4.7 Observed and expected frequency of V2 bomb hits.

No. Hits	0	1	2	3	4	5	6	7	≥8
Observed No. Squares	229	211	93	35	7	0	0	1	0
Expected No. Squares	226.7	211.4	98.5	30.6	7.1	1.3	0.2	0.0	0.0

Having input the data, the author then used the **Stat** menu and **Stat** > **Tables** > **Tally Individual Values…** to check the input. The session commands corresponding to the menu actions appear in the Session window. Finally, **Basic Statistics** > **Display Descriptive Statistics…** from the **Stat** menu was used, with **Mean** and **N total** selected via **Statistics…**, to obtain the mean number of hits.

The mean of 0.9323 can be used to compute Poisson probabilities via **Calc** > **Probability Distributions** > **Poisson…**. The Poisson model gives the probability of no hits in a square to be 0.393 647, so the expected frequency of squares with no hits is $576 \times 0.393\,647 = 226.7$ correct to one decimal place. This is very close to the observed frequency of squares with no hits, 229. The expected frequencies for 1, 2, 3, 4, 5, 6 and 7 hits per square may be calculated in a similar fashion. Use of the **Cumulative probability** facility with 7 entered in the **Input constant:** box yields $F(7) = P(X \leq 7) = 0.999\,99$ so $P(X \geq 8) = 1 - 0.999\,99 = 0.000\,01$. Hence the expected frequency of squares with 8 or more hits is $576 \times 0.000\,01 = 0.0$ correct to one decimal place. Table 4.7 gives the summarized results.

The bar chart in Figure 4.9 highlights how well the Poisson distribution models the situation. (A follow-up exercise will indicate how such charts may be created using Minitab.) The good fit of the Poisson distribution with parameter 0.9323, i.e. of the P(0.9323) distribution in shorthand, indicated that the V2 impacts were occurring at random locations within the city and thus provided evidence to the scientists that the flying bombs were not equipped with a sophisticated guidance system.

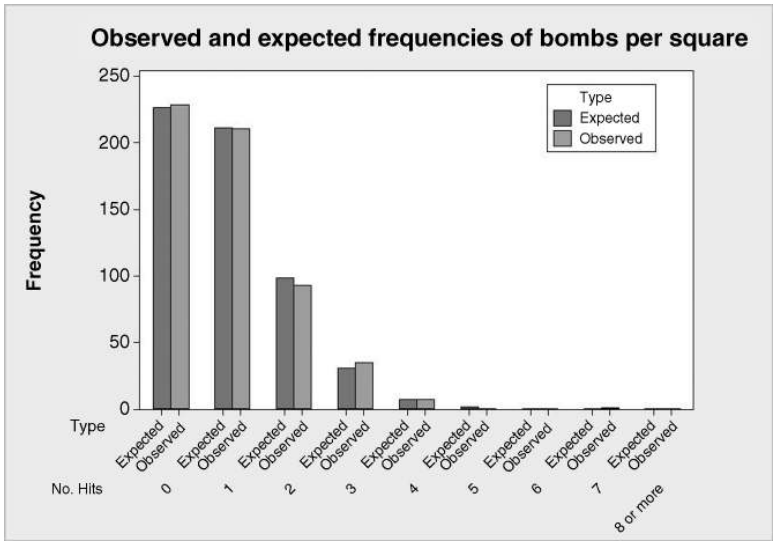

Figure 4.9 Comparison of observed and expected frequencies.

The Poisson distribution often provides a statistical model for counts of events occurring at random in either space or time, e.g. the number of nonconformities on a printed circuit board or the number of line stoppages per month in a factory. As with the binomial distribution, it provides the basis for two very important types of control chart to be introduced in Chapter 5.

4.2.3 Normal (Gaussian) Distribution

The normal distribution is central in the application of statistical methods in quality improvement and in understanding the concept of sigma quality level within Six Sigma programmes. The name 'normal' is unfortunate in that it suggests that such distributions are to be expected as some sort of norm. Some authors refer to it as the Gaussian distribution in honour of the German mathematician, Karl Friedrich Gauss. The normal distribution provides a model for continuous random variables. To introduce the normal distribution use will be made of the bottle weight data that is available in Weight1A.MTW and was displayed in Figure 2.22 in Chapter 2.

In the histogram in Figure 4.10 a different set of bins was used from that used in Figure 2.22. The bins for the histogram in Figure 4.10 were bounded by the values 483.0, 485.0, 487.0, 489.0, 491.0, 493.0 and 495.0. The first bin range from 483.0 to 485.0 has a range of 2 and there was a single bottle from the sample of 25 with weight in this range. The relative frequency of weight in this range for the sample was therefore $1/25 = 0.04$. This relative frequency provides an estimate of the probability that a bottle from the population of bottles sampled has weight in the corresponding bin range. Thus a probability of 0.04 spread over a range of 2 is equivalent to a probability of 0.02 spread over a range of 1; the *probability density* corresponding to the first bin is $0.04/2 = 0.02$. You should verify that the probability densities corresponding to the remaining bins are 0.08, 0.14, 0.16, 0.04 and 0.06. Minitab offers the option of displaying a density histogram of the weight data, in which the height of the bars corresponds to probability density rather than frequency. This histogram is shown in

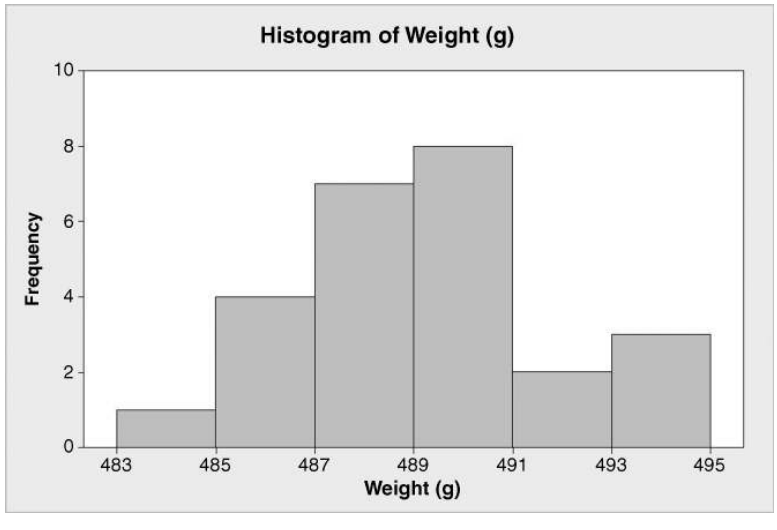

Figure 4.10 Frequency histogram of bottle weights.

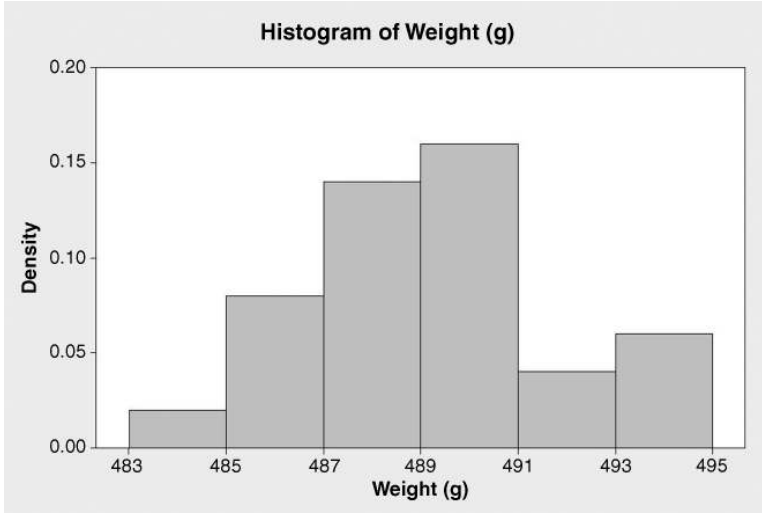

Figure 4.11 Density histogram of bottle weights.

Figure 4.11. Note that the vertical scale now represents density as opposed to frequency. Some properties of the density histogram are given in Table 4.8. The areas of the bars were obtained by multiplying width by height. It should be borne in mind that the histograms are not drawn to scale!

If the probability that a bottle weight is less than or equal to, say, 487 g is required then it may be estimated from the data via the histogram as follows:

$$
\begin{aligned}
&P(\text{Weight} \leq 487)\\
&\quad = P(\text{Either weight lies in bin } 483,\ 485 \text{ or weight lies in bin } 485,\ 487)\\
&\quad = P(\text{Weight lies in bin } 483,\ 485) + P(\text{Weight lies in bin } 485,\ 487)\\
&\quad = 0.04 + 0.16\\
&\quad = 0.20.
\end{aligned}
$$

Thus the estimated probability that a bottle weight is less than or equal to 487 g is the shaded area indicated in the density histogram in Figure 4.12.

Table 4.8 Properties of the density histogram of weight.

Weight bin range	Histogram bar width	Histogram bar height	Histogram bar area
483–485	2	0.02	0.04
485–487	2	0.08	0.16
487–489	2	0.14	0.28
489–491	2	0.16	0.32
491–493	2	0.04	0.08
493–495	2	0.06	0.12
	Total area of histogram		1.00

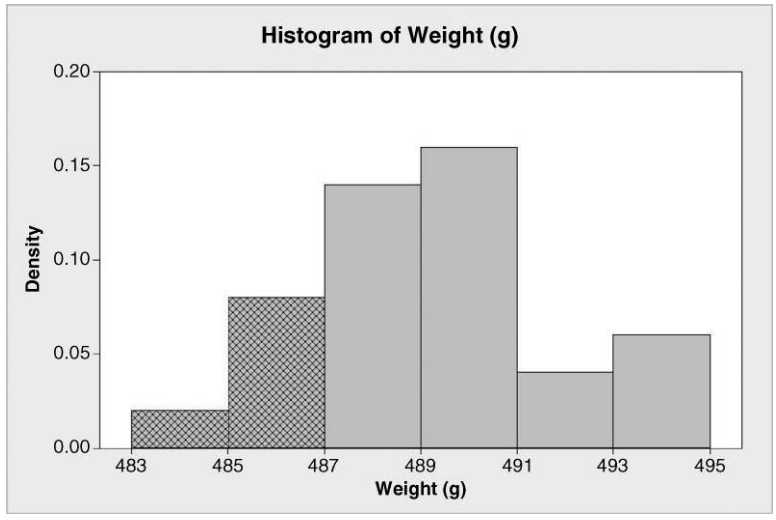

Figure 4.12 Shaded area gives $P(\text{Weight} \leq 487)$.

A further option available for histograms in Minitab is the fitting of a normal distribution. The curve in Figure 4.13 is the probability density function of the fitted normal distribution. The parameters of the fitted distribution are the sample mean and sample variance. It appears that the normal distribution provides a reasonable model for the random variable bottle weight.

The probability density function $f(x)$ for a continuous random variable, X, can never be less than 0 and is such that the total area it encloses with the horizontal axis is 1. This value of 1 represents total probability. The cumulative probability function $F(x)$ gives $P(X \leq x)$. A

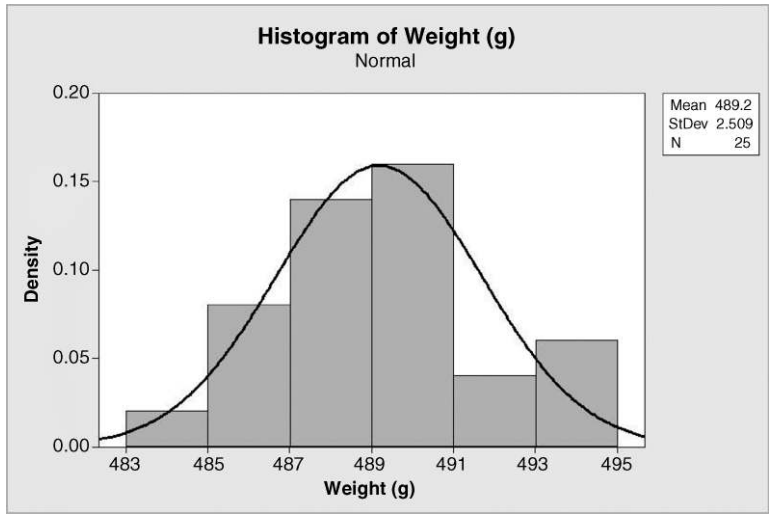

Figure 4.13 Normal distribution fitted to bottle weight data.

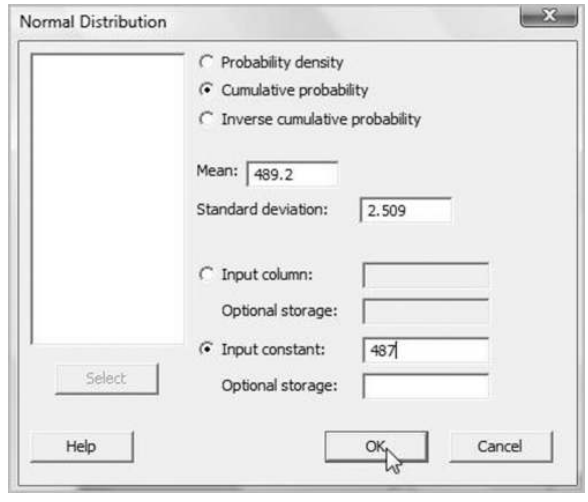

Figure 4.14 Evaluation of cumulative probability function for a normal distribution.

normal distribution is specified by two parameters: the mean and the variance. The normal distribution with mean 0 and variance 1 is referred to as the *standard normal distribution*. Because of the central role in statistics of the standard normal distribution its probability density function and its cumulative probability function are denoted by the special functions $\phi(x)$ and $\Phi(x)$, respectively. The letters ϕ and Φ are the lower- and upper-case Greek letters phi. Tables of the function $\Phi(x)$ are available but with Minitab you will have no need for them.

The sample mean and sample variance for the bottle weight data are 489.2 and 2.509^2 respectively, and these are the parameters used for the fitted normal distribution shown in Figure 4.13. In shorthand this distribution is denoted by $N(489.2, 2.509^2)$. (As with the binomial and Poisson distributions, the letter in front of the brackets indicates the distribution type and the numbers within the brackets are the parameters.) In order to obtain from the model the probability $P(\text{Weight} \leq 487) = F(487)$, the cumulative probability function has to be evaluated for 487. This may be obtained using **Calc** > **Probability Distributions** > **Normal...**. Note that the cumulative probability function and the standard normal distribution are the defaults. The value of interest, 487, is entered in the **Input constant:** window. The appropriate mean and standard deviation must be specified in the **Mean:** and **Standard deviation:** windows respectively. The dialog is completed as shown in Figure 4.14.

The Session window output is shown Panel 4.5. This indicates that the fitted normal distribution model gives $F(487) = 0.190\,286$, i.e. the probability of a bottle weight of 487 g or less is 0.19 to two decimal places. This agrees quite closely with the empirical probability of 0.20 represented by the shaded area in Figure 4.13. In fact the evaluation of the cumulative probability function involves calculation of the area under the probability density function lying to the left of 487. This area is shaded in Figure 4.15.

In Chapter 2 specification limits of 485 and 495 g for bottle weight were indicated. The fitted model may be used to estimate the proportion of bottles meeting those requirements. You should verify, using dialogs similar to that in Figure 4.14, that $F(485) = 0.0471$ and that

Cumulative Distribution Function

```
Normal with mean = 489.2 and standard deviation = 2.509

  x  P( X <= x )
487     0.190286
```

Panel 4.5 Evaluation of cumulative probability function for a normal distribution.

$F(495) = 0.9896$ to four decimal places. The probability that a bottle weight lies between 485 and 495 is therefore $0.9896 - 0.0471 = 0.9425$.

This result may also be obtained using **Graph** > **Probability Distribution Plot...** > **View Probability**. On double-clicking the graphic under **View Probability**, select Normal in the **Distribution:** window and enter **Mean:** 489.2 and **Standard deviation:** 2.509. On the **Shaded Area** tab, with **Define Shaded Area By X Value** selected, click on the graphic under **Middle** and enter the lower and upper specification limits 485 and 495 for bottle weight in the **X_value 1:** and **X_value 2:** boxes, respectively. The graph in Figure 4.16 results, confirming the required probability as 0.9425.

Thus the model predicts that 94.25% of bottles from the process would conform to requirements on weight. The proportion nonconforming is therefore estimated to be 6.75% or 67 500 per million. Reference to Appendix 1 indicates the sigma quality level to be approximately 3. (The sample size of 25 is small, so in practice estimates based on such samples should be viewed with some caution.)

Earlier it was stated that the normal distribution appears to provide a reasonable model for the random variable bottle weight in view of the manner in which the curve (the model) in Figure 4.13 fitted the data (the histogram). In order to make a formal assessment whether or not a normal distribution provides a satisfactory model, the normality test provided in Minitab may be used. The points in the associated plot should be reasonably linear and

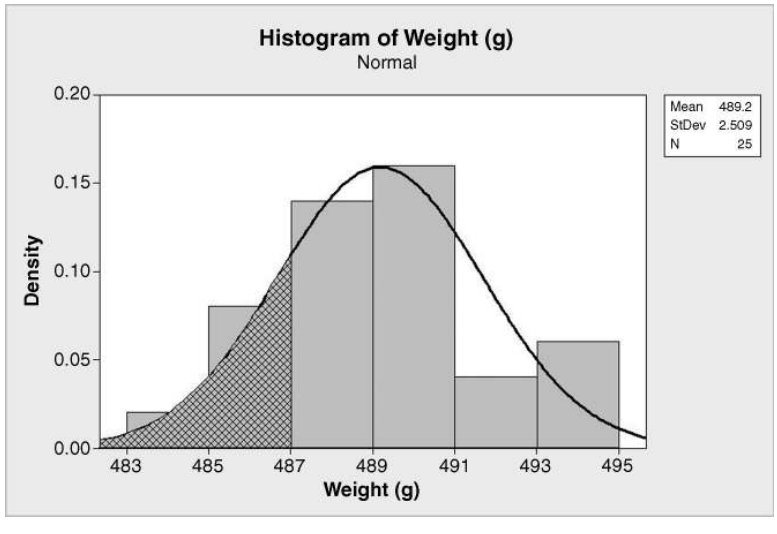

Figure 4.15 Area representing cumulative probability function.

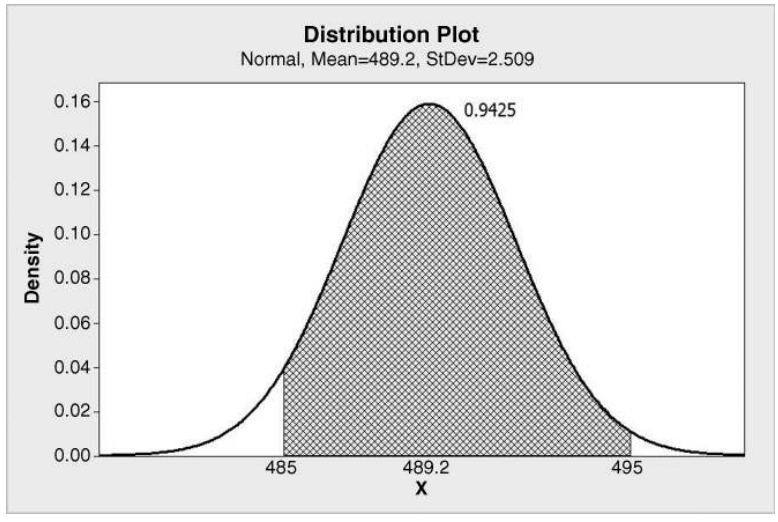

Figure 4.16 Area representing proportion of bottles conforming to weight specifications.

a *P*-value in excess of 0.05 is usually taken to mean that a normal distribution may be accepted as a satisfactory model. The normality test is available via **Stat** > **Basic Statistics** > **Normality Test...**. The output obtained with the default options is shown in Figure 4.17. With *P*-value well in excess of 0.05 the normal distribution model is clearly acceptable. (The AD value quoted is the Anderson–Darling test statistic. Test statistics will be explained in Chapter 7.)

Consider now the conduct of interviews by a researcher undertaking a customer satisfaction survey. Suppose that the duration, in minutes, of interviews can be adequately modelled by the $N(40, 8^2)$ distribution, i.e. by the normal or Gaussian distribution with mean 40 and

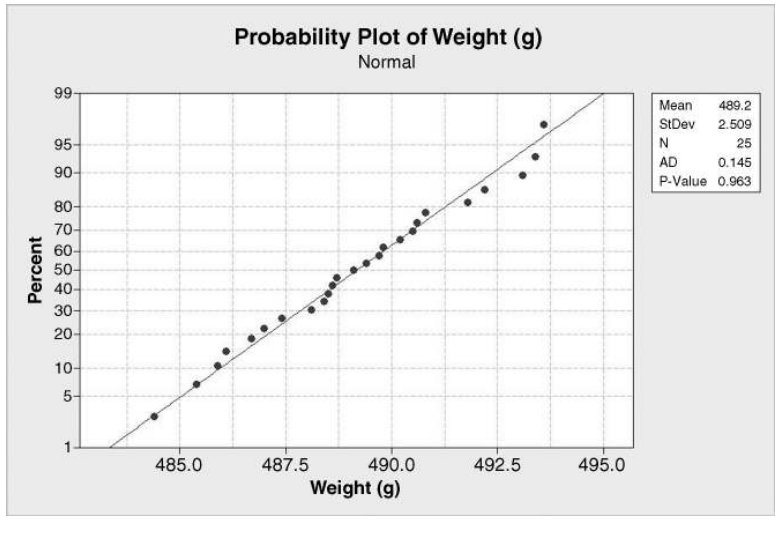

Figure 4.17 Normality test of bottle weight data.

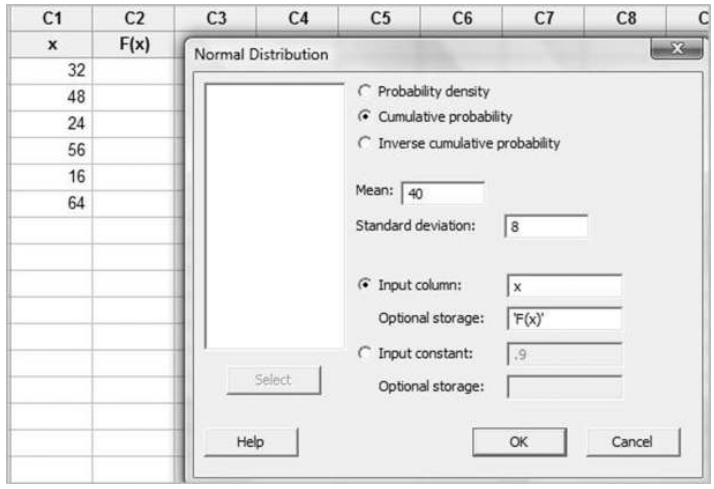

Figure 4.18 Evaluation of a series of normal cumulative probability function values.

variance 8^2, standard deviation 8. Minitab can be used to obtain the probabilities that interviews last between 32 and 48 minutes, between 24 and 56 minutes, and between 16 and 64 minutes.

Here the cumulative probability function is required for six different values of the random variable of interest, the duration of an interview. These six values can be entered into a column in Minitab and the corresponding cumulative probabilities stored in a second column, named F(x) in advance of the calculations being performed. The dialog is displayed in Figure 4.18.

The probability that duration lies between 32 and 48 minutes is

$$\begin{aligned} P(32 < X \leq 48) &= F(48) - F(32) \\ &= 0.841\,345 - 0.158\,655 \text{ (from the F(x) column in the worksheet)} \\ &= 0.6827 \text{ to four decimal places,} \end{aligned}$$

or about two-thirds. The durations of 32 and 48 are one standard deviation below and above the mean respectively. The calculation demonstrates the feature of the normal distribution that approximately two thirds of observed values are within one standard deviation of the mean. You should verify that that approximately 95% of values are within two standard deviations of the mean, i.e. between 24 and 56 (probability 0.9545), and that 99.73% of values are within three standard deviations of the mean, i.e. between 16 and 64 minutes. Thus it would be very rare (probability 0.0027) for an interview to have duration outside the range 16 minutes to 64 minutes, i.e. outside the range $\mu - 3\sigma$ and $\mu + 3\sigma$. The normal distribution properties illustrated by this example are displayed in Figure 4.19.

Suppose the interviewer wishes to know the duration which would be exceeded for one interview in ten in the long run. In other words, the value d is required such that $P(X \leq d) = 9/10 = 0.9$. Duration d may be referred to as the 90th percentile of the distribution. In order to obtain d, use **Calc** > **Probability Distributions** > **Normal...**, select **Inverse cumulative probability** and enter 0.9 in the **Input constant:** window. The appropriate mean and standard deviation must be specified in the **Mean:** and **Standard deviation:** windows, respectively. On execution the Session window displays the text in

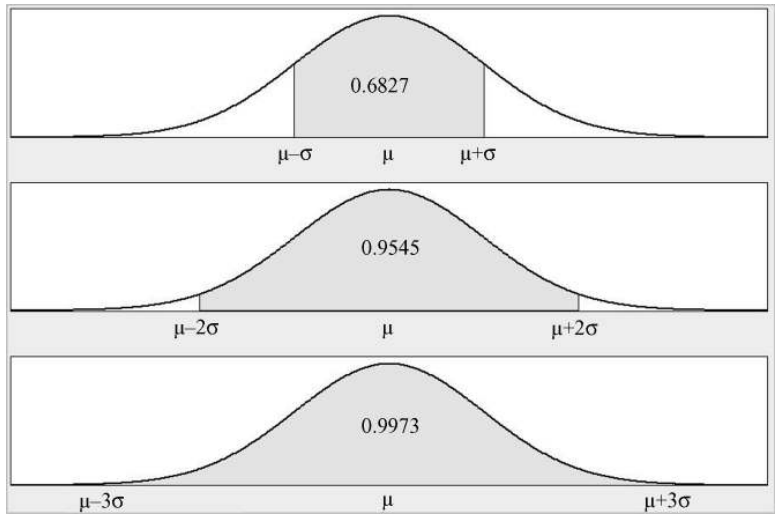

Figure 4.19 Probabilities of values lying within 1, 2 and 3 standard deviations of the mean for a normal distribution.

Panel 4.6. This indicates that d is 50.25 and that approximately one interview in ten would have duration in excess of 50 minutes.

Alternatively the result may be obtained using **Graph > Probability Distribution Plot... > View Probability**. On double-clicking the graphic under **View Probability** select Normal in the **Distribution:** window and enter **Mean:** 40 and **Standard deviation:** 8. On the **Shaded Area** tab, with **Define Shaded Area By Probability** selected, click on the graphic under **Right Tail** and enter **Probability:** 0.1. The graph in Figure 4.20 results, confirming the required duration d to be 50.25 minutes.

In Section 1.1 the concept of sigma quality level was introduced. Minitab can be used to compute the entries in Table 1.1 and in Appendix 1. In order to demonstrate the calculations involved, consider a bottle manufacturing process which is producing bottles with weights (g) which are $N(493, 2^2)$ and for which the specification limits are 486 and 494 g. The target weight can be considered to be 490 g, the weight that is midway between the specification limits. Thus the specification limits here are two standard deviations away from the target and the process is off target, on the high side, by 3 g – equivalent to 1.5 standard deviations. The situation is illustrated in Figure 4.21,created using the **Graph > Probability Distribution Plot... > View Probability** facility, as described earlier for the creation of Figure 4.16, and Graph Annotation Tools.

Inverse Cumulative Distribution Function

```
Normal with mean = 40 and standard deviation = 8

P( X <= x )        x
        0.9  50.2524
```

Panel 4.6 Evaluation of inverse cumulative probability function.

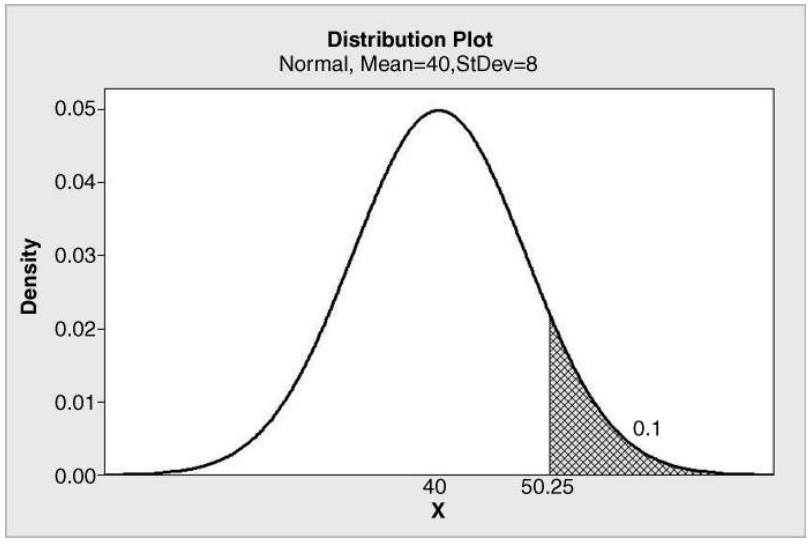

Figure 4.20 Interview duration exceeded on one in ten occasions.

It is evident from the diagram that a small proportion of bottles will be nonconforming due to having weight below the lower specification limit, while a large proportion will be nonconforming due to having weight above the upper specification limit. These proportions are $F(486)$ and $1 - F(494)$ respectively, which you should verify to be 0.000233 and 0.308538 respectively. This gives a total proportion nonconforming of $0.000\,233 + 0.308\,538 = 0.308\,771$ which equates to $0.308\,771 \times 1\,000\,000 = 308\,771$ nonconforming bottles per million. You should confirm that this is the entry in Table 1.1 corresponding to a sigma quality level of 2

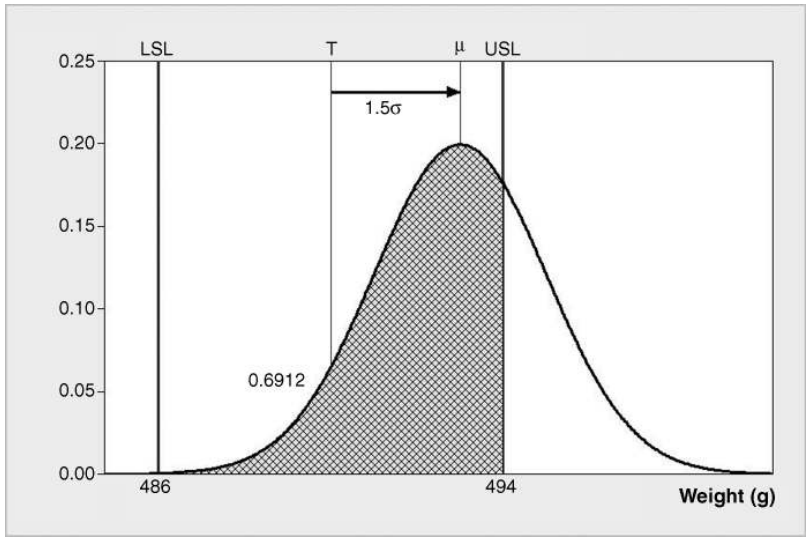

Figure 4.21 Bottle weight distribution with specification limits superimposed.

(apart from a small rounding discrepancy). Note that in Figure 4.21 has displayed probability 0.6912 for conformance corresponding to probability $1 - 0.6912 = 0.3088$ for nonconformance, confirming the above result on rounding.

Because the specification limits in the scenario discussed here are two standard deviations away from the target the process is said to have a sigma quality level of 2. The convention in Six Sigma is to quote the nonconformities per million opportunities when the process is operating with a mean that is 'off target' by 1.5 standard deviations. The use of a 1.5σ 'shift' in the calculation of sigma quality levels is controversial (Ryan, 2000, p. 522). As an exercise you should verify that the proportion is the same when the mean is off target, on the low side, by 1.5 standard deviations. A series of calculations of this type can be used to complete Table 1.1. A more comprehensive table is given in Appendix 1.

Figure 4.22 illustrates a 'Six Sigma' process, i.e. one for which the specification limits are six standard deviations away from the target. As illustrated by the solid curve, the process is operating 'on target'. Clearly with such a process the location could shift by an appreciable amount without any real impact on the proportion of product falling outside the specifications. For this process, a shift of 1.5 standard deviations from the target would lead to 3.4 nonconforming items per million. The dotted curve illustrates the distribution that would occur were the mean to shift upwards by 1.5 standard deviations.

Finally, consider again the $N(40, 8^2)$ distribution used to model the duration, X (minutes), of interviews. You should verify, using Minitab, that $P(X < 0) = 0.000\,000\,3$. Thus, according to the model, there is a very small probability that the duration of an interview could be negative, which is impossible. The probability density function of a normal distribution is defined for all values of the random variable being modelled, whether the values are feasible or not. Thus the model is 'wrong' in the sense that it allows the possibility of an impossible event. Yet the model could prove valuable to the team planning the survey and evaluating the performance of interviewers. Hence the comment quoted at the very beginning of the chapter is relevant.

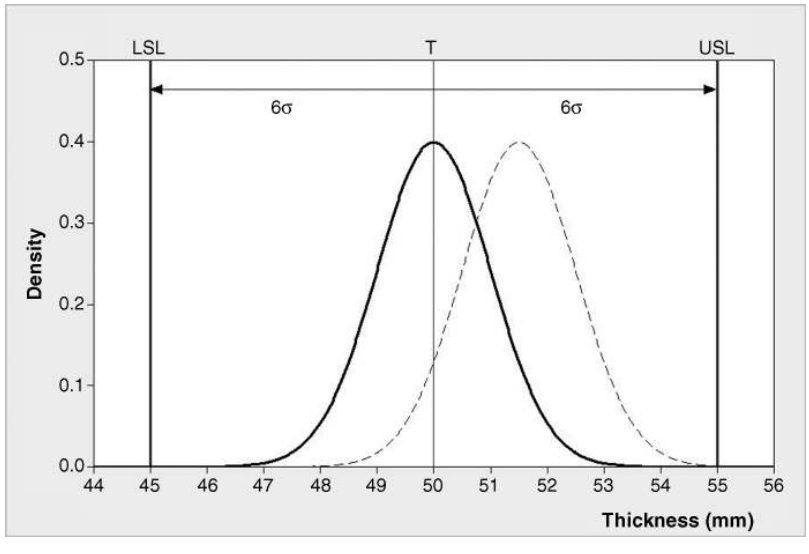

Figure 4.22 A 'Six Sigma' process operating on target.

Suppose that X is a random variable and that a second random variable, V, is defined as $V = kX$, where k is a positive constant. Two important results are:

$$\mu_V = k\mu_X, \quad \sigma_V = k\sigma_X.$$

Thus if k is thought of as a scale factor for converting values of X to values of V then that same scale factor must be used to convert the mean and standard deviation for X into those for V.

If X is normally distributed, then so is V.

Box 4.1 Multiple of a random variable.

4.3 Distribution of means and proportions

4.3.1 Two preliminary results

There are a number of results required for the development of control charts and of other statistical methods of value in quality improvement. These results will be introduced in this section, but theory and mathematical detail will be avoided. Two preliminary results are required.

The first preliminary result (Box 4.1) concerns *multiples of a random variable*. Informal justification of the numerical aspects may be obtained by converting the height data in inches, given in the Minitab pulse data set referred to earlier, to metres. This may readily be done using **Calc** > **Calculator...** to multiply the given heights in inches by the conversion factor 0.0254. The means and standard deviations of the two sets of height measurements are displayed in Panel 4.7. The reader is invited to verify that multiplication of the mean and standard deviation of the heights in inches by 0.0254 yields the mean and standard deviation of the heights in metres.

The second preliminary result (Box 4.2) concerns *sums of independent random variables*. If two random variables are independent then knowledge of the value of one does not yield any information about the value of the other. Independent random variables have zero correlation. However, the converse is not true. In words, the results summarized in Box 4.2 state that:

- the mean of the sum of a set of independent random variables is the sum of the their means;
- the variance of the sum of a set of independent random variables is the sum of their variances.

Descriptive Statistics: Height, Height (m)

Variable	Mean	StDev
Height	68.717	3.659
Height (m)	1.7454	0.0929

Panel 4.7 Descriptive statistics for height measurements.

Let X_1, X_2, X_3, ..., X_p be a set of independent random variables and let $T = X_1 + X_2 + X_3 + \ldots + X_p$, i.e. T is the sum of the set of random variables. Let $X_1, X_2, X_3, \ldots, X_p$ have means $\mu_1, \mu_2, \mu_3, \ldots, \mu_p$ and variances $\sigma_1^2, \sigma_2^2, \sigma_3^2, \ldots, \sigma_p^2$ respectively. Then the mean and variance of T are respectively

$$\mu_T = \mu_1 + \mu_2 + \mu_3 + \ldots + \mu_p,$$
$$\sigma_T^2 = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 \ldots + \sigma_p^2.$$

If the Xs are normally distributed, then T is normally distributed.

Box 4.2 Sum of independent random variables.

Consider an assembly operation for automatic telling machines with three phases: set-up, build and test. Let the durations of each phase have means 15, 90 and 18 minutes, respectively. Suppose that the durations are independent of each other and that all three have uniform distributions with ranges of 6, 12 and 6 minutes respectively. This means that set-up is equally likely to take any time from 12 to 18 minutes, build any time from 84 to 96 minutes and test any time from 15 to 21 minutes. For a uniform distribution the probability density function is constant over the range of possible values and zero elsewhere. Statistical theory tells us that the variance of a uniform distribution is one-twelfth of the square of its range. Hence the variances for the durations of set-up, build and test will be 3, 12 and 3 minutes respectively.

According to the theory in Box 4.2, the mean and variance of the sum of a set of independent random variables are obtained by summing the individual means and the individual variances, respectively. Thus theory predicts that the total duration for the operation will have mean $15 + 90 + 18 = 123$ and variance $3 + 12 + 3 = 18$, which corresponds to a standard deviation of 4.24 minutes.

Minitab enables samples to be generated from uniform distributions so the theory can be checked by simulation of a series of assembly operations. Having named columns C1, C2, C3 and C4 respectively Set-up, Build, Test and Total, one can proceed to simulate data for 10 000 assembly operations as indicated in Figure 4.23. Use is required, three times, of **Calc** > **Random Data** > **Uniform...**. Figure 4.23 shows the procedure about to be implemented for the third time in order to generate 10 000 values for the duration of the Test phase. Note the specification **Lower endpoint:** 15 and **Upper endpoint:** 21. The Total duration can then be computed using **Calc** > **Row Statistics...**, with **Sum** selected as **Statistic**, Set-up, Build and Test entered in the **Input variables:** window and Total entered in the **Store result in:** window.

The descriptive statistics displayed in Panel 4.8 were obtained using **Stat** > **Basic Statistics** > **Display Descriptive Statistics...**. Under **Statistics...** only **Mean**, **Standard deviation** and **Variance** were checked. The mean and standard deviation of the sample of 10 000 values of Total were 122.90 and 4.23 for the simulation carried out by the author. These are close to the values of 123.00 and 4.24 for the population mean and standard deviation given by the results on sums of random variables. The reader is invited to carry out the simulation for her/himself.

The histograms of the simulated data in Figure 4.24 illustrate the uniform distribution of the component times and also that the distribution of the Total duration has the appearance of a

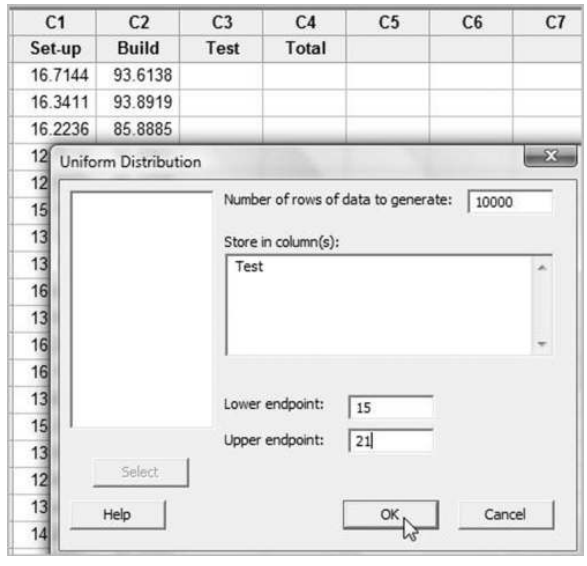

Figure 4.23 Generating data from a uniform distribution.

Descriptive Statistics: Set-up, Build, Test, Total

Variable	Mean	StDev	Variance
Set-up	14.993	1.746	3.048
Build	89.913	3.444	11.859
Test	17.989	1.731	2.996
Total	122.90	4.23	17.86

Panel 4.8 Descriptive statistics for simulated data.

normal distribution – a fitted normal distribution curve is shown. This gives an indication of the importance of the normal distribution – the sum of a series of independent random variables with nonnormal distributions can often be adequately modelled by a normal distribution. If each of the independent random variables is normally distributed then the sum of the random variables is also normally distributed.

4.3.2 Distribution of the sample mean

The sample mean is widely used in control charting and other statistical methods of value in Six Sigma quality improvement. Box 4.3 summarizes the important results required for the sample mean. These results may be obtained by considering the mean of a random sample as a multiple of a sum of independent random variables.

If the reader finds the mathematics a bit daunting then perhaps some further Minitab simulation will aid understanding. A set of 1000 random samples from the $N(60, 2^2)$

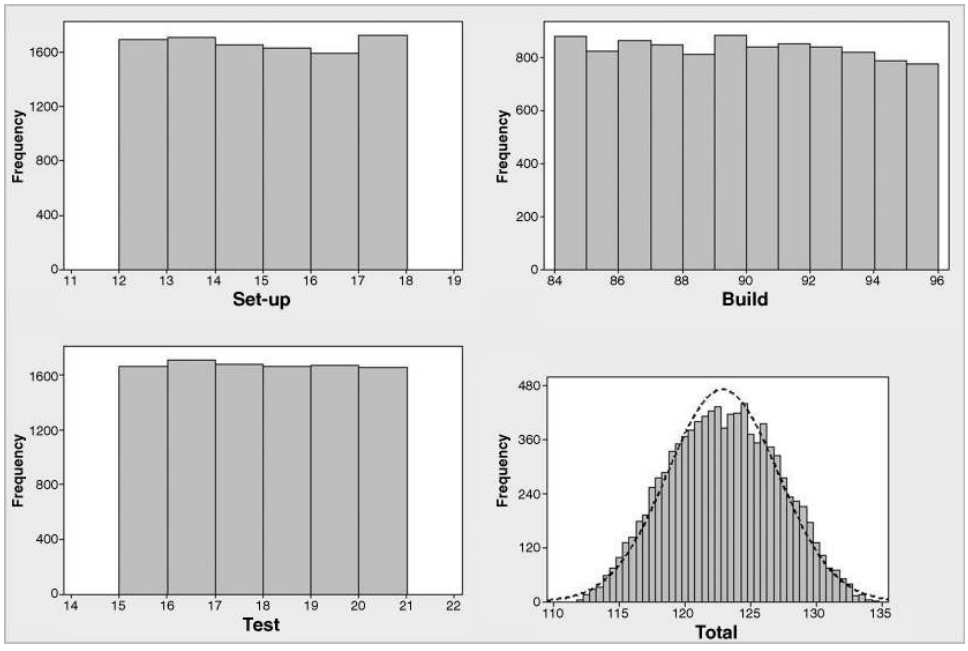

Figure 4.24 Histograms of simulated data.

distribution may be simulated using **Calc > Random Data > Normal...** to create four columns of 1000 values from the specified distribution. Each row of four values may then be considered to be a sample of $n = 4$ values from the $N(60, 2^2)$ distribution. **Calc > Row Statistics...**, with **Mean** as the selected **Statistic,** may then be used to compute the 1000 sample means and to store them in a column named Mean. The dialog involved in this last step is displayed in Figure 4.25.

The theory in Box 4.3 indicates that the mean of the population of sample means will be $\mu_{\bar{X}} = \mu = 60$ and that the standard deviation of the population of sample means will be $\sigma_{\bar{X}} = \sigma/\sqrt{n} = 2/\sqrt{4} = 1$. The theory also indicates that the distribution of the sample mean will be normal.

The histogram of the sample means, with fitted normal curve, shown in Figure 4.26 supports the theory. Note that the mean of the sample means of 60.03 and that the standard deviation of the sample means of 1.007, reported in the text box at the top right-hand corner of the display, are both close to the population values given by theory of 60.00 and 1.000, respectively. Again the reader is invited to carry out the simulation for her/himself.

Box 4.4 gives a very important result, the *central limit theorem*, concerning the distribution of the sample mean when the random variable of interest is *not* normally distributed. As an illustration of the central limit theorem, consider a weaving process that operates continuously and for which filament breaks occur at random at the rate of 2 per hour. This means that the number of breaks occurring per hour will have the Poisson distribution with parameter 2. Statistical theory shows that it follows that the time interval, in minutes, between filament breaks will have the exponential distribution with mean 30. (No details of the

Consider a random sample $X_1, X_2, X_3, \ldots, X_n$ of a random variable X with mean μ and standard deviation σ. The sample mean $\bar{X}$ is also a random variable and is given by:

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1}{n}(X_1 + X_2 + X_3 + \ldots + X_n) = \frac{1}{n}T$$

Thus the sample mean can be considered to be a multiple (scale factor $1/n$) of the total $T = X_1 + X_2 + X_3 + \ldots + X_n$ of n independent random variables.

Application of the two preliminary results yields the mean mean and standard deviation of the sample mean as follows:

$$\begin{aligned}
\mu_{\bar{X}} &= \frac{1}{n}(\mu + \mu + \mu + \ldots + \mu) = \frac{1}{n}n\mu = \mu, \\
\sigma_{\bar{X}} &= \frac{1}{n} \times \sigma_T = \frac{1}{n}\sqrt{\sigma_T^2} \\
&= \frac{1}{n} \times \sqrt{\sigma^2 + \sigma^2 + \sigma^2 + \ldots + \sigma^2} \\
&= \frac{1}{n} \times \sqrt{n\sigma^2} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}.
\end{aligned}$$

Thus the mean and standard deviation of the sample mean are respectively μ and $\sigma/\sqrt{n}$.

Many authors use the phrase 'standard error' in place of standard deviation in this context.

Box 4.3 Distribution of the sample mean.

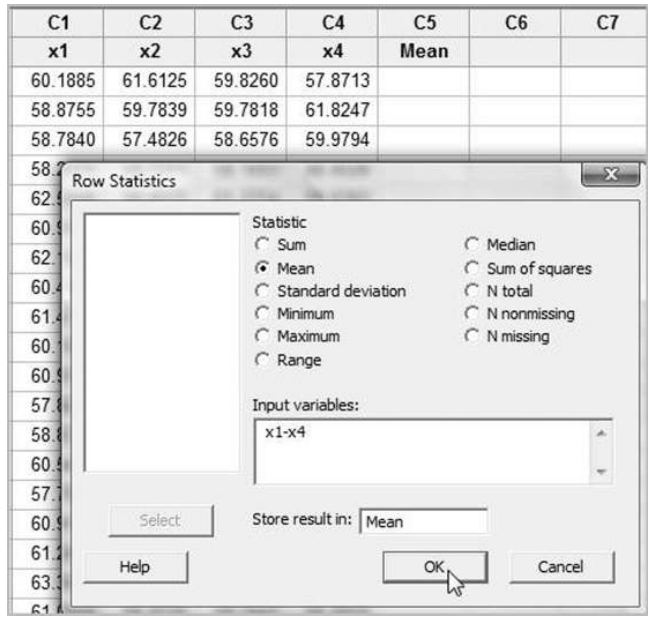

Figure 4.25 Calculation of the sample means.

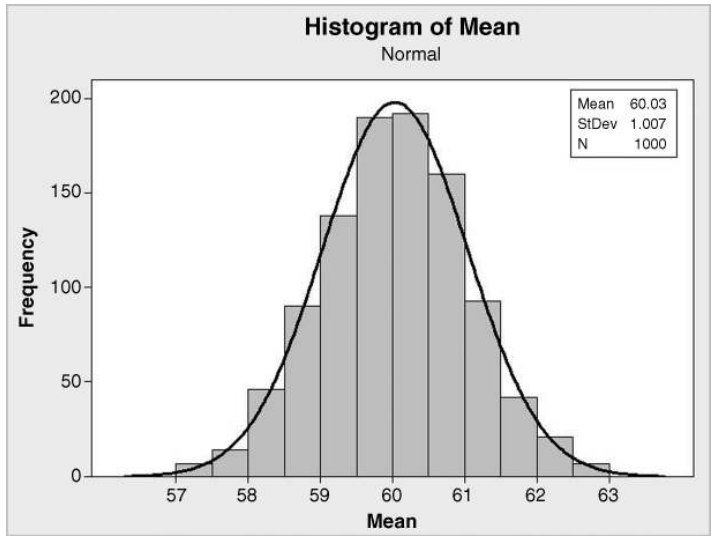

Figure 4.26 Histogram of sample means with fitted normal curve.

exponential distribution are provided in this book.) Scrutiny of log sheets maintained by the process operators yielded 1000 time intervals between filament breaks that are plotted in Figure 4.27 in the form of a density histogram with the probability density function of the exponential distribution with mean 30 superimposed. This type of distribution is positively skewed.

Minitab was used to simulate 1000 samples of $n = 4$ intervals between breaks and also 1000 samples of $n = 25$ intervals between breaks. The sample means were calculated and are displayed in the histograms in Figure 4.28 with fitted normal curves superimposed. With sample size $n = 4$ the distribution of the sample mean is positively skewed and the distribution is not normal. However, with sample size $n = 25$ the distribution of the sample mean is much more symmetrical and the normal distribution appears to provide and adequate model for the distribution of the sample mean.

The major importance of the central limit theorem is that it enables probability statements to be made about means of reasonably large samples regardless of whether or not the distribution of individual values is normal. As an example, consider a type of car tyre with life nonnormally distributed with mean 20 000 miles and standard deviation 1600 miles. Suppose that we wish to obtain the probability that a random sample of 64 of these tyres has mean life of 20 400 miles or greater. The solution is as follows.

Even if the random variable X is *not* normally distributed the sample mean will be *approximately* normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$. The larger the sample size n, the better the approximation. This result is known as the central limit theorem.

Box 4.4 The central limit theorem.

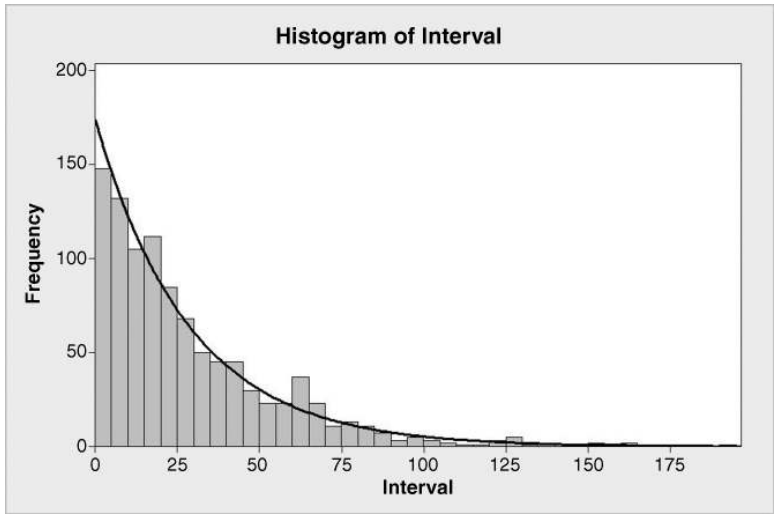

Figure 4.27 Histogram of intervals between breaks with exponential distribution.

Let X denote tyre life and let $\bar{X}$ denote the mean life of random samples of 64 tyres. By the central limit theorem $\bar{X}$ will be approximately normally distributed with mean $\mu_{\bar{X}} = \mu = 20\,000$ and standard deviation $\sigma_{\bar{X}} = \sigma/\sqrt{n} = 1600/\sqrt{64} = 200$. Use of **Calc** > **Probability Distributions** > **Normal...** gives the cumulative probability function value

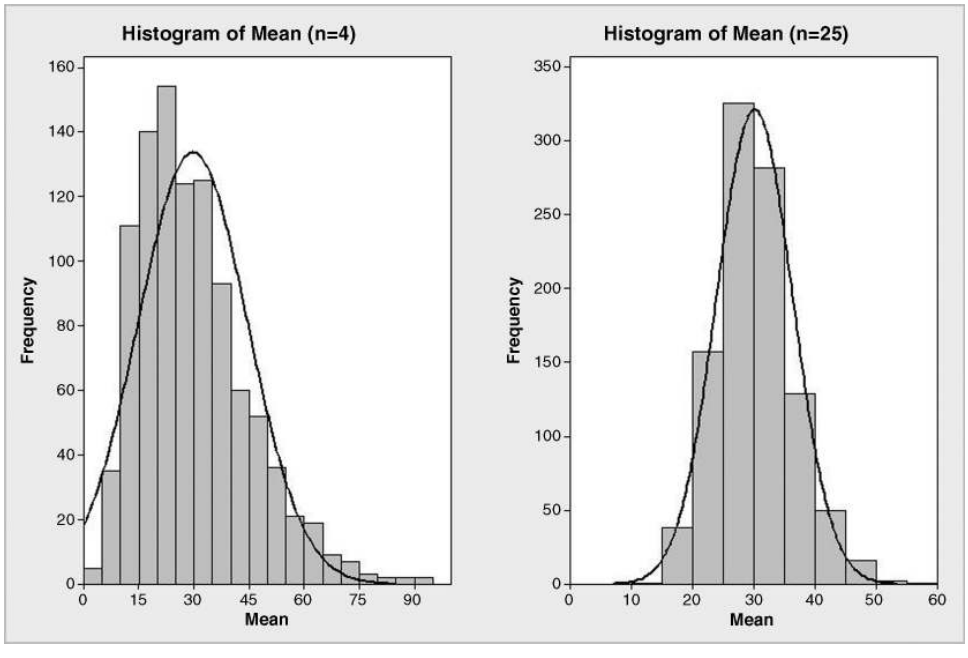

Figure 4.28 Histograms of means of samples of size $n = 4$ and of samples of size $n = 25$.

Let there be constant probability p that an item is nonconforming and let X denote the number of nonconforming items in random samples of n items. Thus X will have the binomial distribution with parameters n and p, i.e. X has the $B(n, p)$ distribution with mean $\mu_X = np$ and standard deviation $\sigma_X = \sqrt{npq}$, where $q = 1 - p$.

The proportion of nonconforming items is given by $V = X/n = kX$, where $k = 1/n$, so the theory in Box 4.1 gives

$$\mu_V = \frac{1}{n} np = p,$$

$$\sigma_V = \frac{1}{n}\sqrt{npq} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p(1-p)}{n}}.$$

The fact that the standard deviation of a proportion is $\sqrt{p(1-p)/n}$ is of fundamental importance in the creation of control charts for proportion of nonconforming items.

Box 4.5 Standard deviation of a proportion.

0.977 25 so the required probability is $1 - 0.977\,25 = 0.022\,75 \approx 1/44$. (Alternatively use can be made of **Graph** > **Probability Distribution Plot...** > **View Probability**.) The conclusion is that there is a 1 in 44 chance of obtaining a mean life of 20 400 or greater for a random sample of 64 of the tyres. Calculations of this nature lie at the heart of hypothesis testing which provides important tools for determining whether or not steps taken to improve a process have been effective. Hypothesis testing is introduced in Chapter 7.

4.3.3 Distribution of the sample proportion

For the development of control charts etc., a formula for the standard deviation of the proportion of nonconforming items in a sample is required. A proportion may be considered as a multiple of a random variable – details are presented in Box 4.5

Consider a scenario where samples of 10 items are taken at regular intervals from a process and checked. Conforming items are denoted by S and nonconforming items by D. The indicator random variable B is defined as having value 0 for a conforming item and value 1 for a nonconforming item. Table 4.9 gives results for one sample. For this sample the proportion of nonconforming items is $3/10 = 0.3$. This is also the mean of the sample of 10 values of the indicator random variable B for the sample. A proportion is a sample mean in disguise! Thus if the probability, p, of a nonconforming item remains constant as successive random samples are taken, the central limit theorem indicates that the series of proportions of nonconforming items

Table 4.9 Conformance record for a sample of 10 items.

Item No.	1	2	3	4	5	6	7	8	9	10
Status	S	D	S	S	D	S	S	S	D	S
Indicator B	0	1	0	0	1	0	0	0	1	0

will have an approximate normal distribution with mean p and standard deviation $\sqrt{p(1-p)/n}$.

In order to illustrate, **Calc > Random Data > Binomial...** was used to generate 1000 random samples from the binomial distribution with parameters $n = 100$ and $p = 0.2$. Thus the number of nonconforming items in random samples of 100 components from a process yielding constant probability 0.2 of a nonconforming item was being simulated for a total of 1000 samples. With **Calc > Calculator...** the simulated counts of nonconforming items were converted to proportions by dividing by 100. According to the above theory the proportions will be approximately normally distributed with mean $p = 0.2$ and standard deviation

$$\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.2 \times 0.8}{100}} = 0.04.$$

The histogram of the proportions displayed in Figure 4.29 was created using **Graph > Histogram...**, the **With Fit** option being used to superimpose a fitted normal curve. The mean and standard deviation of the sample of 1000 proportions of 0.2037 and 0.04065 (displayed to the right of the histogram in Figure 4.29) are close to the population values of 0.20 and 0.04 respectively. The normal distribution also clearly provides an adequate model for the distribution of sample proportions.

The statistical models referred to in this chapter provide the foundations for much of what follows in this book. Readers who wish to gain a deeper and wider understanding of these statistical models would benefit from consulting the books by Montgomery and Runger (2010) and Hogg and Ledolter (1992).

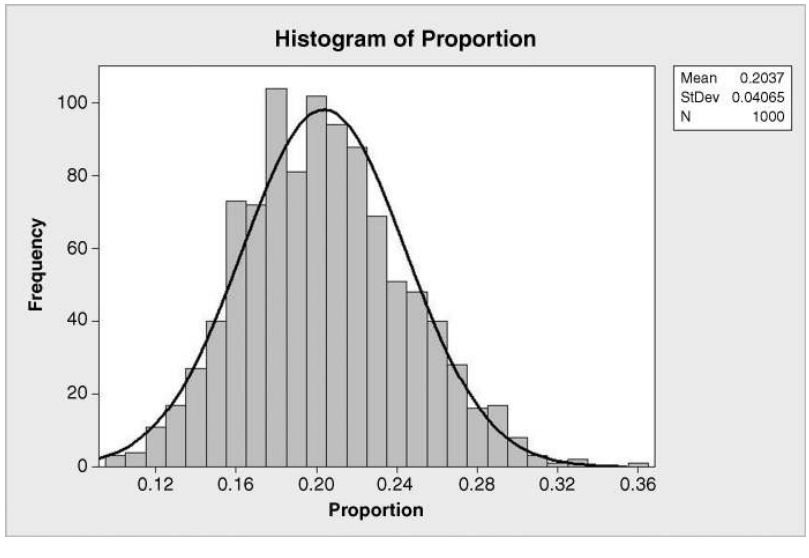

Figure 4.29 Histogram of proportions with fitted normal curve.

4.4 Multivariate normal distribution

The *multivariate normal distribution* provides a statistical model for some scenarios where two, three or more continuous random variables are of interest. For the case of two random variables the multivariate normal distribution is referred to as the *bivariate normal distribution*. A bivariate normal distribution for the two random variables X and Y is specified by the five parameters μ_x, μ_y, σ_x^2, σ_y^2 and σ_{xy}, i.e. by the means, variances and covariance. (The symbol σ_{xy} denotes the population covariance between random variables X and Y.) If X and Y have the bivariate normal distribution specified then X has the $N(\mu_x, \sigma_x^2)$ distribution and Y has the $N(\mu_y, \sigma_y^2)$ distribution. These are known as the marginal distributions of X and Y.

Suppose that a blow moulding process for plastic PET 500 ml bottles yields bottles with weight (g) and diameter (mm) having the bivariate normal distribution with means 25.0 and 72.0, variances 0.04 and 0.05 respectively and covariance -0.03. Thus the covariance matrix is $\begin{pmatrix} 0.04 & -0.03 \\ -0.03 & 0.05 \end{pmatrix}$.

In order to generate some data from this distribution via Minitab, first set up the means in column C1. Second, with commands enabled, use **Calc** > **Matrices** > **Read** to specify that the matrix comprises two rows and two columns and to name it Covariance. The default option **Read from keyboard** is accepted. The dialog box is shown in Figure 4.30. On clicking **OK**, the matrix can be entered following the data prompts as indicated in Panel 4.9. In presenting the data to Minitab one must adhere to the matrix pattern of two rows and two columns.

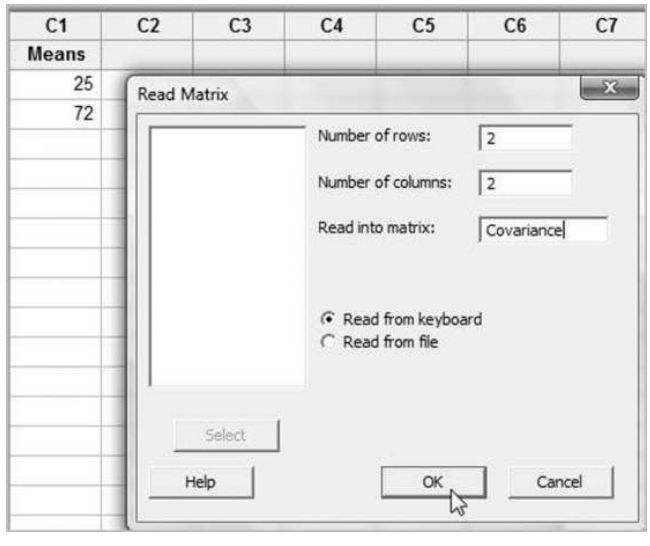

Figure 4.30 Reading data into a matrix.

```
MTB > Name m1 "Covariance"
MTB > Read 2 2 'Covariance'.
DATA> 0.04 -0.03
DATA> -0.03 0.05
2 rows read.
MTB >
```

Panel 4.9 Reading a matrix via the keyboard.

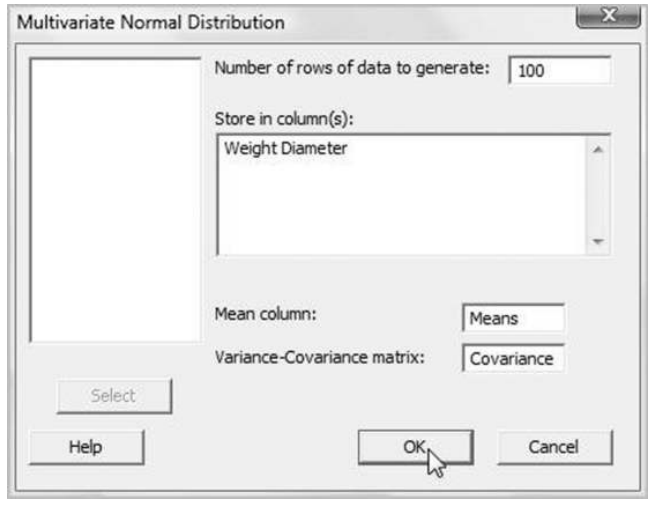

Figure 4.31 Generating data from a bivariate normal distribution.

Minitab now has the necessary information to enable simulated data to be generated for, say, a sample of $n = 100$ bottles. Use **Calc** > **Random Data** > **Multivariate** > **Normal** as shown in Figure 4.31. The marginal plot for the sample generated by the author is displayed in Figure 4.32.

The multivariate normal distribution provides the basis for an important type of control chart for the monitoring of two or more process variables simultaneously.

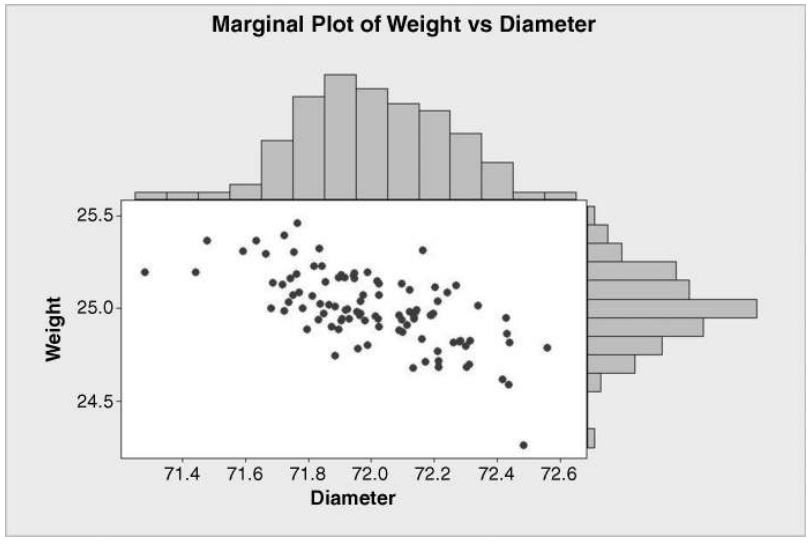

Figure 4.32 Marginal plot of simulated bivariate data.

4.5 Statistical models applied to acceptance sampling

4.5.1 Acceptance sampling by attributes

Acceptance sampling deals with the inspection and classification of a sample of units selected at random from a larger batch or lot and the ultimate decision about disposition of the batch – accept, reject or some other action. Acceptance sampling may be applied to incoming batches of parts from an external supplier that are to be used in a manufacturing process or to batches of incomplete product as they proceed from an internal 'supplier' to the next stage in a manufacturing process. In his chapter on acceptance sampling Montgomery (2009) states that acceptance sampling was a major component of quality improvement activities at the time of the Second World War but that more recently 'it has been typical to work with suppliers to improve their process performance through the use of [statistical process control] and designed experiments'. A brief introduction to acceptance sampling is included here as statistical models discussed earlier in this chapter are applied.

Consider the following single-sampling plan. A random sample of size $n = 400$ units from a batch of size $N = 40\,000$ units are inspected, and if the total number of nonconforming units is less than or equal to the acceptance number $c = 2$ then the batch is accepted; otherwise it is rejected. Suppose that the process that creates the units currently yields 0.5% nonconforming. We would therefore expect a batch of $N = 40\,000$ units to contain $M = 0.5\% \times 40\,000 = 200$ nonconforming units. The probability of acceptance of the batch is the probability that a random sample of $n = 400$ contains 2 or fewer nonconforming units. This probability may be calculated using **Calc** > **Probability Distributions** > **Hypergeometric...** and making the entries displayed in Figure 4.33.

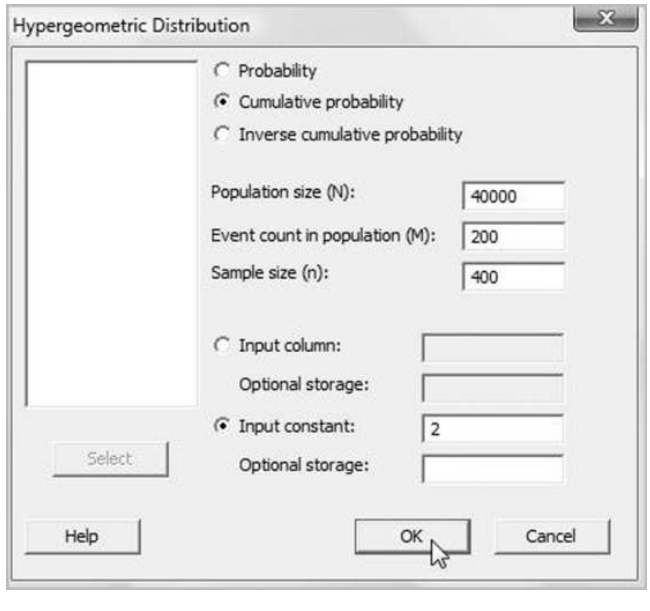

Figure 4.33 Calculation of probability of batch being accepted.

Cumulative Distribution Function

```
Hypergeometric with N = 40000, M = 200, and n = 400

x  P( X <= x )
2     0.676686
```

Panel 4.10 Probability of acceptance of batch with 0.5% nonconforming units.

The result in the Session window, displayed in Panel 4.10, indicates that the probability of obtaining 2 or fewer nonconforming items in the sample is 0.677. Thus the probability of a batch containing 0.5% defective units being accepted is 0.677. (The hypergeometric rather than the binomial distribution is used to do the calculation because the event probability is not constant. However, with large batch sizes and low proportions of nonconforming items the binomial may be used to approximate the hypergeometric. The interested reader will find that the binomial approximation yields 0.676 677 as compared with the 0.676 686 in Panel 4.10.) The reader is invited to verify that the probability of accepting a batch that contains 1% nonconforming units is 0.235 i.e. there is approximately a 1 in 4 chance that a batch containing 1% defective units would be accepted.

A series of calculations yields the table of acceptance probabilities displayed in Figure 4.34. If there are no defects in the lot then the probability of acceptance is 1.000. This means that a batch with no nonconforming items is certain to be accepted. (That is of course desirable! We are, of course, assuming that there are no inspection errors!) The operating characteristic (OC) curve for the sampling plan is a plot of acceptance probability versus proportion of nonconforming items (see Figure 4.35).

In order to design a single-sampling plan two concepts are widely used. The *acceptable quality level* (AQL) is the poorest level of quality for the manufacturing process that the customer would consider acceptable as the average in the long run. The *rejectable quality level* (RQL) is the poorest level of quality the customer is prepared to accept in an individual batch or lot. The RQL is also known as the *lot tolerance percent defective* (LTPD) and the *limiting quality level* (LQL).

Worksheet 1 ***

↓	C1	C2
	Proportion nonconfroming	Probability of acceptance
1	0.000	1.000
2	0.001	0.992
3	0.002	0.954
4	0.003	0.881
5	0.004	0.784
6	0.005	0.677
7	0.006	0.569
8	0.007	0.469
9	0.008	0.378
10	0.009	0.300
11	0.010	0.235

Figure 4.34 Table of acceptance probabilities.

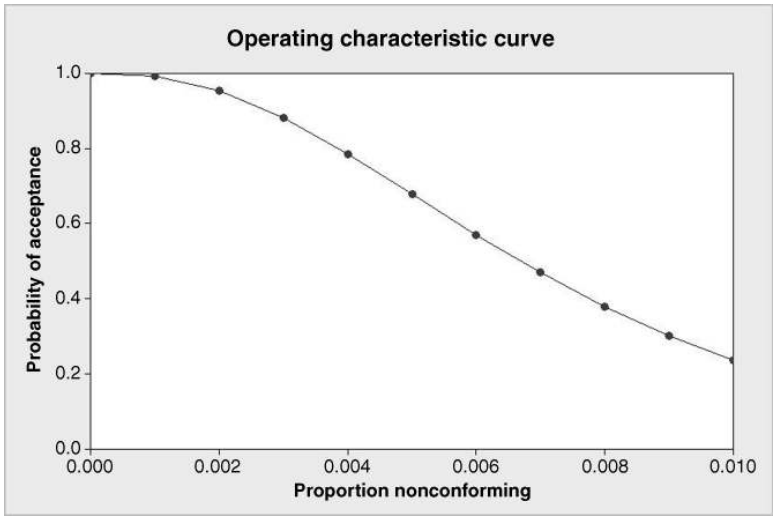

Figure 4.35 Operating characteristic curve.

For the purposes of illustration, suppose that in a particular scenario AQL is 2% and RQL is 8%. Suppose, too, that it considered desirable that there should be probability 0.95 of accepting a lot containing 2% defectives (the AQL) and that there should be probability 0.10 of accepting a lot containing 8% defectives (the RQL). Put another way, this means there would be a 0.05 probability ($1 - 0.95$) of rejecting a lot containing 2% defectives and there would be a 0.90 ($1 - 0.10$) probability of rejecting a lot containing 8% defectives. We can think of the probability 0.05 of rejecting a good batch (with proportion defective equal to the AQL of 2%) as being the *producer's risk* and the probability of and the probability of 0.10 of accepting a bad batch (with proportion defective equal to the RQL of 8%) as being the *consumer's risk*.

It is desirable, then, in this scenario, to have an OC curve which passes through the points (0.02, 0.95) and (0.08, 0.10). Nomograms and tables are available to determine appropriate values for the sample size n and the acceptance number c. Minitab facilitates the calculation using **Stat > Quality Tools > Acceptance Sampling by Attributes...** with the dialog displayed in Figure 4.36. Proportion defective was selected under **Units for quality levels:**, the alternative choices being Percent defective and Defectives per million. Note that it is not necessary to specify a lot (batch) size.

The key part of the Session window output is shown in Panel 4.11. The plan involves taking a random sample of 98 from the batch and acceptance of the batch if the number of nonconforming items found is less than or equal to 4. It was desired to have the OC curve pass through the points (0.02, 0.950) and (0.08, 0.100). The latter part of the Session window output indicates that the best that could be found was the OC curve that passed through (0.02, 0.953) and (0.08, 0.099). The OC curve is also displayed.

The supplied worksheet Unit_Reference_Codes.MTW contains a column named Unit Reference giving reference codes for a batch of 1200 units. In order to select a random sample of 98 units one could use **Calc > Random Data > Sample From Columns...** and enter **Number of rows to sample:** 98, **From columns:** 'Unit Reference', **Store samples in:** Sample. On doing this the author obtained the sequence UO2391, UQ2432, US2428, ...,

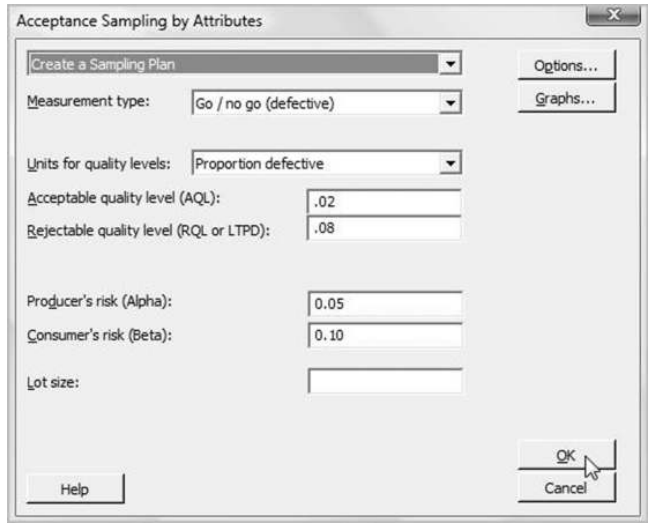

Figure 4.36 Designing an attributes single-sampling plan.

```
Generated Plan(s)

Sample Size        98
Acceptance Number   4

Accept lot if defective items in 98 sampled <= 4; Otherwise reject.
```

Panel 4.11 Specification of the required single-sampling plan.

US2446. The reader is invited to try this for her/himself – it is very unlikely that the same sequence will be obtained!

4.5.2 Acceptance sampling by variables

Rather than simply classifying units as either conforming or nonconforming, it will be possible in some circumstances to base the decision whether or not to accept a batch on the basis of a measurement on each unit in the sample taken. Consider glass bottles for which the specification limits for weight are 485 and 495 g and a scenario where AQL is 0.1% and RQL is 0.5% with producer and consumer risks of 0.05 and 0.10, respectively. Suppose that the standard deviation of bottle weight is unknown but that experience shows that bottle weight may be adequately modelled by a normal distribution. Use of **Stat** > **Quality Tools** > **Acceptance Sampling by Variables** > **Create/Compare...** with the dialog displayed in Figure 4.37 is required.

The key part of the Session window output is shown in Panel 4.12. The plan involves taking a random sample of 160 bottles from the batch, weighing them and calculating the sample mean and standard deviation. Data are provided in supplied worksheet Bottle_Weight_Sample.MTW. The reader is invited to verify that sample mean and standard deviation are respectively 490.94 and 1.18. The two Z-values required in Panel 4.12 may be calculated

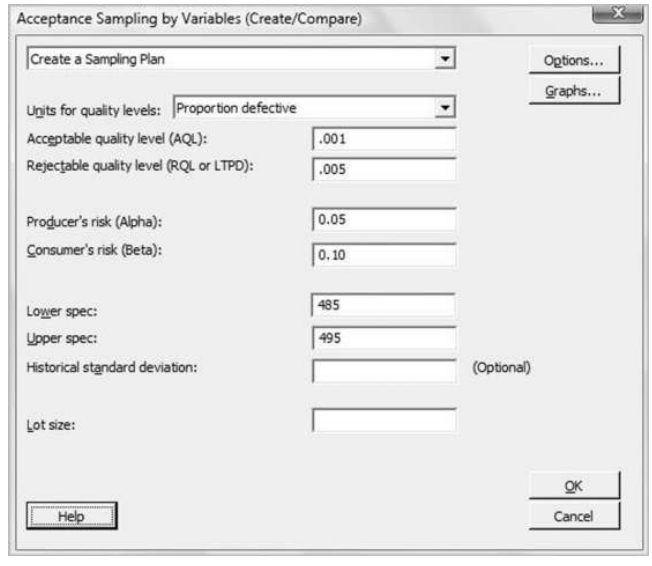

Figure 4.37 Designing a variables single-sampling plan.

```
Generated Plan(s)

Sample Size                      160
Critical Distance (k Value)      2.80110
Maximum Standard Deviation (MSD) 1.65685

Z.LSL = (mean - lower spec)/standard deviation
Z.USL = (upper spec - mean)/standard deviation
Accept lot if standard deviation <= MSD, Z.LSL >= k and Z.USL >= k; otherwise reject.
```

Panel 4.12 Specification of the required single-sampling plan.

as shown in Box 4.6. Since *both* Z-values *exceed* the critical distance (k value) of 2.80 *and* the sample standard deviation (1.18) is *less than* the maximum allowable standard deviation (MSD) of 1.66, the decision would be to accept the batch.

Further information on acceptance sampling for both attributes and variables, including Military Standard plans, may be found in Montgomery (2009) and in the references cited therein.

$$\text{Z.LSL} = \frac{\text{Mean} - \text{Lower spec}}{\text{Standard deviation}} = \frac{490.94 - 485}{1.18} = 5.03$$

$$\text{Z.USL} = \frac{\text{Upper spec} - \text{Mean}}{\text{Standard deviation}} = \frac{495 - 490.94}{1.18} = 3.44$$

Box 4.6 Calculation of the Z-values.

4.6 Exercises and follow-up activities

1. Consider a scenario where a company supplies packaged units to customers. The following probabilities apply:

$$P(\text{Unit conforms to customer requirements}) = 0.95,$$
$$P(\text{Packaging is sound}) = 0.92,$$
$$P(\text{Delivery is on schedule}) = 0.90.$$

Assuming independence, calculate the probability that a customer who orders a unit will receive it free from nonconformities, soundly packaged and delivered on schedule.

2. The file Transaction.MTW contains data on a sample of transactions carried out by two teams, A and B, at a branch office of a major financial institution. Each transaction was classified as having status either conforming (C) or nonconforming (N-C) in terms of the current specifications within the institution. Use **Stat > Tables > Cross Tabulation and Chi-Square...** to summarize the data. Hence, write down estimates of the following probabilities for the population of transactions sampled: $P(A), P(B), P(C), P(\bar{C}), P(\bar{C}|A), P(\bar{C}|B)$. Do you think that status is independent of team?

3. An injection moulding process for the production of digital camera casings yields 10% defective. Denote by X the number of defective casings in random samples of 20 casings selected from the process output at regular intervals.

   (i) State the distribution of the random variable X and its parameters.

   (ii) Obtain $P(X \leq 3)$, $P(X > 3)$, $P(X < 3)$.

   (iii) Use **Calc > Make Patterned Data > Simple Set of Numbers...** to set up the values of x from 0 to 20 in column C1 as indicated in Figure 4.38. (This facility can save much tedious typing!)

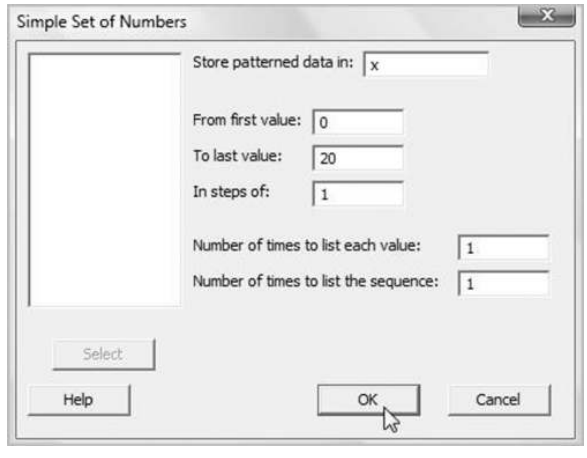

Figure 4.38 Setting up the sequence of values of x.

Table 4.10 Frequencies of goals per match.

Number of goals scored	Number of matches
0	7
1	17
2	13
3	14
4	7
5	5
6	0
7	1
8 or more	0

(iv) Tabulate and display the probability function $f(x)$ and tabulate the cumulative probability function $F(x)$.

(v) Calculate the mean and standard deviation of X.

4. Primary healthcare workers believe that anaphylactic shock reaction to immunization injections occurs with children in one case in two thousand. Immunization teams carry adrenalin packs in order to treat children suffering a shock reaction. Calculate, using Minitab, the probability that two adrenalin packs would be insufficient to treat the shock reactions occurring in a group of 1200 children.

5. Table 4.10 gives the numbers of goals scored in a the soccer matches played in the 2010 FIFA World Cup in South Africa.

   (a) Set up the data in Minitab in a column named Goals using the same method as was used for the V2 flying bomb data and obtain the mean number of goals per match.

   (b) Fit a Poisson distribution to the data. (You should find that you obtain expected frequencies 6.6, 15.0, 17.0, 12.9, 7.3, 3.3, 1.2, 0.4 and 0.2.)

   (c) In order to create the type of display in Figure 4.9 set up columns in Minitab as displayed in Figure 4.39. (Try using **Calc** > **Make Patterned Data** > **Text Values...** to create the column indicating the type of frequency.)

   (d) In order to create the chart use **Graph** > **Bar Chart...** and use the drop-down menu to select the option **Bars represent:** Values from a table.
   Select Cluster under **One column of values** and then complete the dialog boxes as shown in Figure 4.40. The use of the variable Type for attribute assignment under **Data View** means that bars representing observed frequencies in the chart have a different colour from those representing expected frequencies. Observe the good fit of the Poisson model to the observed data and note that this indicates that goals may be considered as random events in a time continuum.

   (e) Minitab actually provides under **Stat** > **Basic Statistics** > **Goodness-of-Fit Test for Poisson...** a formal method for assessing how well a Poisson distribution fits

↓	C1	C2	C3-T
	Goals	Frequency	Type
1	0	7.0	Observed
2	1	17.0	Observed
3	2	13.0	Observed
4	3	14.0	Observed
5	4	7.0	Observed
6	5	5.0	Observed
7	6	0.0	Observed
8	7	1.0	Observed
9	8	0.0	Observed
10	0	6.6	Expected
11	1	15.0	Expected
12	2	17.0	Expected
13	3	12.9	Expected
14	4	7.3	Expected
15	5	3.3	Expected
16	6	1.2	Expected
17	7	0.4	Expected
18	8	0.2	Expected

Figure 4.39 Data for bar chart creation.

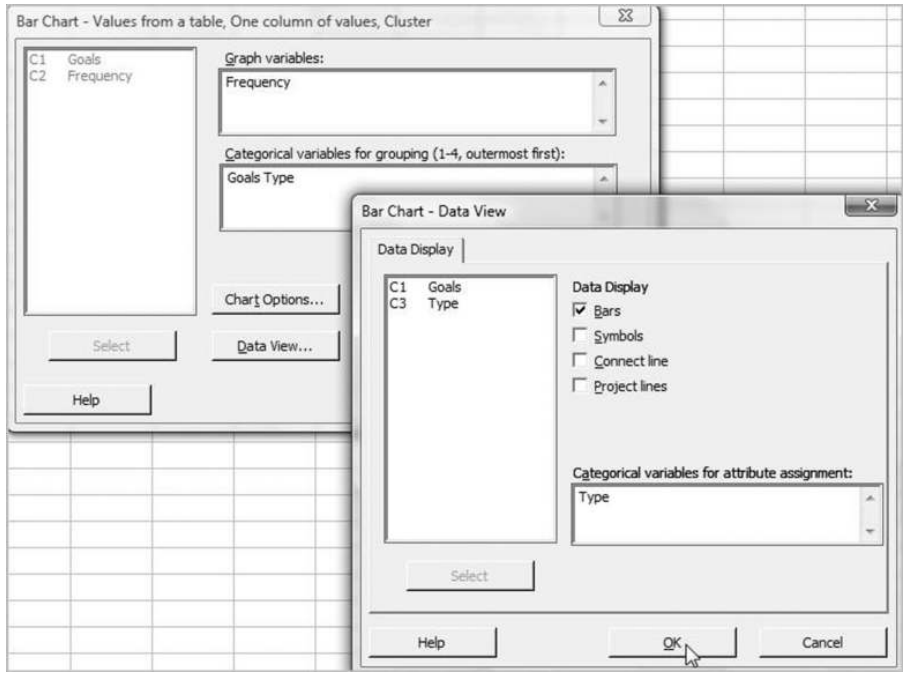

Figure 4.40 Creating a bar chart to compare observed and expected frequencies.

Goodness-of-Fit Test for Poisson Distribution

```
Data column: No. Goals

Poisson mean for No. Goals = 2.26563

No.                    Poisson              Contribution
Goals  Observed  Probability  Expected     to Chi-Sq
0             7     0.103765    6.6410      0.019410
1            17     0.235093   15.0459      0.253777
2            13     0.266316   17.0442      0.959612
3            14     0.201124   12.8720      0.098858
4             7     0.113918    7.2908      0.011595
>=5           6     0.079783    5.1061      0.156476


 N  N*  DF   Chi-Sq  P-Value
64   0   4  1.49973    0.827
```

Panel 4.13 Assessing goodness-of-fit of a Poisson distribution.

the data. Enter **Variable:** Goals, leave the **Frequency:** window blank and under **Graphs...** check only **Bar chart of the observed and the expected values**. A similar bar chart to the one created earlier is produced, along with the Session window output in Panel 4.13.

In the procedure followed by Minitab numbers of goals per match with low expected frequencies are combined so that no expected frequency is less than 5. Informal support for the provision of a satisfactory model by a Poisson distribution comes from the relatively close match between the lengths of the bars representing observed frequencies and the lengths of the bars representing expected frequencies in the chart. Formal acceptance of the Poisson distribution as a satisfactory model is provided by the P-value well in excess of 0.05. Details of the chi-squared goodness-of-fit test used are not provided in this book.

6. Customers of a mortgage bank expect to have mortgage applications processed within 35 working days of submission. Given that the processing time at the bank can be modelled by the $N(28, 3^2)$ distribution, verify that 99% of applications would be processed within 35 working days. Following a successful Six Sigma project aimed at reducing processing time, it was found that, although variability in processing time was unchanged, the mean had dropped to 19 working days. The bank wishes to inform prospective customers that '99 times out 100 we can process your application in q working days'. Assuming that the normal distribution is still an adequate model, calculate q.

7. Use of the symbol Z for a random variable having the standard normal distribution, i.e. the $N(0, 1)$ distribution, is fairly widespread. Many statistical texts use the notation z_p for the value such that $P(Z \leq z_p) = 1 - p$. Thus for the case of $p = 0.1, 1 - p = 0.9$, the required value may be obtained using **Calc > Probability Distributions > Normal...** with the **Inverse cumulative probability** option.

   Use Minitab to generate the worksheet in Figure 4.41. In order to round the values of z_p to two decimal places click on a cell in the column containing the values to be rounded. Next use **Editor > Format Column > Numeric...**, select **Fixed decimal** and change the number of decimal places to 2.

↓	C1	C2	C3
	p	1-p	zp
1	0.1000	0.9000	1.28
2	0.0500	0.9500	1.64
3	0.0250	0.9750	1.96
4	0.0100	0.9900	2.33
5	0.0050	0.9950	2.58
6	0.0025	0.9975	2.81
7	0.0010	0.9990	3.09
8	0.0005	0.9995	3.29

Figure 4.41 Worksheet to be created in Exercise 7.

8. The supplied worksheet Burst.MTW contains burst strength (psi) data for a sequence of bottles taken from mould number 67 at regular intervals during a production run.

   (i) Create a run chart of the data and note how it provides no indication that mould 67 behaved in other than a stable, predictable manner during the production run.

   (ii) Create a histogram of the data with a fitted normal distribution.

   (iii) Assess how well a normal distribution models burst strength by performing a normality test.

   (iv) Estimate the proportion of bottles from mould 67 failing to conform to the requirement that burst strength should be at least 250 psi. (Use the normal distribution with parameters estimated from the data, i.e. with mean 626.2 and standard deviation 117.5, to perform the calculation.)

   (v) Use the table in Appendix 1 to obtain the sigma quality level for mould 67.

9. In addition to columns and matrices, constants may be used in Minitab. With commands enabled the calculation, performed using **Calc** > **Calculator** and described towards the end of Section 4.1, may be performed using constants K1 and K2 as shown in Panel 4.14. Check the calculation involved in Exercise 4.2 using this method.

```
MTB > let k1=0.933189
MTB > let k2=k1**10
MTB > print k2

Data Display

K2    0.500837
```

Panel 4.14 Calculation using constants.

10. Follow the procedure described in Section 4.4 and simulate your own sample of weight and diameter data for 100 bottles. Obtain the correlation between weight and diameter for your sample and compare with the population value calculated from the covariance matrix used in the simulation.

11. Read the Minitab tutorial material at http://www.minitab.com/en-GB/training/tutorials/accessing-the-power.aspx?id=1688&langType=2057 on acceptance sampling.

12. A single-sample acceptance attributes sampling plan is required to have a producer's risk of 0.06 for an acceptable quality level of 0.5% nonconforming, and a consumer's risk of 0.10 for a rejectable quality level of 5% nonconforming. Use Minitab to determine the appropriate plan.

13. It is desirable that the greatest torque required to loosen the cap on a type of food container should be 2.5 N m. For producer's risk of 0.05 for an acceptable quality level of 1% nonconforming, and a consumer's risk of 0.10 for a rejectable quality level of 8% nonconforming, determine the appropriate plan. Use Minitab to confirm that the sample size required is 27 and that the critical distance is 1.81. Data for a sample of 27 containers from a lot is provided in Torque.MTW. Verify that Z. USL = 2.58 and that therefore the lot would be accepted. Confirm this using **Stat > Quality Tools > Acceptance Sampling by Variables > Accept/Reject Lot...**.