original numerical information. For example, no use is made of the fact that two out of the 18 processing times in the above practical problem are actually greater than 16 working hours, i.e. *two* days! Unfortunately, as we have discussed (see page 67) there is no way of using such more detailed information without upsetting the distribution-free property of the procedure. Nevertheless the signed-rank test, published by Wilcoxon (1945), does make some use of the sizes of the differences between the observed values and the hypothesized median. The price to be paid is a restriction on the form of the population distribution for which the procedure is valid. Not only should the distribution be continuous (as usual), but also its probability density function must be *symmetric*. This is, of course, by no means as severe a restriction as that of normality, which is required by the classical statistical methods, but it does mean that the test is not truly distribution-free.

PRACTICAL PROBLEM

We shall illustrate Wilcoxon's signed-rank test with the same practical problem as for the sign test, i.e. using the times that 18 orders took to be processed by an administration system.

HYPOTHESES

As before, we have

$$H_0: \phi = 8 \text{ h.}$$
$$H_1: \phi > 8 \text{ h.}$$

STATISTIC

For each observation we compute the difference between the time taken and the hypothesized median value of 8 h. These differences are shown below:

+8 h 30 min	+6 h 00 min	−2 h 20 min	+1 h 10 min	+3 h 45 min
−3 h 40 min	−0 h 05 min	+2 h 15 min	−0 h 15 min	+8 h 05 min
+2 h 05 min	−0 h 30 min	+1 h 15 min	+3 h 55 min	+1 h 25 min
+2 h 35 min	+0 h 20 min	+2 h 10 min		

To check whether the symmetry condition on the distribution of differences seems reasonable, we could construct the 'dot diagram' shown

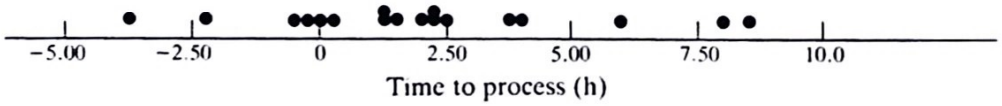


Figure 3.1 Dotplot of the data.

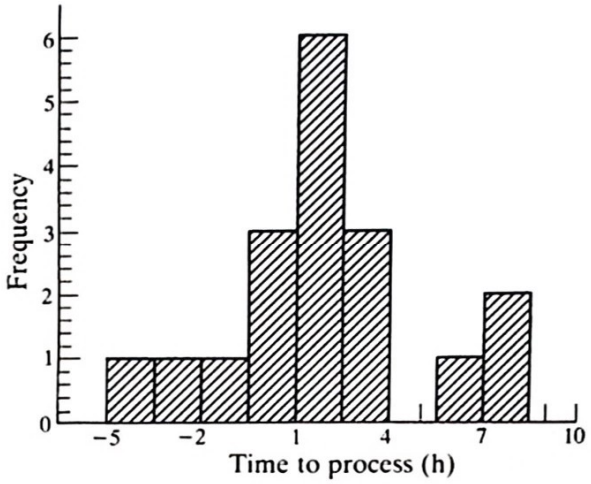


Figure 3.2 Histogram of the data.

in Figure 3.1 or the histogram shown in Figure 3.2. Neither diagram gives cause for concern about the symmetry assumption.

Next, these differences are ranked, *ignoring whether they are positive or negative*, from 1 for the smallest to 18 for the largest. The table below shows these ranks together with the differences:

Difference	Rank	Difference	Rank
+8 h 30 min	18	+8 h 05 min	17
+6 h 00 min	16	+2 h 05 min	8
−2 h 20 min	11	−0 h 30 min	4
+1 h 10 min	5	+1 h 15 min	6
+3 h 45 min	14	+3 h 55 min	15
−3 h 40 min	13	+1 h 25 min	7
−0 h 05 min	1	+2 h 35 min	12
+2 h 15 min	10	+0 h 20 min	3
−0 h 15 min	2	+2 h 10 min	9

The Wilcoxon signed-rank statistic T is then calculated as the sum of the ranks of *either* the positive differences or the negative differences, *whichever the one-sided alternative hypothesis suggests should be the smaller* (see later for the procedure with a two-sided alternative hypothesis). In our problem, H_1 suggests that there should be a smaller number of *negative* differences than positive ones, and so we take T to be the sum of the ranks of the *negative* differences.

Clearly, if H_0 is true, we would expect the sums of the ranks of the negative and positive differences to be roughly equal. On the other hand, if H_1 is true, we would expect one of these sums to be large and the other small. As with Table C for the sign test, the critical values in Table D are

given in terms of *small* values of the signed-rank statistic, which is the reason for the italicized part of the above definition of T.

CRITICAL REGIONS

As we have just indicated, critical regions for the Wilcoxon signed-rank test are of the form

$$T \leqslant \text{critical value}$$

Table D gives critical values of T for sample sizes up to 50 (a more extensive table is on page 29 of *EST*). For $n=18$ and significance levels $\alpha_1=5\%$ and 1%, the critical regions are $T\leqslant 47$ and $T\leqslant 32$ respectively.

SOLUTION

The sum of the ranks for the negative differences (−2 h 20 min, −3 h 40 min, −0 h 05 min, −0 h 30 min and −0 h 15 min) is given by

$$T = 11+13+1+2+4 = 31$$

CONCLUSION

The result is within in the 1% critical region, and so we have strong evidence to support the alternative hypothesis that the median processing time is greater than 8 h.

Discussion

As we have previously stated, the null distribution of T is only valid if the underlying probability density function is symmetric; if this density function is non-symmetric then the given critical values will not be valid. Thus, if we have reason to believe that the underlying distribution is symmetric, then the Wilcoxon signed-rank test may be used instead of the sign test and will usually have superior power. Indeed, in our practical problem, Wilcoxon's test gave a result that was significant at $\alpha_1=1\%$ whereas the result from the sign test was significant at only $\alpha_1=5\%$. Wilcoxon's test should not be used if there is any substantial doubt about the symmetry assumption.

One-sided and two-sided tests

We know that a one-sided alternative hypothesis is supported by small values of a particular one of the two rank sums (i.e. the sums of the ranks of either the negative or positive differences). But if the alternative hypothesis is two-sided then T is defined as the smaller of these two rank sums. In this case, critical regions are again of the form $T\leqslant$ critical value, but now the significance levels are those indicated by α_2 in the tables.

What happens if any of the differences are zero, or if any of the differences are equal to each other?

The answer to the first of these questions is usually the same as for the sign test, i.e. we ignore those data and reduce the sample size accordingly. As regards the second question, the problem is of course that the ranks are not then uniquely defined. The way out of this problem is to average the ranks that would have been taken by these tied observations had we been able to discriminate between them, and then allocate this *average* to *each* of the tied values. This important device for dealing with tied values will also be used in many of the other tests in this book. To illustrate the idea, let us suppose the previous times had been recorded to the *nearest 15 min*; this is a nice way of showing how ties can occur by imprecise measuring (as opposed to being 'really' present in the data). These times are:

16 h 30 min	14 h 00 min	5h 45 min	9 h 15 min	11 h 45 min
4 h 15 min	8 h 00 min	10 h 15 min	7 h 45 min	16 h 00 min
10 h 00 min	7 h 30 min	9 h 15 min	12 h 00 min	9 h 30 min
10 h 30 min	8 h 15 min	10 h 15 min		

which now give the following differences:

+8 h 30 min	+6 h 00 min	−2 h 15 min	+1 h 15 min	+3 h 45 min
−3 h 45 min	0 h 00 min	+2 h 15 min	−0 h 15 min	+8 h 00 min
+2 h 00 min	−0 h 30 min	+1 h 15 min	+4 h 00 min	+1 h 30 min
+2 h 30 min	+0 h 15 min	+2 h 15 min		

In the ranking process, the difference of 0 h 00 min is discarded, leaving just 1 differences to rank. The ranks of these remaining differences are:

Difference	Rank	Difference	Rank
+8 h 30 min	17	+8 h 00 min	16
+6 h 00 min	15	+2 h 00 min	7
−2 h 15 min	9	−0 h 30 min	3
+1 h 15 min	4.5	+1 h 15 min	4.5
+3 h 45 min	12.5	+4 h 00 min	14
−3 h 45 min	12.5	+1 h 30 min	6
0 h 00 min	–	+2 h 30 min	11
+2 h 15 min	9	+0 h 15 min	1.5
−0 h 15 min	1.5	+2 h 15 min	9

The two differences +0 h 15 min and −0 h 15 min are (remembering that signs are ignored) in positions that would normally have been assigned ranks of 1 and 2, and so the average of these ranks, 1.5, is assigned to each of these differences. The two differences of +1 h 15 min are treated similarly, as are the differences +3 h 45 min and −3 h 45 min. The three differences −2 h 15 min, +2 h 15 min and +2 h 15 min are in positions 8, 9 and 10 and so each is assigned the average rank of 9.

The test statistic T, the sum of the ranks of the negative differences, is now

equal to 9+12.5+1.5+3 = 26. Since the effective sample size is now 17, we see that this result falls in the α_1=1% critical region of $T \leqslant 27$.

Null distributions

Under the null hypothesis, the true median of the differences is zero. Thus, assuming that the distribution of the differences is *symmetric* about zero, the chances are *even* ('50–50') that the smallest difference (with rank 1) is positive or negative; the same is true for the next smallest difference (with rank 2); and similarly for every one of the ranked differences. (The theory assumes there are no ties.) It follows that all possible allocations (2^n in all) of signs to ranks are equally likely. So if one generates these 2^n allocations (either manually or by computer) and calculates the value of T for each, a frequency distribution for T can be constructed and then the null probability distribution of T is obtained by simply dividing the frequencies by 2^n.

We now outline this procedure for n=6, and thus confirm the critical values given for this value of n in Table D. We will consider the one-sided case where T is the rank sum of negative differences. With n=6 there are 2^6=64 possible alocations of signs; we show below just a small part of the layout:

Ranks						
1	2	3	4	5	6	Value of T
+	+	+	+	+	+	0
−	+	+	+	+	+	1
+	−	+	+	+	+	2
−	−	+	+	+	+	3
+	+	−	+	+	+	3
⋮	⋮	⋮	⋮	⋮	⋮	⋮
+	−	−	−	−	−	20
−	−	−	−	−	−	21

If you complete the details, you will find that the distribution of T is symmetric (why?). Figure 3.3 shows the histogram of the distribution. In particular notice that

$$\text{Prob}(T=0) = 1/64 = 0.015\,625$$
$$\text{Prob}(T\leqslant 1) = 2/64 = 0.031\,250$$
$$\text{Prob}(T\leqslant 2) = 3/64 = 0.046\,875$$
$$\text{Prob}(T\leqslant 3) = 5/64 = 0.078\,125$$

Thus the best conservative critical regions for α_1=2½% and 5% are T=0 and $T\leqslant 2$ respectively, while there are no critical regions corresponding to significance levels of 1% or ½%; these results confirm the entries for n=6 in Table D.

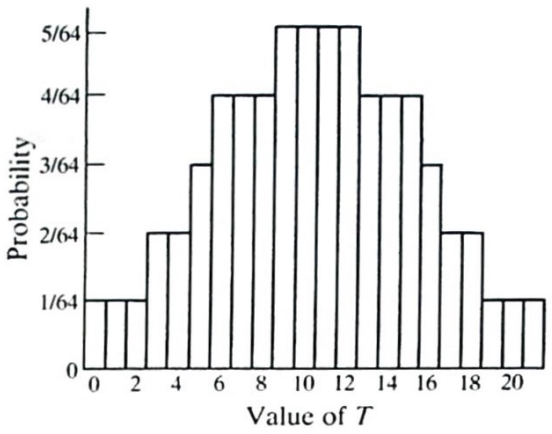


Figure 3.3

Large sample sizes

Figure 3.1 (earlier) showed that for $n=6$ the distribution of T is symmetric. Indeed, as the sample size increases, the shape becomes ever closer to that of the normal distribution. This enables us to obtain approximate critical values for T when n is large. It can be shown that T has mean $\mu=n(n+1)/4$ and standard deviation $\sigma=\sqrt{n(n+1)(2n+1)/24}$.

So approximate critical values of T may be obtained by taking the integer part of $\mu-z\sigma-\frac{1}{2}$, where z is the appropriate percentage point of the standard normal distribution and the '$-\frac{1}{2}$' is the continuity correction.

To illustrate the reliability of this approximation, consider the case when $n=50$ (the largest sample size in Table D). Then $\mu=50\times51/4=637.5$ and $\sigma=\sqrt{50\times51\times101/24}=103.59$, and so the formula for approximate critical values is the integer part of $637.5-103.59z-0.5$. For example, for $\alpha_1=5\%$ we have $z=1.6449$. This gives an approximate critical value of 466.6, and the integer part of this, 466, actually agrees with the entry in Table D.

Critical values for the two-sided test are obtained in a similar manner using the appropriate values of z; for example, for $\alpha_2=5\%$ and 1%, z is 1.96 and 2.5758 respectively.

A note on the computer programs

At the end of this and subsequent chapters we are providing computer programs for most of the statistical procedures discussed in that chapter. The programs are written in BASIC since this language is so widely known and available. However, we are well aware that there are many versions of BASIC in use. Consequently, to make our programs directly usable on as many different types of computer as possible, we have written them in a fairly rudimentary version of BASIC. If your computer supports a more sophisticated BASIC, you may well be able to improve and shorten the programs.

Our programs can be described as 'utilitarian' – they do their job of implementing the test procedures but do not cater for sound, mistakes,

flashing messages, colour, etc.! (No doubt computer buffs will add such effects.) However, in the program that follows, we have included a routine to allow correction of mistakes made in entering the data (this is particularly useful when dealing with large samples) – these few lines of the program can be easily incorporated into other programs if desired.

For small sample sizes, most of the tests in the book are quicker to do 'by hand' than to use the computer – which, of course, is one of the appeals of distribution-free tests. Further, some of the simpler tests (for example, the sign test and Tukey's location test in Chapter 5) can usually be carried out manually in less time than it would take to type in the data, even with large data sizes! Therefore, we have not provided programs for tests of this particularly 'quick-and-easy' type.

Assignments

The first assignment is more suited to the sign test than to Wilcoxon's signed-rank test since it is well known that the distribution of times between arrivals at queues is asymmetric (the sample bears this out!). The second assignment is equally suited for either test and so you may like to try both of them and compare conclusions.

1 Arrivals at check-out

A supermarket chain employed a consultancy firm to investigate service times at their newest store. One aspect that was examined was the time between successive customers arriving at a check-out. Times between arrivals (minutes) were obtained for various periods of a week and those for part of Monday morning are given below:

6.0	7.5	2.7	2.0	2.8	2.6	5.0	11.7	1.3	1.7
3.9	1.2	0.6	0.7	1.7	6.7	2.1	9.6	3.7	4.5
0.5	5.9	1.2	3.1	3.3	0.2	1.9	0.2	2.4	2.0
4.4	0.2	1.2	8.6	0.3	1.9	5.1	0.4	10.4	10.3

Task: Use a histogram or a dot diagram to see whether or not it is reasonable to suggest that these data could have come from a symmetric distribution.

The average time for established stores for a corresponding period is 1.2 min. Is there evidence to suggest that the median time for the new store is greater than this average?

2 Serum sodium levels

The data give the results of analyses of 20 samples of serum measured for their sodium content. The average value for the method of analysis used is 140 ppm.

140	143	141	137	132	157	143	149	118	145
138	144	144	139	133	159	141	124	145	139

(*Source*: National Quality Control Scheme, Queen Elizabeth Hospital, Birmingham, referenced in *Data* by D.F. Andrews and A.M. Herzberg, Springer, 1985.)

Task: Is there evidence that the median level of sodium in this serum is different from 140 ppm?

Program 3.1 The Wilcoxon signed-rank test

```
100 REM *** WILCOXON SIGNED-RANK ONE SAMPLE TEST ***

110 REM ** INPUT THE DATA **
120 INPUT "SAMPLE SIZE ? ";N
130 DIM A(N),D(N),RK(N)
140 PRINT "INPUT THE OBSERVATIONS :"
150 FOR I=1 TO N
160 INPUT A(I)
170 NEXT I

180 REM ** CHECK ENTRIES **
190 INPUT "ARE THE ENTRIES CORRECT ? (Y/N) ";Z$
200 IF Z$="Y" THEN GO TO 300
210 INPUT "WHICH OBSERVATION ? ";L
220 INPUT "NEW OBSERVATION ? ";A(L)
230 PRINT "THE SAMPLE VALUES ARE NOW :"
240 PRINT
250 FOR I=1 TO N
260 PRINT A(I);"  ";
270 NEXT I
280 PRINT
290 GO TO 190

295 REM ** CALCULATION OF THE TEST STATISTIC **
300 INPUT "HYPOTHESISED VALUE OF THE MEDIAN = ";M
310 FOR I=1 TO N
320 LET A(I)=A(I)-M
330 NEXT I

340 REM ** REMOVE ZERO DIFFERENCES **
350 LET NC=0
360 FOR I=1 TO N
370 IF A(I)=0 THEN GO TO 400
380 LET NC=NC+1
390 LET A(NC)=A(I)
400 NEXT I

405 REM ** RANK SAMPLE **
410 LET N=NC
420 FOR I=1 TO N
430 LET D(I)=ABS(A(I))
440 NEXT I
450 GO SUB 1000
```

Continued overleaf

Program 3.1 Continued

```
460 REM ** CALCULATION OF RANK SUMS **
470 LET SN=0
480 LET SP=0
490 FOR I=1 TO N
500 IF A(I)<0 THEN LET SN=SN+RK(I)
510 IF A(I)>0 THEN LET SP=SP+RK(I)
520 NEXT I
530 PRINT
540 PRINT "SUM OF THE POSITIVE RANKS = ";SP
550 PRINT "SUM OF THE NEGATIVE RANKS = ";SN
560 PRINT
570 IF N<=50 THEN STOP
580 REM ** CALCULATION OF APPROXIMATE CRITICAL VALUES (N > 50) **
590 INPUT "PERCENTAGE POINT OF STANDARD NORMAL DISTRIBUTION ";Z
600 LET SD = SQRT(N*(N+1)*(2*N+1)/24)
610 LET MN = N*(N+1)/4
620 PRINT "CRITICAL VALUE = ";INT(MN-Z*SD-.5)

1000 REM ** SORT SAMPLE **
1010 FOR I=1 TO N
1020 LET T=0
1030 LET S=0
1040 LET C=D(I)
1050 FOR J=1 TO N
1060 IF D(J)>C THEN GO TO 1110
1070 IF D(J)<C THEN GO TO 1100
1080 LET S=S+1
1090 GO TO 1110
1100 LET T=T+1
1110 NEXT J
1120 LET RK(I)=T + (S + 1)/2
1130 NEXT I
1140 RETURN
```

4

Goodness-of-fit tests

Introduction

Quite a number of important practical problems require the use of so-called *goodness-of-fit* tests. The following example illustrates the meaning of goodness-of-fit.

A market-research company was given the task of finding out consumer preferences for three new flavours of toothpaste. In the investigation, the new flavours, whisky, brandy and champagne, were tested along with a traditional mint flavour. The four toothpastes were identically packaged; only a letter A, B, C or D identified them. In all, 400 people took part in the experiment and each was asked to state which flavour (as identified by the letter) they preferred. At the end of the experiment, the market researcher would know how many of the 400 people preferred each flavour. Now if there is no general preference for any particular flavour the mean or *expected* number of votes for each flavour would be 400/4=100; the expected* frequencies of votes would be as follows:

Flavour	whisky	brandy	champagne	mint
Number expected in each category	100	100	100	100

What we shall examine is how well the observed data *fit* this expected pattern; hence the term *goodness-of-fit*. Because of natural sample variations, we cannot expect the fit to be exact; but if the fit of the 'observed frequencies' to the 'expected frequencies' is worse than might be reasonably caused by sampling variations, then we have evidence against the equal-preference hypothesis, and so the researcher would then have some specific likes and dislikes to report.

The above example illustrates just one of many different situations

* Although seemingly in universal use, the term *expected* frequency is misleading. Even if there were no general preferences, it would be remarkable indeed if *precisely* 100 votes for each flavour were obtained, and so this could hardly be an 'expected' occurrence in the normal sense of the word. However, in statistics, 'expected' has to come to be used as analogous to 'mean value of'. In the same way, when 1000 000 fair coins are tossed, the mean ('expected') number of heads is 500 000; however, the probability of getting exactly this number is minute.

where a goodness-of-fit test is appropriate. As with other hypothesis tests, it often serves as a preliminary to a larger experiment. For instance, most people experience queues – in the supermarket, at the traffic lights, in the doctors' surgery – and in order to obtain reliable information concerning, say, the average waiting time, or the average queue length, it is very helpful to have some idea of the distributions of arrival times and service times. Data need to be collected and a goodness-of-fit test performed to see whether or not it is reasonable to use a particular probability model.

Another application of goodness-of-fit tests arises when applying classical tests. We have already mentioned (a number of times) that one of the fundamental assumptions of classical statistical tests is that the observations should come from normal distributions. Clearly, it is sensible and important to check whether this assumption seems valid before using such procedures, i.e. we need goodness-of-fit tests to check for normality.

So, in performing a goodness-of-fit test, we are interested in seeing just how well observed data fit a theoretical model or hypothesized distribution. In this chapter we shall examine three popular goodness-of-fit tests: the chi-squared test, the Kolmogorov–Smirnov test and Lilliefors' test (which is a special type of Kolmogorov–Smirnov procedure designed specifically for testing normality).

THE CHI-SQUARED GOODNESS-OF-FIT TEST

Introduction

This versatile and widely used test was introduced by Karl Pearson (1900), which makes it the oldest procedure in this book except for the sign test. It owes its popularity to the fact that it is easy to use and is valid over a very wide range of situations. Indeed, all that the test demands is that the n independent observations are capable of being classified into a number, say k, non-overlapping categories (for example, the preferred flavours of toothpaste, where $k=4$) and that the probabilities of observations falling into these categories can be calculated when the appropriate null hypothesis is assumed true. The procedure has the name 'chi-squared' because (except for small samples) the null distribution of the test statistic is approximately a chi-squared distribution.

PRACTICAL PROBLEM

The sales director of a major pottery company predicted the 1986 sales (as proportions of the total sales) of their five leading designs of bone china dinner sets. The predicted proportions were:

Design	Clementine	Ice rose	Cottage garden	Birds of paradise	Arden
Predicted percentage total sales	35	20	20	10	15

Following usual company practice, at the end of the first quarter, the accounts department samples 400 invoices for dinner sets. The production manager would like to use this sample to indicate whether or not the sales predictions are being fulfilled.

HYPOTHESES

The production manager is interested in seeing whether the sales of the five designs is following the pattern predicted by the sales director. If not, then he will need an indication of the changes so as to take action to adjust production accordingly. The appropriate null hypothesis states that the sales of the five designs is following the predicted pattern and the alternative hypothesis states that this pattern is not being followed. So we have:

H_0: The percentage of sales for Clementine, Ice rose, Cottage garden, Birds of paradise and Arden designs are 35%, 20%, 20%, 10%, and 15% respectively.

H_1: The percentages of sales for the five designs are not 35%, 20%, 20%, 10% and 15%.

DATA

The breakdown of the 400 sampled invoices between the five designs was:

Design	Clementine	Ice rose	Cottage garden	Birds of paradise	Arden
Number	108	102	92	42	56

We shall refer to these data as the *observed* frequencies (Ob. for short) to distinguish them from the frequencies expected (Ex.) under H_0.

STATISTIC

What we need is some method of comparing the observed frequencies (Ob.) with the mean or *expected* frequencies (Ex.) when H_0 is true, i.e. we seek a statistic that measures the discrepancies between the observed and expected frequencies. Now if H_0 is true, then the mean or expected number of Clementine orders is 35% of 400, i.e. 140. Likewise, the

expected frequencies for the remaining designs are $20\% \times 400 = 80$, $20\% \times 400 = 80$, $10\% \times 400 = 40$ and $15\% \times 400 = 60$, respectively. These observed and expected frequencies are shown in the following table:

Frequency	Clementine	Ice rose	Cottage garden	Birds of paradise	Arden
Ob.	108	102	92	42	56
Ex.	140	80	80	40	60

If H_0 *is* true, then the observed frequencies should roughly fit the expected frequencies so that the differences Ob.−Ex. will be relatively small. It is in fact these differences that form the basis of the χ^2 (chi-squared) statistic. Although it might at first sight seem reasonable to form a test statistic by simply summing these differences, this turns out to be useless since the negative differences always cancel out the positive ones to give a sum of zero. Pearson overcame this problem by squaring the differences before performing a summation; in fact his actual test statistic was

$$\chi^2 = \sum \frac{(\text{Ob.} - \text{Ex.})^2}{\text{Ex.}}$$

where the sum is over the k categories. In this formula, each squared difference $(\text{Ob.}-\text{Ex.})^2$ is divided by the appropriate Ex. before summing; this compensates for the effect that the differences (Ob.−Ex.) naturally tend to be larger for large values of Ex. even when H_0 is true.

Therefore, if H_0 is true, the observed frequencies will tend to be relatively close to the expected frequencies, resulting in a corresponding small value of χ^2; on the other hand, if there are large discrepancies between the observed and expected frequencies, indicating that H_1 is true, then χ^2 will tend to be large.

CRITICAL REGIONS

We have seen that, if H_1 is true, then χ^2 will tend to take large values. This immediately leads us to critical regions of the form

$$\chi^2 \geqslant \text{critical value}$$

Critical values are usually obtained from the chi-squared distribution (Table B) since for large samples (see the section entitled 'A note on small expected frequencies' below for more particulars) the null distribution of χ^2 is approximately chi-squared. This approximation is very reliable for large sample sizes, which is fortunate since, in general,

the exact null distribution of χ^2 cannot be calculated. The appropriate number of degrees of freedom of the null distribution is $k-1$, where k is the number of categories.

In our problem, we have $k=5$, and so the approximate critical values for χ^2 are found by consulting the chi-squared distribution (Table B) with four degrees of freedom. Doing this gives the 5% and 1% critical regions as $\chi^2 \geqslant 9.488$ and $\chi^2 \geqslant 13.277$ respectively.

SOLUTION

The calculation of χ^2 is most conveniently performed by carrying out the following sequence:

(1) For each category, calculate the differences Ob.−Ex.
(2) Square the differences to get $(\text{Ob.}-\text{Ex.})^2$.
(3) Divide each $(\text{Ob.}-\text{Ex.})^2$ by the appropriate Ex.
(4) Sum the quantities $(\text{Ob.}-\text{Ex.})^2/\text{Ex}$ to obtain the value of χ^2.

These stages are shown in the following table for the data of our practical problem. (Notice that the sum of the differencs Ob.−Ex. does indeed equal zero as we mentioned above.)

	Clementine	Ice rose	Cottage garden	Birds of paradise	Arden
Ob.	108	102	92	42	56
Ex.	140	80	80	40	60
Ob.−Ex.	−32	22	12	2	−4
$(\text{Ob.}-\text{Ex.})^2$	1024	484	144	4	16
$(\text{Ob.}-\text{Ex.})^2/\text{Ex.}$	1024/140 = 7.31	484/80 = 6.05	144/80 = 1.80	4/40 = 0.10	16/60 = 0.27

The value of χ^2 is now easily obtained by summing the quantities $(\text{Ob.}-\text{Ex.})^2/\text{Ex.}$, thus

$$\chi^2 = 7.31 + 6.05 + 1.80 + 0.10 + 0.27$$
$$= 15.53$$

This value of χ^2 falls in the $\frac{1}{2}$% critical region, which well and truly indicates that we have very strong evidence to support the alternative hypothesis that the percentage of sales does not agree with the sales director's predictions.

Rather than leaving the problem at this stage, it is usually fruitful (whenever H_1 is supported) to see what else can be deduced from the data. Looking at the discrepancies between the observed and expected frequencies – and, in particular, at the individual values of

$(Ob.-Ex.)^2/Ex.$ – we see that the sales for the last three categories are more or less on target, while those for Clementine and Ice rose are not. It would seem that the sales for these two designs are roughly equal rather than reflecting the predicted sales pattern.

A note on small expected frequencies

The chi-squared approximation for the null distribution of χ^2 runs into trouble when the expected frequencies are small; this can happen when some proportions predicted under H_0 are small and/or the sample size is small. The approximation may then be poor and so in this case it is necessary to modify the procedure. As a rough guide:

(a) no expected frequency should be less than 1, and
(b) no more than a fifth of the expected frequencies should be less than 5.

But what should we do if this criterion is not satisfied? The usual approach to this problem is to combine some of the categories in order to avoid the small expected frequencies. To illustrate this idea, suppose that in our practical problem the predicted sales had been 40%, 25%, 20%, 8% and 7%, and that only 50 sales invoices had been sampled, yielding observed frequencies 14, 13, 11, 5 and 7 for the five designs. The table of observed and expected frequencies would now be:

Frequency	Clementine	Ice rose	Cottage garden	Birds of paradise	Arden
Ob.	14	13	11	5	7
Ex.	20	12.5	10	4	3.5

The expected frequencies in the last two categories fail to satisfy condition (b) above – both are less than 5. In this case, the solution is simple; we combine these two offending categories to give the modified table of frequencies:

Frequency	Clementine	Ice rose	Cottage garden	Birds of paradise + Arden
Ob.	14	13	11	12
Ex.	20	12.5	10	7.5

We can now proceed with the usual calculation for χ^2, happy in the

knowledge that the critical values from the chi-squared distribution (now based on $k-1 = 4-1 = 3$ degrees of freedom) are reliable. In fact, the reader may verify that the value of χ^2 for these data is 4.62, which does not fall in the 5% critical region of $\chi^2 \geq 7.815$, and so these results would not give cause to reject the production manager's predictions.

Discussion

Whenever there is a natural choice of categories into which the data can be classified, the chi-squared test is the obvious 'goodness-of-fit' procedure. However, even when natural categories do not exist (as when the distribution specified under H_0 is continuous), the test may still be used by creating 'artificial' categories. On the other hand, in these circumstances, one may prefer the Kolmogorov–Smirnov test (see later in this chapter), which is designed specifically for use with continuous distributions. Nevertheless, since it is still quite common practice to use the chi-squared test on continuous distributions, we now give two examples of this kind of application. We have chosen the same scenario for both examples; in each case we are interested in whether it is likely that a sample could have been taken from a normal distribution – the difference between the two cases is that in the first one the mean and the standard deviation are specified in the null hypothesis, whilst in the second case they are not.

Example on the normal distribution with mean and standard deviation specified

PRACTICAL PROBLEM

A major elevator manufacturer bases its safety standards on the principle that the weights of adults are normally distributed with a mean 62 kg and a standard deviation 11 kg, i.e. N(62,121). However, it was recently brought to their notice that this information dates from over 45 years ago! Naturally, their R&D department wish to find out whether this distribution of weights is still appropriate; for, if the distribution has changed, it might necessitate a revision in the recommendation for the maximum number of people permitted in their elevators.

HYPOTHESES

The R&D department's concern is whether there has been a change in the distribution of weights of the adult population. So the alternative hypothesis relevant to them is that the distribution of weights is not normal with $\mu=62$ kg and $\sigma \leq 11$ kg.

H_0: The sample is from the normal distribution with mean 62 and standard deviation 11.

H_1: The sample is not from the normal distribution with mean 62 and standard deviation 11.

DATA

In the actual survey, the R&D team sampled 500 adults. However, for our purpose in illustrating the chi-squared technique (and later the Kolmogorov–Smirnov test), we will use just 50 of these weights. The weights (kg) were:

72.2	64.0	53.4	76.8	86.3	58.1	63.2	73.1	78.0	44.3
85.1	66.6	80.4	76.0	68.8	76.8	58.9	58.1	74.9	72.2
73.1	39.3	52.8	54.2	65.3	74.0	63.2	64.7	68.8	85.1
62.2	76.0	70.5	48.9	78.0	66.6	58.1	32.5	63.2	64.0
68.8	65.3	71.9	72.2	63.2	72.2	70.5	80.4	45.4	59.6

SOLUTION

Our first job is to group these data into intervals (which we have arbitrarily chosen to be roughly 5 kg in width) to form the 'grouped' frequency table:

Interval	Frequency
less than 40 kg	2
40–49.9 kg	3
50–54.9 kg	3
55–59.9 kg	5
60–64.9 kg	8
65–69.9 kg	7
70–74.9 kg	11
75–79.9 kg	6
80–89.9 kg	5
90 kg or more	0

Next, based on the supposition that the null hypothesis is correct, i.e. the data are from the N(62,121) distribution, we calculate the expected proportion of the distribution in each interval; the expected frequencies are then obtained by multiplying these proportions by the sample size, 50. Using the relation $Z=(X-\mu)/\sigma$ ($= (X-62)/11$, in our case), which we first used in Chapter 1, we compute the value of the standard normal variable corresponding to each interval boundary in the above frequency table. The following table shows the details:

Interval	Left-hand boundary, x	$z=(x-62)/11$	$\Phi(z)$	Expected proportion in interval	Expected frequency in interval
					$50 \times 0.0228 =$
< 40				0.0228	1.140
40–49.9	40	−2.00	0.0228	0.1151	5.755
50–54.9	50	−1.09	0.1379	0.1232	6.160
55–59.9	55	−0.64	0.2611	0.1675	8.375
60–64.9	60	−0.18	0.4286	0.1778	8.890
65–69.9	65	0.27	0.6064	0.1609	8.045
70–74.9	70	0.73	0.7673	0.1137	5.685
75–79.9	75	1.18	0.8810	0.0685	3.425
80–89.9	80	1.64	0.9495	0.0451	2.255
$\geqslant 90$	90	2.55	0.9946	0.0054	0.270

To satisfy the requirements of the chi-squared approximation, we combine the last three categories; then only one expected frequency – the first one – out of the remaining eight is less than 5, and so the conditions mentioned previously are satisfied. The value of χ^2 is obtained in the usual manner from the resulting eight categories:

Interval	< 40	40–49.9	50–54.9	55–59.9	60–64.9	65–69.9	70–74.9	$\geqslant 75$
Ob.	2	3	3	5	8	7	11	11
Ex.	1.140	5.755	6.160	8.375	8.890	8.045	5.685	5.950
$(\text{Ob.}-\text{Ex.})^2/\text{Ex.}$	0.649	1.319	1.621	1.360	0.089	0.136	4.969	4.286

Hence

$$\chi^2 = 0.649 + 1.319 + 1.621 + 1.360 + 0.089 + 0.136 + 4.969 + 4.286 = 14.429$$

We have not been able to specify the decision rule before now since we need to know the number k of categories available for calculating χ^2, and this was unknown until the expected frequencies were computed.

In this problem, we have $k=8$ categories and hence $k-1=7$ degrees of freedom. Hence, from Table B, we see that we reject H_0 in favour of H_1 at the 5% and 1% levels of significance if $\chi^2 \geqslant 14.067$ and $\chi^2 \geqslant 18.475$ respectively. We see that the calculated value of $\chi^2=14.429$ lies in the 5%, but not in the 1%, critical region, and so we have some (but not strong) evidence to support H_1. In other words, it seems unlikely that the sample is from N(62,121) – indeed, it may not be from a normal distribution at all or it could be from a normal distribution but having mean and standard deviation different from 62 and 11 respectively.

In the next example of the use of the chi-squared test, we again look at the problem where the sample is hypothesized to come from a normal

distribution but now where the mean and standard deviation are not specified in H_0, i.e. we are now concerned with the general 'parentage' of our sample but not in the particular member of the family!

Example involving the normal distribution – mean and standard deviation not specified

PRACTICAL PROBLEM AND DATA

We will use the same problem and data as in the previous example.

HYPOTHESES

This time the R&D department's concern is simply whether or not it is reasonable to claim that the sample came from some normal distribution, without assuming anything about the mean and standard deviation of the distribution. So we have:

H_0: The sample of adult weights is from some normal distribution.
H_1: The sample is not from a normal distribution.

SOLUTION

The solution follows the same pattern as before, except now the mean and standard deviation have to be estimated from the sample using the formulae from Chapter 1. The sample mean $\bar{X}=66.35$ estimates μ while σ is estimated by the modified sample standard deviation, $s=\sqrt{\sum(X-\bar{X})^2/(n-1)} = 11.61$. So, by using the relation $Z=(X-66.35)/11.61$ (from $Z=(X-\mu)/\sigma$ with μ and σ replaced by their estimates $\bar{X}$ and s respectively) we can now calculate the value of the standard normal variable corresponding to each interval boundary. The details of this calculation and that for the expected frequencies is given in the following table.

Interval	x	$z=\frac{(x-66.35)}{11.61}$	$\Phi(z)$	Expected proportion	Expected frequency
					$50 \times 0.0116 =$
< 40				0.0116	0.580
40–49.9	40	−2.27	0.0116	0.0677	3.385
50–54.9	50	−1.41	0.0793	0.0842	4.210
55–59.9	55	−0.98	0.1635	0.1277	6.385
60–64.9	60	−0.55	0.2912	0.1610	8.050
65–69.9	65	−0.12	0.4522	0.1695	8.475
70–74.9	70	0.31	0.6217	0.1487	7.435
75–79.9	75	0.75	0.7734	0.1076	5.380
80–89.9	80	1.18	0.8810	0.0983	4.915
≥ 90	90	2.04	0.9793	0.0207	1.035

This time, to satisfy the validity conditions for the chi-squared approximation, we combine the first three and the last two categories. The value of χ^2 is calculated in the usual manner from the resulting seven categories:

Interval	< 55	55–59.9	60–64.9	65–69.9	70–74.9	75–79.9	≥ 80
Ob.	8	5	8	7	11	6	5
Ex.	8.175	6.385	8.050	8.475	7.435	5.380	5.950
$(\text{Ob.}-\text{Ex.})^2/\text{Ex.}$	0.004	0.300	0.000	0.257	1.709	0.071	0.152

from which $\chi^2=2.493$.

Recall that in our previous examples, the number of degrees of freedom has been $k-1$. However, in this problem, both the mean and the standard deviation of the population have had to be estimated from the data in order to compute the expected frequencies. This tends to make the expected frequencies closer to the observed frequencies than would otherwise be the case. Fortunately, this effect is easily compensated by *reducing* the number of degrees of freedom by the number (two, in this example) of such estimated quantities. So our decision rule is now based on $k-1-2=7-1-2=4$ degrees of freedom, and rejects H_0 in favour of H_1 at the 5% and 1% levels of significance if $\chi^2 \geq 9.488$ and $\chi^2 \geq 13.277$ respectively.

The calculated value of $\chi^2=2.493$ is nowhere near any of these critical regions. Consequently, there is no evidence at all of the population being non-normal.

KOLMOGOROV–SMIRNOV TEST

Introduction

This test was devised by A. N. Kolmogorov (1933), and a similar technique was developed by N. V. Smirnov (1939); consequently it is commonly known as the Kolmogorov–Smirnov test. It is specifically designed for analysing goodness-of-fit situations where the underlying distribution is continuous. One immediate consequence is that we are saved the awkwardness of creating arbitrary categories as needed in the chi-squared test. Another useful feature of their test is that it can be carried out using either an arithmetical method or an equivalent graphical approach; the latter may not be any quicker, but it does aid the understanding of the principles behind the technique.

In Chapter 1 we saw that the sample cdf can be expected to reflect the main features of the population cdf (unless the sample is very small). Basically, the Kolmogorov–Smirnov procedure compares the sample cdf

(denoted here by $F_n(x)$ with n referring to the sample size) with the hypothesized population cdf (denoted here by $F_0(x)$ with the '0' indicating that this is the cdf predicted under H_0). If H_0 were true, we would expect $F_0(x)$ and $F_n(x)$ to stay reasonably close together; on the other hand, large differences between $F_0(x)$ and $F_n(x)$ constitute evidence against H_0. There are several tests available based on comparing $F_0(x)$ and $F_n(x)$, essentially differing only in their methods of comparison. The Kolmogorov–Smirnov test is the best-known of these tests, and its measure of difference between the cdfs is simply the maximum (vertical) distance between $F_0(x)$ and $F_n(x)$.

It is worth noting that, although this test was primarily designed for continuous distributions, nevertheless it can be applied to discrete distributions, though in this case the critical values may be over-conservative (i.e. the true α may be less than the nominal α).

PRACTICAL PROBLEM

We have chosen to use the same example as before, i.e. the elevator problem. This will give us the advantage of seeing how the Kolmogorov–Smirnov test performs relative to the chi-squared test, which if you recall is *not* ideally suited for dealing with continuous distributions.

HYPOTHESES

The R&D department's concern is whether there has been a change in the distribution of the weights of the adult population. So the alternative hypothesis to them is that the distribution of weights is not normal with $\mu=62$ kg, $\sigma=11$ kg; clearly this is a *two-sided* alternative hypothesis since the alternative cdf could lie above or below that of the hypothesized distribution. Thus, the hypotheses may be written as follows:

H_0: The sample is from the normal distribution with mean 62 and standard deviation 11.
H_1: The sample is not from the normal distribution with mean 62 and standard deviation 11.

Should the evidence favour H_1, then, of course, further analysis will be needed to examine what kind of model distribution of weights might now be appropriate.

Note that we have specifically stated that H_1 is a two-sided hypothesis rather than saying 'nothing' as we did when we used the chi-squared test to analyse this problem. As we discuss later, the Kolmogorov–Smirnov test has the advantage of being easily adaptable for use with either one-sided and two-sided alternative hypotheses as appropriate.

DATA

For convenience we reproduce below the data for this problem:

72.2	64.0	53.4	76.8	86.3	58.1	63.2	73.1	78.0	44.3
85.1	66.6	80.4	76.0	68.8	76.8	58.9	58.1	74.9	72.2
73.1	39.3	52.8	54.2	65.3	74.0	63.2	64.7	68.8	85.1
62.2	76.0	70.5	48.9	78.0	66.6	58.1	32.5	63.2	64.0
68.8	65.3	71.9	72.2	63.2	72.2	70.5	80.4	45.4	59.6

STATISTIC

If H_0 is true, then we would expect the sample cdf to bear some resemblance to the cdf of the normal distribution N(62,121). On the other hand, if H_1 is true, then the two cdfs may differ markedly in some respect. For example, they may have similar shapes but with one of them displaced from the other, indicating that only the means differ; or, one may be steeper than the other, indicating a difference in spread; or, the shapes may be very different, indicating that it is unlikely that the sample comes from any normal distribution.

The closeness, or otherwise, of the two cdfs is measured by calculating the greatest vertical difference, D_n, between them; if D_n is large, then we can claim it is unlikely that the sample comes from the hypothesized distribution, i.e. N(62,121). We recall that the sample cdf, $F_n(x)$, is given by

$$F_n(x) = \text{proportion of sample observations less than or equal to } x$$

i.e.

$$F_{50}(x) = \frac{\text{number of observations} \leqslant x}{50}$$

The sample cdf, $F_{50}(x)$, for our problem is calculated in Table 4.1, where you should note that we have rearranged the observations into ascending order. Also, whether one opts for the graphical or the arithmetical method, it is advantageous to convert the observed values to standardized values, z, using the transformation $Z = (X-\mu)/\sigma$, where μ and σ are the values given in H_0, i.e. we use $z=(x-62)/11$. We then obtain the results in Table 4.1.

Since the hypothesized distribution is normal, we may make use of graph paper on which the standard normal distribution has already been drawn (such as the one on page 27 of *EST*). Then the 'standardized' sample cdf (i.e. using the z values just calculated) may be drawn and the largest vertical difference, D_{50}, found. Figure 4.1 shows the result. We have indicated the greatest vertical difference, D_{50}, on Figure 4.1, it

Table 4.1

Observation, x	Standardized value, z	Frequency	Cumulative frequency	$F_{50}(x)$
32.5	−2.68	1	1	0.02
39.1	−2.06	1	2	0.04
44.3	−1.61	1	3	0.06
45.4	−1.51	1	4	0.08
48.9	−1.19	1	5	0.10
52.8	−0.84	1	6	0.12
53.4	−0.78	1	7	0.14
54.2	−0.71	1	8	0.16
58.1	−0.35	3	11	0.22
58.9	−0.28	1	12	0.24
59.6	−0.22	1	13	0.26
62.2	0.02	1	14	0.28
63.2	0.11	4	18	0.36
64.0	0.18	2	20	0.40
64.7	0.25	1	21	0.42
65.3	0.30	2	23	0.46
66.6	0.42	2	25	0.50
68.8	0.62	3	28	0.56
70.5	0.77	2	30	0.60
72.2	0.93	5	35	0.70
73.1	1.01	2	37	0.74
74.0	1.09	1	38	0.76
74.9	1.17	1	39	0.78
76.0	1.27	2	41	0.82
76.8	1.35	2	43	0.86
78.0	1.45	2	45	0.90
80.4	1.67	2	47	0.94
85.1	2.10	2	49	0.98
86.3	2.21	1	50	1.00

occurs at just before $z=0.11$ (corresponding to $x=63.2$) and is approximately equal to 0.26.

As we mentioned in the introduction, an alternative to the graphical method is an equivalent arithmetical procedure, the details of which are given below. First note that the maximum vertical distance must occur at or just before one of the steps of $F_{50}(x)$. So we only need to find $F_0(x)$ ($\Phi(z)$ in this case) at each of the data values of $F_{50}(x)$ – using either the bottom or the top of the step in each case, whichever gives the greater distance. The details of these computations are given in Table 4.2. From this table we see immediately that the value of D_{50} is 0.2638.

CRITICAL REGIONS

Since large values of the Kolmogorov–Smirnov statistic D_n indicate that the sample cdf $F_n(x)$ is not similar to the hypothesized cdf $F_0(x)$, the

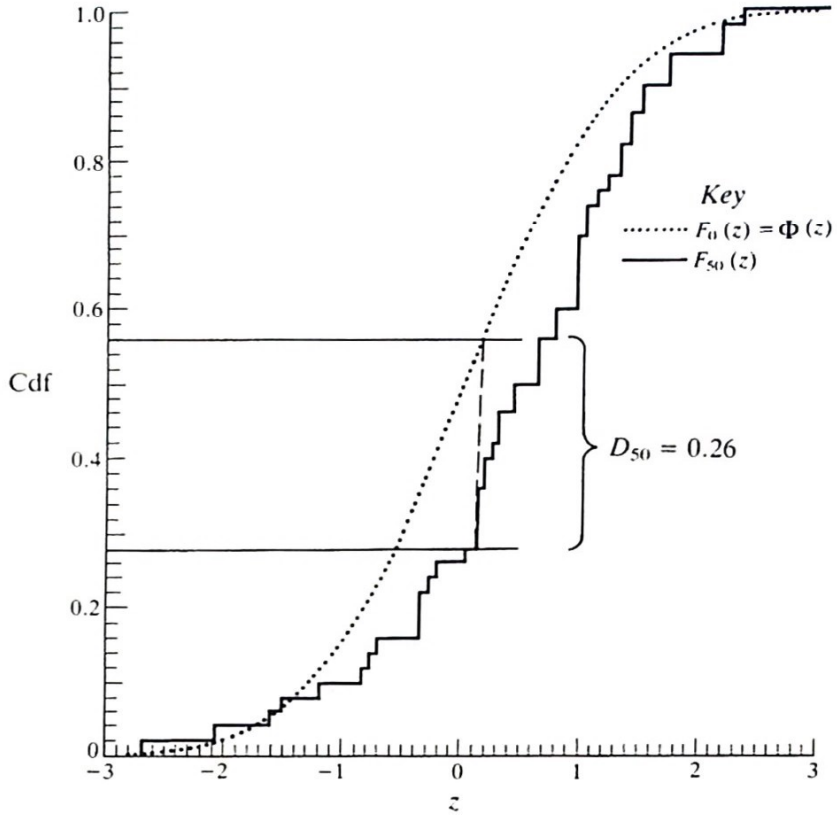


Figure 4.1

critical regions must have the form

$$D_n \geqslant \text{critical value}$$

Consulting Table E (or page 26 of *EST*) we find that for $n=50$, the $\alpha_2=5\%$, 2% and 1% critical regions are $D_{50}\geqslant 0.1884$, $D_{50}\geqslant 0.2107$ and $D_{50}\geqslant 0.2260$ respectively.

SOLUTION

We found that $D_{50}=0.2638$, which falls well inside the $\alpha_2=1\%$ critical region. In other words, we have strong evidence to support the alternative hypothesis that the data are not from the normal distribution N(62,121). A glance at the graph in Figure 4.1 shows that the sample cdf roughly follows what would be expected of a normal distribution, but it seems to be shifted to the right. So one possibility is that the observations are from a normal distribution whose mean is greater than 62.

Discussion

It is worth bearing in mind that the Kolmogorov–Smirnov test does not look *specifically* for changes in location, or dispersion (or in anything so specific). It

Table 4.2

Observation, x	Standardized value, z	$F_{50}(x)$	$\Phi(z)=F_0(x)$	Differences: At step	Differences: Before step
< 32.5		0			
32.5	−2.68	0.02	0.0037	0.0163	0.0037
39.1	−2.06	0.04	0.0197	0.0203	0.0003
44.3	−1.61	0.06	0.0537	0.0063	0.0137
45.4	−1.51	0.08	0.0655	0.0145	0.0055
48.9	−1.19	0.10	0.1170	0.0170	0.0370
52.8	−0.84	0.12	0.2005	0.0805	0.1005
53.4	−0.78	0.14	0.2117	0.0717	0.0917
54.2	−0.71	0.16	0.2389	0.0789	0.0989
58.1	−0.35	0.22	0.3632	0.1432	0.2032
58.9	−0.28	0.24	0.3897	0.1497	0.1697
59.6	−0.22	0.26	0.4129	0.1529	0.1729
62.2	0.02	0.28	0.5080	0.2280	0.2480
63.2	0.11	0.36	0.5438	0.1838	0.2638*
64.0	0.18	0.40	0.5714	0.1714	0.2114
64.7	0.25	0.42	0.5987	0.1787	0.1987
65.3	0.30	0.46	0.6179	0.1579	0.1979
66.6	0.42	0.50	0.6628	0.1628	0.2028
68.8	0.62	0.56	0.7324	0.1724	0.2324
70.5	0.77	0.60	0.7793	0.1793	0.1193
72.2	0.93	0.70	0.8238	0.1238	0.2238
73.1	1.01	0.74	0.8438	0.1038	0.1438
74.0	1.09	0.76	0.8621	0.1021	0.1221
74.9	1.17	0.78	0.8790	0.0990	0.1190
76.0	1.27	0.82	0.8980	0.0780	0.1180
76.8	1.35	0.86	0.9115	0.0515	0.0915
78.0	1.45	0.90	0.9265	0.0265	0.0665
80.4	1.67	0.94	0.9525	0.0125	0.0525
85.1	2.10	0.98	0.9821	0.0021	0.0421
86.3	2.21	1.00	0.9864	0.0136	0.0064

was designed to signal *any* kind of substantial difference in distribution from that proposed in H_0. Sometimes though, as appears to be the case in our current problem, the general shape of the sample cdf is roughly the same as that of the theoretical distribution, in which case it may be possible to be more specific about the kind of difference present.

It should be noted that had we worked in terms of the original x values instead of the standardized z values, we would have finished up with precisely the same picture after an appropriate adjustment in scale, and so would still have obtained precisely the same value of D_{50}.

Also, the Kolmogorov–Smirnov test has given us a much stronger conclusion than did the χ^2 test for the same problem. This follows from examining the levels at which significance has been attained. In the chi-squared test, χ^2 only just fell into the 5% critical region, thus indicating only moderate evidence against H_0. But here, D_{50} has finished up in the 1% critical region, indicating very strong evidence against H_0. This is not really surprising, for as we have already

remarked, the Kolmogorov–Smirnov procedure is specifically designed for dealing with situations where the underlying distribution is continuous.

One-sided alternatives

Although our problem involved a two-sided alternative hypothesis (i.e. comprising all kinds of differences from $F_0(x)$), the Kolmogorov–Smirnov procedure can be easily applied to one-sided alternatives. All we do, instead of considering *all* the differences (positive *and* negative ones), is to consider *only* the positive or the negative differences, *whichever are expected under the one-sided alternative hypothesis*. So, for example, suppose the alternative hypothesis was:

H_1: The population from which the sample is taken has a cdf which is *greater* than the hypothesized cdf, $F_0(x)$, for at least part of the range of possible x values.

For evidence that this particular H_1 is true, we need $F_n(x)$ to be substantially higher than $F_0(x)$ for some value(s) of x. Such an alternative hypothesis might be chosen because it is suspected that the population under investigation has a lower average than the distribution in H_0 or, more generally, that the overall characteristics of the population are such as to give mainly lower values than would be expected under H_0. In such cases, $F_n(x)$ will tend to rise earlier (i.e. for those lower values of x) than $F_0(x)$; consequently the differneces $F_n(x)-F_0(x)$ will tend to be *positive*. The appropriate test statistic which reflects this behaviour is:

$$D_n^+ = max[F_n(x)-F_0(x)]$$

(where the '+' sign in D_n^+ *denotes that* $F_n(x)$ is expected to be greater than $F_0(x)$).

Conversely, suppose the chosen alternative hypothesis were:

H_1: The population being sampled has a cdf which is *less* than $F_0(x)$ for all or at least some values of x.

Then the differences $F_n(x)-F_0(x)$ will tend to be negative. This suggests that the appropriate test statistic is:

$$D_n^- = \max[F_0(x)-F_n(x)]$$

(where the '−' sign in D_n^- denotes that $F_n(x)$ is expected to be less than $F_0(x)$).

In either case, the larger the value of the statistic, the stronger is the evidence for that particular H_1. In other words, critical regions are of the form $D_n^+ \geqslant$ critical value or $D_n^- \geqslant$ critical value, respectively, where the critical values are obtained from Table E (or page 26 in *EST*). As usual, since we now have a one-sided test, it is the α_1 significance levels that are appropriate, so that, for example, the $\alpha_1=5\%$ critical value for a sample of size 50 is 0.1696.

We will illustrate the one-sided version of the Kolmogorov–Smirnov test with the same problem and data as previously for the two-sided case. The null hypothesis is the same as before:

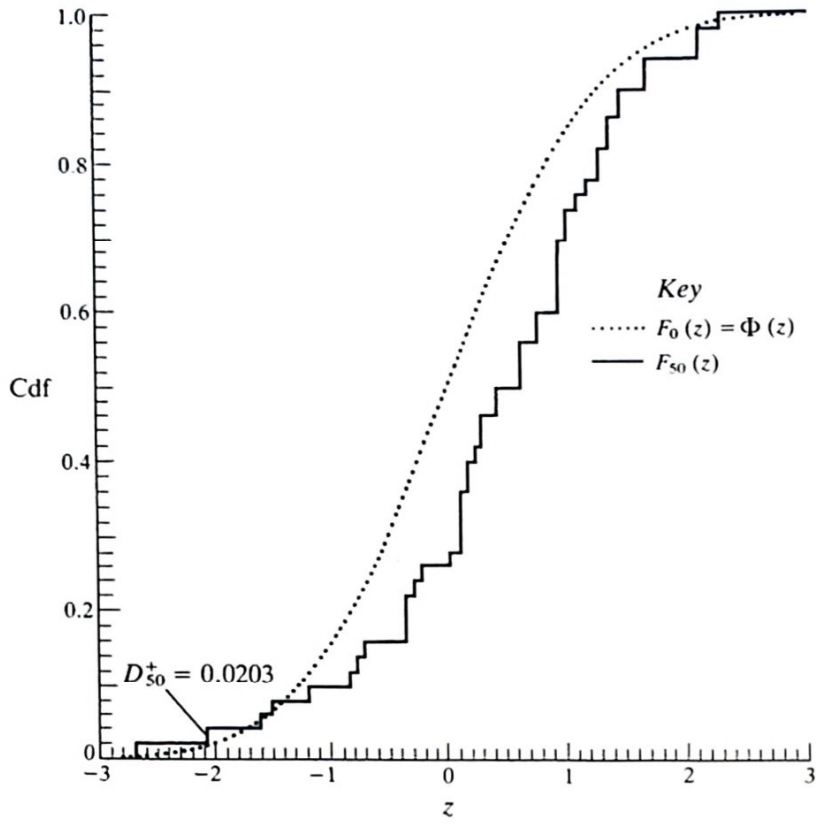


Figure 4.2

H_0: The sample is from a normal distribution with mean 62 and standard deviation 11.

However, whereas previously H_1 comprised essentially *all* other distributions, we can now have a more selective H_1. There are two possibilities. (Imagine, for the moment, that you have not yet seen the data, so that we are still in the early stages of formulating the type of hypothesis test appropriate for particular versions of the practical problem.) First, suppose that in order to make his product more marketable, the manufacturer is considering the possibility of *raising* the permitted number of people in the elevator. He may only do this if there is evidence that adult weights are generally *less* than expected under H_0. This gives rise to the one-sided alternative hypothesis:

H_1: The population being sampled has a cdf which, for some or all values of x, is *higher* than $F_0(x)$, the cdf of N(62,121).

The statistic that would then be expected to give large values when this one-sided H_1 is true is $D_n^+ = \max[F_n(x) - F_0(x)]$. But, of course, when we consider the graphs of $F_n(x)$ and $F_0(x)$ in Figure 4.2, we see that the step function $F_n(x)$ is mostly below $F_0(x)$, i.e. $F_n(x) - F_0(x)$ is mostly *negative*. It does creep above $F_0(x)$ very occasionally but the furthest it gets above $F_0(x)$ is by an amount 0.0203 (as indicated in Fig. 4.2), and so the value of the statistic D_{50}^+ is 0.0203. Obviously this

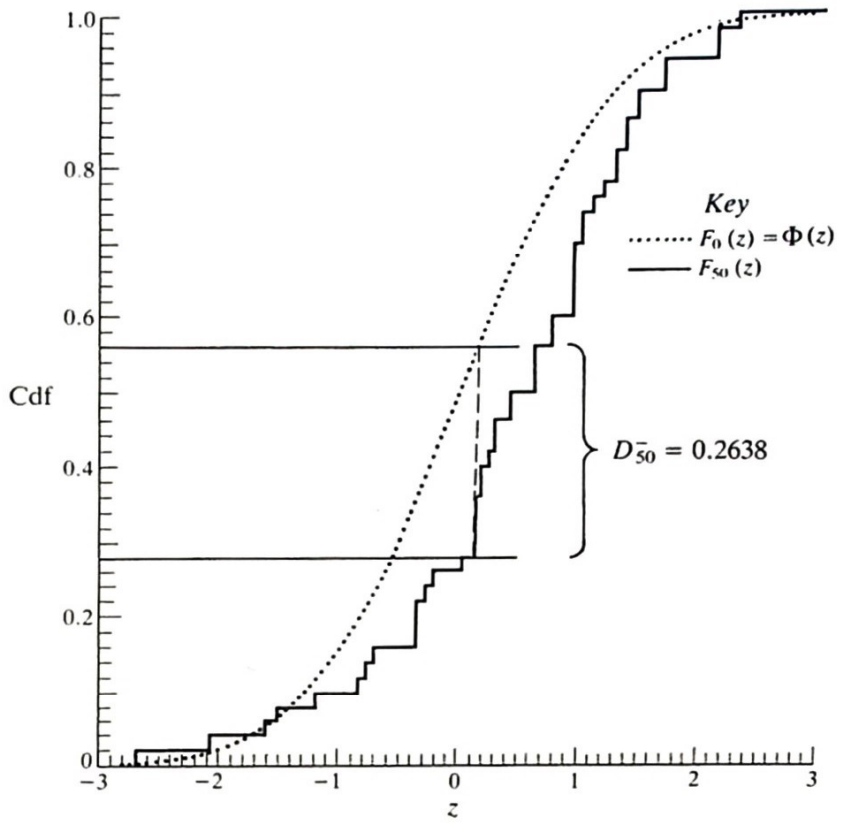


Figure 4.3

is nowhere near the α_1=5% critical region of $D_{50}^+ \geqslant 0.1696$ and so we have no evidence at all to support this one-sided H_1.

However, now consider the different practical situation where the manufacturer is worried that the weights may now be generally *higher* than indicated in H_0, so that he may have to *reduce* the number of people permitted in the elevator. In this case his alternative hypothesis would be:

H_1: The population being sampled has a cdf $F_0(x)$ which, for some or all values of x, is *lower* than that of the normal distribution in H_0.

Now the appropriate test statistic is $D_n^- = \max[F_0(x) - F_n(x)]$, since this is the one which would be expected to be large if this H_1 were true. We are now inspecting those parts of the graphs where the step function $F_{50}(x)$ is *below* $F_0(x)$, i.e. virtually the whole picture! So the value of D_{50}^- is 0.2638, the same as D_{50} in the two-sided case. Figure 4.3 shows the two cdfs and D_{50}^-. However, as usual when comparing one-sided and two-sided versions of tests, the significance level has been halved, with the result that $D_{50}^- = 0.2638$ is well into the $\alpha_1 = \frac{1}{2}\%$ critical region, giving practically conclusive evidence of a general increase in adult weights.

Large sample sizes

It was shown by L.A. Goodman (1954) that the null distribution of $4nD_n^{+2}$ (or $4nD_n^{-2}$) is approximately chi-squared with two degrees of freedom. To see how

this can be used to obtain approximate critical values of D_n^+, D_n^- or D_n, consider $\alpha_1=5\%$. From Table B, the $\alpha_1=5\%$ critical value of the chi-squared distribution with two degrees of freedom is 5.991. Identifying this with $4nD_n^{+2}$, for example, the approximate $\alpha_1=5\%$ critical value of D_n^+ (or D_n^-) is $\frac{1}{2}\sqrt{5.991/n} = 1.2238/\sqrt{n}$.

The table below gives the multipliers of $1/\sqrt{n}$ required to obtain approximate critical values for the usual levels of significance:

α_1	5%	2½%	1%	½%
α_2	10%	5%	2%	1%
	1.2238	1.3581	1.5174	1.6276

To see how good these approximations are, we consider $n=50$, which is the largest sample size in Table E. The above formula approximates the $\alpha_1=5\%$ critical value as $\frac{1}{2}\sqrt{5.991/50} = 0.1731$, compared with the value of 0.1619 given in Table E. So this approximation method, though useful, is not wonderfully accurate for moderate sample sizes, and the reader is recommended to use the table in *EST* on page 26 for all n up to 100.

Discrete distributions

Although specifically designed for continuous distributions, the Kolmogorov–Smirnov test can be used successfully with discrete distributions. However, in this case the tabulated critical values are conservative; for example, a tabulated 5% critical value might in actual fact represent, say, an actual 4% level of significance.

We will briefly illustrate the procedure for discrete distributions with a set of data from 'Is goal scoring a Poisson process?' By D.J. Colwell and J.R. Gillett (1981, *Mathematical Gazette*). In this article, the scores for the 1978–79 season of 22 football teams are analysed to see whether the Poisson distribution could be a reasonable model for the results. The hypotheses for this problem are:

H_0: The number of goals scored by QPR were from a Poisson distribution with mean 1.
H_1: The number of goals scored by QPR were not from a Poisson distribution with mean 1.

The goals per match scored by QPR in the 1978–79 season were as follows:

Goals scored	0	1	2	3	4	5	6+
Frequency	11	23	4	3	0	1	0

The decision rule for this two-sided test is, at a (nominal) 5% significance level, to reject H_0 in favour of H_1 if $D_{42} \geqslant 0.2052$.

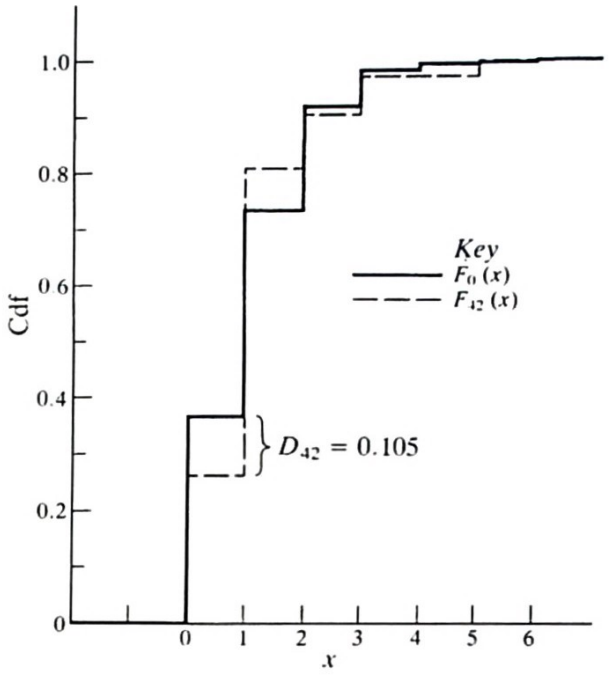


Figure 4.4

The hypothesized cdf, $F_0(x)$, is calculated from

$$F_0(x) = \sum_{r=0}^{x} e^{-1}/r!$$

where $x = 0,1,2, \ldots$ (alternatively the table on page 14 in *EST* may be used to obtain $F_0(x)$). The following table gives the sample cdf $F_{42}(x)$ and $F_0(x)$:

Score	0	1	2	3	4	5	6+
Frequency	11	23	4	3	0	1	0
Cumulative frequency	11	34	38	41	41	42	42
$F_{42}(x)$	0.262	0.810	0.905	0.976	0.976	1	1
$F_0(x)$	0.343	0.710	0.906	0.976	0.995	0.999	1

Figure 4.4 shows the graphs of $F_{42}(x)$ and $F_0(x)$ and also the maximum difference D_{42} equal to 0.105 (approximately). A more accurate value of D_{42} can be obtained from the above table and is $F_0(0)-F_{42}(0)=0.3679-0.2619 = 0.1060$. This value of D_{42} does not lie in the $\alpha_2=5\%$ critical region and so we have no evidence at all to support H_1. (As a matter of interest, the reader might like to check that the chi-squared test also reaches the same conclusion.)

LILLIEFORS' TEST FOR NORMALITY

We have already remarked on the importance of the problem of testing for normality when μ and σ are unknown, and have shown how the chi-squared test may be used in this context. However, there are two big disadvantages of using the chi-squared test for this problem: (a) it involves an arbitrary choice of categories into which the data are classified, and (b) it needs a fairly large sample size.

In its standard form, the Kolmogorov–Smirnov test cannot be used unless the hypothesized distribution is *completely* specified (for example, normal with mean $\mu=62$ and standard deviation $\sigma=11$); it cannot be used for testing the null hypothesis that the population simply has some unspecified normal distribution. However, W. H. Lilliefors (1967) produced a special table of critical values for a Kolmogorov–Smirnov type of test for this more general situation. His procedure was to estimate μ by the sample mean $\overline{X}$, and σ by the 'classical' estimator

$$s = \sqrt{\frac{\sum_{i=1}^{n} (X-\overline{X})^2}{n-1}}$$

(where the sample is denoted by $X_1, X_2, \ldots, X_n$), and then to calculate a D_n statistic with reference to that normal distribution with mean $\overline{X}$ and standard deviation s. (This is, of course, similar in principle to how we adapted the chi-squared test in this general case.) Critical regions for the resulting D_n are obtained from a special table of critical values rather than from the ones we have been using so far. (Unfortunately, Lilliefors' table of critical values was rather inaccurate. A far more accurate table is given in Table F (or on page 27 of *EST*).)

PRACTICAL PROBLEM AND DATA

We will illustrate Lilliefors' test for normality with the same problem and data as before, i.e. the elevator manufacturer.

HYPOTHESES

The manufacturer is now solely concerned with whether or not the data are from some normal distribution. This leads to the more general kind of hypotheses:

H_0: The sample is from some normal distribution.
H_1: The sample is not from a normal distribution.

Here H_1 is a two-sided hypothesis since no interest is being shown as to

whether the cdf $F(x)$ of the population is higher or lower than the hypothesised cdf $F_0(x)$.

CRITICAL REGIONS

As with the Kolmogorov–Smirnov test, critical regions are of the form

$$D_n \geqslant \text{critical value}$$

Table F gives the $\alpha_2=5\%$ and 1% critical regions as $D_{50}\geqslant 0.1246$ and $D_{50}\geqslant 0.1454$ respectively.

SOLUTION

As before, the easiest way to calculate the statistic is to use standardized values, which in this case are calculated using the relation $Z=(X-\overline{X})/s$. For the data of our problem $\overline{X}=66.35$ and $s=11.61$. These standardized values, arranged in ascending order, together with the sample cdf, $F_{50}(x)$, are given in Table 4.3. (Note that F_{50} is the same as before.)

Figure 4.5 shows the graphs of $F_{50}(x)$ and $F_0(x)$. The value of D_{50} is approximately 0.12 and occurs just before $z=-0.27$ (corresponding to

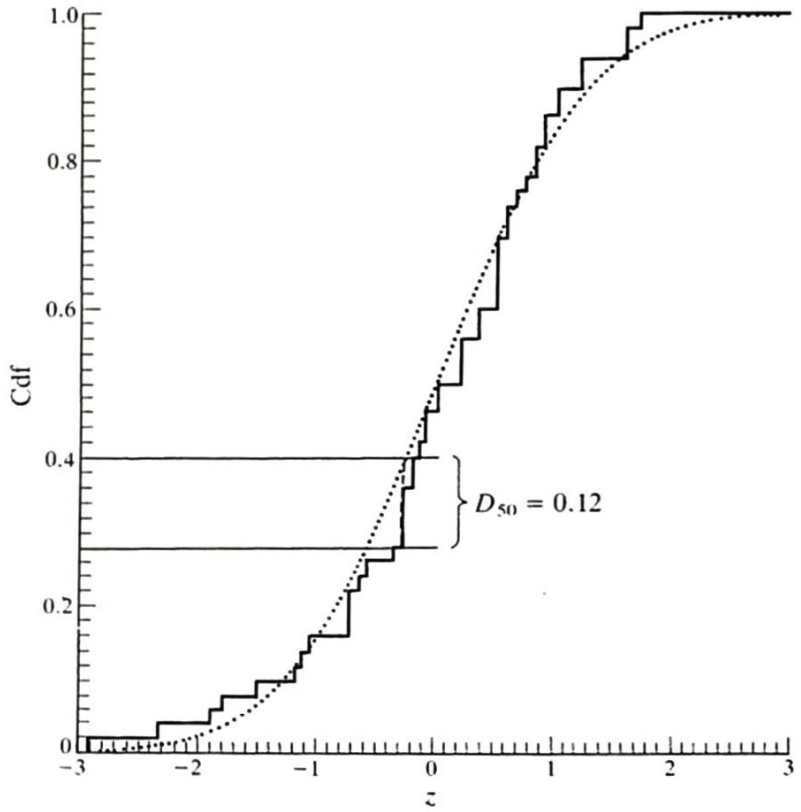


Figure 4.5

Table 4.3

Observation, x	Standardized value, z	$F_{50}(x)$
32.5	−2.92	0.02
39.1	−2.33	0.04
44.3	−1.90	0.06
45.4	−1.80	0.08
48.9	−1.50	0.10
52.8	−1.17	0.12
53.4	−1.12	0.14
54.2	−1.05	0.16
58.1	−0.71	0.22
58.9	−0.64	0.24
59.6	−0.58	0.26
62.2	−0.36	0.28
63.2	−0.27	0.36
64.0	−0.20	0.40
64.7	−0.14	0.42
65.3	−0.09	0.46
66.6	0.02	0.50
68.8	0.21	0.56
70.5	0.36	0.60
72.2	0.50	0.70
73.1	0.58	0.74
74.0	0.66	0.76
74.9	0.74	0.78
76.0	0.83	0.82
76.8	0.90	0.86
78.0	1.00	0.90
80.4	1.21	0.94
85.1	1.61	0.98
86.3	1.72	1.00

x=63.2). Using the normal cdf tables (Table A), a more accurate value of D_{50} is found to be $\Phi(-0.27)-0.28 = 0.3936-0.28 = 0.1136$.

CONCLUSION

The calculated value D_{50}=0.1136 does not lie in the α_2=5% critical region, and so these data certainly contain no evidence of non-normality. It would seem that the company can be fairly confident that, at least on the evidence of the sample, the population of adult weights is something like normally distributed.

Large sample sizes
Approximate critical values may be found by dividing the constants in the following table by $\sqrt{n}$:

α_1	5%	2½%	1%	1%
α_2	10%	5%	2%	½%
	0.8255	0.8993	0.9885	1.0500

So, for example, with $n=50$ (which is the largest sample size in Table F) the $\alpha_2=5\%$ approximate critical value is given by $0.8993/\sqrt{50} = 0.1272$, which is only 0.0026 different from the value given in Table F. Again, if possible, the table on page 27 of *EST* is recommended if n is less than 100.

Assignments

1 Genetics
In 1982, a group of our college students performed an experiment to examine some genetic characteristics of flies. The characteristics were:

(a) length of wing – (i) vestigial and (ii) long;
(b) eye colour – (i) sepia and (ii) red;
(c) body colour – (i) pale and (ii) ebony;

giving eight possible mutations:

(1) long, red and pale;
(2) long, sepia and pale;
(3) long, red and ebony;
(4) vestigial, red and pale;
(5) long, sepia and ebony;
(6) vestigial, sepia and pale;
(7) vestigial, red and ebony;
(8) vestigial, sepia and ebony.

If all these mutations are independent of each other, then the respective proportion of each, according to genetic theory, should be 27 : 9 : 9 : 9 : 3 : 3 : 3 : 1. In fact, the number of flies in each of the eight mutations categories was respectively:

83 21 47 18 14 3 2 1

Task: Apply a goodness-of-fit test to these data to see whether the genetic theory is a reasonable explanation of the results.

2 Bus arrival times

As part of a public transport survey, the times at which buses actually arrived at a particular stop were recorded. The following data give the number of minutes 40 buses were late or early ('early' being denoted by a negative value).

−0.5	0.7	0.1	1.4	0.6	1.6	−0.5	1.6
1.5	2.7	1.2	1.1	0.6	−1.7	1.7	0.2
0.4	0.1	0.9	−2.3	−1.6	5.5	0.9	−1.3
−1.0	0.1	−3.6	1.9	−0.4	−1.6	2.0	0.6
−1.4	−2.9	1.5	−1.3	−0.2	0.9	−1.5	−3.9

Task: Use Lilliefors' test to see whether it is reasonable to suppose that these times could be normally distributed.

Program 4.1 Chi-squared goodness-of-fit test

```
100 REM *** CALCULATION OF CHI-SQUARED GOODNESS-OF-FIT STATISTIC ***

110 REM ** INPUT NUMBER OF CATEGORIES **
120 INPUT "NUMBER OF CATEGORIES ? ";N
130 DIM OB(N),EX(N)

140 REM ** INPUT OBSERVED AND EXPECTED FREQUENCIES **
150 LET L=0
160 LET Z=0
170 PRINT "INPUT THE OBSERVED FREQ. AND THE CORRESPONDING EXPECTED FREQ."
180 FOR I=1 TO N
190 INPUT OB(I),EX(I)
200 IF EX(I)<5 THEN LET L=L+1
210 IF EX(I)=0 THEN LET Z=Z+1
220 NEXT I
230 PRINT

240 REM ** CHECK FOR ZERO AND LOW EXPECTED FREQUENCIES **
250 IF Z=0 THEN GO TO 280
260 PRINT "THERE ARE ";Z;" EXPECTED FREQUENCIES EQUAL TO ZERO!"
270 STOP
280 IF L=0 THEN GO TO 310
290 PRINT "WARNING ... THERE ARE ";L;" EXPECTED FREQUENCIES LESS THAN 5"

300 REM ** CALCULATION OF TEST STATISTIC **
310 Q=0
320 FOR I=1 TO N
330 Q=Q+(OB(I)-EX(I))^2/EX(I)
340 NEXT I
350 PRINT
360 PRINT "VALUE OF THE TEST STATISTIC = ";INT(1000*Q+.5)/1000
```

Program 4.2 Lilliefors' test for normality

```
          PROGRAM 4.2: LILLIEFORS' TEST FOR NORMALITY

100 REM *** LILLIEFORS' STATISTICS FOR NORMALITY ***

110 REM ** INPUT THE SAMPLE SIZE **
120 INPUT "NUMBER OF X-VALUES ? ";N
130 DIM F1(N),F(N),S(N),D(N)
140 LET SUM=0
150 LET SS=0
160 LET TF=0

170 REM ** INPUT VALUES  & FREQUENCIES **
180 PRINT "INPUT THE ";N;" VALUES & THEIR FREQUENCIES"
190 FOR I=1 TO N
200 INPUT D(I),F(I)
210 NEXT I
220 GO SUB 1000

230 REM ** CALCULATION OF SAMPLE CDF, MEAN & SD **
240 FOR I=1 TO N
250 LET TF=TF+F(I)
260 LET S(I)=TF
270 LET SUM=SUM+D(I)*F(I)
280 LET SS=SS+D(I)^2*F(I)
290 NEXT I
300 FOR I=1 TO N
310 S(I)=S(I)/TF
320 NEXT I
330 LET M=SUM/TF
340 LET SD=SQRT((SS/TF-M^2)*TF/(TF-1))
350 FOR I=1 TO N
360 LET D(I)=(D(I)-M)/SD
370 NEXT I

375 REM ** CALCULATION OF THE NORMAL CDF **
380 FOR I=1 TO N
390 LET X=ABS(D(I))
400 LET X1=1/(X*.231642+1)
410 LET X2=(((((X1*.530701-.726578)*X1+.710708)*X1-.142248)*X1+.12742)
*X1)*EXP(-X*X/2)
420 IF D(I)>0 THEN LET F1(I)=1-X2
430 IF D(I)<=0 THEN LET F1(I)=X2
440 NEXT I
450 LET DF=0
460 FOR I=1 TO N
470 LET AD=ABS(S(I)-F1(I))
480 IF AD>DF THEN LET DF=AD
490 LET AD=ABS(S(I-1)-F1(I))
500 IF AD>DF THEN LET DF=AD
510 NEXT I
520 PRINT "VALUE OF LILLIEFORS' STATISTIC = ";DF
530 STOP

1000 REM ** SORT INTO ASCENDING ORDER **
1010 LET N1=N
1020 LET N1=INT(N1/2)
1030 IF N1=0 THEN GO TO 1190
1040 FOR I=1 TO N-N1
1050 FOR J=I TO 1 STEP -N1
```

Continued overleaf

Program 4.2 Continued

```
1060 LET J1=J+N1
1070 IF D(J)>D(J1) THEN GO TO 1100
1080 LET J=0
1090 GO TO 1160
1100 LET T=D(J)
1110 LET T1=F(J)
1120 LET D(J)=D(J1)
1130 LET F(J)=F(J1)
1140 LET D(J1)=T
1150 LET F(J1)=T1
1160 NEXT J
1170 NEXT I
1180 GO TO 1020
1190 RETURN
```

PART III

Tests involving two samples

This part of the book is involved with comparisons of various kinds between two populations, the comparisons being made on the basis of samples drawn from the populations. In each of these four chapters in this part, we present two tests: one that has fairly high power, and one where the emphasis is more on speed and simplicity.

The tests in Chapter 5 are designed for detecting differences in 'location' between the populations, i.e. differences between their averages (in particular, their medians). On the other hand, the tests in Chapter 6 are designed to detect differences in the *spread* (or dispersion) of the populations. Remember that by 'spread' we mean a measure of the variability of observations about the average value; thus if the observations tend to be close to the median then the population has small spread, whereas if many of them tend to be a long way from the median then the population has large spread.

The tests in Chapters 5 and 6 work best if the difference (in location or spread respectively) is the only substantial difference between the populations; in particular, the ability of the tests in Chapter 6 to recognize a difference in spread can be severely impaired if there is a difference in location as well. However, with many processes it is often the case that location and spread are related; if one is high, the other tends to be high as well. A quick test presented in Chapter 7 is particularly effective in this situation, and the other test in that chapter has power to detect just about any kind of substantial difference between the populations.

Chapter 8 deals with data in the form of so-called 'matched pairs'. This is where each observation in one sample has some kind of natural link

with an observation in the other sample. An obvious example is where the two samples consist of observations on the same set of patients before and after some treatment. It turns out that the analysis of such data is quite unlike the analysis of other two-sample situations and, in fact, uses techniques that you have already seen in Part II.

5

Two independent samples – tests for differences in location

THE MANN–WHITNEY TEST

Introduction

This extremely good and widely used test was published by H. B. Mann and D. R. Whitney (1947). As long as the samples are not too large, the computations required are neither excessive nor difficult. Also, the test has high power; indeed, in the special case of sampling from normal distributions having equal variances – the only circumstances when the two-sample *t*-test (which is the standard test from classical statistics for this particular two-sample problem) is valid – the Mann–Whitney test is virtually as powerful as the *t*-test. In addition, as with all the distribution-free procedures, it retains its validity over a vastly broader range of sampling situations.

A version of the Mann–Whitney test was introduced two years earlier by Wilcoxon (1945) in the same article as that in which his signed-rank test (see Ch. 3) first appeared. However, Wilcoxon only dealt with equal sample sizes and used a rank-sum calculation which involves slightly heavier arithmetic than the Mann–Whitney method. Nevertheless, we shall cover the Wilcoxon version also, not only for historic reasons but also because it is closely related to a test for dispersion differences that we shall present in Chapter 6.

PRACTICAL PROBLEM

In 1980 a consumer magazine published a table showing the percentage of new cars, of a wide selection of makes and models, that broke down at least once in the first year's motoring. One aim of the survey was to investigate the truth (or otherwise) of the widely held opinion that Japanese cars are more reliable than British cars. The data for the Japanese models and those models wholly or mainly manufactured in Britain were extracted from the table to enable appropriate comparisons to be made.

HYPOTHESES

Presumably, the null hypothesis H_0 of equal reliabilities of Japanese and British cars implies approximately equal overall percentages of Japanese and British cars suffering breakdowns during their first year. However, this would not be the case if there were some tendency for the usage of the two nations' cars to differ appreciably and, in practice, this question may need considering first. If this is a problem, one might have to collect different data (e.g. the numbers of cars suffering breakdowns during their first 10 000 miles, say). In any case, the data obviously concern only this one particular aspect of reliability, and any conclusions to be drawn from the analysis should of course be limited to this alone.

The survey is investigating the truth of the idea that Japanese cars are more reliable than British ones, and so this, or rather its implication in the current context, is the alternative hypothesis H_1 that we shall use. Therefore, our hypotheses are:

H_0: On average, equal proportions of Japanese and British cars suffer breakdowns during their first year.

H_1: On average, a higher proportion of British cars than Japanese cars suffer breakdowns during their first year.

DATA

The survey included eight Japanese models and 12 British models. The percentages of cars of each type breaking down during their first year were as follows:

Japanese	3	7	15	10	4	6	4	7				
British	19	11	36	8	25	23	38	14	17	41	25	21

Regrettably for the British manufacturers, these data tell their own story without much need for a formal statistical test! Nevertheless, we shall pursue the detailed analysis in order to illustrate the two tests of this chapter; besides, it will be interesting to see whether the weight of evidence formally assessed by the tests agrees with our intuitive interpretation of the data.

STATISTIC

Just suppose for the moment that H_0 were true so that, on average, Japanese and British cars are equally reliable, at least during their first year. Then, if we were to select a Japanese model and a British model at random from the lists available, there should be an even chance as to which will suffer the greater percentage of breakdowns. On the other hand, if Japanese cars really are the more reliable on average then,

although it is of course not certain that the Japanese model selected would fare better than the British model in this particular survey, the odds will be in favour of this happening. So it seems reasonable that, if several such comparisons were made, it must be highly likely that a *majority* of these comparisons would favour the Japanese models. The Mann–Whitney test is based directly on this idea.

For consistency with later work, let us denote the data on the eight Japanese models as sample A and the data on the 12 British models as sample B. The sizes of the samples will be denoted n_A and n_B, so that $n_A=8$ and $n_B=12$ in our problem. The Mann–Whitney test compares every item in sample A in turn with every item in sample B, a record being kept of the number of times, say, that the item from sample A is the greater. In all, there are of course n_An_B comparisons to be made.

Even with the fairly small samples concerned in our example, this procedure would seem at first to involve a rather hefty amount of work – for, with $n_A=8$ and $n_B=12$, we have 96 comparisons to make. Fortunately there is a simple method, which we now describe, that carries out this computation extremely efficiently.

The first task is to arrange all the data into ascending order, though with the samples still retaining their separate identities. An effective way of doing this is to place the values roughly on a linear scale as shown in Figure 5.1.

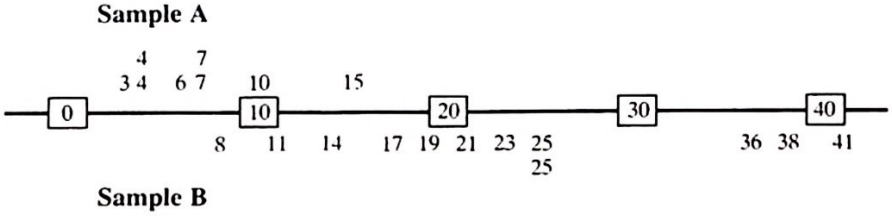


Figure 5.1

It is now easy to write down a list of 'A's and 'B's corresponding to the origins of the numbers in the ordered sequence, thus obtaining the *letter sequence*:

A A A A A A B A B B A B B B B B B B B B

This indicates that the six lowest readings came from sample A, the next lowest from sample B, the next from sample A, the next two from sample B, and so on. Next let us write under each A the number of 'B's that precede it in the sequence:

A	A	A	A	A	A	B	A	B	B	A	B	B	B	B	B	B	B	B
0	0	0	0	0	0		1			3								

Each of these numbers is actually the result of 12 (i.e. n_B) comparisons,

obtained by comparing the corresponding A with every one of the 'B's. The numbers recorded show how many of the B values are exceeded by the A value, i.e. in relation to our particular example, the number of cases in which the Japanese model suffered more breakdowns than a British model. Note especially that these numbers are expected to be *small* if our one-sided alternative hypothesis H_1 is true. Since there are eight (i.e. $n_A=8$) such numbers, we have effectively carried out all the required 96 (i.e. $n_A n_B=8\times12$) comparisons. To determine how many of these 96 comparisons have the A value *exceeding* the B value we simply add together these eight figures, giving a value which we denote by U:

$$U = 0 + 0 + 0 + 0 + 0 + 0 + 1 + 3 = 4$$

Note that the same total would of course have been obtained had we instead written under each B the number of 'A's that *follow* it in the sequence; we would then have computed U as

$$U = 2 + 1 + 1 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 = 4$$

Of course, had the alternative hypothesis been that British cars are more reliable on average than Japanese, then U would have been defined as the number of comparisons in which the B value *exceeds* the A value. This would have been computed either by counting the number of 'A's preceding each B in the letter sequence or the number of 'B's following each A. The important thing is to ensure that U is defined to be *whichever type of count would be expected to be small when H_1 is true.*

CRITICAL REGIONS

Since it is *small* values of U that are significant, critical regions are of the form

$$U \leqslant \text{critical value}$$

The critical values for U are tabulated on page 30 of *EST* for all sample sizes n_1, $n_2\leqslant25$ and also for $n_1=n_2\leqslant50$; here n_1 represents the smaller of n_A and n_B, and n_2 the larger. Table G is an abbreviated version of this table. As we are carrying out a one-sided test in this example, the significance levels are indicated by α_1 in the tables.

Hence, for sample sizes 8 and 12, we see that the critical regions for significance levels $\alpha_1=5\%$, $2\frac{1}{2}\%$, 1% and $\frac{1}{2}\%$ are $U\leqslant26$, $U\leqslant22$, $U\leqslant17$ and $U\leqslant15$ respectively.

SOLUTION

Not unexpectedly, the value of the statistic U obtained from the breakdowns data, i.e. $U=4$, is well inside even the $\frac{1}{2}\%$ critical region. It is

indeed obvious that, if the tables had given details for still smaller values of α_1, this value of U would have remained significant even at much more stringent levels. The Mann–Whitney test has therefore confirmed our initial impressions about the data, and has judged there to be essentially conclusive evidence that the Japanese models have better average reliability (certainly during the first year) than the British models.

WILCOXON'S VERSION OF THE TEST

As with his one-sample test, the statistic for Wilcoxon's version of this two-sample test is computed by adding together certain ranks (which explains why his version is often referred to as the *rank-sum* test). The *ranks* of the observations in the two samples are obtained by simply numbering the letters in the letter sequence from 1 to N, where N is the total number of observations. For our problem, $N = n_A+n_B=20$, and we obtain:

A	A	A	A	A	A	B	A	B	B	A	B	B	B	B	B	B	B	B	B
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

The rank-sum statistic is then simply either

the sum R_A of the ranks of the 'A's

or

the sum R_B of the ranks of the 'B's

The choice between R_A and R_B must be the one that you would expect to be smaller when H_1 is true than when H_0 is true. Our H_1 is that the Japanese cars are the more reliable, which would imply that the 'A's monopolize the start of the letter sequence and the 'B's monopolize the end (which of course does happen in this case) – this will result in R_A being smaller than expected under H_0 and R_B being larger. Thus R_A is the required rank sum in this case; you should check that $R_A=40$.

Special tables of critical values for the rank-sum statistic have been produced elsewhere. However, it is very easy to convert R_A and R_B to the value of U found by the Mann–Whitney method. We shall prove later that

$$U=R_A-\tfrac{1}{2}n_A(n_A+1)$$

or, had R_B been the appropriate rank sum, by

$$U = R_B - \tfrac{1}{2}n_B(n_B+1)$$

In this case, we have $U = R_A - \frac{1}{2}n_A(n_A+1) = 40 - \frac{1}{2}\times 8\times 9 = 40-36 = 4$, as before.

Discussion

As we have said previously, the Mann–Whitney procedure really is an excellent test for the two-sample location problem, being reasonably easy to execute, having power almost reaching that of the classical t-test under normality, and of course being much more widely applicable than the t-test. In fact, it can be shown that, when sampling from other distributions, the Mann–Whitney test is often more powerful than the t-test would be even if valid critical values for t were then available. Perhaps the only real drawback of the Mann–Whitney test is that it does require knowledge of the complete letter sequence; and, if the samples are large, ordering the whole set of observations can be quite a time-consuming operation; use of Program 5.1 is then recommended. However, as we shall shortly see, the other test to be described in this chapter contrasts sharply with the Mann–Whitney test in that it only requires knowledge of what is happening in the very extremes of the letter sequence, so that complete ordering is unnecessary. As a result, this other test is very quick to execute however large the samples may be.

One-sided and two-sided tests

The relation between one-sided and two-sided Mann–Whitney tests is analogous to that with the sign test and Wilcoxon's signed-rank test of Chapter 3. Previously we have seen that, in the one-sided case, U has been defined as either

> the number (U_A, say) of comparisons having the A value less than the B value

or

> the number (U_B, say) with the A value exceeding the B value

according to which is expected to be the *smaller when the one-sided H_1 is true*. With the Wilcoxon rank-sum version, there was a similar choice to be made between R_A and R_B.

However, if the alternative hypothesis is *two-sided*, both eventualities must be catered for, since unusually low values of either U_A or U_B (or equivalently, of either R_A or R_B) would indicate the truth of H_1. In this case, we proceed as follows:

Perform any one of the pairs counts described in the 'Statistic' section above but, for the moment, denote the answer by U^* instead of U. In other words, U^* is obtained by adding together either

(a) the number of 'B's preceding each A (or equivalently, the number of 'A's following each B), or
(b) the number of 'A's preceding each B (or equivalently, the number of 'B's following each A).

Similarly, if using Wilcoxon's version, compute either $R_A - \frac{1}{2}n_A(n_A+1)$ or $R_B - \frac{1}{2}n_B(n_B+1)$, and again call the answer U^*.

In general, there are two possible values that can be obtained for U^*: either

the number of pairs with the A greater than the B (case (a) above)

or

the number of pairs with the B greater than the A (case (b) above)

As you know, the tables of critical values are appropriate only for small values of the statistic. Thus we must ensure that we finish with the smaller of the two possible values of U^*.

Now let us be clever. Note that, having computed one value of U^*, the other possible value must be $n_A n_B - U^*$, since the two values have to add up to the total number of comparisons made, which is $n_A n_B$. Consequently, for whichever value of U^* has been computed, U is defined simply as the smaller of U^* and $n_A n_B - U^*$. Critical regions are then of the form $U \leqslant$ critical value, with the significance levels now being denoted by α_2.

Incidentally, if you are smart, you would probably have chosen to calculate what turns out to be the smaller value of U^* in any case – for the arithmetic is then easier! You can soon see this if you try out computation (b) on the breakdowns data, i.e. one of the counts which gives $U^*=92$. If you need to check whether you have indeed obtained the smaller value of U^*, the easiest way is to compare it with $\frac{1}{2}n_A n_B$ (=48 in our problem); as long as your U^* does not exceed $\frac{1}{2}n_A n_B$, then it is the required value, i.e. U is equal to that U^* value.

How do we deal with ties?

Tied (that is, equal) observations in the same sample cause no difficulty; indeed there are three such pairs of tied values in the breakdowns data, but that has been of no consequence in the above work. However, if ties occur between observations from both samples, the letter sequence is no longer uniquely defined, and so we then need to generalize the definition of the test statistic.

We have noted that, if we are sampling from continuously distributed populations and record the data with sufficient accuracy, we are extremely unlikely to find any ties. However, ties do occur in practice, either because the populations have discrete distributions or, in the case of continuous distributions, because the data may only have been recorded (for instance) to one or two decimal places, or even just to the nearest whole number. We have already seen an example of this situation in Chapter 3 when using the Wilcoxon signed-rank test in the problem of times to process orders. There we found that, although there happened to be no problems with tied observations in the original version of the data, ties did occur when the directions were later rounded to the nearest 15 min. This very phenomenon motivates a very reasonable method of dealing with ties, not only in the Mann–Whitney test but quite generally. Imagine that the data are remeasured with sufficient accuracy to break all the ties. It is surely possible to claim that all possible ways of breaking the ties are about equally likely. Thus, if we consider the set of letter sequences obtained by breaking the ties in all possible ways, and calculate the statistic for each, the mean of the values