EXAMPLE 1

A certain computer program is supposed to furnish random digits. If the program is accomplishing its purpose, the computer prints out digits (2, 3, 7, 4, etc.) that seem to be observations on independent and identically distributed random variables, where each digit 0, 1, 2, . . . , 8, 9 is equally likely (probability 0.1) to be obtained. One way of testing

H_0: The numbers appear to be random digits

against the alternative

H_1: Some digits are more likely than others

is to count how many times each digit appears. Three hundred digits are generated with the following results.

1578748416	4705188926	6936349612
4653843213	0282868892	3928057043
5101259393	9837006785	3011679938
7122863085	6528271107	2956427027
2671728075	9759178719	9373309535
8363265100	2546793732	2212122529
9453087720	3976759377	9593511031
5605373242	1819898287	3872181027
3494768396	9296177240	8620774591
4659773922	9246724287	8326143939

Each digit is equally likely under the null hypothesis, so the expected number of occurrences is 30 for each of the ten digits. But the digit 2 occurs 41 times while the digit 4 occurs only 19 times. Is this what one could expect from random fluctuation?

The complete list of observed counts is as follows:

Digit:	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*Total*
Observed Frequency	22	28	41	35	19	25	25	40	30	35	300
Expected Frequency	30	30	30	30	30	30	30	30	30	30	300

The test statistic

$$T = \sum_{i=1}^{10} \frac{O_i^2}{E_i} - N = 317 - 300 = 17 \tag{4}$$

is compared with the 0.95 quantile of a chi-squared random variable with 9 degrees of freedom, which is given in Table A2 as 16.92. Therefore the null

of the grouped data, and the resulting k equations are solved simultaneously for the k unknown parameters. The following example and the subsequent comment section should help to clarify the above procedure.

EXAMPLE 2

Efron and Morris (1975) presented data on the first 18 major league baseball players to have 45 times at bat in 1970. The players' names and the number of hits they got in their 45 times at bat are given as follows.

Clemente	18	Kessinger	13	Scott	10
F. Robinson	17	L. Alvarado	12	Petrocelli	10
F. Howard	16	Santo	11	E. Rodriguez	10
Johnstone	15	Swoboda	11	Campaneris	9
Berry	14	Unser	10	Munson	8
Spencer	14	Williams	10	Alvis	7

We will test the null hypothesis that these data follow a binomial distribution with $n = 45$. But first we need to estimate $p = P(\text{hit})$ for each time at bat.

A good estimate for p is the overall relative frequency of getting a hit based on these data.

$$\hat{p} = \frac{\text{total number of hits}}{\text{total number of at-bats}} = \frac{215}{810} = 0.2654 \tag{5}$$

Using $n = 45$ and $p = 0.2654$ the binomial probabilities are calculated,

$$P(X = i) = \binom{45}{i} (0.2654)^i (0.7346)^{45-i} \qquad i = 0, 1, \ldots, 45 \tag{6}$$

The expected cell counts are

$$E_i = 18 \cdot P(X = i) \qquad i = 0, 1, \ldots, 45 \tag{7}$$

Cells with expected values less than 0.5 are combined to avoid problems of having a poor approximation by the chi-squared distribution. The resulting cells, after combining, are given as follows.

	No. of hits												
	≤7	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*≥18*	*Total*
Observed	1	1	1	5	2	1	1	2	1	1	1	1	18
Expected	1.10	1.06	1.57	2.04	2.35	2.40	2.20	1.82	1.36	0.92	0.57	0.61	18

The test statistic is

$$T = \sum_{i=1}^{12} \frac{O_i^2}{E_i} - N = 24.73 - 18 = 6.73 \tag{8}$$

which is compared with the 0.95 quantile from the chi-squared distribution with $12 - 1 - 1 = 10$ degrees of freedom, which is given by Table A2 as 18.31, so the null hypothesis is accepted at $\alpha = 0.05$. In fact, a comparison of the observed value 6.73 with quantiles from Table A2, 10 degrees of freedom, shows the p-value to be much larger than 0.25, so the fit to the binominal distribution is quite good. Note that 1 degree of freedom was subtracted because the parameter p was estimated from the data. ■

Comment

The parameter p in Example 2 was estimated using the total number of hits divided by the total number of times at bat for all 18 players. This is a good estimator, but it may not be the one that minimizes the value of T, in accordance with the asymptotic theory that allows 1 degree of freedom to be subtracted because one parameter is estimated. Two comments need to be considered here.

First, the p-value is already much larger than 0.25 in Example 2, and the null hypothesis is easily accepted. There is no need to find the minimum value of T, which will further increase the p-value. The conclusion will remain the same, that the data follow the binomial distribution fairly well. Therefore the extra work required to find the minimum chi-squared statistic is not necessary unless the p-value is small and the decision is in doubt.

Second, the theory that justifies subtraction of 1 degree of freedom for each parameter estimated using the minimum chi-squared method is an asymptotic theory, as the sample size goes to infinity and the expected values in each cell also go to infinity. This by itself is no guarantee that the minimum chi-squared method results in a more accurate approximation for small sample sizes, the kind we encounter in real samples from the real world. Therefore we can be comfortable using the usual estimators for unknown parameters, such as moment estimators, or maximum likelihood estimators, knowing that for the sample being examined the chi-squared approximation may be as good as if the minimum chi-squared method were being used. For more discussion of this topic see Yule and Kendall (1950), Chernoff and Lehmann (1954), and Berkson (1980).

The chi-squared goodness-of-fit test is not limited to discrete random variables, as the previous two examples might suggest. It can also be used to test whether the data come from a specified continuous distribution, where some of the unknown parameters may be estimated from the data as in Example 2. The first step is to "discretize" the continuous random variable by forming intervals, which then become the classes described in the test. The number of observations in each interval O_j is compared with the expected number in each interval

$$E_j = N \cdot P(X \text{ is in interval } j) \tag{9}$$

when the null hypothesis is true.

The following example illustrates the chi-squared goodness-of-fit test to a continuous distribution where two parameters are estimated from the data. Note that the formation of intervals is somewhat arbitrary, and is therefore a weakness in applying the chi-squared goodness-of-fit test to any continuous distribution.

EXAMPLE 3

Fifty two-digit numbers were drawn at random from a telephone book, and the chi-squared test for goodness of fit is used to see if they could have been observations on a normally distributed random variable. The numbers, after being arranged in order from the smallest to the largest, are as follows.

23	23	24	27	29	31	32	33	33	35
36	37	40	42	43	43	44	45	48	48
54	54	56	57	57	58	58	58	58	59
61	61	62	63	64	65	66	68	68	70
73	73	74	75	77	81	87	89	93	97

The null hypothesis is

H_0: These numbers are observations on a normally distributed random variable

The normal distribution has two parameters (Definition 1.5.3), both of which are unspecified by H_0, and must be estimated before the goodness-of-fit test may be applied. For illustration, the procedure is divided into steps.

Step 1 *Divide the observations into intervals of finite length.* We arbitrarily choose the intervals, 20 to 40, 40 to 60, 60 to 80, and 80 to 100, not including the upper limit of each interval.

	Interval				
	20 to 40	**40 to 60**	**60 to 80**	**80 to 100**	**Total**
Number of Observations	12	18	15	5	50

Step 2 *Estimate μ and σ with the sample mean $\overline{X}$ and sample standard deviation S of the grouped data.* The 12 observations in the interval 20 to 40 are treated as if they all equal the middle point 30. The 18 observations from 40 to 60 are all considered to be 50, and so on. These are the numbers used for computing $\overline{X}$ and S, using the equations of Definition 2.2.3.

$$\begin{aligned}\overline{X} &= \frac{1}{N}\sum_{i=1}^{N} X_i \\ &= \frac{1}{50}[12(30) + 18(50) + 15(70) + 5(90)] \\ &= 55.2\end{aligned} \tag{10}$$

$$\begin{aligned}S = \sqrt{S^2} &= \sqrt{\frac{1}{N}\sum_{i=1}^{N} X_i^2 - \overline{X}^2} \\ &= \left\{\frac{1}{50}[12(30)^2 + 18(50)^2 + 15(70)^2 + 5(90)^2] - (55.2)^2\right\}^{1/2} \\ &= 18.7\end{aligned} \tag{11}$$

Therefore our estimates of μ and σ are 55.2 and 18.7, respectively.

Step 3 *Using the estimated parameters from Step 2, compute the E_js for the groups in Step 1 and for the "tails."*

Class Boundaries b_j	$(b_j - \overline{X})/S = x_p$	$F(x_p)$	*Interval*	p_j^*
$b_1 = 20$	−1.88	0.03	<20	0.03
$b_2 = 40$	−0.813	0.21	20 to 40	0.18
$b_3 = 60$	+0.256	0.60	40 to 60	0.39
$b_4 = 80$	+1.33	0.91	60 to 80	0.31
$b_5 = 100$	+2.40	0.99	80 to 100	0.08
			≥ 100	0.01

To find the hypothesized probabilities of being in the various classes, when the hypothesized distribution is the normal distribution with mean 55.2 and standard deviation of 18.7, the class boundaries (column 1 in the table) are considered to be the quantiles of the hypothesized distribution. These quantiles are converted to quantiles of a standard normal random variable (column 2) by Equation 1.5.3 in order to find out which quantile the boundaries represent (column 3). Subtraction of the items in column 3 then yields the probabilities p_j^* of being in the various intervals under the hypothesized distribution. The E_js equal 50 p_j^*, from Equation 1, and are given below.

	Class					
	<20	20–40	40–60	60–80	80–100	≥100
Expected Number E_j	1.5	9.0	19.5	15.5	4	0.5
Observed Number O_j	0	12	18	15	5	0

Because of the small E_js, the first and last cells are combined with the cells adjacent to them.

	Class			
	<40	40–60	60–80	≥80
Expected Number E_i	10.5	19.5	15.5	4.5
Observed Number O_i	12	18	15	5

Step 4 *Compute* T. The test statistic is now computed using Equation 2.

$$T = \frac{(12-10.5)^2}{10.5} + \frac{(18-19.5)^2}{19.5} + \frac{(15-15.5)^2}{15.5} + \frac{(5-4.5)^2}{4.5}$$
$$= 0.401 \tag{12}$$

The critical region of size 0.05 corresponds to values of T greater than 3.841, the 0.95 quantile of a chi-squared random variable with $c - 1 - k = 4 - 1 - 2 = 1$ degree of freedom. Therefore H_0 is accepted, with a p-value well above 0.25.

Usually a modification called Sheppard's correction is used when the variance is being estimated from grouped data and when the interior intervals are of equal width, say h. Sheppard's correction consists of subtracting $h^2/12$ from S^2 in order to obtain a better estimate of variance. In this example $h = 20$ (the width of each interval), so $(20)^2/12 = 33.33$ could have been subtracted in Step 2 before extracting the square root. The result is $S = 17.8$, a smaller estimate for σ. This smaller estimate of σ results in a larger value of T in this example and, since our objective is to obtain estimates that give the smallest possible value for T, the correction was not used. In most situations we can expect a smaller T when the correction is used.

Another peculiarity of this example is the fact that a smaller value of T (0.279) may be obtained by using $\bar{X} = 55.04$ and $s = 19.0$ as estimates of μ and σ. These estimates are the sample moments obtained from the original observations, before grouping. No matter how they are obtained, the estimates to use are the estimates that result in the smallest value of T. The procedure described in this example can be relied on to provide a value of T not far from its minimum value in most cases. Therefore it is the recommended procedure. ■

In the preceding example the test statistic just happened to be smaller when μ and σ were estimated using the sample mean and standard deviation based on the original, rather than the grouped, observations. This procedure may be used and is even recommended by Yule and Kendall (1950), but it is usually