Chapter 7

Chi-Square Tests

7.1 Introduction

The two sections of this Chapter correspond to the two different kinds of test based on the Chi-square distribution. Although we have already met this distribution in Chapter 6 concerning inference on the population variance based on a Normal random sample, the tests here do not depend on *any* distributional assumption (other than random sampling). Indeed the first kind of Chi-square test can be used to test the distributional assumptions themselves, although not very powerfully in the case of the Normality assumption.

Computing for this Chapter You are *not* expected to use MINITAB for this part of the course. Nearly all calculations can easily be performed by hand or using a calculator, so MINITAB does not serve here. The TABLE command allows one to construct a table of frequencies and perform a chi-square test on them, but not to store them for further manipulation. The CHISQUARE command operates on columns of frequencies but unless the data appears as frequencies these columns have to be obtained by hand. Use of COPY and TALLY will produce these frequencies.

7.2 Goodness-of-Fit Tests

The question we are attempting to answer here is: Do the data fit the assumed or postulated distribution? More precisely we ask: Is there any evidence against the null hypothesis H_0, say, that the data are a random sample1from some particular distribution, such as the Normal or Exponential (for continuous variables) and the Binomial or Poisson (for discrete variables.) Note that no particular alternative is specified: often H_0 represents the simplest explanation of the data and only if H_0 is rejected with a low P-value would we consider elaborating our original hypothesis. This is following the philosophical principle of *Occam's Razor* or the principle of *Parsimony:* seeking the simplest explanation of variation.

Example 10 :In 116 randomly selected families with two children, 42 have no girls, 52 have one girl and only 22 have two girls. Assuming births of either sex are equally likely, do these data conflict with the hypothesis that the sexes of successive births are independent? If the hypothesis is true, then the number of girls in any family of two children follows a Binomial distribution with parameters $n_0 = 2$ and p = 1/2 (we reserve *n* for the sample size, the number of families, here 116). We can construct a *table of frequencies:*

Number of girls *k*	0	1	2	Total
Observed Frequency *Ok*	42	52	22	116
Probability (under H_0)	$(1/2)^2$	2 x 1/2 x ½	$(1/2)^2$	1
Expected Frequency *Ek*	29	58	29	116

where the expected frequencies are obtained by multiplying the total frequency or sample size by the probability under H_0. *They need not be integers.*

Do the discrepancies between observed and expected frequencies provide sufficient evidence to cast doubt on H_0?
One of *two* possible test statistics studied here, and the most common one, is *Pearson's* X^2 *statistic:*

$$X^2 = \sum_k \frac{(O_k - E_k)^2}{E_k}$$

The statistic is also called, somewhat confusingly, (Pearson's) 'Chi-Square(d)'. Clearly large values of X^2 indicate departures from H_0. We can show that, for large n, X^2 approximately follows a chi-square (sampling) distribution with degrees of freedom equal to the number of classes minus one (but see later).

Example 10 revisited

$$X^2 = \frac{(42-29)^2}{29} + \frac{(52-58)^2}{58} + \frac{(22-29)^2}{29} = 8.14$$

Referring this observed value to Table 3, we see that $\chi_2^2(0.025) = 7.38$, $\chi_2^2 = (O.Ol0) = 9.21$ so P, which is as usual the probability of obtaining a value at least as extreme as the one actually observed, lies in the range
$0.010 < P < 0.025$. Thus there is moderate evidence *against* the independence of the sexes of successive births, if we assume that births of either sex are equally likely.

We can test this assumption also by testing the broader hypothesis that the number of girls follows a Binomial distribution with unspecified p, that

is the independence hypothesis *per se.* We must estimate p from the data:

$$\hat{p} = \frac{total\ no\ of\ girls}{total\ no\ of children} = \frac{2\times 22 + 52}{2\times 116} = \frac{96}{232}$$

We now substitute this estimated value of p to calculate new probabilities and resulting expected frequencies but now it can be shown that the best approximation to the sampling distribution of X^2 is provided again by the chi—square distribution, but now with degrees of freedom reduced by one (as we have estimated a single parameter.) The test itself is left as an exercise.

Example 11 The number of accidents in a month is observed over a period of ten years. If these accidents occur randomly and at a uniform rate, then the data should follow a Poisson

distribution. The data are

Number of Accidents	Observed Frequency O_k	Probability	Expected Frequency E_k
0	41	0.30119	36.14
1	40	0.36144	43.37
2	22	0.21686	26.02
3	10	0.08674	10.41
4	6	0.02602	3.12
5	0	0.00625	0.75
6	1	0.00125	0.15
7 or more	0	0.00025	0.03
Total	120	1.0000	120

The above table contains both the observed data and the probabilities and expected frequencies calculated under the Poisson assumption. The unknown population parameter which is the mean number of accidents \ per month is estimated by

$$\hat{\lambda} = \bar{x} = \frac{(41\times 0) + (40\times 1) + ... + (0\times 7)}{120} = \frac{144}{120} = 1.20$$

so we can use Table 2 without interpolation to calculate the probabilities in the third column. The observed value of Pearson's X^2 is then referred to Table 8.

There is, however, a further complication regarding the approximation of the sampling distribution of X^2 by the chi-square distribution. The approximation for large *n* is valid only if the *expected* frequencies are sufficiently large. Just how large depends on the accuracy of approximation desired, the sample size and the distribution under test. As we require a probability (the P-value) accurate to say, only three significant figures, we will adopt *two* alternative conditions, either of which must be satisfied to make the approximation 'valid'.

The first condition (Condition *A)* states that *all* expected frequencies must be 5 or more and that to satisfy it, merging of (neighbouring) classes may be necessary.

Example 11 revisited To satisfy Condition *A*, we must merge the last *five* classes so that the last class is '3 or more' with *Ok* = 17 and *Ek* 14.46. Thus

$$X^2 = \frac{(41-36.14)^2}{36.14} + \frac{(40-43.47)^2}{43.47} + \frac{(22-26.02)^2}{26.02} + \frac{(17-14.46)^2}{14.46}$$

The degrees of freedom **ii** of the appropriate chi-square distribution are given by:
ν = no. of classes — no. of parameters estimated — 1

where the classes are counted *after* merging.

Example 11 revisited: $\nu = 4-1-1 = 2$ degrees of freedom. Table5 gives

$\chi^2_2(0.500) = 1.38629,$ and $\chi^2_2(0.250) = 2.7759$ so $0.25 < P < 0.5$ and there is no evidence against the null hypothesis of a Poisson distribution, that is no evidence against the hypothesis that accidents occur randomly.

The second, less stringent condition (Condition *B)* states that not more than 20% of all the *Ek* can be less than 5, but that *no* E_k must be less than 1.

Example 11 revisited : To satisfy Condition *B*, we need only merge the last *four* classes so the last class is now '4 or more' with $O_k = 7$ and $E_k = 4.05 < 5$, and this class represents exactly 20% of the new number of classes. Now $X^2 = 3.72$ on v = 5 — 1 — 1 = 3 degrees of freedom so again $0.5 > P > 0.25$ and there is no evidence against randomness. An alternative statistic to X^2 is the *log-likelihood ratio statistic:*

$$Y^2 = 2\sum_k O_k \ln(O_k / E_k)$$

Like X^2, Y^2 has a large sample chi-square distribution with the *same* degrees of freedom. The *same* merging rules should be obeyed when calculating Y^2.
Often the observed values of the two statistics are very close. You should be able to use either.
Exercise Expand Y^2 using Taylor's theorem to get X^2 as the first approximation.

7.3 Tests of Independence

Here we ask: Is there any evidence in the data for *association* between two *categorical* variables? This question is answered by a (chi-square) test of Independence but the tests of Homogeneity and Similarity are formally identical so we will deal with all three together.

Example 12 A survey of smoking habits in a sixth form sampled 50 boys and 40 girls at random and the frequencies were noted in the following *contingency table:*

	Non-Smokers	Light Smokers	Heavy Smokers	
Boys	16	20	14	50
Girls	24	10	6	40
Total	40	30	20	90

Is there evidence of differences between the sexes? We are comparing two distributions (over smoking habits) so test is one of SIMILARITY. The null hypothesis is that the population proportions of boys and girls in each smoking category are the same.

Example 13 : In a study of migrant birds, nestlings were 'ringed' in four different locations *A*— *D.* One year later, birds were recaptured at each location and the number of ringed birds noted. The data were:

	A	B	C	D	Total
Recovered	30	75	24	31	160
Not Recovered.	150	225	63	202	640
Total	180	300	87	233	800

Is there evidence for differences in the four recovery rates? We are comparing four proportions so the test is one of HOMOGENEITY. The null hypothesis is that the proportion of recovered birds is the same for the four locations.

Example 14 227 randomly selected males were classified by eye and hair colour:

		Red/Fair	Brown	Black	Total
EYES	Blue	65	26	8	99
	Grey/Green	32	41	24	97
	Brown	5	16	10	31

Is there evidence for association, or lack of independence, between the two factors (at three levels)? Do the proportions (or probabilities) of the three eye colours differ among the sub-populations comprising the three hair colours? Equivalently do the proportions (or probabilities) of the three hair colours differ among the three eye colours? This is a test of INDEPENDENCE, sometimes (confusingly) called a test of association. The null hypothesis is that for each pair of eye and hair colours

P(eye colour and hair colour) = P(eye colour) x P(hair colour)

The data from all three of the above examples have the general form of an *r x c contingency table* with *r* rows, *c* columns, (both with appropriate labels), row and column totals, with the observed frequencies O_k in the 'cells' of the table. The overall sample size appears as an overall total in the bottom right hand corner. In order to use Pearson's X^2 or the log-likelihood ratio statistic Y^2. we need to calculate expected frequencies E_k under the null hypotheses.
In each case we use

$$E_k = \frac{\text{row total x column total}}{\text{overall sample size (n)}}$$

Example 12 revisited The method of sampling fixes the row totals and under H_0

$$\frac{E_k}{RowTotal} = \frac{ColumnTotal}{n}$$

i.e. the row proportions or probabilities are the same for each row.

Example 13 revisited The method of sampling fixes the column totals and under H_0

$$\frac{E_k}{Column\ \ Total} = \frac{Row\ Total}{n}$$

i.e. the column proportions are the same for each column.

Example 14 revisited The method of sampling fixes only the overall sample size and under H_0

$$\frac{E_k}{n} = \frac{Row\,total}{n} \times \frac{Column\,Total}{n}$$

For large *n*, the distribution of X^2 and Y^2 is approximately chi-square with ν degrees of freedom determined by

ν = (no. of rows — 1)(no. of cols. — 1)

As with goodness-of-fit tests the approximation is valid only if either Condition *A* or Condition *B* holds. If merging is necessary we should merge only complete rows or columns.

Example 12 revisited The table of expected frequencies is

	Non-Smokers	Light Smokers	Heavy Smokers	Total
Boys	22.2	16.7	11.11	50.0
Girls	7.8	13.3	8.9	40.0
Total	40.0	30.0	20.0	90.0

For example the top left hand cell is 22.2 = $\frac{50 \times 40}{90}$ The observed value of X^2 is (summing over cells reading across rows) 1.73 + 0.65 + 0.76 + 0.94 + 2.16 + 0.82 = 7.06, which we refer to Table 8 with ν = (2 —1) x (3 — 1) = 2: χ_2^2 (0.05) = 5.99146 and χ_2^2 (0.025) = 7.37776 so 0.025 <*P* <0.05 and there is evidence that smoking habits differ between the sexes. For a. 2 x 2 table

a	b	a+b
c	d	c+d
a+c	b+d	

straightforward algebra shows that

$$\chi^2 = \frac{(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

but a better approximation (of the distribution of X^2 to a χ^2 distribution) is achieved by using continuity correction.

$$\chi^2 = \frac{(|\,ad - bc\,| - 1/2)^2\, n}{(a+b)(c+d)(a+c)(b+d)}$$

Example 12 revisited: If we collapse the original 2 x 3 table into 2 x 2

	Non Smokers	Smokers	Total
Boys	16	34	50
Girls	24	16	40
Total	40	50	90

Then, using the formula with continuity correction, $X^2 = 7.04$ which on referred to χ_1^2 gives $0.005 < P < 0.01$, very strong evidence for differences originally observed seem to be due to the incidence of smoking rather than the amount of it, once an individual is a smoker.

7.4 Exercise

(1) Four seeds were planted in each of one hundred pots under identical conditions as part of an experimental investigation into seed germination. After a fixed period of time the number of seeds germinating in each pot was noted and the frequency table was as follows:

No. of seeds germinating	0	1	2	3	4
Number of pots	12	24	39	22	3

If seeds germinate independently under these conditions then the number germinating should follow a Binomial distribution. Test this hypothesis using a goodness-of-fit statistic. Use an alternative statistic and comment on the results of both your tests.

(2) (a) In the migrant birds study (Example 13) test the hypothesis that the probability of recovering a ringed bird after one year is constant over the four different locations.

(b) Test the hypothesis that eye and hair colour are independent characteristics in the survey of British males (Example 14).