Chapter 11

Chi–square Tests

11.1 Introduction

In this chapter we will consider the use of chi–square tests (χ^2–tests) to determine whether hypothesized models are consistent with observed data. These tests are based on the χ^2–square statistic which serves as an index of discrepancy between a collection of observed frequencies and a hypothesized collection of expected frequencies. The χ^2–statistic summarizes the differences between the values actually observed and the values we would expect to see if the hypothesized model was correct; with a large χ^2 value indicating that the hypothesized model is not consistent with the observed data. The first step in forming the χ^2–statistic is to find the observed frequencies with which each possible value occurs in the data and the expected frequencies with which each possible value should occur according to the hypothesized model. For each value the difference between the observed frequency and the expected frequency is computed, this difference is then squared and this squared difference is divided by the corresponding expected frequency. These standardized squared differences are then added yielding the χ^2–statistic

$$\chi^2 = \sum \frac{(\text{observed frequency } - \text{ expected frequency})^2}{\text{expected frequency}} ,$$

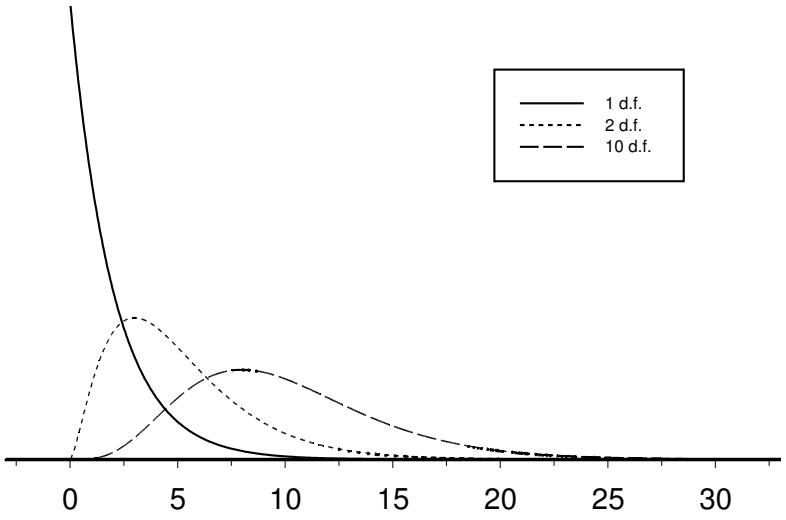
where the sum is over all of the possible values. Large values of this χ^2–statistic indicate evidence that, at least some of, the observed frequencies do not agree with the hypothesized expected frequencies and thus that the hypothesized model may not be correct. That is, large values of the χ^2–statistic indicate that the observed data are not consistent with the hypothesized model.

The χ^2 distributions are skewed to the right with density curves which are positive only for positive values of the variable. Density curves for representative χ^2 distributions are provided in Figure 1. The χ^2 distributions for 1 and 2 degrees of freedom have their mode at zero; for larger degrees of freedom (d.f.) the mode of the χ^2 distribution is located at d.f. $- 2$. Notice that the variability in the χ^2 distribution increases as the degrees of freedom increases. For the χ^2–tests discussed in this chapter a large value of the χ^2–statistic constitutes evidence against the null hypothesis and the P–values for these tests are areas under the appropriate χ^2 density curve to the right of the observed value χ^2_{calc} of the χ^2–statistic.

The χ^2–tests and associated P–values discussed in this chapter are based on large sample approximations which require reasonably large expected frequencies. One rule of

thumb regarding this requirement says that no more than 20% of the expected frequencies should be less than 5 and all of the expected frequencies should be at least 1. If these conditions are not satisfied, you can combine some categories (values of the variable) to increase the expected frequencies which are too small.

Figure 1. Chi–square distribution density curves.



We will consider three different applications of χ^2–tests in this chapter. In Section 11.2 we will consider χ^2–tests for goodness of fit. These goodness of fit tests can be viewed as extensions of the Z–test for $H_0 : p = p_0$ versus $H_1 : p \neq p_0$ of Chapter 5 to populations with more than two possible classifications. In Section 11.3 we will consider χ^2–tests for homogeneity. These tests of homogeneity can be viewed as extensions of the Z–test for $H_0 : p_1 = p_2$ versus $H_1 : p_1 \neq p_2$ of Chapter 6 to two or more populations with two or more possible classifications. Finally in Section 11.4 we will consider χ^2–tests for independence. These tests for independence are used to examine the relationship between two or more classification factors.

Throughout this chapter we will provide details of the computations involved in computing χ^2–statistics. This does not indicate that you should compute these statistics by hand; however, if you choose to do so be sure to avoid rounding at intermediate stages. Some calculators and most statistical programs will compute χ^2–statistics, associated P–values, and other relevant information.

11.2 Chi–square Tests for Goodness of Fit

A χ^2–test for goodness of fit is used to determine whether the outcomes predicted by a hypothesized model are consistent with observed data. The hypothesized model is used to determine the outcomes we would expect to observe and the χ^2–statistic is used to quantify the agreement between the observed outcomes and the expected outcomes. A

small value of the χ^2–statistic indicates that the observed outcomes are in agreement with the outcomes predicted by the hypothesized model (the data are consistent with the model) and a large value indicates inconsistency between the observed data and the hypothesized model.

First consider the χ^2–test for goodness of fit for situations where the hypothesized model completely specifies the probabilities for each of the possible outcomes. More formally, consider a situation where the population units can be categorized into k mutually exclusive and exhaustive classifications and where the model completely specifies the probabilities, $p_1, p_2, \ldots, p_k$, of belonging to these k classifications. The χ^2–test of goodness of fit is used to test the null hypothesis that the k probabilities specified by the model are correct versus the alternative hypothesis that these probabilities are not all correct. The χ^2–test is most easily presented in terms of the observed frequencies (observed counts), $f_1, f_2, \ldots, f_k$, of the k classifications and the hypothetical expected frequencies (expected counts), $F_1, F_2, \ldots, F_k$, predicted by the model. Assuming that the data correspond to a random sample of size n, we can express the expected frequencies in terms of the model probabilities as $F_1 = np_1, F_2 = np_2, \ldots, F_k = np_k$. We will develop the χ^2–test in the context of several examples.

Example. Inheritance in peas (flower color). In Section 5.3 we described a simple Mendelian inheritance model for the color of pea plant flowers arising from crossing two first generation plants. This model hypothesizes that the probability that a plant has red flowers is $p_R = 3/4$ and the probability that a plant has white flowers is $p_W = 1 - p_R = 1/4$. Mendel observed $n = 929$ pea plants arising from a cross of two first generation plants of which 705 plants had red flowers and 224 plants had white flowers. Under the hypothesized model we would expect to see red flowers 3/4 of the time and white flowers 1/4 of the time. Thus, for Mendel's experiment with a total of 929 plants we would expect to see about 696.75 plants with red flowers and about 232.25 plants with white flowers.

We can test the consistency of this model with the data by comparing the observed frequencies of red and white flowered plants to the corresponding expected frequencies. The first step in this comparison is to find the differences between the observed and expected frequencies of plants for each of the two colors. In this example we have differences of $705 - 696.75 = 8.25$ (red) and $224 - 232.25 = -8.25$ (white.) These differences add to zero, since both the observed and expected frequencies add to 929. The second step is to square each difference and standardize it by dividing by the corresponding expected frequency. This standardization gives $68.0625/696.75 = .0977$ (red) and $68.0625/232.25 = .2931$ (white.) Adding these standardized squared differences gives the χ^2–statistic $\chi^2_{calc} = .0977 + .2931 = .3908$. These computations are summarized in Table 1.

Table 1. Pea plant flower color example.

flower color	observed frequency	expected frequency	obs − exp	$(\text{obs}-\text{exp})^2/\text{exp}$
red	705	696.75	8.25	.0977
white	224	232.25	-8.25	.2931
total	929	929		$\chi^2_{calc} = .3908$

A large value of the χ^2–statistic indicates evidence against the null hypothesis that the model is valid $H_0 : p_R = 3/4$ and $p_W = 1/4$ and in favor of the alternative hypothesis that the model is not valid H_1 : it is not true that $p_R = 3/4$ and $p_W = 1/4$. We can determine whether $\chi^2_{calc} = .3908$ is large by computing the relevant P–value. The P–value for this χ^2–test is the probability of observing a value of χ^2 as large or larger than the calculated value $\chi^2_{calc} = .3908$ computed using the appropriate χ^2 distribution. In a situation like the present example, where there are k categories or classifications and the model completely specifies the k corresponding probabilities, the appropriate χ^2 distribution is the χ^2 distribution with $k-1$ degrees of freedom. In this example there are $k = 2$ possible classifications (red or white) and the model completely specifies the two corresponding probabilities ($p_R = 3/4$ and $p_W = 1/4$), so the χ^2 distribution with $k - 1 = 1$ degree of freedom is used to compute the P–value. With $\chi^2_{calc} = .3908$ and one degree of freedom we get the P–value $P(\chi^2 \geq .3908) = .5310$. This P–value is quite large and we are not able to reject the null hypothesis; therefore, we conclude that the observed data are consistent with Mendel's model.

In this example there are $k = 2$ classifications and $p_W = 1 - p_R$, thus the hypotheses specified in terms of p_R and p_W above can be written more simply as $H_0 : p_R = 3/4$ and $H_1 : p_R \neq 3/4$. In section 5.3 we used the normal approximation to perform a Z–test for these hypotheses. The χ^2–test presented above is actually equivalent to this Z–test. To see this equivalence consider the Z–test for $H_0 : p = p_0$ versus $H_1 : p \neq p_0$; for this test we have

$$\begin{aligned} Z^2 &= \frac{(\hat{p}-p_0)^2}{p_0(1-p_0)/n} = \frac{(n\hat{p}-np_0)^2}{np_0(1-p_0)} \\ &= [(1-p_0)+p_0]\frac{(n\hat{p}-np_0)^2}{np_0(1-p_0)} \\ &= \frac{(n\hat{p}-np_0)^2}{np_0} + \frac{(n[1-\hat{p}]-n[1-p_0])^2}{n(1-p_0)} = \chi^2; \end{aligned}$$

and a value of Z which is far away from zero corresponds to a large value of $Z^2 = \chi^2$. Furthermore, it can be shown that the square of a standard normal variable follows the χ^2 distribution with one degree of freedom; and thus these two approaches are equivalent.

The χ^2–test is of more interest when there are three or more classifications, since there is no Z–test in these cases. The χ^2–test for an example with $k = 4$ classifications is developed below.

Example. Inheritance in peas (seed shape and color). We will now consider the Mendelian inheritance model for two independently inherited characteristics. In particular we will consider the characteristics seed shape, with possible shapes of round (R, dominant) and wrinkled (r, recessive), and seed color, with possible colors of yellow (Y, dominant) and green (y, recessive). If a $RRYY$ genotype plant with round yellow seeds is crossed with a $rryy$ genotype plant with wrinkled green seeds, the offspring will all have round yellow seeds and genotype $RrYy$. If two of the resulting $RrYy$ genotype plants with round yellow seeds are crossed, there are 16 equally likely possible genotypes. The nine genotypes $RRYY, RRYy, RRyY, RrYY, RrYy, RryY, rRYY, rRYy, rRyY$ yield round yellow seeds; the three genotypes $rrYY, rrYy, rryY$ yield wrinkled yellow seeds; the three genotypes $RRyy, Rryy, rRyy$ yield round green seeds; and, the single genotype $rryy$ yields wrinkled green seeds. The facts that these 16 possible genotypes are equally likely and each plant possesses only one genotype yield the probability distribution summarized in Table 2.

Table 2. Pea plant seed shape/color distribution.

shape/color	probability
round yellow	9/16
wrinkled yellow	3/16
round green	3/16
wrinkled green	1/16

The results of one of Mendel's experiments regarding seed shape and color, with $n = 556$ plants, are summarized in Table 3. Table 3 also contains the expected frequencies, computed using the distribution of Table 2, and the computations leading to the χ^2 statistic. In this example there are $k = 4$ classifications and the P–value $P(\chi^2 \geq .4700) = .9254$

Table 3. Pea plant seed shape and color example.

shape/color	observed frequency	expected frequency	obs − exp	$(\text{obs} - \text{exp})^2/\text{exp}$
round yellow	315	312.75	2.25	.0162
wrinkled yellow	101	104.25	-3.25	.1013
round green	108	104.25	3.75	.1349
wrinkled green	32	34.75	-2.75	.2176
total	556	556		$\chi^2_{calc} = .4700$

for the χ^2–test is computed using the χ^2 distribution with $k-1=3$ degrees of freedom. In this example, the P–value is quite large and we are not able to reject the null hypothesis; therefore, we conclude that the observed data are consistent with Mendel's model.

In both of the preceding examples the data are consistent with the hypothesized model. The following example, which is also concerned with a Mendelian inheritance model, illustrates a situation where the data are not consistent with the model.

Example. Inheritance in maize (leaf characteristics). This example is taken from Snedecor and Cochran (1980), the original source is Lindstrom (1918) *Cornell Agr. Exp. Sta. mem. 13*. Lindstrom crossed two types of maize (corn) plants and classified the resulting plants into four categories based on the appearance of the leaves. The Mendelian model for this example is analogous to the model of the pea plant seed shape and color example with respective probabilities of $9/16, 3/16, 3/16$, and $1/16$. Thus the model predicts that the four leaf types should occur in a ratio of $9:3:3:1$. The data and the computations are summarized in Table 4.

Table 4. Maize leaf type example.

leaf type	observed frequency	expected frequency	obs − exp	$(\text{obs}-\text{exp})^2/\text{exp}$
green	773	731.813	41.1875	2.3181
golden	231	243.938	-12.9375	0.6862
green–striped	238	243.938	-5.9375	0.1445
green–golden–striped	59	81.313	-22.3125	6.1226
total	1301	1301		$\chi^2_{calc} = 9.2714$

In this example $\chi^2_{calc} = 9.2714$ is large indicating disagreement between the model and the data. The P–value .0259, computed using the χ^2 distribution with 3 degrees of freedom, is small enough to allow us to conclude that Lindstrom's data are not consistent with the Mendelian model which predicts frequencies in the ratio of $9:3:3:1$.

Examination of the four terms we added to get χ^2_{calc} indicates that the green–golden–striped term 6.1226 is large relative to the other terms. Thus the evidence against the model seems to be due to the fact that the observed frequency of green–golden–striped plants 59 is much smaller than the expected frequency 81.313. Lindstrom argued that this discrepancy could be explained by "the weakened condition of the last three classes due to their chlorophyll abnormality". In particular, he noted that the plants in the green–golden–striped class were not very vigorous (did not grow well). This suggests that the evidence against the model may be due to the fact that some of the green–golden–striped plants did not survive long enough to be counted. Therefore, we might wonder whether

our rejection of the 9 : 3 : 3 : 1 model can be attributed to the poor survivorship of the green–golden–striped plants. We will now perform an exploratory analysis to address this question.

First consider the 1242 plants in the first three classifications. According to the model the frequencies for these three classifications should be in the ratio 9 : 3 : 3. The computations for a χ^2 test for this subset of the original data are demonstrated in Table 5. For this subset of the original data we have $\chi^2_{calc} = 2.6914$ with $3 - 1 = 2$ degrees of freedom which gives a P–value of .2604. Therefore, there is evidence that the frequencies in the first three classes are consistent with the predicted ratio of 9 : 3 : 3.

Table 5. Maize leaf type example, 9:3:3 model.

leaf type	observed frequency	expected frequency	obs − exp	$(\text{obs} - \text{exp})^2/\text{exp}$
green	773	745.2	27.8	1.0371
golden	231	248.4	-17.4	1.2188
green–striped	238	248.4	-10.4	0.4354
total	1242	1242		$\chi^2_{calc} = 2.6914$

This test and the fact that the observed frequency of green–golden–striped plants is much smaller than expected suggest that the reason that the original data do not agree with the model may be poor survivorship of the green–golden–striped plants, since the data for the other classes do agree with the model.

In some situations, like the two examples which follow, the hypothesized model does not completely specify the probabilities for the k possible outcomes and it is necessary to estimate these probabilities before performing the χ^2 goodness of fit test.

Example. Radioactive disintegrations. This example is taken from Feller (1957), p. 149 and Cramér (1946) p. 436. In a famous experiment by Rutherford, Chadwick, and Ellis (*Radiations from Radioactive Substances*, Cambridge, 1920) a radioactive substance was observed during 2608 consecutive time intervals of length 7.5 seconds each. The number of particles reaching a counter was recorded for each period giving the results summarized in Table 6.

The Poisson distribution, as discussed in Chapter 4a, provides a plausible model for the number of particles, X, observed in this experiment. Therefore, we will perform a χ^2 goodness of fit test to see whether a Poisson distribution is suitable as a model for the distribution of the observed number of particles in this experiment. The Poisson model places no upper bound on the number of particles which could be observed; so, for this

test, we will use "10 or more particles" as the largest possible "value" of the variable. The Poisson distribution with parameter λ specifies probabilities of the form

$$P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

for $x = 0, 1, 2, \ldots$ Notice that this probability function does not completely specify the

Table 6. Radioactive disintegrations data.

number	observed frequency
0	57
1	203
2	383
3	525
4	532
5	408
6	273
7	139
8	45
9	27
10	10
11	4
12	2
total	2608

probabilities of the possible values of X, since there is an unknown parameter λ in the formula. Therefore, to perform the χ^2–test we first need to use the data to estimate λ. Since λ is the mean of the Poisson distribution, we can use the sample mean 3.8704 as an estimate of λ and use the formula from above to determine the expected frequencies. Thus, for $x = 0, 1, \ldots, 9$, we compute the expected frequencies using the formula

$$2608 \left(\frac{3.8704^x}{x!} e^{-3.8704} \right)$$

and we subtract the sum of these expected frequencies from 2608 to find the expected frequency for $X \geq 10$. The observed and expected frequencies and the terms used to calculate the χ^2–statistic are summarized in Table 7. Because we estimated the parameter λ of the hypothesized Poisson distribution we need to reduce the degrees of freedom for the χ^2–test by one. For this example we have $\chi^2_{calc} = 12.8815$ with $k - 2 = 9$ degrees of freedom which gives a P–value of .1680. Since this P–value is not small we can conclude that a Poisson model with $\lambda = 3.8704$ provides a reasonable model for the number of radioactive disintegrations observed in this experiment.

Table 7. Radioactive disintegrations.

number	observed frequency	expected frequency	obs − exp	$(\text{obs} - \text{exp})^2/\text{exp}$
0	57	54.3769	2.6231	0.1265
1	203	210.4604	-7.4604	0.2645
2	383	407.2829	-24.2829	1.4478
3	525	525.4491	-.4491	0.0004
4	532	508.4244	23.5756	1.0932
5	408	393.5610	14.4390	0.5297
6	273	253.8730	19.1270	1.4410
7	139	140.3700	-1.3700	0.0134
8	45	67.9110	-22.9110	7.7294
9	27	29.2047	-2.2047	0.1664
≥ 10	16	17.0865	-1.0865	0.0691
total	2608	2608		$\chi^2_{calc} = 12.8815$

Example. Bacteria counts. This example is taken from Feller (1957), p.153. The original source is T. Matuszewsky, J. Supinska, and J. Neyman (1936), *Zentralblatt für Bakteriologie, Parasitenkunde und Infektionskrankrankheiten*, II Abt., **95**.

Table 8. Bacteria counts data.

number	observed frequency
0	5
1	19
2	26
3	26
4	21
5	13
6	8
total	118

A Petri dish with bacteria colonies was examined under a microscope. The dish was divided into small squares and the number of bacteria colonies, visible as dark spots, was recorded for each square. The data are given in Table 8. If the bacteria colonies were randomly distributed over the Petri dish, without being clustered together, then the Poisson model should hold. The sample mean number of bacteria colonies is 2.9322 and, as in the preceding example, we can use this sample mean to estimate the parameter λ of the Poisson model.

Table 9. Bacteria counts.

number	observed frequency	expected frequency	obs − exp	$(\text{obs} - \text{exp})^2/\text{exp}$
0	5	6.2870	-1.2870	.2635
1	19	18.4347	.5653	.0173
2	26	27.0272	-1.0272	.0390
3	26	26.4164	-.4164	.0066
4	21	19.3645	1.6354	.1381
5	13	11.3562	1.6438	.2380
≥ 6	8	9.1140	-1.1141	.1362
total	118	118		$\chi^2_{calc} = .8386$

The observed and expected frequencies and the terms used to calculate the χ^2–statistic are summarized in Table 9. Again, since we estimated the parameter λ of the hypothesized Poisson distribution, we need to reduce the degrees of freedom for the χ^2–test by one. For this example we have $\chi^2_{calc} = .8386$ with $k - 2 = 5$ degrees of freedom which gives a P–value of .9745. Since this P–value is very large, we can conclude that a Poisson model with $\lambda = 2.9322$ provides a reasonable model for the number of bacteria colonies as observed in this experiment. This indicates that the conjecture that the bacteria colonies are randomly distributed over the Petri dish, without being clustered together, is consistent with the observations.

11.3 Chi–square Tests for Homogeneity

We will now consider χ^2–tests for the homogeneity of two or more population distributions. These tests can be viewed as generalizations of the Z–test of equality of two population proportions of Section 6.2 to allow for more than two populations or more than two possible classifications.

A probability distribution for a qualitative variable with k possible values corresponding to k mutually exclusive classifications can be represented by a collection $\mathbf{p} = (p_1, p_2, \ldots, p_k)$ of k probabilities which sum to one. Given m such probability distributions, we have m collections of probabilities $\mathbf{p}_1 = (p_{11}, p_{12}, \ldots, p_{1k})$, $\mathbf{p}_2 = (p_{21}, p_{22}, \ldots, p_{2k})$, ..., and $\mathbf{p}_m = (p_{m1}, p_{m2}, \ldots, p_{mk})$ as shown in Table 10. The null hypothesis of homogeneity of these m distributions, $H_0 : \mathbf{p}_1 = \mathbf{p}_2 = \cdots = \mathbf{p}_m$, specifies that the probability of observing a unit in a particular classification is the same for all m of the populations, *i.e.*, for each classification $j = 1, 2, \ldots, k$, we have $p_{1j} = p_{2j} = \cdots = p_{mj}$.

Table 10. Notation for m populations and k classifications.

	classification probabilities				
population	1	2	...	k	sum
1	p_{11}	p_{12}	...	p_{1k}	1
2	p_{21}	p_{22}	...	p_{2k}	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
m	p_{m1}	p_{m2}	...	p_{mk}	1

Suppose that m independent random samples of sizes $n_1, n_2, \ldots,$ and n_m are obtained from these m population distributions and let f_{ij} denote the observed frequency of units in classification j for the sample from population i as shown in Table 11. Under the null hypothesis we would expect the m collections of k observed frequencies in each row (sample) of Table 11 to be the same (no difference from row to row).

We can use the combined observed frequencies $F_1, F_2, \ldots, F_k$ in the last line of Table 11, obtained by adding the corresponding frequencies in the respective columns, and the combined sample size $n = n_1 + n_2 + \cdots + n_m$ to form estimates of the frequencies we would expect to observe under the null hypothesis of homogeneity. We first compute the estimates $\hat{p}_1 = F_1/n$, $\hat{p}_2 = F_2/n$, ..., and $\hat{p}_k = F_k/n$ of the assumed common classification probabilities $p_1, p_2, \ldots,$ and p_k and then we multiply this collection of $\hat{p}'s$ by the respective sample sizes to get the expected frequencies for each population (row of the table). The P–value for the resulting χ^2–statistic, which is based on the $m \times k$ observed and expected frequencies, is obtained from the χ^2 distribution with $(m-1)(k-1)$ degrees of freedom.

Table 11. Data for m populations and k classifications.

population	observed frequencies				sample
	1	2	...	k	size
1	f_{11}	f_{12}	...	f_{1k}	n_1
2	f_{21}	f_{22}	...	f_{2k}	n_2
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
m	f_{m1}	f_{m2}	...	f_{mk}	n_m
combined	F_1	F_2	...	F_k	n

We will first apply this χ^2–test of homogeneity to an example with $m = 2$ dichotomous $(k = 2)$ populations.

Example. Cocaine addiction. This example is based on a study of D.M. Barnes (1988), *Science*, **241**, 1029–1030, as described in Moore (1995). This study was conducted to compare two antidepressants as treatments for cocaine addiction. In particular, the researchers wanted to compare the effects of the antidepressant desipramine with the effects of lithium (a standard treatment for cocaine addiction.) A group of 48 chronic cocaine users was randomly divided into two groups of 24. One group was treated with desipramine and the other was treated with lithium. The subjects were tracked for three years and the number of subjects who relapsed into cocaine use during this period was recorded. The data are summarized as observed frequencies in Table 12.

Table 12. Cocaine example: observed and expect frequencies.

observed frequency				expected frequency			
	relapsed				relapsed		
treatment	yes	no	total	treatment	yes	no	total
desipramine	10	14	24	desipramine	14	10	24
lithium	18	6	24	lithium	14	10	24
combined	28	20	48				

For this example we can view the data as independent random samples of size 24 from dichotomous populations with population success probabilities p_D and p_L, where p_D is the probability that one of these 48 cocaine users would relapse into cocaine use if all 48 users were treated with desipramine and p_L is the analogous probability assuming that all 48 users were treated with lithium. We can use a χ^2–test to test the null hypothesis $H_0 : p_D = p_L$ that the probability of relapse is the same for both treatments versus the alternative hypothesis $H_1 : p_D \neq p_L$ of different probabilities of relapse. Under the null hypothesis we would expect to observe the same relapse proportions under each treatment; furthermore, since 28 of the 48 users suffered a relapse we can use the combined sample relapse proportion $\hat{p} = 28/48$ as our estimate of the common relapse probability we would expect to observe under the null hypothesis. The expected frequencies in Table 12 were computed using this $\hat{p}$ as the estimated common relapse probability and the sample sizes, which are both 24. The differences between the observed and expected frequencies and the four components of the χ^2–statistic are given in Table 13.

The P–value for $\chi^2_{calc} = 192/35 = 5.487$, computed using the χ^2 distribution with $(2-1)(2-1) = 1$ degrees of freedom, is $P(\chi^2 \geq 5.487) = 0.0192$. This small P–value allows us to reject the null hypothesis of homogeneity and conclude that $p_D \neq p_L$ indicating that the probability of relapse is not the same when a user is treated with desipramine as when the user is treated with lithium. Since this example involves two dichotomous populations, we could have used the Z–test of Section 6.2, which is equivalent to the χ^2 test from above

in this situation, to perform this test. More importantly, since we have two dichotomous populations, we can use the Z–interval of Section 6.1 to quantify the size and direction of the difference between p_D and p_L. The sample success proportions are $\hat{p}_D = .4167$ and $\hat{p}_L = .75$ and the 95% confidence interval for $p_L - p_D$ is $(.0708, .5959)$. Hence, we are 95% confident that treating one of these 48 cocaine users with desipramine instead of lithium would reduce the probability of relapse by at least .0708 and as much as .5959.

Table 13. Cocaine example: chi–square computations.

obs - exp			$(\text{obs - exp})^2/\text{ exp}$		
	relapsed			relapsed	
treatment	yes	no	treatment	yes	no
desipramine	-4	4	desipramine	16/14	16/10
lithium	4	-4	lithium	16/14	16/10
				$\chi^2_{calc} = 192/35 = 5.487$	

The next example with $m = 3$ populations and $k = 3$ categories will be used to demonstrate the extension of the χ^2–test of homogeneity to situations with three or more populations and three or more categories.

Example. Attitudes of School Children. This example is based on a study described by Chase and Dummer (1992), *Research Quarterly for Exercise and Sport*, **63**, 418–424, as described in DeGroot and Schervish (2002). This study was conducted to examine the attitudes of school–aged children in Michigan. Three independent random samples of children were obtained. A sample of 149 children from rural areas, a sample of 151 children from suburban areas, and a sample of 178 children from urban areas. Each child was asked which of the following was most important to them: good grades, athletic ability, or popularity. The observed frequencies are given in Table 14 and the expected frequencies, based on the combined probability estimates $247/478 = .5167$, $90/478 = .1883$, and $141/478 = .2950$ and the sample sizes 149, 151, and 178 are given in Table 15.

Table 14. Attitude example: observed frequencies.

sample	good grades	athletic ability	popularity	sample size
rural	57	42	50	149
suburban	87	22	42	151
urban	103	26	49	178
combined	247	90	141	478

Table 15. Attitude example: expected frequencies.

sample	good grades	athletic ability	popularity	sample size
rural	76.9937	28.0544	43.9519	149
suburban	78.0272	28.4310	44.5418	151
urban	91.9791	33.5146	52.5063	178

The differences between the observed and expected frequencies and the nine components of the χ^2–statistic are given in Table 16. The P–value for $\chi^2_{calc} = 18.8276$, computed using the χ^2 distribution with $(3-1)(3-1) = 4$ degrees of freedom, is $P(\chi^2 \geq 18.8276) = 0.0008$. This very small P–value indicates very strong evidence that the attitude distributions (the three probabilities for the three choices given to these children) are not the same for the three areas.

Table 16. Attitude example: chi–square computations.

	obs - exp			(obs - exp)^2/exp		
sample	good grades	athletic ability	popularity	good grades	athletic ability	popularity
rural	-19.9937	13.9456	6.0481	5.19197	6.93225	0.83227
suburban	8.9728	-6.4310	-2.5418	1.03184	1.45466	0.14505
urban	11.0209	-7.5146	-3.5063	1.32053	1.68493	0.23414

The two largest (obs - exp)^2/exp terms, 5.19197 for the rural–good grades category and 6.93225 for the rural–athletic ability category, are much larger than the other terms. This fact and the observed relative frequencies given in Table 17 suggest that the attitude distributions might be the same for the suburban and urban children but different for the rural children.

Table 17. Attitude example: observed relative frequencies.

sample	good grades	athletic ability	popularity
rural	.3826	.2819	.3356
suburban	.5762	.1457	.2781
urban	.5787	.1461	.2753

The χ^2–statistic based on the data for suburban and urban children only is $\chi^2_{calc} = .0034$ with $(2-1)(3-1) = 2$ degrees of freedom, which gives a P–value of .9983 and supports the contention that the attitude distribution is the same for the suburban children as it is for the urban children. Furthermore, if we combine the suburban sample and the urban

sample to form a nonrural sample of size 329, the χ^2–statistic for comparing the rural and nonrural samples is $\chi^2_{calc} = 18.8243$ with $(2-1)(3-1) = 2$ degrees of freedom and the P–value is less than .0001, confirming our conjecture that the attitude distribution for the rural children is not the same as that for the nonrural children.

11.4 Chi–square Tests for Independence

A χ^2–test for independence is used to determine whether two or more qualitative classification factors are independent. In this section we will restrict our attention to crossed classifications of units with respect to two qualitative classification factors. Two classification factors, A and B, are said to be independent, if the conditional probabilities for the levels of factor A (respectively, factor B), obtained by fixing the level of factor B (factor A), are the same regardless of the level at which factor B (factor A) is fixed. To avoid complex notation we will describe independence and develop the χ^2–test for independence in the context of the following example.

Example. Hawaiian blood types. This example uses data from A.E. Mourant, *et al.*, *The Distribution of Blood Groups and Other Polymorphisms*, Oxford University Press, London, 1976. The Blood Bank of Hawaii cross classified 145,057 individuals according to their blood type (A, AB, B, O) and their ethnic group (Hawaiian, Hawaiian–Chinese, Hawaiian–White, White). The frequencies for each of the 16 combinations of the 4 levels of these two qualitative classification factors are given in Table 18.

Table 18. Blood type and ethnic group observed frequencies.

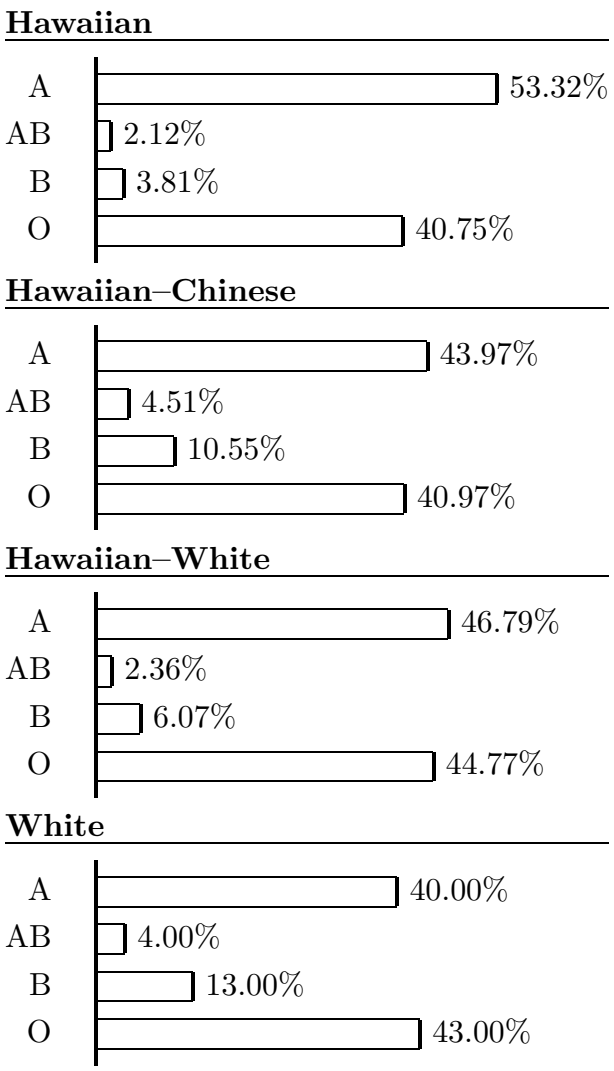
	ethnic group				
blood type	Hawaiian	Hawaiian–Chinese	Hawaiian–White	White	total
A	2490	2368	4671	50008	59537
AB	99	243	236	5001	5579
B	178	568	606	16252	17604
O	1903	2206	4469	53759	62337
total	4670	5385	9982	125020	145057

The question we want to consider here is whether the distribution of blood types is independent of the distribution of the ethnic groups. If the distribution of blood types is the same for each of the four ethnic groups, then classification with respect to blood type is independent of classification with respect to ethnic group. Furthermore, independence of two factors is symmetric so that if the distribution of blood types is independent of the

distribution of ethnic groups, then it also follows that the distribution of ethnic groups is independent of the distribution of blood types.

Under the hypothesis of independence the theoretical conditional distributions of blood type are the same for each ethnic group. The conditional distributions of blood type for each ethnic group summarized in Figure 2 show some evidence that the distributions of blood type are not the same for these ethnic groups indicating dependence between classification with respect to blood type and classification with respect to ethnic group.

Figure 2. Conditional distributions of blood type by ethnic group.



We can compute the expected frequencies for this example the same way we did for the χ^2–tests of homogeneity in Section 11.3. The deviations between the observed and expected frequencies and the 16 terms which are summed to give the χ^2–statistic are given in Table 19. In this example several of the χ^2 terms are large indicating where the hypothesis of independence is not supported by these data. The χ^2–statistic for testing the

independence of blood type and ethnic group is $\chi^2_{calc} = 1078.6036$ with $(4-1)(4-1) = 9$ degrees of freedom and the P–value is less than .0001. Therefore, there is very strong evidence against the null hypothesis of independence. We can conclude that the data collected by the Blood Bank of Hawaii are clearly inconsistent with the hypothesis of independence and that the distribution of blood types is not the same for these four ethnic groups.

Table 19. Hawaiian blood type example chi–square information.

The first number is the deviation (obs - exp) and the number in parentheses is the χ^2 term (obs - exp)^2/exp.

	ethnic group			
blood type	Hawaiian	Hawaiian–Chinese	Hawaiian–White	White
A	573.25	157.79	574	-1305
	(171.45)	(11.265)	(80.419)	(33.191)
AB	-80.61	35.889	-147.9	192.64
	(36.179)	(6.2189)	(56.989)	(7.7177)
B	-388.7	-85.52	-605.4	1079.7
	(266.65)	(11.191)	(302.56)	(76.83)
O	-103.9	-108.2	179.32	32.729
	(5.3783)	(5.055)	(7.4962)	(.0199)