Topic 21

Goodness of Fit

21.1 Fit of a Distribution

Goodness of fit tests examine the case of a sequence of independent observations each of which can have 1 of k possible categories. For example, each of us has one of 4 possible of blood types, O, A, B, and AB. The local blood bank has good information from a national database of the fraction of individuals having each blood type,

$$\pi_O, \pi_A, \pi_B, \text{ and } \pi_{AB}.$$

The actual fraction p_O, p_A, p_B, and p_{AB} of these blood types in the community for a given blood bank may be different than what is seen in the national database. As a consequence, the local blood bank may choose to alter its distribution of blood supply to more accurately reflect local conditions.

To place this assessment strategy in terms of formal hypothesis testing, let $\pi = (\pi_1, \ldots, \pi_k)$ be postulated values of the probability

$$P_\pi\{\text{individual is a member of } i\text{-th category}\} = \pi_i$$

and let $\mathbf{p} = (p_1, \ldots, p_k)$ denote the possible states of nature. Then, the parameter space is

$$\Theta = \{\mathbf{p} = (p_1, \ldots, p_k); p_i \geq 0 \text{ for all } i = 1, \ldots, k, \sum_{i=1}^{k} p_i = 1\}.$$

This parameter space has $k - 1$ free parameters. Once these are chosen, the remaining parameter value is determined by the requirement that the sum of the p_i equals 1. Thus, $\dim(\Theta) = k - 1$.

The hypothesis is

$$h01gofH_0 : p_i = \pi_i, \text{ for all } i = 1, \ldots, k \quad \text{versus} \quad H_1 : p_i \neq \pi_i, \text{ for some } i = 1, \ldots, k. \tag{21.1}$$

The parameter space for the null hypothesis is a single point $\pi = (\pi_1, \ldots, \pi_k)$. Thus, $\dim(\Theta_0) = 0$. Consequently, the likelihood ratio test will have a chi-square test statistic with $\dim(\Theta) - \dim(\Theta_0) = k - 1$ degrees of freedom. The data $\mathbf{x} = (x_1, \ldots, x_n)$ are the categories for each of the n observations.

Let's use the likelihood ratio criterion to create a test for the distribution of human blood types in a given population. For the data

$$\mathbf{x} = \{O, B, O, A, A, A, A, A, O, AB\}$$

for the blood types of tested individuals, then, in the case of independent observations, the likelihood is

$$L(\mathbf{p}|\mathbf{x}) = p_O \cdot p_B \cdot p_O \cdot p_A \cdot p_A \cdot p_A \cdot p_A \cdot p_A \cdot p_O \cdot p_{AB} = p_O^3 p_A^5 p_B p_{AB}.$$

Notice that the likelihood has a factor of p_i whenever an observation take on the value i. In other words, if we summarize the data using

$$n_i = \#\{\text{observations from category } i\}$$

to create $\mathbf{n} = (n_1, n_2, \cdots, n_k)$, a vector that records the number of observations in each category, then, the likelihood function

$$L(\mathbf{p}|\mathbf{n}) = p_1^{n_1} \cdots p_k^{n_k}. \tag{21.2}$$

The **likelihood ratio** is the ratio of the maximum value of the likelihood under the null hypothesis and the maximum likelihood for any parameter value. In this case, the numerator is the likelihood evaluated at π.

$$\Lambda(\mathbf{n}) = \frac{L(\pi|\mathbf{n})}{L(\hat{\mathbf{p}}|\mathbf{n})} = \frac{\pi_1^{n_1}\pi_2^{n_2}\cdots\pi_k^{n_k}}{\hat{p}_1^{n_1}\hat{p}_2^{n_2}\cdots\hat{p}_k^{n_k}} = \left(\frac{\pi_1}{\hat{p}_1}\right)^{n_1} \cdots \left(\frac{\pi_k}{\hat{p}_k}\right)^{n_k}. \tag{21.3}$$

To find the maximum likelihood estimator $\hat{p}$, we, as usual, begin by taking the logarithm in (21.2),

$$\ln L(\mathbf{p}|\mathbf{n}) = \sum_{i=1}^{k} n_i \ln p_i.$$

Because not every set of values for p_i is admissible, we cannot just take derivatives, set them equal to 0 and solve. Indeed, we must find a maximum under the constraint

$$s(\mathbf{p}) = \sum_{i=1}^{k} p_i = 1.$$

The maximization problem is now stated in terms of the method of **Lagrange multipliers**. This method tells us that at the maximum likelihood estimator $(\hat{p}_1, \ldots, \hat{p}_k)$, the gradient of $\ln L(\mathbf{p}|\mathbf{n})$ is proportional to the gradient of the constraint $s(\mathbf{p})$. To explain this briefly, recall that the gradient of a function is a vector that is perpendicular to a level set of that function. In this case,

$$\nabla_{\hat{\mathbf{p}}} s(\mathbf{p}) \quad \text{is perpendicular to the level set} \quad \{\mathbf{p}; s(\mathbf{p}) = 1\}.$$

Now imagine walking along the set of parameter values of $\mathbf{p}$ given by the constraint $s(\mathbf{p}) = 1$, keeping track of the values of the function $\ln L(\mathbf{p}|\mathbf{n})$. If the walk takes us from a value of this function below ℓ_0 to values above ℓ_0 then (See Figure 21.1.), the level surfaces

$$\{\mathbf{p}; s(\mathbf{p}) = 1\}$$

and

$$\{\ln L(\mathbf{p}|\mathbf{n} = \ell_0\}$$

intersect. Consequently, the gradients

$$\nabla_{\hat{\mathbf{p}}} s(\mathbf{p}) \quad \text{and} \quad \nabla_{\hat{\mathbf{p}}} \ln L(\mathbf{p}|\mathbf{n})$$

point in different directions on the intersection of these two surfaces. At a local maximum or minimum of the log-likelihood function, the level surfaces are tangent and the two gradients are parallel. In other words, the these two gradients vectors are related by a constant of proportionality, λ, known as the **Lagrange multiplier**. Consequently, at extreme values,

$$\nabla_{\mathbf{p}} \ln L(\hat{\mathbf{p}}|\mathbf{n}) = \lambda \nabla_{\hat{\mathbf{p}}} s(\mathbf{p}).$$

$$\left(\frac{\partial}{\partial p_1} \ln L(\hat{\mathbf{p}}|\mathbf{n}), \ldots, \frac{\partial}{\partial p_k} \ln L(\hat{\mathbf{p}}|\mathbf{n})\right) = \lambda \left(\frac{\partial}{\partial p_1} s(\mathbf{p}), \ldots, \frac{\partial}{\partial p_k} s(\mathbf{p})\right)$$

$$\left(\frac{n_1}{\hat{p}_1}, \ldots, \frac{n_k}{\hat{p}_k}\right) = \lambda(1, \ldots, 1)$$

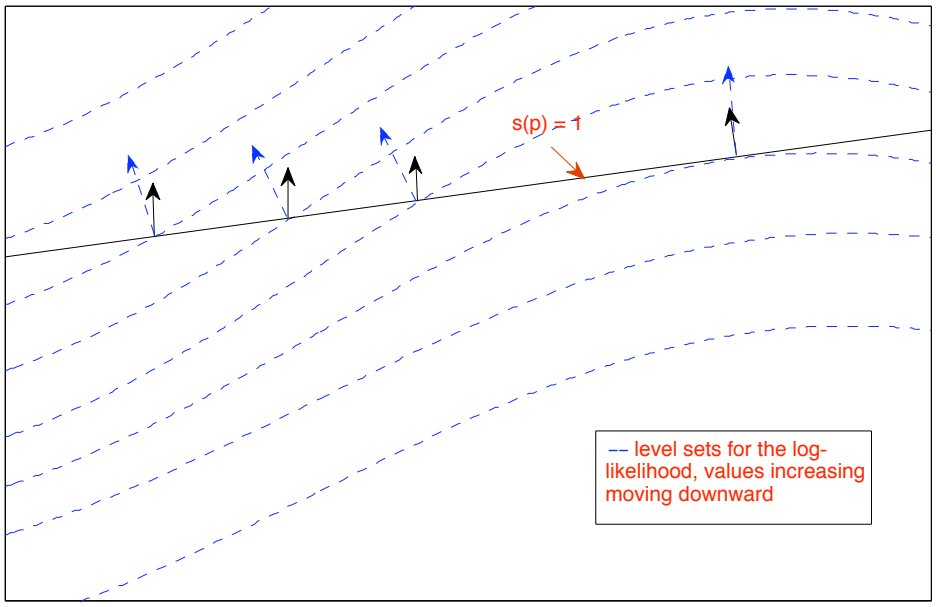

Figure 21.1: Lagrange multipliers Level sets of the log-likelihood function.shown in dashed blue. The level set $\{s(\mathbf{p}) = 1\}$ shown in black. The gradients for the log-likelihood function and the constraint are indicated by dashed blue and black arrows, respectively. At the maximum, these two arrows are parallel. Their ratio λ is called the Lagrange multiplier. If we view the blue dashed lines as elevation contour lines and the black line as a trail, crossing contour line indicates walking either up or down hill. When the trail reaches its highest elevation, the trail is tangent to a contour line and the gradient for the hill is perpendicular to the trail.

Each of the components of the two vectors must be equal. In other words,

$$\frac{n_i}{\hat{p}_i} = \lambda, \quad n_i = \lambda \hat{p}_i \quad \text{for all } i = 1, \dots, k. \tag{21.4}$$

Now sum this equality for all values of i and use the constraint $s(\mathbf{p}) = 1$ to obtain

$$n = \sum_{i=1}^{k} n_i = \lambda \sum_{i=1}^{k} \hat{p}_i = \lambda s(\hat{\mathbf{p}}) = \lambda.$$

Returning to (21.4), we have that

$$\frac{n_1}{\hat{p}_i} = n \quad \text{and} \quad \hat{p}_i = \frac{n_i}{n}.$$

This is the answer we would guess - the estimate for p_i is the fraction of observations in category i. Thus, for the introductory example,

$$\hat{p}_O = \frac{3}{10}, \quad \hat{p}_A = \frac{5}{10}, \quad \hat{p}_B = \frac{1}{10}, \quad \text{and} \quad \hat{p}_O = \frac{1}{10}.$$

Next, we substitute the maximum likelihood estimates $\hat{p}_i = n_i/n$ into the likelihood ratio (21.3) to obtain

$$\Lambda(\mathbf{n}) = \frac{L(\pi|\mathbf{n})}{L(\hat{\mathbf{p}}|\mathbf{n})} = \left(\frac{\pi_1}{n_1/n}\right)^{n_1} \cdots \left(\frac{\pi_k}{n_k/n}\right)^{n_k} = \left(\frac{n\pi_1}{n_1}\right)^{n_1} \cdots \left(\frac{n\pi_k}{n_k}\right)^{n_k}. \tag{21.5}$$

Recall that we reject the null hypothesis if this ratio is too low, i.e, the maximum likelihood under the null hypothesis is sufficiently smaller than the maximum likelihood under the alternative hypothesis.

Let's review the process. the random variables $X_1, X_2, \dots, X_n$ are independent, taking values in one of k categories each having distribution π. In the example, we have 4 categories, namely the common blood types O, A, B, and AB. Next, we organize the data into

$$N_i = \#\{j; X_j = i\},$$

the number of observations in category i. Next, create the vector $\mathbf{N} = (N_1, \ldots, N_k)$ to be the vector of observed number of occurrences for each category i. In the example we have the vector (3,5,1,1) for the number of occurrences of the 4 blood types.

When the null hypothesis holds true, $-2\ln\Lambda(\mathbf{N})$ has approximately a χ^2_{k-1} distribution. Using (21.5) we obtain the the likelihood ratio test statistic

$$-2\ln\Lambda(\mathbf{N}) = -2\sum_{i=1}^{k} N_i \ln\frac{n\pi_i}{N_i} = 2\sum_{i=1}^{k} N_i \ln\frac{N_i}{n\pi_i}$$

The last equality uses the identity $\ln(1/x) = -\ln x$ for the logarithm of reciprocals.

The test statistic $-2\ln\Lambda_n(\mathbf{n})$ is generally rewritten using the notation $O_i = n_i$ for the number of **observed** occurrences of i and $E_i = n\pi_i$ for the number of **expected** occurrences of i as given by H_0. Then, we can write the test statistic as

$$-2\ln\Lambda_n(\mathbf{O}) = 2\sum_{i=1}^{k} O_i \ln\frac{O_i}{E_i} \tag{21.6}$$

This is called the G^2 **test statistic**. Thus, we can perform our inference on the hypothesis (**??**) by evaluating G^2. The p-value will be the probability that the a χ^2_{k-1} random variable takes a value greater than $-2\ln\Lambda_n(\mathbf{O})$

The traditional method for a test of goodness of fit, we use, instead of the G^2 statistic, the chi-square statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(E_i - O_i)^2}{E_i}. \tag{21.7}$$

This was introduced between 1895 and 1900 by Karl Pearson and consequently has been in use for longer that the concept of likelihood ratio tests. We establish the relation between (21.6) and (21.7), through the following two exercises.

Exercise 21.1. *Define*

$$\delta_i = \frac{O_i - E_i}{E_i} = \frac{O_i}{E_i} - 1.$$

Show that

$$\sum_{i=1}^{k} E_i\delta_i = 0 \quad \textit{and} \quad E_i(1+\delta_i) = O_i.$$

Exercise 21.2. *Show the relationship between the G^2 and χ^2 statistics in (21.6) and (21.7) by applying the quadratic Taylor polynomial approximation for the natural logarithm,*

$$\ln(1+\delta_i) \approx \delta_i - \frac{1}{2}\delta_i^2$$

and keeping terms up to the square of δ_i

To compute either the G^2 or χ^2 statistic, we begin by creating a table.

i	1	2	$\cdots$	k
observed	O_1	O_2	$\cdots$	O_k
expected	E_1	E_2	$\cdots$	E_k

We show this procedure using a larger data set on blood types.

Example 21.3. *The Red Cross recommends that a blood bank maintains 44% blood type O, 42% blood type A, 10% blood type B, 4% blood type AB. You suspect that the distribution of blood types in Tucson is not the same as the recommendation. In this case, the hypothesis is*

$$H_0 : p_O = 0.44, p_A = 0.42, p_B = 0.10, p_{AB} = 0.04 \quad \textit{versus} \quad H_1 : \textit{at least one } p_i \textit{ is unequal to the given values}$$

Based on 400 observations, we observe 228 for type O, 124 for type A, 40 for type B and 8 for type AB by computing $400 \times p_i$ using the values in H_0. This gives the table

type	O	A	B	AB
observed	228	124	40	8
expected	176	168	40	16

Using this table, we can compute the value of either (21.6) and (21.7). The `chisq.test` *command in* R *uses (21.7). The program computes the expected number of observations.*

```
> chisq.test(c(228,124,40,8),p=c(0.44,0.42,0.10,0.04))

Chi-squared test for given probabilities

data:  c(228, 124, 40, 8)
X-squared = 30.8874, df = 3, p-value = 8.977e-07
```

The number of degrees of freedom is $4 - 1 = 3$. Note that the p-value is very low and so the distribution of blood types in Tucson is very unlikely to be the same as the national distribution.

We can also perform the test using the G^2-statistic in (21.6):

```
> O<-c(228,124,40,8)
> E<-sum(O)*c(0.44,0.42,0.10,0.04)
> G2stat<-2*sum(O*log(O/E))
> G2stat
[1] 31.63731
> 1-pchisq(G2stat,3)
[1] 6.240417e-07
```

One way to visualize the discrepancies from from the null hypothesis is to display them with a **hanging chi-gram**. *This plots category i with a bar of height of the standardized residuals*

$$\frac{O_i - E_i}{\sqrt{E_i}}. \tag{21.8}$$

Note that these values can be either positive or negative.

```
> resid<-(O-E)/sqrt(E)
> barplot(resid, names.arg=c("O","A","B","AB"),
  xlab="chigram for blood donation data")
```

Example 21.4. *Is sudden infant death syndrome seasonal (SIDS)? Here we are hypothesizing that 1/4 of each of the occurrences of sudden infant death syndrome take place in the spring, summer, fall, and winter. Let p_1, p_2, p_3, and p_4 be the respective probabilities for these events. Then the hypothesis takes the form*

$$H_0 : p_1 = p_2 = p_3 = p_4 = \frac{1}{4}, \quad \textit{versus} \quad H_1 : \textit{at least one } p_i \textit{ is unequal to } \frac{1}{4}.$$

To test this hypothesis, public health officials from King County, Washington, collect data on $n = 322$ cases, finding

$$n_1 = 78, \quad n_2 = 71, \quad n_3 = 87, \quad n_4 = 86$$

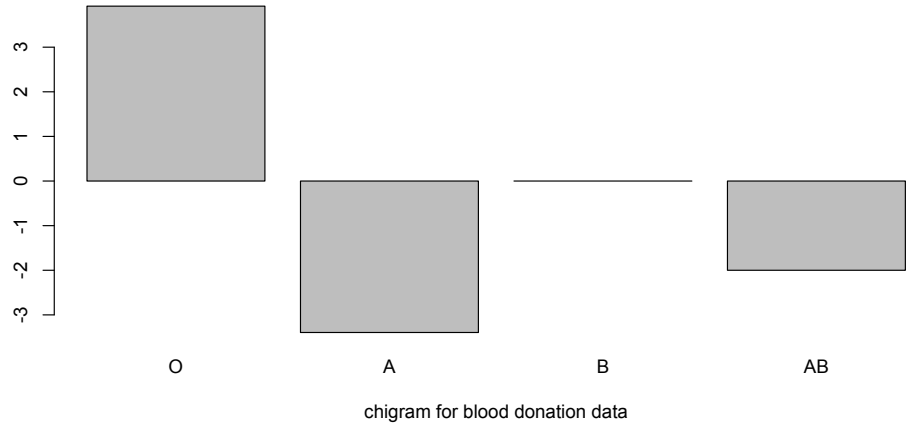

Figure 21.2: The heights of the bars for each category are the standardized scores (21.8). Thus, blood type O is overrepresented and types A and AB are underrepresented compare to the expectations under the null hypothesis.s

for deaths in the spring, summer, fall, and winter, respectively. Thus, we find more occurrences of SIDS in the fall and winter. Is this difference statistical significant or are these difference better explained by chance fluctuations?

We carry out the chi square test. In this case, each of the 4 categories is equally probable. Because this is the default value in R*, we need not include this in the command.*

```
> chisq.test(c(78,71,87,86))

Chi-squared test for given probabilities

data:  c(78, 71, 87, 86)
X-squared = 2.0994, df = 3, p-value = 0.552
```

This p-value is much to high to reject the null hypothesis.

Example 21.5 (Hardy-Weinberg equilibrium)**.** *As we saw with Gregor Mendel's pea experiments, the two-allele Hardy-Weinberg principle states that after two generations of random mating the genotypic frequencies can be represented by a binomial distribution. So, if a population is segregating for two alleles A_1 and A_2 at an autosomal locus with frequencies p_1 and p_2, then random mating would give a proportion*

$$p_{11} = p_1^2 \text{ for the } A_1A_1 \text{ genotype}, \; p_{12} = 2p_1p_2 \text{ for the } A_1A_2 \text{ genotype, and } p_{22} = p_2^2 \text{ for the } A_2A_2 \text{ genotype}. \quad (21.9)$$

Then, with both genes in the homozygous genotype and half the genes in the heterozygous genotype, we find that

$$p_1 = p_{11} + \frac{1}{2}p_{12} \quad p_2 = p_{22} + \frac{1}{2}p_{12}. \quad (21.10)$$

Our parameter space $\Theta = \{(p_{11}, p_{12}, p_{22}); p_{11} + p_{12} + p_{22} = 1\}$ is 2 dimensional. Θ_0, the parameter space for the null hypothesis, are those values p_1, p_2 that satisfy (21.10). With the choice of p_1, the value p_2 is determined because $p_1 + p_2 = 1$. Thus, $\dim(\Theta_0) = 1$. Consequently, the chi-square test statistic will have 2-1=1 degree of freedom. Another way to see this is the following.

McDonald et al. (1996) examined variation at the CVJ5 locus in the American oyster, Crassostrea virginica*. There were two alleles, L and S, and the genotype frequencies in Panacea, Florida were 14 LL, 21 LS, and 25 SS. So,*

$$\hat{p}_{11} = \frac{14}{60}, \quad \hat{p}_{12} = \frac{21}{60}, \quad \hat{p}_{22} = \frac{25}{60}.$$

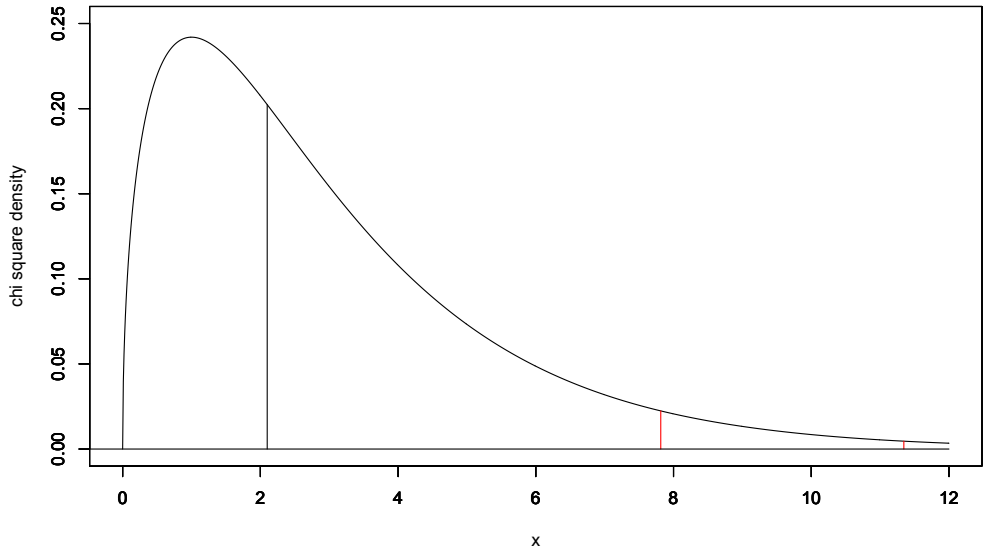

Figure 21.3: Plot of the chisquare density function with 3 degrees of freedom. The black vertical bar indicates the value of the test statistic in Example 21.3. The area 0.552 under the curve to the right of the vertical line is the p-value for this test. This is much to high to reject the null hypothesis. The red vertical lines show the critical values for a test with significance $\alpha = 0.05$ (to the left) and $\alpha = 0.01$ (to the right). Thus, the area under the curve to the right of these vertical lines is 0.05 and 0.01, respectively. These values can be found using `qchisq(1-α,3)`. We can also see that the test statistic value of 30.8874 in Example 21.3 has a very low p-value.

So, the estimate of p_1 and p_2 are

$$\hat{p}_1 = \hat{p}_{11} + \frac{1}{2}\hat{p}_{12} = \frac{49}{120}, \quad \hat{p}_2 = \hat{p}_{22} + \frac{1}{2}\hat{p}_{12} = \frac{71}{120}.$$

So, the expected number of observations is

$$E_{11} = 60\hat{p}_1^2 = 10.00417, \quad E_{12} = 60 \times 2\hat{p}_1\hat{p}_2 = 28.99167, \quad E_{22} = 60\hat{p}_2^2 = 21.00417.$$

The chi-square statistic

$$\chi^2 = \frac{(14-10)^2}{10} + \frac{(21-29)^2}{29} + \frac{(25-21)^2}{21} = 1.600 + 2.207 + 0.762 = 4.569$$

The p-value

```
> 1-pchisq(4.569,1)
[1] 0.03255556
```

Thus, we have moderate evidence against the null hypothesis of a Hardy-Weinberg equilibrium. Many forces may be the cause of this - non-random mating, selection, or migration to name a few possibilities.

Exercise 21.6. *Perform the chi-squared test using the G^2 statistic for the example above.*

21.2 Contingency tables

Contingency tables, also known as **two-way tables** or **cross tabulations** are a convenient way to display the frequency distribution from the observations of two categorical variables. For an $r \times c$ contingency table, we consider two factors A and B for an experiment. This gives r categories

$$A_1, \ldots A_r$$

for factor A and c categories

$$B_1, \ldots B_c$$

for factor B.

Here, we write O_{ij} to denote the number of occurrences for which an individual falls into both category A_i and category B_j. The results is then organized into a two-way table.

	B_1	B_2	$\cdots$	B_c	total
A_1	O_{11}	O_{12}	$\cdots$	O_{1c}	$O_{1\cdot}$
A_2	O_{21}	O_{22}	$\cdots$	O_{2c}	$O_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_r	O_{r1}	O_{r2}	$\cdots$	O_{rc}	$O_{r\cdot}$
total	$O_{\cdot 1}$	$O_{\cdot 2}$	$\cdots$	$O_{\cdot c}$	n

Example 21.7. *Returning to the study of the smoking habits of 5375 high school children in Tucson in 1967, here is a two-way table summarizing some of the results.*

	student smokes	student does not smoke	total
2 parents smoke	400	1380	1780
1 parent smokes	416	1823	2239
0 parents smoke	188	1168	1356
total	1004	4371	5375

For a contingency table, the null hypothesis we shall consider is that the factors A and B are independent. To set the parameters for this model, we define

$$p_{ij} = P\{\text{an individual is simultaneously a member of category } A_i \text{ and category } B_j\}.$$

Then, we have the parameter space

$$\Theta = \{\mathbf{p} = (p_{ij}, 1 \le i \le r, 1 \le j \le c); p_{ij} \ge 0 \text{ for all } i, j = 1, \sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij} = 1\}.$$

Write the **marginal distribution**

$$p_{i\cdot} = \sum_{j=1}^{c} p_{ij} = P\{\text{an individual is a member of category } A_i\}$$

and

$$p_{\cdot j} = \sum_{i=1}^{r} p_{ij} = P\{\text{an individual is a member of category } B_j\}.$$

The null hypothesis of independence of the categories A and B can be written

$$H_0 : p_{ij} = p_{i\cdot}p_{\cdot j}, \text{ for all } i, j \quad \text{versus} \quad H_1 : p_{ij} \ne p_{i\cdot}p_{\cdot j}, \text{ for some } i, j.$$

Follow the procedure as before for the goodness of fit test to end with a G^2 and its corresponding χ^2 test statistic. The G^2 statistic follows from the likelihood ratio test criterion. The χ^2 statistics is a second order Taylor series approximation to G^2.

$$-2\sum_{i=1}^{r}\sum_{j=1}^{c} O_{ij} \ln \frac{E_{ij}}{O_{ij}} \approx \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

The null hypothesis $p_{ij} = p_{i\cdot}p_{\cdot j}$ can be written in terms of observed and expected observations as

$$\frac{E_{ij}}{n} = \frac{O_{i\cdot}}{n}\frac{O_{\cdot j}}{n}.$$

or

$$E_{ij} = \frac{O_{i\cdot}O_{\cdot j}}{n}.$$

The test statistic, under the null hypothesis, has a χ^2 distribution. To determine the number of degrees of freedom, consider the following. Start with a contingency table with no entries but with the prescribed marginal values.

	B_1	B_2	$\cdots$	B_c	total
A_1					$O_{1\cdot}$
A_2					$O_{2\cdot}$
$\vdots$					$\vdots$
A_r					$O_{r\cdot}$
total	$O_{\cdot 1}$	$O_{\cdot 2}$	$\cdots$	$O_{\cdot c}$	n

The number of degrees of freedom is the number of values that we can place in the contingency table before all the remaining values are determined. To begin, fill in the first row with values $E_{11}, E_{12}, \ldots, E_{1,c-1}$. The final value $E_{1,c}$ in this determined by the other values in the row and the constraint that the row sum must be $O_{1\cdot}$. Continue filling the rows, noting that the value in column c is determined by the constraint on the row sum. Finally, when the time comes to fill in the bottom row r, notice that all the values are determined by the constraint on the row sums $O_{\cdot j}$. Thus, we can fill $c-1$ values in each of the $r-1$ rows before the remaining values are determined. Thus, the number of degrees of freedom is $(r-1)\times(c-1)$,

Example 21.8. *Returning to the data set on smoking habits in Tucson, we find that the expected table is*

	student smokes	student does not smoke	total
2 parents smoke	332.49	1447.51	1780
1 parent smokes	418.22	1820.78	2239
0 parents smoke	253.29	1102.71	1356
total	1004	4371	5375

For example,

$$E_{11} = \frac{O_{1\cdot}O_{\cdot 1}}{n} = \frac{1780\cdot 1004}{5375} = 332.49.$$

To compute the chi-square statistic

$$\begin{aligned}
& \tfrac{(400-332.49)^2}{332.49} + \tfrac{(1380-1447.51)^2}{1447.51} \\
+ \; & \tfrac{(416-418.22)^2}{418.22} + \tfrac{(1823-1820.78)^2}{1820.78} \\
+ \; & \tfrac{(188-253.29)^2}{253.29} + \tfrac{(1168-1102.71)^2}{1102.71} \\
= \; & 13.71 + 3.15 \\
+ \; & 0.012 + 0.003 \\
+ \; & 16.83 + 3.866 \\
= \; & 37.57
\end{aligned}$$

The number of degrees of freedom is $(r-1)\times(c-1)=(3-1)\times(2-1)=2$*. This can be seen by noting that one the first two entries in the "student smokes" column is filled, the rest are determined. Thus, the* p*-value*

```
> 1-pchisq(37.57,2)
[1] 6.946694e-09
```

is very small and leads us to reject the null hypothesis. Thus, we conclude that children smoking habits are not independent of their parents smoking habits. An examination of the individual cells shows that the children of parents who do not smoke are less likely to smoke and children who have two parents that smoke are more likely to smoke. Under the null hypothesis, each cell has a mean approximately 1 and so values much greater than 1 show contribution that leads to the rejection of H_0*.*

R *does the computation for us using the* `chisq.test` *command*

```
> smoking<-matrix(c(400,416,188,1380,1823,1168),nrow=3)
> smoking
     [,1] [,2]
[1,]  400 1380
[2,]  416 1823
[3,]  188 1168
> chisq.test(smoking)

Pearson's Chi-squared test

data:  smoking
X-squared = 37.5663, df = 2, p-value = 6.959e-09
```

We can look at the residuals $(O_{ij}-E_{ij})/\sqrt{E_{ij}}$ *for the entries in the* χ^2 *test as follows.*

```
> smokingtest<-chisq.test(smoking)
> residuals(smokingtest)
           [,1]        [,2]
[1,]  3.7025160 -1.77448934
[2,] -0.1087684  0.05212898
[3,] -4.1022973  1.96609088
```

Notice that if we square these values, we obtain the entries found in computing the test statistic.

```
> residuals(smokingtest)^2
            [,1]       [,2]
[1,] 13.70862455 3.14881241
[2,]  0.01183057 0.00271743
[3,] 16.82884348 3.86551335
```

Exercise 21.9. *Make three horizontally placed chigrams that summarize the residuals for this* χ^2 *test in the example above.*

Exercise 21.10 (two-by-two tables)**.** *Here is the contingency table can be thought of as two sets of Bernoulli trials as shown.*

	group 1	group 2	total
successes	x_1	x_2	x_1+x_2
failures	n_1-x_1	n_2-x_2	$(n_1+n_2)-(x_1+x_2)$
total	n_1	n_2	n_1+n_2

Show that the chi-square test is equivalent to the two-sided two sample proportion test.

21.3 Applicability and Alternatives to Chi-squared Tests

The chi-square test uses the central limit theorem and so is based on the ability to use a normal approximation. One criterion, the **Cochran conditions** requires no cell has count zero, and more than 80% of the cells have counts at least 5. If this does not hold, then **Fisher's exact test** uses the hypergeometric distribution (or its generalization) directly rather than normal approximation.

For example, for the 2×2 table,

	B_1	B_2	total
A_1	O_{11}	O_{12}	$O_{1\cdot}$
A_2	O_{21}	O_{22}	$O_{2\cdot}$
total	$O_{\cdot 1}$	$O_{\cdot 2}$	n

The idea behind Fisher's exact test is to begin with an empty table:

	B_1	B_2	total
A_1			$O_{1\cdot}$
A_2			$O_{2\cdot}$
total	$O_{\cdot 1}$	$O_{\cdot 2}$	n

and a null hypothesis that uses equally likely outcomes to fill in the table. We will use as an analogy the model of mark and recapture. Normally the goal is to find n, the total population. In this case, we assume that this population size is known and will consider the case that the individuals in the two captures are independent. This is assumed in the mark and recapture protocol. Here we test this independence.

In this regard,

- A_1 - an individual in the first capture and thus tagged.
- A_2 - an individual not in the first capture and thus not tagged.
- B_1 - an individual in the second capture.
- B_2 - an individual not in the second capture

Then, from the point of view of the A classification:

- We have $O_{1\cdot}$ from a population n with the A_1 classification (tagged individuals). This can be accomplished in
$$\binom{n}{O_{1\cdot}} = \frac{n!}{O_{1\cdot}!O_{2\cdot}!}$$
ways. The remaining $O_{2\cdot} = n - O_{1\cdot}$ have the A_2 classification (untagged individuals). Next, we fill in the values for the B classification
- From the $O_{\cdot 1}$ belonging to category B_1 (individuals in the second capture), O_{11} also belong to A_1 (have a tag). This outcome can be accomplished in
$$\binom{O_{\cdot 1}}{O_{11}} = \frac{O_{\cdot 1}!}{O_{11}!O_{21}!}$$
ways.
- From the $O_{\cdot 2}$ belonging to category B_2 (individuals not in the second capture), O_{12} also belong to A_1 (have a tag). This outcome can be accomplished in
$$\binom{O_{\cdot 2}}{O_{21}} = \frac{O_{\cdot 2}!}{O_{12}!O_{22}!}$$
ways.

Under the null hypothesis that every individual can be placed in any group, provided we have the given marginal information. In this case, the probability of the table above has the formula from the hypergeometric distribution

$$\frac{\binom{O_{1\cdot}}{O_{11}}\binom{O_{2\cdot}}{O_{21}}}{\binom{n}{O_{\cdot 1}}} = \frac{O_{\cdot 1}!/(O_{11}!O_{21}!)\cdot O_{\cdot 2}!/(O_{12}!O_{22}!)}{n!/(O_{1\cdot}!O_{2\cdot}!)} = \frac{O_{\cdot 1}!O_{\cdot 2}!O_{1\cdot}!O_{2\cdot}!}{O_{11}!O_{12}!O_{21}!O_{22}!n!}. \tag{21.11}$$

Notice that the formula is symmetric in the column and row variables. Thus, if we had derived the hypergeometric formula from the point of view of the B classification we would have obtained exactly the same formula (21.11).

To complete the exact test, we rely on statistical software to do the following:

- compute the hypergeometric probabilities over all possible choices for entries in the cells that result in the given marginal values, and
- rank these probabilities from most likely to least likely.
- Find the ranking of the actual data.
- For a one-sided test of too rare, the p-value is the sum of probabilities of the ranking lower than that of the data.

A similar procedure applies to provide the Fisher exact test for $r \times c$ tables.

Example 21.11. *As a test of the assumptions for mark and recapture. We examine a small population of 120 fish. The assumption are that each group of fish are equally likely to be capture in the first and second capture and that the two captures are independent. This could be violated, for example, if the tagged fish are not uniformly dispersed in the pond.*

Twenty-five are tagged and returned to the pond. For the second capture of 30, seven are tagged. With this information, given in red in the table below, we can complete the remaining entries.

	in 2nd capture	not in 2nd capture	total
in 1st capture	7	18	25
not in 1st capture	23	72	95
total	30	90	120

Fisher's exact test show a much too high p-value to reject the null hypothesis.

```
> fish<-matrix(c(7,23,18,72),ncol=2)
> fisher.test(fish)

Fisher's Exact Test for Count Data

data:  fish
p-value = 0.7958
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.3798574 3.5489546
sample estimates:
odds ratio
  1.215303
```

Exercise 21.12. *Perform the χ^2 test on the data set above and report the findings.*

Example 21.13. *We now return to a table on hemoglobin genotypes on two Indonesian islands. Recall that heterozygotes are protected against malaria.*

genotype	AA	AE	EE
Flores	128	6	0
Sumba	119	78	4

We noted that heterozygotes are rare on Flores and that it appears that malaria is less prevalent there since the heterozygote does not provide an adaptive advantage. Here are both the chi-square test and the Fisher exact test.

```
> genotype<-matrix(c(128,119,6,78,0,4),nrow=2)
> genotype
     [,1] [,2] [,3]
[1,]  128    6    0
[2,]  119   78    4
> chisq.test(genotype)

Pearson's Chi-squared test

data:  genotype
X-squared = 54.8356, df = 2, p-value = 1.238e-12

Warning message:
In chisq.test(genotype) : Chi-squared approximation may be incorrect
```

and

```
> fisher.test(genotype)

Fisher's Exact Test for Count Data

data:  genotype
p-value = 3.907e-15
alternative hypothesis: two.sided
```

Note that R cautions against the use of the chi-square test with these data.

21.4 Answer to Selected Exercise

21.1. For the first identity, using $\delta_i = (O_i - E_i)/E_i$.

$$\sum_{i=1}^{k} E_i \delta_i = \sum_{i=1}^{k} E_i \frac{O_i - E_i}{E_i} = \sum_{i=1}^{k} (O_i - E_i) = n - n = 0$$

and for the second

$$E_i(1 + \delta_i) = E_i \left(\frac{E_i}{E_i} + \frac{O_i - E_i}{E_i} \right) = E_i \frac{O_i}{E_i} = O_i.$$

21.2. We apply the quadratic Taylor polynomial approximation for the natural logarithm,

$$\ln(1 + \delta_i) \approx \delta_i - \frac{1}{2}\delta_i^2,$$

and use the identities in the previous exercise. Keeping terms up to the square of δ_i, we find that

$$\begin{aligned}
-2\ln\Lambda_n(\mathbf{O}) &= 2\sum_{i=1}^{k} O_i \ln\frac{O_i}{E_i} = 2\sum_{i=1}^{k} E_i(1+\delta_i)\ln(1+\delta_i) \\
&\approx 2\sum_{i=1}^{k} E_i(1+\delta_i)(\delta_i - \frac{1}{2}\delta_i^2) \approx 2\sum_{i=1}^{k} E_i(\delta_i + \frac{1}{2}\delta_i^2) \\
&= 2\sum_{i=1}^{k} E_i\delta_i + \sum_{i=1}^{k} E_i\delta_i^2 \\
&= 0 + \sum_{i=1}^{k} \frac{(E_i - O_i)^2}{E_i}.
\end{aligned}$$

21.6. Here is the R output.

```
> O<-c(14,21,25)
> phat<-c(O[1]+O[2]/2,O[3]+O[2]/2)/sum(O)
> phat
[1] 0.4083333 0.5916667
> E<-sum(O)*c(phat[1]^2,2*phat[1]*phat[2],phat[2]^2)
> E
[1] 10.00417 28.99167 21.00417
> sum(E)
[1] 60
> G2stat<-2*sum(O*log(O/E))
> G2stat
[1] 4.572896
> 1-pchisq(G2stat,1)
[1] 0.03248160
```

21.9. Here is the R output

```
> resid<-residuals(smokingtest)
> colnames(resid)<-c("smokes","does not smoke")
> par(mfrow=c(1,3))
> barplot(resid[1,],main="2 parents",ylim=c(-4.5,4.5))
> barplot(resid[2,],main="1 parent",ylim=c(-4.5,4.5))
> barplot(resid[3,],main="0 parents",ylim=c(-4.5,4.5))
```

21.10. The table of expected observations

	group 1	group 2	total
successes	$\frac{n_1(x_1+x_2)}{n_1+n_2}$	$\frac{n_2(x_1+x_2)}{n_1+n_2}$	x_1+x_2
failures	$\frac{n_1((n_1+n_2)-(x_1+x_2))}{n_1+n_2}$	$\frac{n_2((n_1+n_2)-(x_1+x_2))}{n_1+n_2}$	$(n_1+n_2)-(x_1+x_2)$
total	n_1	n_2	n_1+n_2

Now, write $\hat{p}_i = x_i/n_i$ for the sample proportions from each group, and

$$\hat{p}_0 = \frac{x_1+x_2}{n_1+n_2}$$

for the pooled sample proportion. Then we have the table of observed and expected observations

Figure 21.4: Chigram for the data on teen smoking in Tucson, 1967. R commands found in Exercise 21.9.

observed	group 1	group 2	total
successes	$n_1\hat{p}_1$	$n_2\hat{p}_2$	$(n_1+n_2)\hat{p}_0$
failures	$n_1(1-\hat{p}_1)$	$n_2(1-\hat{p}_2)$	$(n_1+n_2)(1-\hat{p}_0)$
total	n_1	n_2	n_1+n_2

expected	group 1	group 2	total
successes	$n_1\hat{p}_0$	$n_2\hat{p}_0$	$(n_1+n_2)\hat{p}_0$
failures	$n_1(1-\hat{p}_0)$	$n_2(1-\hat{p}_0)$	$(n_1+n_2)(1-\hat{p}_0)$
total	n_1	n_2	n_1+n_2

The chi-squared test statistic

$$\begin{array}{rcl}
 & \frac{(n_1(\hat{p}_1-\hat{p}_0))^2}{n_1\hat{p}_0} & + \quad \frac{(n_2(\hat{p}_2-\hat{p}_0))^2}{n_2\hat{p}_0} \\
+ & \frac{(n_1((1-\hat{p}_1)-(1-\hat{p}_0)))^2}{n_1(1-\hat{p}_0)} & + \quad \frac{(n_2((1-\hat{p}_2)+(1-\hat{p}_0)))^2}{n_2(1-\hat{p}_0)} \\
= & n_1\frac{(\hat{p}_1-\hat{p}_0)^2}{\hat{p}_0} & + \quad n_2\frac{(\hat{p}_2-\hat{p}_0)^2}{\hat{p}_0} \\
+ & n_1\frac{(\hat{p}_1-\hat{p}_0)^2}{(1-\hat{p}_0)} & + \quad n_2\frac{(\hat{p}_2-\hat{p}_0)^2}{(1-\hat{p}_0)} \\
= & n_1(\hat{p}_1-\hat{p}_0)^2\frac{1}{\hat{p}_0(1-p_0)} & + \quad n_2(\hat{p}_2-\hat{p}_0)^2\frac{1}{\hat{p}_0(1-p_0)} \\
= & \frac{n_1(\hat{p}_1-\hat{p}_0)^2+n_2(\hat{p}_2-\hat{p}_0)^2}{\hat{p}_0(1-p_0)} & = \quad -\ln\Lambda(\mathbf{x}_1,\mathbf{x}_2)
\end{array}$$

from the likelihood ratio computation for the two-sided two sample proportion test.

21.12. The R commands follow:

```
> chisq.test(fish)

Pearson's Chi-squared test with Yates' continuity correction

data:  fish
X-squared = 0.0168, df = 1, p-value = 0.8967
```

The p-value is notably higher for the χ^2 test.