17

Chi-Square Tests

The chi-square statistic for testing hypotheses concerning multinomial distributions derives its name from the asymptotic approximation to its distribution. Two important applications are the testing of independence in a two-way classification and the testing of goodness-of-fit. In the second application the multinomial distribution is created artificially by grouping the data, and the asymptotic chi-square approximation may be lost if the original data are used to estimate nuisance parameters.

17.1 Quadratic Forms in Normal Vectors

The *chi-square distribution* with k degrees of freedom is (by definition) the distribution of $\sum_{i=1}^{k} Z_i^2$ for i.i.d. $N(0, 1)$-distributed variables $Z_1, \ldots, Z_k$. The sum of squares is the squared norm $\|Z\|^2$ of the standard normal vector $Z = (Z_1, \ldots, Z_k)$. The following lemma gives a characterization of the distribution of the norm of a general zero-mean normal vector.

17.1 ***Lemma.*** *If the vector X is $N_k(0, \Sigma)$-distributed, then $\|X\|^2$ is distributed as $\sum_{i=1}^{k} \lambda_i Z_i^2$ for i.i.d. $N(0, 1)$-distributed variables $Z_1, \ldots, Z_k$ and $\lambda_1, \ldots, \lambda_k$ the eigenvalues of Σ.*

Proof. There exists an orthogonal matrix O such that $O\Sigma O^T = \operatorname{diag}(\lambda_i)$. Then the vector OX is $N_k\big(0, \operatorname{diag}(\lambda_i)\big)$-distributed, which is the same as the distribution of the vector $(\sqrt{\lambda_1} Z_1, \ldots, \sqrt{\lambda_k} Z_k)$. Now $\|X\|^2 = \|OX\|^2$ has the same distribution as $\sum(\sqrt{\lambda_i} Z_i)^2$. ■

The distribution of a quadratic form of the type $\sum_{i=1}^{k} \lambda_i Z_i^2$ is complicated in general. However, in the case that every λ_i is either 0 or 1, it reduces to a chi-square distribution. If this is not naturally the case in an application, then a statistic is often transformed to achieve this desirable situation. The definition of the Pearson statistic illustrates this.

17.2 Pearson Statistic

Suppose that we observe a vector $X_n = (X_{n,1}, \ldots, X_{n,k})$ with the multinomial distribution corresponding to n trials and k classes having probabilities $p = (p_1, \ldots, p_k)$. The *Pearson*

statistic for testing the null hypothesis $H_0 : p = a$ is given by

$$C_n(a) = \sum_{i=1}^{k} \frac{(X_{n,i} - na_i)^2}{na_i}.$$

We shall show that the sequence $C_n(a)$ converges in distribution to a chi-square distribution if the null hypothesis is true. The practical relevance is that we can use the chi-square table to find critical values for the test. The proof shows why Pearson divided the squares by na_i and did not propose the simpler statistic $\|X_n - na\|^2$.

17.2 ***Theorem.*** *If the vectors X_n are multinomially distributed with parameters n and $a = (a_1, \ldots, a_k) > 0$, then the sequence $C_n(a)$ converges under a in distribution to the χ^2_{k-1}-distribution.*

Proof. The vector X_n can be thought of as the sum of n independent multinomial vectors $Y_1, \ldots, Y_n$ with parameters 1 and $a = (a_1, \ldots, a_k)$. Then

$$\mathrm{E}Y_i = a, \qquad \mathrm{Cov}\, Y_i = \begin{pmatrix} a_1(1-a_1) & -a_1a_2 & \cdots & -a_1a_k \\ -a_2a_1 & a_2(1-a_2) & \cdots & -a_2a_k \\ \vdots & \vdots & & \vdots \\ -a_ka_1 & -a_ka_2 & \cdots & a_k(1-a_k) \end{pmatrix}.$$

By the multivariate central limit theorem, the sequence $n^{-1/2}(X_n - na)$ converges in distribution to the $N_k(0, \mathrm{Cov}\, Y_1)$-distribution. Consequently, with $\sqrt{a}$ the vector with coordinates $\sqrt{a_i}$,

$$\left(\frac{X_{n,1} - na_1}{\sqrt{na_1}}, \ldots, \frac{X_{n,k} - na_k}{\sqrt{na_k}} \right) \rightsquigarrow N(0, I - \sqrt{a}\sqrt{a}^T).$$

Because $\sum a_i = 1$, the matrix $I - \sqrt{a}\sqrt{a}^T$ has eigenvalue 0, of multiplicity 1 (with eigenspace spanned by $\sqrt{a}$), and eigenvalue 1, of multiplicity $(k-1)$ (with eigenspace equal to the orthocomplement of $\sqrt{a}$). An application of the continuous-mapping theorem and next Lemma 17.1 conclude the proof. ■

The number of degrees of freedom in the chi-squared approximation for Pearson's statistic is the number of cells of the multinomial vector that have positive probability. However, the quality of the approximation also depends on the size of the cell probabilities a_j. For instance, if 1001 cells have null probabilities $10^{-23}, \ldots, 10^{-23}, 1 - 10^{-20}$, then it is clear that for moderate values of n all cells except one are empty, and a huge value of n is necessary to make a χ^2_{1000}-approximation work. As a rule of thumb, it is often advised to choose the partitioning sets such that each number na_j is at least 5. This criterion depends on the (possibly unknown) null distribution and is not the same as saying that the number of observations in each cell must satisfy an absolute lower bound, which could be very unlikely if the null hypothesis is false. The rule of thumb means to protect the level.

The Pearson statistic is oddly asymmetric in the observed and the true frequencies (which is motivated by the form of the asymptotic covariance matrix). One method to symmetrize

the statistic leads to the *Hellinger statistic*

$$H_n^2(a) = 4\sum_{i=1}^{k} \frac{(X_{n,i} - na_i)^2}{(\sqrt{X_{n,i}} + \sqrt{na_i})^2} = 4\sum_{i=1}^{n} (\sqrt{X_{n,i}} - \sqrt{na_i})^2.$$

Up to a multiplicative constant this is the Hellinger distance between the discrete probability distributions on $\{1, \ldots, k\}$ with probability vectors a and X_n/n, respectively. Because $X_n/n - a \xrightarrow{P} 0$, the Hellinger statistic is asymptotically equivalent to the Pearson statistic.

17.3 Estimated Parameters

Chi-square tests are used quite often, but usually to test more complicated hypotheses. If the null hypothesis of interest is composite, then the parameter a is unknown and cannot be used in the definition of a test statistic. A natural extension is to replace the parameter by an estimate $\hat{a}_n$ and use the statistic

$$C_n(\hat{a}_n) = \sum_{i=1}^{k} \frac{(X_{n,i} - n\hat{a}_{n,i})^2}{n\hat{a}_{n,i}}.$$

The estimator $\hat{a}_n$ is constructed to be a good estimator if the null hypothesis is true. The asymptotic distribution of this modified Pearson statistic is not necessarily chi-square but depends on the estimators $\hat{a}_n$ being used. Most often the estimators are asymptotically normal, and the statistics

$$\frac{X_{n,i} - n\hat{a}_{n,i}}{\sqrt{n\hat{a}_{n,i}}} = \frac{X_{n,i} - na_{n,i}}{\sqrt{n\hat{a}_{n,i}}} - \frac{\sqrt{n}(\hat{a}_{n,i} - a_{n,i})}{\sqrt{\hat{a}_{n,i}}}$$

are asymptotically normal as well. Then the modified chi-square statistic is asymptotically distributed as a quadratic form in a multivariate-normal vector. In general, the eigenvalues determining this form are not restricted to 0 or 1, and their values may depend on the unknown parameter. Then the critical value cannot be taken from a table of the chi-square distribution. There are two popular possibilities to avoid this problem.

First, the Pearson statistic is a certain quadratic form in the observations that is motivated by the asymptotic covariance matrix of a multinomial vector. If the parameter a is estimated, the asymptotic covariance matrix changes in form, and it is natural to change the quadratic form in such a way that the resulting statistic is again chi-square distributed. This idea leads to the Rao-Robson-Nikulin modification of the Pearson statistic, of which we discuss an example in section 17.5.

Second, we can retain the form of the Pearson statistic but use special estimators $\hat{a}$. In particular, the maximum likelihood estimator based on the multinomial vector X_n, or the *minimum–chi square estimator* $\bar{a}_n$ defined by, with $\mathcal{P}_0$ being the null hypothesis,

$$\sum_{i=1}^{k} \frac{(X_{n,i} - n\bar{a}_{n,i})^2}{n\bar{a}_{n,i}} = \inf_{p \in \mathcal{P}_0} \sum_{i=1}^{k} \frac{(X_{n,i} - np_i)^2}{np_i}.$$

The right side of this display is the "minimum–chi square distance" of the observed frequencies to the null hypothesis and is an intuitively reasonable test statistic. The null hypothesis

is rejected if the distance of the observed frequency vector X_n/n to the set $\mathcal{P}_0$ is large. A disadvantage is greater computational complexity.

These two modifications, using the minimum–chi square estimator or the maximum likelihood estimator based on X_n, may seem natural but are artificial in some applications. For instance, in goodness-of-fit testing, the multinomial vector is formed by grouping the "raw data," and it is more natural to base the estimators on the raw data rather than on the grouped data. On the other hand, using the maximum likelihood or minimum–chi square estimator based on X_n has the advantage of a remarkably simple limit theory: If the null hypothesis is "locally linear," then the modified Pearson statistic is again asymptotically chi-square distributed, but with the number of degrees of freedom reduced by the (local) dimension of the estimated parameter.

This interesting asymptotic result is most easily explained in terms of the minimum–chi square statistic, as the loss of degrees of freedom corresponds to a projection (i.e., a minimum distance) of the limiting normal vector. We shall first show that the two types of modifications are asymptotically equivalent and are asymptotically equivalent to the likelihood ratio statistic as well. The likelihood ratio statistic for testing the null hypothesis $H_0 : p \in \mathcal{P}_0$ is given by (see Example 16.1)

$$L_n(\hat{a}_n) = \inf_{p \in \mathcal{P}_0} L_n(p), \qquad L_n(p) = 2\sum_{i=1}^{k} X_{n,i} \log \frac{X_{n,i}}{np_i}.$$

17.3 ***Lemma.*** *Let $\mathcal{P}_0$ be a closed subset of the unit simplex, and let $\hat{a}_n$ be the maximum likelihood estimator of a under the null hypothesis $H_0 : a \in \mathcal{P}_0$ (based on X_n). Then*

$$\inf_{p \in \mathcal{P}_0} \sum_{i=1}^{k} \frac{(X_{n,i} - np_i)^2}{np_i} = C_n(\hat{a}_n) + o_P(1) = L_n(\hat{a}_n) + o_P(1).$$

Proof. Let $\overline{a}_n$ be the minimum–chi square estimator of a under the null hypothesis. Both sequences of estimators $\overline{a}_n$ and $\hat{a}_n$ are $\sqrt{n}$-consistent. For the maximum likelihood estimator this follows from Corollary 5.53. The minimum–chi square estimator satisfies by its definition

$$\sum_{i=1}^{k} \frac{(X_{n,i} - n\overline{a}_{n,i})^2}{n\overline{a}_{n,i}} \le \sum_{i=1}^{k} \frac{(X_{n,i} - na_i)^2}{na_i} = O_P(1).$$

This implies that each term in the sum on the left is $O_P(1)$, whence $n|\overline{a}_{n,i} - a_i|^2 = O_P(\overline{a}_{n,i}) + O_P\big(|X_{n,i} - na_i|^2/n\big)$ and hence the $\sqrt{n}$-consistency.

Next, the two-term Taylor expansion $\log(1 + x) = x - \frac{1}{2}x^2 + o(x^2)$ combined with Lemma 2.12 yields, for any $\sqrt{n}$-consistent estimator sequence $\hat{p}_n$,

$$\begin{aligned}\sum_{i=1}^{k} X_{n,i} \log \frac{X_{n,i}}{n\hat{p}_{n,i}} &= -\sum_{i=1}^{k} X_{n,i}\left(\frac{n\hat{p}_{n,i}}{X_{n,i}} - 1\right) + \frac{1}{2}\sum_{i=1}^{k} X_{n,i}\left(\frac{n\hat{p}_{n,i}}{X_{n,i}} - 1\right)^2 + o_P(1) \\ &= 0 + \frac{1}{2}\sum_{i=1}^{k} \frac{(X_{n,i} - n\hat{p}_{n,i})^2}{X_{n,i}} + o_P(1).\end{aligned}$$

In the last expression we can also replace $X_{n,i}$ in the denominator by $n\hat{p}_{n,i}$, so that we find the relation $L_n(\hat{p}_n) = C_n(\hat{p}_n)$ between the likelihood ratio and the Pearson statistic, for

every $\sqrt{n}$-consistent estimator sequence $\hat{p}_n$. By the definitions of $\overline{a}_n$ and $\hat{a}_n$, we conclude that, up to $o_P(1)$-terms, $C_n(\overline{a}_n) \le C_n(\hat{a}_n) = L_n(\hat{a}_n) \le L_n(\overline{a}_n) = C_n(\overline{a}_n)$. The lemma follows. ■

The asymptotic behavior of likelihood ratio statistics is discussed in general in Chapter 16. In view of the preceding lemma, we can now refer to this chapter to obtain the asymptotic distribution of the chi-square statistics. Alternatively, a direct study of the minimum–chi square statistic gives additional insight (and a more elementary proof).

As in Chapter 16, say that a sequence of sets H_n converges to a set H if H is the set of all limits $\lim h_n$ of converging sequences h_n with $h_n \in H_n$ for every n and, moreover, the limit $h = \lim_i h_{n_i}$ of every converging subsequence h_{n_i} with $h_{n_i} \in H_{n_i}$ for every i is contained in H.

17.4 ***Theorem.*** *Let $\mathcal{P}_0$ be a subset of the unit simplex such that the sequence of sets $\sqrt{n}(\mathcal{P}_0 - a)$ converges to a set H (in $\mathbb{R}^k$), and suppose that $a > 0$. Then, under a,*

$$\inf_{p \in \mathcal{P}_0} \sum_{i=1}^{k} \frac{(X_{n,i} - np_i)^2}{np_i} \rightsquigarrow \inf_{h \in H} \left\| X - \frac{1}{\sqrt{a}} H \right\|^2,$$

for a vector X with the $N(0, I - \sqrt{a}\sqrt{a}^T)$-distribution. Here $(1/\sqrt{a})H$ is the set of vectors $(h_1/\sqrt{a_1}, \ldots, h_k/\sqrt{a_k})$ as h ranges over H.

17.5 ***Corollary.*** *Let $\mathcal{P}_0$ be a subset of the unit simplex such that the sequence of sets $\sqrt{n}(\mathcal{P}_0 - a)$ converges to a linear subspace of dimension l (of $\mathbb{R}^k$), and let $a > 0$. Then both the sequence of minimum–chi square statistics and the sequence of modified Pearson statistics $C_n(\hat{a}_n)$ converge in distribution to the chi-square distribution with $k-1-l$ degrees of freedom.*

Proof. Because the minimum–chi square estimator $\overline{a}_n$ (relative to $\overline{\mathcal{P}}_0$) is $\sqrt{n}$-consistent, the asymptotic distribution of the minimum–chi square statistic is not changed if we replace $n\overline{a}_{n,i}$ in its denominator by the true value na_i. Next, we decompose,

$$\frac{X_{n,i} - np_i}{\sqrt{na_i}} = \frac{X_{n,i} - na_i}{\sqrt{na_i}} - \frac{\sqrt{n}(p_i - a_i)}{\sqrt{a_i}}.$$

The first vector on the right converges in distribution to X. The (modified) minimum–chi square statistics are the distances of these vectors to the sets $H_n = \sqrt{n}(\mathcal{P}_0 - a)/\sqrt{a}$, which converge to the set $H/\sqrt{a}$. The theorem now follows from Lemma 7.13.

The vector X is distributed as $Z - \Pi_{\sqrt{a}} Z$ for $\Pi_{\sqrt{a}}$ the projection onto the linear space spanned by the vector $\sqrt{a}$ and Z a k-dimensional standard normal vector. Because every element of H is the limit of a multiple of differences of probability vectors, $1^T h = 0$ for every $h \in H$. Therefore, the space $(1/\sqrt{a})H$ is orthogonal to the vector $\sqrt{a}$, and $\Pi\,\Pi_{\sqrt{a}} = 0$ for Π the projection onto the space $(1/\sqrt{a})H$. The distance of X to the space $(1/\sqrt{a})H$ is equal to the norm of $X - \Pi X$, which is distributed as the norm of $Z - \Pi_{\sqrt{a}} Z - \Pi Z$. The latter projection is multivariate normally distributed with mean zero and covariance matrix the projection matrix $I - \Pi_{\sqrt{a}} - \Pi$ with $k - l - 1$ eigenvalues 1. The corollary follows from Lemma 17.1 or 16.6. ■

17.6 ***Example (Parametric model).*** If the null hypothesis is a parametric family $\mathcal{P}_0 = \{p_\theta : \theta \in \Theta\}$ indexed by a subset Θ of $\mathbb{R}^l$ with $l \leq k$ and the maps $\theta \mapsto p_\theta$ from Θ into the unit simplex are differentiable and of full rank, then $\sqrt{n}(\mathcal{P}_0 - p_\theta) \to \dot{p}_\theta(\mathbb{R}^l)$ for every $\theta \in \mathring{\Theta}$ (see Example 16.11). Then the chi-square statistics $C_n(\hat{p}_\theta)$ are asymptotically χ^2_{k-l-1}-distributed.

This situation is common in testing the goodness-of-fit of parametric families, as discussed in section 17.5 and Example 16.1. □

17.4 Testing Independence

Suppose that each element of a population can be classified according to two characteristics, having k and r levels, respectively. The full information concerning the classification can be given by a $(k \times r)$ *table* of the form given in Table 17.1.

Often the full information is not available, but we do know the classification $X_{n,ij}$ for a random sample of size n from the population. The matrix $X_{n,ij}$, which can also be written in the form of a $(k \times r)$ table, is multinomially distributed with parameters n and probabilities $p_{ij} = N_{ij}/N$. The null *hypothesis of independence* asserts that the two categories are independent: $H_0 : p_{ij} = a_i b_j$ for (unknown) probability vectors a_i and b_j.

The maximum likelihood estimators for the parameters a and b (under the null hypothesis) are $\hat{a}_i = X_{n,i.}/n$ and $\hat{b}_j = X_{n,.j}/n$. With these estimators the modified Pearson statistic takes the form

$$C_n(\hat{a}_n \otimes \hat{b}_n) = \sum_{i=1}^{k} \sum_{j=1}^{r} \frac{(X_{n,ij} - n\hat{a}_i\hat{b}_j)^2}{n\hat{a}_i\hat{b}_j}.$$

The null hypothesis is a $(k + r - 2)$-dimensional submanifold of the unit simplex in $\mathbb{R}^{kr}$. In a shrinking neighborhood of a parameter in its interior this manifold looks like its tangent space, a linear space of dimension $k + r - 2$. Thus, the sequence $C_n(\hat{a}_n \otimes \hat{b}_n)$ is asymptotically chi square–distributed with $kr - 1 - (k + r - 2) = (k - 1)(r - 1)$ degrees of freedom.

Table 17.1. *Classification of a population of N elements according to two categories, N_{ij} elements having value i on the first category and value j on the second. The borders give the sums over each row and column, respectively.*

N_{11}	N_{12}	$\cdots$	N_{1r}	$N_{1.}$
N_{21}	N_{22}	$\cdots$	N_{1r}	$N_{2.}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
N_{k1}	N_{k2}	$\cdots$	N_{1r}	$N_{k.}$
$N_{.1}$	$N_{.2}$	$\cdots$	$N_{.r}$	N

17.7 ***Corollary.*** *If the $(k \times r)$ matrices X_n are multinomially distributed with parameters n and $p_{ij} = a_i b_j > 0$, then the sequence $C_n(\hat{a}_n \otimes \hat{b}_n)$ converges in distribution to the $\chi^2_{(k-1)(r-1)}$-distribution.*

Proof. The map $(a_1, \ldots, a_{k-1}, b_1, \ldots, b_{r-1}) \mapsto (a \times b)$ from $\mathbb{R}^{k+r-2}$ into $\mathbb{R}^{kr}$ is continuously differentiable and of full rank. The true values $(a_1, \ldots, a_{k-1}, b_1 \ldots, b_{r-1})$ are interior to the domain of this map. Thus the sequence of sets $\sqrt{n}(\mathcal{P}_0 - a \times b)$ converges to a $(k + r - 2)$-dimensional linear subspace of $\mathbb{R}^{kr}$. ■

*17.5 Goodness-of-Fit Tests

Chi-square tests are often applied to test goodness-of-fit. Given a random sample $X_1, \ldots, X_n$ from a distribution P, we wish to test the null hypothesis $H_0 : P \in \mathcal{P}_0$ that P belongs to a given class $\mathcal{P}_0$ of probability measures. There are many possible test statistics for this problem, and a particular statistic might be selected to attain high power against certain alternatives. Testing goodness-of-fit typically focuses on no particular alternative. Then chi-square statistics are intuitively reasonable.

The data can be reduced to a multinomial vector by "grouping." We choose a partition $\mathcal{X} = \cup_j \mathcal{X}_j$ of the sample space into finitely many sets and base the test only on the observed numbers of observations falling into each of the sets $\mathcal{X}_j$. For ease of notation, we express these numbers into the empirical measure of the data. For a given set A we denote by $\mathbb{P}_n(A) = n^{-1}(1 \le i \le n : X_i \in A)$ the fraction of observations that fall into A. Then the vector $n\big(\mathbb{P}_n(\mathcal{X}_1), \ldots, \mathbb{P}_n(\mathcal{X}_k)\big)$ possesses a multinomial distribution, and the corresponding modified chi-square statistic is given by

$$\sum_{i=1}^{k} \frac{n\big(\mathbb{P}_n(\mathcal{X}_j) - \hat{P}(\mathcal{X}_j)\big)^2}{\hat{P}(\mathcal{X}_j)}.$$

Here $\hat{P}(\mathcal{X}_j)$ is an estimate of $P(\mathcal{X}_j)$ under the null hypothesis and can take a variety of forms.

Theorem 17.4 applies but is restricted to the case that the estimates $\hat{P}(\mathcal{X}_j)$ are based on the frequencies $n\big(\mathbb{P}_n(\mathcal{X}_1), \ldots, \mathbb{P}_n(\mathcal{X}_k)\big)$ only. In the present situation it is more natural to base the estimates on the original observations $X_1, \ldots, X_n$. Usually, this results in a non–chi square limit distribution. For instance, Table 17.2 shows the "errors" in the level of a chi-square test for testing normality, if the unknown mean and variance are estimated by the sample mean and the sample variance but the critical value is chosen from the chi-square distribution. The size of the errors depends on the numbers of cells, the errors being small if there are many cells and few estimated parameters.

17.8 ***Example (Parametric model).*** Consider testing the null hypothesis that the true distribution belongs to a regular parametric model $\{P_\theta : \theta \in \Theta\}$. It appears natural to estimate the unknown parameter θ by an estimator $\hat{\theta}_n$ that is asymptotically efficient under the null hypothesis and is based on the original sample $X_1, \ldots, X_n$, for instance the maximum likelihood estimator. If $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P_\theta)$ denotes the empirical process, then efficiency entails the approximation $\sqrt{n}(\hat{\theta}_n - \theta) = I_\theta^{-1} \mathbb{G}_n \dot{\ell}_\theta + o_P(1)$. Applying the delta method to

Table 17.2. *True levels of the chi-square test for normality using $\chi^2_{k-3,\alpha}$-quantiles as critical values but estimating unknown mean and variance by sample mean and sample variance. Chi square statistic based on partitions of $[-10, 10]$ into $k = 5, 10,$ or 20 equiprobable cells under the standard normal law.*

	$\alpha = 0.20$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
$k = 5$	0.30	0.15	0.08	0.02
$k = 10$	0.22	0.11	0.06	0.01
$k = 20$	0.21	0.10	0.05	0.01

Note: Values based on 2000 simulations of standard normal samples of size 100.

the variables $\sqrt{n}\big(P_{\hat\theta}(\mathcal{X}_j) - P_\theta(\mathcal{X}_j)\big)$ and using Slutsky's lemma, we find

$$\frac{\sqrt{n}\big(\mathbb{P}_n(\mathcal{X}_j) - P_{\hat\theta}(\mathcal{X}_j)\big)}{\sqrt{P_{\hat\theta}(\mathcal{X}_j)}} = \frac{\mathbb{G}_n 1_{\mathcal{X}_j} - (P_\theta 1_{\mathcal{X}_j} \dot\ell_\theta)^T I_\theta^{-1} \mathbb{G}_n \dot\ell_\theta}{\sqrt{P_\theta(\mathcal{X}_j)}} + o_P(1).$$

(The map $\theta \mapsto P_\theta(A)$ has derivative $P_\theta 1_A \dot\ell_\theta$.) The sequence of vectors $(\mathbb{G}_n 1_{\mathcal{X}_j}, \mathbb{G}_n \dot\ell_\theta)$ converges in distribution to a multivariate-normal distribution. Some matrix manipulations show that the vectors in the preceding display are asymptotically distributed as a Gaussian vector X with mean zero and covariance matrix

$$I - \sqrt{a_\theta}\sqrt{a_\theta}^T - C_\theta^T I_\theta^{-1} C_\theta, \qquad (a_\theta)_j = P_\theta(\mathcal{X}_j), \qquad (C_\theta)_{ij} = \frac{P_\theta 1_{\mathcal{X}_j} \dot\ell_{\theta,i}}{\sqrt{(a_\theta)_j}}.$$

In general, the covariance matrix of X is not a projection matrix, and the variable $\|X\|^2$ does not possess a chi-square distribution.

Because $P_\theta \dot\ell_\theta = 0$, we have that $C_\theta \sqrt{a_\theta} = 0$ and hence the covariance matrix of X can be rewritten as the product $(I - \sqrt{a_\theta}\sqrt{a_\theta}^T)(I - C_\theta^T I_\theta^{-1} C_\theta)$. Here the first matrix is the projection onto the orthocomplement of the vector $\sqrt{a_\theta}$ and the second matrix is a positive-definite transformation that leaves $\sqrt{a_\theta}$ invariant, thus acting only on the orthocomplement $\sqrt{a_\theta}^\perp$. This geometric picture shows that $\operatorname{Cov}_\theta X$ has the same system of eigenvectors as the matrix $I - C_\theta^T I_\theta^{-1} C_\theta$, and also the same eigenvalues, except for the eigenvalue corresponding to the eigenvector $\sqrt{a_\theta}$, which is 0 for $\operatorname{Cov}_\theta X$ and 1 for $I - C_\theta^T I_\theta^{-1} C_\theta$. Because both matrices $C_\theta^T I_\theta^{-1} C_\theta$ and $I - C_\theta^T I_\theta^{-1} C_\theta$ are nonnegative-definite, the eigenvalues are contained in $[0, 1]$. One eigenvalue (corresponding to eigenvector $\sqrt{a_\theta}$) is 0, $\dim \mathrm{N}(C_\theta) - 1$ eigenvalues (corresponding to eigenspace $\mathrm{N}(C_\theta) \cap \sqrt{a_\theta}^\perp$) are 1, but the other eigenvalues may be contained in $(0, 1)$ and then typically depend on θ. By Lemma 17.1, the variable $\|X\|^2$ is distributed as

$$\sum_{i=1}^{\dim \mathrm{N}(C_\theta)-1} Z_i^2 + \sum_{i=\dim \mathrm{N}(C_\theta)}^{k-1} \lambda_i(\theta) Z_i^2.$$

This means that it is stochastically "between" the chi-square distributions with $\dim \mathrm{N}(C_\theta) - 1$ and $k - 1$ degrees of freedom.

The inconvenience that this distribution is not standard and depends on θ can be remedied by not using efficient estimators $\hat\theta_n$ or, alternatively, by not using the Pearson statistic.

The square root of the matrix $I - C_\theta^T I_\theta^{-1} C_\theta$ is the positive-definite matrix with the same eigenvectors, but with the square roots of the eigenvalues. Thus, it also leaves the vector $\sqrt{a_\theta}$ invariant and acts only on the orthocomplement $\sqrt{a_\theta}^\perp$. It follows that this square root commutes with the matrix $I - \sqrt{a_\theta}\sqrt{a_\theta}^T$ and hence

$$\big(I - C_{\hat\theta}^T I_{\hat\theta}^{-1} C_{\hat\theta}\big)^{-1/2} \frac{\sqrt{n}\big(\mathbb{P}_n(\mathcal{X}_j) - P_{\hat\theta}(\mathcal{X}_j)\big)}{\sqrt{P_{\hat\theta}(\mathcal{X}_j)}} \rightsquigarrow N_k\big(0, I - \sqrt{a_\theta}\sqrt{a_\theta}^T\big).$$

(We assume that the matrix $I - C_\theta^T I_\theta^{-1} C_\theta$ is nonsingular, which is typically the case; see problem 17.6). By the continuous-mapping theorem, the squared norm of the left side is asymptotically chi square–distributed with $k - 1$ degrees of freedom. This squared norm is the *Rao-Robson-Nikulin statistic*. □

It is tempting to choose the partitioning sets $\mathcal{X}_j$ dependent on the observed data $X_1, \ldots, X_n$, for instance to ensure that all cells have positive probability under the null hypothesis. This is permissible under some conditions: The choice of a "random partition" typically does not change the distributional properties of the chi-square statistic. Consider partitioning sets $\hat{\mathcal{X}}_j = \mathcal{X}_j(X_1, \ldots, X_n)$ that possibly depend on the data, and a further modified Pearson statistic of the type

$$\sum_{i=1}^k \frac{n\big(\mathbb{P}_n(\hat{\mathcal{X}}_j) - \hat{P}(\hat{\mathcal{X}}_j)\big)^2}{\hat{P}(\hat{\mathcal{X}}_j)}.$$

If the random partitions settle down to a fixed partition eventually, then this statistic is asymptotically equivalent to the statistic for which the partition had been set equal to the limit partition in advance. We discuss this for the case that the null hypothesis is a model $\{P_\theta : \theta \in \Theta\}$ indexed by a subset Θ of a normed space. We use the language of Donsker classes as discussed in Chapter 19.

17.9 ***Theorem.*** *Suppose that the sets $\hat{\mathcal{X}}_j$ belong to a P_{θ_0}-Donsker class $\mathcal{C}$ of sets and that $P_{\theta_0}(\hat{\mathcal{X}}_j \triangle \mathcal{X}_j) \xrightarrow{\mathrm{P}} 0$ under P_{θ_0}, for given nonrandom sets $\mathcal{X}_j$ such that $P_{\theta_0}(\mathcal{X}_j) > 0$. Furthermore, assume that $\sqrt{n}\|\hat\theta - \theta_0\| = O_P(1)$, and suppose that the map $\theta \mapsto P_\theta$ from Θ into $\ell^\infty(\mathcal{C})$ is differentiable at θ_0 with derivative $\dot P_{\theta_0}$ such that $\dot P_{\theta_0}(\hat{\mathcal{X}}_j) - \dot P_{\theta_0}(\mathcal{X}_j) \xrightarrow{\mathrm{P}} 0$ for every j. Then*

$$\sum_{i=1}^k \frac{n\big(\mathbb{P}_n(\hat{\mathcal{X}}_j) - P_{\hat\theta}(\hat{\mathcal{X}}_j)\big)^2}{P_{\hat\theta}(\hat{\mathcal{X}}_j)} = \sum_{i=1}^k \frac{n\big(\mathbb{P}_n(\mathcal{X}_j) - P_{\hat\theta}(\mathcal{X}_j)\big)^2}{P_{\hat\theta}(\mathcal{X}_j)} + o_P(1).$$

Proof. Let $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P_{\theta_0})$ be the empirical process and define $\mathbb{H}_n = \sqrt{n}(P_{\hat\theta} - P_{\theta_0})$. Then $\sqrt{n}\big(\mathbb{P}_n(\hat{\mathcal{X}}_j) - P_{\hat\theta}(\hat{\mathcal{X}}_j)\big) = (\mathbb{G}_n - \mathbb{H}_n)(\hat{\mathcal{X}}_j)$, and similarly with $\mathcal{X}_j$ replacing $\hat{\mathcal{X}}_j$. The condition that the sets $\mathcal{X}_j$ belong to a Donsker class combined with the continuity condition $P_{\theta_0}(\hat{\mathcal{X}}_j \triangle \mathcal{X}_j) \xrightarrow{P} 0$, imply that $\mathbb{G}_n(\hat{\mathcal{X}}_j) - \mathbb{G}_n(\mathcal{X}_j) \xrightarrow{\mathrm{P}} 0$ (see Lemma 19.24). The differentiability of the map $\theta \mapsto P_\theta$ implies that

$$\sup_C \big|P_{\hat\theta}(C) - P_{\theta_0}(C) - \dot P_{\theta_0}(C)(\hat\theta - \theta_0)\big| = o_P\big(\|\hat\theta - \theta_0\|\big).$$

Together with the continuity $\dot P_{\theta_0}(\hat{\mathcal{X}}_j) - \dot P_{\theta_0}(\mathcal{X}_j) \xrightarrow{\mathrm{P}} 0$ and the $\sqrt{n}$-consistency of $\hat\theta$, this

shows that $\mathbb{H}_n(\hat{\mathcal{X}}_j) - \mathbb{H}_n(\mathcal{X}_j) \xrightarrow{P} 0$. In particular, because $P_{\theta_0}(\hat{\mathcal{X}}_j) \xrightarrow{P} P_{\theta_0}(\mathcal{X}_j)$, both $P_{\hat{\theta}}(\hat{\mathcal{X}}_j)$ and $P_{\hat{\theta}}(\mathcal{X}_j)$ converge in probability to $P_{\theta_0}(\mathcal{X}_j) > 0$. The theorem follows. ■

The conditions on the random partitions that are imposed in the preceding theorem are mild. An interesing choice is a partition in sets $\mathcal{X}_j(\hat{\theta})$ such that $P_\theta\big(\mathcal{X}_j(\theta)\big) = a_j$ is independent of θ. The corresponding modified Pearson statistic is known as the *Watson-Roy statistic* and takes the form

$$\sum_{i=1}^{k} \frac{n\Big(\mathbb{P}_n\big(\mathcal{X}_j(\hat{\theta})\big) - a_j\Big)^2}{a_j}.$$

Here the null probabilities have been reduced to fixed values again, but the cell frequencies are "doubly random." If the model is smooth and the parameter and the sets $\mathcal{X}_j(\theta)$ are not too wild, then this statistic has the same null limit distribution as the modified Pearson statistic with a fixed partition.

17.10 ***Example (Location-scale).*** Consider testing a null hypothesis that the true underlying measure of the observations belongs to a location-scale family $\big\{F_0\big((\cdot - \mu)/\sigma\big) : \mu \in \mathbb{R}, \sigma > 0\big\}$, given a fixed distribution F_0 on $\mathbb{R}$. It is reasonable to choose a partition in sets $\hat{\mathcal{X}}_j = \hat{\mu} + \hat{\sigma}(c_{j-1}, c_j]$, for a fixed partition $-\infty = c_0 < c_1 < \cdots < c_k = \infty$ and estimators $\hat{\mu}$ and $\hat{\sigma}$ of the location and scale parameter. The partition could, for instance, be chosen equal to $c_j = F_0^{-1}(j/k)$, although, in general, the partition should depend on the type of deviation from the null hypothesis that one wants to detect.

If we use the same location and scale estimators to "estimate" the null probabilities $F_0\big((\hat{\mathcal{X}}_j - \mu)/\sigma\big)$ of the random cells $\hat{\mathcal{X}}_j = \hat{\mu} + \hat{\sigma}(c_{j-1}, c_j]$, then the estimators cancel, and we find the fixed null probabilities $F_0(c_j) - F_0(c_{j-1})$. □

*17.6 Asymptotic Efficiency

The asymptotic null distributions of various versions of the Pearson statistic enable us to set critical values but by themselves do not give information on the asymptotic power of the tests. Are these tests, which appear to be mostly motivated by their asymptotic null distribution, sufficiently powerful?

The asymptotic power can be measured in various ways. Probably the most important method is to consider local limiting power functions, as in Chapter 14. For the likelihood ratio test these are obtained in Chapter 16. Because, in the local experiments, chi-square statistics are asymptotically equivalent to the likelihood ratio statistics (see Theorem 17.4), the results obtained there also apply to the present problem, and we shall not repeat the discussion.

A second method to evaluate the asymptotic power is by Bahadur efficiencies. For this nonlocal criterion, chi-square tests and likelihood ratio tests are not equivalent, the second being better and, in fact, optimal (see Theorem 16.12).

We shall compute the slopes of the Pearson and likelihood ratio tests for testing the simple hypothesis $H_0 : p = a$. A multinomial vector X_n with parameters n and $p = (p_1, \ldots, p_k)$ can be thought of as n times the empirical measure $\mathbb{P}_n$ of a random sample of size n from the distribution P on the set $\{1, \ldots, k\}$ defined by $P\{i\} = p_i$. Thus we can view both the

Pearson and the likelihood ratio statistics as functions of an empirical measure and next can apply Sanov's theorem to compute the desired limits of large deviations probabilities. Define maps C and K by

$$C(p,a) = \sum_{i=1}^{k} \frac{(p_i - a_i)^2}{a_i},$$

$$K(p,a) = -P \log \frac{a}{p} = \sum_{i=1}^{k} p_i \log \frac{p_i}{a_i}.$$

Then the Pearson and likelihood ratio statistics are equivalent to $C(\mathbb{P}_n, a)$ and $K(\mathbb{P}_n, a)$, respectively.

Under the assumption that $a > 0$, both maps are continuous in p on the k-dimensional unit simplex. Furthermore, for t in the interior of the ranges of C and K, the sets $B_t = \{p : C(p,a) \geq t\}$ and $\tilde{B}_t = \{p : K(p,a) \geq t\}$ are equal to the closures of their interiors. Two applications of Sanov's theorem yield

$$\frac{1}{n} \log \mathrm{P}_a\big(C(\mathbb{P}_n, a) \geq t\big) \to -\inf_{p \in B_t} K(p,a),$$

$$\frac{1}{n} \log \mathrm{P}_a\big(K(\mathbb{P}_n, a) \geq t\big) \to -\inf_{p \in \tilde{B}_t} K(p,a) = -t.$$

We take the function $e(t)$ of (14.20) equal to minus two times the right sides. Because $\mathbb{P}_n\{i\} \to p_i$ by the law of large numbers, whence $C(\mathbb{P}_n, a) \xrightarrow{\mathrm{P}} C(P,a)$ and $K(\mathbb{P}_n, a) \xrightarrow{\mathrm{P}} K(P,a)$, the Bahadur slopes of the Pearson and likelihood ratio tests at the alternative $H_1 : p = q$ are given by

$$2 \inf_{p : C(p,a) \geq C(q,a)} K(p,a)$$

and

$$2K(q,a).$$

It is clear from these expressions that the likelihood ratio test has a bigger slope. This is in agreement with the fact that the likelihood ratio test is asymptotically Bahadur optimal in any smooth parametric model. Figure 17.1 shows the difference of the slopes in one particular case. The difference is small in a neighborhood of the null hypothesis a, in agreement with the fact that the Pitman efficiency is equal to 1, but can be substantial for alternatives away from a.

Notes

Pearson introduced his statistic in 1900 in [112] The modification with estimated parameters, using the multinomial frequencies, was considered by Fisher [49], who corrected the mistaken belief that estimating the parameters does not change the limit distribution. Chernoff and Lehmann [22] showed that using maximum likelihood estimators based on the original data for the parameter in a goodness-of-fit statistic destroys the asymptotic chi-square distribution. They note that the errors in the level are small in the case of testing a Poisson distribution and somewhat larger when testing normality.

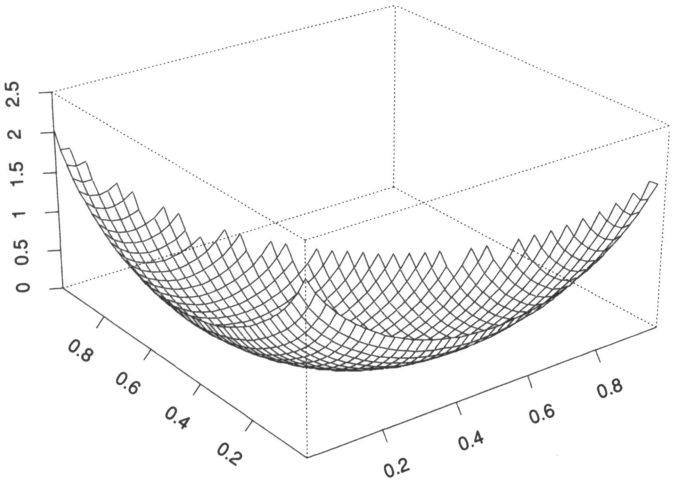

Figure 17.1. The difference of the Bahadur slopes of the likelihood ratio and Pearson tests for testing $H_0 : p = (1/3, 1/3, 1/3)$ based on a multinomial vector with parameters n and $p = (p_1, p_2, p_3)$, as a function of (p_1, p_2).

The choice of the partition in chi-square goodness-of-fit tests is an important issue that we have not discussed. Several authors have studied the optimal number of cells in the partition. This number depends, of course, on the alternative for which one desires large power. The conclusions of these studies are not easily summarized. For alternatives p such that the likelihood ratio p/p_{θ_0} with respect to the null distribution is "wild," the number of cells k should tend to infinity with n. Then the chi-square approximation of the null distribution needs to be modified. Normal approximations are used, because a chi-square distribution with a large number of degrees of freedom is approximately a normal distribution. See [40], [60], and [86] for results and further references.

PROBLEMS

1. Let $N = (N_{ij})$ be a multinomial matrix with success probabilities p_{ij}. Design a test statistic for the null hypothesis of symmetry $H_0 : p_{ij} = p_{ji}$ and derive its asymptotic null distribution.
2. Derive the limit distribution of the chi-square goodness-of-fit statistic for testing normality if using the sample mean and sample variance as estimators for the unknown mean and variance. Use two or three cells to keep the calculations simple. Show that the limit distribution is not chi-square.
3. Suppose that X_m and Y_n are independent multinomial vectors with parameters $(m, a_1, \ldots, a_k)$ and $(n, b_1, \ldots, b_k)$, respectively. Under the null hypothesis $H_0 : a = b$, a natural estimator of the unknown probability vector $a = b$ is $\hat{c} = (m+n)^{-1}(X_m + Y_n)$, and a natural test statistic is given by

$$\sum_{i=1}^{k} \frac{(X_{m,i} - m\hat{c}_i)^2}{m\hat{c}_i} + \sum_{i=1}^{k} \frac{(Y_{n,i} - n\hat{c}_i)^2}{n\hat{c}_i}.$$

Show that $\hat{c}$ is the maximum likelihood estimator and show that the sequence of test statistics is asymptotically chi square–distributed if $m, n \to \infty$.

4. A matrix Σ^- is called a *generalized inverse* of a matrix Σ if $x = \Sigma^- y$ solves the equation $\Sigma x = y$ for every y in the range of Σ. Suppose that X is $N_k(0, \Sigma)$-distributed for a matrix Σ of rank r. Show that

(i) $Y^T \Sigma^- Y$ is the same for every generalized inverse Σ^-, with probability one;

(ii) $Y^T \Sigma^- Y$ possesses a chi-square distribution with r degrees of freedom;

(iii) if $Y^T C Y$ possesses a chi-square distribution with r degrees of freedom and C is a nonnegative-definite symmetric matrix, then C is a generalized inverse of Σ.

5. Find the limit distribution of the *Dzhaparidze-Nikulin statistic*

$$n \frac{\big(\mathbb{P}_n(\mathcal{X}_j) - P_{\hat\theta}(\mathcal{X}_j)\big)}{\sqrt{P_{\hat\theta}(\mathcal{X}_j)}} \Big(I - C_{\hat\theta}^T \big(C_{\hat\theta} C_{\hat\theta}^T\big)^{-1} C_{\hat\theta}\Big) \frac{\big(\mathbb{P}_n(\mathcal{X}_j) - P_{\hat\theta}(\mathcal{X}_j)\big)}{\sqrt{P_{\hat\theta}(\mathcal{X}_j)}}.$$

6. Show that the matrix $I - C_\theta^T I_\theta^{-1} C_\theta$ in Example 17.8 is nonsingular unless the empirical estimator $\big(\mathbb{P}_n(\mathcal{X}_1), \ldots, \mathbb{P}_n(\mathcal{X}_k)\big)$ is asymptotically efficient. (The estimator $\big(P_{\hat\theta}(\mathcal{X}_1), \ldots, P_{\hat\theta}(\mathcal{X}_k)\big)$ is asymptotically efficient and has asymptotic covariance matrix diag $(\sqrt{a_\theta}) C_\theta^T I_\theta^{-1} C_\theta$ diag $(\sqrt{a_\theta})$; the empirical estimator has asymptotic covariance matrix diag $(\sqrt{a_\theta})(I - \sqrt{a_\theta}\sqrt{a_\theta}^T)$ diag $(\sqrt{a_\theta})$.)