Chi-squared tests
Math 6070, Spring 2013

Davar Khoshnevisan
University of Utah

February 14, 2013

Contents

"Chi-squared tests" refers to a family of statistical test whose large-sample asymptotics are related in one way or another to the χ^2 family of distributions. These are our first examples of "non-parametric procedures." Let me mention one particularly useful chi-squared test here before we get started properly. Consider a population whose distribution is believed to follow a known density function f—say, exponential(12). Then there is a chi-squared test that can be used to check our hypothesis. This chi-squared test is easy to implement, and works well for large-sized samples. First, we study a related parametric estimation problem. Then we find a consistent MLE. And finally, we use our estimate to devise a test for our hypothesis. This will be done in several stages. It is likely that you already know something of the resulting "χ^2-test" (page 8). But here you will find a rigorous proof, in line with the spirit of this course, of the fact that the said χ^2-test actually works. A [convincing but] non-rigorous justification of the same fact can be found in pages 314–316 of the excellent monograph of P. Bickel and K. Doksum [*Mathematical Statistics*, Holden Day, First edition, 1977]. There, a likelihood principle argument is devised, which "usually" works.

1 MLE for goodness-of fit

Suppose $X_1, \ldots, X_n$ is an independent sample from a finite population. More precisely, let $t_1, \ldots, t_m$ be m distinct numbers (or "types"), and $\boldsymbol{\theta} := (\theta_1, \ldots, \theta_m)$ an m-dimensional probability vector. Consider i.i.d. random variables $X_1, \ldots, X_n$ such that $\mathrm{P}\{X_1 = t_i\} = \theta_i$ for all $1 \le i \le m$. That is, each X is of type t_i with probability θ_i. We wish to find a (actually, "the") MLE for the mass function of the X_i's. That is, the MLE for the vector $\boldsymbol{\theta} := (\theta_1, \ldots, \theta_m)$. Note that, here, Θ is the collection of all $\boldsymbol{\theta}$ such that $0 \le \theta_1, \ldots, \theta_m \le 1$ and $\theta_1 + \cdots + \theta_m = 1$.

This is a kind of parametric problem, of course, and we proceed by computing the MLE.

Define $\boldsymbol{N} := (N_1, \ldots, N_m)$ to be the count statistics for the various types. Using column-vector notation, this means that

$$\boldsymbol{N} = \begin{pmatrix} \sum_{j=1}^n \mathbf{I}\{X_j = t_1\} \\ \vdots \\ \sum_{j=1}^n \mathbf{I}\{X_j = t_m\} \end{pmatrix}. \tag{1}$$

In words, N_k denotes the number of occurances of type k in the sample $\boldsymbol{X} := (X_1, \ldots, X_n)$. Note that $\sum_{j=1}^m N_j = n$, so the N_j's are obviously dependent rv's. Let us calculate their joint mass function.

Let $\boldsymbol{n} = (n_1, \ldots, n_m)$ be an m-vector of positive integers such that $\sum_{j=1}^m n_j = n$. Such $\boldsymbol{n}$'s are the possible values of $\boldsymbol{N}$. Moreover,

$$\mathrm{P}_{\boldsymbol{\theta}}\{\boldsymbol{N} = \boldsymbol{n}\} = \sum \mathrm{P}_{\boldsymbol{\theta}}\{X_1 = s_1, \ldots, X_n = s_n\}, \tag{2}$$

where the sum is taken over all $\boldsymbol{s} = (s_1, \ldots, s_n)$ such that n_1 of them are of type t_1, n_2 of type t_2, etc. For all such $\boldsymbol{s}$,

$$\mathrm{P}_{\boldsymbol{\theta}}\{X_1 = s_1, \ldots, X_n = s_n\} = \theta_1^{n_1} \cdots \theta_m^{n_m}. \tag{3}$$

Moreover, the number of all such $\boldsymbol{s}$'s is exactly

$$\binom{n}{n_1, \ldots, n_m} := \frac{n!}{n_1! \cdots n_m!}. \tag{4}$$

Therefore, we have derived the following.

Lemma 1. *The random n-vector $\boldsymbol{N}$ has the multinomial distribution with parameters $(n, \boldsymbol{\theta})$. That is, for all integers $n_1, \ldots, n_m \ge 1$ such that $n_1 + \cdots + n_m = n$,*

$$\mathrm{P}_{\boldsymbol{\theta}}\{\boldsymbol{N} = \boldsymbol{n}\} = \binom{n}{n_1, \ldots, n_m} \theta_1^{n_1} \cdots \theta_m^{n_m}. \tag{5}$$

Consider the special case $n = 2$. Then, $N_1 \sim \text{binomial}(n, \theta)$, and $N_2 = n - N_1$. So the multinomial distribution is an extension of the binomial in this sense. We write "$\boldsymbol{N} \sim \text{Multinomial}(n, \boldsymbol{\theta})$" in place of the statement that "$\boldsymbol{N}$ is multinomial with parameters $(n, \boldsymbol{\theta})$."

It turns out that we can compute the MLE for $\boldsymbol{\theta}$ by using $\boldsymbol{N}$ as the basic data. To do so, we use Lemma 1 and first find the log-likelihood as follows:

$$L(\boldsymbol{\theta}) = A + \sum_{i=1}^{m} N_i \ln \theta_i, \tag{6}$$

where A does not depend on $\boldsymbol{\theta}$. The case where some $\theta_i \in \{0\,, 1\}$ is trivial and can be ruled out easily. Therefore, we may assume that $0 < \theta_i < 1$ for all $i = 1, \ldots, m$. Thus, for all $l = 1, \ldots, m$,

$$\frac{\partial}{\partial \theta_l} L(\boldsymbol{\theta}) = \sum_{i=1}^{m} \left(\frac{N_i}{\theta_i} \cdot \frac{\partial \theta_i}{\partial \theta_l} \right). \tag{7}$$

As $\boldsymbol{\theta}$ varies in Θ, $\theta_1, \ldots, \theta_m$ roam freely subject to the single condition that $\theta_m = 1 - \sum_{i=1}^{m-1} \theta_i$. In other words, $(\partial \theta_i / \partial \theta_l) = 0$ for all $i = 1, \ldots, m-1$ different from l, but $(\partial \theta_m / \partial \theta_l) = -1$. This proves that $(\partial L(\boldsymbol{\theta})/\partial \theta_l) = (N_m/\theta_m) - (N_l/\theta_l)$ for all $l = 1, \ldots, m-1$. Set it equal to zero (for MLE) to see that there must exist γ such that $N_j = \gamma \theta_j$ for all $j = 1, \ldots, m$. But $\sum_{j=1}^{m} N_j = n$, whereas $\sum_{j=1}^{m} \theta_j = 1$. Therefore, $\gamma = n$. We have proved the following.

Lemma 2. *The MLE for $\boldsymbol{\theta}$ based on $N_1, \ldots, N_m$ is given by*

$$\hat{\boldsymbol{\theta}} := \begin{pmatrix} N_1/n \\ \vdots \\ N_m/n \end{pmatrix}. \tag{8}$$

[To prove that $\partial L / \partial \theta_l = 0$ indeed yields a maximum, you take second derivatives and apply similar arguments.]

2 The Multivariate normal distribution

We have seen already the multinomial distribution, which is a probability distribution for discrete m-dimensional random vectors. You should note that the multinomial distribution is a generalization of the binomial distribution. And, as it turns out, multinomial distributions have a multivariate normal approximation that is a generalization of the better-known central limit theorem for univariate random variables. In this section, I will hash out a rough outline of the theory of multivariate normal distributions [and random vectors]. The mathematical underpinnings of this theory is explained in math. 6010, and should not be missed.

Definition 3. Let $\boldsymbol{X} := (X_1\,, \ldots, X_m)'$ denote an m-dimensional random vector. It *mean vector* is defined as

$$\mathrm{E}(\boldsymbol{X}) := \begin{pmatrix} \mathrm{E}(X_1) \\ \vdots \\ \mathrm{E}(X_m) \end{pmatrix},$$

and its covariance matrix is defined as the matrix whose $(i\,,j)$th entry is $\text{Cov}(X_i\,,X_j)$; since $\text{Cov}(X_i\,,X_i) = \text{Var}(X_i)$,

$$\text{Cov}(\boldsymbol{X}) := \begin{pmatrix} \text{Var}(X_1) & \text{Cov}(X_1\,,X_2) & \cdots & \text{Cov}(X_1\,,X_m) \\ \text{Cov}(X_2\,,X_1) & \text{Var}(X_2) & \cdots & \text{Cov}(X_2\,,X_m) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_m\,,X_1) & \text{Cov}(X_m\,,X_2) & \cdots & \text{Var}(X_m) \end{pmatrix}.$$

Of course, we are tacitly assuming that the means, variances, and covariances are finite.

If $\boldsymbol{x}$ is an m-dimensional nonrandom vector, then $\boldsymbol{x}'\,\text{Cov}(\boldsymbol{X})\boldsymbol{x}$ is always a scalar. One can show that, in fact, that scalar is equal to the variance of the random scalar quantity $\boldsymbol{x}'\boldsymbol{X}$. That is,

$$\text{Var}\,(\boldsymbol{x}'\boldsymbol{X}) = \boldsymbol{x}'\,\text{Cov}(\boldsymbol{X})\boldsymbol{x}.$$

Since the variance of a random variable is always ≥ 0, it follows that if $\boldsymbol{\Sigma} := \text{Cov}(\boldsymbol{X})$, then

$$\boldsymbol{x}'\boldsymbol{Q}\boldsymbol{x} \geq 0 \qquad \text{for all } \boldsymbol{x}.$$

Any matrix that has the preceding property is called *positive semidefinite*, and we just saw that all covariance matrices are positive semidefinite. The converse is also true, but harder to prove.

Proposition 4. *Let $\boldsymbol{Q}$ be a symmetric $m\times m$ matrix. Then $\boldsymbol{Q}$ is the covariance of a random vector $\boldsymbol{X}$ if and only if $\boldsymbol{Q}$ is positive semidefinite. Also, a symmetric $m\times m$ matrix $\boldsymbol{Q}$ is positive semidefinite if and only if all of eigenvalues are ≥ 0.*

We say that $\boldsymbol{Q}$ is *positive definite* if

$$\boldsymbol{x}'\boldsymbol{Q}\boldsymbol{x} > 0 \quad \text{for all } \boldsymbol{x}.$$

Proposition 5. *A symmetric $m \times m$ matrix $\boldsymbol{Q}$ is positive definite if and only if all of its eigenvalues are > 0. In particular, such a matrix is invertible.*

Definition 6. Given an m-dimensional vector $\boldsymbol{\mu}$ and an $m \times m$-dimensional covariance matrix $\boldsymbol{Q}$, the *multivariate normal pdf* with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{Q}$ is

$$f(\boldsymbol{x}) := \frac{1}{(2\pi)^{m/2}|\det \boldsymbol{Q}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{Q}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right).$$

This is a proper pdf on $\mathbf{R}^m$. We say that $\boldsymbol{X} \sim \text{N}_m(\boldsymbol{\mu}\,,\boldsymbol{Q})$, when $\boldsymbol{X} = (X_1\,,\ldots,X_m)'$, and the joint pdf of $(X_1\,,\cdots,X_m)$ is f.

The preceding function f is indeed a proper joint density in the sense that $f(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x} := (x_1\,,\ldots,x_m)'$ and

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(x_1\,,\ldots,x_m)\,\text{d}x_1\cdots\text{d}x_m = 1.$$

So it makes sense to say "$\boldsymbol{X} \sim \text{N}_m(\boldsymbol{\mu}\,,\boldsymbol{Q})$."

Proposition 7. *If* $\boldsymbol{X} \sim \mathrm{N}_m(\boldsymbol{\mu}\,, \boldsymbol{Q})$ *for some vector* $\boldsymbol{\mu} := (\mu_1\,, \ldots, \mu_m)'$ *and matrix* $\boldsymbol{Q}$*, then*

$$\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{X}), \quad \textit{and} \quad \mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{Q}.$$

Thus, we refer to $\boldsymbol{\mu}$ *and* $\boldsymbol{Q}$ *respectively as the* mean vector *and* covariance matrix *of the* $\mathrm{N}_m(\boldsymbol{\mu}\,, \boldsymbol{Q})$ *distribution.*

The following states that a random vector is distributed as a multivariate normal if and only if all of possible [nonrandom] linear combinations of its components are distributed as [univariate] normals.

Proposition 8. $\boldsymbol{X} \sim \mathrm{N}_m(\boldsymbol{\mu}\,, \boldsymbol{Q})$ *if and only if* $\boldsymbol{t}'\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{t}'\boldsymbol{\mu}\,, \boldsymbol{t}'\boldsymbol{Q}\boldsymbol{t})$ *for all nonrandom* m*-vectors* $\boldsymbol{t} := (t_1\,, \ldots, t_m)'$*.*

In other words, the preceding states that the property about the linear combinations of the components characterizes completely the multivariate normal distribution. An advantage of the preceding characterization is that the nonsingularity of $\boldsymbol{Q}$ is not mentioned anywhere explicitly. Therefore, we can define the following slight generalization of multivariate normals to the case that $\boldsymbol{Q}$ is positive semidefinite [as opposed to being positive definite]:

Definition 9. Let $\boldsymbol{X}$ be a random m-vector. We say that $\boldsymbol{X} \sim \mathrm{N}_m(\boldsymbol{\mu}\,, \boldsymbol{Q})$, for some $\boldsymbol{\mu} = (\mu_1\,, \ldots, \mu_m)'$ and some $m \times m$ symmetric positive semidefinite matrix $\boldsymbol{Q}$, when

$$\boldsymbol{t}'\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{t}'\boldsymbol{\mu}\,, \boldsymbol{t}'\boldsymbol{Q}\boldsymbol{t}) \qquad \text{for all } \boldsymbol{t} := (t_1\,, \ldots, t_m)'.$$

3 Central limit theorems

Choose and fix a probability vector $\boldsymbol{\theta}$, and assume it is the true one so that we write P, E, etc. in place of the more cumbersome $\mathrm{P}_{\boldsymbol{\theta}}$, $\mathrm{E}_{\boldsymbol{\theta}}$, etc.

We can use (1) to say a few more things about the distribution of $\boldsymbol{N}$. Note first that (1) is equivalent to the following:

$$\boldsymbol{N} = \sum_{j=1}^{n} \begin{pmatrix} \mathbf{I}\{X_j = t_1\} \\ \vdots \\ \mathbf{I}\{X_j = t_m\} \end{pmatrix} := \sum_{j=1}^{n} \mathbf{Y}_j. \tag{9}$$

We can compute directly to find that

$$\mathrm{E}\boldsymbol{Y_1} = \boldsymbol{\theta}, \tag{10}$$

and

$$\begin{aligned} \mathrm{Cov}\boldsymbol{Y}_1 &= \begin{pmatrix} \theta_1(1-\theta_1) & -\theta_1\theta_2 & \cdots & -\theta_1\theta_m \\ -\theta_1\theta_2 & \theta_2(1-\theta_2) & \cdots & -\theta_2\theta_m \\ \vdots & \vdots & \ddots & \vdots \\ -\theta_m\theta_1 & -\theta_m\theta_2 & \cdots & \theta_m(1-\theta_m) \end{pmatrix} \\ &:= \boldsymbol{Q}_0. \end{aligned} \tag{11}$$

Because $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ are i.i.d. random vectors, the following is a consequence of the preceding, considered in conjunction with the multidimensional CLT.

Theorem 10. *As $n \to \infty$,*

$$\frac{\boldsymbol{N} - n\boldsymbol{\theta}}{\sqrt{n}} \xrightarrow{d} \mathrm{N}_m\left(\mathbf{0}\,, \boldsymbol{Q}_0\right). \tag{12}$$

There is a variant of Theorem 10 which is more useful to us. Define

$$\widetilde{\boldsymbol{N}} := \begin{pmatrix} (N_1 - n\theta_1)/\sqrt{\theta_1} \\ \vdots \\ (N_m - n\theta_m)/\sqrt{\theta_m} \end{pmatrix}. \tag{13}$$

This too is a sum of n i.i.d. random vectors $\widetilde{\boldsymbol{Y}}_1, \ldots, \widetilde{\boldsymbol{Y}}_n$, and has

$$\mathrm{E}\widetilde{\boldsymbol{Y}}_1 := \mathbf{0} \quad \text{and} \quad \mathrm{Cov}\widetilde{\boldsymbol{Y}}_1 = \boldsymbol{Q}, \tag{14}$$

where $\boldsymbol{Q}$ is the $(m \times m)$-dimensional covariance matrix given below:

$$\boldsymbol{Q} = \begin{pmatrix} 1-\theta_1 & -\sqrt{\theta_1\theta_2} & \cdots & -\sqrt{\theta_1\theta_m} \\ -\sqrt{\theta_1\theta_2} & 1-\theta_2 & \cdots & -\sqrt{\theta_2\theta_m} \\ \vdots & \vdots & \ddots & \vdots \\ -\sqrt{\theta_m\theta_1} & -\sqrt{\theta_m\theta_2} & \cdots & 1-\theta_m \end{pmatrix}. \tag{15}$$

Then we obtain,

Theorem 11. *As $n \to \infty$,*

$$\frac{1}{\sqrt{n}}\widetilde{\boldsymbol{N}} \xrightarrow{d} \mathrm{N}_m\left(\mathbf{0}\,, \boldsymbol{Q}\right). \tag{16}$$

Remark 12. One should be a little careful at this stage because $\boldsymbol{Q}$ is singular. [For example, consider the case $m = 2$.] However, $\mathrm{N}_m(\mathbf{0}\,, \boldsymbol{Q})$ makes sense even in this case. The reason is this: If $\boldsymbol{C}$ is a non–singular covariance matrix, then we know what it means for $\boldsymbol{X} \sim \mathrm{N}_m(\mathbf{0}\,, \boldsymbol{C})$, and a direct computation reveals that for all real numbers t and all m-vectors $\boldsymbol{\xi}$,

$$\mathrm{E}\left[\exp\left(it\boldsymbol{\xi}'\boldsymbol{X}\right)\right] = \mathrm{E}\left[\exp\left(-\frac{t^2}{2}\boldsymbol{\xi}'\boldsymbol{C}\boldsymbol{\xi}\right)\right]. \tag{17}$$

This defines the characteristic function of the random variable $\boldsymbol{\xi}'\boldsymbol{X}$ uniquely. There is a theorem (Mann–Wold device) that says that a unique description of the characteristic function of $\boldsymbol{\xi}'\boldsymbol{X}$ for all $\boldsymbol{\xi}$ also describes the distribution of $\boldsymbol{X}$ uniquely. This means that any random vector $\boldsymbol{X}$ that satisfies (17) for all $t \in \mathbf{R}$ and $\boldsymbol{\xi} \in \mathbf{R}^m$ is $\mathrm{N}_m(\mathbf{0}\,, \boldsymbol{C})$. Moreover, this approach does not require $\boldsymbol{C}$ to be invertible. To learn more about multivariate normals, you should attend Math. 6010.

As a consequence of Theorem 11 we have the following: As $n \to \infty$,

$$\frac{1}{\sqrt{n}} \left\| \widetilde{\boldsymbol{N}} \right\|^2 \xrightarrow{d} \left\| \mathrm{N}_m(\boldsymbol{0}\,, \boldsymbol{Q}) \right\|^2. \tag{18}$$

[Recall that $\|\boldsymbol{x}\|^2 = x_1^2 + \cdots + x_m^2$.] Next, we identify the distribution of $\|\mathrm{N}_m(\boldsymbol{0}\,, \boldsymbol{Q})\|^2$. A direct (and somewhat messy) calculation shows that $\boldsymbol{Q}$ is a *projection matrix*. That is, $\boldsymbol{Q}^2 := \boldsymbol{Q}\boldsymbol{Q} = \boldsymbol{Q}$. It follows fairly readily from this that its eigenvalues are all zeros and ones. Here is the proof: Let λ be an eigenvalue and $\boldsymbol{x}$ an eigenvector. Without loss of generality, we may assume that $\|\boldsymbol{x}\| = 1$; else, consider the eigenvector $\boldsymbol{y} := \boldsymbol{x}/\|\boldsymbol{x}\|$ instead. This means that $\boldsymbol{Q}\boldsymbol{x} = \lambda \boldsymbol{x}$. Therefore, $\boldsymbol{Q}^2\boldsymbol{x} = \lambda^2 \boldsymbol{x}$. Because $\boldsymbol{Q}^2 = \boldsymbol{Q}$, we have proven that $\lambda = \lambda^2$, whence it follows that $\lambda \in \{0\,, 1\}$, as promised. Write $\boldsymbol{Q}$ in its spectral form. That is,

$$\boldsymbol{Q} = \boldsymbol{P}' \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_m \end{pmatrix} \boldsymbol{P}, \tag{19}$$

where $\boldsymbol{P}$ is an orthogonal $(m \times m)$ matrix. [This is because $\boldsymbol{Q}$ is positive semidefinite. You can learn more on this in Math. 6010.] Define

$$\boldsymbol{\Lambda} := \begin{pmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\lambda_m} \end{pmatrix} \boldsymbol{P}. \tag{20}$$

Then, $\boldsymbol{Q} = \boldsymbol{\Lambda}'\boldsymbol{\Lambda}$, which is why $\boldsymbol{\Lambda}$ is called the *square root* of $\boldsymbol{Q}$. Let $\boldsymbol{Z}$ be an m-vector of i.i.d. standard normals and consider $\boldsymbol{\Lambda}\boldsymbol{Z}$. This is multivariate normal (check characteristic functions), $\mathrm{E}[\boldsymbol{\Lambda}\boldsymbol{Z}] = \boldsymbol{0}$, and $\mathrm{Cov}(\boldsymbol{\Lambda}\boldsymbol{Z}) = \boldsymbol{\Lambda}'\mathrm{Cov}(\boldsymbol{Z})\boldsymbol{\Lambda} = \boldsymbol{\Lambda}'\boldsymbol{\Lambda} = \boldsymbol{Q}$. In other words, we can view $\mathrm{N}_m(\boldsymbol{0}\,, \boldsymbol{Q})$ as $\boldsymbol{\Lambda}\boldsymbol{Z}$. Therefore,

$$\left\| \mathrm{N}_m(\boldsymbol{0}\,, \boldsymbol{Q}) \right\|^2 \stackrel{d}{=} \left\| \boldsymbol{\Lambda}\boldsymbol{Z} \right\|^2 = \sum_{i=1}^{m} \lambda_j Z_i^2. \tag{21}$$

Because $\lambda_i \in \{0\,, 1\}$, we have proved that $\|\mathrm{N}_m(\boldsymbol{0}\,, \boldsymbol{Q})\| \sim \chi_r^2$, where $r = \mathrm{rank}(\boldsymbol{Q})$. Another messy but elementary computation shows that the rank of $\boldsymbol{Q}$ is $m - 1$. [Roughly speaking, this is because the θ_i's have only one source of linear dependence. Namely, that $\theta_1 + \cdots + \theta_m = 1$.] This proves that

$$\frac{1}{\sqrt{n}} \left\| \widetilde{\boldsymbol{N}} \right\|^2 \xrightarrow{d} \chi_{m-1}^2. \tag{22}$$

Write out the left-hand side explicitly to find the following:

Theorem 13. *As $n \to \infty$,*

$$\sum_{i=1}^{m} \frac{(N_i - n\theta_i)^2}{n\theta_i} \xrightarrow{d} \chi_{m-1}^2. \tag{23}$$

It took a while to get here, but we will soon be rewarded for our effort.

4 Application to goodness-of-fit

4.1 Testing for a finite distribution

In light of the development of Section 1, can paraphrase Theorem 13 as saying that under $\mathrm{P}_{\boldsymbol{\theta}}$, as $n \to \infty$,[1]

$$n\mathbb{X}_n(\boldsymbol{\theta}) \stackrel{d}{\to} \chi^2_{m-1}, \quad \text{where} \quad \mathbb{X}_n(\boldsymbol{\theta}) := \sum_{i=1}^{m} \frac{\left(\hat{\theta}_i - \theta_i\right)^2}{\theta_i}. \tag{24}$$

Therefore, we are in a position to derive an asymptotic $(1-\alpha)$-level confidence set for $\boldsymbol{\theta}$. Define for all $a > 0$,

$$C_n(a) := \left\{\boldsymbol{\theta} \in \Theta : \ \mathbb{X}_n(\boldsymbol{\theta}) \le \frac{a}{n}\right\}. \tag{25}$$

Equation (24) shows that for all $\boldsymbol{\theta} \in \Theta$,

$$\lim_{n\to\infty} \mathrm{P}_{\boldsymbol{\theta}} \left\{\boldsymbol{\theta} \in C_n(a)\right\} = \mathrm{P}\left\{\chi^2_{m-1} \le a\right\}. \tag{26}$$

We can find (e.g., from χ^2-tables) a number $\chi^2(\alpha)$ such that

$$\mathrm{P}\left\{\chi^2_{m-1} > \chi^2_{m-1}(\alpha)\right\} = \alpha. \tag{27}$$

It follows that

$$C_n\left(\chi^2_{m-1}(\alpha)\right) \text{ is an asymptotic level } (1-\alpha) \text{ confidence interval for } \boldsymbol{\theta}. \tag{28}$$

The statistics $\mathbb{X}_n(\boldsymbol{\theta})$ is a little awkward to work with, and this makes the computation of $C_n(\chi^2_{m-1}(\alpha))$ difficult. Nevertheless, it is easy enough to use $C_n(\chi^2_{m-1}(\alpha))$ to test simple hypotheses for $\boldsymbol{\theta}$.

Suppose we are interested in the test

$$H_0 : \ \boldsymbol{\theta} = \boldsymbol{\theta}_0 \quad \text{versus} \quad H_1 : \ \boldsymbol{\theta} \ne \boldsymbol{\theta}_0, \tag{29}$$

where $\boldsymbol{\theta}_0 = (\theta_{01}, \ldots, \theta_{0m})$ is a (known) probability vector in Θ. Then we can reject H_0 (at asymptotic level $(1-\alpha)$) if $\boldsymbol{\theta}_0 \notin C_n(\chi^2_{m-1}(\alpha))$. In other words,

$$\text{we reject } H_0 \text{ if } \mathbb{X}_n(\boldsymbol{\theta}_0) > \frac{\chi^2_{m-1}(\alpha)}{n}. \tag{30}$$

Because $\mathbb{X}_n(\boldsymbol{\theta}_0)$ is easy to compute, (29) can be carried out with relative ease. Moreover, $\mathbb{X}_n(\boldsymbol{\theta}_0)$ has the following interpretation: Under H_0,

$$\mathbb{X}_n(\boldsymbol{\theta}_0) = \sum_{i=1}^{m} \frac{(\mathrm{Obs}_i - \mathrm{Exp}_i)^2}{\mathrm{Exp}_i}, \tag{31}$$

[1] The statistic $\mathbb{X}_n(\boldsymbol{\theta})$ was discovered by K. Pearson in 1900, and is therefore called the "Pearson χ^2 statistic." Incidentally, the term "standard deviation" is also due to Karl Pearson (1893).

where Obs_i denotes the number of observations in the sample that are of type i, and $\text{Exp}_i := \theta_{0i}$ is the expected number—under H_0—of sample points that are of type i.

Next we study two examples.

Example 14. A certain population contains only the digits 0, 1, and 2. We wish to know if every digit is equally likely to occur in this population. That is, we wish to test

$$H_0 : \boldsymbol{\theta} = (\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{3}) \quad \text{versus} \quad H_1 : \boldsymbol{\theta} \neq (\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{3}). \tag{32}$$

To this end, suppose we make $1,000$ independent draws from this population, and find that we have drawn 300 zeros ($\hat{\theta}_1 = 0.3$), 340 ones ($\hat{\theta}_2 = 0.34$), and 360 twos ($\hat{\theta}_3 = 0.36$). That is,

Types	**Obs**	**Exp**	$(\textbf{Obs} - \textbf{Exp})^2/\textbf{Exp}$
0	0.30	1/3	3.33×10^{-3}
1	0.34	1/3	1.33×10^{-4}
2	0.36	1/3	2.13×10^{-3}
Total	1.00	1	0.00551 (approx.)

Suppose we wish to test (32) at the traditional (asymptotic) level 95%. Here, $m = 3$ and $\alpha = 0.05$, and we find from chi-square tables that $\chi^2_2(0.05) \approx 5.9991$. Therefore, according to (30), we reject if and only if $\mathbb{X}_n(\boldsymbol{\theta}_0) \approx 0.00551$ is greater than $\chi^2_2(0.05)/1000 \approx 0.005991$. Thus, **we do not reject the null hypothesis** at asymptotic level 95%. On the other hand, $\chi^2_2(0.1) \approx 4.605$, and $\mathbb{X}_n(\boldsymbol{\theta}_0)$ *is* greater than $\chi^2_2(0.1)/1000 \approx 0.004605$. Thus, had we performed the test at asymptotic level 90% we would indeed reject H_0. In fact, this discussion has shown that **the "asymptotic *P*-value" of our test is somewhere between 0.05 and 0.1**. It is more important to report/estimate P-values than to make blanket declarations such as "accept H_0" or "reject H_0." Consult your Math. 5080–5090 text(s).

To illustrate this example further, suppose the table of observed vs. expected remains the same but $n = 2000$. Then, $\mathbb{X}_n(\boldsymbol{\theta}_0) \approx 0.00551$ is a good deal greater than $\chi^2_2(0.05)/n \approx 0.00299955$ and we reject H_0 soundly at the asymptotic level 95%. In fact, in this case the P-value is somewhere between 0.005 and 0.0025 (check this!).

4.2 Testing for a Density

Suppose we wish to know if a certain data-set comes from the exponential(12) distribution (say!). So let $X_1, \ldots, X_n$ be an independent sample. We have

$$H_0 : X_1 \sim \text{exponential}(12) \quad \text{versus} \quad H_1 : X_1 \not\sim \text{exponential}(12). \tag{33}$$

Now take a fine mesh $0 < z_1 < z_2 < \ldots < z_m < \infty$ of real numbers. Then,

$$\mathrm{P}_{H_0}\{z_j \le X_1 \le z_{j+1}\} = 12\int_{z_j}^{z_{j+1}} \mathrm{e}^{-12r}\,\mathrm{d}r = \mathrm{e}^{-12z_j} - \mathrm{e}^{-12z_{j+1}}. \tag{34}$$

These are θ_{0j}'s, and we can test to see whether or not $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, where θ_j is the true probability $\mathrm{P}\{z_j \le X_1 \le z_{j+1}\}$ for all $j = 1, \ldots, m$. The χ^2-test of the previous section handles just such a situation. That is, let $\hat{\theta}_j$ denote the proportion of the data $X_1, \ldots, X_n$ that lies in $[z_j\,, z_{j+1}]$. Then (30) provides us with an asymptotic $(1-\alpha)$-level test for H_0 that the data is exponential(12). Other distributions are handled similarly, but θ_{0j} is in general $\int_{z_j}^{z_{j+1}} f_0(x)\,dx$, where f_0 is the density function under H_0.

Example 15. Suppose we wish to test (33). Let $m = 5$, and define $z_j = (j-1)/10$ for $j = 1, \ldots, 5$. Then,

$$\theta_{0j} = \mathrm{P}_{H_0}\left\{\frac{j-1}{10} \le X_1 \le \frac{j}{10}\right\} = \exp\left(-\frac{6(j-1)}{5}\right) - \exp\left(-\frac{6j}{5}\right). \tag{35}$$

That is,

$\boldsymbol{j}$	$\boldsymbol{\theta_{0j}}$
1	0.6988
2	0.2105
3	0.0634
4	0.0191
5	0.0058

Consider the following (hypothetical) data: $n = 200$, and

j	θ_{0j}	$\hat{\theta}_j$
1	0.6988	0.7
2	0.2105	0.2
3	0.0634	0.1
4	0.0191	0
5	0.0058	0

Then, we form the χ^2-table, as before, and obtain

Types	**Exp**	**Obs**	$\textbf{(Obs} - \textbf{Exp)}^2\textbf{/Exp}$
1	0.6988	0.7	2.06×10^{-6}
2	0.2105	0.2	0.00052
3	0.0634	0.1	0.0211
4	0.0191	0	0.0191
5	0.0058	0	0.0058
Total	0.9976	1	0.04655 (approx.)

Thus, $\mathbb{X}_n(\boldsymbol{\theta}_0) \approx 0.04655$. You should check that $0.05 \leq P\text{-value} \leq 0.1$. In fact, the P-value is much closer to 0.1. So, at the asymptotic level 95%, we do not have enough evidence against H_0.

Some things to think about: In the totals for "Exp," you will see that the total is 0.9976 and not 1. Can you think of why this has happened? Does this phenomenon render our conclusions false (or weaker than they may appear)? Do we need to fix this by including more types? How can we address it in a sensible way? Would addressing this change the final conclusion?