THE χ^2 TEST OF SIMPLE AND COMPOSITE HYPOTHESES

1. Multinomial distributions

Suppose we have a multinomial $(n, \pi_1, ..., \pi_k)$ distribution, where π_j is the probability of the jth of k possible outcomes on each of n independent trials. Thus $\pi_j \geq 0$ and $\sum_{j=1}^k \pi_j = 1$. Let X_j be the number of times that the jth outcome occurs in n independent trials. Then for any integers $n_j \geq 0$ such that $n_1 + \cdots + n_k = n$, we have

$$P(X_j = n_j,\ j = 1, ..., k) = \binom{n}{n_1, ..., n_k} \pi_1^{n_1} \pi_2^{n_2} \cdots \pi_k^{n_k}.$$

Recall that multinomial coefficients are defined by $\binom{n}{n_1,...,n_k} = \frac{n!}{n_1! n_2! \cdots n_k!}$ if $n_j \geq 0$ are integers with $\sum_{j=1}^k n_j = n$, or 0 if $\sum_{j=1}^k n_j \neq n$. In statistics, the values $\pi_1, ..., \pi_k$ are unknown. If we make no further hypothesis about them, we have the "full multinomial model" which has dimension $d = k - 1$ due to the one constraint $\sum_{j=1}^k \pi_j = 1$.

A random variable X is binomial(n, p) if and only if $(X, n - X)$ is multinomial $(n, p, 1 - p)$. On the other hand if $(X_1, ..., X_k)$ is multinomial $(n, p_1, ..., p_k)$ then for each j, X_j is binomial (n, p_j).

2. Simple multinomial hypotheses

Suppose we have a simple hypothesis H_0 specifying the π_j, namely $\pi_j = p_j$ for $j = 1, ..., k$. For example, in rolling a die, there are $k = 6$ possible outcomes (faces of a cube) numbered from 1 to 6, and a simple hypothesis would be that the dice are "fair," namely that $\pi_j = 1/6$ for $j = 1, ..., 6$. In Weldon's dice data, in 315672 individual dice throws, the outcome "5 or 6" occurred 106602 times. For a fair die the probability of "5 or 6" is 1/3, but from Weldon's data the point estimate of $\pi_5 + \pi_6$ is about 0.3377 and the 99% confidence interval (which can be found in this case by the plug-in method) excludes 1/3. In fact for fair dice, the probability of "5 or 6" occurring 106602 or more times is $E(106602, 315672, 1/3) \doteq 1.02 \cdot 10^{-7}$. (On real dice, the faces are marked by hollowed-out pips, so the higher-numbered 5 and

Date: 18.650, Oct. 14, 2015.

6 faces are a little lighter than the others, and the opposite 1 and 2 faces a little heavier, unless some compensation is made.)

Or, for a human birth, consider the two possible outcomes female or male. A simple hypothesis was that each had probability 1/2, but for a large enough n, it has been estimated that the natural probability of a female birth is about 0.488. (The fraction may vary with time or between populations, according to Web sources.) Both these examples reduced to binomial probabilities.

2.1. **The χ^2 test of a simple multinomial hypothesis.** How can one test a simple hypothesis about multinomial probabilities for general k? The chi-squared test is as follows.

If values $X_1, X_2, ..., X_k$ are observed, and a simple hypothesis H_0 specifies values $\pi_j = p_j$ with $p_j > 0$ for all $j = 1, ..., k$, then the X^2 statistic for testing H_0 is

$$X^2 = \sum_{j=1}^{k} \frac{(X_j - np_j)^2}{np_j}.$$

A shorthand notation for X^2 is $\sum \frac{(\mathcal{O}-\mathcal{E})^2}{\mathcal{E}}$ where $\mathcal{O}$ = "observed" and $\mathcal{E}$ = "expected."

Theorem. If the hypothesis H_0 is true, then as $n \to \infty$, the distribution of X^2 converges to that of $\chi^2(k-1)$, i.e. χ^2 with $k-1$ degrees of freedom.

Rule for application: A widely accepted rule is that the approximation of X^2 by a $\chi^2(k-1)$ distribution is good enough if all the expected numbers np_j are at least 5.

Remarks. For each j, the (marginal) distribution of X_j is binomial (n, π_j), where $\pi_j = p_j$ under H_0. Thus $EX_j = np_j$ and $E((X_j - np_j)^2) = np_j(1-p_j)$. In order for X_j to be approximately normal, we need $np_j(1-p_j)$ to be large enough and so np_j to be large enough. Another way to see that np_j should not be small is that if it is, since X_j has integer values, there will be relatively wide gaps between adjacent possible values of $(X_j - np_j)^2/(np_j)$, making the distribution of X^2 too discrete, and so not close to the continuous distribution of χ^2.

The quantities $X_j - np_j$ are not linearly independent, since $\sum_{j=1}^k X_j - np_j = n - n = 0$. We have $E_0(X^2) = \sum_{j=1}^k 1 - p_j = k - 1$, which equals the expectation of a $\chi^2(k-1)$ random variable.

Proof. Under H_0, the random vector $(X_1, ..., X_k)$ has a multinomial $(n, p_1, ..., p_k)$ distribution. Let's find the covariance of X_i and X_j for

$i \neq j$. If we can do that for $i = 1$ and $j = 2$ we can extend the result to any i and j.

Let $q_1 := 1 - p_1$. Given X_1, the conditional distribution of X_2 is binomial $(n - X_1, p_2/q_1)$. Thus $E(X_2|X_1) = (n - X_1)p_2/q_1$ and

$$E(X_1X_2) = E(X_1E(X_2|X_1)) = n^2p_1p_2/q_1 - p_2q_1^{-1}EX_1^2.$$

Since $EX_1^2 = \mathrm{Var}(X_1) + (EX_1)^2 = np_1q_1 + n^2p_1^2$ we get

$$E(X_1X_2) = \frac{n^2p_1p_2 - n^2p_1^2p_2}{q_1} - np_1p_2 \ = \ (n^2 - n)p_1p_2,$$

which is symmetric in p_1 and p_2 as it should be. It follows that $\mathrm{Cov}(X_1, X_2) = -np_1p_2$. It's natural that this covariance should be negative since for larger X_1, X_2 will tend to be smaller. For $1 \leq i < j \leq n$ we have likewise $\mathrm{Cov}(X_i, X_j) = -np_ip_j$.

Let $Y_j := (X_j - np_j)/\sqrt{np_j}$ for $j = 1, ..., k$. Then $X^2 = Y_1^2 + \cdots + Y_k^2$. For each j we have $EY_j = 0$ and $EY_j^2 = q_j := 1 - p_j$. We also have for $i \neq j$

$$EY_iY_j = \mathrm{Cov}(Y_i, Y_j) = \mathrm{Cov}(X_i, X_j)/(n\sqrt{p_ip_j}) = -\sqrt{p_ip_j}.$$

Recall that $\delta_{ij} = 1$ for $i = j$ and 0 for $i \neq j$. As a matrix, $I_{ij} = \delta_{ij}$ gives the $k \times k$ identity matrix. We have

$$C_{ij} := EY_iY_j = \mathrm{Cov}(Y_i, Y_j) = \delta_{ij} - \sqrt{p_ip_j}$$

for all $i, j = 1, ..., k$. Let u_p be the column vector $(\sqrt{p_1}, ..., \sqrt{p_k})'$. This vector has length 1. We can then write $C = I - u_pu_p'$ as a matrix. Let's make a change of basis in which u_p becomes one of the basis vectors, say the first, e_1, and $e_2, ..., e_k$ are any other vectors of unit length perpendicular to each other and to e_1. In this basis C becomes $D = I - e_1e_1'$ which is a diagonal matrix, in other words $D_{ij} = 0$ for $i \neq j$. Also $D_{11} = 0$, and $D_{jj} = 1$ for $j = 2, ..., k$.

Let the vector $Y = (Y_1, ..., Y_k)$ in the new coordinates become $Z = (Z_1, ..., Z_k)$, where $EZ_j = 0$ for all j and the Z_j have covariance matrix D.

As $n \to \infty$, by the multidimensional central limit theorem (proved in 18.175, e.g. in Professor Panchenko's OCW version of the course, Spring 2007), $(Z_1, Z_2, ..., Z_k)$ asymptotically have a multivariate normal distribution with mean 0 and covariance matrix D, in other words $Z_1 \equiv 0$ and $Z_2, ..., Z_k$ are asymptotically i.i.d. $N(0, 1)$. Thus $X^2 = |Y|^2 = |Z|^2 = Z_2^2 + \cdots + Z_k^2$ has asymptotically a $\chi^2(k-1)$ distribution as $n \to \infty$, Q.E.D.

3. Chi-squared tests of composite hypotheses

In doing a chi-squared test of a composite hypothesis H_0: $\pi_j = p_j(\theta)$ indexed by an m-dimensional parameter θ, two kinds of adjustment may be made. If we estimate θ by some $\widehat{\theta}$ and find the chi-squared statistic

$$\hat{X}^2 = \sum_{j=1}^{k} \frac{(X_j - np_j(\widehat{\theta}))^2}{np_j(\widehat{\theta})},$$

the usual rule is that if H_0 holds, for n large enough, this should have approximately a χ^2 distribution with $k-1-m$ degrees of freedom. For that to be valid, we need that all expected numbers $np_j(\widehat{\theta}) \geq 5$, and that $\widehat{\theta}$ is a suitable function of the statistics $X_1, ..., X_k$. Two suitable estimators for this are the minimum chi-squared estimate, where $\widehat{\theta}$ is chosen to minimize $\hat{X}^2$, or the maximum likelihood estimate $\widehat{\theta}_{MLE}$ based on the given data $X_1, ..., X_k$.

As the dimension d of the full multinomial model is $k-1$, the $\chi^2(d-m)$ distribution is the same as the asymptotic distribution for large n of the Wilks statistic for testing an m-dimensional hypothesis included in an assumed d-dimensional model. We will see in another handout that this is not just a coincidence.

In a file called "χ^2 tests for composite hypotheses – asymptotic distributions," posted on the course website as compos-chisqpfs.pdf, Theorem 1 proves under some assumptions, so that $p_j(\theta)$ depend in a suitably smooth way on θ, that the distribution of $\widehat{X}^2 = \widehat{X}^2_{MLE}$ using $\widehat{\theta}_{MLE}$ does converge to that of $\chi^2(k-1-m)$ as $n \to \infty$. Theorem 11 of that file proves that moreover, for any $\widehat{\theta}_{\min}$ (depending on n) that minimize(s) $\widehat{X}^2$ for the given $(X_1, ..., X_n)$, giving $\widehat{X}^2_{\min}$, the difference between the two statistics $\widehat{X}^2_{MLE} - \widehat{X}^2_{\min}$ converges to 0 in probability as $n \to \infty$, meaning that for any $\varepsilon > 0$, the probability that $|\widehat{X}^2_{MLE} - \widehat{X}^2_{\min}| > \varepsilon$ converges to 0 as $n \to \infty$. It follows that the distribution of $\widehat{X}^2_{\min}$ also converges to that of $\chi^2(k-1-m)$. Although $\widehat{\theta}_{\min}$ is usually not easy to compute, we know that for an arbitrary estimate $\widehat{\theta}$ of θ, the $\widehat{X}^2$ based on $\widehat{\theta}$ is at least as large as $\widehat{X}^2_{\min}$, and we will use that.

Another adjustment that's made is that if the expected numbers $np_j(\widehat{\theta})$ in some categories are less than 5, we can combine categories until all the expectations are larger than 5. When the categories are subintervals (or half-lines) of the line or of the nonnegative integers, only adjacent intervals should be combined, so that the categories remain intervals.

4. Grouped vs. ungrouped data

Suppose we combine categories, which certainly will happen if we start with infinitely many possible outcomes, as in a Poisson or geometric distribution where the outcome can be any nonnegative integer. Then when we come to do the test, the X_j will no longer be the original observations $V_1, \ldots, V_n$, which may be called the ungrouped data, but they'll be what are called grouped data.

Another way data can be grouped is that $V_1, ..., V_n$ might be real numbers, for example, and we want to test by χ^2 whether they have a normal $N(\mu, \sigma^2)$ distribution, so we have an $m = 2$ dimensional parameter. One can decompose the real line into k intervals (the leftmost and rightmost being half-lines) and do a χ^2 test. Here the numbers X_i of observations in each interval are already grouped data. (This way of testing normality is outdated now that we have the Shapiro–Wilk test.)

It tends to be very convenient to estimate the parameters based on ungrouped data, for all the cases mentioned (normal, Poisson, geometric) and hard to estimate them using grouped data. Unfortunately though, using estimates based on ungrouped data, but doing a chi-squared test on grouped data, violates the conditions for the X^2 statistic to have a χ^2 distribution with $k - 1 - m$ degrees of freedom, as many textbooks have claimed it does, although Rice, third ed., p. 359, correctly points out the issue. He also says "it seems rather artificial and wasteful of information to group continuous data." The problem is that the ungrouped data have more information in them than the grouped data do, and consequently, if the hypothesis H_0 is true, an estimate $\tilde{\theta}$ based on the ungrouped data tends to be closer to the true value θ_0 of the parameter than the estimate $\widehat{\theta}$ based on the grouped data would be, and consequently farther from the observations, in the sense measured by the X^2 statistic.

Let $\tilde{\theta}$ be an estimate of θ based on ungrouped data and let

$$\tilde{X}^2 = \sum_{j=1}^{k} \frac{(X_j - np_j(\tilde{\theta}))^2}{np_j(\tilde{\theta})}.$$

Chernoff and Lehmann (1954, Theorem 1) prove, under some regularity conditions, the following. Let $\tilde{\theta}$ be the maximum likelihood estimator of θ based on the ungrouped data, and suppose the given composite hypothesis H_0 that $\{\pi_j\}_{j=1}^k = \{p_j(\theta_0)\}_{j=1}^k$ for some θ_0 is true. Then as

$n \to \infty$, the distribution of $\tilde{X}^2$ converges to that of

$$\xi^2(\theta_0) = \sum_{j=1}^{k-m-1} Z_j^2 + \sum_{j=k-m}^{k-1} \lambda_j(\theta_0) Z_j^2$$

where $Z_1, ..., Z_{k-1}$ are i.i.d $N(0,1)$ and $0 \le \lambda_j(\theta_0) \le 1$ for $j = k-m$, $k-m+1, ..., k-1$. The values of λ_j all satisfy $0 < \lambda_j < 1$ in an example given by Chernoff and Lehmann, p. 586. In general we have

$$\chi^2(k-m-1) = \sum_{j=1}^{k-m-1} Z_j^2 \le \xi^2(\theta_0) \le \sum_{j=1}^{k-1} Z_j^2 = \chi^2(k-1),$$

and so for the $1-\alpha$ quantiles,

$$\chi^2_{1-\alpha}(k-m-1) \le \xi^2_{1-\alpha}(\theta_0) \le \chi^2_{1-\alpha}(k-1). \tag{1}$$

It is hard to get any information about the quantiles $\xi^2_{1-\alpha}(\theta_0)$ better than (1) because of the dependence on the unknown θ_0. From (1) we can conclude:

If $\tilde{X}^2 > \chi^2_{1-\alpha}(k-1)$, it follows that $\tilde{X}^2 > \xi^2_{1-\alpha}(\theta_0)$, so H_0 should be rejected.

On the other hand if $\tilde{X}^2 < \chi^2_{1-\alpha}(k-m-1)$, it follows that $\tilde{X}^2 < \xi^2_{1-\alpha}(\theta_0)$, so we can decide not to reject H_0. Another way to see this is that by definition of minimum chi-squared estimate $\widehat{\theta}$ based on the grouped data, we know that $\hat{X}^2_{\min} \le \tilde{X}^2$, and under H_0, $\hat{X}^2$ has as $n \to \infty$ a $\chi^2(k-1-m)$ distribution, so using $\hat{X}^2_{\min}$ we wouldn't reject H_0. This is true if $\theta \in H_0$ is estimated by any method, not only by maximum likelihood based on ungrouped data.

If $\tilde{X}^2$ is in an intermediate range

$$\chi^2_{1-\alpha}(k-m-1) < \tilde{X}^2 < \chi^2_{1-\alpha}(k-1)$$

then one is uncertain whether H_0 should be rejected, in other words whether the p-value of the test is less than α or not. Then one might do more computation, to evaluate the MLE or minimum chi-squared estimate $\widehat{\theta}$ of the parameter θ based on the grouped data X_j. It seems that these estimates may be difficult to compute by methods based on derivatives such as Newton's method or gradient descent. One may then use a search method with randomization, such as simulated annealing, but we won't go into that in this course.

If the computation is done, and all categories still have expected numbers $np_j(\widehat{\theta})$ at least 5, then $\hat{X}^2$ will have approximately a $\chi^2(k-$

$1-m$) distribution and one can do the test. If one is unlucky, some category may now have an expected number less than 5. Then I suppose one should stop and say we cannot reject H_0.

Another possibility is to gather more data and redo the test.

Historical Notes. Karl Pearson in 1900 first proposed the χ^2 test of a simple hypothesis for a multinomial with k categories and stated that the limiting distribution of X^2 as $n \to \infty$ is $\chi^2(k-1)$. According to Lancaster (1966), Bienaymé in 1838 had "very nearly anticipated K. Pearson's work on the normal approximation to the multinomial. Bienaymé (1852) used the gamma variable to obtain the distribution of a sum of squares in the least squares theory" i.e., apparently, to show that a $\chi^2(d)$ distribution is $\Gamma(d/2, 1/2)$.

Egon Pearson, of the Neyman–Pearson Lemma, was the son of Karl Pearson who invented the χ^2 test of fit.

REFERENCES

*Bienaymé, J. (1838). Sur la probabilité des résultats moyens des observations; démonstration directe de la règle de Laplace. *Mém. Sav. Étranger Acad. Sci.*, Paris, **5**, 513–558; repr. *ibid.* 1868 615-663.

*Bienaymé, J. (1852) *Mém. Sav. Étranger Acad. Sci.*, Paris, **15**, 615–663.

Chernoff, H., and Lehmann, E. L. (1954). The use of maximum likelihood estimates in χ^2 tests of goodness of fit. *Ann. Math. Statist.* **25**, 579–586.

Lancaster, H. O. (1966). Forerunners of the Pearson χ^2. *Austral. J. Statist.* **8**, 117–126.

Pearson, K. (1900). On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Phil. Mag.* (Ser. 5) **50**, 157–175.

(*) I have not seen these papers in the original.