Lecture 23 — Hypothesis testing for categorical data

23.1 Test of independence

Last lecture, we introduced the generalized likelihood ratio test, and we applied it to an example of testing the hypothesis of Hardy-Weinberg equilibrium in a population at a single diallelic locus. This was an example of testing whether the parameters of a multinomial model satisfy certain additional constraints.

Here is a second example of this type of hypothesis testing problem:

Example 23.1 (Independence test)**.** The following table (from the GSS 2008) cross-classifies a random sample of 1972 people by gender and by political party identification:

	dem	indep	repub
female	422	381	273
male	299	365	232

In this sample, approximately 39% of females identified as democrat and 25% identified as republican, while approximately 33% of males identified as democrat and 26% identified as republican. Is this significant evidence of an association between gender and party identification in the population from which this sample was drawn?

Denote the observed counts by N_{ij} for $i = 1, 2$ and $j = 1, 2, 3$. We may model these counts as multinomial with $n = 1972$ total observations and with outcome probabilities p_{ij} for $i = 1, 2$ and $j = 1, 2, 3$. Denote by $p_{i\cdot} = \sum_j p_{ij}$ and $p_{\cdot j} = \sum_i p_{ij}$ the marginal row and column probabilities. If there is no association between gender and party identification, then $p_{ij} = p_{i\cdot}p_{\cdot j}$. Hence we wish to test the **independence null hypothesis**

$$H_0 : p_{ij} = p_{i\cdot}p_{\cdot j} \text{ for all } i = 1, 2 \text{ and } j = 1, 2, 3.$$

The dimension of this sub-model may be determined as follows: The five row and column marginal probabilities $p_{1\cdot}, p_{2\cdot}, p_{\cdot 1}, p_{\cdot 2}, p_{\cdot 3}$ specify all of the multinomial cell probabilities under H_0. However, they satisfy the constraints $p_{1\cdot} + p_{2\cdot} = 1$ and $p_{\cdot 1} + p_{\cdot 2} + p_{\cdot 3} = 1$, so this sub-model has dimension $5 - 2 = 3$. The full multinomial model has dimension 5, so the generalized likelihood ratio statistic has approximate null distribution χ^2_2 (since $5 - 3 = 2$).

To derive the form of the likelihood ratio statistic in the above example, suppose more generally that we observe $(N_1, \ldots, N_k) \sim \text{Multinomial}(n, (p_1, \ldots, p_k))$, and we wish to test the null hypothesis $H_0 : (p_1, \ldots, p_k) \in \Omega_0$, where Ω_0 represents some sub-model. The multinomial likelihood is given by

$$\text{lik}(p_1, \ldots, p_k) = \binom{n}{N_1, \ldots, N_k} \prod_{i=1}^{k} p_i^{N_i}.$$

Letting $\hat{p}_{0,i}$ denote the MLEs in this sub-model Ω_0 and $\hat{p}_i$ denote the MLEs in the full multinomial model, the generalized likelihood ratio is

$$\Lambda = \frac{\text{lik}(\hat{p}_{0,1}, \ldots, \hat{p}_{0,k})}{\text{lik}(\hat{p}_1, \ldots, \hat{p}_k)} = \prod_{i=1}^{k} \left(\frac{\hat{p}_{0,i}}{\hat{p}_i} \right)^{N_i},$$

so

$$-2 \log \Lambda = 2 \sum_{i=1}^{k} N_i \log \frac{\hat{p}_i}{\hat{p}_{0,i}}.$$

Recall that the full model MLEs are given by $\hat{p}_i = N_i/n$, by Example 13.4 of Lecture 13. Let us write $E_i = \hat{p}_{0,i} n$, which denotes the "expected count" for outcome i corresponding to the sub-model MLE $\hat{p}_{0,i}$. Then we obtain the simple formula

$$-2 \log \Lambda = 2 \sum_{i=1}^{k} N_i \log \frac{N_i}{E_i}. \tag{23.1}$$

Example 23.2 (Independence test (cont'd))**.** Applying equation (23.1) to Example 23.1, we must compute the sub-model MLEs. Under H_0, the likelihood as a function of the row and column marginal probabilities is

$$\begin{aligned} \text{lik}(p_{1\cdot}, p_{2\cdot}, p_{\cdot 1}, p_{\cdot 2}, p_{\cdot 3}) &= \binom{n}{N_{11}, \ldots, N_{23}} \prod_{i=1}^{2} \prod_{j=1}^{3} (p_{i\cdot} p_{\cdot j})^{N_{ij}} \\ &= \binom{n}{N_{11}, \ldots, N_{23}} \prod_{i=1}^{2} p_{i\cdot}^{N_{i\cdot}} \prod_{j=1}^{3} p_{\cdot j}^{N_{\cdot j}}, \end{aligned}$$

where $N_{i\cdot} = \sum_j N_{ij}$ and $N_{\cdot j} = \sum_i N_{ij}$ are the row and column marginal counts. Taking the logarithm and introducing Lagrange multipliers for the constraints, we wish to maximize

$$\log \binom{n}{N_{11}, \ldots, N_{23}} + \sum_{i=1}^{2} N_{i\cdot} \log p_{i\cdot} + \sum_{j=1}^{3} N_{\cdot j} \log p_{\cdot j} + \lambda \left(\sum_{i=1}^{2} p_{i\cdot} - 1 \right) + \mu \left(\sum_{j=1}^{3} p_{\cdot j} - 1 \right).$$

Setting the derivatives with respect to $p_{i\cdot}$ and $p_{\cdot j}$ equal to 0 yields the equations $N_{i\cdot}/p_{i\cdot} + \lambda = 0$ and $N_{\cdot j}/p_{\cdot j} + \mu = 0$, so $p_{i\cdot} = -N_{i\cdot}/\lambda$ and $p_{\cdot j} = -N_{\cdot j}/\mu$. Picking the Lagrange multipliers $\lambda = -n$ and $\mu = -n$ enforces the constraints, and we obtain the MLEs $\hat{p}_{i\cdot} = N_{i\cdot}/n$ and $\hat{p}_{\cdot j} = N_{\cdot j}/n$. Then the sub-model MLEs for p_{ij} are given by $\hat{p}_{0,ij} = (N_{i\cdot}/n)(N_{\cdot j}/n)$.

For the data of Example 23.1, the row and column marginal counts are given by $N_{1\cdot} = 1076$, $N_{2\cdot} = 896$, $N_{\cdot 1} = 721$, $N_{\cdot 2} = 746$, and $N_{\cdot 3} = 505$. Computing the sub-model MLEs $\hat{p}_{0,ij}$ and multiplying by n, we obtain the table of expected counts E_{ij}:

	dem	indep	repub
female	393.4	407.0	275.5
male	327.6	339.0	229.5

Applying equation (23.1) with the 6 observed and expected counts yields $-2 \log \Lambda = 8.31$. Letting F denote the CDF of the χ^2_2 distribution, we obtain a p-value for the generalized likelihood ratio test of $1 - F(8.31) = 0.016$, so there is reasonably strong evidence of an association between gender and party identification.

23.2 Test of homogeneity

Consider now a slightly different problem: We have independent count observations from 2 multinomial distributions, each with k outcomes: $(N_1, \ldots, N_k) \sim \text{Multinomial}(n, (p_1, \ldots, p_k))$ and $(M_1, \ldots, M_k) \sim \text{Multinomial}(m, (q_1, \ldots, q_k))$, where n and m are known sample sizes. We wish to test the **homogeneity null hypothesis**

$$H_0 : p_i = q_i \text{ for all } i = 1, \ldots, k.$$

Example 23.3 (Homogeneity test)**.** This example is from Rice Section 13.3. When Jane Austen died, she left the novel *Sandition* partially completed. An admirer finished the novel, attempting to emulate Jane Austen's style. The following table counts the occurrences of six different short words in Chapters 1 and 6 of *Sandition*, written by Austen, and in Chapters 12 and 24 of *Sandition*, written by the admirer:

	a	an	this	that	with	without
Ch. 1 and 6	101	11	15	37	28	10
Ch. 12 and 24	83	29	15	22	43	4

Is there a significant difference between the relative frequencies of these words between the two authors?

Let us model the counts from Chapters 1 and 6 as $\text{Multinomial}(202, (p_1, \ldots, p_6))$ and those from Chapters 12 and 24 as $\text{Multinomial}(196, (q_1, \ldots, q_6))$. Then we wish to test the homogeneity null hypothesis that $p_i = q_i$ for all $i = 1, \ldots, 6$.

To derive the generalized likelihood ratio test, note that the joint likelihood of all parameters is the product of the two multinomial likelihoods:

$$\text{lik}(p_1, \ldots, p_k, q_1, \ldots, q_k) = \binom{n}{N_1, \ldots, N_k} \prod_{i=1}^{k} p_i^{N_i} \times \binom{m}{M_1, \ldots, M_k} \prod_{i=1}^{k} q_i^{M_i}. \tag{23.2}$$

Let $\hat{p}_i$ and $\hat{q}_i$ denote the full model MLEs, and let $\hat{p}_{0,i} = \hat{q}_{0,i}$ denote the sub-model MLEs. Then the generalized likelihood ratio statistic is

$$\Lambda = \prod_{i=1}^{k} \left(\frac{\hat{p}_{0,i}}{\hat{p}_i} \right)^{N_i} \prod_{i=1}^{k} \left(\frac{\hat{q}_{0,i}}{\hat{q}_i} \right)^{M_i},$$

so

$$-2 \log \Lambda = 2 \sum_{i=1}^{k} \left(N_i \log \frac{\hat{p}_i}{\hat{p}_{0,i}} + M_i \log \frac{\hat{q}_i}{\hat{q}_{0,i}} \right). \tag{23.3}$$

In the full model with two independent and unconstrained multinomial distributions, the MLEs are simply $\hat{p}_i = N_i/n$ and $\hat{q}_i = M_i/m$. Letting $E_i = \hat{p}_{0,i} n$ and $F_i = \hat{q}_{0,i} m$ denote the expected counts in the sub-model, we may write the above in the simple form

$$-2 \log \Lambda = 2 \sum_{i=1}^{k} \left(N_i \log \frac{N_i}{E_i} + M_i \log \frac{M_i}{F_i} \right).$$

To compute the above statistic, we need to compute the sub-model MLEs $\hat{p}_{0,i} = \hat{q}_{0,i}$. Under H_0, the likelihood in equation (23.2) simplifies to

$$\text{lik}(p_1, \ldots, p_k) = \binom{n}{N_1, \ldots, N_k} \binom{m}{M_1, \ldots, M_k} \prod_{i=1}^{k} p_i^{N_i + M_i}.$$

Taking the logarithm and introducing a Lagrange multiplier for the constraint $p_1 + \ldots + p_k = 1$, we wish to maximize

$$\log\left(\binom{n}{N_1, \ldots, N_k}\binom{m}{M_1, \ldots, M_k}\right) + \sum_{i=1}^{k}(N_i + M_i)\log p_i + \lambda\left(\sum_{i=1}^{k} p_i - 1\right).$$

Setting the derivatives with respect to p_i equal to 0, we obtain the equations $(N_i + M_i)/p_i + \lambda = 0$, so $p_i = -(N_i + M_i)/\lambda$. Choosing the Lagrange multiplier $\lambda = -(n + m)$ enforces the constraints, and we obtain the sub-model MLEs $\hat{p}_{0,i} = \hat{q}_{0,i} = (N_i + M_i)/(n + m)$.

Example 23.4 (Homogeneity test (cont'd))**.** In the data of Example 23.3, we have the marginal word counts $N_1 + M_1 = 184$, $N_2 + M_2 = 40$, $N_3 + M_3 = 30$, $N_4 + M_4 = 59$, $N_5 + M_5 = 71$, and $N_6 + M_6 = 14$. This yields the table of expected counts

	a	an	this	that	with	without
Ch. 1 and 6	93.4	20.3	15.2	29.9	36.0	7.1
Ch. 12 and 24	90.6	19.7	14.8	29.1	35.0	6.9

Applying equation (23.3) with the observed and expected counts, we obtain $-2\log\Lambda = 19.8$. The dimensionality of the sub-model in this example is 5 (6 parameters minus 1 constraint), and the dimensionality of the full model is 10 (12 parameters minus 2 constraints), so the null distribution of $-2\log\Lambda$ is approximately χ^2_5. Letting F denote the CDF of the χ^2_5 distribution, we obtain a p-value of $1 - F(19.8) = 0.0014$, so there is significant evidence of a difference in writing style between Austen and her admirer.

Remark 23.5. In both Examples 23.1 and 23.3, we wanted to test whether there is a significant difference in the relative frequencies between the two rows. The distinction between these examples is only in the sampling design/modeling assumption: In Example 23.1, we treated the counts from all rows as observations from a single multinomial distribution, because (we believe that) the GSS 2008 survey sampled a fixed total number of people rather than a fixed number of people of each gender. In Example 23.3, we modeled each row as a separate multinomial distribution with a fixed row sum.

In fact, the table of expected counts, generalized likelihood ratio statistic, and degrees of freedom for the test are all the same under the two different modeling assumptions (although we derived them in two different ways), so the tests of independence and of homogeneity are procedurally the same, and the distinction between these is sometimes blurred in practice.