Goodness of Fit

April 10, 2008

A goodness of fit test examine the case of a sequence if independent experiments each of which can have 1 of k possible outcomes. In terms of hypothesis testing, let $\pi = (\pi_1, \ldots, \pi_k)$ be postulated values of the probability

$$P_\pi\{\text{experiment takes on the } i\text{-th outcome}\} = \pi_i$$

and let $\mathbf{p} = (p_1, \ldots, p_n)$ denote the actual state of nature. Then, the parameter space is the $n-1$ **simplex**

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

The hypothesis test is

$$H_0 : p_i = \pi_i, \text{ for all } i = 1, \ldots, k \quad \text{versus} \quad H_1 : p_i \neq \pi_i, \text{ for some } i = 1, \ldots, k,$$

The data $\mathbf{x}$ is the outcome of the n experiments. A **sufficient statistic** is $\mathbf{n} = (n_1, \ldots, n_k)$ where n_i is the number of time that outcome i occurs in n experiments. Thus,

$$n = \sum_{i=1}^k n_i.$$

The likelihood function

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

Its logarithm

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

We maximize this using the method of Lagrange multipliers with constraint

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

Thus, at the maximum likelihood estimator $(\hat{p}_1, \ldots, \hat{p}_k)$,

$$\begin{aligned} \nabla_{\mathbf{p}} \ln L(\hat{\mathbf{p}}|\mathbf{n}) &= \lambda \nabla_{\hat{\mathbf{p}}} s(\mathbf{p}). \\ \left(\frac{n_1}{\hat{p}_1}, \ldots, \frac{n_k}{\hat{p}_k}\right) &= \lambda(1, \ldots, 1) \end{aligned}$$

So, $n_i/\hat{p}_i = \lambda, n_i = \lambda\hat{p}_i$. Now sum on i to obtain

$$\sum_{i=1}^{k} n_i = \lambda \sum_{i=1}^{k} \hat{p}_i \quad \text{and} \quad n = \lambda.$$

Consequently,

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

The **likelihood ratio test**

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

Recall that as the number of experiments $n \to \infty$,

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

converges to a χ^2_{k-1} random variable. Here $N = (N_1, \ldots, N_k)$ is the observed number of occurrences of outcome i.

The traditional method was introduced between 1985 and 1900 by Karl Pearson and consequenttly has been in use for longer that the idea of likelihood ratio tests. To show the connection between the two tests, recall that

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

is the quadratic Taylor polynomial approximation of $\ln a$. Apply this to the logarithm of the likelihood ratio, we find that

$$\begin{aligned}
-2\ln\Lambda_n(N) &= -2\sum_{i=1}^{k} N_i \left(\left(\frac{n\pi_i}{N_i} - 1\right) - \frac{1}{2}\left(\frac{n\pi_i}{N_i} - 1\right)^2\right) \\
&= -2\sum_{i=1}^{k} (n\pi_i - N_i) + \sum_{i=1}^{k} N_i \left(\frac{n\pi_i}{N_i} - 1\right)^2 \\
&= 0 + \sum_{i=1}^{k} \frac{(n\pi_i - N_i)^2}{N_i}
\end{aligned}$$

The is generally rewritten by writing $O_i = N_i$ to be the number of **observed** occurrences of i and $E_i = n\pi_i$ to be the number of **expected** occurrences of i as given by H_0. The data can be stored in a table

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

Then,

$$\sum_{i=1}^{k} \frac{(n\pi_i - N_i)^2}{N_i} \approx \sum_{i=1}^{k} \frac{(n\pi_i - N_i)^2}{n\pi_i} \approx \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}.$$

1 Contingency tables

For an $r \times c$ contingency table, we consider two classifications for an experiment. Thus, we can partition the outcome of each experiment into two groups:

$$A_1, \ldots A_c \quad \text{and} \quad B_1, \ldots B_r.$$

Here, we write O_{ij} to denote the number of occurences of the outcome $A_i \cap B_j$ are organize the results in a two-way table.

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

The null hypothesis is that the classifications A and B are independent . To set the parameter space for this model, we have the $rc - 1$ simplex

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

Write

$$p_{i\cdot} = \sum_{j=1}^{c} p_{ij} \quad \text{and} \quad p_{\cdot j} = \sum_{i=1}^{r} p_{ij}.$$

The hypothesis test is

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

Follow the procedure as before for the goodness of fit test to end with the test statistic

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

where

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