Topic 19: Goodness of Fit

November 24, 2009

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 $\mathbf{p} = (p_1, \ldots, p_m)$ denote the actual state of nature. Then, the parameter space is

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

The hypothesis test is

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

The data $\mathbf{x} = (x_1, \ldots, x_n)$ is the outcome of the n experiments. Set

$$n_i = \#\{j; x_j = i\}$$

to be the number of times the outcome i occurs in the data.

The likelihood function

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

Its logarithm

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

We maximize this using the method of Lagrange multipliers with constraint

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

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

$$\nabla_{\mathbf{p}} \ln L(\hat{\mathbf{p}}|\mathbf{n}) = \lambda \nabla_{\hat{\mathbf{p}}} s(\mathbf{p}).$$
$$\left(\frac{n_1}{\hat{p}_1}, \ldots, \frac{n_m}{\hat{p}_m}\right) = \lambda(1, \ldots, 1)$$

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

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

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 1895 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,

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

Example 1 (Hardy-Weinberg equilibrium)**.** *The two allele Hardy Weinberg principle states that after one generation of random mating the genotypic frequencies can be represented by a binomial distribution. So, if a population is segregating for two alleles A_1 and A_2 at an autosomal locus with frequencies p_1 and p_2. Then, random mating would give a proportion*

$$p_{11} = p_1^2 \textit{ for the } A_1A_1 \textit{ genotype}, \quad p_{12} = 2p_1p_2 \textit{ for the } A_1A_2 \textit{ genotype, and} \quad p_{22} = p_2^2 \textit{ for the } A_2A_2 \textit{ genotype}.$$

Then,

$$p_1 = p_{11} + \frac{1}{2}p_{12} \quad p_2 = p_{22} + \frac{1}{2}p_{12}.$$

This give us 5 parameters: $p_1, p_2, p_{11}, p_{12}, p_{22}$, plus the fact $p_1 + p_2 = p_{11} + p_{12} + p_{22} = 1$. Thus $n = 3$. The two restrictions from the Handy-Weinberg equilibrium above give $k = 2$.

McDonald et al. (1996) examined variation at the CVJ5 locus in the American oyster, Crassostrea virginica. *There were two alleles, L and S, and the genotype frequencies in Panacea, Florida were 14 LL, 21 LS, and 25 SS. So,*

$$\hat{p}_{11} = \frac{14}{60}, \quad \hat{p}_{12} = \frac{21}{60}, \quad \hat{p}_{22} = \frac{25}{60}.$$

So, the estimate of p_1 and p_2 are

$$\hat{p}_1 = \frac{49}{120}, \quad \hat{p}_2 = \frac{71}{120}.$$

So, the expected number of observations is

$$E_{11} = 60\hat{p}_1^2 = 10.00417, \quad E_{12} = 60 \times 2\hat{p}_1\hat{p}_2 = 28.99167, \quad E_{22} = 60\hat{p}_2^2 = 21.00417.$$

The chi-square statistic

$$\approx \frac{(14-10)^2}{10} + \frac{(21-29)^2}{29} + \frac{(25-21)^2}{21} = 1.600 + 2.207 + 0.762 = 4.569$$

The p-value

```
> 1-pchisq(4.569,1)
[1] 0.03255556
```

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:

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

Here, we write O_{ij} to denote the number of occurences of the outcome $A_i \cap B_j$ and 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

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

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

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

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

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

The null hypothesis $p_{ij} = p_{i\cdot}p_{\cdot j}$ can be written in terms of observed and expected observations as

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

or

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

Example 2. *We have the following data on malaria in three regions of the world.*

	Asia	South America	Africa	Total
Malaria Type A	31	45	14	90
Malaria Type B	2	53	5	60
Malaria Type C	53	2	45	100
Total	86	100	64	250

The expected table is

	Asia	South America	Africa	Total
Malaria Type A	30.96	36.00	23.04	90
Malaria Type B	20.64	24.00	15.36	60
Malaria Type C	34.40	40.00	25.60	100
Total	86	100	64	250

To compute the chi-square statistic

$$\begin{array}{rlllll}
 & \frac{(31-30.96)^2}{30.96} & + & \frac{(45-36.00)^2}{36.00} & + & \frac{(14-23.04)^2}{23.04} \\
+ & \frac{(2-20.64)^2}{20.64} & + & \frac{(53-24.00)^2}{24.00} & + & \frac{(5-15.36)^2}{15.36} \\
+ & \frac{(53-34.40)^2}{34.40} & + & \frac{(2-40.00)^2}{40.00} & + & \frac{(45-25.60)^2}{25.60} \\
= & 0.00005 & + & 2.250 & + & 3.564 \\
+ & 16.830 & + & 35.040 & + & 6.990 \\
+ & 10.060 & + & 36.100 & + & 14.700 \\
= & 125.516 & & & &
\end{array}$$

The degrees of freedom are $(c-1)(r-1) = 2 \times 2 = 4$. *In R, we have*

```
> malaria<-matrix(c(31,2,53,45,53,2,14,5,45),nrow=3,ncol=3)
> malaria
     [,1] [,2] [,3]
[1,]   31   45   14
[2,]    2   53    5
[3,]   53    2   45
> chisq.test(malaria)

Pearson's Chi-squared test

data:  malaria
X-squared = 125.5186, df = 4, p-value < 2.2e-16
```

2 Applicability of Chi-squared Tests

The chi-square test uses the central limit theorem and so is based on the ability to us a normal approximation. On criterion, the **Cochran conditions** requires no cell has count zero, and more than 80% of the cells have counts at least 5. If this does not hold, then **Fisher's exact test** uses the hypergeometric distribution directly rather than normal approximation.