

Balan

biomedical inc.

Document Title: Testing for adherence to a specified distribution, a parametric approach

Kenneth C. Syracuse, Ph.D., CQE
Chief Scientist
Balan Biomedical, Inc.

A Balan Biomedical, Inc. WHITE paper

April, 2009

Purpose: Often times we are presented with data from an unknown distribution. The purpose of this paper is to provide one means of testing the data's adherence to a specified parametric distribution.

Chi-squared test (χ^2)

Let X_1, X_2, \dots, X_n be a series of events within which an event can happen in just one way out of k different ways in a sequence of n independent trials. Allow that $P(X_i) = p_i$ for each trial with $\sum_{i=1}^k p_i = 1$. Then the probability that X_1 will occur N_1 times, X_2 will occur N_2 times, and, ..., X_k will occur N_k times is given by:

$$P(N_1, N_2, \dots, N_k) = \frac{n!}{\prod_{i=1}^k N_i!} \prod_{i=1}^k p_i^{N_i}$$

N_i is binomially distributed with parameters n, p_i so that

$$E(N_i) = np_i, \text{ and } \text{Var}(N_i) = np_i(1-p_i)$$

Now allow that we wish to test the hypothesis:

$$H_0: p_0=p_{10}, p_2=p_{20}, \dots, p_k=p_{k0},$$

where $p_{10}, p_{20}, \dots, p_{k0}$ are specified values of $P(X_1), P(X_2), \dots, P(X_k)$

satisfying $\sum_{i=1}^k p_{i0} = 1$. The likelihood of a set of observed values

N_1, N_2, \dots, N_k , under the null hypothesis H_0 , is

$$\lambda(N_1, N_2, \dots, N_k) = \frac{n!}{\prod_{i=1}^k N_i!} \prod_{i=1}^k p_{i0}^{N_i} .$$

Next, allow the alternative hypotheses to include all possible sets of values p_1, p_2, \dots, p_k , satisfying $p_i \geq 0$ and $\sum_{i=1}^k p_i = 1$.

The goal is to find the set of values p_1, p_2, \dots, p_k maximizing the likelihood relative to the hypothesis. That is, maximize:

$$H_{\max} = \max[\lambda(N_1, N_2, \dots, N_k | H)] = \max \left[\frac{n!}{\prod_{i=1}^k N_i!} \prod_{i=1}^k p_i^{N_i} \right]$$

subject to the constraints given above. Taking natural logarithms of both sides gives:

$$L = \ln \lambda = \ln \left(\frac{n!}{\prod_{i=1}^k N_i!} \right) + \sum_{i=1}^k N_i \ln p_i.$$

Because of the constraint, only $(k-1)$ terms can be determined independently and

$$p_k = 1 - [p_1 + p_2 + \dots + p_{k-1}].$$

Differentiating with respect to p_i gives,

$$\frac{\partial L}{\partial p_i} = \frac{N_i}{p_i} - \frac{N_k}{p_k}.$$

Setting this equal to 0, and recalling $\sum_{i=1}^k p_i = 1$ gives $p_i = \frac{N_i}{n}$.

That p_i is a maximum is confirmed by setting $\frac{\partial^2 L}{\partial p_i^2} = 0$. Substituting this into the previous result gives:

$$\lambda(N_1, N_2, \dots, N_k | H_{\max}) = \frac{n!}{n^n \prod_{i=1}^k N_i!} \prod_{i=1}^k N_i^{N_i}$$

Writing the likelihood ratio in terms of H_0 and H_{\max} gives:

$$\frac{\lambda(N_1, N_2, \dots, N_k | H_0)}{\lambda(N_1, N_2, \dots, N_k | H_{\max})} = \prod_{i=1}^k \left(\frac{np_{i_0}}{N_i} \right)^{N_i},$$

as the test statistic. By defining:

$np_{i_0} =$ the expected number of events $X_i = E_i$, and

N_i = the observed number of events $X_i = O_i$ the resultant likelihood ratio (above) can be rewritten as:

$$\prod_{i=1}^k \left(\frac{E_i}{O_i} \right)^{O_i} .$$

Applying a linearization transformation gives:

$$\sum_{i=1}^k O_i \ln \left(\frac{E_i}{O_i} \right) = - \sum_{i=1}^k [(O_i - E_i) + E_i] \ln \left[1 + \frac{(O_i - E_i)}{E_i} \right]$$

Recall that $E(O_i) = np_{io} = E$, and the standard deviation $\sqrt{np_{io}(1 - np_{io})}$

which is on the order of $\sqrt{E_i}$, then $(O_i - E_i)$ will be on the order of

$\sqrt{E_i}$ or \sqrt{n} . This result implies that the right hand side of the

above expression need only be extended to retain terms up to \sqrt{n} .

Since $\sum_i (O_i - E_i) = 0$, we have:

$$\begin{aligned} & \sum_{i=1}^k O_i \ln \left(\frac{E_i}{O_i} \right) \\ &= - \sum_{i=1}^k [(O_i - E_i) + E_i] \left[\frac{(O_i - E_i)}{E_i} - \frac{(O_i - E_i)^2}{2E_i^2} + \frac{(O_i - E_i)^3}{3E_i^3} - \dots \right] \\ &= - \frac{1}{2} \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} + \frac{1}{6} \sum_{i=1}^k \frac{(O_i - E_i)^3}{E_i^2} + \dots \end{aligned}$$

The first term is used to replace the test's likelihood function and

thus if the higher order terms are small:

$$\sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \sim \chi_{k'}^2, \text{ where}$$

k' is equal to k minus the number of distinct linear relationships. And so by counting the number of observations falling within a given interval and comparing this observed value to that of the hypothesized distribution, a goodness of fit measurement is achieved.