Goodness of fit test

Let X be a random variable. Given i.i.d. copies of X, the goodness of fit test determines whether X has a certain distribution (e.g. normal, uniform, Student's T).

Multinomial distribution

The multinomial distribution is a generalization of the binomial distribution. A multinomial distribution with K modalities has K possible outcomes (for the binomial distribution, $K = 2$).

The parameters are as follows:

- n' is the number of trials.
- p_i is the probability of observing the ith outcome in a single trial. $\sum_{i=1}^Kp_i$ must equal 1.

Let $\textbf{p} = [p_1 \ p_2 \ \ldots \ p_K]^T$. The multinomial distribution can be represented as a random vector $\textbf{N} \in \mathbb{Z}$, where $N^{(i)}$ represents the count of outcome i. The multinomial pdf for all $\textbf{n}$ such that $\sum_{i=1}^{K}n^{(i)}=n', n^{(i)} \geq 0, i = 1,\ldots,K$, and $n^{(i)} \in \mathbb{Z}, i = 1,\ldots,K$ is given by:

$$p_N(N^{(1)}=n^{(1)}, \ldots ,N^{(K)}=n^{(K)}) = \frac{n'!}{n^{(1)}!n^{(2)}! \ldots n^{(K)}!}\prod_{i=1}^{K}p_i^{n^{(i)}}$$

The likelihood function of a sequence of n trials $X_1,X_2, \ldots , X_n \sim X$ is

$$L_n(X_1, \ldots ,X_n, p_1, \ldots , p_K) = p_1^{N_1}p_2^{N_2} \ldots p_K^{N_K}$$

From this result, the maximum likelihood estimator for each of the probabilities is:

$$\hat{p}_i = \frac{N_i}{n'}$$

χ^2 test for multinomial distribution

The χ^s test tests if a discrete RV comes from a certain PMF.

Theorem

$$n\sum_{j=1}^K\frac{(\hat{\textbf{p}}-\textbf{p}_j^0)^2}{\textbf{p}_j^0}\rightarrow_{n\rightarrow\infty}^{(d)}\chi^2_{K-1}$$

For a test of the form

$$H_0: \textbf{p} = \textbf{p}^0$$ $$H_1: \textbf{p} \neq \textbf{p}^0$$

the test statistic is $$T_n=n\sum_{j=1}^K\frac{(\hat{p}_j-p_j^0)^2}{p_j^0}$$

A χ^2 test with asymptotic level α is

$$\psi = \mathbf{1}\left\{ T_ n > q_\alpha^{K-1} \right\}$$

The asymptotic p-value is then given by $P(\chi^2_{K-1} > T_n)$. **Note that the degrees of freedom is given by $K-1$, the number of modalities minus one, not by $N-1$, the number of observations minus one.** This can be calculated in MATLAB using the following command:

```
chi2cdf(T_n, K-1, 'upper')
```