TWO EXAMPLES OF CHI-SQUARE GOODNESS-OF-FIT TEST

William W. S. Chen
Department of Statistics
The George Washington University
Washington D.C. 20013

ABSTRACT

In this paper we present two examples of what is typically referred to as Chi-Squared Goodness of fit test. The first example is multinormial parametric distribution, while the second is Kruskall Wallis Pearson nonparametric statistics by ranking. We show that both distributions have Chi-Squared distributions with k-r degree of freedom.

Mathematical Subject Classification : 62H10

Keywords and phases: Chi-Squared Goodness of fit test, Kruskall Wallis Pearson nonparametric statistics, nonparametric statistics, multinormial parametric distribution, parametric statistics, ranking.

INTRODUCTION

Since most of the standard statistical techniques are based on the assumption of normality, the methods for judging the normality of a set of data are of importance. To this end we usually apply the method called the Chi-Squared goodness of fit test. We test the hypothesis H_0: the data comes from normal distribution versus H_a the data does not come from normal distribution. We proceed the following steps to complete this test. Step 1 : we grouped the data into classes to form a frequency distribution and calculated the sample mean , $\bar{x}$, and standard deviation, s. Step 2 : from these quantities, a normal distribution is fitted and the expected frequencies in each class are obtained. Let us use f_i to denote the observed frequencies and E_i the expected frequencies. Step 3: for each class, we compute and record the quantity $(f_i - E_i)^2 / E_i = (obs - expected)^2 / expected$. Step 4: the final test criterion is $\sum (f_i - E_i)^2 / E_i \sim \chi^2_{(k-r)}$ summed over all the classes. It has shown that this test statistics has a $\chi^2_{(k-r)}$ distribution with (k-r) degree of freedom, where k is the number of classes used in computing $\chi^2_{(k-r)}$ and r is the number of parameters that have been replaced by sample estimates. If the data actually come from a normal distribution, this quantity approximately follows the theoretical $\chi^2_{(k-r)}$ distribution with (k-r) degree freedom. If the data came from some other distribution, the observed f_i will tend to agree poorly with the values of E_i that are expected on the assumption of normality and computed $\chi^2_{(k-r)}$ becomes large. Consequently, large values of $\chi^2_{(k-r)}$ cause us to reject the hypothesis of normality. The book of Snedeco G.W. Cochran W.G.[1] is the best source related to this topic.

Multinomial Distribution Parametric Application

Let p_i denote Pr[an object falling into cell i] such that $p_1 + ... + p_k = 1$
x_i denote the number of object out of n, observed in cell i, such that
$x_i + ... + x_k = n.$ then

$$f(x_1 .....x_k) = \frac{n!}{x_1! x_2!...x_k!} p_1^{x_1} p_2^{x_2} ......p_k^{x_k} \qquad where\ x_i = 0,1,2..... .for\ all\ i$$

Then $E(x_i) = np_i$, $Var(x_i) = np_i(1 - p_i)$, $Cov(x_i , x_j) = -np_i p_j$
Then the variance covariance matrix of $\underline{x}$, $say\ \Sigma\ is$

$$\Sigma = \begin{bmatrix} np_1(1 - p_1) & -np_1p_2 & . & . & -np_1p_k \\ -np_1p_2 & np_2(1 - p_2) & . & . & -np_2p_k \\ . & . & . & . & . \\ . & . & . & . & . \\ -np_1p_k & -np_2p_k & . & . & np_k(1 - p_k) \end{bmatrix}$$

Then Σ is singular with rank deficiency 1.

$$Proof : \Sigma = nV = n[D_p - \underline{p}\,\underline{p}'] \ \ where\ \underline{p}' = [p_1 ,,......, p_k]$$

$$then\ \Sigma\ \underline{j} = n[D_p - \underline{p}\,\underline{p}']\underline{j} = n[\underline{p} - \underline{p}\,\underline{p}'\,\underline{j}] = n[\underline{p} - \underline{p}] = n\underline{0} = \underline{0}$$

Σ *is singular.* $\underline{j}$ *is a column vector of 1.*

Then conditional inverse of Σ *is* $\Sigma^{(-1)} = D_{np}^{-1} = D_{\frac{1}{np}}$

$\underline{X} \sim Multinormial$ *; then any* $x_i \sim Binomial.$ x_i , x_j *are correlated,*

$$so\ x_i \sim normal,\ \ y_i = \frac{x_i - np_i}{\sqrt{np_i(1 - np_i)}} \rightarrow N(0,1)$$

$$then\ \ y_i^2 \sim \chi^2_{(1)}.$$

$$so\ \ q = (\underline{x} - n\underline{p})' \Sigma^{(-1)} (\underline{x} - n\underline{p}) \rightarrow \chi^2_{(k-1)}$$

then q can be express as a sum

$$q = (\underline{x} - n\underline{p})' D_{np}^{-1} (\underline{x} - n\underline{p})$$

$$= \sum \frac{(x_i - np_i)^2}{np_i} \rightarrow \chi^2_{(k-1)}$$

Finally, we can rename

$$\underline{x_i} : \underline{o_i} \; ; \; and \;\; np_i = E_i \; (expected\, number)$$

$$then \quad q = \sum \frac{(x_i - np_i)^2}{np_i} = \sum \frac{(O_i - E_i)^2}{E_i} \rightarrow \chi^2_{(k-1)}$$

This result shows that q has Chi-Square distribution with k-1 degree of freedom.

Kruskal Wallis Pearson Statistics

Observations are available as rank only. R_i is the sum of the rank in group i (i=1,2,…k).

$$R_1 + R_2 + \ldots + R_k = \sum_{i=1}^{n} i = \frac{n(n+1)}{2} \; ,$$

We want to find the standard quadratic form associated with $R_1, R_2, \ldots R_k$
(in the central limit theorem sense) This quadratic form will have χ^2 distribution with (k-1) degree of freedom. Let there be n_i observations in group i. Let x_{ij} denote the rank associated with the jth observation in group i. If there is no difference between group mean then

$$E(x_{ij}) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij} / n = \frac{n+1}{2},$$

$$E(R_i) = \frac{n_i(n+1)}{2},$$

$$E(x_{ij}, x_{kl}) = \sum \sum_{i \neq j} \frac{ij}{n(n-1)} = [\sum \sum ij - \sum i^2] / n(n-1)$$

$$= [\sum i \sum j - \sum i^2] / n(n-1)$$

$$= \{ (\frac{n(n+1)}{2})^2 - \frac{n(n+1)(2n+1)}{6} \} / n(n-1)$$

$$= \frac{(n+1)(3n^2 - n - 2)}{12(n-1)} = \frac{(n+1)(3n+2)}{12}$$

$$Cov(x_{ij}, x_{kl}) = E(x_{ij}, x_{kl}) - E(x_{ij})E(x_{kl})$$

$$= \frac{(n+1)(3n+2)}{12} - (\frac{n+1}{2})^2 = - \frac{n+1}{12}$$

$$E(x_{ij}^2)=\sum_{i=1}^{n}\frac{i^2}{n}=\frac{(n+1)(2n+1)}{6},$$

$$Var(x_{ij})=E(x_{ij}^2)-(E(x_{ij}))^2=\frac{(n+1)(2n+1)}{6}-\frac{(n+1)^2}{4}$$

$$=\frac{(n+1)(n-1)}{12}$$

$$Var(R_i)=Var(\sum_{j=1}^{n_i}x_{ij})=n_i\frac{(n+1)(n-1)}{12}-n_i\frac{(n+1)(n_i-1)}{12}$$

$$=\frac{n_i(n+1)(n-n_i)}{12}$$

$$Cov(R_i,R_j)=\sum_{j=1}^{n_i}\sum_{i=1}^{n_k}Cov(x_{ij},x_{kl})=\frac{-n_in_k(n+1)}{12}$$

Conclude that

$$Var(R_i)=\frac{(n+1)(nn_i-n_i^2)}{12},\ Cov(R_i,R_j)=\frac{-n_in_k(n+1)}{12}.$$

$$Var(R)=\frac{n+1}{12}\begin{vmatrix} nn_1-n_1^2 & -n_1n_2 & . & . & -n_1n_k \\ -n_1n_2 & nn_2-n_2^2 & . & . & -n_2n_k \\ . & . & . & . & . \\ . & . & . & . & . \\ -n_1n_k & -n_2n_k & . & . & nn_k-n_k^2 \end{vmatrix}$$

Singular rank deficiency by 1.

$$(Var(R))^{(-1)}=(Var(R)+nn')^{-1};\ add\ nn'\ where\ n=\begin{vmatrix} n_1 \\ * \\ n_k \end{vmatrix}$$

$$\Sigma^{(-1)}=\frac{12}{n+1}\begin{vmatrix} nn_1 & 0 & . & 0 & 0 \\ 0 & nn_2 & . & 0 & 0 \\ . & . & . & . & . \\ 0 & 0 & . & . & 0 \\ 0 & 0 & . & 0 & nn_k \end{vmatrix}^{-1}=\frac{12}{n(n+1)}D_{n_i}^{-1}$$

$$Recall\ x'D_wx=\sum_{i=1}^{k}w_ix_i^2;$$

$$x=R-\frac{n+1}{2}n;\quad i.e.\ \ x_i=R_i-\frac{n+1}{2}n_i;\ w_i=\frac{1}{n_i}$$

$$Q=\frac{12}{n(n+1)}\sum_{i=1}^{k}\frac{(R_i-\frac{n+1}{2}n_i)^2}{n_i}=\frac{6}{n}\sum_{i=1}^{k}\frac{(R_i-\frac{n+1}{2}n_i)^2}{\frac{n+1}{2}n_i}$$

$$\rightarrow \chi^2_{(k-1)}\ d.f\ \text{in the central limit}$$

The above result has demonstrated that Q has Chi-Square distribution with k-1 degree of freedom.

CONCLUSION

We have presented two examples that related to Chi-Squared distribution. Both examples are commonly used in elementary statistics. In Kendall M. and Stuart A. book[2] there are more related theory and application in multinomial distribution. In Lehmann E.L. and D'Abrera H.J.M. book [3] there are the richest source of theory and applications of nonparametric statistics.

REFERENCES

[1] Snedecor G.W. and Cochran W.G. (1955) Statistical Methods. Sixth Edition, The Iowa State University Press, Ames, Iowa, U.S.A.
[2] Kendall, M. and Stuart A. (1976) The Advanced Theory of Statistics. Volume 1. 4th Edition. Macmillan Publishing Company, Inc.
[3] Lehmann E.L. and D'Abrera H.J.M. (1975) Nonparametrics Statistical Methods Based on Ranks. Holden-Day, Inc. San Francisco.