Poisson Transformation for ANOVA

Dechavudh Nityasuddhi[1)], Prachoom Suwattee[2)]

[1)] Department of Biostatistics, Faculty of Public health, Mahidol University,
402/1 Rajvithee, Rachathewi, Bangkok 10400, Thailand (phdnt@mahidol.ac.th)
[2)]School of Applied Statistics, National Institute of Development Administration,
118 SeriThai, Bangkapi, Bangkok 10240, Thailand (prachoom@as.nida.ac.th)

Abstract

Most of the insurance researches are concerned with rare events such as deaths or accidents. Unknown parameters need to be estimated, the hypotheses about the parameters tested and conclusions made for decision-making on goal setting and policy planning. When the data are from Poisson distributions the likelihood ratio test with chi-square approximation is usually applied to compare more than two population means. This test uses asymptotic property, and sample size tends to infinity, so it is called an approximation test. For the data from normal distributions with homogeneity of variance, ANOVA, the most powerful test for complicated analysis, is used. Therefore, in this research Poisson data are appropriately transformed to fit the assumptions for ANOVA. The following transformation is proposed:

$$y_{ij} = \begin{cases} \dfrac{(x_{ij}+\frac{1}{2})^{\gamma} + \bar{x}_{i.}^{\gamma}}{\gamma}, & \gamma \neq 0, \\ \log(x_{ij}+\frac{1}{2}), & \gamma = 0, \end{cases}$$

where x_{ij} is a random variable value in the j^{th} trial from the i^{th} Poisson population with mean $\bar{x}_{i.}$; y_{ij} the transformed variable in the j^{th} trial from the i^{th} group ; and γ, an unknown constant to be estimated so that it will maximize the likelihood function, and the transformed variables are normal with homogeneous variances [1]. But the estimate of γ cannot be solved easily. A numerical method called a half partition method is used for the estimation. The algorithm starts with any initial value of γ and is increased or decreased one unit at a time until the turning point of the likelihood is reached and then the half partition method is applied. The result shows that the transformed data follow the two required assumptions for ANOVA, and the proposed test is expected to be more powerful than the ordinary chi-square approximation of the likelihood ratio test.

1. Introduction

Data on accidents are important for making decisions in insurance. Most of the unexpected events in insurance are considered as rare events with Poisson distribution. Researchers usually apply the likelihood ratio test for testing the hypotheses about the Poisson parameter [2]. The logarithmic likelihood ratio test statistic has chi-square approximation when the sample size is large enough [3]. But many researchers often use this test for medium and small sample size of data without serious consideration about the power of the test. In the case of testing the difference of more than two Poisson population means, the approximation chi-square for likelihood ratio test is still applied. Since ANOVA is a most powerful test for normal populations with homogenous variances [4,5], ANOVA is recommended instead of the likelihood ratio test. An appropriate transformation from the Poisson data to normally distributed data with homogeneity of variances is needed if ANOVA will be applied for testing the difference between more than two population means.

2. Transformation

Bartlett [6] and Cochran [7] suggested the square-root transformation for Poisson data in order to obtain approximately normal distributions with equal variances. This transformation is appropriate for large sample sizes but does not use any characteristics of the Poisson data through its expected value. The transformation is valid for general situations where the data are greater than zero [1]. However, the Poisson data can have zero values, too. Thus a new transformation is proposed which should be suitable for any Poisson data.

$$y_{ij} = \begin{cases} \dfrac{(x_{ij}+\frac{1}{2})^{\gamma} + \bar{x}_{i.}^{\gamma}}{\gamma}, & \gamma \neq 0, \\ \log(x_{ij}+\frac{1}{2}), & \gamma = 0, \end{cases}$$

where x_{ij} is a random variable in the j^{th} trial from the i^{th} Poisson population with mean $\bar{x}_{i.}$; y_{ij}, the transformed variable of x_{ij} ; and γ, an unknown constant to be estimated, so that y_{ij} will meet the required properties.

The maximum likelihood method is now applied for the estimation of γ. The likelihood function of the transformed data should be normally distributed with homo geneity of variances. So

$$\log L(\gamma|N,H) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \frac{1}{n\gamma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}+\tfrac{1}{2})^{2\gamma} - \sum_{i=1}^{k}\frac{1}{n_i}\sum_{j=1}^{n_i}(x_{ij}+\tfrac{1}{2})^{\gamma}{}^2$$

$$-\frac{n}{2}+(\gamma-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i}\log(x_{ij}+\tfrac{1}{2}),$$

where $n=\sum_{i=1}^{k} n_i$, and other parameters except γ have been estimated by the maximum likelihood method. The value of γ that gives the maximum value of the $\log L(\gamma|N,H)$ must be estimated. However, this cannot be solved easily with partial differentiation. Alternatively, we will use a numerical method called the half partition method to find such a γ value for a specific set of data.

3. The half partition method

To find the value of γ that gives the maximum value of the $\log L(\gamma|N,H)$, γ , an initial value for γ_0 needs to be set. Then the initial value of γ must be decrease or increase by one unit of measurement for monotone increasing function $\log L(\gamma|N,H)$. This step must be repeated until the value of $\log L(\gamma|N,H)$ increased. Then the half partition method is performed. Consider the last three values of γ namely S_1, S_3 and S_5. Next, S_2 and S_4 are calculated by partitioned half way between S_1 and S_3, and between S_3 and S_5 respectively, i.e., $S_2=\frac{S_1+S_3}{2}$ and $S_4=\frac{S_3+S_5}{2}$. S_2, S_3 and S_4 must be concentrated on. If the $\log L(\gamma|N,H)$ at $\gamma = S_2 <$ $\log L(\gamma|N,H)$ at $\gamma = S_3 <$ $\log L(\gamma|N,H)$ at $\gamma = S_4$, then the maximum value of $\log L(\gamma|N,H)$ should come from the value between S_3 and S_5. If the $\log L(\gamma|N,H)$ at $\gamma = S_2 >$ $\log L(\gamma|N,H)$ at $\gamma = S_3 >$ $\log L(\gamma|N,H)$ at $\gamma = S_4$ then the maximum value of $\log L(\gamma|N,H)$ should come from the value of γ between S_1 and S_3. Otherwise, the maximum value of $\log L(\gamma|N,H)$ should come from the value between S_2 and S_4. From the last three points of the possible range to make $\log L(\gamma|N,H)$ maximum, the half partition method must be repeated again until the absolute value of the difference between two values of $\log L(\gamma|N,H)$ from the last three points is less than any predetermined small value. Thus γ equals to the point S that gives the maximum value of $\log L(\gamma|N,H)$. The following algorithm can summarize the half partition process:

1. Let $\gamma_1 = \gamma_0$, where γ_0 = initial value of γ
2. Let γ_2 = increased by one unit of γ_1
3. If $[l(\gamma_2)>l(\gamma_1)]$, then go to 4
   else if $[l(\gamma_2)=l(\gamma_1)]$, then $s_1=\gamma_1$, $s_5=\gamma_2$, go to 6
   else if $[l(\gamma_2)<l(\gamma_1)]$, then go to 10
4. Let $\gamma_3 = \gamma_2 +1$
5. If $[l(\gamma_3)>l(\gamma_2)]$, then $\gamma_1=\gamma_2$, $\gamma_2 = \gamma_3$, go to 4
   else if $[l(\gamma_3)=l(\gamma_2)]$, then $s_1=\gamma_2$, $s_5=\gamma_3$, go to 6
   else if $[l(\gamma_3)<l(\gamma_2)]$, then $s_1=\gamma_1$, $s_5=\gamma_3$, go to 6
6. Let $s_3=(s_1+s_5)/2$, $s_2=(s_1+s_3)/2$, $s_4=(s_3+s_5)/2$
7. If $[(l(s_2)<l(s_3)).and.(l(s_3)<l(s_4))]$, then $s_1=s_3$
   else if $[(l(s_4)<l(s_3)).and.(l(s_3)<l(s_2))]$, then $s_5=s_3$
   else $s_1=s_2$; $s_5=s_4$
8. Let $s_3=(s_1+s_5)/2$
9. If $[\min(l(s_3)-l(s_1), l(s_3)-l(s_5))<\varepsilon]$, then $l(s_{max})=\max[l(s_1),l(s_3),l(s_5)]$,
   $\gamma_m = s_{max}$, go to 13
   else go to 6
10. Interchange value between γ_1 and γ_2

11. Let γ_3 = decreased by one unit of γ_2
12. If [l(γ_3)>l(γ_2)], then $\gamma_3 = \gamma_2$, $\gamma_2 = \gamma_3$, go to 10
    else if [l(γ_3)=l(γ_2)], then $s_1 = \gamma_2$, $s_5 = \gamma_3$, go to 6
    else if [l(γ_3) < l(γ_2)], then $s_1 = \gamma_3$, $s_5 = \gamma_3$, go to 6
13. Let $\gamma = \gamma_m$

4. Example

To give a demonstration we use IMSL library to simulate two Poisson data sets. First, Poisson data with equal in means are generated. Six groups of the size of 50 with the same Poisson population mean, $\lambda = 20$, were tested. The half partition method was used to figure out the estimate γ equal to 0.54687494. Shapiro-Wilks test was used to check for normality, (that all groups were normally distributed). Then Bartlett test was applied to check for homogeneity of variances.

The gamma value
Gamma = 0.54687494, Max. log-lh = -870.0544
Test for normality
G 1; Lambda=20, S-W= 0.9537, p= 0.084076
G 2; Lambda=20, S-W= 0.9741, p= 0.509959
G 3; Lambda=20, S-W= 0.9608, p= 0.169080
G 4; Lambda=20, S-W= 0.9793, p= 0.704457
G 5; Lambda=20, S-W= 0.9663, p= 0.278335
G 6; Lambda=20, S-W= 0.9850, p= 0.890000
Test for homogeneity of variances
Bartlett test (correction) = 0.09960890, p-value=0.99983925

The second data set was generated for the case of Poisson data with difference in mean. Six groups of Poisson data of the size of 50 with different means -- 1, 5, 10, 20, 50 and 100 -- were generated. The derived value of γ is 0.44848543. The transformed data were checked for normality and homogeneity of variances, and again the Shapiro-Wilks and Bartlett tests were used. The results were accepted as expected.

The gamma value
Gamma = 0.44848543, Max. log-lh = -795.1220
Test for normality
G 1; Lambda =1 , S-W= 0.8991, p= 0.070938
G 2; Lambda =5 , S-W= 0.9565, p= 0.111939
G 3; Lambda =10 , S-W= 0.9592, p= 0.135283
G 4; Lambda =20 , S-W= 0.9626, p= 0.199536
G 5; Lambda =50 , S-W= 0.9892, p= 0.971022
G 6; Lambda =100, S-W= 0.9755, p= 0.560295
Test for homogeneity of variances
Bartlett test (correction) = 2.69390059, p-value=0.74705124

5. Conclusion

The proposed transformation needs to estimate one parameter γ from the log-likelihood of normal distribution and homogeneity of variances. The numerical method with half partition is applied to figure out the estimates. The results show that the transformed data have a normal distribution and homogeneity of variances. From these properties of the transformed data, ANOVA may be applied to test the differences of the population means. Practically, it is easy to find γ before ANOVA can be applied by using a computer program to call subroutine from IMSL library.

References

[1] Box G.E.P. and Cox, D.R.; An Analysis of Transformations, Journal of the Royal Statistical Society B, Vol.26, pp211-243, 1964.
[2] Vu, H.T.V. and Maller, R.A.; The Likelihood Ratio Test for Poisson versus Binomial Distributions, Journal of the American Statistical Association, Vol.91, pp818-824, 1996.
[3] Kendall, M.G. and Stuart, A.; The Advanced Theory of Statistics, London: Griffin, pp234-267, 1973.
[4] Eisenhart, C.; The Assumptions Underlying the Analysis of Variance, Biometrics, Vol.3,

pp1-21, 1947.
[5] Curtiss, J.H.; On Transformations used in the Analysis of Variance, Annals of Mathematical Statistics, Vol.14, pp107-122, 1943.
[6] Bartlett, M.S.; The Square Root Transformation in Analysis of Variance, Supplement to the Journal of the Royal Statistical Society, Vol.3, pp68-78, 1936.
[7] Cochran, W.G.; The Analysis of Variance where Experimental Errors follow the Poisson or Binomial laws, Annals of Mathematical Statistics, Vol.11, pp335-347, 1940.