APPLIED AND ENVIRONMENTAL MICROBIOLOGY, May 1985, p. 1304–1306
0099-2240/85/051304-03$02.00/0
Vol. 49, No. 5
Copyright © 1985, American Society for Microbiology

Some Goodness-of-Fit Methods for the Poisson Plus Added Zeros Distribution

A. H. EL-SHAARAWI

National Water Research Institute, Canada Centre for Inland Waters, Burlington, Ontario, Canada L7R 4A6

Received 22 October 1984/Accepted 31 January 1985

Methods for making inferences about the Poisson plus added zeros distribution and the truncated Poisson distribution are presented and illustrated with bacteriological data. Some of the methods are designed for testing the compatability of the zero frequency with the Poisson distribution, whereas others are given for testing the goodness of fit for the truncated Poisson. In particular, a modified form of the Fisher index of dispersion is presented which is suitable for the truncated case. It is shown that the use of the usual expression of the index of dispersion for testing the adequacy of the truncated Poisson is not correct and leads to accepting inadequate fits more frequently than expected on the basis of test of significance. Furthermore, three test statistics are presented for testing the compatability of the zero frequency with the Poisson distribution. The results of the simulation show that two test statistics, one due to Cochran (W. G. Cochran, Biometrics 10:417–451, 1954) and the other to Rao and Chakravarti (C. R. Rao and I. M. Chakravarti, Biometrics 12:264–282, 1956), are preferable to those from the likelihood ratio test.

The use of the Poisson distribution as a model for the variation of the number of bacteria in equal samples from completely homogeneous material is well known (3, 6–9). However, the conditions needed for the application of the Poisson model are rarely fulfilled in practice (4). The reasons range from the natural heterogeneity of the material sampled to problems associated with performing the bacteriological analysis. To accommodate for the situation with greater heterogeneity than that expected under the Poisson model, Fisher (5) introduced the negative binomial as a more general model for the dispersion of bacteria. To derive the negative binomial, Fisher assumed that the mean of the Poisson distribution is not a fixed quantity but is itself a random variable with a two-parameter gamma distribution. One of the parameters can be used as a measure for the deviation from complete homogeneity. Another model for accommodating the bacteriological heterogeneity is the Poisson plus added zeros (PZ), which was used by Christian and Pipes (1). According to this model the number of bacteria found in parallel samples follows a Poisson distribution with the exception of the zero frequency. The underlying assumption in the derivation of the PZ model is that the material sampled consists of two portions. The first is completely free of bacterial contamination, whereas the distribution of bacteria in the second is random and hence can be modeled by the truncated Poisson distribution (TP).

The aims of this paper are to present a number of methods for testing the adequacy of the PZ and TP models. Specifically, the following are given: (i) three different tests for determining whether the zero frequency is actually responsible for the lack of fit of the Poisson model (the first of these tests is the likelihood ratio test, the second was proposed by Cochran [*C*] [2], and the third was developed by Rao and Chakravarti [*R*] [11]); (ii) an index of dispersions for testing the adequacy of the TP; and (iii) numerical illustrations with data reported by Christian and Pipes (1).

MATERIALS AND METHODS

PZ. Let θ (Table 1 defines all terms used in the paper) be the proportion of the material that is free of bacterial contamination and $(1 - \theta)$ be the proportion of contaminated material. Then, under the assumption that bacteria are distributed at random in the contaminated part, the probability of observing r bacteria in the sample is

$$P(r) = \begin{cases} (1 - \theta)e^{-\lambda} + \theta & \text{when } r = 0 \\ (1 - \theta)e^{-\lambda}\,(\lambda^r/r!) & \text{when } r \neq 0 \end{cases} \tag{1}$$

where λ is the bacterial density in the contaminated part. Equation 1 is the PZ model. There are two unknown parameters θ and λ in equation 1. When $\theta = 0$ the PZ model reduces to the Poisson model, and when $\theta = 1$ the material is free of bacterial contamination.

TP. When the frequency of zeros is the source of the lack of fit, one might consider ignoring the zero frequency in the analysis. In this case the appropriate model is the TP which is

$$P(r) = (e^{-\lambda}\lambda^r)/[r!(1 - e^{-\lambda})] \text{ when } r > 0 \tag{2}$$

The mean and the variance of the TP model are $\mu = \lambda/(1 - e^{-\lambda})$ and $\sigma^2 = \mu(1 - e^{-\lambda}\,\mu)$, respectively. From this it appears that the mean is always larger than the variance, but the difference between σ^2 and μ decreases with increasing λ, and when $\lambda \rightarrow \infty$ then $\sigma^2 = \mu$.

Estimation of λ and θ. Suppose that bacteriological analysis was performed on $n + n_0$ samples and that n_0 of these samples have zero counts, whereas the remaining n samples have the counts $r_1, r_2, \ldots, r_n$. The maximum likelihood estimates $\hat{\lambda}$ and $\hat{\theta}$ of λ and θ, respectively, are shown by

$$\hat{\lambda} = \bar{r}_t\,(1 - e^{-\hat{\lambda}}) \tag{3}$$

and

$$\hat{\theta} = (P_{\text{obs}} - e^{-\hat{\lambda}})/(1 - e^{-\hat{\lambda}}) \tag{4}$$

where $\bar{r}_t = (r_1 + r_2 + \ldots + r_n)/n$ and $P_{\text{obs}} = n_0/(n + n_0)$. The solution for $\hat{\lambda}$ can be obtained iteratively. It is easy to show that $(\bar{r} - 1) < \hat{\lambda} < \bar{r}$, and hence the search for $\hat{\lambda}$ should be restricted to the interval $(\bar{r} - 1, \bar{r}_t)$. The iterative solution $\hat{\lambda}$ can be obtained in the following manner. Take $\hat{\lambda}_0 = \bar{r}_t$ as a

TABLE 1. Table of definitions

Term	Definition
θ	Proportion of material that is free of bacterial contamination
$1 - \theta$	Proportion of contaminated material
λ	Bacterial density in the contaminated part
PZ	PZ
TP	TP
μ	Mean of TP
σ^2	Variance of TP
∞	Infinity
n	No. of contaminated samples
n_0	No. of uncontaminated samples
$N = n + n_0$	Total no. of samples
$\hat{\lambda}$	Maximum likelihood estimate for λ
$\hat{\theta}$	Maximum likelihood estimate for θ
$\bar{r}$	Mean of bacterial counts (all data)
$\bar{r}_t$	Mean of bacterial counts for contaminated samples
P_{obs}	Proportion of uncontaminated samples
$-2\ln\Lambda$	Likelihood ratio test
C	C test
R	R test
T	Sum of all the observations
D^2	Fisher index of dispersion
D_t^2	Index of dispersion for TP

first approximation for $\hat{\lambda}$ and compute a better approximation of $\hat{\lambda}_1$ as

$$\hat{\lambda}_1 = \bar{r}_t\,(1 - e^{-\hat{\lambda}_0})$$

Then replace $\hat{\lambda}_0$ by $\hat{\lambda}_1$ and repeat the process until two successive values of $\hat{\lambda}_1$ are almost identical. This value of $\hat{\lambda}_1$ will be the required estimate of $\hat{\lambda}$. It should be pointed out that $\hat{\lambda}$ is the maximum likelihood estimate for λ under the TP and PZ models.

Discrimination between the Poisson model and the PZ model. The problem of discriminating between the Poisson and the PZ models is reduced to testing the hypothesis H_0 of $\theta = 0$ against the alternative hypothesis H_1 of $\theta \neq 0$. There are three test statistics for doing this: (i) the likelihood ratio test, (ii) the C test, and (iii) the R test (Table 2).

Goodness of fit for the TP. If the hypothesis $\theta = 0$ is rejected, then a Poisson distribution is not appropriate for fitting all the data. However, this does not imply that the PZ model is the appropriate model. One way of determining that is to test whether the positive samples can be represented by the TP model. A test for doing this was obtained by Rao and Chakravarti (1/1) and is the index of dispersion for the TP.

TABLE 2. Tests for discriminating between the Poisson model and the PZ model

Test	Formula[a]
Likelihood ratio	$-2\ln\Lambda = 2\{n_0 \ln P_{\text{obs}} + n \ln [(1 - P_{\text{obs}})\bar{r}_t]/(\lambda) - n\hat{\lambda} + n\bar{r}_t \ln \lambda + N\bar{r}\,(1 - \ln \bar{r})\}$
C	$C = (n_0 - Ne^{-\bar{r}})/[Ne^{-\bar{r}}\,(1 - e^{-\bar{r}} - \bar{r}e^{-\bar{r}})]^{1/2}$
R	$R = \{n_0 - [(N-1)/N]^T N\}/\{N[(N-1)/N]^T - N^2[(N-1)/N]^{2T} + N(N-1)[(N-2)/N]^T\}^{1/2}$

[a] ln, Natural logarithm; $\bar{r}$, mean of all the data. For large N the distribution of $-2\ln\Lambda$ is chi square with one degree of freedom, and the distribution of C and R is normal with a mean of 0 and unit variance.

$$D_t^2 = \Sigma(r_i - \bar{r}_t)^2/[\bar{r}_t(1 + \hat{\lambda} - \bar{r}_t)] \qquad (5)$$

The asymptotic distribution of D_t^2 is chi square with $(n - 1)$ degrees of freedom.

Confidence interval for λ. If λ is the parameter of interest in the PZ model or the TP model, then the approximate 95% confidence limits for λ are

$$\hat{\lambda} \pm 1.96\,\hat{\lambda}(1 - e^{-\hat{\lambda}})/\sqrt{n(1 - e^{-\hat{\lambda}} - \hat{\lambda}e^{-\hat{\lambda}})} \qquad (6)$$

RESULTS AND DISCUSSION

Simulation study. The behavior of the three test statistics $-2\ln\Lambda$, C, and R are compared by experimental sampling or simulation. The Control Data Corporation Cyber 171 computer was used to generate samples from the Poisson distribution with a mean λ of 5. Two groups of experiments were generated. In the first group, 1,000 sets of samples of 15 observations were generated. The three test statistics were computed for each of the 1,000 samples. For each test the proportions of the 1,000 sets of samples for which the null hypothesis was rejected at the 5 and 1% levels were recorded. The proportions estimate the significance levels of the various tests. In the second group of experiments, the above procedure was repeated, except that a number of zeros were added to the 15 observations. In one set of 1,000 experiments two zeros were added, and in the other set five zeros were added. Hence in these experiments the null hypothesis ($\theta = 0$) is not true; then the proportion of the 1,000 samples for which the null hypothesis was rejected at the 5 and 1% levels represents the empirical powers of various tests. Moreover, the above simulation was repeated with samples of 50 observations so that the effect of increasing the sample size on the estimate of the significance level can be determined.

The results of these simulations are given in Table 3. It appears that $-2\ln\Lambda$ gives the closest estimate to the true significance level; however, the power of this test is much lower than that of the other two tests. For samples of 15 observations the significance levels with the two tests C and R are overestimated; however, these levels are closer to the true significance levels for samples of 50 observations. For a fixed sample size, the power of the tests increases as the proportion of zeros increases. As an illustration, the power of the likelihood ratio test ($-2\ln\Lambda$) increases from 0.153 to 0.946 as θ increases from 2/15 = 0.13 to 5/15 = 0.33. Furthermore, the power increases when the sample size

TABLE 3. Empirical power functions for testing the zero frequency

True significance level (α)	Tests for zero frequency	Sample size					
		15			50		
		0[a]	2[a]	5[a]	0[a]	2[a]	5[a]
0.05	$-2\ln\Lambda$	0.040	0.153	0.946	0.050	0.138	0.528
	C	0.075	0.722	0.992	0.040	0.266	0.905
	R	0.085	0.794	0.996	0.043	0.275	0.922
0.01	$-2\ln\Lambda$	0.010	0.055	0.613	0.016	0.027	0.191
	C	0.037	0.418	0.945	0.025	0.170	0.619
	R	0.055	0.556	0.956	0.026	0.180	0.652

[a] Number of zeros added.

TABLE 4. Tests for the zero frequency, TP, ML, and the confidence interval for three data sets (1)

Data set	No. of locations sampled (locations with coliforms)	Mean current CFU/100 ml		Test for the zero frequency			Maximum likelihood estimate	95% confidence interval	Index of dispersion	
		$\bar{r}$	$\bar{r}_t$	$-2\ln\Lambda$	C	R			D^2	D_t^2
CV	225 (10)	0.11	2.52	76.32	12.621	12.962	2.256	(0.714–3.798)	36	48.91
WHI	66 (6)	0.40	4.42	86.90	9.550	9.774	4.363	(0.792–7.934)	29	30.73
BBI	92 (11)	0.51	4.27	132.78	11.347	11.439	4.206	(1.657–6.755)	287	306.50

increases while θ remains fixed. For example, the powers of $-2\ln\Lambda$ are 0.153 for samples of 15 observations and $\theta = 0.13$ and 0.528 for samples of 50 observations and $\theta = 5/50 = 0.10$. On the basis of power calculation, the C and R statistics are preferable to the $-2\ln\Lambda$ statistics, with R slightly better than C. From these results, the use of test statistics C and R are recommended.

Example. Three data sets reported by Christian and Pipes (1) are used to illustrate the methods of this paper. Table 4 gives the values of the three test statistics $-2\ln\Lambda$, C, and R for testing whether the observed frequency of zeros is compatible with that expected for the Poisson model, the maximum likelihood and the confidence limits for λ, the index of dispersion D^2 which was calculated by Christian and Pipes, and the index of dispersion D_t^2. The R values exceed those of C in the three cases. This agrees with the simulation results which indicate that the R test is more sensitive than the C test (i.e., more significant results will be obtained if the R test is used instead of the C test). The statistic $-2\ln\Lambda$ can be compared with C^2 and R^2 since the square of the R and C results will have asymptotically the same distribution as $-2\ln\Lambda$. The comparison does not indicate that $-2\ln\Lambda$ is consistently larger or smaller than C^2. In fact $-2\ln\Lambda$ is smaller than C^2 for the data sets CV and WHI and larger than R^2 for the BBI case. All three tests indicated that the frequency of zero is significantly different from that expected for the Poisson model. The last two columns of Table 3 give the values of D_t^2 for the TP model and D^2, i.e., the correct index of dispersion. The values of D_t^2 are larger than those of D^2. In fact, this is always true and can be shown immediately by noting that

$$\begin{aligned} D^2 &= \Sigma(r_i - \bar{r}_t)^2/\bar{r}_t \\ &= D^2{}_t\,(1 + \lambda - \bar{r}_t) \end{aligned}$$

and by observing that if $\bar{r}_t - 1 < \hat{\lambda} < \bar{r}_t$, then $D^2 < D_t^2 < \infty$. This means that significant values obtained by calculating D^2 are also significant when D_t^2 is calculated, but the inverse is not true.

In conclusion the statistic D_t^2 should always be used instead of D^2 for testing the goodness of fit of the TP distribution. For testing the adequacy of the zero frequency the C and R statistics are preferable since they have higher power than $-2\ln\Lambda$. The C statistic has the added advantage of being easier to compute.

LITERATURE CITED

1. **Christian, R. R., and W. O. Pipes.** 1983. Frequency distribution of coliforms in water distribution systems. Appl. Environ. Microbiol. **45:**603–609.
2. **Cochran, W. G.** 1954. Some methods of strengthening χ^2 tests. Biometrics **10:**417–451.
3. **El-Shaarawi, A. H., S. R. Esterby, and B. J. Dutka.** 1981. Bacterial density in water determined by Poisson or negative binomial distributions. Appl. Environ. Microbiol. **41:**107–116.
4. **El-Shaarawi, A. H., and W. O. Pipes.** 1982. Enumeration and statistical inferences. *In* W. O. Pipes (ed.), Bacterial indicators of pollution. CRC Press, Inc., Boca Raton, Fla.
5. **Fisher, R. A.** 1941. The negative binomial distribution. Ann. Eugenics **11:**182–187.
6. **Fisher, R. A., H. G. Thornton, and W. A. Mackenzie.** 1922. The accuracy of the plating method of estimating the density of bacterial populations with particular reference to the use of Thronton's agar medium with soil samples. Ann. Appl. Biol. **9:** 325–359.
7. **Gosset, W. S.** 1907. On the error of counting with a haemocytometer. Biometrika **5:**357.
8. **McCrady, M. H.** 1915. The numerical interpretation of fermentation tube results. J. Infect. Dis. **17:**183.
9. **Pipes, W. O., P. Ward, and S. H. Ahn.** 1977. Frequency distributions for coliform bacteria in water. J. Am. Water Works Assoc. **69:**644–647.
10. **Rao, C. R.** 1973. Linear statistical inference and its applications. John Wiley & Sons, Inc., New York.
11. **Rao, C. R., and I. M. Chakravarti.** 1956. Some small sample tests of significance for a Poisson distribution. Biometrics **12:** 264–282.