Journal of Modern Applied Statistical Methods

Volume 7 | Issue 1 Article 22

5-1-2008

Test for Spatio-Temporal Counts Being Poisson

Haiyan Chen
University of Maryland, hchen@umaryland.edu

Howard H. Stratton
SUNY, hstratton2001@yahoo.com

Journal of Modern Applied Statistical Methods
May, 2008, Vol. 7, No. 1, 259-274

1538 – 9472/08/$95.00

Test for Spatio-Temporal Counts Being Poisson

Haiyan Chen
University of Maryland

Howard H. Stratton
SUNY

The new Log-Linear Test (T_L) is proposed to identify when the Poisson model fails for a collection of count random variables. T_L is shown to have better rejection rate with small sample size and essentially the same power compared to a classical Fisher-Bohning's Statistic T_F for standard alternatives to Poisson.

Key words: Fisher-Bohning's Statistic, log-linear test, over-dispersion to Poisson, Negative binomial, Zero-inflated Poisson.

Introduction

Human disease data are often in the form of count data and its associated rate. Examples include disease incidence, prevalence, and/or mortality (Lindsey, 1995, Hinde & Demetrio, 1998). The Poisson distribution is a traditional probability model for count data (Hinde & Demetrio, 1998), and has the property that its expected value equals its variance, i.e., $E(Y) = var(Y)$. Thus, count data for which $E(Y) < var(Y)$ indicate over-dispersion relative to the Poisson. The extra disparity could be due to heterogeneity in the population, or an overabundance of certain specific values, e.g., excess zeros (Tiago de Oliveria, 1965, Bohning, 1994, Lambert & Roeder, 1995, Lindsey, 1995, Hinde & Demetrio, 1998, Brown & Zhao, 2002, Smyth, 2003). The term 'over-dispersion' is reserved for over-dispersion relative to the Poisson, i.e., any random variable that has variance-mean ratio greater than one is called over-dispersed. Failure to take account of this over-dispersion can lead to serious underestimation of standard errors and misleading inference for the regression parameters.

Suppose that I independent count variables Y_i (i = 1, …, I) are each observed N times. The associated I sample means are given by $\overline{Y}_i = \sum_j Y_{ij} / N$ and I sample variances are given by $S_i^2 = \sum_j (Y_{ij} - \overline{Y})^2 / (N-1)$, where j = 1, …, N. Hypothesis tests that an individual count variable is Poisson have been developed (Hoel, 1943, Tiago de Oliveria, 1965, Cameron & Trivedi, 1990, Bohning, 1994, Lambert & Roeder, 1995, Lindsey, 1995, Hinde & Demetrio, 1998, Brown & Zhao, 2002, Smyth, 2003), however, little development of tests of hypothesis that a group of count variables are simultaneously distributed as Poisson has been done. This paper investigates three possible hypothesis tests that a group of variables are simultaneously Poisson vs. over-dispersion to the Poisson. It will be seen that only one of these tests is feasible in terms of both test size and test power for relatively small number of observations for each of the variables.

The main data set used to illustrate the proposed tests in this paper, named *NYSLD*, was derived from the New York State Department of Health Lyme Registry Surveillance System. Only confirmed cases using the Lyme disease

Haiyan Chen is a Research Assistant Professor in the Division of Health Services Research in the Department of Health Promotion and Policy in the Dental School. His research interests include Bayesian methods applied to clinical trial design, spatio-temporal data analysis, and survey design and data analysis. Email: hchen@umaryland.edu. Howard Stratton is a Professor in the Department of Epidemiology and Biostatistics in the School of Public Health, SUNY. His research interests include measurement error models, selection bias correction, spatial statistics, mixture models, and hidden Markov chains. Email: hstratton2001@yahoo.com

(LD) surveillance definition (White, Chang, Benach, et al., 1991, CDC, 1997) were selected. For each case, county of residence and year of onset were used. Cases with either of these two pieces of information missing were excluded. The LD data of three neighboring states of New York were accessed online from the Connecticut Department of Public Health (Connectcut State Depatment of Health website, 2004), New Jersey Department of Health and Senior Services (New Jersey Department of Health and Senior Services website, 2004), and Pennsylvania Department of Health (Pennsylvania Department of Health website, 2004) websites.

A standard reference statistic: the Fisher-Bohning Statistic

A simple diagnostic test for over-dispersion of a single variable has been a long sought goal for deciding whether a further investigation of latent heterogeneity is necessary. Tiago de Oliveira (1965) approached this via the difference of $D_i = (S_i^2 - \overline{Y}_i)$ for an individual random variable Y_i. They argued that D_i' variance under the null hypothesis $Y_{ij} \sim Poisson(\lambda_i)$ is given by $(1 - 2\lambda_i^{1/2} + 3\lambda_i)/N$, which can be estimated by $(1 - 2\overline{Y}_i^{1/2} + 3\overline{Y}_i)/N$. The proposed test statistic, $O_T = N^{1/2}(S_i^2 - \overline{Y}_i)/(1 - 2\overline{Y}_i^{1/2} + 3\overline{Y}_i)^{1/2}$, was treated as if it had a standard Normal limiting null distribution.

However, Dankmar Bohning (1994) showed by simulation that the limiting distribution of Tiago de Oliveira's statistic under the Poisson null hypothesis is neither a standard normal nor is it independent of λ_i. Bohning noted that the failure of Tiago de Oliveira's test is due to incorrect computation of the standard deviation of D_i and showed its correct variance to be $2\lambda_i^2/(N-1)$. The corrected test statistic,

$$O_T^n = \frac{S_i^2 - \overline{Y}_i}{\sqrt{2\overline{Y}_i^2/(N-1)}} = \{(N-1)/2\}^{1/2}\{(S_i^2/\overline{Y}_i) - 1\}$$

is asymptotically N(0, 1).

To address the multiple comparison problem in this paper, the overall *p*-value for the *I* independent Bohning's over-dispersion tests is calculated using the Fisher's statistic for combining independent tests (Hedges & Olkin, 1995) and is named the Fisher-Bohning's Statsitc (T_F).

If p_i is *p*-value of *i*th individual test for a continuous test statistic, p_i has a uniform (0, 1) distribution when the null hypothesis H_{0i} of the test is true. Fisher's procedure then uses the fact that -2log p_i has a χ^2 distribution with two degrees of freedom. Because the sum of independent χ^2 variables has a χ^2 distribution with degree of freedom equal to sum of the degrees of freedom of each individual χ^2, the Fisher-Bohning's Statistic, $T_F = -2\log(p_1 p_2 \cdots p_n) = -2\sum_{i=1}^{I} \log p_i$, has a χ^2 distribution with $2n$ degrees of freedom under the null hypothesis H_0. Although the null hypothesis H_{0i}: $Y_i \sim$ Poisson involves a discrete distribution, it is well approximated by continuous normal distributions if the expected values of the corresponding Poissons are sufficiently large (Johnson, Kotz, & Kemp, 1992). Thus the χ^2 null distribution for Fisher-Bohning's statistic T_F is applicable to the Poisson null hypothesis in this case.

Two new test statistics

Before presenting new test procedures, there are some general concepts and theorems that need to be introduced. First, the concepts 'corresponding zero-inflated variable' and 'corresponding zero-inflated distribution' are defined.

Let Y be a random variable with probability function $p(Y)$ and ω be a value between 0 and 1. If a random variable $\widetilde{Y}$ has

$$\widetilde{Y} = \begin{cases} 0 \text{ with probability } \omega \\ Y \text{ with probability } 1\text{-}\omega \end{cases}$$

then $\widetilde{Y}$ has density function:

$$p(\tilde{Y}) = \begin{cases} \omega + (1-\omega)p(0) & \tilde{Y} = 0 \\ (1-\omega)p(y) & \tilde{Y} = y > 0 \end{cases}$$

and is called the corresponding zero-inflated variable to Y with zero-inflated distribution, $p(\tilde{Y})$.

Theorem 1: If $E(Y^k)$ exists, then $E(\tilde{Y}^k) = (1-\omega)E(Y^k)$, and $\frac{\text{var}(\tilde{Y})}{E(\tilde{Y})} = \frac{\text{var}(Y)}{E(Y)} + \frac{\omega}{1-\omega}E(\tilde{Y})$ for $E(Y) > 0$, $E(\tilde{Y}) > 0$.

Corollary 1: If $\frac{\text{var}(Y)}{E(Y)} = 1 + \eta E(Y)$ holds, then $\frac{\text{var}(\tilde{Y})}{E(\tilde{Y})} = 1 + \frac{\omega + \eta}{1-\omega}E(\tilde{Y})$.

The theorem and its corollary presented above provide a basis on which new tests of whether a distribution with the property $\frac{\text{var}(Y)}{E(Y)} = 1 + \eta E(Y)$ is over-dispersed to the Poisson, i.e. $\eta > 0$ are developed. More precisely the two new proposed tests of hypothesis deal with:

1. $\{Y_i, i = 1, \ldots, I\}$ which are I independent random variables.
2. For each Y_i, N independent records were observed.
3. Test Y_i being simultaneously Poisson (λ_i) by the null hypothesis

   H_0: $\frac{\text{var}(Y_i)}{E(Y_i)} = 1$ for $E(Y_i) > 0$

   for all i

   versus over-dispersion to the Poisson

   H_1: $\frac{\text{var}(Y_i)}{E(Y_i)} = 1 + \eta E(Y_i)$ for

   $E(Y_i) > 0$, and $\eta > 0$.

That is, a test of $\eta = 0$ vs. $\eta > 0$.

A test based on a linear regression of sample variance-mean ratio on the sample mean

Under H_1, $\frac{E(y_{ij} - E(Y_i))^2}{E(Y_i)} - 1 = \eta E(Y_i)$. If $E(Y_i)$ is known, a test for over-dispersion is a t-test for η in the least-squares (LS) regression

$$\frac{S_{Y_i}^2}{E(Y_i)} - 1 = \eta E(Y_i) + \varepsilon_i,$$

where the error term is defined by $\varepsilon_i = \frac{S_{Y_i}^2}{E(Y_i)} - \frac{\text{var}(Y_i)}{E(Y_i)}$. Since $E(Y_i)$ is unknown, it is estimated by $\overline{Y}_i$. A linear regression of the sample variance-mean ratio on sample mean $\frac{S^2}{\overline{Y}} = \beta_0 + \beta_1 \overline{Y} + \varepsilon$ is made so that a test for over-dispersion or under-dispersion is a test of whether $\beta_1 = 0$ by treating the test statistic, $T_L = \frac{\hat{\beta}_1 - 1}{\sqrt{\hat{\text{var}}(\beta_1)}}$, as N(0, 1) under H_0. WLS is used to estimate regression coefficients and t-test is used to draw statistical inferences. Note that above justification has been intuitively developed rather than by strict logic. It will be shown to be unreliable.

A test based on a linear regression of the log-sample variance on the log-sample mean

An alternative to T_S is suggested by re-expressing the alternative H_1, $\frac{\text{var}(Y)}{E(Y)} = 1 + \eta E(Y)$ for $E(Y_i) > 0$, via a logarithmic transformation to give $\log(\text{var}(Y)) = \log(1 + \eta\, E(Y)) + \log(E(Y))$. The Poisson condition of $\eta = 0$ is then equivalent to $\log(\text{var}(Y)) = \log(E(Y))$.

The unknown var(Y) and E(Y) are estimated by S^2 and $\overline{Y}$. In order to test under/over-dispersion to Poisson, a least square fit of $\log S^2 = \beta_0 + \beta_1 \log \overline{Y} + \varepsilon$ is made. Rejecting either $\beta_0 = 0$ or $\beta_1 = 1$ is sufficient to reject the Poisson. Here a test of whether $\beta_1 = 1$ is proposed by treating the test statistic, $T_L = \frac{\hat{\beta}_1 - 1}{\sqrt{\hat{\text{var}}(\beta_1)}}$, as N(0, 1) in rejecting H_0. The

validity using this distribution under H_0 will be confirmed by simulation studies.

Simulation

General design for the simulations

Simulations were conducted to examine and compare the size and power characteristic of the three proposed tests T_S, T_L, and T_F. Two alternative hypotheses to a null hypothesis of Poisson that will be tested are 1) H_1: $Y_i \sim$ Negative Binomial, i.e. NB (μ_i, ν), and 2) H_1: $Y_i \sim$ Zero Inflated Poisson, ZIP (μ_i, ω).

Four sample sizes, $N = 5$, 11, 50, or 100 are used to resemble data of small, moderate, and large sample size. Each simulation experiment is based on 500 replications. Two nominal α levels, $\alpha = 0.01$, or $\alpha = 0.05$ are evaluated.

In all cases, μ_i is set equal to ith NYS county's observed average annual incidence rates (per 100,000 population) of LD, where $i = 1, \ldots, 57$, in order to have a practical sense of how the tests perform relative to the NYSLD data. Note that during the 11-year studied time period, none of the 57 counties in NYS had a zero average annual incidence rate of LD, indicating that every of those counties had at least one case reported in some year. Figure 1a-b graph the empirical distributions of μ_i and $\log(\mu_i)$, respectively. Range of μ_i is from 0.55 to 349.00 and this covers a relative wide range. The distribution of $\log(\mu_i)$ (skewness = 1.40) is much less positively skewed than that of μ_i (skewness = 3.86).

Analysis of test size

Four sets of data from H_0: $Y_i \sim$ Poisson (μ_i) were generated corresponding to four sample sizes, $N = 5$, 11, 50, 100 for each i ($i = 1, \ldots, 57$, and see Section 'General design for the simulations' for the values assigned to μ_i). Percentages of rejections of H_0: $Y_i \sim$ Poisson (μ_i) were calculated for the two αs: $\alpha = 0.01$ and $\alpha = 0.05$ in order to evaluate whether sizes of tests were sufficiently close to their nominal αs. The results are summarized in Table 1.

The size of the test T_S for all four sample sizes turns out to be considerably greater than both the nominal sizes of 1% and 5% (Table 1). The size of the test T_L for all four samples sizes is statistically indistinguishable from both the nominal sizes of 1% and 5%.

The match between the actual and nominal sizes for T_F is different for different sample sizes: for small sample size as $N = 5$, the size of the test T_F turns out to be smaller than both the nominal sizes of 1% and 5%; while for moderate and large sample size ($N = 10$, 50, or 100), the size of the test T_F is statistically indistinguishable from both the nominal sizes of 1% and 5%.

In summary, the match between the actual and nominal size is worst for T_S and best for T_L. When sample size is adequately large, T_F as well as T_L have consistent test sizes.

Power analysis

Because T_S does not have consistent test sizes but T_L and T_F, essentially do, in the following Sections, power investigations are only made for T_L and T_F as a function of increasing sample size or the degree of over-dispersion.

Power analysis under alternative hypothesis H_1: $Y_i \sim$ NB

Under the alternative hypothesis H_1: $Y_i \sim$ NB (μ_i, ν), the probability density is

$$p(Y_i) = \frac{\Gamma(y_i + \nu)}{\Gamma(\nu)\Gamma(y_i + 1)} \left(\frac{\nu}{\nu + \mu_i}\right)^{\nu} \left(\frac{\mu_i}{\nu + \mu_i}\right) \text{an}$$

d its variance-mean ratio is $\frac{\text{var}(Y)}{E(Y)} = 1 + \frac{E(Y)}{\nu}$.

Taking the logarithm, this equation becomes:

$$\log(\text{var}(Y)) = \log(E(Y)) + \log(1 + E(Y)/\nu)$$
$$= g(E(Y)).$$

In this experiment, the test power is set up as a function of ν, the dispersion parameter, with 25 different values for ν set discretely from 1 to 5000. This simulates the degree of over-dispersion from large ($1 \le \nu < 50$), moderate ($50 \le \nu < 500$) to small ($500 \le \nu < 5{,}000$) correspondently. At each value of ν, the experiment described is performed.

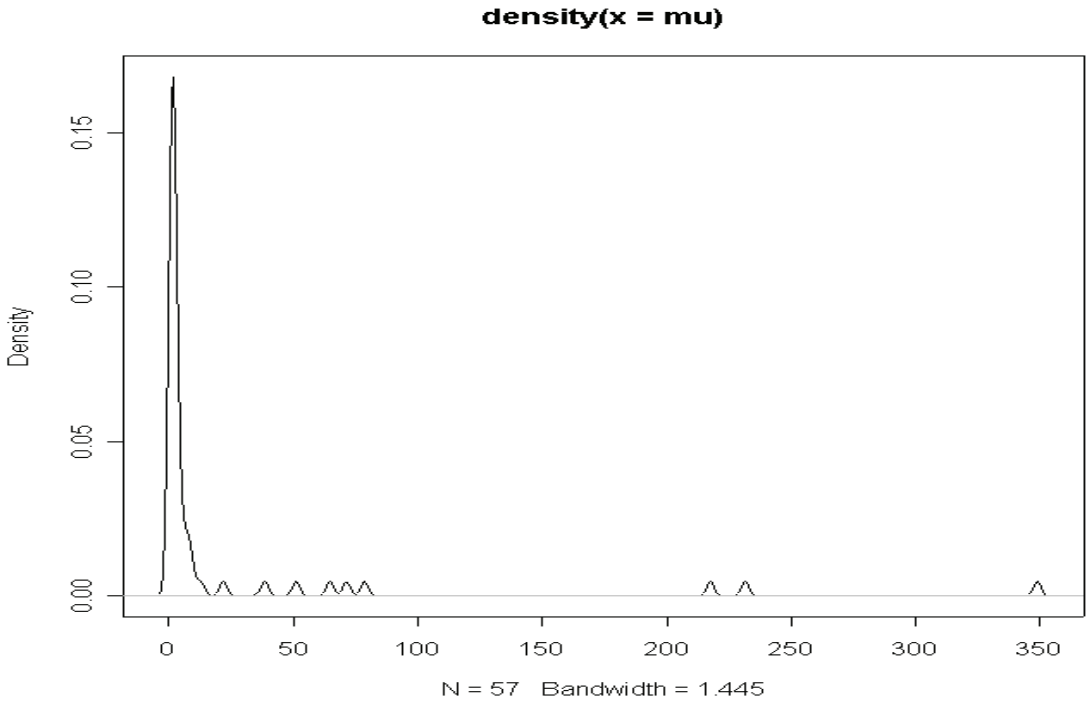


Figure 1.a Empirical distribution of μ_is

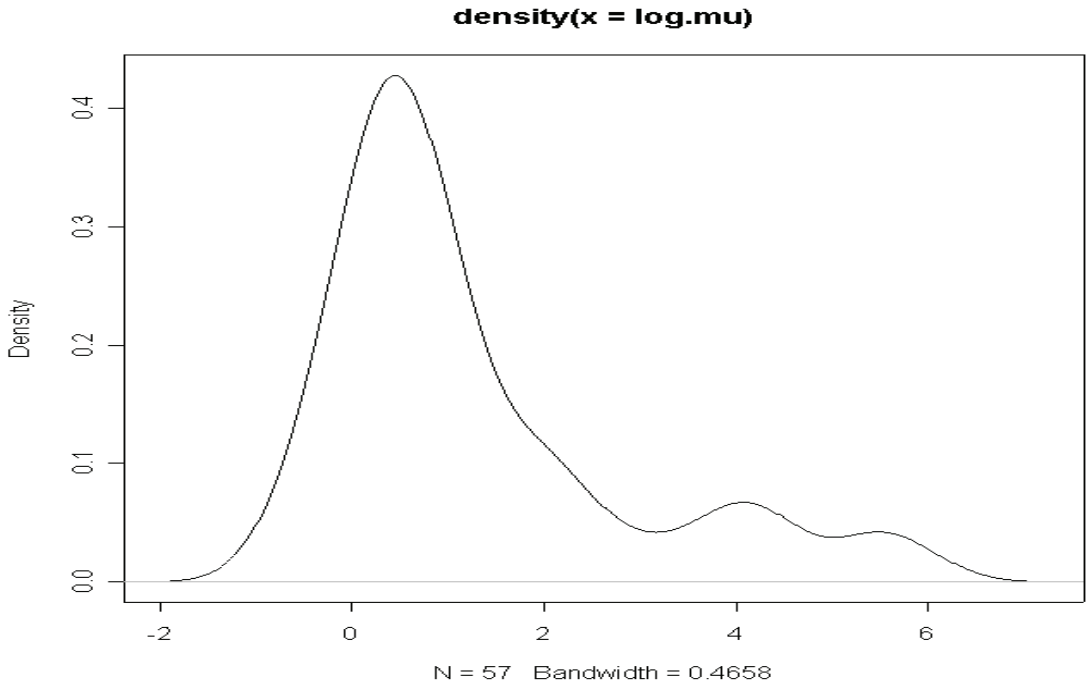


Figure 1b. Empirical distribution of log(μ_i)s

The empirical power curves of the two tests (T_L, T_F) as functions of ν for two nominal test sizes (α = 1%, α = 5%) are presented in Figure 2a-d for sample sizes N = 5, 11, 50, and 100. Figure 2a-d indicate that the powers of the two tests are nonlinear monotonically decreasing functions of ν, which represents the degree of over-dispersion. When the degree of over-dispersion to Poisson is large (i.e., ν is 50 or less), both tests T_L and T_F have fairly high powers for all four samples sizes and for two nominal test sizes, ranging from 65.4% to 100%. The power decreases dramatically when the degree of over-dispersion decreases (i.e., ν increases from 50 to 1000). When ν is as big as 5,000, the test powers are very low, ranging from 0.8 to 21.6 (Fig. 2a–d). Sample size seems to have less influence on powers of the tests than the degree of over-dispersion does. When sample size is increased from small (N = 5), moderate (N =11, 50) to large (N = 100), the corresponding test powers only slightly increase.

With small sample sizes, the ratios between the power of T_L and the power of T_F with the increase of values of ν are relative unstable (TL1 vs. TF1 and TL5 vs. TF5 in Fig. 2a-b). With lager sample size, the power of T_L decreases fast then the power of T_F with the increase of values of ν (TL1 vs. TF1 and TL5 vs. TF5 in Fig. 2c-d), indicating T_L is more sensitive than T_F to the degree of over-dispersion. This is especially true for moderate degree of over-dispersion.

Power analysis under alternative hypothesis H_1: $Y_i \sim$ ZIP

Under the alternative hypothesis $\mathbf{H}_1$: $Y_i \sim$ ZIP (μ_i, ω), the probability density is

$$\Pr(Y=y)=\begin{cases}\omega+(1-\omega)\exp(-\mu) & \\ \\ (1-\omega)\exp(-\mu)\mu^{y}/y! & \end{cases}.$$

The variance-mean ratio is $\frac{\mathrm{var}(Y)}{E(Y)}=1+\frac{\omega E(Y)}{1-\omega}$. Taking logarithms gives,

$$\log(\mathrm{var}(Y)) = \log(E(Y))+\log(1+\omega E(Y)/(1-\omega)) = g(E(Y))$$

In this simulation, the test power is again set up as a function of ω, the dispersion parameter. The empirical power curves of the two tests (T_L, T_F) as functions of ω for two nominal test sizes (α = 1%, α = 5%) are presented in Figure 3a-d for sample sizes N = 5, 11, 50, 100 respectively. Total 99 different values, ω = {0.01, 0.02, …, 0.98, 0.99}, was used while only test powers for ω = {0.01, 0.02, …, 0.49, 0.50} are presented in Figure 3a-d. At each value of ω, the experiment described in 'General design for the simulation' was performed.

Again, it appears that the test power is a nonlinear monotone increasing function of the degree of over-dispersion to Poisson, which is represented here by ω (Smaller values of ω index smaller degree of over-dispersion). And with small sample size, the ratios of powers of the two tests (T_L, T_F) are unstable with the decrease of values of ω (TL1 vs. TF1 and TL5 vs. TF5 in Fig. 3a-b). With lager sample size, the power of T_L decreases fast then the power of T_F with the increase of values of ν (TL1 vs. TF1 and TL5 vs. TF5 in Fig. 3c-d), indicating T_L is more sensitive than T_F to the degree of over-dispersion.

In summary, the simulation experiments demonstrate that among the three evaluated tests, T_S is ruled out due to unacceptable test size; the power characteristic of T_L is empirically superior to T_F in terms of sensitivity to degree of over-dispersion. Thus, only T_L is used in the four states (New York, Connecticut, New Jersey, and Pennsylvania) LD data including the *NYSLD* data.

Table 1. Percentage rejections of H_0: $Y_i \sim$ Poisson (μ_i)

For nominal alpha = 1%

sample size	percentage of reject H0 (95%CI)		
	T_S	T_L	T_F
5	12.4(11.6, 13.2)	1.2(0.4, 2.0)	0.2(0.0, 1.0)
11	12.0(11.2, 12.8)	1.4(0.0, 1.4)	0.8(0.0, 1.6)
50	13.2(12.4, 14.0)	0.6(0.0, 1.4)	0.6(0.0, 1.4)
100	13.6(12.8, 14.4)	0.6(0.0, 1.4)	1.0(0.2, 1.8)

For nominal alpha = 5%

sample size	percentage of reject H0 (95%CI)		
	T_S	T_L	T_F
5	18.4(16.4, 20.4)	6.6(4.6, 8.6)	2.6(0.6, 4.6)
11	20.0(18.0, 22.0)	6.0(2.4, 6.4)	5.0(3.0, 7.0)
50	20.0(18.0, 22.0)	4.4(2.4, 6.4)	5.2(3.2, 7.2)
100	22.2(20.2, 24.2)	4.4(2.4, 6.4)	5.0(3.0, 7.0)

Applications

The new test statistic T_L is applied to the *NYSLD* data, as well as LD data of Connecticut State, New Jersey State, and Pennsylvania State in this section. The reasons that these other three states have been chosen are: 1) They are geographical neighbors to NYS; 2) In these three states, LD was present and incidence rates (per 100,000 population) at county level have been recorded roughly over same period as the *NYSLD* data.

In the following section, descriptions are first given to the LD data for the three 'neighboring' states to NYS. The geographic relations of the three states to NYS are displayed in Figure 4. The time period from which the data for each state were available and the number of counties per state are summarized in Table 2. The results from the tests are also given.

LD data of Connecticut, New Jersey, and Pennsylvania

The time period during which yearly LD counts and incidence rates were available at the county level for all eight counties in Connecticut State was from 1991 to 2002, for all 21 counties in New Jersey State was from 1990 to 2000, and for all 67 counties in Pennsylvania State was from 1990 to 2001 (Table 2).

Test results of LD data for the four states

The relationships between county sample mean, $\overline{Y}$, and its sample variance-mean ratio, $S^2/\overline{Y}$, for New York, Connecticut, New Jersey, and Pennsylvania are displayed in Figure 5a-d. The relationship between log-sample mean, $\log \overline{Y}$, and log-sample variance, $\log S^2$, for the four states are displayed in Figure 6a-d. The results of the test T_L for each of the four states are summarized in Table 3. Note that the T_L test is developed under the assumption that j observations of Y_i are identical independent distributions, which is not the case in the *NYSLD* data. Figure 7 shows the LD incidence curves of each county over the years from 1990 to 2000, with a small map of NYS to indicate geographic locations of the counties. For example,

1. Westchester County's incidence rate, the green curve, was high in 1990, but decreased over time.

2. Putnam County's, the pink curve, was high in 1990, increased from 1990 to 1996,
   and has decreased since then.
3. Dutchess County's, the blue curve, was high in 1990 and kept increasing over time.
4. Columbia County's, the red curve, was very low in 1990 and then gradually increased from 1990 to 1995. It has increased very rapidly since 1996. In 2000, Columbia County had the highest LD incidence rate in the United States (CDC, 2002).
5. Rensselaer County's, the black curve, was low in 1990 and stayed the same till 1998. Then it increased slightly from 1998 to 2000.

The finding above indicated that LD occurrence in some of the NYS counties had strong time trends. To adjust for this violation to the independence assumption, the T_L test is also applied to the partial *NYSLD* data after taking out counties having significant time trends. Figure 8a displays the relationships between county sample mean, $\overline{Y}$, and its sample variance-mean ratio, $S^2/\overline{Y}$, and Figure 8b displays the relationship between log-sample mean, $\log \overline{Y}$, and log-sample variance, $\log S^2$, for the partial *NYSLD* data. The result of the test T_L for it is summarized in the row NY_p of Table 3.

Note that in Figure 5a-d, 6a-d, and 8a-b, the same axis scales are used for plots displaying relationship between $\overline{Y}$ and $S^2/\overline{Y}$, so are for plots displaying relationships between $\log \overline{Y}$ and $\log S^2$, for convenience of comparisons.

The scatter plots in Figure 5a-d and Figure 6a-d show that the relation between $\log S^2$ and $\log \overline{Y}$ has a much clearer linear form than the relation between $S^2/\overline{Y}$ and $\overline{Y}$ for all the four states. In contrast to Figure 5a and Figure 6a for the entire *NYSLD* data, Figure 8a and b show that NYS counties with significant time trends tend to have larger LD counts and sample variances than those without time trends.

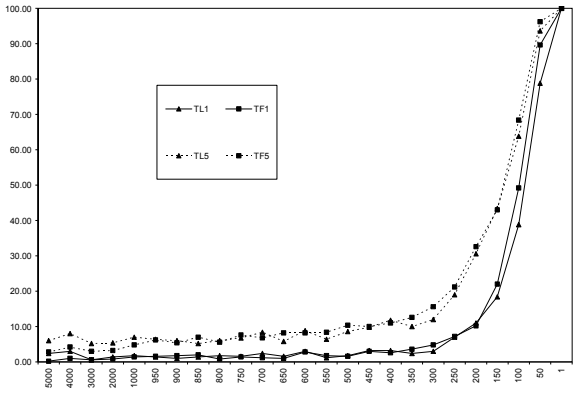
The facts that *p*-values of T_L were close to zero and 95% Confidence Interval (95%CI) of β_1 did not include one for the data from all the four states and for the partial *NYSLD* data give strong evidence of over-dispersion to Poisson. The *p*-value of T_L for Connecticut was less than 0.05 but greater than 0.01. This may be caused by larger variation due to the fewer number of observations. The consistent test results of β_1 for both the entire and the partial *NYSLD* data indicate that the T_L test is robust to time trend in Poisson data. Note that the values of β_1 for New York, Connecticut, and New Jersey are very close to each other (ranging from 1.26 to 1.72).

Conclusion

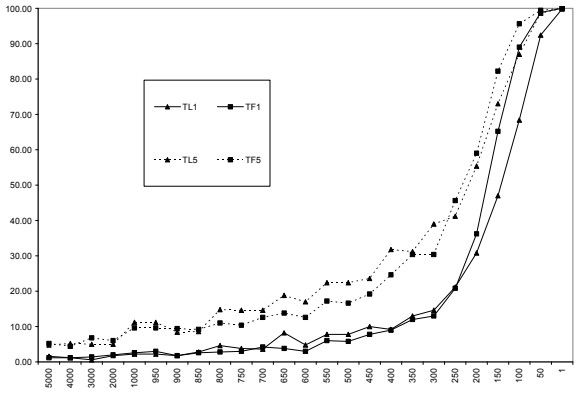
In this paper, two simple, easy to implement tests (T_S and T_L) are proposed for assessment of simultaneous over-dispersion of a group of random variables to the Poisson model. These tests specify a relationship between the mean and variance. However, they do not require specification of the distribution under the alternative. The tests are easy to implement: the T_S is computed as the *t*-test from an WLS regression, and the T_L from an OLS regression. In this sense, the two tests can be given the name 'regression-based' tests.

Simulation experiments implemented in samples of small (5), moderate (11 and 50), and large (100) sizes shows that the empirical test size matches the nominal size well for T_L, but is unacceptably liberal for T_S for all experimented sample sizes, which suggests the log-transformation makes T_L less possible to break assumptions of linear regression. It is noted that the reference test T_F has unfit empirical test size when sample size is small but performs fine with moderate or bigger sample sizes. This may be due to the fact that the Fisher statistic is strictly correct for continuous variables, which becomes more realistic for Poisson as the sample size *N* increases. The power simulation experiments performed on T_L and T_F treat power as a function of dispersion parameter of alternative distributions. The empirical comparisons of power curves suggest that although both tests have adequate power even for small sample size, the power characteristic of T_L empirically superior to T_F in terms of sensitivity to degree of over-dispersion. This is especially true when over-dispersion is moderate.

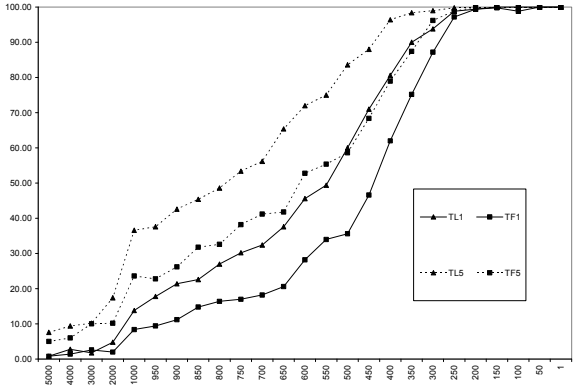
a. sample size = 5



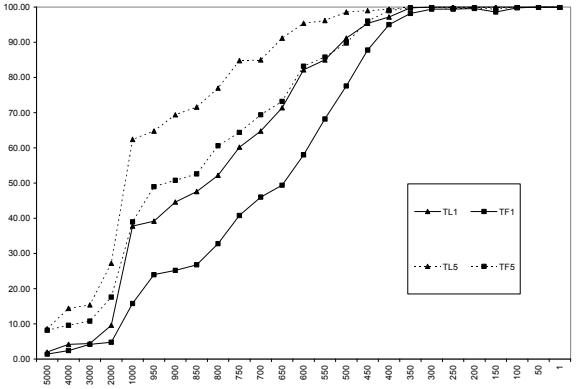
b. sample size = 11



c. sample size = 50



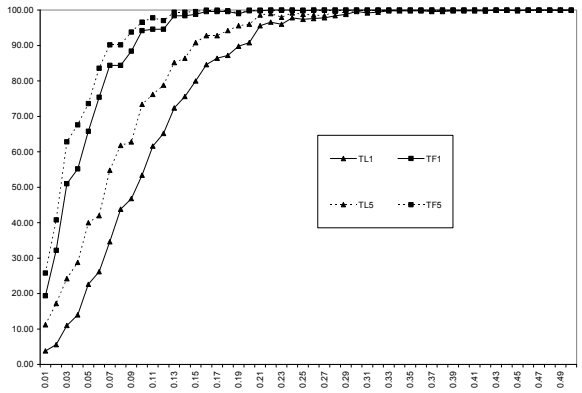
d. sample size = 100



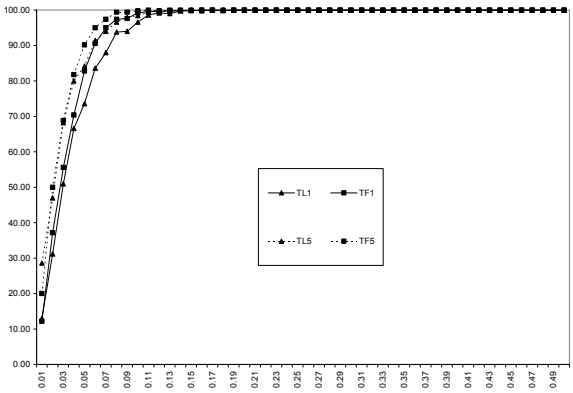
(Notes: 1. x-axis is ν and y-axis is the percentage of reject H_0 for plots a-d.
2. TL1 = T_L for α =0.01; TF1 = T_F for α =0.01; TL5 = T_L for α =0.05; TF5 = T_F for α =0.05.)

Figure 2. The empirical power curve of T_L, and T_F under alternative hypothesis H_1: $Y_i \sim$ NB

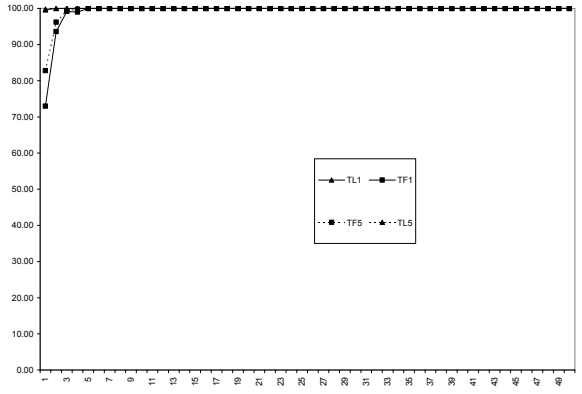
a. sample size = 5



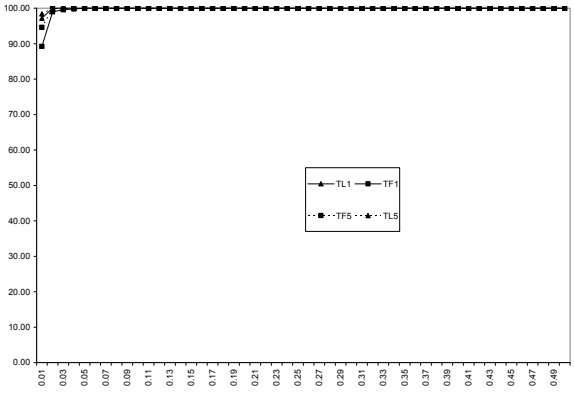
b. sample size = 11



c. sample size = 50



d. sample size = 100



(notes: 1. x-axis is ω and y-axis is the percentage of reject H_0 for plots a-d.
2. TL1 = T_L for α=0.01; TF1 = T_F for α=0.01; TL5 = T_L for α=0.05; TF5 = T_F for α=0.05.)

Figure 3. The empirical power curve of T_L, and T_F under alternative hypothesis H_1: $Y_i \sim$ ZIP

Figure 4. Geographic location of the four states

The most commonly used probability models for discrete data are binomial (Bin), Poisson (Pois), and Negative Binomial (NB) (series 1) and their corresponding zero-inflated models: zero-inflated binomial (ZIBin), zero-inflated Poisson(ZIP), and zero-inflated Negative Binomial (ZINB) (series 2). Each of them has different flexibility to model over-dispersion to Poisson. Table 4 summarizes their variance-mean relationships (VMR) and relative over/under dispersion to Poisson.

Table 4 reveals that the Bin probability model only allows under-dispersion to Poisson, while the ZIP, NB, and ZINB probability models only allow over-dispersion to Poisson. Among these six probability models, ZIBin is the most flexible model. It allows all the three situations (under-dispersion, over-dispersion, and none) based on different relative values of ω and n.

On the other hand, after we assess the over or under-dispersion of a data set using the test T_L, different choices of probability models can be recommended based on different estimated values of β_1 (Table 5). For example, if a test result indicates that the estimated β_1 is statistically significantly greater than one ($\beta_1 > 1$), probability models that allows over-dispersion will be recommended such as ZIBin, ZIP, NB, or ZINB.

Although the motivation and essential theory of these tests exploits only the equality between mean and variance, this approach can be extended to tests of other relationships between mean and variance.

Equal observation points (N) for each variable are assumed in this study. Future research can be done by studying a group of variables with unequal observation points (i.e., I independent variables, each with N_i observation points). In this paper, when the linear regression is applied to mean and variance-mean ratio, a common regression coefficient (β_1) is assumed for the group of variables. In the future research, individual regression coefficient (β_{1i}) can be given to each variable and Bayesian approaches can be used to estimate the parameters of interest.

Applications of the T_L test to the *NYSLD* and the LD data for Connecticut, New Jersey, and Pennsylvania suggest that the Poisson model is not statistically consistent with these count data and a natural alternative is the Negative binomial model. The fact that the values of β_1 for New York, Connecticut, and New Jersey are close to each other (ranging from 1.26 to 1.72)

Table 2. Summary of time periods and locations studied

	NY	CT	NJ	PA
Starting year	1990	1991	1990	1990
Ending year	2000	2002	2000	2001
No. year	11	12	11	12
No.county	57	8	21	67

Table 3. Summary of the tests T_L on LD data for the four states

State	T_L	
	beta_1(95%CI)	*p*-value
NY	1.70 (1.61, 1.79)	0.00
NY_p	1.34 (1.16, 1.51)	0.00
CT	1.72 (1.22, 3.23)	0.03
NJ	1.72 (1.40, 2.05)	0.00
PA	1.26 (1.10, 1.43)	0.00

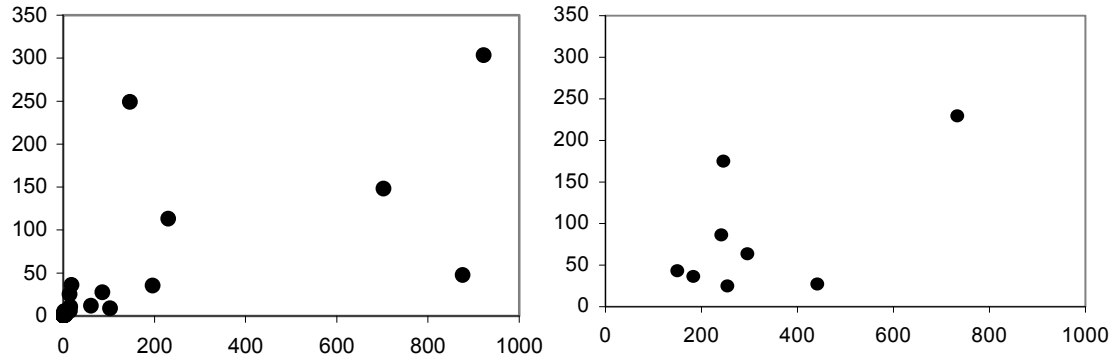
Table 4. Variance-mean relationships of six commonly used probability models for discrete data and their relative over/under-dispersion to Poisson

series 1	Variance-Mean Relation	dispersion
Bin	$\text{var}(Y)/E(Y) = 1 - E(Y)/n$	under
Pois	$\text{var}(Y)/E(Y) = 1$	none
NB	$\text{var}(Y)/E(Y) = 1 + E(Y)/\nu$	over
series 2	**Variance-Mean Relation**	**dispersion**
ZIBin	$\frac{\text{var}(\tilde{Y})}{E(\tilde{Y})} = 1 + \frac{\omega - 1/n}{1-\omega} E(\tilde{Y})$	dependent*
ZIP	$\frac{\text{var}(\tilde{Y})}{E(\tilde{Y})} = 1 + \frac{\omega}{1-\omega} E(\tilde{Y})$	over
ZINB	$\frac{\text{var}(\tilde{Y})}{E(\tilde{Y})} = 1 + \frac{\omega + 1/\nu}{1-\omega} E(\tilde{Y})$	over

* The over or under-dispersion to Poisson is dependent on the relative values between ω and n: if ω > 1/ n, then this model is over-dispersion to Poisson; if ω = 1/ n, then this model is neither over nor under-dispersion to Poisson; if ω > 1/ n, then this model is under-dispersion to Poisson.

a. New York

b. Connecticut



c. New Jersey

d. Pennsylvania

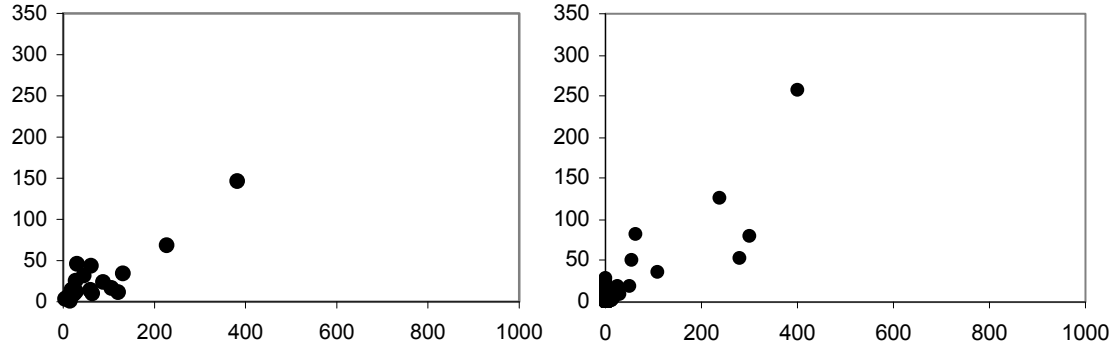
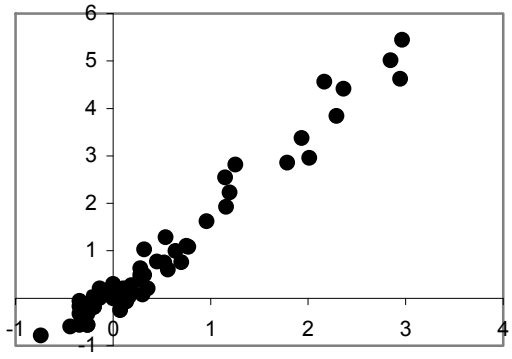
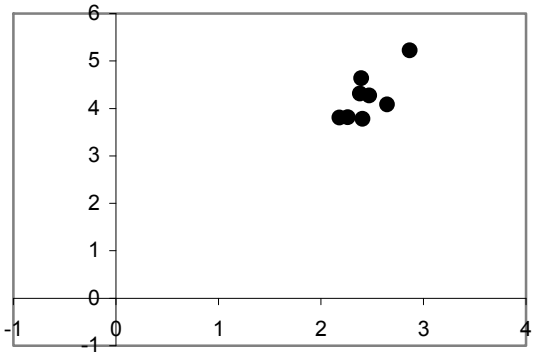


Figure 5. Scatter plots of $S^2 / \overline{Y}$ vs. $\overline{Y}$ for the four states
(note: x-axis is $\overline{Y}$ and y-axis is $S^2 / \overline{Y}$ for plots a-d.)

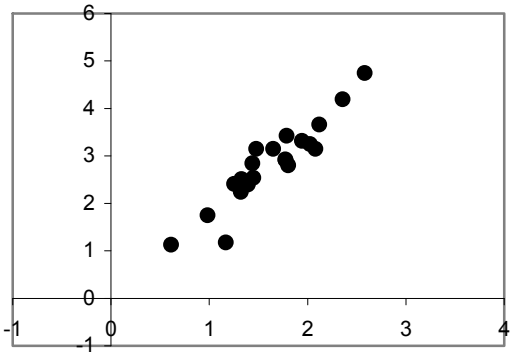
a. New York



b. Connecticut



c. New Jersey



d. Pennsylvania

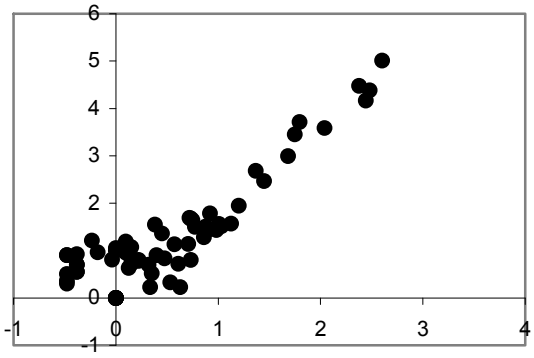


Figure 6. Scatter plots of $\log S^2$ vs. $\log \overline{Y}$ for the four states
(note: x-axis is $\log \overline{Y}$ and y-axis is $\log S^2$ for plots a-d.)

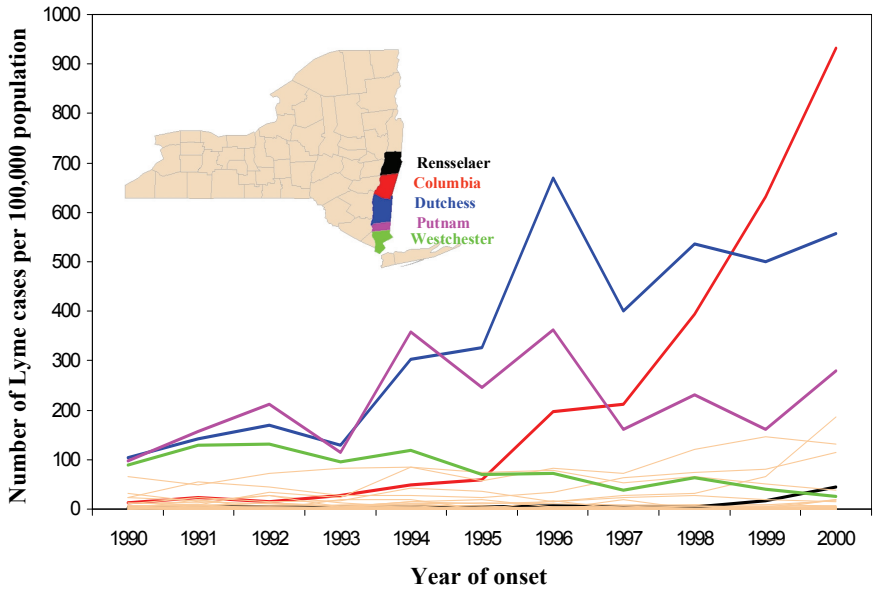


Figure 7. Fifty-seven NYS county annual incidence rates from 1990 to 2000

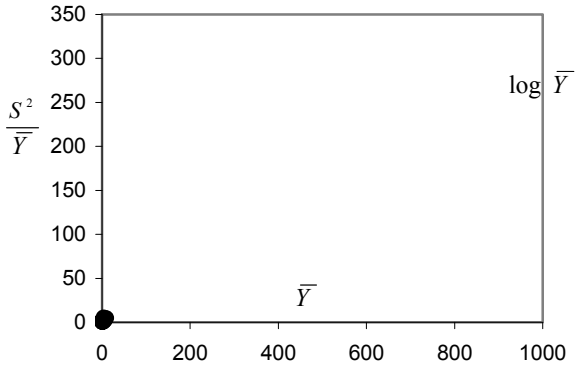


Figure 8a. Scatter plots of $S^2 / \overline{Y}$ vs. $\overline{Y}$ for the partial NYS

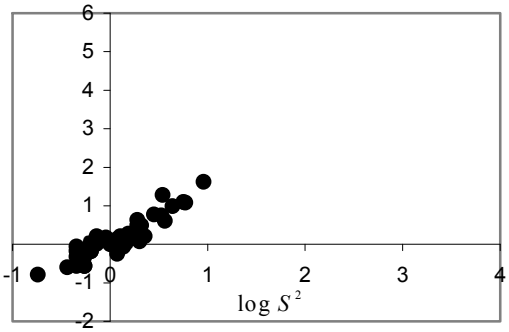


Figure 8b. Scatter plots of $\log S^2$ vs. $\log \overline{Y}$ for the partial NYS

Table 5. Recommended probability models based on estimated coefficient for regression of log-sample variance on log-sample mean

est. Beta1	suggested models					
	Bin	ZIBin	Pois	ZIP	NB	ZINB
> 1		+		+	+	+
1		+	+			
< 1	+	+				

may suggest a general pattern of LD existing in the studied geographic area.

Results from the *NYSLD* data suggested that the new test statistic T_L seems robust to data with time trend in Poisson model. This is probably related to the fact that sums involved in the averages of individual Poissons are also Poissons. However, more systematic studies are needed before making any determinant conclusions.

References

Bohning, D. (1994). A note on a test for Poisson overdispersion. *Biometrika 81(2)*, 418-9.

Brown, L. D., & Zhao, L. H. (2002). A test for the Poisson distribution. Sankhya: *The Indian Journal of Statistics 64(A3)*, 611-25.

Cameron, A. C., & Trivedi, P. K. (1990). Regression based tests for overdispersion in the Poisson model. *Journal of Econometrics 46*, 347-64.

Centers for Disease Control and Prevention (CDC). (1997). Cases definitions for infectious conditions under public health surveillance. *Morbidity and Mortality Weekly Report 46(RR-10)*, 1-55.

Centers for Disease Control and Prevention (CDC) (2002). Summary of notifiable diseases—United States, 2000. *Morbidity and Mortality Weekly Report 49(53)*, 1-16. Connecticut state department of health website (http://www.dph.state.ct.us), downloaded date: 06-11-04.

Hedges, L. V., & Olkin, I. (1995). *Statistical Methods for Meta-Analysis*. Academic Press: Orlando, FL.

Hinde, J., & Demetrio, C. (1998). Overdispersion: models and estimation. *Computational Statistics and Data Analysis 27*, 151-170.

Hoel, P. G. (1943) On indices of dispersion. *Ann Math Statist 14*, 155-62.

Johnson, N.L., Kotz, S., & Kemp A.W. (1992). *Univariate Discrete Distributions*, second edition. John Wiley & Sons Inc.: New York.

Lambert, D., & Roeder, K. (1995). Overdispersion diagnostics for generalized linear models. *Journal of American Statistical Association 90*, 1225-36.

Lindsey, J. K. (1995). *Modeling frequency and count data*. Clarendon Press: Oxford. New Jersey Department of Health and Senior Services website (*http://www.state.nj.us/health*), downloaded date: 06-11-04.

Pennsylvania Department of Health website (*http://www.dsf.health.state.pa.us*), downloaded date: 06-11-04.

Smyth, G. K. (2003). Pearson's goodness of fit statistic as a score test statistic. In: *Science and Statistics: A Festschrift for Terry Speed*. DR Goldstein (ed.). IMS Lecture Notes–Monograph Series, Volume 40, Institute of Mathematical Statistics, Beachwood, Ohio, pp115-26.

Tiago de Oliveria, J. (1965). Some elementary tests of mixtures of discrete distributions. In: *Classical and Contagious Discrete Distributions*. GP Patil (ed.). Pergamon Press: New York, pp379-84.

White, D. J., Chang, H-G., Benach, J. L., et al. (1991). Geographic spread and temporal increase of the Lyme disease epidemic. *JAMA 266*, 1230-6.