Spinned Poisson distribution with health management application

Ramalingam Shanmugam

Received: 30 November 2010 / Accepted: 28 March 2011 / Published online: 26 April 2011
© Springer Science+Business Media, LLC 2011

Abstract Consider a data collection setup during a spread of an infectious disease. Examples include severe acute respiratory syndrome (SARS) or influenza A virus H3N2. The health management in such scenarios quickly removes the infected cases before the data collection is completed. The estimate of the incidence rate or the chance for anyone to be immune requires using a correct probability distribution. The usual Poisson distribution is inappropriate because of the impact of removing infected cases on the incidence rate as well as observing a new case. An appropriate new probability distribution is derived and is named as *spinned Poisson* distribution. The statistical properties of the spinned Poisson distribution are developed and illustrated using infectious smallpox incidence in Abakaliki, Nigeria.

Keywords Health management count distributions · Likelihood ratio test · Non-central chi-squared distribution · Sampling bias

1 Motivation

Let Y be a *random number* of cases in an observable set $\mathbb{S}$ of non-negative integers with unknown parameters $\Theta \in \{\theta, \rho\}$ in a Poisson type chance mechanism where the infected cases are quickly removed by the health management even before the completion of data. The analysis of data has to be carefully done with a selection of an appropriate underlying probability distribution. The usual Poisson distribution is inappropriate to use because of the impact of removing

infected cases on the incidence rate and the chance of observing a new case. The parameters $\theta > 0$ and $\rho \geq 0$ portray respectively the *incidence rate* and the impact level of *removing infected cases*. Consequently, both the random observation count Y and the incidence rate, θ become size/length biased. Published articles in the statistical literature deal with only the size/length bias on observation, Y in a marginal sense excluding the impact on the parameter (see pages 149–150 and the reference section of Johnson et al. [3] for a full list of published articles on size/length biased sampling). There is no article in the literature discussing the sampling bias on the incidence rate, θ . The sampling bias on both the observation and the incidence rate should be dealt with simultaneously to be realistic in modeling the situation in which the infected cases are removed.

For this purpose, assume that the observation Y is proportional to a *sampling weight factor* $w(y, \rho, \theta) = \frac{(1+\rho y)}{(1+\rho\theta)}$. An insight for this weight factor arises from the following reasons. The observation y is impacted by the removal bias $\rho \geq 0$ due to the removal. A scale shift on one in the weight factor is necessary to avoid degeneration of probabilities to zero when $\rho=0$. The denominator $(1+\rho\theta)$ in the weight factor echoes the impact of the removal bias on the incidence parameter θ and it is a necessity for the expression in (1) to be a bona fide probability distribution.

In this weighted Poisson sampling framework, when $\rho=0$, it nullifies the impact of removing cases. When $\rho=1$, it is called *size/length biased sampling*. That is found in count distribution literature. Since the seminal work of Cox [2], many articles have appeared about size/length biased sampling. These articles considered only the bias in the observation, Y , but not the bias in the parameter. This article is the first one to suggest that the selection bias could influence the parameter(s) as much as the observation.

R. Shanmugam (✉)
School of Health Administration, Texas State University,
San Marcos, TX 78666, USA
e-mail: rs15@txstate.edu

A new Poisson distribution in (1) incorporates the chance mechanism of rare events in which the infected cases are removed.

$$p(y, |\rho, \theta) = \Pr[Y = y] = \frac{(1 + \rho y)}{(1 + \rho \theta)} e^{-\theta} \theta^y / y! \quad (1)$$

$$y = 0, 1, 2, \dots, \infty; 0 < \theta < \infty; 0 \leq \rho < \infty$$

The probability distribution in (1) is named the *spinned Poisson distribution* (SPD). Is expression in (1) a bona fide probability distribution? The answer is affirmative. The following arguments prove it.

For the specified parameter space $0 < \theta < \infty$; $0 \leq \rho < \infty$ and the sample space $y=0,1,2,\dots,\infty$, the right side of the expression (1) has clearly non-negative valued functions. In addition, the sum of their non-negative values is equal to one and it is proved below.

$$\begin{aligned} \sum_{y=0}^{\infty} p(y, |\rho, \theta) &= \frac{1}{(1 + \rho \theta)} \sum_{y=0}^{\infty} (1 + \rho y) e^{-\theta} \theta^y / y! \\ &= \left(1 + \rho \theta \sum_{z=y-1=0}^{\infty} e^{-\theta} \theta^z / z! \right) \\ &= \frac{1}{(1 + \rho \theta)} (1 + \rho \theta) \\ &= 1 \end{aligned}$$

The SPD is versatile enough to describe data from engineering, finance and economics in which the removal of cases occurs before completing the data collection. For example, in finance studies, a defaulted homeowner may be removed from a potential home mortgage insurance list.

The SPD has several interesting statistical properties as explained below. The estimation of its parameters, assessment of its survival function, and the application of the hypothesis testing procedures for the SPD are of interest. Results are demonstrated, using Bailey's [1] data on Nigeria's smallpox incidences in Section 3. The Table 1 displays data. The Fig. 1 displays the relationship between $1/(1+\rho\theta Y)$ in terms of Y to capture their correlation (= the linear relationship). The Fig. 2 displays the survival function of the fitted usual Poisson and the fitted spinned Poisson distributions along with the empirical survival function of the data in Table 1. The Fig. 2 helps to compare the goodness of fit test results.

Table 1 Y = the number of days waited to remove a smallpox case with the frequency f

$Y \rightarrow$	0	1	2	3	4	≥ 5	N	$\bar{Y}$	S_Y^2
$f \rightarrow$	7	6	4	4	1	7	19	2.14	3.69

Fig. 1 Scatterplot of $\frac{1}{1+\rho Y}$ versus Y

2 Main results: spinned Poisson distribution

In this section, several statistical properties are derived and interpreted. The expected value $\mu = E(Y)$ is obtained in a straightforward manner as stated below. That is,

$$\begin{aligned} \mu = E(Y) &= \sum_{y=0}^{\infty} y \frac{(1 + \rho y)}{(1 + \rho \theta)} e^{-\theta} \theta^y / y! \\ &= \frac{\theta[1 + \rho + \rho \theta]}{1 + \rho \theta} = \theta \left[1 + \frac{\rho}{1 + \rho \theta} \right] \quad (2) \\ &= \theta + 1 - \frac{1}{1 + \rho \theta} \approx \theta + 1 \end{aligned}$$

neglecting the fraction $\frac{1}{1+\rho\theta} < 1$. When the impact of removing infected cases is negligible (that is, $\rho=0$), the mean in (2) reduces to just the expected value, θ of the usual Poisson distribution. From (2), note that

$$\rho = \left(\frac{\mu}{\theta} - 1 \right) / (1 - [\mu - \theta])$$

Fig. 2 Empirical, fitted Poisson and spinned Poisson survival functions

and it attests to a property that there are triangular relationships among the mean μ , parameters θ and ρ . In a similar manner the variance

$$\sigma^2 = E(Y[Y-1]) - \mu(\mu-1)$$

is implicitly obtained as in

$$\begin{aligned} E(Y[Y-1]) &= \sigma^2 + \mu(\mu-1) \\ &= \begin{cases} \theta^2 & \rho = 0 \\ \theta \left[\mu + 1 - \frac{1}{1+\rho\theta} \right], & \text{if } \rho \neq 0. \end{cases} \end{aligned} \quad (3.a)$$

The expression (3.a) is the second factorial moment. The expression (3.a) could be rewritten equivalently, for $\rho \neq 0$,

$$\begin{aligned} \sigma^2 &= \text{var}(Y) \\ &= \theta \left[\mu + \frac{\rho\theta}{1+\rho\theta} \right] - \mu(\mu-1) \\ &= \mu[\theta - (\mu-1)] + \frac{\rho\theta^2}{1+\rho\theta} \\ &= \frac{\mu(\frac{\mu}{\theta} - 1)}{\rho} + \frac{\rho\theta^2}{1+\rho\theta} \\ &= (\mu - \theta) \left[\frac{\mu}{\rho\theta} + \theta \right] \end{aligned} \quad (3.b)$$

The expression (3.b) illustrates a mean-variance relationship like the one for the usual Poisson distribution in which the mean and variance are equal (that is, $\sigma^2 = \mu$).

When there is negligible impact of removing the infected cases (that is, $\rho=0$), the result in (3.a) and (3.b) reduce respectively to the second factorial moment and variance of the usual Poisson distribution. Otherwise, there is an intrinsic relationship among the incidence rate, impact of removing infected cases and the expected number of infected cases. That intrinsic relationship is $\rho = \frac{\mu(\mu-\theta)}{\theta[\sigma^2 - \theta(\mu-\theta)]}$. This result confirms that $\rho \rightarrow 0$ when $\mu \rightarrow \theta$, a property of the usual Poisson distribution.

From (1) note that $p(Y=0, |\rho, \theta) = \frac{e^{-\theta}}{(1+\rho\theta)}$ is the threshold probability for the event $Y=0$. However, a recurrence relationship among the spinned Poisson probabilities exists and it is

$$\begin{aligned} (y+1)(1+\rho y)p(y+1, |\rho, \theta) &= \theta(1+\rho[y+1])p(y, |\rho, \theta) \end{aligned} \quad (4)$$

The recurrence relationship in (4) of the SPD implies that

$$\begin{aligned} p(Y=1, |\rho, \theta) &= \theta(1+\rho)p(Y=0, |\rho, \theta) \\ p(Y=2, |\rho, \theta) &= \frac{1}{2(1+\rho)}\theta(1+2\rho)p(Y=1, |\rho, \theta), \\ p(Y=3, |\rho, \theta) &= \frac{1}{3(1+2\rho)}\theta(1+3\rho)p(Y=2, |\rho, \theta) \end{aligned}$$

and so on. These relationships point out that the data-collection mechanism that includes the removal of infected cases indeed imposes a sampling bias on both the observation and the incidence rate.

When the impact of removing infected cases turns out to be negligible (that is, $\rho \rightarrow 0$), the SPD in (1) reduces to the usual Poisson distribution

$$p(y, |\theta) = \Pr[Y=y] = e^{-\theta}\theta^y/y!.$$

See Johnson et al. [3] for properties of the usual Poisson distribution. Now, other properties of the SPD are derived.

First, note incidentally that the statistic $\frac{1}{1+\rho Y}$ is an unbiased estimate of the parametric function $\frac{1}{1+\rho\theta}$ because

$$\begin{aligned} E\left(\frac{1}{1+\rho Y}\right) &= \sum_{y=0}^{\infty} \left(\frac{1}{1+\rho y}\right) \frac{(1+\rho y)}{(1+\rho\theta)} e^{-\theta}\theta^y/y! \\ &= \frac{1}{(1+\rho\theta)} \end{aligned}$$

Furthermore, the variance of the statistic $\frac{1}{1+\rho Y}$ is obtained using a formula (see Stuart and Ord [4] for details) that

$$\text{Var}\left(\frac{U}{V}\right) = \frac{\mu_u^2}{\mu_v^2} \left(\frac{\sigma_u^2}{\mu_u^2} + \frac{\sigma_v^2}{\mu_v^2} - \frac{2\text{cov}[U, V]}{\mu_u\mu_v} \right) \quad (5)$$

where μ_j^2 and σ_j^2 denote the mean and variance respectively of $j = U$ or V . In this set up, note that $U=1$ and $V=1+\rho Y$. Hence, the variance of the statistic $\frac{1}{1+\rho Y}$ is $\text{Var}\left(\frac{1}{1+\rho Y}\right) = \frac{\rho^2\sigma^2}{(1+\rho\mu)^4}$. Also, note that

$$\begin{aligned} E\left(\frac{Y}{1+\rho Y}\right) &= \sum_{y=0}^{\infty} \left(\frac{y}{1+\rho y}\right) \frac{(1+\rho y)}{(1+\rho\theta)} e^{-\theta}\theta^y/y! \\ &= \frac{\theta}{1+\rho\theta} \end{aligned}$$

Consequently, an expression for computing the correlation coefficient between two sample statistics Y and $\frac{1}{1+\rho Y}$ is obtained and utilized to confirm whether the SPD matches well the given data in Section 3. The correlation coefficient $\text{corr}\left(Y, \frac{1}{1+\rho Y}\right)$ is zero when $\rho=0$. Otherwise (that is for $\rho \neq 0$),

$$\begin{aligned} & \text{corr}\left(Y, \frac{1}{1+\rho Y}\right) \\ &= \frac{E\left(\frac{Y}{1+\rho Y}\right) - E(Y)E\left(\frac{1}{1+\rho Y}\right)}{\sqrt{\text{Var}(Y)\text{Var}\left(\frac{1}{1+\rho Y}\right)}} \\ &= (1+\rho\mu)^2(\mu-\theta-1) \\ &= -\frac{(1+\rho\mu)^2}{(1+\rho\theta)} \end{aligned} \quad (6)$$

using (2) and (3.b) and the plot of the data Y versus a statistic $\frac{1}{1+\rho Y}$ exhibits a trend between them. Their linear relationship is captured by the correlation coefficient. The plot and the correlation coefficient of the collected data reveal how much a deviation might have occurred from the usual Poisson distribution due to removing the infected cases before the completion of data collection.

The maximum likelihood estimate (MLE) of the correlation coefficient is

$$\widehat{\text{corr}}\left(Y, \frac{1}{1+\rho Y}\right) \approx -\frac{\widehat{\theta}_{mle}}{s^2} \left(\frac{1+\widehat{\rho}_{mle}\bar{y}}{1+\widehat{\rho}_{mle}\widehat{\theta}_{mle}} \right)^2 \quad (7)$$

using the MLE of the dispersion σ^2 and the mean μ .

Consider a random sample $y_1, y_2, \dots, y_n$ from the SPD in Eq. 1. The MLEs are asymptotically efficient. The log likelihood function is $\ln L = \sum_{i=1}^n \ln p(y_i, |\rho, \theta)$. The first derivatives of the log likelihood function with respect to the parameters θ and ρ give their score functions. Equating the score functions $\partial_\theta \ln L$ and $\partial_\rho \ln L$ to zero gives the MLEs. The exact value of the MLEs $\widehat{\theta}_{mle}$ and $\widehat{\rho}_{mle}$ could be obtained by solving iteratively the equations

$$\bar{y} = \widehat{\theta}_{mle} \left[1 + \widehat{\rho}_{mle} \left(1 + \widehat{\rho}_{mle} \widehat{\theta}_{mle} \right)^{-1} \right] \quad (8)$$

and

$$\Phi(\widehat{\rho}_{mle}) = \sum_{i=1}^n \frac{y_i - \widehat{\theta}_{mle}}{1 + \widehat{\rho}_{mle} y_i} = 0. \quad (9)$$

The process is tedious for the exact values. The process can be eased if approximate values are acceptable. For this

purpose, the expression (9) is approximated using Taylor's series approximation. That is,

$$\begin{aligned} & \Phi(\widehat{\rho}_{mle}) \\ & \approx \Phi(\widehat{\rho}_{mle} = 0) + \widehat{\rho}_{mle} \partial_{\widehat{\rho}_{mle}=0} \Phi(\widehat{\rho}_{mle}) \\ & = n - \widehat{\rho}_{mle} \sum_{i=1}^n (y_i - \widehat{\theta}_{mle}) y_i \end{aligned}$$

where the notation $\partial_{\widehat{\rho}_{mle}=0} \Phi(\widehat{\rho}_{mle})$ portrays the derivative of $\Phi(\widehat{\rho}_{mle})$ with respect to $\widehat{\rho}_{mle}$ evaluated at $\widehat{\rho}_{mle} = 0$ and $\widehat{\theta}_{mle}$ is the MLE of the incidence rate. Hence, the approximate values of the MLEs are

$$\widehat{\theta}_{mle} \approx \bar{y} - 1 \quad (10)$$

and

$$\widehat{\rho}_{mle} \approx \left[\left(1 - \frac{1}{n} \right) s^2 - \bar{y} (\widehat{\theta}_{mle} - \bar{y}) \right]^{-1} \quad (11)$$

where $\bar{y}$ is the sample mean and s is the sample standard deviation.

Next, an expression for the survival function of the SPD is derived. The cumulative distribution function of the usual Poisson distribution is expressed in terms of the cumulative chi-squared distribution function in Johnson et al. [3], which is tabulated extensively in books. Following this line of thinking, a similar relationship between the cumulative distribution functions (CDF) of the SPD and the chi-squared distribution can be established. For this purpose, let $CDF(m, \theta, \rho) = \Pr[Y \leq m]$ be CDF of the SPD for a pre-specified m in the sample space. Implicitly, the survival function $Sf(m+1, \theta, \rho) = 1 - CDF(m, \theta, \rho)$ of the spinned Poisson distribution is, after algebraic simplifications,

$$\begin{aligned} Sf(m+1, \theta, \rho) &= \Pr[Y \geq m+1] = 1 - \Pr[Y \leq m] \\ &= 1 - \frac{1}{(1+\rho\theta)} \sum_{y=0}^m (1+\rho y) e^{-\theta} \theta^y / y! \\ &= 1 - \frac{1}{(1+\rho\theta)} \left[\sum_{y=0}^m e^{-\theta} \theta^y / y! + \rho\theta \sum_{z=y-1=0}^{m-1} z e^{-\theta} \theta^z / z! \right] \\ &= 1 - \frac{1}{(1+\rho\theta)} \left[\Pr(\chi_{2(m+1),df}^2 > 2\theta) + \rho\theta \Pr(\chi_{2m,df}^2 > 2\theta) \right] \\ &= \frac{[CDF_{\chi_{2(m+1),df}^2}(2\theta) + \rho\theta CDF_{\chi_{2m,df}^2}(2\theta)]}{(1+\rho\theta)} \end{aligned} \quad (12)$$

in which $CDF_{\chi_{2m,df}^2}(2\theta)$ denotes the cumulative chi-squared distribution function with $2m$ degrees of freedom (df) and the percentile (2θ) .

An *immediate removal* means that the number of days to remove a next infected case is $Y=0$. The probability for an immediate removal to occur, with $m=0$, is

$$\begin{aligned} \Pr[Y=0] &= 1 - Sf[1, \theta, \rho] \\ &= CDF_{\chi^2_{2,df}}(2\theta)/(1 + \rho\theta). \end{aligned} \quad (13)$$

The scenario, $\rho=0$ is indicative of no impact of the removing the infected cases. Substituting $\rho=0$ in (12) and (13), they reduce to the results for the usual Poisson distribution. On the contrary, the scenario, $\rho=1$ is indicative of the presence of impact on the observation count Y and on the incidence rate θ due to removing the infected cases. Otherwise (that is when $\rho>0$), there is a *significant impact* on the observation and on the incidence rate due to removing infected cases. A hypothesis testing procedure about the impact parameter ρ is therefore worthwhile.

The likelihood ratio concept can be used to test the null hypothesis $H_0: \rho = \rho_o$ where ρ_o is zero against a research/alternative hypothesis $H_1: \rho = \rho^* \neq \rho_o$. Note that the likelihood ratio under the null hypothesis is

$$A_{\rho_o} = \frac{L(y_1, y_2, \dots, y_n; \hat{\theta}_{\rho_o=0}, \rho_o = 0)}{L(y_1, y_2, \dots, y_n; \hat{\theta}_{\rho_{mle}}, \hat{\rho}_{mle} \neq 0)}$$

and is

$$A_{\rho^*} = \frac{L(y_1, y_2, \dots, y_n; \hat{\theta}_{\rho^*}, \rho = \rho^*)}{L(y_1, y_2, \dots, y_n; \hat{\theta}_{\rho_{mle}}, \hat{\rho}_{mle} \neq 0)}$$

under a research/alternative hypothesis. The MLE of the incidence rate is $\hat{\theta}_{mle, \rho_o=0} = \bar{y}$ under the null hypothesis and $\hat{\theta}_{mle, \rho=\rho^*} \approx \bar{y} - 1$ under the alternative/research hypothesis. The MLE of the impact parameter is $\hat{\rho}_{mle} = \left[\left(1 - \frac{1}{n}\right)s^2 - \bar{y}(\hat{\theta}_{mle} - \bar{y}) \right]^{-1}$ under the alternative/research hypothesis. Under the null hypothesis, the minus log likelihood ratio is

$$\begin{aligned} -\ln A_{\rho_o} &= -n \left[\ln \left(1 + \frac{\hat{\rho}_{mle}\bar{y}}{(1 + \hat{\rho}_{mle})} \right) - \frac{\bar{y}}{(1 + \hat{\rho}_{mle})} \right. \\ &\quad \left. + \bar{y} \ln \left(\frac{\hat{\rho}_{mle}}{1 + \hat{\rho}_{mle}} \right) \right] + \sum_{i=1}^{i=n} \ln(1 + \hat{\rho}_{mle}y_i) \end{aligned} \quad (14)$$

and it follows a non-central chi-squared distribution with one df and the non-centrality parameter

$$\hat{\delta}_{\rho_o} = \frac{(\hat{\rho}_{mle} - \rho_o)}{\text{var}(\hat{\rho}_{mle})}. \quad (15)$$

See Wald [5] for definition and properties of the non-central chi squared distribution. A unique property of the MLE is that the MLE of a function is the function of the MLE of the parameters. In addition, the variance-covariance matrix of the MLE of the parameters is the inverse of the information matrix

$$\begin{aligned} I &= \begin{bmatrix} -E(\partial_{\theta\theta}^2 \ln L) & -E(\partial_{\rho\theta}^2 \ln L) \\ -E(\partial_{\theta\rho}^2 \ln L) & -E(\partial_{\rho\rho}^2 \ln L) \end{bmatrix} \\ &= \begin{bmatrix} \frac{\mu}{\theta^2} - \left(\frac{\rho}{1 + \rho\theta}\right)^2 & \left(\frac{1}{1 + \rho\theta}\right)^2 \\ \left(\frac{1}{1 + \rho\theta}\right)^2 & \frac{\sigma^2}{(1 + \rho\mu)^4} \end{bmatrix} \end{aligned} \quad (16)$$

using the result

$$\begin{aligned} E\left(\frac{Y}{1 + \rho Y}\right)^2 &= \text{var}\left(\frac{Y}{1 + \rho Y}\right) + \left[E\left(\frac{Y}{1 + \rho Y}\right)\right]^2 \\ &= \frac{\sigma^2}{(1 + \rho\mu)^4} + \left(\frac{\theta}{1 + \rho\theta}\right)^2 \end{aligned}$$

and (2). The determinant of the information matrix in (16) is

$$D = \frac{\sigma^2}{(1 + \rho\mu)^4} \left[\frac{\mu}{\theta^2} - \left(\frac{\rho}{1 + \rho\theta}\right)^2 \right] - \left(\frac{1}{1 + \rho\theta}\right)^4$$

whose MLE is

$$\begin{aligned} \hat{D}_{mle} &= \frac{\bar{y} - \left(\frac{\hat{\rho}_{mle}\bar{y}}{1 + \hat{\rho}_{mle} + \hat{\rho}_{mle}\bar{y}}\right)^2}{(1 + \hat{\rho}_{mle}\bar{y})^4} \left[\frac{(1 + \hat{\rho}_{mle})^2}{\bar{y}} \right. \\ &\quad \left. - \left(\frac{\hat{\rho}_{mle}(1 + \hat{\rho}_{mle})}{1 + \hat{\rho}_{mle} + \hat{\rho}_{mle}\bar{y}}\right)^2 \right] \\ &\quad - \left(\frac{1 + \hat{\rho}_{mle}}{1 + \hat{\rho}_{mle} + \hat{\rho}_{mle}\bar{y}}\right)^4 \end{aligned}$$

using (3.5) and (10). The variance-covariance matrix of the MLE of the parameters is

$$I^{-1} = \begin{bmatrix} \text{var}(\hat{\theta}_{mle}) & \text{cov}(\hat{\theta}_{mle}, \hat{\rho}_{mle}) \\ \text{cov}(\hat{\theta}_{mle}, \hat{\rho}_{mle}) & \text{var}(\hat{\rho}_{mle}) \end{bmatrix}$$

$$= \frac{1}{\hat{D}_{mle}} \begin{bmatrix} \bar{y} - \left(\frac{\hat{\rho}_{mle}\bar{y}}{1 + \hat{\rho}_{mle} + \hat{\rho}_{mle}\bar{y}} \right)^2 & - \left(\frac{1 + \hat{\rho}_{mle}}{1 + \hat{\rho}_{mle} + \hat{\rho}_{mle}\bar{y}} \right)^2 \\ - \left(\frac{1 + \hat{\rho}_{mle}}{1 + \hat{\rho}_{mle} + \hat{\rho}_{mle}\bar{y}} \right)^2 & \frac{(1 + \hat{\rho}_{mle})^2}{\bar{y}} - \left(\frac{1 + \hat{\rho}_{mle}}{1 + \hat{\rho}_{mle} + \hat{\rho}_{mle}\bar{y}} \right)^2 \end{bmatrix}$$

Note that

$$\hat{\delta}_{\rho_o} = \frac{(\hat{\rho}_{mle} - \rho_o)}{\text{var}(\hat{\rho}_{mle})}$$

$$\hat{\rho}_{mle} \left\{ \frac{\bar{y} - \left(\frac{\hat{\rho}_{mle}\bar{y}}{1 + \hat{\rho}_{mle} + \hat{\rho}_{mle}\bar{y}} \right)^2}{(1 + \hat{\rho}_{mle}\bar{y})^4} \left[\frac{(1 + \hat{\rho}_{mle})^2}{\bar{y}} - \left(\frac{\hat{\rho}_{mle}(1 + \hat{\rho}_{mle})}{1 + \hat{\rho}_{mle} + \hat{\rho}_{mle}\bar{y}} \right)^2 \right] - \left(\frac{1 + \hat{\rho}_{mle}}{1 + \hat{\rho}_{mle} + \hat{\rho}_{mle}\bar{y}} \right)^4 \right\}$$

$$= \frac{\left[\frac{(1 + \hat{\rho}_{mle})^2}{\bar{y}} - \left(\frac{1 + \hat{\rho}_{mle}}{1 + \hat{\rho}_{mle} + \hat{\rho}_{mle}\bar{y}} \right)^2 \right]}{\left[\frac{(1 + \hat{\rho}_{mle})^2}{\bar{y}} - \left(\frac{1 + \hat{\rho}_{mle}}{1 + \hat{\rho}_{mle} + \hat{\rho}_{mle}\bar{y}} \right)^2 \right]}$$

with the MLE of $\hat{\theta}_{mle}$ and $\hat{\rho}_{mle}$. In addition, the non-central chi squared distribution with one df and non-centrality parameter δ approximately follows $\left(1 + \frac{\delta}{1 + \delta}\right)$ times a central chi squared distribution with $\frac{(1 + \delta)^2}{(1 + 2\delta)}$ df (see Stuart and Ord [4] for details of this equivalence). This means that the null hypothesis $H_0: \rho = \rho_o$ where ρ_o is zero will be rejected in favor of the research/alternative hypothesis $H_1: \rho \neq \rho_o$ if

$$- \ln A_{\rho_o=0} > \left(1 + \frac{\hat{\delta}_{\rho_o}}{1 + \hat{\delta}_{\rho_o}}\right) \chi^2_{\frac{(1 + \hat{\delta}_{\rho_o})^2}{(1 + 2\hat{\delta}_{\rho_o})}, \alpha} \quad (18)$$

where the right side is the critical value based on the 100 $(1 - \alpha)^{th}$ percentile of the central chi squared distribution with $\frac{(1 + \hat{\delta}_{\rho_o})^2}{(1 + 2\hat{\delta}_{\rho_o})}$ df and a significance level $\alpha \in (0, 1)$.

The statistical power of the test statistic in (14) can be examined with a selection of a specific value for ρ in the research/alternative hypothesis. Let $H_1: \rho = \rho^* \neq \rho_o$ in which ρ_o is zero. It then suggests that the statistical power is the probability of rejecting the null hypothesis $H_0: \rho = \rho_o$ in favor of the research/alternative hypothesis $H_1: \rho = \rho^* \neq \rho_o$. That is, under the research hypothesis, the minus log likelihood ratio is

$$- \ln A_{\rho^*}$$

$$= n \ln \left(\frac{1 + \hat{\rho}_{mle} \hat{\theta}_{\rho_{mle}}}{(1 + \rho^* \hat{\theta}_{\rho^*})} \right) + n (\hat{\theta}_{\rho_{mle}} - \hat{\theta}_{\rho^*})$$

$$+ n \bar{y} \ln \left(\frac{\hat{\theta}_{\rho^*}}{\hat{\theta}_{\rho_{mle}}} \right) + \sum_{i=1}^n \ln \left(\frac{1 + \rho^* y_i}{1 + \hat{\rho}_{mle} y_i} \right) \quad (19)$$

In addition, it follows a non-central chi-squared distribution with one df and non-centrality parameter

$$\hat{\delta}_{\rho^*} = \frac{(\hat{\rho}_{mle} - \rho^*)}{\text{var}(\hat{\rho}_{mle})}$$

$$(\hat{\rho}_{mle} - \rho^*) \left\{ \frac{\bar{y} - \left(\frac{\hat{\rho}_{mle}\bar{y}}{1 + \hat{\rho}_{mle} + \hat{\rho}_{mle}\bar{y}} \right)^2}{(1 + \hat{\rho}_{mle}\bar{y})^4} \left[\frac{(1 + \hat{\rho}_{mle})^2}{\bar{y}} - \left(\frac{\hat{\rho}_{mle}(1 + \hat{\rho}_{mle})}{1 + \hat{\rho}_{mle} + \hat{\rho}_{mle}\bar{y}} \right)^2 \right] - \left(\frac{1 + \hat{\rho}_{mle}}{1 + \hat{\rho}_{mle} + \hat{\rho}_{mle}\bar{y}} \right)^4 \right\}$$

$$= \frac{\left[\frac{(1 + \hat{\rho}_{mle})^2}{\bar{y}} - \left(\frac{1 + \hat{\rho}_{mle}}{1 + \hat{\rho}_{mle} + \hat{\rho}_{mle}\bar{y}} \right)^2 \right]}{\left[\frac{(1 + \hat{\rho}_{mle})^2}{\bar{y}} - \left(\frac{1 + \hat{\rho}_{mle}}{1 + \hat{\rho}_{mle} + \hat{\rho}_{mle}\bar{y}} \right)^2 \right]} \quad (20)$$

This non-central chi squared distribution with one df and non-centrality parameter $\hat{\delta}_{\rho^*}$ approximately follows $\left(1 + \frac{\hat{\delta}_{\rho^*}}{1 + \hat{\delta}_{\rho^*}}\right)$ times a central chi squared distribution with

$\frac{(1 + \hat{\delta}_{\rho^*})^2}{(1 + 2\hat{\delta}_{\rho^*})}$ df. The power is the probability of accepting the research/alternative hypothesis H_1 when $\rho = \rho^*$. That is,

$$\begin{aligned}
\text{Power} &= \Pr \left[\text{Accept } H_1 \text{ when } \rho = \rho^* \right] \\
&= \Pr \left[\left(1 + \frac{\hat{\delta}_{\rho_o}^*}{1 + \hat{\delta}_{\rho_o}^*} \right) \chi^2_{\left(\frac{1 + \hat{\delta}_{\rho_o}^*}{1 + 2\hat{\delta}_{\rho_o}^*} \right)^2, \alpha} < \frac{\left(1 + \frac{\hat{\delta}_{\rho_o}^*}{1 + \hat{\delta}_{\rho_o}^*} \right) \chi^2_{\left(\frac{1 + \hat{\delta}_{\rho_o}^*}{1 + 2\hat{\delta}_{\rho_o}^*} \right)^2, \alpha}}{(-\ln \Lambda_{\rho_o=0})} \right] \\
&= \Pr \left[\chi^2_{\left(\frac{1 + \hat{\delta}_{\rho_o}^*}{1 + 2\hat{\delta}_{\rho_o}^*} \right)^2, \alpha} < \frac{\left(1 + \frac{\hat{\delta}_{\rho_o}^*}{1 + \hat{\delta}_{\rho_o}^*} \right) \chi^2_{\left(\frac{1 + \hat{\delta}_{\rho_o}^*}{1 + 2\hat{\delta}_{\rho_o}^*} \right)^2, \alpha}}{\left(1 + \frac{\hat{\delta}_{\rho_o}^*}{1 + \hat{\delta}_{\rho_o}^*} \right) (-\ln \Lambda_{\rho_o=0})} \right].
\end{aligned} \tag{21}$$

3 Illustrations

In this section, the data consisting of smallpox cases that are displayed in Bailey [1, p.105] are considered for illustration of the results in Section 2. Bailey reported the number, Y of days waited to remove a smallpox case in a closed community of 100 individuals in Abakaliki, Nigeria (as in the Table 1 below). Two graphs and one table are included.

The MLEs are $\hat{\rho}_{mle} \approx 0.48$ and $\hat{\theta}_{mle} \approx 0.93$ using (11) and (10) respectively. It is clear that there is an impact of removing infected smallpox cases on the incidence rate, θ and on the observation, y as well. The usual Poisson distribution $p(y, |\theta) = \Pr[Y = y] = e^{-\theta} \theta^y / y!$ is therefore inappropriate for this data with the incidence rate θ . Under the usual Poisson distribution, the estimate of incidence rate would have been $\hat{\theta}_{\rho=0} \approx \bar{y} = 1.39$. The graph of Y versus $\frac{1}{1+\rho Y}$ in Fig. 1 confirms the appropriateness of SPD for the given data. Their linear relationship is captured by the correlation coefficient and it is estimated to be $\widehat{corr}\left(Y, \frac{1}{1+\rho Y}\right) = -0.79$ due to (6) which is quite different from zero. If the data were to be simple Poisson with no sampling bias on y and no impact on the incidence rate due to the removal of infected cases, the configuration in Fig. 1 would have been horizontal with zero correlation. But, it did not happen.

The cumulative distribution function, $CDF(m, \theta, \rho) = \Pr[Y \leq m]$ or implicitly its survival function, $Sf(m+1, \theta, \rho) = 1 - CDF(m, \theta, \rho)$ of the spinned Poisson distribution is after algebraic simplifications,

$$\begin{aligned}
&Sf\left(m+1, \hat{\theta}_{mle}, \hat{\rho}_{mle}\right) \\
&= \Pr[Y > m+1] \\
&= \frac{\left[CDF_{\chi^2_{2(m+1),df}}(1.86) + 0.45 CDF_{\chi^2_{2m,df}}(1.86) \right]}{1.45}
\end{aligned}$$

in which $CDF_{\chi^2_{2m,df}}(1.86)$ is the cumulative chi-squared distribution function with $2m$ df and percentile (1.86).

The Fig. 2 illustrates a graphical comparison of the survival functions of the usual Poisson, spinned Poisson and empirical distributions of the data in Table 1. The vertical and horizontal axes represent survival probability and $m+1$ respectively. The survival function of the spinned Poisson survival function rather than the survival function of the usual Poisson distribution is closer to the empirical survival function.

The immediate removal occurs when the number of waiting days to remove an infected case is zero, meaning that $Y=0$. The probability for an *immediate removal* situation to occur is

$$\begin{aligned}
&\Pr[Y = 0] \\
&= 1 - Sf\left[1, \hat{\theta}_{mle}, \hat{\rho}_{mle}\right] \\
&= \left[CDF_{\chi^2_{2,df}}(1.86) \right] / 1.45 \\
&= 0.59.
\end{aligned}$$

The null hypothesis is $H_0: \rho = \rho_o = 0$ and the research/alternative hypothesis is $H_1: \rho = \rho^* = 1$. These are now tested using the likelihood ratio test. Under the null hypothesis, the minus log likelihood ratio is $-\ln \Lambda_{\rho_o=0}^{\rho_{mle}} = 14.14$ and the non-centrality parameter $\hat{\delta}_{\rho_o=0} = 0.008$. This non-central chi squared value with one df and non-centrality parameter is approximately $\left(1 + \frac{\hat{\delta}_{\rho_o=0}}{1 + \hat{\delta}_{\rho_o=0}} \right) = 1.0079$ times a central chi squared distribution with $\frac{(1 + \hat{\delta}_{\rho_o=0})^2}{(1 + 2\hat{\delta}_{\rho_o=0})} \approx 1.00006$ df. The critical chi-squared score is 3.87 with $\alpha = 0.05$. Because $-\ln \Lambda_{\rho_o=0}^{\rho_{mle}} = 14.14 > 3.87$, notice that the null hypothesis $H_0: \rho = \rho_o = 0$ is rejected with a 95% confidence level. Under the research hypothesis, the minus log likelihood ratio is $-\ln \Lambda_{\rho^*=1}^{\rho_{mle}} = 4.85$ and the non-centrality parameter $\hat{\delta}_{\rho^*=1} = 0.0084$. This non-central chi squared value with one df and non-centrality parameter is approximately $\left(1 + \frac{\hat{\delta}_{\rho^*=1}}{1 + \hat{\delta}_{\rho^*=1}} \right) = 1.008$ times a central chi squared distribution with $\frac{(1 + \hat{\delta}_{\rho^*=1})^2}{(1 + 2\hat{\delta}_{\rho^*=1})} \approx 1.000069$ df.

Using the likelihood ratio criterion, the statistical power of rejecting the null hypothesis in favor of the research hypothesis is calculated using Eq. 21. That is, $\text{Power} = \Pr[\chi^2_{0.991df} < 0.34] = 0.55$ when the research hypothesis $H_1: \rho = \rho^* = 1$ is true.

4 Comments and conclusions

The results of this article are applicable to analyzing data from business, engineering, economics, and sociologic

studies in which some cases are removed from the study population due to a screening criterion. The results based on the spinned Poisson distribution could become the foundation for performing further reliability or validity analysis. These and other aspects are currently under investigation and the findings will be reported later.

Acknowledgement The author sincerely thanks all three reviewers and the editor for comments and suggestions, which helped to greatly improve the manuscript.

References

1. Bailey NTJ (1975) The mathematical theory of infectious diseases and its applications, 2nd edn. Griffin, London
2. Cox DR (1969) Some sampling problems in technology. In: Johnson NL, Smith H Jr (eds) New development in survey sampling. Wiley Interscience, New York, pp 506–517
3. Johnson N, Kemp A, Kotz S (2005) Univariate discrete distributions. Wiley Publication, New York
4. Stuart A, Ord K (1994) Kendall's advanced theory of statistics, volume I. Griffin Publication, London
5. Wald A (1943) Tests of statistical hypotheses concerning several parameters when the number of observations is large. Trans Am Math Soc 54:416