Ann. Inst. Statist. Math.
Vol. 50, No. 3, 567–585 (1998)

WEIGHTED POISSON DISTRIBUTIONS FOR OVERDISPERSION AND UNDERDISPERSION SITUATIONS

JOAN DEL CASTILLO[1] AND MARTA PÉREZ-CASANY[2]

[1] *Dpt. Matemàtiques, Universitat Autònoma de Barcelona, 08193 Cerdanyola, Barcelona, Spain*
[2] *Dpt. Matemàtica Aplicada II, Universitat Politècnica de Catalunya, 08028 Barcelona, Spain*

(Received October 21, 1996; revised October 6, 1997)

Abstract. The main goal of this paper is to introduce new exponential families, that come from the concept of weighted distribution, that include and generalize the Poisson distribution. In these families there are distributions with index of dispersion greater than, equal to or smaller than one. This property makes them suitable to fit discrete data in overdispersion or underdispersion situations. We study the statistical properties of the families and we provide a useful interpretation of the parameters. Two classical examples are considered in order to compare the fits with some other distributions. To obtain the fits with the new family, the study of the profile log-likelihood is required.

Key words and phrases: Weighted version, dispersion index, stochastic order relations, exponential family, scoring method.

1. Introduction

Because the Poisson distribution is the limit of Binomial distributions, it has been widely used in life and social sciences, specially in fitting the distributions of the number of events that only occur with low probability. Nevertheless, the mean and the variance are equal in Poisson distribution. Equivalently, the index of dispersion, a measure of aggregation or repulsion, is always equal to one. This is sometimes too restrictive to fit practical data, which usually present indexes of dispersion greater (aggregation) or smaller (repulsion) than one. Many distributions have been introduced in order to weaken this assumption. The Negative Binomial distribution appears naturally as a mixture of Poisson distribution when the parameter follows a Gamma distribution and this has been used since Greenwood and Yule (1920) to fit data with aggregation. Other important generalizations are the double Poisson, which is also a mixture, and the generalized Poisson distribution proposed by Consul. More information on this subject can be found in Haight (1967), Johnson *et al.* (1992), Kendall and Stuart (1979) and Consul (1989).

In this paper we provide a new approach to the previous problem using the concept of weighted distribution introduced by C. R. Rao in 1963 (see Rao (1965)).

This concept has been used for the last thirty years in the selection of the appropriate model for observed data, especially when samples are recorded without a sampling frame that enables random samples to be drawn. The weight function that usually appears in the scientific and statistical literature is $w(k) = k$, which provides the size-biased version of the random variable. The size-biased version of order r, which corresponds to the weight $w(k) = k^r$, for r any real positive number, has also been widely used. The weight proposed in this paper is $w(k) = (k+a)^r$, where a is a positive displacement parameter. The weighted Poisson distribution (WPD) that results from the modification of the Poisson distribution with this weight can also be considered as a mixture of the size-biased versions of the Poisson distribution, as we show in Subsection 3.3. This new family of probability distributions can be studied from the theoretical and the numerical points of view. Moreover, they can be seen as stochastic models to describe and analyse empirical data of the type refered to above.

First, it is important to remark that given $a > 0$, the Poisson distribution is nested in the corresponding sub-family, which we denote by WPD_a (it corresponds to $r = 0$). Thus, WPD_a is a two-parameter family of probability distributions containing the Poisson. Let us observe that this property is not verified for the Negative Binomial family of distributions. WPD_a includes overdispersion ($r < 0$) and underdispersion ($r > 0$) distributions (see Theorem 2.1 to interpret parameter r as a repulsion parameter). Figure 2.1 shows that positive values of r concentrate the probabilities around the mean, and negative values increase the dispersion, with respect to the Poisson distribution.

Given $a > 0$, the sub-families WPD_a are regular exponential families, which are the distributions of maximum entropy given a sufficient statistic. In this case, the sufficient statistic is defined by the arithmetic and the geometric means of the data shifted by a. Hence, the classical theory of likelihood works perfectly in these families and the scoring method provides useful solutions to the likelihood equations. We refer to Barndorff-Nielsen (1978) for a general theory of exponential families. See also Efron (1978), Letac (1992), Brown (1986) and Castillo (1994).

With respect to the three-parameter model, the stochastic order relations between two probabilities of the family are established. From the inference point of view, we propose to find the maximum likelihood estimator associated to the family WPD_a and to study the profile log-likelihood with respect to parameter a (see Section 4).

Finally, we give two examples of fitting practical data by a distribution of the family WPD. The first comes from a paper in which Greenwood and Yule (1920) introduced the Negative Binomial as an alternative to the Poisson distribution, in overdispersion situations. The second is an example of underdispersion. This is one of the examples used by Consul (1989) to show the good properties of his Poisson generalization. In both cases, WPD gives better fittings (see Tables 1 and 2).

Moreover, the weighted Poisson distribution can also be used in generalized linear models with covariates (see McCullagh and Nelder (1989)), in particular, in log-linear models where the variance function is assumed to be linear. Obviously, the linearity of the variance function is verified for the Poisson distribution, but for a WPD the variance function is "almost" linear in a large range of the mean parameters. In Fig. 2.2, we compare the variance function for the Negative Binomial distribution and the WPD for the Greenwood and Yule example. It is also possible to appreciate this quasi linearity in Fig. 4.1, since the behaviour of the dispersion index is almost constant when the repulsion parameter r is fixed.

2. Weighted poisson distributions

In this section we introduce new families of probability distributions related to the Poisson distribution. We give an interpretation of the parameters, we study the stochatic order relations between the different families and we show that those families are applicable in overdispersion (agregation) and underdispersion (repulsion) situations.

2.1 *Point probability functions*

Following the concept of weighted distribution introduced by Rao in 1963 (see Rao (1965)), and given a random variable X with Poisson distribution with parameter λ, we consider for $a \geq 0$, $r \in \Re$, the weight $w(k) = (k+a)^r$ and the weighted Poisson distribution (WPD), with point probability functions given by:

$$P(k; \lambda, r, a) = \frac{(k+a)^r \lambda^k e^{-\lambda}}{E_\lambda[(X+a)^r]k!} \quad (k = 0, 1, \ldots), \tag{2.1}$$

where we denote by $E_\lambda[\cdot]$ the mean value with respect to the Poisson distribution with parameter λ.

Let us remark that, when $a > 0$, it is possible to express (2.1) in the following way:

$$P(k; \lambda, r, a) = \frac{\exp(k \log \lambda + r \log(k+a))}{k! C(\lambda, r, a)}, \tag{2.2}$$

where:

$$C(\lambda, r, a) = e^\lambda E_\lambda[(X+a)^r] = \sum_{k=0}^{\infty} \frac{\lambda^k (k+a)^r}{k!}. \tag{2.3}$$

From the quotient criterion, it follows that series (2.3) converges for all $\lambda > 0$, $a > 0$ and $r \in \Re$ and $\lambda > 0$, $a = 0$ and $r \in \Re^+$ (more details on the convergence of this series will be considered in Section 5). From (2.2) we see that, given $a > 0$, the corresponding sub-family of probabilities, which we denote by WPD_a, is a *full exponential family* on $\mathcal{Z}^+$. The reference measure is $\mu = \sum_{k=0}^{\infty} \delta_k / k!$, where δ_k means the evaluation at k and $T_a(k) = (k, \log(k+a))$ is the sufficient statistic. Moreover, it is a *regular* exponential family with *natural parameters* $(\theta, r) \in \Re^2$, where $\theta = \log(\lambda)$ (see Barndorff-Nielsen (1978)).

Note that if in (2.1) we take $r = 0$, we obtain the point probability function of a Poisson distribution with parameter λ. Therefore, the Poisson family of

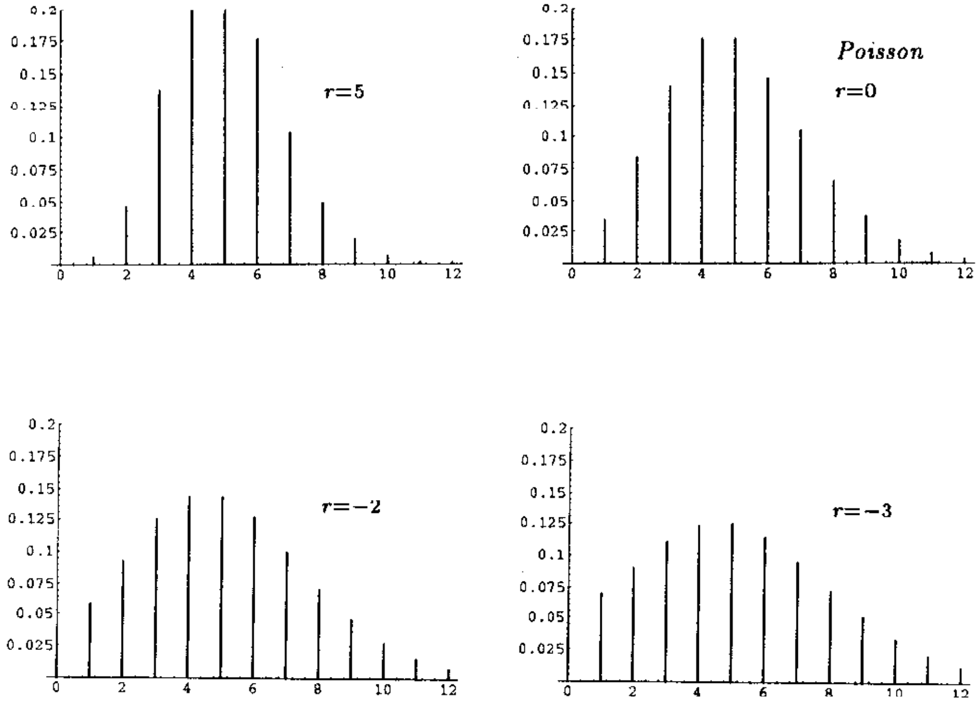

Fig. 2.1. Graphs of probability distributions for the WPD model, when $a = 1$, the mean $\mu = 5$ and the repulsion parameter $r = 5, 0, -2, -3$.

distributions is nested in all the sub-families WPD_a and it corresponds to a linear subspace in the *natural parameter space*.

In order to understand the behaviour of the point probability function WPD varying the values of the parameters, we advance some ideas that will be developed later. First, parameter λ is closely related to the mean value. When $r = 0$, λ is just the mean value and Fig. 4.1 shows that given λ, the mean value increases with r (see also Corollary 2.2). The interpretation of parameter r as a repulsion parameter is given in Theorem 2.1. Moreover, Fig. 2.1 shows that positive values of r concentrate the probabilities around the mean, and negative values increase the dispersion around the mean, with respect to the Poisson distribution ($r = 0$). Finally, parameter a is a mesure of the proximity to the Poisson distribution (see Section 3). Moreover, small perturbations of this parameter give an oscillatory behaviour of the probabilities $P(k, \lambda, r, a)$. This property will be appreciated in Section 5.

2.2 *Stochastic order relations*

In this subsection we are going to study the stochastic order relations between two random variables which have a WPD with different parameters. In particular we are going to compare the Poisson distribution with a general WPD.

First, it is important to notice that given two random variables with a WPD with different parameters it is always possible to consider one a weighted version of the other. From now on, we will denote by WPD(λ, r, a) the weighted Poisson distribution with parameter (λ, r, a). In particular, WPD$_a$ is equivalent to WPD$(\cdot, \cdot, a)$. If X is a random variable with a WPD(λ_1, r, a), then the weighted version X^{w_1}, where $w_1(k) = (\lambda_2/\lambda_1)^k$, is a WPD$(\lambda_2, r, a)$. Similary, the weights $w_2(k) = (k+a)^{(r_2-r_1)}$ and $w_3(k) = (k + a_2/k + a_1)^r$ transform a WPD(λ, r_1, a) into a WPD(λ, r_2, a), and a WPD(λ, r, a_1) into a WPD(λ, r, a_2) respectively.

Patil *et al.* (1986) have proved that given a random variable X, the weighted version X^w is stocastically greater or smaller than the original random variable X according as the weight function $w(x)$ is monotone increasing or decreasing in x. Now, if $0 < \lambda_1 < \lambda_2$ and $r_1 < r_2$ the weights w_1 w_2 are increasing functions of k. On the other hand, if $r \leq 0 (\geq 0)$ and $0 < a_1 < a_2$, the weight w_3 is also an increasing (decreasing) function of k. So, this gives the following proposition:

PROPOSITION 2.1. *Assume that X_i has a WPD(λ_i, r_i, a_i) for $i = 1, 2$. If $a_1 = a_2$ and $(\lambda_1, r_1) \leq (\lambda_2, r_2)$ then X_2 is stochastically greater than X_1. If $a_1 < a_2$ and $(\lambda_1, r_1) = (\lambda_2, r_2)$, if $r \leq 0 (\geq 0)$ then X_2 is stochastically greater (smaller) than X_1.*

COROLLARY 2.1. *If $r > 0 (< 0)$ a random variable with a WPD(λ, r, a) is stochastically greater (smaller) than a random variable with a Poisson distribution with parameter λ.*

2.3 *Mean, variance and moments*

Let X be a random variate with a WPD(λ, r, a), $r \subset \Re$, $\lambda > 0$, $a > 0$. Denote the mean values with respect to this distribution by $E[\cdot]$. We now give some properties of monotony related to the mean values of any monotonous function of X.

From Ross (1983), a random variable X is stochastically greater than a random variable Y if and only if for every nondecreasing function g, $E[g(X)] \geq E[g(Y)]$. Consequently:

COROLLARY 2.2. *Let $g(x)$ be a real-valued nondecreasing (nonincreasing) function. Given any value of the displacement parameter a, the mean values $E[g(X)]$ are nondecreasing (nonincreasing) functions of the parameters (λ, r), $\lambda > 0$, $r \in \Re$. Moreover, $E[g(X)]$ is also a nondecreasing (nonincreasing) function of the parameter a if $r \leq 0$.*

In particular, the moments about the origin are nondecreasing functions of (λ, r), $\lambda > 0$, $r \in \Re$ and also a nondecreasing functions of a if $r \leq 0$.

We can write the mean and the variance of a WPD in terms of the normalizing function $C(\lambda, r, a)$. Given $s \in \Re$,

$$E[(X+a)^s] = \frac{C(\lambda, r+s, a)}{C(\lambda, r, a)} \quad (s \in \Re). \tag{2.4}$$

Clearly, from (2.3), the function $C(\lambda, r, a)$ verifies the recurrent property

$$C(\lambda, r+1, a) = \lambda C(\lambda, r, a+1) + aC(\lambda, r, a), \tag{2.5}$$

that permits us to reduce r in a numerical approach to $C(\lambda, r, a)$ (see Section 5). Also, taking $s = 1$ and $s = 2$ into (2.4), we have the following expressions for the mean and the variance:

$$E(X) = \lambda \frac{C(\lambda, r, a+1)}{C(\lambda, r, a)} \tag{2.6}$$

and

$$V(X) = \frac{C(\lambda, r+2, a)C(\lambda, r, a) - C^2(\lambda, r+1, a)}{C^2(\lambda, r, a)}. \tag{2.7}$$

2.4 *Index of dispersion*

Given a random variable X, the index of dispersion $I(X) = V(X)/E(X)$ is a measure of aggregation or repulsion. Usually, the events cluster in time or in space because of the environment. When this happens, $I(X)$ is greater than one. Alternatively, the absence of aggregation gives values for $I(X)$ less than one. If X has a Poisson distribution then $I(X) = 1$. For this reason, $I(X)$ has been used as an indicator of the degree of departure from the Poisson distribution.

When the empirical index of dispersion is greater than one (and only in this case), it is often better to use the Negative Binomial distribution rather than the Poisson distribution in order to fit empirical data (see Johnson *et al.* (1992)). As we will see, the WPD admits indexes of dispersion greater than, equal to, or smaller than one. This fact makes the WPD suitable to fit empirical data in different fields, even in repulsion situations.

From (2.6) and (2.7), the index of dispersion for the WPD is expressed as:

$$I(X) = \frac{C(\lambda, r+2, a)C(\lambda, r, a) - C^2(\lambda, r+1, a)}{C(\lambda, r, a)C(\lambda, r+1, a) - aC^2(\lambda, r, a)}. \tag{2.8}$$

Moreover, from (2.5) we deduce that the denominator of (2.8) is positive. Thus, saying that $I(X) > 1$ is equivalent to saying that

$$C^2(\lambda, r+1, a) - aC^2(\lambda, r, a) < (C(\lambda, r+2, a) - C(\lambda, r+1, a))C(\lambda, r, a)$$

and applying (2.5) several times, we obtain:

$$I(X) > 1 \quad \text{iff} \quad C^2(\lambda, r, a+1) < C(\lambda, r, a)C(\lambda, r, a+2). \tag{2.9}$$

This inequality allows us to characterize the aggregation or repulsion of the WPD in the following result and, in consequence, to interpret the parameter r as a repulsion parameter.

THEOREM 2.1. *Suppose that the random variable X is distributed with a WPD(λ, r, a), $r \in \Re$, $\lambda > 0$, $a > 0$. Therefore, the index of dispersion is $I(X) = 1$ if and only if $r = 0$ or, equivalently, X has a Poisson distribution. Moreover, $I(X) > 1$ ($I(X) < 1$) if and only if $r < 0$ ($r > 0$).*

PROOF. See the Appendix.

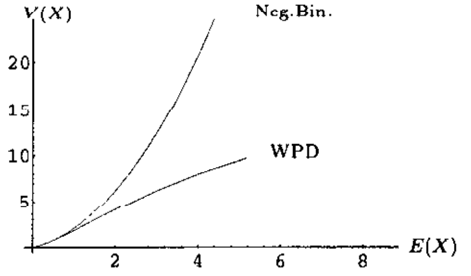

Fig. 2.2. Graphs of the variance function of the Neg. Bin. and the WPD, for the Greenwood and Yule data (Example 1).

3. Properties related to the Poisson distribution

In this section, the Poisson distribution will be used to obtain a new expresion for the normalization constant. An alternative estimation method will be analyzed. Further, the concepts of weighted version and mixture will be related for a general distribution. Finally, an example will be used to compare the variance function for the WPD and the Negative Binomial distribution.

3.1 *Closed form for the normalization constant*

The Poisson distribution is closely related to the WPD. As we saw in Subsection 2.1, the Poisson distribution belongs to the sub-family WPD_a, for all values of a. In fact, it corresponds to the subspace $r = 0$ of the natural parameter space. So, the Poisson distribution is a linear subfamily of WPD_a. Further, as a consequence of the last theorem, we find that a WPD has index of dispersion equal to one if and only if it is a Poisson distribution. Moreover, we can use the normalization constant $C(\lambda, r, a)$ to compute moments for a translation of a Poisson variable and we can compute $C(\lambda, r, a)$, for $r \in Z^+$, from the moment generating function of a Poisson distribution. More precisely, let X be a random variable with Poisson distribution of parameter λ, and denote by $E_\lambda[\cdot]$ the mean value with respect to this distribution. Therefore, given $a > 0$ and $r \in \Re$ we can compute

$$E_\lambda[(X+a)^r] = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k (k+a)^r}{k!} = e^{-\lambda} C(\lambda, r, a).$$

Moreover, if $r \in Z^+$, using the moment generating function of $X + a$, it follows the next closed form for the normalization constant:

$$C(\lambda, r, a) = \frac{d^r}{dt^r} \exp[at + \lambda e^t] \,|_{t=0} \qquad (r = 0, 1, 2, \ldots). \tag{3.1}$$

This result can be used as an alternative way to compute $C(\lambda, r, a)$ for positive integer values of r. In this case, the normalization constant is equal to the product of e^λ by a polynomial of degree r in λ.

3.2 *Alternative method of estimation*

Denote by $P(k;\lambda)$, $(k = 0, 1, 2, \ldots)$ the point probability function of a Poisson distribution with parameter λ. Then we have $P(k+1;\lambda)/P(k;\lambda) = \lambda/(k+1)$. Analogously, by (2.1), the ratio of two successive probabilities of a WPD verifies:

$$\frac{P_{k+1}}{P_k} = \frac{\lambda}{k+1}\left(1 + \frac{1}{k+a}\right)^r, \tag{3.2}$$

where $P_k = P(k;\lambda, r, a)$. Note that the second factor in the right hand of (3.2) tends to one, for every $a > 0$, $r \in \Re$, when k tends to infinity. Then, the tails of a WPD with parameter λ are similar to the tails of the Poisson distribution with parameter λ.

We now deduce an estimation method for the parameters of the model using (3.2). Fixing an integer non-negative value for k and applying (3.2) for k, $k+1$, and $k+2$, we find that the theoretic probabilities given by the model verify:

$$\frac{P_k P_{k+2}}{P_{k+1}^2} = \frac{k+1}{k+2}\left(\frac{(k+a)(k+a+2)}{(k+a+1)^2}\right)^r, \tag{3.3}$$

for every integer positive value of k. Applying (3.3) for k and $k+1$ and taking logarithms we can avoid parameters λ, r and concentrate in the following equation in a:

$$\frac{\log\left(\frac{k+2}{k+1}\frac{P_k P_{k+2}}{P_{k+1}^2}\right)}{\log\left(\frac{k+3}{k+2}\frac{P_{k+1} P_{k+3}}{P_{k+2}^2}\right)} = \frac{\log\left(\frac{(k+a)(k+a+2)}{(k+a+1)^2}\right)}{\log\left(\frac{(k+a+1)(k+a+3)}{(k+a+2)^2}\right)} \tag{3.4}$$

Instead of putting the theoretical probabilities P_k in (3.4), we can put the observed ones and find a possible value for parameter a numerically. We can use this value to obtain values for parameters λ and r from (3.2) and (3.3) using, again, the observed probabilities instead of the theoretical ones.

The main advantage of this estimation method is its simplicity. It does not depend on $C(\lambda, r, a)$ and it only requires the solution of one uniparametric equation. As a disadvantage, it does not always work. Let us denote the right hand side of (3.4) by $f_k(a)$. In order to apply this method it is necessary that the left hand side of (3.4) evaluated with the observed probabilities lies between values 1 and $f_k(0)$ for at least one integer non-negative value of k, and this is not always verified. Nevertheless, if parameter a is fixed, the solutions of (3.3) and (3.2) are good starting points in order to find the maximum likelihood estimations using the *scoring method* (see, Section 4), specially when k is close to the mode of the empirical distribution.

3.3 *Weighted versions and mixtures*

Patil and Rao (1978) pointed out the importance of the *size-biased version*, X^*, of a random variable X. That is the weighted version with weight $w(k) = k$. In particular, it is interesting to remark that the next property is verified $E(X^*) - E(X) = I(X)$. They show that many classical discrete distributions have a *size-biased version* of the same form with the variable reduced by unity. For instance, this is true for the Poisson, Binomial, Negative Binomial, Hypergeometric and Binomial Beta distributions. The same result is also true for the WPD. More specifically, if X has a WPD(λ, r, a) then, X^* has a WPD$(\lambda, r, a+1)$ and it takes values $k = 1, 2, \ldots$.

It is important to notice that the weighted version with weight $w_0(k) = k + a$ of a random variable X with a Poisson distribution is a mixture of X and X^*. More exactly,

$$\Pr\{X^{w_0} = k\} = \left(1 - \frac{\lambda}{\lambda + a}\right)\Pr\{X = k\} + \frac{\lambda}{\lambda + a}\Pr\{X^* = k\}.$$

It is easy to observe that the greater the value of parameter a, the nearer we are to the Poisson distribution. Consequently, we can interpret parameter a as an indicator of the proximity to the Poisson distribution.

More generally, given any discrete random variable X, its weighted version X^w with $w(k) = (k+a)^r$, $r \in \mathcal{Z}^+$, is a mixture of the successive *size-biased versions of order s* of X for $s = 0, 1, 2, \ldots, r$, which we denote by X^{*s}. More exactly,

$$\Pr\{X^w = k\} = \sum_{s=0}^{r} \binom{r}{s} \frac{a^{r-s}\mu_s}{b} \Pr\{X^{*s} = k\},$$

where μ_s is the s-th moment about the origin of X and $b = \sum_{s=0}^{r} \binom{r}{s} a^{r-s}\mu_s$. Notice that if $r \in \mathcal{Z}^+$, the WPD is a mixture of the successive *size-biased versions* of a Poisson distribution.

3.4 *Weighted Poisson distribution and the generalized linear models*

The generalized linear models, introduced by Nelder and Wedderburn (1972), are determined by the variance function of the error distributions (see McCullagh and Nelder (1989)), and the *link* between means and covariates. The log-linear models assume linear variance function and a logarithmic relation between means and covariates. The Poisson distribution is, of course, the most characteristic error distribution with these properties. In many practical situations the overdispersion problem appears, i.e., $I(X)$ is significatively greater than one. In these situations the Negative Binomial distribution is generally used instead of the Poisson distribution to model the data. Nevertheless, the variance function of the negative Binomial distribution is a quadratic function, which is far from the assumption of linearity.

In Fig. 2.2 we show the variance function for the Negative Binomial distribution and the WPD related to the data of the Example 1 in Section 5. We can see that for the WPD the variance function is almost linear which shows its usefulness in log-linear models. This property is not only verified for the Example 1 data.

Figure 4.1 also shows that $I(X)$ is almost constant when parameter r is fixed and then, $V(X)$ is almost linear in terms of $E(X)$.

In fact, using (2.7) and the recurrent formula (2.5), the variance of the WPD with parameters (λ, r, a) may be written as:

$$V_{(\lambda,r,a)}(X) = (E_{(\lambda,r,a)}(X) + a)(E_{(\lambda,r+1,a)}(X) - E_{(\lambda,r,a)}(X)).$$

In the particular case that $r = n \in \mathcal{Z}^+$, (3.1) allows us to write:

$$C(\lambda, n, a) = e^{\lambda}\left(\lambda^n + \left(na + \frac{n(n-1)}{2}\right)\lambda^{n-1} + O(\lambda^{n-2})\right).$$

Using this expression and (2.6), it is possible to prove that $E_{(\lambda,r+1,a)}(X) - E_{(\lambda,r,a)}(X)$ tends to one when λ tends to $+\infty$. Consequently, the variance of the WPD is approximatly equal to the linear function of the mean $E_{(\lambda,r,a)}(X) + a$ when parameter λ is big enough and r is a non negative integer. Numerically, it is possible to see that this property is verified for any real value of r. Unfortunatly, we do not have a general proof for it.

4. Statistical inference

In Section 3, an alternative estimation method has been introduced. However, it is not always applicable. In this section we deal with the maximum likelihood estimator which, of course, is the best way to do inference. If parameter a is suposed to be known, the problem is simpler to study because the distribution is a two-parameter exponential family, as we have said in Section 2. So, if parameter a is unknown, what we suggest is to study the profile log-likelihood varing parameter a in order to find the best parameter estimation for the three parameter-model.

4.1 *Likelihood equations*

The concept of weighted distribution arises from sampling with unequal chances to observations to be recorded (see Rao (1965)). In such a situation, it is necessary to modify the original distribution with a certain weight so that the sample comes from the new probability distribution. Thus, the weight is closely related to the way in which the sample is obtained and, sometimes it is possible for the researcher who collects the sample to give advice about a possible value for parameter a. Note that the positive displacement a is required in order to avoid the evaluation of the logarithm of zero.

If we are not in the last situation, a good way to find a maximum of the log-likelihood function of the three-parameter model is to find for several values of a the maximum likelihood estimations of the two-parameter model and to plot the profile log-likelihood varying a. In particular, it is interesting to study the neighbourhood of $a = 1/2$ (see Figs. 5.1 and 5.2 in Section 5). The reason is as follows, let us suppose that X is a random variable with a Poisson(λ) distribution, the equation in a $E_\lambda[\log(X + a)] = \log(\lambda)$ has only one solution that is smaller than and asymptotically equal to $1/2$. From now on, we are going to suppose that the value of parameter a is fixed and we are going to work with the two-parameter model.

Let $x = (x_1, x_2, \dots, x_n)^\dagger$ be a random sample of a $\mathrm{WPD}_a(\lambda, r)$. Let us define

$$t_1 = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \text{and} \qquad t_2 = \frac{1}{n}\sum_{i=1}^{n} \log(x_i + a).$$

Note that, t_1, t_2 are respectively the sample mean, and the *log-geometric* mean of the sample shifted by a.

The log-likelihood function for the WPD_a model and the sample x is

$$l_a(\lambda, r; x) = n[\log(\lambda)t_1 + rt_2 - K(\lambda, r, a)], \tag{4.1}$$

where $K(\lambda, r, a) = \log C(\lambda, r, a)$.

The likelihood equations may be written as:

$$\left.\begin{aligned} \lambda\partial_\lambda K(\lambda, r, a) = E[X] = t_1 \\ \partial_r K(\lambda, r, a) = E[\log(X + a)] = t_2 \end{aligned}\right\}, \tag{4.2}$$

where we denote $\partial_\lambda = \partial/\partial\lambda$ and $\partial_r = \partial/\partial r$.

If sample x is drawn from a $\mathrm{WPD}_a(\lambda, r)$, solving the likelihood equations is equivalent to finding (λ, r) such that, for the random variable $X + a$, the *arithmetic* and the *log-geometric* means for the population and the sample are equal.

As we have observed, given $a > 0$ the corresponding sub-family of probabilities WPD_a is a full exponential family on $\mathcal{Z}^+$ with respect to the measure $\mu = \sum_{k=0}^{\infty} \delta_k/k!$. The natural parameters are (θ, r), where $\theta = \log(\lambda)$, and the natural parameter space is $\mathcal{R}^2$. Because $\mathcal{R}^2$ is an open set, WPD_a is a regular exponential family (see Barndorff-Nielsen (1978)). Let T_a be the statistic defined in Section 2 by $T_a(k) = (k, \log(k + a))$. Because $S_a = T_a(\mathcal{Z}^+)$ is not included in an affine subspace of $\mathcal{R}^2$, $\overline{T_a} = (t_1, t_2)$ is a minimal and sufficient statistic for the family WPD_a.

Let $\mathcal{T}_a$ be the interior of the convex hull of S_a. There is a one-to-one transformation between $\Re^2$ and the *domain of means* $\mathcal{T}_a$, because WPD_a is a regular exponential family. This transformation may be expressed in terms of $\tau(\theta, r) = E_{(\theta, r)}(T_a) = (\tau_1(\theta, r), \tau_2(\theta, r))^\dagger$ (see Barndorff-Nielsen (1978)).

So, if (t_1, t_2) belongs to $\mathcal{T}_a$, there is a solution for system (4.2) that is unique. To verify this assumption it is sufficient that sample x has more than two different values, or, if it has only two values, it is sufficient if those values are not consecutive.

To sum up, if the value a is known and nonzero, we can find the maximum likelihood estimations for parameters (λ, r) solving (4.2). Then, it is necessary to study the profile log-likelihood $l_p(a) = l(\hat{\lambda}_a, \hat{r}_a, a)$ where $(\hat{\lambda}_a, \hat{r}_a)$ are the solutions of (4.2). Usually $l_p(a)$ is unimodal, and sometimes its maximum will be obtained as a limit when a tends to zero (see Example 2 in Section 5).

4.2 *The method of scoring*

Equation (4.2) is highly nonlinear and must be solved using numerical methods. A general procedure to deal with this equations is to use the *scoring method*.

Let $F_a(\theta, r)$ be the Fisher matrix information at point $(\theta, r) \in \Re^2$. Because WPD_a is an exponential family we have:

$$F_a(\theta, r) = (\partial_i \partial_j K_a(\theta, r))_{i,j=1,2} = \mathrm{Cov}_{(\theta,r)}(T_a, T_a).$$

Now, let us define:

$$D^s(\lambda, r, a) = \sum_{k=0}^{\infty} \frac{\lambda^k (k+a)^r \log^s(k+a)}{k!}, \tag{4.3}$$

the elements $F_{i,j}$ $(i, j = 1, 2)$ expressed in function of λ, are given by:

$$F_{11} = \frac{\lambda^2 (C(\lambda, r, a+2) C(\lambda, r, a) - C^2(\lambda, r, a+1)) + \lambda C(\lambda, r, a+1) C(\lambda, r, a)}{C^2(\lambda, r, a)},$$

$$F_{12} = F_{21} = \frac{\lambda (C(\lambda, r, a) D^1(\lambda, r, a+1) - C(\lambda, r, a+1) D^1(\lambda, r, a))}{C^2(\lambda, r, a)}, \qquad \text{and:}$$

$$F_{22} = \frac{C(\lambda, r, a) D^2(\lambda, r, a) - (D^1(\lambda, r, a))^2}{C^2(\lambda, r, a)}.$$

Also, using this notation, $\tau_1(\theta, r)$ and $\tau_2(\theta, r)$ are expressed as:

$$\tau_1(\theta, r) = \lambda \frac{C(\lambda, r, a+1)}{C(\lambda, r, a)}, \qquad \tau_2(\theta, r) = \frac{D^1(\lambda, r, a)}{C(\lambda, r, a)}. \tag{4.4}$$

Given a trial estimate of the parameters (θ_n, r_n), we can update to (θ_{n+1}, r_{n+1}) by

$$\begin{pmatrix} \theta_{n+1} \\ r_{n+1} \end{pmatrix} = \begin{pmatrix} \theta_n \\ r_n \end{pmatrix} + F_a(\theta_n, r_n)^{-1} \begin{pmatrix} t_1 - \tau_1(\theta_n, r_n) \\ t_2 - \tau_2(\theta_n, r_n) \end{pmatrix}. \tag{4.5}$$

We propose two ways to obtain a starting point. The first one, which is very simple, is to use equations (3.2) and (3.3) with the observed probabilities instead of the theoretical ones and k close to the empirical mode, as suggested in Subsection 3.2. The other way, which is more precise, consists in drawing the parametric plot $(E[X], I[X])$, for a fixed value of a, where the mean and the dispersion index are associated with the parameters (λ, r, a), for different values of parameters λ and r. If we design all those parametric patterns simultaneously, the result is a reticle (see Fig. 4.1). To use this reticle, we estimate the mean and the dispersion index from the sample. That is, we calculate the point $(\bar{x}, s^2/\bar{x})$ where:

$$\bar{x} = t_1, \qquad \text{and} \qquad s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 / (n-1).$$

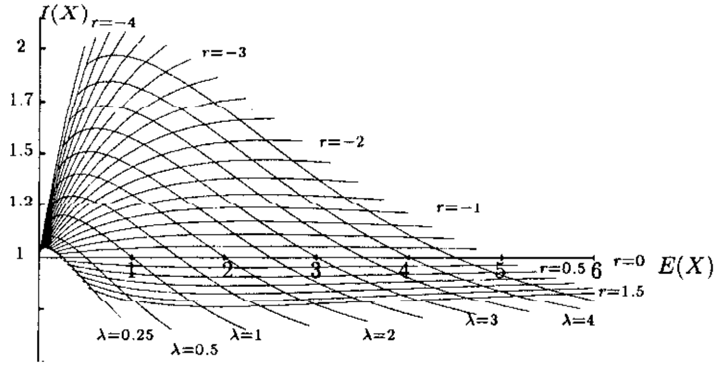

Fig. 4.1. Parametric plots of $(E(X), I(X))$ for $a = 1$ and different values for parameters λ and r.

Locating this point in the reticle, two values for parameter θ and two values for parameter r are determined. Those values define the zone where the point $(\bar{x}, s^2/\bar{x})$ is situated. The middle points of those values constitute a very good starting point, not only for the *scoring method* but also for any numerical method applicable to this situation. If parameter a is lightly modified, it is not necessary to recalculate the starting point. The same point is useful for different values of parameter a which do not differ greatly.

The equation (4.2), using the formulas (4.4), can also be solved by standard numerical methods such as the function *FindRoot* in MATHEMATICA (see Wolfram (1991)). Finally, note that the method of scoring in exponential families is equivalent to the Newton Raphson algorithm.

4.3 *Statistical testing*

It is important to test whether a sample comes from a Poisson distribution or not. Therefore, in this subsection we describe a test in large samples using the fact that the theoretical mean and the variance are equal for the Poisson distribution.

Assume $x = (x_1, x_2, \ldots, x_n)^\dagger$ is a sample of size n from a population with a Poisson distribution, with estimated mean and variance $\bar{x} = t_1$, and s^2.

A general test, which can be used to detect overdispersion or heterogeneity in the population sampled, is based on the statistic

$$S_R = \sqrt{n-1}(\bar{x} - s^2)/(\sqrt{2}\bar{x}),$$

which, if n is large, can be used as a standard normal deviate, (see Rao (1973), p. 439).

5. Fitting empirical data

In this section we give some examples of fitting practical data by a distribution of family *WPD*. We use two classical examples in order to compare the results with the fits given by some other distributions. The first was used to introduce the Negative Binomial distribution, as an alternative to the Poisson distribution, by Greenwood and Yule (1920). The second is one of the examples considered by Consul (1989) to introduce his Poisson generalization. The differences between observed and expected values are computed by the $\mathcal{X}^2$ test of Pearson.

First, we shall solve some numerical aspects in the computation of series $C(\lambda, r, a)$ and $D^s(\lambda, r, a)$ given in (2.3) and (4.3) respectively. Applying (2.5) several times, we can reduce the computation of $C(\lambda, r, a)$ to the case $r < 0$. The ratio of two successive terms in series (2.3) is:

$$\frac{\lambda}{k+1}\left(1+\frac{1}{k+a}\right)^r \leq \frac{\lambda}{k+1}.$$

Taking $k > 2\lambda$, the ratio of two successive terms is less than $1/2$, then the rest of the series is smaller than the k-term; that is, if $k_0 > 2\lambda$, then:

$$\sum_{k=k_0+1}^{\infty} \frac{\lambda^k (k+a)^r}{k!} \leq \frac{\lambda^{k_0}(k_0+a)^r}{k_0!} \leq \frac{\lambda^{k_0}}{k_0!}.$$

Using the Stirling formula, the right hand in the last expression can be dominated, if $k_0 > 3\lambda$, by

$$\frac{\lambda^{k_0}}{k_0!} \leq \frac{\lambda^{3\lambda}}{(3\lambda)!} \approx \frac{1}{\sqrt{6\pi\lambda}}\left(\frac{e}{3}\right)^{3\lambda} \approx 0.23\frac{(0.9)^{3\lambda}}{\sqrt{\lambda}}.$$

Consequently, in computing $C(\lambda, r, a)$ we use at least $k > 3\lambda$ terms.

The series $D^s(\lambda, r, a)$ also verifies the recurrent formula

$$D^s(\lambda, r+1, a) = \lambda D^s(\lambda, r, a+1) + aD^s(\lambda, r, a).$$

This fact allows us to compute $D^s(\lambda, r, a)$ in a similar way as we have explained above, to compute $C(\lambda, r, a)$. In fact, if $s > 0$, which is the important case, it is enough to apply the recurrence until $r < 0$ and $-r \geq s$ and to use at least $k > 3\lambda$ terms.

The Negative Binomial distribution with parameters p and r is defined by $P(k; p, r) = \binom{-r}{k} p^r (-1)^k (1-p)^k$, for $k = 0, 1, \ldots$. The values of p and r that appear in the Negative Binomial columns of Table 1 are the estimations by the moments method, given in the original references. Because we do not know the value of parameter a, we have studied the profile log-likelihood varying this parameter. For a given value of a, the solution of (4.2) has been computed using the methods of Subsection 4.2. In both cases, good starting points can be found solving equations (3.2) and (3.3) with $k = 0$ and the empirical probabilities instead of the theoretical ones. Finally, the parameter λ of the Poisson distribution has been estimated by the sample mean.

Example 1. Table 1 contains data given by Greenwood and Yule (1920). The data show the distribution of the number of accidents among 647 machine operators in a fixed period of time. The ratio of the sample variance to the sample mean is greater than one. This is an example with absence of repulsion, which is

Table 1. Number of accidents for machine operators. Greenwood and Yule (1920).

No. of accidents	Obs.	WPD $a = 0.8$ $\lambda = 2.16572$ $r = -2.47553$	WPD $a = 1$ $\lambda = 2.46942$ $r = -3.07412$	Neg. Bin. $p = 0.673427$ $r = 0.959349$	POISSON $\lambda = 0.465224$
0	447	447.158	446.874	442.768	406.312
1	132	130.084	131.032	138.717	189.026
2	42	47.182	46.517	44.380	43.969
3	21	15.993	15.813	14.296	6.818
4	3	4.856	4.916	4.621	0.793
5	2	1.725	1.848	2.218	0.082
	647	$\chi^2 = 2.917$	$\chi^2 = 2.906$	$\chi^2 = 4.227$	$\chi^2 = 101.871$

translated in a negative value for parameter r when we fit the data using a WPD. The data were proposed as an example of data following a Negative Binomial distribution rather than a Poisson distribution. Notwithstanding, the first and second columns of Table 1 indicate that the WPD fits better than the Negative Binomial. Not only is the value of $\mathcal{X}^2$ smaller but also most of the expected values are close to the observed ones.

The estimated parameters that appear in Table 1 were computed from the statistics:

$$t_1 = 0.46522, \quad \text{and} \quad s^2 = 0.6919$$

defined in Subsections 4.1 and 4.2 and, the corresponding value of t_2 depending on the value of a. Moreover, in order to test if the data came from a Poisson distribution, we have computed the statistic S_R defined in Subsection 4.3. This, for the Greenwood and Yule data, and for $a = 1$ is $S_R = -8.75^{**}$.

Figure 5.1 shows the behaviour of the log-likelihood function for the Greenwood and Yule data. In the x-coordinate we represent the value of parameter a and in the y-coordinate we represent the value of the profile log-likelihood $l_a(\hat{\lambda}_a, \hat{r}_a, a)$, according with (4.1), where $(\hat{\lambda}_a, \hat{r}_a)$ is the solution of (4.2), i.e., the maximum likelihood estimators of the two-parameter model. In this case the maximum is attached at $a = 0.8$ approximatively. It is possible to appreciate, in the first two columns of Table 1, that there are not big differences between two approximations corresponding to close values of parameter a.

Example 2. P. C. Consul (1989) proposed a generalized Poisson distribution defining a bi-parametrical model which is not an exponential family. The data in Table 2 were obtained from Consul but, as is mentioned in Consul (1989), they come from a well-known paper on natural laws in the social sciences written by Kendall (1961). The data correspond to the observed data on the number of outbreaks of strikes in 4-week periods, in a coal mining industry in the United Kingdom during 1948–1959. This is a repulsion case because the ratio of the sample variance to the sample mean is less than one. Consequently, $r > 0$ when

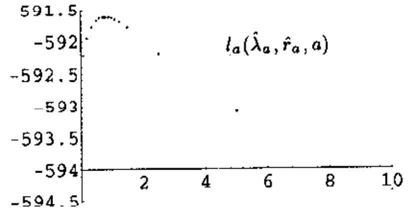

Fig. 5.1. Graph of the profile log-likelihood associated with the data of Table 1. In the x-coordinate we represent the value of parameter a, in the y-coordinate we represent the value of the log-likelihood in the point $(\hat\lambda_a, \hat r_a, a)$.

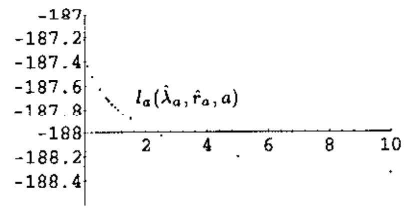

Fig. 5.2. Graph of the Profile-loglikelihood associated with the data of Table 2. In the x-coordinate we represent the value of parameter a, in the y-coordinate we represent the value of the log-likelihood in the point $(\hat\lambda_a, \hat r_a, a)$.

Table 2. Number of outbreaks of strikes. Kendall, M. G. (1961).

Outbr.	Obs.	GPD	WPD $a = 0.5$	WPD $a = 0.00001$	POISSON
		$\theta = 1.14$	$\lambda = 0.454975$	$\lambda = 0.70107$	$\lambda = 0.99359$
		$\lambda = -0.15$	$r = 1.12$	$r = 0.07268$	
0	46	50.01	46.607	46.030	57.758
1	76	66.77	72.572	74.514	57.388
2	24	31.23	29.252	27.469	28.510
3	9	7.23	6.466	6.611	9.442
≥ 4	1	0.76	1.103	1.374	2.902
	156	$\chi^2 = 4.520$	$\chi^2 = 2.115$	$\chi^2 = 1.432$	$\chi^2 = 5.057$

we fit the data with a WPD. In this case, the generalized Poisson distribution of Consul fits better than the Poisson distribution. As in Example 1, the WPD not only gives the smallest value for $\mathcal{X}^2$ but also gives expected values very close to those observed.

The estimated parameters of Table 2 were computed from the statistics

$$t_1 = 0.99359 \qquad \text{and} \qquad s^2 = 0.741894$$

defined in Subsections 4.1 and 4.2, and the corresponding value of t_2 depending on the value of parameter a. In order to test if the data come from a Poisson distribution we have computed the S_R statistic from Kendall's data and $a = 0.00001$ and it is $S_R = 2.23^{**}$.

The main difference between this example and the previous one is that the smaller the value of parameter a, the larger the value of the log-likelihood function. Nevertheless, the maximum is not going to lie in the boundary, that is $a = 0$, because $p(0; \lambda, r, 0) = 0$ and, from the data it is possible to see that $k = 0$ has a probability different from zero. In fact, it seems that $\hat{r}_a$ tends to zero when a tends to zero and in this case, it is not easy to specify the limit distribution, because it depends on the rates of convergence to zero.

Acknowledgements

We would like to thank the Associate Editor and referees for their helpful comments and suggestions.

Appendix. Proof of the theorems

PROOF OF THEOREM 2.1. We prove the theorem by distinguishing the sign of the repulsion parameter r.

Case (i) $r = 0$. For this case X has a Poisson distribution. In consequence, the mean and the variance are equal and therefore $I(X) = 1$.

Case (ii) $r < 0$. For this case we prove that for $\lambda > 0$ and $a > 0$

$$C^2(\lambda, r, a+1) < C(\lambda, r, a)C(\lambda, r, a+2), \tag{A.1}$$

which is equivalent to $I(X) > 1$, as we saw at (2.9). Applying Fubini's theorem, (A.1) is the same as:

$$\sum_{k_1,k_2=0}^{\infty} \frac{\lambda^k (k_1+a+1)^r (k_2+a+1)^r}{k_1!k_2!} < \sum_{k_1,k_2=0}^{\infty} \frac{\lambda^k (k_1+a)^r (k_2+a+2)^r}{k_1!k_2!},$$

where $k = k_1 + k_2$. By symmetry, the last inequality is equivalent to:

$$\sum_{k_1,k_2} \frac{\lambda^k 2(k_1+a+1)^r (k_2+a+1)^r}{k_1!k_2!} < \sum_{k_1,k_2} \frac{\lambda^k [(k_1+a)^r (k_2+a+2)^r + (k_2+a)^r (k_1+a+2)^r]}{k_1!k_2!}. \tag{A.2}$$

Moreover, it is easy to see that:

$$\frac{(k_1+a)(k_2+a+2) + (k_2+a)(k_1+a+2)}{2(k_1+a+1)(k_2+a+1)} < 1$$

and since $r < 0$, $\phi_r(x) = x^r$ is a decreasing function, then:

$$(k_1+a+1)^r (k_2+a+1)^r < \left(\frac{(k_1+a)(k_2+a+2) + (k_2+a)(k_1+a+2)}{2}\right)^r.$$

Taking into account that $\phi_r(x)$ is also a convex function in $\Re^1$ for any negative value of r, we have:

$$2(k_1+a+1)^r(k_2+a+1)^r < (k_1+a)^r(k_2+a+2)^r + (k_2+a)^r(k_1+a+2)^r,$$

and consequently, (A.1) is proved.

Case (iii) $r > 0$. We distinguish if r is larger or smaller than one, in order to prove the opposite of (A.1). If $0 < r \leq 1$, the proof is similar to case (ii). It is only necessary to observe that now the function $\phi_r(x) = x^r$ is increasing and concave in $\Re^+$, and to change the inequalities. To prove the opposite to (A.1) for values or r larger than one, it is enough to see that if for any $a > 0$, $\lambda > 0$ and given a fixed value of r:

$$C(\lambda,r,a)C(\lambda,r,a+2) < C^2(\lambda,r,a+1), \tag{A.3}$$

then (A.3) is also verified by $r+1$, which is

$$C(\lambda,r+1,a)C(\lambda,r+1,a+2) < C^2(\lambda,r+1,a+1). \tag{A.4}$$

Writing hypothesis (A.3) for $a+1$ we have:

$$C(\lambda,r,a+1)C(\lambda,r,a+3) < C^2(\lambda,r,a+2). \tag{A.5}$$

Multiplying (A.3) and (A.5) we obtain:

$$C(\lambda,r,a)C(\lambda,r,a+3) < C(\lambda,r,a+1)C(\lambda,r,a+2). \tag{A.6}$$

Applying (2.5), (A.4) becomes:

$$\begin{aligned}
&\lambda^2 C(\lambda,r,a+1)C(\lambda,r,a+3)\\
&\qquad + \lambda[(a+2)C(\lambda,r,a+2)C(\lambda,r,a+1) + aC(\lambda,r,a+3)C(\lambda,r,a)]\\
&\qquad + a(a+2)C(\lambda,r,a)C(\lambda,r,a+2)\\
&\quad < \lambda^2 C^2(\lambda,r,a+2) + \lambda[2(a+1)C(\lambda,r,a+2)C(\lambda,r,a+1)]\\
&\qquad + (a+1)^2 C^2(\lambda,r,a+1).
\end{aligned}$$

By (A.5) and (A.6) the coefficients of λ^2 and λ on the left hand side are smaller than the respective coefficients on the right hand side. By (A.3), the independent term on the left hand side is also smaller than the independent term on the right hand side. Consequently, the proof is complete. □

References

Barndorff-Nielsen, O. (1978). *Information and Exponential Families in Statistical Theory*, Wiley, Norwich.

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families*, Lecture Notes-Monograph Series, Institute of Mathematical Statistics, California.

Castillo, J. (1994). The singly truncated Normal distribution a non-steep exponential family, *Ann. Inst. Statist. Math.*, **46** (1), 57–66.

Consul, P. C. (1989). *Generalized Poisson Distributions*, Marcel Dekker, New York.

Efron, B. (1978). The geometry of exponential families, *Ann. Statist.*, **6**, 362–376.

Greenwood, M. and Yule, U. (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurence of multiple attacks of disease or of repeated accidents, *J. Roy. Statist. Soc. Ser. A*, **83**, 255–279.

Haight, Frank A. (1967). *Handbook of the Poisson Distribution*, Wiley, New York.

Johnson, N. L., Kotz, S. and Kemp, A. W. (1992). *Univariate Discrete Distributions*, Wiley, New York.

Kendall, M. G. (1961). Natural law in social sciences, *J. Roy. Statist. Soc. Ser. A*, **124**, 1–19.

Kendall, M. and Stuart, A. (1979). *The Advanced Theory of Statistics*, Macmillan, New York.

Letac, G. (1992). *Lectures on Natural Exponential Families and Their Variance Functions*, Instituto de Matemática Pura y Aplicada, Rio de Janeiro.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed., University Press, Cambridge.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models, *J. Roy. Statis. Soc. Ser. A*, **135**, 370–384.

Patil, G. P. and Rao, C. R. (1978). Weighted distributions ans size-biased sampling with applications to wildlife populations and human families, *Biometrics*, **34**, 179–189.

Patil, G. P., Rao, C. R. and Ratnaparkhi, M. V. (1986). On discrete weighted distributions and their use in model for observed data, *Communication in Statistics Theory and Methods*, **15** (3), 907–918.

Rao, C. R. (1965). On discrete distributions arising out of methods of ascertainment, *Sankhyā Ser. A*, 311–324.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*, Wiley, New York.

Ross, S. M. (1983). *Stochastic Processes*, Wiley, New York.

Wolfram, S. (1991). *Mathematica. A System for Doing Mathematics by Computer*, Addison-Wesley, New York.