econometrics

MDPI

Article

Distributions You Can Count On . . . But What's the Point? †

Brendan P. M. McCabe [1] and Christopher L. Skeels [2,*]

[1] School of Management, University of Liverpoo, Liverpool L69 7ZH, UK; Brendan.McCabe@liverpool.ac.uk
[2] Department of Economics, The University of Melbourne, Carlton VIC 3053, Australia
* Correspondence: Chris.Skeels@unimelb.edu.au

† An early version of this paper was originally prepared for the Fest in Celebration of the 65th Birthday of Professor Maxwell King, hosted by Monash University. It was started while Skeels was visiting the Department of Economics at the University of Bristol. He would like to thank them for their hospitality and, in particular, Ken Binmore for a very helpful discussion. We would also like to thank David Dickson and David Harris for useful comments along the way.

Received: 2 September 2019; Accepted: 25 February 2020; Published: 4 March 2020

Abstract: The Poisson regression model remains an important tool in the econometric analysis of count data. In a pioneering contribution to the econometric analysis of such models, Lung-Fei Lee presented a specification test for a Poisson model against a broad class of discrete distributions sometimes called the Katz family. Two members of this alternative class are the binomial and negative binomial distributions, which are commonly used with count data to allow for under- and over-dispersion, respectively. In this paper we explore the structure of other distributions within the class and their suitability as alternatives to the Poisson model. Potential difficulties with the Katz likelihood leads us to investigate a class of point optimal tests of the Poisson assumption against the alternative of over-dispersion in both the regression and intercept only cases. In a simulation study, we compare score tests of 'Poisson-ness' with various point optimal tests, based on the Katz family, and conclude that it is possible to choose a point optimal test which is better in the intercept only case, although the nuisance parameters arising in the regression case are problematic. One possible cause is poor choice of the point at which to optimize. Consequently, we explore the use of Hellinger distance to aid this choice. Ultimately we conclude that score tests remain the most practical approach to testing for over-dispersion in this context.

Keywords: Katz family of distributions; binomial distribution; negative binomial distribution; point optimal test; regression; score test; Hellinger distance

JEL Classification: C12; C25; C46

1. Introduction

The well-known Pearson family of continuous distributions, originally explored by Pearson (1895), is comprised of any solution to a particular differential equation. In his PhD thesis, Katz (1945) explored a family of discrete distributions that are solutions to a difference equation analogous to the Pearson differential equation.[1] The Pearson family is a collection of four-parameter distributions and specializations

[1] An abstract to this thesis appeared in Katz (1946).

thereof. Katz (1965, p. 175) observes that certain specializations 'produce simpler and more manageable classes' and restricts attention to a set of one- and two-parameter distributions. In particular, his restrictions result in a family of distributions that nest the two-parameter binomial and negative binomial (or Pascal) distributions, together with the one-parameter Poisson distribution.[2] A defining characteristic of these distributions is that they arise when certain parameters, or parameter ratios, take integer values and so represent a set of measure zero in respect of the set of family members, which are defined in terms of real-valued parameters. The Katz family of distributions has proved important in the analysis of count data. It provides a framework within which practitioners can extend simple Poisson models to models that allow for individual heterogeneity, using the Poisson regression model (PRM). The PRM can, in turn, be extended to models that allow for either over-dispersion, using the negative binomial regression model (NBRM), or under-dispersion, using the binomial regression model. We shall, for the most part, defer consideration of under-dispersion to another time.

The problems of modelling and testing for over-dispersion have proved important in the count data literature. Essentially concurrently, papers by Cameron and Trivedi (1986); Lee (1986) and Lawless (1987ab) made substantial contributions to the literature on inference in the PRM, the NBRM, and testing for over-dispersion, with both Cameron and Trivedi (1986) and Lee (1986), in particular, couching substantial parts of their analysis within the context of the Katz family of distributions. This class of distributions is interesting because the binomial and negative binomial distributions are alternative specifications to the Poisson that allow under- and over-dispersion, respectively. Subsequent contributions to this literature include Dean (1992); Dean and Lawless (1989); Qu et al. (1990) and Fang (2003). Collectively they have explored likelihood ratio (LR), Lagrange multiplier (LM), and Wald tests, together with tests based on generalised method of moments (GMM) estimators, for over-dispersion in the PRM. Fang (2003) concludes that his preferred GMM test is that based on the fewest over-identifying assumptions offering essentially the same power as tests based on more over-identifying restrictions but having the greatest ease of calculation.[3] Interestingly, this preferred test is that originally proposed by Katz (1965), on an ad hoc basis.

In this paper, we investigate a new family of tests for over-dispersion in the PRM by exploring point optimal tests where the alternative hypothesis lies in the Katz family of distributions. An analysis of the Katz likelihood reveals that maximum likelihood estimation may be problematic in the over-dispersed case, suggesting that the use of point optimal tests may have value. For overviews of the use of point optimal tests in econometrics see King (1987) and King and Srianantha kumar (2015). To the best of our knowledge they have not previously been used in the context of testing for over-dispersion.

This paper can be thought of as being comprised of three main parts. The first part provides a very brief description of the family of distributions introduced by Katz (1965), the second explores the role that these distributions can play in extending the PRM to allow for over-dispersion and, finally, we introduce a new class of point-optimal tests for over-dispersion. Specifically, in Section 2 we explore the Katz family of distributions, although most of the analysis is relegated to the Appendix A while Section 3 explores the PRM and NBRM. In particular, we highlight that the typical treatments of the NBRM really have little to do with what might be thought of as the canonical negative binomial distribution. Section 4 then focuses on the problem of testing for over-dispersion and the structure of the Katz likelihood. It is here that we

---

[2] This family of distributions, and extensions to it, have proved important in the actuarial modelling of claims; see, for example, Hess et al. (2002); Panjer (1981); Sundt and Jewell (1981); Willmot (1988), and Pestana and Velosa (2004). Johnson et al. (1993, chp. 2) provides an extensive discussion of both the Katz family and various other, often related, families of discrete distributions. Although, in respect of the Katz family of distributions alone, the treatment in Johnson and Kotz (1969, chp. 2.4) is more complete; see also Gurland (2006) for a more recent treatment.

[3] The one caveat to this observation is that the use of higher order moments may provide some power against models which share low order moments, thereby creating a class of implicit null hypotheses (Davidson and MacKinnon 1987).

introduce our family of point optimal tests and explore their small sample characteristics relative to some existing tests via a simulation study. We find that it is possible to choose a point optimal test which is better in the intercept only case, although the regression case proves problematic. One possible source of weakness in our point optimal tests is the choice of 'point' at which to optimize. In Section 5 we explore the use of Hellinger distance as a device to assist in choice of point. Although exact calculation of the Hellinger distance in this context is not tractable, it is straight-forward to obtain bounds on the distance. Using the upper bound, we find that the implied optimal points are extremely close to zero, implying that use of the score test is close to the optimal strategy in this context and so our advice to practitioners is to continue to use score tests to test for over-dispersion in this context. Section 6 concludes.

2. The Katz Family of Distributions

Among his many and varied interests, Karl Pearson was concerned with the problem of modelling (possibly) asymmetric empirical distributions. To this end, he developed a four-parameter family of skewed continuous distributions as solutions to a particular differential equation (Pearson 1895). The idea being that the distributions might be fitted to any data set using the method of moments approach that he had developed earlier (Pearson 1894). Perhaps surprisingly, the motivation for the choice of differential equation came from a difference equation that could be used to generate the hypergeometric distribution (Pearson 1895, pp.360–361); that is, from a discrete distribution. If we let $p(y) \equiv P[Y = y]$ denote the probability that the discrete random variable Y takes a value $y \in \mathbb{Y}$, where $\mathbb{Y}$ denotes the support of the distribution of Y, then the form of this difference equation was

$$\frac{p(y) - p(y-1)}{p(y-1)} = \frac{a - y}{b_0 + b_1 y + b_2 y^2}, \tag{1}$$

with a, b_0, b_1, and b_2 denoting the various parameters of the distribution. We note that this expression is of the form

$$p(y)/p(y-1) = P(y)/Q(y),$$

where P and Q are polynomials in y, and remark in passing that the sequence of probabilities so-defined are hypergeometric in that the ratio of adjacent terms in the sequence can be expressed as a ratio of polynomials in the index y.

Pearson did not pursue a discrete analogue to his family of distributions. Indeed, apart from some incidental investigations along these lines, Katz (1945) provided the first detailed analysis of the family of distributions arising from (1) although, apart from some abstracts (Katz 1946, 1948), it was not until Katz (1965) that this material was published. In the event, Katz (1965) focussed on a two-parameter special case of (1),[4] which he expressed in the form

$$\frac{p(y+1)}{p(y)} = \frac{\lambda + y\gamma}{y+1}, \quad \lambda \in \Lambda \subseteq \mathbb{R},\ \gamma \in \Gamma \subseteq \mathbb{R}, \tag{2}$$

[4] Numerous extensions soon followed; see, for example, Bardwell and Crow (1964); Crow and Bardwell (1965); Ord (1967a, 1967b); Staff (1964, 1967) and Kemp (1968). Here we only briefly sketch some key ideas. For a more complete treatment of such families of distributions see, for example, any of Johnson et al. (1993, chp. 2.3), Ord (1972, chp. 5), or Dacey (1972).

with $y \in \mathbb{Y} \subseteq \mathbb{Z}_0$, where $\mathbb{Z}_0$ denotes the set of non-negative integers and subject to the usual axiomatic properties of probability:

$$0 \leq p(y) \leq 1 \tag{3}$$

$$\sum_{y \in \mathbb{Y}} p(y) = 1. \tag{4}$$

As we demonstrate in the Appendix A, there are circumstances where both Λ and Γ may include values that are positive, negative, or zero.

Although Katz himself included zero in $\mathcal{Y}$, subsequent literature has not always done so, choosing instead to focus on the difference equation (2), whilst still referring to the resulting distributions as members of the Katz family; see, for example, Sundt and Jewell (1981); Willmot (1988) and Miller (1998). We too shall proceed in this latter manner, focussing on the what we call left-truncated Katz distributions, that include the original definition of Katz (1965) as the special case of no left-truncation. We relegate the technical analysis of these distributions to the Appendix A, which also gathers a number of other properties of this family of distributions.

Two members of this family that will be of particular interest to us are the Poisson and negative binomial distributions, which are commonly encountered in the modelling of counts and the possibility of over-dispersion. The probability mass functions (pmfs) of these distributions are the form:

$$p(y) = \begin{cases} \dfrac{\lambda^y e^{-\lambda}}{y!}, & \gamma = 0, \\ \dfrac{(1-\gamma)^{\lambda/\gamma} \gamma^y \left(\frac{\lambda}{\gamma}\right)_y}{y!}, & 0 < \gamma < 1, \end{cases} \qquad y \in \{0, 1, 2, \ldots\}, \tag{5}$$

respectively.[5] Evidently, when $\gamma = 0$, $p(y)$ is the pmf of the Poisson distribution. When $\gamma > 0$, if $\lambda/\gamma = r$ is integer then $p(y)$ yields a standard representation of the negative binomial pmf, where the probability of success in any given trial is $\pi = 1 - \gamma$. Even if $\lambda/\gamma = \tau$ is not integer, $p(y)$ is still the pmf of a negative binomial distribution — see, for example, (9) — although the interpretation of λ/γ differs between the two cases.[6] We shall, hereafter, denote the Poisson distribution with parameter λ, $\mathcal{P}(\lambda)$, and the negative binomial with parameters τ and π, $\mathcal{NB}(\tau, \pi)$.

Before moving on, let us consider the well-known Poisson approximation to the negative binomial. A common statement of this result is $\mathcal{NB}(\tau, \pi) \to \mathcal{P}(\lambda)$ as $\tau \to \infty$ provided $\lambda = \tau(1 - \pi)$ remains fixed. That is, $\pi \to 1$ at the same rate as τ diverges. One advantage of the parameterization adopted in (5) is that the somewhat convoluted requirement on how the parameters evolve in the approximation readily reduces to $\gamma \to 0^+$ for fixed λ.[7]

---

[5] Observe that the Pochhammer symbol $(r)_y = \Gamma(y + r)/\Gamma(r)$, where y is a non-negative integer. Note that r can be negative. If r is a negative integer then $(r)_y = 0$ for all $y > r$. If r is a positive integer then $(r)_y = (y + r - 1)!/(r - 1)!$.

[6] When λ/γ is integer the resulting pmfs are sometimes referred to as those of Pascal distributions, with the term negative binomial reserved for the more general case of λ/γ not necessarily integer.

[7] Similarly, the Poisson approximation to the Binomial reduces to $\gamma \to 0^-$ for fixed λ, which is also a more intuitive statement of how parameters must evolve for the approximation to work than is typically encountered.

3. The Poisson, Negative Binomial, and Katz Regression Models

3.1. The Poisson Regression Model

The PRM extends the Poisson distribution to allow for individual heterogeneity. It has played an important role in the analysis of count data in both econometrics and statistics — early references include Gart (1964); Jorgenson (1961), and Haight (1967, chp. 5) — and is readily available in standard software such as MATLAB, Stata, and R. The use of the PRM in econometrics became increasingly widespread following the significant contributions of Gilbert (1979, 1982) and Hausman et al. (1984). Recent summaries can be found in Greene (2007); Winkelmann (2008) and Cameron and Trivedi (2013).

The PRM is obtained from the Poisson distribution by replacing the fixed parameter λ with a function, denoted λ_i say, of the k-vector of characteristics x_i that can vary across individuals. Specifically, in the language of generalized linear models (GLIMs), we have the link function

$$\ln \lambda_i = x_i^\top \beta, \tag{6}$$

with regression coefficients β. The work of Nelder and Wedderburn (1972) and Frome et al. (1973) shows how iterated least squares methods can be used to obtain maximum likelihood estimates of β; see also McCullagh and Nelder (1989).

One shortcoming of the PRM is the implied equality of mean and variance that is characteristic of the Poisson distribution. Specifically, on replacing λ with λ_i in (A8), we obtain[8]

$$\mathrm{E}\left[Y_i \mid \lambda_i\right] = \mathrm{V}\left[Y_i \mid \lambda_i\right] = \lambda_i. \tag{7}$$

This is at odds with the observation that variability typically exceeds location in real world data, a feature known as over-dispersion. A common response to concerns about over-dispersion has been to explore extensions to the Poisson model that allow for different means and variances. To the extent that the Poisson regression model can be nested in such generalizations, this approach provides a framework within which one might test for either over-dispersion or underdispersion, although we will not explore this latter case here.

The fundamental characteristic of the PRM is that it is a function of the linear index, $x_i^\top \beta$, only through the 'parameter' λ_i, as per (7). In the next two sub-sections we will consider different extensions to this model, the first being the classical NBRM and the second being what we dub the Katz regression model (KRM). Both models extend the PRM by nesting it within a richer model with an additional 'parameter'. An important distinction between the NBRM and the KRM is the role of λ_i. In the case of the NBRM, λ_i remains the conditional mean of the count Y_i, whereas this is not the case in the KRM. A second distinction between the models is that the additional 'parameter' is typically treated as being a function of the linear index in the NBRM whereas in our treatment of the KRM it is not, it is a genuine parameter, although it is easy to envisage extensions where that requirement is relaxed.

---

[8] We shall persist with the abuse of notation inherent in expressions like $\mathrm{E}\left[Y_i \mid \lambda_i\right]$ rather than, say, a more complete notation along the lines of $\mathrm{E}\left[Y_i \mid \beta; x_i\right]$, for the sake of the notational economy it affords.

3.2. The Classical Negative Binomial Regression Model

There are numerous paths leading to what might reasonably be called a negative binomial regression model.[9] This is due, at least in part, to the variety of ways in which one might generate a negative binomial distribution. For example, Boswell and Patil (1970) provide 15 different derivations and, of course, there is a variety of parameterizations of the negative binomial distribution that can also lead to differences. Below we explore a fairly commonly adopted approach and consider some of its implications.

Our starting point is the following observation, originally due to Greenwood and Yule (1920). Suppose that $Y \mid \theta \sim \mathcal{P}(\theta)$, where θ is a random variable whose distribution is gamma with *shape* (τ) and *rate* (η) parameters, written $\theta \sim \mathcal{G}(\eta, \tau)$, so that the corresponding density function is,[10]

$$g(\theta; \eta, \tau) = \eta^{\tau} \theta^{\tau-1} \exp\{-\theta\eta\} / \Gamma(\tau), \quad \tau > 0, \eta > 0, \theta > 0, \tag{8}$$

with $\mathrm{E}[\theta] = \tau/\eta$ and $\mathrm{V}[\theta] = \tau/\eta^2$.[11] Then, we obtain an unconditional distribution for Y on averaging with respect to $\theta > 0$, so that

$$\begin{aligned} \mathrm{Prob}(Y = y \mid \eta, \tau) &= \int_{\theta>0} f(y \mid \theta)\, g(\theta; \eta, \tau)\, \mathrm{d}\,\theta \\ &= \frac{\Gamma(y+\tau)}{y!\,\Gamma(\tau)} \left(\frac{1}{1+\eta}\right)^{y} \left(\frac{\eta}{1+\eta}\right)^{\tau}. \end{aligned} \tag{9}$$

If one imposes the restriction $\tau \equiv r \in \mathbb{N}^{+}$, where $\mathbb{N}^{+}$ denotes the set of positive integers (or the natural numbers), then this is simply a form of the negative binomial (Pascal) pmf, with $\pi = \eta/(1+\eta)$, see (A5). Note that

$$\mathrm{E}[Y] = \tau/\eta = \lambda \text{ (say)},$$

the same as for the gamma distribution (8), and that

$$\mathrm{V}[Y] = \frac{\tau}{\eta} + \frac{\tau}{\eta^2} = \mathrm{E}[Y] + \tau^{-1}(\mathrm{E}[Y])^2 = \lambda + \tau^{-1}\lambda^2.$$

One posible path to a NBRM is to extend the analysis of Greenwood and Yule (1920) to allow for individual heterogeneity; we follow the treatment of Cameron and Trivedi (1986, p. 32). Specifically, we replace θ by θ_i, where

$$\ln \theta_i = x_i^{\top} \beta + \epsilon_i, \tag{10}$$

with ϵ_i a disturbance term reflecting unobservables. Cameron and Trivedi (1986) then assume that either ϵ_i, or 'equivalently' θ_i, have a gamma distribution, conditional on the regressors. Their analysis then proceeds under the latter assumption, which is completely analogous to the developments of Greenwood and Yule (1920). Specifically, letting $\theta_i \mid x_i \sim \mathcal{G}(\eta_i, \tau_i)$ yields

---

[9] The NBRM was explored in Adamidis (1999); Greene (2008); Lawless (1987a, 1987b), and Raschke and Greene (2010); Hilbe (2011) and Hilbe (2014) provide useful recent surveys of the NBRM.

[10] Common variants of this argument include: (i) Lee (1986), who specifies the gamma distribution in terms of the shape and *scale* (or *inverse rate*) ($\xi = 1/\eta$) parameters, that is, $\theta \sim \mathcal{G}(1/\xi, \tau)$, and (ii) Cameron and Trivedi (1986), who use the so-called index form of the gamma distribution, which is specified in terms of the shape and *mean* ($\phi = \tau/\eta$) parameters, that is, $\theta \sim \mathcal{G}(\tau/\phi, \tau)$. Cameron and Trivedi (1986) call the shape parameter (τ) the *index* or *precision* parameter.

[11] Moments for the gamma distribution specifications given in Footnote 10 follow immediately on making the appropriate substitution for η.

$$\text{Prob}\left(Y_i = y \mid x_i; \eta_i, \tau_i\right) = \frac{\Gamma(y+\tau_i)}{y!\,\Gamma(\tau_i)} \left(\frac{1}{1+\eta_i}\right)^{y} \left(\frac{\eta_i}{1+\eta_i}\right)^{\tau_i}. \tag{11}$$

Moreover,

$$\text{E}\left[Y_i \mid x_i; \eta_i, \tau_i\right] = \frac{\tau_i}{\eta_i} = \lambda_i \text{ (say)}, \tag{12}$$

and

$$\text{V}\left[Y_i \mid x_i; \eta_i, \tau_i\right] = \frac{\tau_i}{\eta_i} + \frac{\tau_i}{\eta_i^2} = \lambda_i + \tau_i^{-1}\lambda_i^2. \tag{13}$$

It is immediately obvious that, in the final analysis, the functional form of (10) is a complete irrelevance, with only the parameters of the mixing Gamma distribution of any importance and we have made no assumptions about them beyond allowing the possibility of varying at the individual level. From here, Cameron and Trivedi (1986) argue that a variety of models are available on defining

$$\tau_i = \alpha^{-1}(\text{E}\left[Y_i \mid x_i\right]; \eta_i, \tau_i)^k \tag{14}$$

for $\alpha > 0$ and arbitrary constant k, so that

$$\text{V}\left[Y_i \mid x_i; \eta_i, \tau_i\right] = \text{E}\left[Y_i \mid x_i; \eta_i, \tau_i\right] + \alpha(\text{E}\left[Y_i \mid x_i; \eta_i, \tau_i\right])^{2-k}.$$

Special cases of importance are then the Negbin I model (obtained when $k = 1$) and the Negbin II model ($k = 0$),[12] of which the latter is probably the more popular in the literature. This model nests the PRM as a limiting case where $\alpha \to 0$ from above, a testable proposition, which is equivalent to τ_i diverging to ∞ for all i.

The specification (10) becomes more relevant if, instead, we assume that $h_i = e^{\epsilon_i} \mid x_i \sim \mathcal{G}(\eta_i, \tau_i)$ rather than θ_i. Define $\delta_i = \exp\{x_i^\top \beta\}$, so that $\theta_i = \delta_i h_i$. Thus, conditional on x_i, θ_i is a scaled Gamma random variate which is, itself, a Gamma random variate. From the properties of the Gamma distribution we have immediately that $\theta_i | x_i \sim \mathcal{G}(\eta_i/\delta_i, \tau_i)$. Moreover, analogs of results (11)–(13) are immediately available in this case on replacing η_i by η_i/δ_i. In short, the differing distributional assumptions are 'equivalent' in that the structure of the results is the same in both cases, however, it is only in this latter case that (10) has any relevance, through the presence of δ_i in the various expressions.

The attraction of the formulation (11)–(13) of Cameron and Trivedi (1986) is its close resemblance to a GLIM, which simplifies estimation.[13] As noted above, the null of a PRM obtains as $\tau_i \to \infty$, however, results in a relatively odd PRM with a potentially unbounded mean, unless η_i is diverging to infinity at the same rate as is τ_i. Moreover, there is a Davies-type problem relating to the separate identification of both τ and η when the null is true. Greene (2008) camouflages this difference by imposing the restriction $\eta = \tau$.[14] He refers to this restriction as being mean preserving, by which is meant that when $\eta = \tau$, $\text{E}\left[Y_i \mid x_i; \eta, \tau\right] = \delta_i$, as would be the case in the PRM. We should note that, in order to generate the same class of models as do Cameron and Trivedi (1986), Greene (2008) also allows τ to be replaced by τ_i,

---

[12] Other values of k yield the Negbin P, or NBP, model (Greene 2008).

[13] Strictly, it is not a generalized linear model as it stands but, conditioning on one of the parameters allows it to be treated so. This parameter can then be estimated conditional on the remaining parameters, which yields a two-step iterative estimation procedure. See, for example, either Hilbe (2011) or Hilbe (2014) for a discussion of the steps involved.

[14] This latter model, of course, corresponds to the Negbin II model of Cameron and Trivedi (1986), and so provides a somewhat stronger theoretical basis for that model, which may explain some of its popularity in the literature.

as defined by (14), but this means that η must be replaced by τ_i too.[15] Of course, the restriction that $\tau_i = \eta_i$ reduces the two parameter mixing gamma distribution to a single parameter distribution with the loss of modelling flexibility that implies. However, without this restriction, the conditional mean of Y_i is other than δ_i.

3.3. The Katz Regression Model

The fundamental difference between the NBRM, as described in the above, and what we refer to as the Katz regression model (KRM) lies in the generation of the underlying distribution. Specifically, the Katz family of distributions is not generated via a mixing argument and so, in contrast to the NBRM, the probabilistic quantities of interest (pmfs and moments) are not functions of the parameters of the mixing distribution; see (11)–(13). In this sense, the parameterization of the Katz family is more natural than that of the NBRM. Directly analogously with the PRM, the KRM can be generated from (A6) simply by replacing λ by λ_i, as per (6), which is analogous to our earlier development of the PRM.[16] Equally, one might explore models that see γ replaced by functions of regressors, γ_i say, although we will not. Note that the conditional mean and variance of this distribution are given by (A8), with λ replaced by λ_i. Contrast this structure with that for the NBRM described above. There we saw that the conditional mean of the dependent variable was not varying with γ, being a function of the linear index $x_i^\top \beta$ alone. Similarly, by construction, the variance exceeded the mean of the dependent variable, but the reduction to a Poisson model requires the shape parameter τ of the mixing gamma distribution to be unbounded, which yields a degenerate distribution for given rate parameter η.

It is clear that it is not necessarily desirable to preserve the mean, in Greene's sense of equating τ_i and η_i (Greene 2008), because, as γ increases, the mean for both the NBRM and the KRM should be decreasing relative to that of the PRM.

We note in passing that this is the model that underlies the generalized event count (GEC) model of King (1989); this model was also considered by Ghahfarokhi et al. (2008). That they obtained more complicated models than that proposed here, resulting in the models being less popular than the NBRM in practice, stems from the fact that they did not have (A6) as the pmf implied by (2), which in part is due to working with (2) rather than (A1).

---

[15] Specifically, Greene (2008) discusses the broader class of models obtained when k is allowed to take values other than 0 or 1 in (14). He dubs this broad model the NBP model, seemingly because his notation uses p rather than the k used by Cameron and Trivedi (1986) (and here).

[16] Alternatively, using similar averaging arguments to those seen previously for the NBRM, if we average $\mathcal{P}(\theta)$ with respect to $\mathcal{G}\left(\theta; \frac{\pi}{1-\pi}, n\right)$, where $\pi = 1-\gamma$ and $n = \lambda/\gamma$, then we obtain a more common form of the negative binomial pmf.

$$\begin{aligned}\text{Prob}\left(Y = y \mid \lambda, \gamma\right) &= \int_{\theta>0} \mathcal{P}(\theta)\, g\left(\theta; \frac{1-\gamma}{\gamma}, \frac{\lambda}{\gamma}\right) \mathrm{d}\,\theta, \quad \lambda > 0,\ 0 < \gamma < 1 \\ &= \frac{(1-\gamma)^{\lambda/\gamma}}{y!\,\gamma^{\lambda/\gamma}\,\Gamma(\lambda/\gamma)} \int_{\theta>0} \theta^{y+\lambda/\gamma-1} e^{-\theta/\gamma}\, \mathrm{d}\,\theta \\ &= \frac{(1-\gamma)^{\lambda/\gamma}\,\gamma^{y}\ (\lambda/\gamma)_y}{y!}.\end{aligned}$$

Note that the mean and variance of this distribution are given by (A8). In contrast with the developments of (11), there is nothing in this model that requires that both the parameters of the mixing gamma distribution vary with the index i. Nor need they be linked in any restrictive way. Specifically, if we were to follow the developments of Greene (2008) who equates the parameters of the mixing distribution, we find that

$$\frac{1-\gamma}{\gamma} = \frac{\lambda}{\gamma} \implies \lambda = 1-\gamma,$$

which constrains $0 < \lambda < 1$ and, as $\lambda = \exp\{x_i^\top \beta\}$, this implies that $x_i^\top \beta < 0$. As a general statement, this would appear to be a very odd restriction to want to impose.

4. Testing for Over-dispersion in Poisson Regression model

There is a vast literature addressing the problem of over-dispersion and how to test for it. We will not attempt to provide a comprehensive survey of this literature, focussing instead on a few key contributions, although it should be noted that most of the references cited so far will have some discussion of the problem. We shall break our discussion into two parts. First, we shall restrict attention to the case where the only regressor in the model is an intercept, so that $\lambda_i = \lambda$ is a constant. Then we will extend the analysis to allow for additional regressors. In each case, the null hypothesis will be that the data have been generated by the PRM. We investigate the performance of tests whose preferred alternative is, variously, that the data have come from one of the Negbin I, Negbin II, or Katz regression models.[17]

4.1. The Katz Likelihood

For any positive real number $n = \lambda/\gamma$, with $\lambda > 0$ and $0 < \gamma < 1$, the Katz pmf is given by

$$\text{Prob}\,(Y = k) = n(n+1)(n+2)\ldots(n+k-1)\frac{\gamma^k(1-\gamma)^n}{k!}$$

and hence

$$\begin{aligned} L\,(y_1,\ldots,y_N \mid \lambda,\gamma) &= \prod_{i=1}^{N}\left[n(n+1)(n+2)\ldots(n+y_i-1)\frac{\gamma^{y_i}(1-\gamma)^n}{y_i!}\right] \\ &= \left[\prod_{i=1}^{N}\prod_{s=0}^{y_i-1}(n+s)\right](1-\gamma)^{Nn}\,\gamma^{\sum_{i=1}^{N}y_i}\left[\prod_{i=1}^{N}y_i!\right]^{-1}, \end{aligned} \tag{15}$$

where products of the form $\prod_{s=0}^{y_i-1} = 1$ for $y_i = 0$. In textbook cases, where n is *known*, the first and last terms in (15) are functions of the data only and $\sum_{i=1}^{N} y_i$ is sufficient by the factorization theorem. But in the current context, when n is not known, there is no reduction to a fixed dimensional sufficient statistic. Even if λ is known there is no sufficiency reduction; the ratio n is required. Only the entire sample (or the order statistics) are sufficient but even they are are not complete so that different parameter configurations may give rise to the same data. We may surmise this from the likelihood (15) since any combination of λ and γ that preserves n gives the same likelihood. Adopting the convention that $\sum_{s=0}^{y_i-1}\log(n+s) = 0$ when $y_i = 0$, the log likelihood is

$$\begin{aligned} \log L &= \sum_{i=1}^{N}\sum_{s=0}^{y_i-1}\log\,(n+s) + \sum_{i=1}^{N}\log\left(\frac{\gamma^{y_i}}{y_i!}\right) + \sum_{i=1}^{N}\log\left((1-\gamma)^n\right) \\ &= \sum_{i=1}^{N}\sum_{s=0}^{y_i-1}\log\,(\lambda+s\gamma) + N\log\,(1-\gamma)^{\frac{\lambda}{\gamma}} - \sum_{i=1}^{N}\log\,(y_i!)\,, \end{aligned} \tag{16}$$

where the second line follows by substituting for n and simplifying.

In fact, (nonlinear) maximum likelihood estimators for n do not exist when the sample variance is less than the mean, that is, $s_y^2 \leq \bar{y}$ (see Al-Khasawneh (2010) and the references therein). Note that, even when drawing from a negative binomial, which by definition is over-dispersed, many individual samples

[17] We note that Yang et al. (2007) and Yang et al. (2009) pursue a similar exercise against variants of the generalized Poisson distribution, see the discussions in Consul (1989) and Joe and Zhu (2005), although we shall not pursue these models further.

will exhibit under-dispersion. We can see the difficulty explicitly by looking at simple moment estimates for λ and γ, that is, solve (A8) to get

$$\hat{\lambda} = \bar{y}\left(1 - \hat{\gamma}\right)$$
$$\hat{\gamma} = 1 - \left(\bar{y}/s_y^2\right).$$

Hence, problems arise when $s_y^2 \leq \bar{y}$ since then $\hat{\gamma} \leq 0$, which is illegtimate when investigating over-dispersion. Even if $s_y^2 \geq \bar{y}$, convergence issues arise if the difference is not great. This suggests that test procedures that use maximum likelihood estimates, such as Wald or Likelihood Ratio tests, can be problematic and that there may be a role for point optimal approaches.

4.2. Point Optimal Tests

Point optimal tests have had a long and varied career in econometrics; see King (1987) and King and Srianthakumar (2015) for an overview. These tests optimize power at a particular parameter value under the alternative, the idea being to have good power at a point where incorrectly accepting the null really matters. This is in contrast to, say, a score test that is locally best, in that it has the steepest power function local to the null hypothesis. Although not an undesireable property in any way, the practical difference between a null model and some other model local to the null is often, although not always, vanishingly small. So, optimizing the ability to distiguish between such null model and another local to it is not necessarily all that desireable a property. Moreover, there is implicit in such an approach the notion that the power function will be monotonically increasing, which ideally it should be, and that it will remain near the power envelope as the data generating process diverges from the null. In many cases this is indeed what happens, although we know that power functions are likely to cross, as otherwise the test would be uniformly most powerful, which is a very rare property indeed. The divergence between the power function of a score test and the power envelope is then something that requires exploration on a case by case basis and we will explore this below.

The log likelihood ratio of the $Katz - NB$ alternative to the Poisson (P) null is written

$$LLR\left(\lambda_1, \lambda_0, \gamma\right) = \sum_{i=1}^{N}\sum_{s=0}^{y_i-1}\log\left(\lambda_1 + s\gamma\right) + N\log\left(1-\gamma\right)^{\frac{\lambda_1}{\gamma}} - \left(\log\lambda_0 \sum_{i=1}^{N} y_i - N\lambda_0\right).$$

Assuming that both distributions are fully specified (with $\lambda = \lambda_0 = \lambda_1$), the Neyman-Pearson Lemma states that the UMP test of $\gamma = 0$ versus $\gamma = \gamma_1$ is given by $LLR\left(\lambda, \lambda, \gamma_1\right)$. Hence assuming λ known, the so-called power envelope is determined by computing $LLR\left(\lambda, \lambda, \gamma\right)$ over a range of values of $\gamma \in (0, 1)$. A PO test is constructed by choosing a fixed $\gamma = \gamma_{PO}$ to be a 'representative' value under the alternative $Katz - NB$ distribution, giving $LLR\left(\lambda, \lambda, \gamma_{PO}\right)$. It is desirable that γ_{PO} be chosen so that the power of the test $LLR\left(\lambda, \lambda, \gamma_{PO}\right)$ is as close as possible to the power of the family of tests $LLR\left(\lambda, \lambda, \gamma\right)$, $\gamma \in (0, 1)$. That is, ideally, γ_{PO} is chosen so that the power function of the resulting test is as close to the power envelope as possible.

4.3. Score Test

A common alternative to likelihood ratio approaches, which does not require maximum likelihood estimation of the parameter of interest, is to construct optimal tests local to the null $\gamma = 0$. The so-called eficient score tests, or simply score test, are derived by differentiating the log likelihood with respect to γ

and then setting $\gamma = 0$. Such score tests are easily found for the Katz family using (16); see, for example, Katz (1965) and Lee (1986). Specifically, the score test is

$$S(\lambda) = \frac{1}{2\lambda} \sum_{i=1}^{N} \left[y_i (y_i - 1) - \lambda^2 \right]. \tag{17}$$

This test was originally proposed by Katz (1965) on heuristic grounds and by Lee (1986) as a formal score test.[18,19]

4.4. Simulation Experiments

4.4.1. The Unconditional Model

In this section we simulate the powers of the point optimal and score tests and compare them to the benchmark power envelope. First we give details for the power envelope and this is followed by a description of the operational tests. We present results for a sample size of $N = 50$ throughout.

In Figures 1–3 below, we consider a range of values for $\lambda \in (0, 8)$ and $\gamma \in (0, 1)$ and look at the *relative* performance of the *PO* and *S* tests and the power envelope. For each (λ, γ) pair we generate samples from the Katz family, $K(\lambda, \gamma)$. This is efficiently accomplished using $p(0) = (1-\gamma)^{\frac{\lambda}{\gamma}}$ along with the defining recurrence $p(y+1) = (\lambda + \gamma y)/(y+1)p(y)$ and then sampling from the inverse cumulative distribution function. Setting $\gamma = \epsilon$, for some very small positive ϵ, effectively simulates from the Poisson null. The null critical values are computed by simulation to avoid asymptotic approximations. This means that the sizes of tests are accurate (up to simulation error) and hence that the power comparisons are meaningful in smaller sample sizes.

For the power envelope, we simulate from the null $P(\lambda) = K(\lambda, \epsilon)$ distribution, compute 10,000 values of $LLR(\lambda, \lambda, \gamma)$ and extract the 95% quantile as a critical value, cv. The DGP is the Katz family $K(\lambda, \gamma)$ and we simulate $10,000$ replicates from the DGP and count the percentage of times the PO statistic, $LLR(\lambda, \lambda, \gamma)$, exceeded the cv to calculate the power envelope.

The operational PO test estimates λ_0, λ_1 and fixes γ at γ_{PO}. To do this, we use simple moment based estimators, that is, compute $\hat{\lambda}_{NB} = \hat{\lambda}_P(1 - \hat{\gamma}_{NB})$, where $\hat{\lambda}_P = \bar{y}$ and $\hat{\gamma}_{NB} = 1 - \left(\bar{y}/s_y^2\right)$ using the mean and variance of the data at hand. Should $\hat{\gamma}_{NB}$ stray negative, we truncate and set $\hat{\gamma}_{NB} = \epsilon$. The PO test is computed as $LLR\left(\hat{\lambda}_P, \hat{\lambda}_{NB}, \gamma_{PO}\right)$ while the score test, using (17), is $S\left(\hat{\lambda}_P\right)$. The null is Poisson $P(\lambda) = K(\lambda, \epsilon)$ and we simulate from the null, computing $LLR\left(\hat{\lambda}_P, \hat{\lambda}_{NB}, \gamma_{PO}\right)$ and $S\left(\hat{\lambda}_P\right)$ for each realization, to get 5% cv's. We calculate the power by simulating from the DGP $K(\lambda, \gamma)$.

---

[18] Strictly, Katz (1965) adopted an approach more in keeping with a method of moments test. Specifically, he looked at the difference between estimators for the mean and variance, which should be equal under the null and then scaled this difference appropriately to obtain a distribution under the null. In any event, the statistic so obtained is the same as the one proposed by Lee (1986) that we consider here.

[19] Lee (1986) proposed other tests than the one considered here, although he did not compare them numerically The results recorded in Miller (1998) suggests that those involving third order moments may have better power properties. For now we are primarily concerned with proof of concept and do not explore these other tests in light of the simplicity of (17).

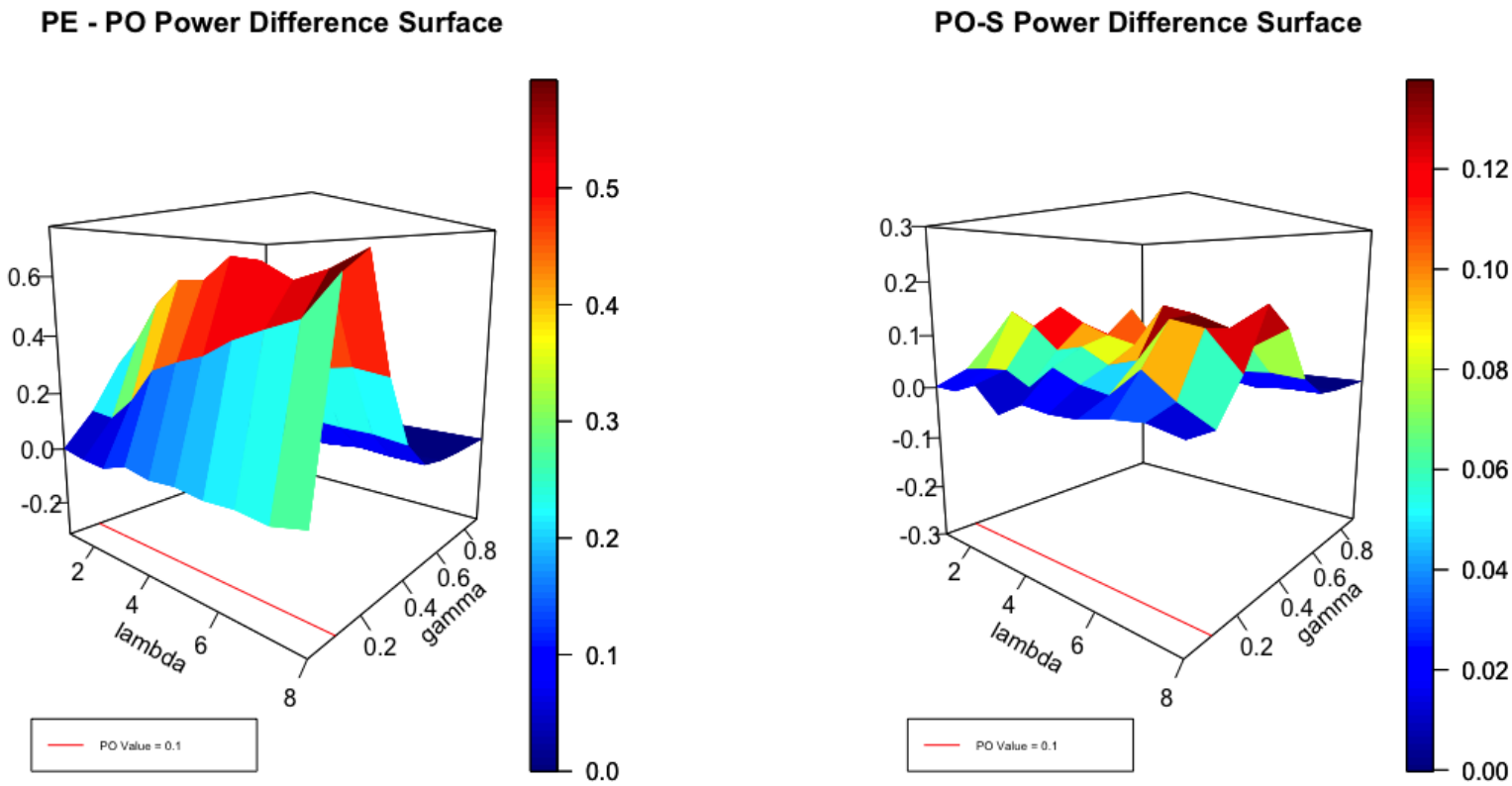


Figure 1. Difference in Power of the Envelope and the PO test as well as the Difference between the PO and Score Tests at $\gamma_{PO} = 0.1$.

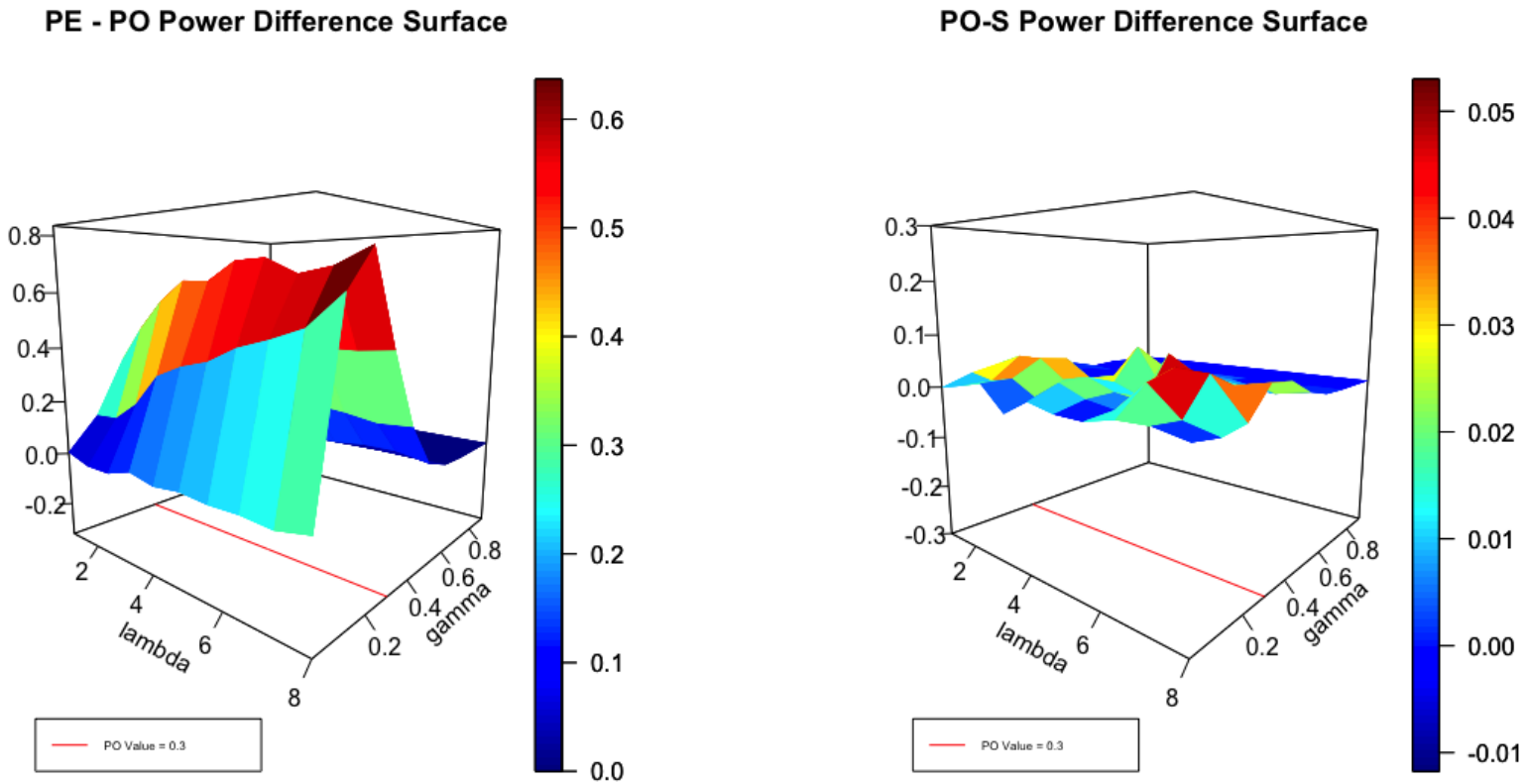


Figure 2. Difference in Power of the Envelope and the PO test as well as the Difference between the PO and Score Tests at $\gamma_{PO} = 0.3$.

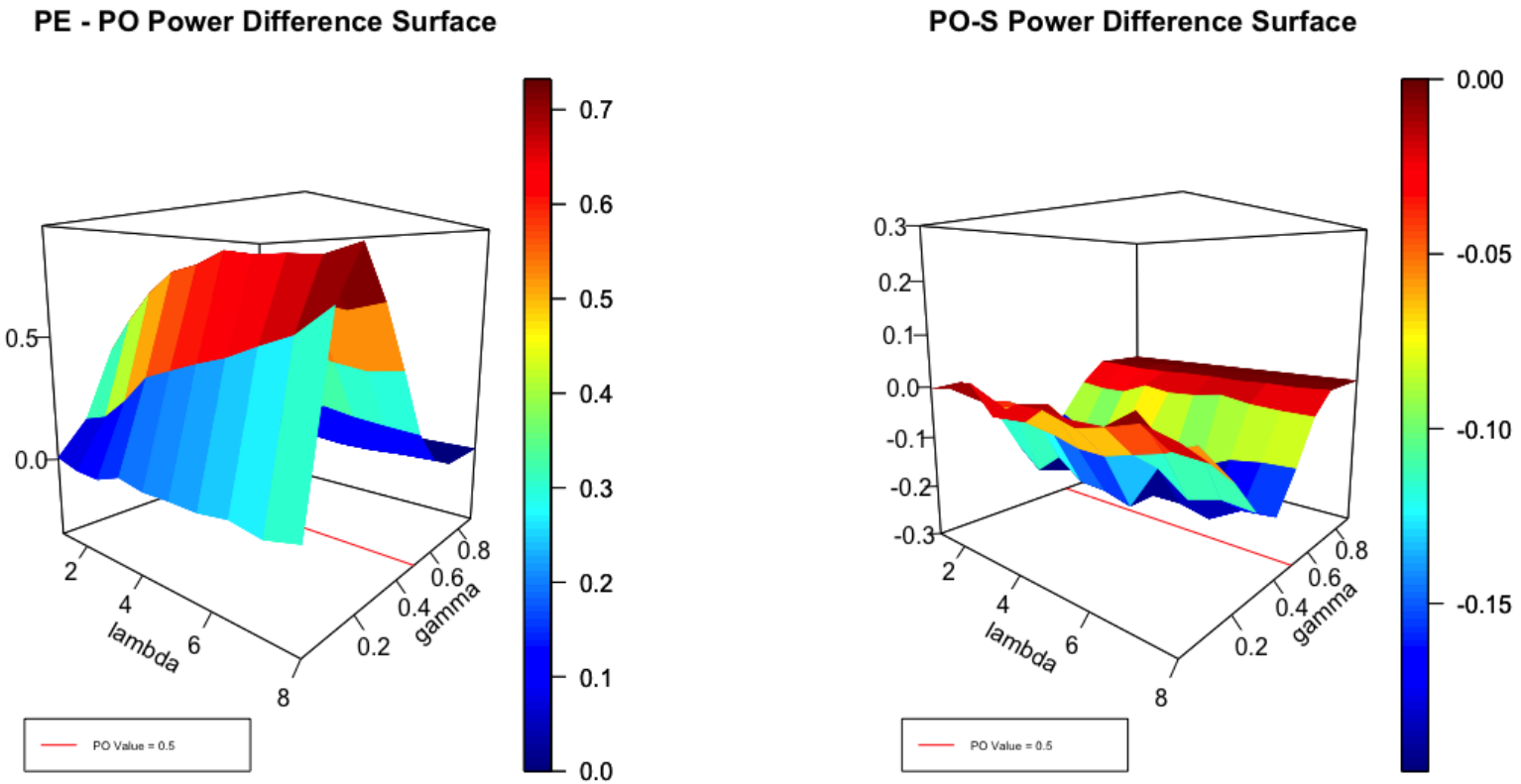


Figure 3. Difference in Power of the Envelope and the PO test as well as the Difference between the PO and Score Tests at $\gamma_{PO} = 0.5$.

It is helpful to view Figures 1–3 in the light of the cross sections displayed in Figure 4. It is clear that, for small values of γ, no test can be expected to perform well close to the null. Equally, for large values of γ, with high degrees of over-dispersion, all reasonable tests can be expected to be powerful. Thus, for small and large degrees of over-dispersion we expect to see little difference in the performance of the envelope and the *PO* test as the left panel of Figure 1 attests. For moderate values of γ, the power of the *PO* test can be significantly smaller than the envelope as the coloured scale suggests. In the left panel the *difference* in power between the *PO* and score tests is plotted for $\gamma_{PO} = 0.1$. The differences are not large but the *PO* test uniformly dominates as the scale indicates. Figure 2 plots the same surfaces for $\gamma_{PO} = 0.3$ and the interesting feature is that both tests perform similarly but none dominates the other. In Figure 3, $\gamma_{PO} = 0.5$ and the score test dominates by a small margin except where the *DGP* corresponds to γ_{PO}.

Since the shapes of the surfaces are quite smooth over λ, we plot a cross-section at $\lambda = 5$ to look at *absolute* performance. There are three panels in Figure 4 each corresponding to a value of $\gamma_{PO} = 0.1, 0.3, 0.5$. The power envelope is shown in red, with those of the PO test in blue and the Score test in orange. Also shown (vertically) is the PO point γ_{PO}.

The power envelope reaches unity at around $\gamma = 0.2$. This corresponds to a degree of over-dispersion, in the $Katz - NB$ distribution, of $\sigma^2/\mu = 1/(1-\gamma) = 1.25$. For the tests to reach equivalent power requires $\gamma = 0.6$ with $\sigma^2/\mu = 2.5$, roughly, and $\gamma = 0.9$ is required at $\sigma^2/\mu = 10$. So, neither test can match the envelope unless the degree of over-dispersion is quite large. For γ_{PO} less than 0.3 the *PO* test performs better uniformly, at 0.3 they perform equally well and for $\gamma_{PO} > 0.3$ the score test is better. Thus, a choice of γ_{PO} which is small will uniformly dominate.

4.4.2. The Katz Regression

In practice, the analysis of over-dispersion often takes place when covariates need to be taken into account. As explained in Section 3 there are many ways in which this may be approached. We work directly from the definition of the Katz family rather than mix over a kernel Poisson distribution. The log of the likelihood takes the form

$$LL = \sum_{i=1}^{N} \sum_{s=0}^{y_i - 1} \log(\lambda_i + s\gamma) + \sum_{i=1}^{N} \log(1-\gamma)^{\frac{\lambda_i}{\gamma}},$$

with $\lambda_i = \exp(\beta_1 x_i)$ varying and γ fixed. This gives $\mathrm{E}[Y_i|x_i] = \mu_i = \lambda_i/(1-\gamma)$ and $\mathrm{V}[Y_i|x_i] = \lambda_i/(1-\gamma)^2$.

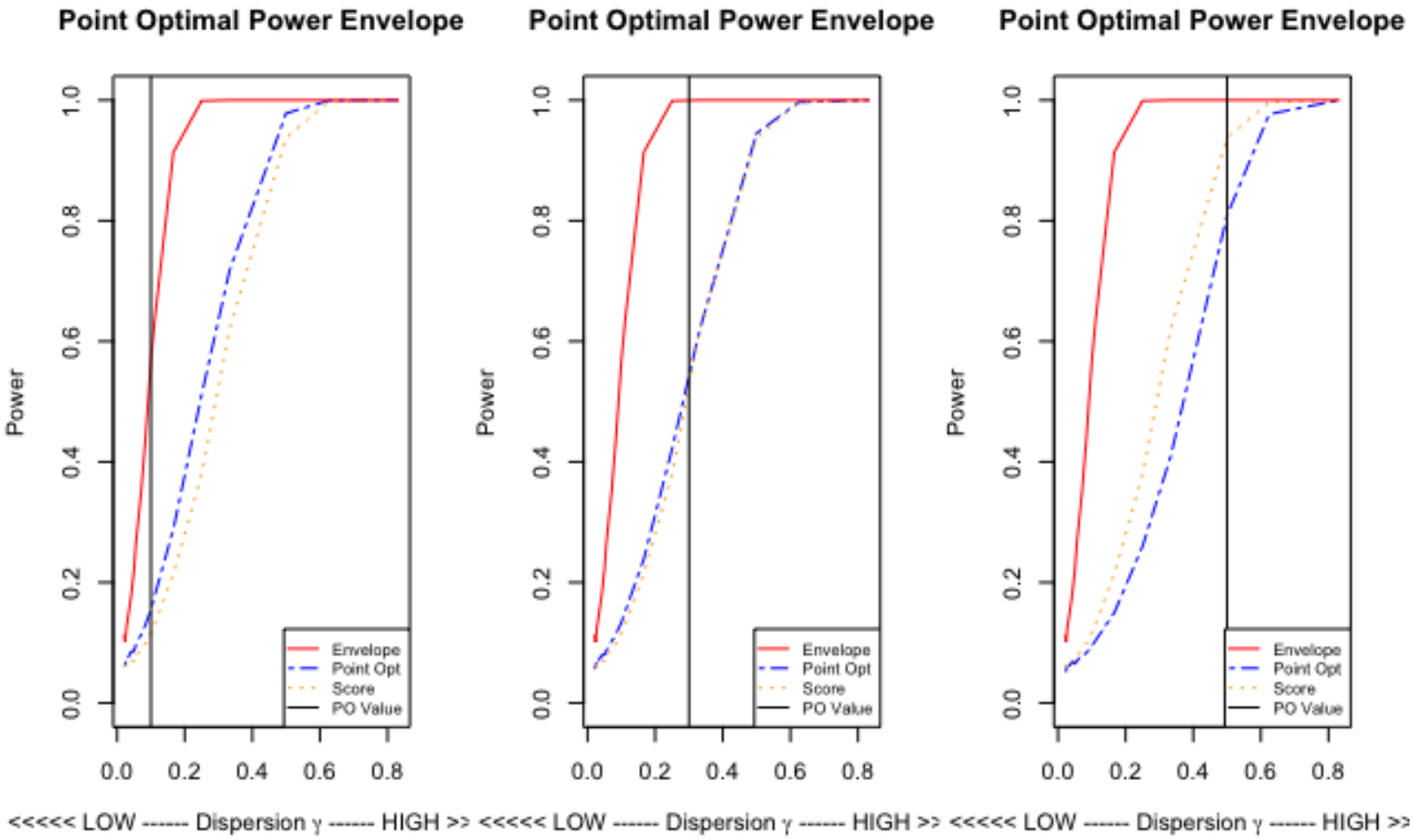


Figure 4. Cross Section of the Power Envelope Surface and Power Surfaces of the Score and Point Optimal tests at $\lambda = 5$.

The *PO* test, using γ_{PO}, is based on the log likelihood ratio,

$$LLR(\lambda_0, \lambda_1, \gamma_{PO}) = \sum_{i=1}^{N} \sum_{s=0}^{y_i-1} \log(\lambda_{1,i} + s\gamma_{PO}) + \sum_{i=1}^{N} \frac{\lambda_{1,i}}{\gamma_{PO}} \log(1-\gamma_{PO}) - \left(\sum_{i=1}^{N} y_i \log \lambda_{0i} - \sum_{i=1}^{N} \lambda_{0i} \right)$$

and the *PO* test needs to estimate the parameters $\breve{}_0$ and $\breve{}_1$. Estimating $\lambda_{1,i} = \exp(\beta_1 x_i)$ may be problematic as trying to fit a *NB* regression when the data is Poisson can lead to identification/convergence problems, exacerbated by the fact that fitting these types of regressions requires nonlinear maximum likelihood estimation. We avoided this issue in the last sub-section by using Katz moment estimators. Here, we use a regression version of the same idea. First, estimate a Poisson regression $P\left(\mu_i^*(x_i)\right)$ (including a constant) which will return the mean estimate $\exp(b_0 + b_1 x_i)$. Noting that we can write $\mu_i = \exp(\beta_1 x_i) / (1-\gamma) = \exp(\beta_0 + \beta_1 x_i)$, where $\beta_0 = -\log(1-\gamma)$, we can set $\hat{\beta}_1 = b_1$ and hence $\hat{\lambda}_{1i} = \exp\left(\hat{\beta}_1 x_i\right)$ to give the vector $\mathring{}_1$. To get $\hat{\lambda}_{0,i}$ we fit the Poisson without the constant term which returns the estimate $\exp(b_2 x_i)$, which gives $\hat{\lambda}_{0,i} = \exp(b_2 x_i)$, and hence $LLR\left(\mathring{}_0, \mathring{}_1, \gamma_{PO}\right)$.

As a comparator to the *PO* statistic, in the regression setting, we use the score test of Dean and Lawless (1989), which avoids the potential difficulties associated with maximum likelihood estimation. Thus, we estimate the Poisson regression $P(\mu(x_i))$ (with a constant) to get the vector of predictors $\hat{\mu}_i$ and, using $z_i = (y_i - \hat{\mu}_i)^2 - y_i$, the test $S\left(\hat{}\right)$ is computed as the t-statistic in the regression of z_i on $\hat{\mu}_i$.[20] Again critical values are computed by simulation. The null is generated as $P(\lambda_i)$, with $\lambda_i = \exp(\beta_1 x_i)$ used to keep the means of the counts low. We used $x_i \sim^i \log(P(2) + 1)$ and the x_i are kept fixed under replication. We generate simulated critical values, based on 10,000 replications, for the tests $S\left(\hat{}\right)$ and $LLR\left(\mathring{}_0, \mathring{}_1, \gamma_{PO}\right)$. To compute powers, the DGP $K(\lambda_i, \gamma) = NB(n_i, p)$ is used, where $n_i = \lambda_i/\gamma$ and γ takes a selection of values in $(0, 1)$. As usual, $\pi = 1 - \gamma$. The results are presented in Figure 5.

The *PO* test performs badly for very high degrees of over-dispersion when γ_{PO} is less than 0.5 approximately and is dominated by the score test for γ_{PO} greater than 0.5. However, the choice $\gamma_{PO} = 0.5$ does lead to superior *PO* performance albeit by not a great margin.

4.4.3. Summary

Our experimental results are mixed. In the unconditional model, the point optimal tests appeared to work best when γ was small, with their performance deteriorating relative to the score test as γ increased. In the regression model, the score test outperformed the point optimal tests suggesting that the null distribution of the score tests was more robust to the presence of nuisance parameters than was that of the point optimal tests, with none of the test statistics being pivotal. However, these rankings were also sensitive to the choice of point. This begs the question as to whether or not we are choosing the 'point' for the point optimal tests in a sensible way. It is to this question that we turn in the next section.

---

[20] We also considered an alternative test based on the t-statistic in the regression of z_i on a constant but there was little difference in performance. These tests correspond to the Negbin I and II cases above.

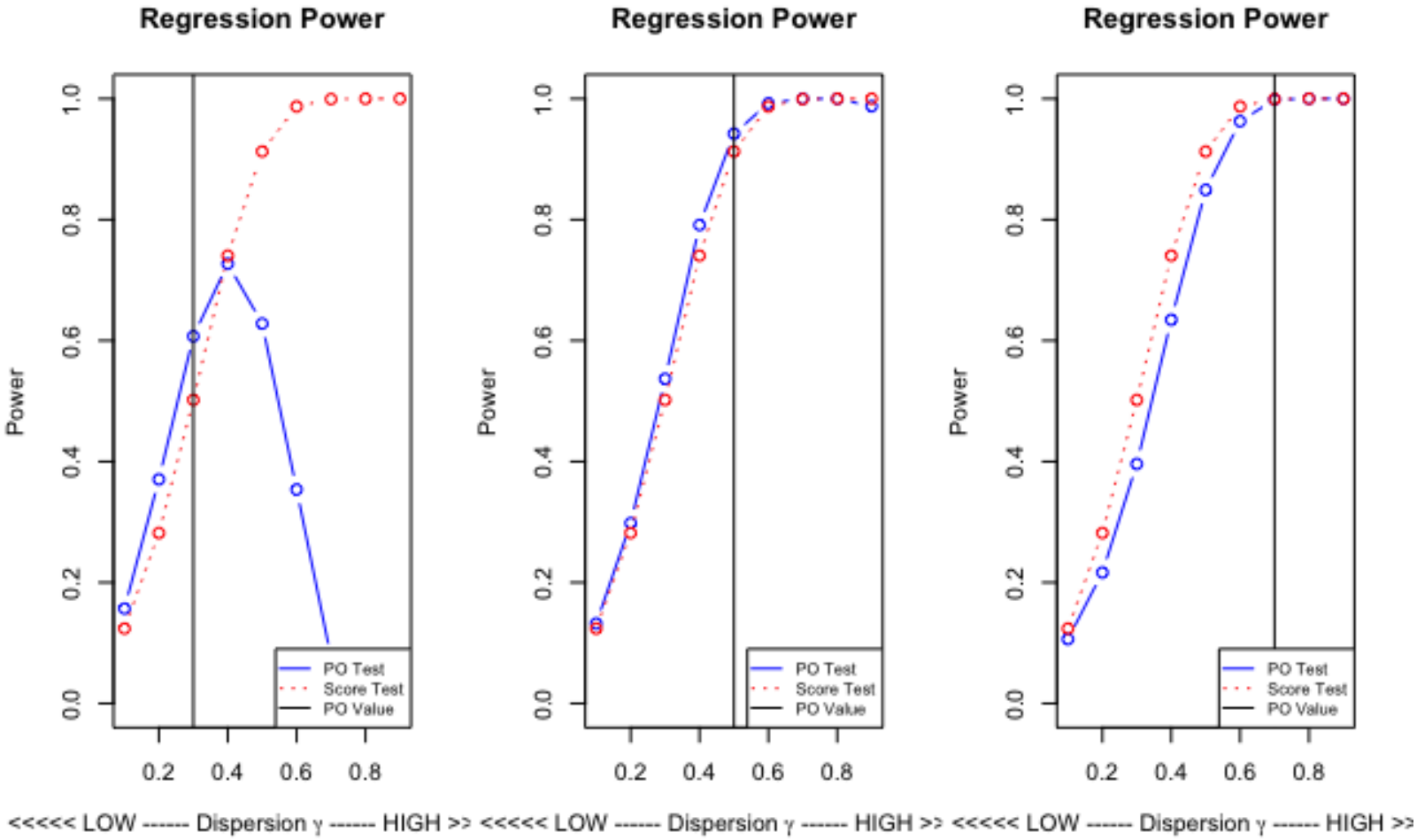


Figure 5. Powers of the Regression Score and Point Optimal tests.

5. Hellinger Distance

The reasoning behind the use of point optimal tests is to put power where it is of greatest practical use. The immediate problem facing the use of point optimal tests is where to place the 'point'. Sometimes the testing problem suggests a solution. Other times the choice is less clear and is often based an the outcome from a simulation study 'run-off', making the results somewhat ad hoc. The attraction of point optimal test in the context of testing for over-dispersion is that the parameter space of interest, namely that of γ, is bounded and so there is some hope of finding an appropriate point. One way of defining appropriate in this context is where the distribution under the alternative starts to depart from that under the null in some substantial way. The questions then reduces to one of how we might measure such a departure. In this section we explore the use of Hellinger distance $(\mathcal{H})$ for this purpose. We think that this is a novel use of such a distance measure and is of independent interest. We do not, however, assert that Hellinger distance is the only choice or even the best choice in this context, but it does yield some interesting results.

To begin, various definitions of Hellinger distance are available.[21] Originally proposed in an integral form by Hellinger (1909), we will work with the following discrete variant:

Definition 1 (Hellinger Distance for Discrete Random Variables)**.** *The squared Hellinger distance between these two discrete distributions P and Q is*

$$\mathcal{H}^2 = \frac{1}{2}\sum_{j=1}^{k}\left(\sqrt{p_j}-\sqrt{q_j}\right)^2 = 1-\sum_{j=1}^{k}\sqrt{p_j q_j}, \tag{18}$$

where $P=(p_1,\ldots,p_k)$ *and* $Q=(q_1,\ldots,q_k)$.

We note in passing that the Hellinger distance is bounded, $0\leq\mathcal{H}\leq 1 \implies 0\leq\mathcal{H}^2\leq 1$. $\mathcal{H}=1$ iff P assigns zero probability to anywhere that Q assigns positive probability and $\mathcal{H}=0$ iff $P=Q$.

5.1. The Poisson Distribution

By way of example, to illustrate the basic idea and to help calibrate the procedure, suppose that we choose as our base case a Poisson distribution with parameter λ_0 so that the implied standard deviation is $\sqrt{\lambda_0}$. Writing Prob $(X=x\mid\lambda)\equiv\mathcal{P}(\lambda)$, we are going to explore the behaviour of $\mathcal{H}$ as we compare $\mathcal{P}(\lambda_0)$ with $\mathcal{P}(\lambda_1)$ for various (λ_0,λ_1). When comparing Poisson distributions, the squared Hellinger distance is readily shown to

$$\mathcal{H}^2 = 1-\exp\left\{-\frac{1}{2}\left(\sqrt{\lambda_1}-\sqrt{\lambda_0}\right)^2\right\}. \tag{19}$$

Figure 6 provides some insight into the sensitivity of Poisson pmfs to changes in parameter values when the parameters are small and includes examples that are variously skewed to the right, (roughly) symmetric, and skewed to the left. Observe that, here we have used $\lambda=1$ as the base case and that, as λ increases, it is by one standard deviation each time and so these changes are quite dramatic.

In Figure 7 we present values for $\mathcal{H}$ for various λ_0 and λ_1. The dashed and dotted lines correspond to Hellinger distances of 0.1 and 0.05, respectively.

---

[21] See, for example, https://en.wikipedia.org/wiki/Hellinger_distance.

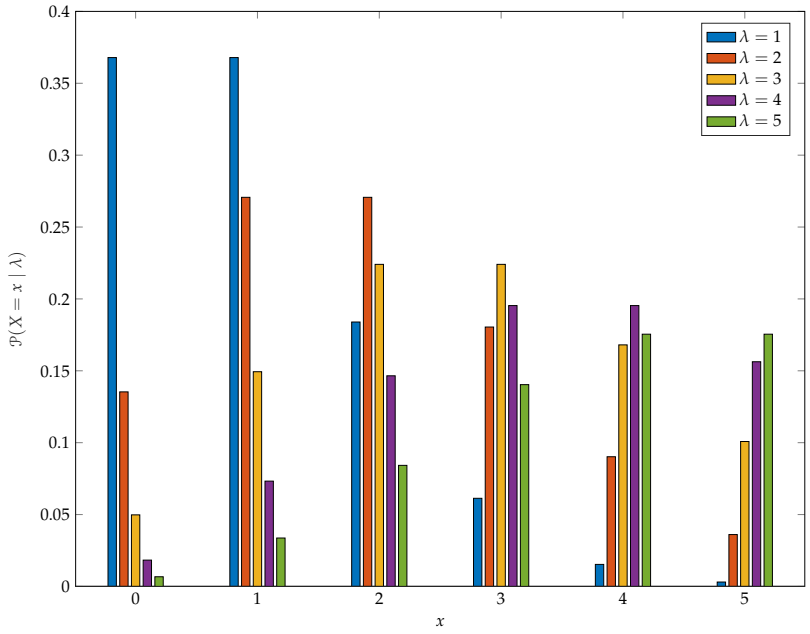


Figure 6. Selected Poisson Probability Mass Functions.

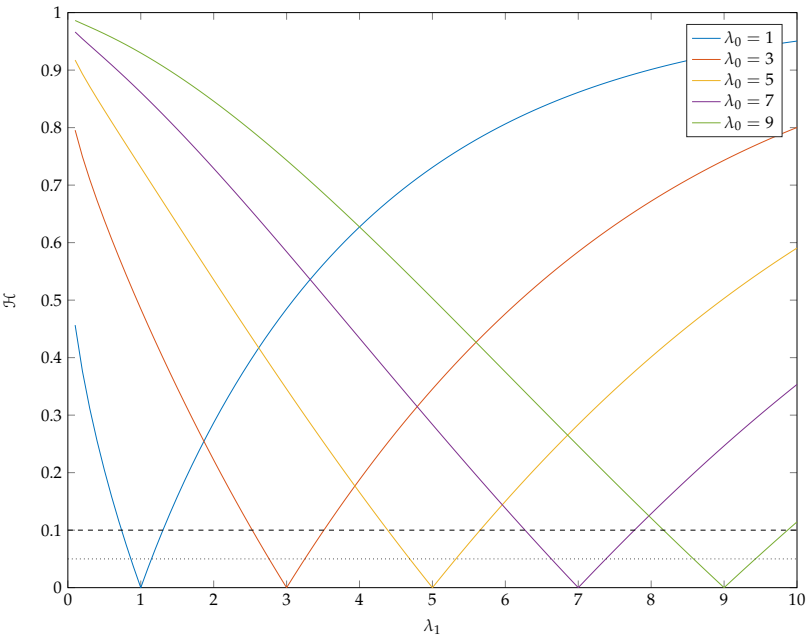


Figure 7. Hellinger Distances ($\mathcal{H}$) Between $\mathcal{P}(\lambda_0)$ and $\mathcal{P}(\lambda)$.

We observe that $\mathcal{H}$ is asymmetric in λ for all λ_0 considered, which reflects the skewed nature of Poisson distributions. Note that, as λ_0 increases, so too does standard deviation of the base distribution. As this happens a given value of $\mathcal{H}$ will admit great differences between λ_0 and λ_1. For example, when $\lambda_0 = 1$ a Hellinger distance of 0.1 or greater is achieved for any $0.8 \approx L \leq \lambda_1 \leq U \approx 1.3$. In contrast, when $\lambda_0 = 9$, $\mathcal{H} \leq 0.1$ for all $8.2 \approx L \leq \lambda_1 \leq U \approx 9.9$, which is a much wider interval than the previous case. Given that we are seeking to construct point optimal tests that compete with locally best tests, these results suggest that we need to be looking at points for which the Hellinger distance is quite small.

5.2. The Katz Distribution

We will take the Poisson ($\gamma = 0$) as our base model. Moreover, as Poisson-ness, or otherwise, is completely determined by the value of γ, we will hold λ fixed across models. Here the support under both null (equi-dispersion) and alternative (over-dispersion) is $y \in \mathbb{Y} = \{0, 1, 2, \ldots\}$ and so

$$\begin{aligned}\mathcal{H}^2 &= 1 - \sum_{y=0}^{\infty} \sqrt{\frac{e^{-\lambda}\lambda^y}{y!} \frac{\left(\frac{\lambda}{\gamma}\right)_y \gamma^y (1-\gamma)^{\lambda/\gamma}}{y!}} \\ &= 1 - \left[e^{-1}(1-\gamma)^{1/\gamma}\right]^{\lambda/2} \sum_{y=0}^{\infty} \frac{1}{y!} \sqrt{\left(\frac{\lambda}{\gamma}\right)_y (\lambda\gamma)^y}.\end{aligned}$$

Although not amenable to direct solution we notice that

$$\left(\frac{\lambda}{\gamma}\right)_y \gamma^y = \gamma^y \prod_{j=0}^{y-1} \left(\tfrac{\lambda}{\gamma} + j\right) = \prod_{j=0}^{y-1} (\lambda + j\gamma) > \lambda^y, \quad \text{for all } 0 < \gamma < 1. \tag{20}$$

Therefore,

$$\begin{aligned}\mathcal{H}^2 &> 1 - \left[e^{-1}(1-\gamma)^{1/\gamma}\right]^{\lambda/2} \sum_{y=0}^{\infty} \frac{1}{y!} \sqrt{\left[\left(\frac{\lambda}{\gamma}\right)_y \gamma^y\right]^2} \\ &= 1 - e^{-\lambda/2}(1-\gamma)^{-\lambda/(2\gamma)} = h_L^2, \text{ (say)}.\end{aligned} \tag{21}$$

We can solve this non-linear equation for γ_L numerically for given λ and h_L. Some results are reported in Table 1.

Table 1. Values for γ_L Obtained From (21) For Given h_L and λ (scaled by a factor of 10^{12}).

$h_L \backslash \lambda$	1	2	3	4	5
0.02	0.0348	0.0493	0.1261	0.1556	0.0147
0.04	0.0019	0.0121	0.0131	0.0250	0.0148
0.06	0.0032	0.0014	0.0115	0.0109	0.0136
0.08	0.0008	0.0033	0.0018	0.0061	0.0140
0.1	0.0027	0.0006	0.0022	0.0057	0.0072

We see that all values of γ_L are positive, albeit extremely to zero. Alternatively, from (20) we also have the result

$$\mathcal{H}^2 < 1 - \left[e^{-1}(1-\gamma)^{1/\gamma}\right]^{\lambda/2} \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} = 1 - e^{\lambda/2}(1-\gamma)^{\lambda/(2\gamma)} = h_U^2, \text{ (say)}. \tag{22}$$

Solutions to (22) for various λ and h_U are given in Table 2.

Table 2. Values for γ_U Obtained From (22) For Given h_U and λ.

$h_U \setminus \lambda$	**1**	**2**	**3**	**4**	**5**
0.02	0.0016	0.0008	0.0005	0.0004	0.0003
0.04	0.0064	0.0032	0.0021	0.0016	0.0013
0.06	0.0143	0.0072	0.0048	0.0036	0.0029
0.08	0.0252	0.0127	0.0085	0.0064	0.0051
0.1	0.0391	0.0198	0.0133	0.0100	0.0080

We see that γ_U is monotonically increasing in h_U but monotonically decreasing in λ. That is, once λ becomes sufficiently large, even small departures of the hellinger distance from zero are consistent with $\gamma > 0$.

All of the above said, however, the over-riding conclusion is that the optimal 'points' are going to be sufficiently close to zero that it is not clear that there is much benefit over just using the the score test, which is essentially point optimal at $\gamma = 0$. The main reason for such a conclusion is our earlier results indicating that, in the regression context, the score test is much less subject to the influence of the regression coefficients, which are nuisance parameters in this testing problem.

6. Conclusions

At a fundamental level, this paper explores the use of point optimal tests in the problem of testing for over dispersion. Our basis of comparison is the score test of Lee (1986), which is the same as the earlier method of moments test proposed by Katz (1965). Our findings are somewhat disappointing and we are unable to recommend that practitioners change their current practices as the performance of the point optimal tests is, at best, mixed. It may be possible to improve the performance of the point optimal tests by a more refined analysis of (i) the problem of nuisance parameters and (ii) the construction of p-values, along the lines suggested by King and Srianantha kumar (2015). This we leave for further work.

Along the way, the paper has made two other contributions. First, in the Appendix A we have provided a reasonably exhaustive treatment of the family of distributions consistent with the difference equation of Katz (1965). To the best of our knowledge this treatment extends all known earlier results by allowing for arbitrary points of left truncation. This expands the class of distributions originally considered by Katz (1965), which can be characterized as including zero in the support of the count variable. The treatment is closest to that of Willmot (1988), although there are differences in the mode of analysis and he restricts attention to extensions where only zero is omitted from the support of the count variable. We note in passing that right truncation is a much easier problem to deal with as it neither expands nor contracts the members in the family, in the way that left-truncation does. Its only consequence is the introduction of a scale factor equal to $1 - R$, where R denotes the upper tail probability that has been truncated.

The other contribution that we have made is to introduce the use of Hellinger distance as a metric by which one might settle on the 'points' characterizing point optimal tests. This is novel and allows a more systematic treatment than the grid searches that have characterized such choices in the past.

Author Contributions: Both authors have contributed equally to all aspects of the preparation and writing of this paper. They have both read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. On Left-Truncated Katz Distributions

Certain properties of the family of distributions defined by (2)–(4) are available on inspection. In particular, the support of Y is, in certain circumstances, parameter dependent. Here we characterize those circumstances.

To begin, let us establish some notation. Our count variable is $Y \in \mathbb{Y}_L \equiv \{L, L+1, L+2, \ldots, n\} \subseteq \mathbb{Z}_0$, where n may be either infinitely large or some finite integer and L is a non-negative integer integer. We will restrict $L < n$ because the case where $L = n$ yields a probability mass function degenerate at L, which is statistically uninteresting and shall, hereafter, be ignored. Next, write (2) as

$$p(y+1) = \frac{(\lambda + y\gamma)\, p(y)}{(y+1)}, \quad y \in \mathbb{Y}_L. \tag{A1}$$

This latter formulation has the advantage of being untroubled by the prospect of $p(y) = 0$. It will also prove convenient to be able to express all probabilities in terms of $p(L)$, which we can do via back substitution in (A1). Thus, for all $y \in \{L+1, \ldots, n\}$,

$$p(y) = \frac{p(L)\prod_{j=L}^{y-1}(\lambda + j\gamma)}{\prod_{k=L+1}^{y} k} = \frac{p(L)L!}{y!}\prod_{j=L}^{y-1}(\lambda + j\gamma). \tag{A2}$$

Moving forward we shall break up our observations into three categories: (i) those relating to the support of the random variable and the parameter space of the associated distributions, (ii) statements of the probability mass functions belonging to the family, and (iii) certain properties of the various distributions. The results are ultimately the same as those of Willmot (1988) in the special case where $L = 1$, although our mode of analysis is different and we extend his results by allowing for arbitrary $L > 0$.[22]

Appendix A.1. Support and Parameter Spaces

From (A2), we see that the sequence of probabilities generated by the difference equation (A1) is governed by $p(L)$ and by the terms $(\lambda + j\gamma)$, $j = L, L+1, \ldots, y$, as the ratio of factorials $L!/(y+1)!$ is a scale factor in the interval $(0, 1]$. Our subsequent analysis revolves around the behaviour of these quantities and the implications for $p(y+1)$ of these behaviours. With the exception of [1], we shall hereafter assume that $p(y) > 0$.

[1] **$0 < p(L) < 1$**
From (A1) we see that if $p(y) = 0$ for any $y \in \mathbb{Y}_L$ then $p(y+r) = 0$ for all $r \in \mathbb{N}$. In particular, if $p(L) = 0$ then $p(y) = 0$ for all $y \in \mathbb{Y}_L$. But this leads to violation of (4), that is, probabilities do not sum to unity, and so we exclude $p(L) = 0$ from further consideration. Equally, if $p(L) = 1$, so that the pmf of Y is degenerate at L, which is a case that we have already excluded from further analysis. Hereafter, we assume that $0 < p(L) < 1$.

[2] **$\gamma = 0 \implies \lambda > 0$ and $n = \infty$**
If $p(y) > 0$ and $\gamma = 0$ then we have a pmf degenerate at L unless $\lambda > 0$, which will be assumed hereafter. In this case there is no implied restriction on the upper bound of $\mathbb{Y}_L$, that is, $n = \infty$.

---

[22] In their extensions to this class of distributions, Panjer (1981); Sundt and Jewell (1981) and Willmot (1988) adopt a slightly different parameterization, specifically $p_y - (a + b/y)p_{y-1} = 0$, $y \in \{2, 3, 4, \ldots\}$. Equivalence with (A1) is seemingly established on setting $a = \gamma$ and $b = \lambda - \gamma$, although there are differences in the support of the resulting variables. In particular, $Y = 0$ is specifically excluded from this definition and hence many of the probability distributions claimed to satisfy the recursion in this form are not completely defined by it.

Of course, the concern when generating an infinite sequence of probabilities is to ensure that the associated series, $\sum_{y\in\mathbb{Y}} p(y)$, converges. This can be examined by considering the quantity

$$r_y(\gamma) = \frac{p(y+1)}{p(y)} = \frac{\lambda + y\gamma}{y+1} = \frac{\gamma + \lambda/y}{1+1/y}$$

and noting that

$$R(0) = \lim_{y\to\infty} r_y(0) = 0.$$

From the limit version of d'Alembert's ratio test we see that the series converges because $R(0) < 1$.

[3] $\boldsymbol{L = 0 \implies \lambda > 0}$
Similar in effect to the previous case, if $L = 0$ then $\lambda + L\gamma = \lambda$. Given $p(L) > 0$, as assumed above, $p(L+1) > 0$ if and only if $\lambda > 0$ which will be assumed, hereafter, for all cases where $L = 0$.

[4] $\boldsymbol{\lambda > 0, \gamma > 0 \implies 0 < \gamma < 1}$ **and** $\boldsymbol{n = \infty}$
Because $y \geq L \geq 0$, if $\lambda > 0$ and $\gamma > 0$ we see that $\lambda + y\gamma > 0$ for all $y \in \{L, L+1, \ldots\}$ and so here the support of the pmf of Y is unbounded from above and independent of the values taken by λ and γ. Again, we can establish convergence of the corresponding series. Here

$$R(\gamma) = \lim_{y\to\infty} r_y(\gamma) = \lim_{y\to\infty} \frac{\gamma + \lambda/y}{1+1/y} = \gamma > 0.$$

Appealing again to the limit version of d'Alembert's ratio test we see that the series converges if $\gamma < 1$, diverges if $\gamma > 1$, but the test is inconclusive if $\gamma = 1$. Expanding the denominator of $r_y(1)$ in power series yields

$$r_y(1) = \frac{1+\lambda/y}{1+1/y} = (1+\lambda/y)\sum_{j=0}^{\infty}\left(-\frac{1}{y}\right)^j = 1 - \frac{1-\lambda}{y} + O(y^{-2}).$$

Applying Gauss's test,[23] we see that the series will converge absolutely if and only if $1 - \lambda > 1$ but will otherwise diverge. Here we have assumed that $\lambda > 0$ and so $1 - \lambda < 1$. Hence, the series is divergent for $\gamma \geq 1$.

[5] $\boldsymbol{\lambda < 0, \gamma < 0}$
In this case there is no value of y that satisfies $\lambda + y\gamma > 0$ and so $y + 1$ cannot belong to $\mathbb{Y}_L$. Moreover, this statement remains true even if $y = L$. Consequently, in this case, the pmf of Y is degenerate at L, a situation that we have chosen to exclude from further consideration.

[6] $\boldsymbol{\lambda}$ **and** $\boldsymbol{\gamma}$ **of different sign**
In this case we see that $\lambda + y\gamma$ can change sign as y increases, unlike the situation of the previous two cases. Let n denote the smallest value of y such that $\lambda + y\gamma \leq 0$. Then n is the largest value in $\mathbb{Y}_L$. There are only two cases to consider here (having treated that of $\gamma = 0$ above): (i) $\lambda \geq 0, \gamma < 0$, and (ii) $\lambda \leq 0, \gamma > 0$.

(a) $\boldsymbol{\lambda \geq 0, \gamma < 0 \implies n = \lceil -\lambda/\gamma \rceil}$
If $\lambda \leq -L\gamma$ then the pmf of Y will be degenerate at L which, as explained above, is statistically uninteresting and a situation that we will assume away. That is, if $\gamma < 0$ then we will assume that $\lambda > -L\gamma$. In particular, if $L = 0$ then this requirement reduces to $\lambda > 0$. As y increases,

[23] See, for example, Weisstein (2019).

$\lambda + y\gamma$ will approach zero from above. That value of y for which $\lambda + y\gamma$ is first less than or equal to zero is the largest value of y in $\mathbb{Y}_L$ and shall be denoted by n, so that $p(n)$ is well-defined but $p(n+1)$ is not.[24] That is, n is the smallest integer greater than or equal to $-\lambda/\gamma$. This is the definition of the so-called ceiling function, written $n = \lceil -\lambda/\gamma \rceil$. In summary, if $\gamma < 0$ then we see that the upper bound on the support of the pmf of Y is a function of the parameters λ and γ, with the space of λ subject to the constraint $\lambda > -L\gamma$.

(b) $\boldsymbol{\lambda \leq 0, \gamma > 0 \implies \lambda > -L\gamma, L > 0, n = \infty,}$ **and** $\boldsymbol{0 < \gamma \leq 1}$
Here $\lambda + y\gamma$ is an increasing function of y but the pmf of Y is non-degenerate at L if and only if $\lambda > -L\gamma$. As we have already excluded from further consideration pmfs degenerate at L we here assume this to be the case. In particular, when $L = 0$ we have a contradiction as we are assuming both $\lambda > 0$, which is required when $L = 0$ (see [3]), and $\lambda \leq 0$; we conclude that $\lambda \leq 0$ and $\gamma > 0$ can only arise when $L \geq 1$. As $\lambda + \psi\gamma > 0$ for all $y \in \mathbb{Y}$, $\mathbb{Y}_L$ will be unbounded from above provided that the series of probabilities so formed is convergent. Using the analysis outlined in [4], applying the ratio test we find convergence for all $-L\gamma < \lambda \leq 0$ provided that $0 < \gamma < 1$. Moreover, if $\gamma = 1$, Gauss's test gives convergence provided that λ is strictly negative, that is, $-L\gamma < \lambda < 0$.

We summarize these findings in Table A1 and note in passing that, when $L = 0$, the only valid parameter configurations are those found in the row $\lambda > 0$.

Table A1. Parameter Configurations When $L > 0$.

$\lambda \setminus \gamma$	$\gamma < 0$	$\gamma = 0$	$0 < \gamma < 1$	$\gamma = 1$
$-L\gamma < \lambda < 0$	n/a	n/a	$n = \infty$	$n = \infty$
$\lambda = 0$	n/a	n/a	$n = \infty$	n/a
$\lambda > 0$	$n = \lceil -\lambda/\gamma \rceil$	$n = \infty$	$n = \infty$	n/a

Appendix A.2. Probability Mass Functions and Their Properties

Having established the various restrictions on the parameter space and the support for the family of distributions generated by (A1), we now turn attention to the resulting pmfs and their properties. To begin, we will distinguish between two classes of distributions: (i) $L = 0$, the class originally explored by Katz (1965), and (ii) $L > 0$, which has subsequently been explored by others. In order to explore these pmfs, our first task is to evaluate $p(L)$ which forms part of the normalizing constant in (A2).

---

[24] In essence, this is the same as adopting the convention that any negative probabilities are set to zero. It might be argued that this is at odds with Katz's original assumptions and should be excluded. Our justification for the inclusion in our analysis of these distributions where λ/γ is non-integer, is that Katz himself included them.

> The class of distributions so defined includes the Poisson distributions, the two-parameter binomial (Bernoulli) distributions, and the two-parameter negative binomial (Pascal) distributions. Aside from these, the class contains only the mild generalizations obtained for the latter two of these types by permitting the parameter n (number of "trials" in direct sampling) and the parameter r (number of failures in inverse sampling) to take any positive real values. (Katz 1965, p. 175).

Appendix A.2.1. $\boldsymbol{L = 0}$

In the previous section we established that, when $L = 0$, we require $\lambda > 0$. Moreover, we also required that $\gamma < 1$, with $\mathbb{Y}_L$ unbounded from above if $0 \leq \gamma < 1$ but that an upper bound of $n = \lceil -\lambda/\gamma \rceil$ exists if $\gamma < 0$. Summing the right-most side of (A2) over all $y \in \mathbb{Y}_L$ and adding $p(0)$ yields

$$\begin{aligned} 1 &= p(0) + \sum_{y=1}^{n} \frac{p(0)0!}{y!} \prod_{j=0}^{y-1} (\lambda + j\gamma) = p(0) \left[1 + \sum_{y=1}^{n} \frac{1}{y!} \prod_{j=0}^{y-1} (\lambda + j\gamma) \right] \\ &= p(0) \left[\sum_{y=0}^{n} \frac{1}{y!} \prod_{j=0}^{y-1} (\lambda + j\gamma) \right]. \end{aligned}$$

where we have adopted the convention of

$$\prod_{j=a}^{b} x(j) = 1, \qquad b < a. \tag{A3}$$

Recall that if $\gamma \geq 0$ then $n \equiv \infty$, otherwise $n = \lceil -\lambda/\gamma \rceil$. Thus, $p(0)S(n) = 1 \implies p(0) = [S(n)]^{-1}$, where

$$S(n) = \begin{cases} \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} = e^{\lambda}, & \gamma = 0, \\ \sum_{y=0}^{\infty} \frac{(\lambda/\gamma)_y \, \gamma^y}{y!} = (1-\gamma)^{-\lambda/\gamma}, & 0 < \gamma < 1, \\ \sum_{y=0}^{n} \frac{(\lambda/\gamma)_y \, \gamma^y}{y!}, & \gamma < 0, \end{cases}$$

where we have used the Pochhammer symbol $(a)_n$ to denote the rising factorial function

$$(a)_n = a(a+1)(a+2)\ldots(a+n-1) = \Gamma(a+n)/\Gamma(a),$$

a polynomial of order n (n a non-negative integer) in a, with $(a)_0 = 1$ (including $(0)_0 = 1$), and where $\Gamma(a)$ denotes the usual Gamma function.[25] Note that the argument of the Pochhammer symbol can be negative and is in certain cases considered below. In the event that 'a' is a negative integer, the Pochhammer symbol will equal zero for all $n > a$. The resulting pmfs are

$$p(y) = \begin{cases} \frac{e^{-\lambda}\lambda^y}{y!}, & \gamma = 0 \quad \text{(Poisson)}, \\ \frac{(\lambda/\gamma)_y \, \gamma^y (1-\gamma)^{\lambda/\gamma}}{y!}, & 0 < \gamma < 1 \quad \text{(Negative Binomial)}, \\ \frac{(\lambda/\gamma)_y \, \gamma^y / y!}{\sum_{j=0}^{n} (\lambda/\gamma)_j \, \gamma^j / j!}, & \gamma < 0, \end{cases} \tag{A4}$$

[25] A useful collection of results on Pochhammer symbols can be found in Slater (1966, Appendix I).

There are two simplifications that arise when λ/γ is integer. First, if one restricts attention to the case where λ/γ is integer, r say, and $0 < \gamma < 1$ then

$$\frac{(r)_y}{y!} = \frac{(r+y-1)!}{(r-1)!\,y!} = \binom{r+y-1}{y}.$$

On setting $\pi = 1-\gamma$, the pmf reduces to

$$p(y) = \binom{r+y-1}{y}(1-\pi)^y \pi^r. \tag{A5}$$

This form of the negative binomial distribution, also known as the Pascal distribution, admits an inverse sampling interpretation is available. Specifically, Y can be interpreted as a count of the number of failures in a sequence of independent Bernoulli trials, each with probability of success π, before the rth success is observed. Interestingly, we note that

$$\lim_{\gamma\to 0}(1-\gamma)^{\lambda/\gamma} = e^{-\lambda},$$

and so the negative binomial representation in (A4) can be thought of as valid for all cases $\gamma \geq 0$, recognizing that the case $\gamma = 0$ must be thought of as a limit. Finally, when λ/γ is non-integer, the pmf in (A4) still gives the probability that $Y = y$ given the parameters λ and γ, it just no longer admits the inverse sampling interpretation usually ascribed to a count variable with a negative binomial distribution.

Second, if $\gamma < 0$ and $n = \lambda/\gamma$ is a negative integer, so that $\lceil -\lambda/\gamma \rceil = -\lambda/\gamma$, then

$$\frac{(\lambda/\gamma)_y\,\gamma^y}{y!} = \frac{(-n)_y\,\gamma^y}{y!} = \frac{n!(-\gamma)^y}{(n-y)!\,y!},$$

so that

$$S(n) = \sum_{y=0}^{n} \frac{n!(-\gamma)^y}{(n-y)!\,y!} = (1-\gamma)^n$$

and

$$p(y) = \binom{n}{y}(-\gamma)^y(1-\gamma)^{-n}.$$

On setting $\pi = -\gamma/(1-\gamma)$, so that $\gamma = -\pi/(1-\pi)$, we can recognize the resulting pmf

$$p(y) = \binom{n}{y}\pi^y(1-\pi)^{n-y}$$

as that of a binomial random variable where, again, π denotes the success of a single Bernoulli trial and $p(y)$ gives the probability of y successes in a sequence of n independent Bernoulli trials. That is, $Y \sim \text{Binomial}(n, \pi)$. These findings are summarized in Figure A1.

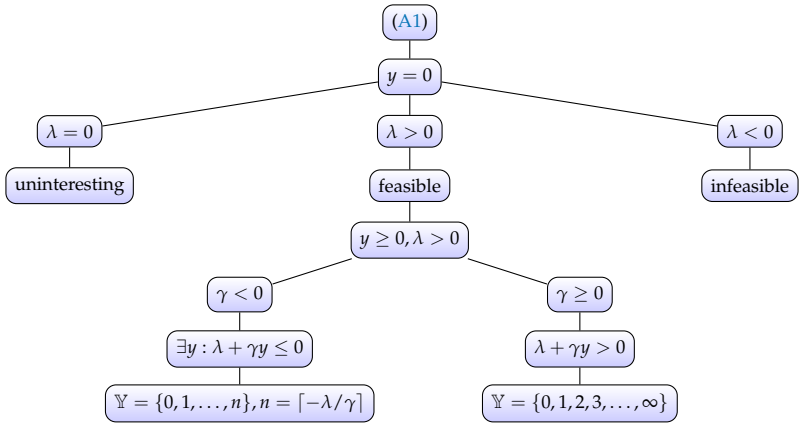
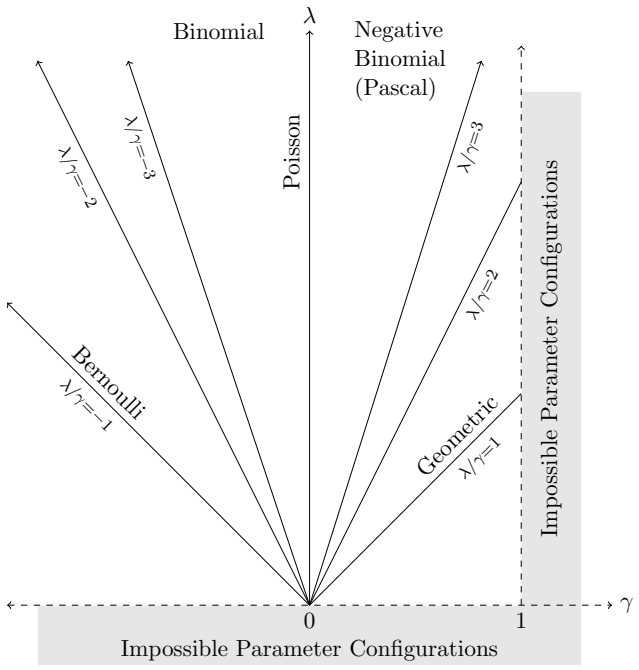


Figure A1. Restrictions on Parameters and Support Implied by (A1).

The Poisson, Pascal, and Binomial distributions, being those cases where λ/γ is integer, were the cases originally explored in Katz (1965). Figure A2, which is a variant of Katz (1965, fig. 1), provides a graphical representation of these distributions.



The set of permissible parameter combinations is represented by rays radiating from the origin, a subset of which are depicted, all other possibilities are ignored. The dashed lines are not included within this set. The area shaded light grey represents the boundary of the set of impossible parameter configurations.

Figure A2. Parameter Space for the Katz Family in Special Cases.

Kemp (1968) observed that the family of distributions depicted in Figure A2 could all be expressed in terms of hypergeometric functions on noting that

$$(1-\gamma)^{-\lambda/\gamma} = {}_1F_0\left(\frac{\lambda}{\gamma}; \gamma\right)$$

and that, specifically,

$$\lim_{\gamma \to 0} {}_1F_0\left(\frac{\lambda}{\gamma}; \gamma\right) = e^{\lambda},$$

so that

$$p(y) = \frac{\left(\frac{\lambda}{\gamma}\right)_y \gamma^y}{y!\, {}_1F_0\left(\frac{\lambda}{\gamma}; \gamma\right)}, \quad y \in \mathbb{Y}, \gamma < 1, \lambda > 0, \tag{A6}$$

subject to the requirement that λ/γ is integer if $\gamma < 0$. This characterization of the probability function makes two things clear. First, the restriction that $\gamma < 1$ follows immediately from the standard convergence criteria for hypergeometric functions; see, inter alios, Abadir (1999, p. 292). Second, it is clear that, for $\gamma > 0$, the restriction that λ/γ be integer is completely unnecessary as the probability function is perfectly well defined for non-integer values of this ratio.[26]

It is straight-forward to show that the probability generating function for this family of distributions is of the form

$$G(t) = \frac{\sum_{y=0}^{n} \frac{\left(\frac{\lambda}{\gamma}\right)_y (t\gamma)^y}{y!}}{\sum_{j=0}^{n} \frac{\left(\frac{\lambda}{\gamma}\right)_j \gamma^j}{j!}}.$$

In the special cases where either $\gamma \geq 0$ or where $-\lambda/\gamma$ is a positive integer, $G(t)$ reduces to

$$G(t) = {}_1F_0\left(\frac{\lambda}{\gamma}; t\gamma\right) / {}_1F_0\left(\frac{\lambda}{\gamma}; \gamma\right).$$

Moments for all members of the family can be calculated directly from (A1), without reference to the exact form of the pmf. A slight re-arrangement of (A1) allows us to sum over $\mathbb{Y}$, the support of Y, thus

$$\sum_{y \in \mathbb{Y}} (y+1)p(y+1) = \sum_{y \in \mathbb{Y}} (\lambda + y\gamma)p(y). \tag{A7}$$

The left-hand side of (A7) can be written

$$\begin{aligned} \sum_{y \in \mathbb{Y}} (y+1)p(y+1) &= 0 \times p(0) + \sum_{y \in \mathbb{Y}} (y+1)p(y+1) \\ &= \sum_{y \in \mathbb{Y}} yp(y) = \mathrm{E}[Y] = \mu \text{ (say)}. \end{aligned}$$

The right-hand side becomes

$$\lambda \sum_{y \in \mathbb{Y}} p(y) + \gamma \sum_{y \in \mathbb{Y}} yp(y) = \lambda + \gamma\mu.$$

Solving for μ yields

$$\mathrm{E}[Y] = \frac{\lambda}{1-\gamma} = \mu \text{ (say)}, \tag{A8a}$$

[26] We note that Katz (1965) was perfectly well aware of the possibility of non-integer values of λ/γ, see the quote in Footnote 24.

and similar arguments lead to

$$\mathrm{V}[Y] = \frac{\lambda}{(1-\gamma)^2} = \sigma^2 \text{ (say).} \tag{A8b}$$

From Katz (1965, p. 176) we have the following inverse parametric relationships

$$\lambda = \frac{\mu^2}{\sigma^2} \quad \text{and} \quad \gamma = 1 - \frac{\mu}{\sigma^2},$$

which yields a potentially useful alternative parameterization of the distributions in terms of mean and variance rather than the somewhat more nebulous λ and γ. Observe that if $\gamma = 0$ then $\mathrm{E}[Y] = \mathrm{V}[Y]$, a situation termed equi-dispersion. If $0 < \gamma < 1$, then $\mathrm{V}[Y] > \mathrm{E}[Y]$, which is called over-dispersion and, if $\gamma < 0$ then $\mathrm{V}[Y] < \mathrm{E}[Y]$, which is called under-dispersion. Importantly, if we consider the ratio $\mathrm{V}[Y]/\mathrm{E}[Y] = 1/(1-\gamma)$ then we see that under-dispersion, equi-dispersion, and over-dispersion are determined by the value of γ alone, and so λ is a nuisance parameter for the testing problems of interest in this paper. Finally, observe that $\mathrm{E}[Y]$ is an increasing function of γ. Specifically,

$$\mathrm{E}[Y] \begin{cases} < \lambda, & \text{if } \gamma < 0, \\ = \lambda, & \text{if } \gamma = 0, \text{ and} \\ > \lambda, & \text{if } 0 < \gamma < 1. \end{cases}$$

Appendix A.2.2. $\boldsymbol{L > 0}$

This case differs from that of $L = 0$ in two key ways: (i) there are three more cases to consider, all related to $\lambda \leq 0$ and, obviously, (ii) zero is no longer in $\mathbb{Y}_L$. To begin the analysis, let us first determine $p(L)$ by summing over (A2). Noting that n may be infinite (depending on parameter configuration) and adopting the convention (A3), we see that

$$1 = \sum_{y=L}^{n} p(y) = p(L) + \sum_{y=L+1}^{n} \frac{p(L)L!}{y!} \prod_{j=L}^{y} (\lambda + j\gamma),$$

so that

$$p(L)L! = \left[\sum_{y=L}^{n} \frac{1}{y!} \prod_{j=L}^{y-1} (\lambda + j\gamma) \right]^{-1} = \mathcal{I}^{-1}, \text{ say.}$$

The exact definition of $p(L)$, as noted above, is parameter dependent. Hence,

(i) if $\gamma = 0, \lambda > 0$ then

$$\mathcal{I} = \sum_{y=L}^{\infty} \frac{\lambda^{y-L+1}}{y!} = \frac{e^{\lambda}}{\lambda^{L-1}} \left(1 - e^{-\lambda} \sum_{j=0}^{L-1} \frac{\lambda^j}{j!} \right);$$

(ii) if $0 < \gamma < 1, \lambda > 0$ then, on noting that $(\lambda/\gamma + L)_{y-L} = (\lambda/\gamma)_y \,/\, (\lambda/\gamma)_L$,

$$\mathcal{I} = \frac{1}{\left(\frac{\lambda}{\gamma}\right)_L \gamma^L} \sum_{y=L}^{\infty} \frac{\left(\frac{\lambda}{\gamma}\right)_y \gamma^y}{y!} = \frac{(1-\gamma)^{-\lambda/\gamma}}{\left(\frac{\lambda}{\gamma}\right)_L \gamma^L} \left[1 - (1-\gamma)^{\lambda/\gamma} \sum_{j=0}^{L-1} \frac{\left(\frac{\lambda}{\gamma}\right)_j \gamma^j}{j!} \right];$$

(iii) if $\gamma < 0, \lambda > 0$ then

$$\begin{aligned}\mathcal{I} &= \frac{1}{\left(\frac{\lambda}{\gamma}\right)_L \gamma^L} \sum_{y=L}^{\lceil \lambda/\gamma \rceil} \frac{\left(\frac{\lambda}{\gamma}\right)_y \gamma^y}{y!} \\ &= \frac{\sum_{k=0}^{\lceil \lambda/\gamma \rceil} \left(\frac{\lambda}{\gamma}\right)_k \gamma^k / k!}{\left(\frac{\lambda}{\gamma}\right)_L \gamma^L} \left[1 - \frac{\sum_{j=0}^{L-1} \left(\frac{\lambda}{\gamma}\right)_j \gamma^j / j!}{\sum_{m=0}^{\lceil \lambda/\gamma \rceil} \left(\frac{\lambda}{\gamma}\right)_m \gamma^m / m!}\right].\end{aligned}$$

These first three results correspond to those examined in the $L = 0$ case and they have the same simplifications for λ/γ integer as mentioned in that case.[27] The structure of the result is clear, with the normalizing constant scaled by a factor of $1 - \text{Prob}\,(Y < L)$, so that the resulting probabilities are simply left-truncated versions of those encountered previously. In particular, we see that, for $L > 0$ and $\lambda > 0$,

$$p(y) = \begin{cases} \dfrac{e^{-\lambda}\lambda^y}{y!\left(1 - e^{-\lambda}\sum_{j=0}^{L-1} \frac{\lambda^j}{j!}\right)}, & \gamma = 0, \\ \dfrac{(\lambda/\gamma)_y \, \gamma^y (1-\gamma)^{\lambda/\gamma}}{y!\left[1 - (1-\gamma)^{\lambda/\gamma} \sum_{j=0}^{L-1} \left(\frac{\lambda}{\gamma}\right)_j \gamma^j / j!\right]}, & 0 < \gamma < 1, \\ \dfrac{(\lambda/\gamma)_y \, \gamma^y / y!}{\sum_{k=0}^{n} (\lambda/\gamma)_k \, \gamma^k / k! \left[1 - \frac{\sum_{j=0}^{L-1} \left(\frac{\lambda}{\gamma}\right)_j \gamma^j / j!}{\sum_{m=0}^{n} \left(\frac{\lambda}{\gamma}\right)_m \gamma^m / m!}\right]}, & \gamma < 0, n = \lceil \lambda/\gamma \rceil. \end{cases} \tag{A9}$$

Before moving it is worth reminding ourselves of cases that we need not consider further. If $L > 0$ and $\gamma \leq 0$ then the only case leading to valid, non-degenerate distributions are those where $\lambda > 0$. The next three cases have no corresponding result when $L = 0$.

(iv) If $0 < \gamma < 1, \lambda = 0$ then

$$\begin{aligned}\mathcal{I} &= \sum_{y=L}^{\infty} \frac{(y-1)!\gamma^{y-L}}{(L-1)!\, y!} \\ &= -\frac{\sum_{y=1}^{\infty} (-1)^{y+1}(-\gamma)^y / y}{\gamma^L (L-1)!} \left[1 - \frac{\sum_{j=1}^{L-1} (-1)^{j+1}(-\gamma)^j / j}{\sum_{y=1}^{\infty} (-1)^{y+1}(-\gamma)^y / y}\right] \\ &= -\frac{\ln(1-\gamma)}{\gamma^L (L-1)!} \left[1 - \frac{\sum_{j=1}^{L-1} (-1)^{j+1}(-\gamma)^j / j}{\ln(1-\gamma)}\right],\end{aligned}$$

where the final equality follows on recognising the Mercator series and

$$p(y) = -\frac{\gamma^y}{y \ln(1-\gamma)\left[1 + \frac{\sum_{j=1}^{L-1} \gamma^j / j}{\ln(1-\gamma)}\right]}. \tag{A10}$$

[27] The condition λ/γ integer obviously requires $\gamma \neq 0$.

In the special case $L = 1$, the quantity in the square brackets reduces to unity and

$$p(y) = -\frac{\gamma^y}{y \ln(1-\gamma)},$$

which is the pmf of a logarithmic distribution. If $L > 1$ then (A10) is recognizable as a left-truncated logarithmic distribution.

(v) If $0 < \gamma < 1, -L\gamma < \lambda < 0$ then

$$\begin{aligned}\mathcal{I} &= \frac{1}{\left(\frac{\lambda}{\gamma}\right)_L \gamma^L} \sum_{y=L}^{\infty} \frac{\left(\frac{\lambda}{\gamma}\right)_y \gamma^y}{y!} = \frac{\sum_{k=0}^{\infty} \left(\frac{\lambda}{\gamma}\right)_k \gamma^k / k!}{\left(\frac{\lambda}{\gamma}\right)_L \gamma^L} \left[1 - \frac{\sum_{j=0}^{L-1} \left(\frac{\lambda}{\gamma}\right)_j \gamma^j / j!}{\sum_{k=0}^{\infty} \left(\frac{\lambda}{\gamma}\right)_k \gamma^k / k!}\right] \\ &= \frac{1}{\left(\frac{\lambda}{\gamma}\right)_L \gamma^L (1-\gamma)^{\lambda/\gamma}} \left[1 - (1-\gamma)^{\lambda/\gamma} \sum_{j=0}^{L-1} \left(\frac{\lambda}{\gamma}\right)_j \gamma^j / j!\right]\end{aligned}$$

and

$$p(y) = \frac{\left(\frac{\lambda}{\gamma}\right)_y \gamma^y (1-\gamma)^{\lambda/\gamma} / y!}{1 - (1-\gamma)^{\lambda/\gamma} \sum_{j=0}^{L-1} \left(\frac{\lambda}{\gamma}\right)_j \gamma^j / j!}.$$

A comparison of this expression with that at (A9) reveals a remarkable similarity to the case where $\gamma < 0$ and $\lambda > 0$. As in the earlier case we see that (i) the ratio λ/γ is negative, (ii) there is a scale factor reflecting left-truncation, with the only substantial difference being that whereas here we have a series reducing to the term $(1-\gamma)^{-\lambda/\gamma}$, in the earlier case we had a sum that only offers a similar simplification when λ/γ is integer.

(vi) The final case to consider is that where $\gamma = 1$ and $-L < \lambda < 0$. Here

$$\begin{aligned}\mathcal{I} &= \sum_{y=L}^{\infty} \frac{(\lambda)_y}{(\lambda)_L \, y!} = \frac{1}{(\lambda)_L} \left[{}_1F_0(\lambda; 1) - \sum_{j=0}^{L-1} (\lambda)_j / j!\right] \\ &= \frac{1}{(\lambda)_L} \left[(1-1)^{-\lambda} - \sum_{j=0}^{L-1} (\lambda)_j / j!\right] = -\frac{1}{(\lambda)_L} \sum_{j=0}^{L-1} \frac{(\lambda)_j}{j!}\end{aligned}$$

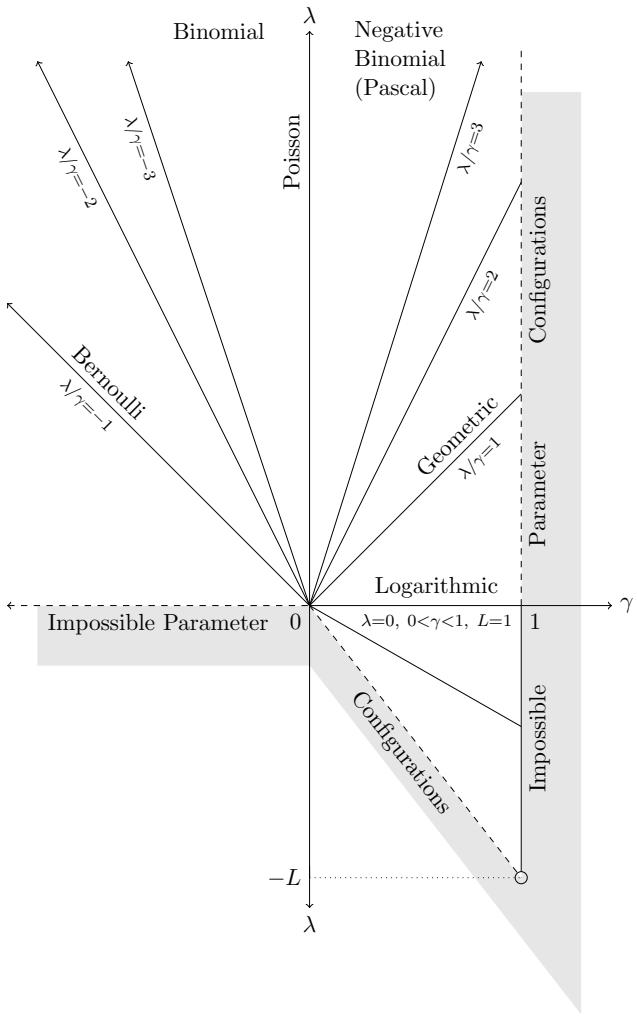
where the third equality is valid because $\lambda < 0$, and

$$p(y) = -\frac{(\lambda)_y / y!}{\sum_{j=0}^{L-1} (\lambda)_j / j!}, \quad y = 1, 2, 3, \ldots$$

This somewhat surprising result reduces to that of Willmot (1988) when $L = 1$, in which case the denominator reduces to unity.

We will not go through all the properties considered in the case $L = 0$, although we note in passing that $y = 0$ contributes nothing to any of the expectations used to calculate either the mean or variance of Y and so the expressions provided remain valid, except in the special case of $\gamma = 1$ where finite moments do not appear to exist. We can, however, update Figure A2 to reflect what we have learned in these cases

where $\{0\} \notin \mathcal{Y}$, see Figure A3.[28] In essence, the major change is that the parameter space now admits non positive values of λ, provided that they exceed $-L\gamma$ and $0 < \gamma < 1$, but only when $\{0\} \notin \mathcal{Y}$.



The set of permissible parameter combinations is represented by rays radiating from the origin, a subset of which are depicted, all other possibilities are ignored. The dashed lines are not included within this set and rays ending in a circle do not include the point depicted by the circle, specifically, if $\gamma = 1$ then $\lambda > -L$. The area shaded light grey represents the boundary of the set of impossible parameter configurations.

Figure A3. Parameter Space for the Extended Katz Family For $L \geq 0$.

References

Abadir, Karim M. 1999. An introduction to hypergeometric functions for economists. *Econometric Reviews* 18: 287–330. [CrossRef]

Adamidis, Konstantinos. 1999. An EM algorithm for estimating negative binomial parameters. *Australian & New Zealand Journal of Statistics* 41: 213–21. doi:10.1111/1467-842X.00075. [CrossRef]

[28] Note that Sundt and Jewell (1981, fig. 1) provide a similar diagram although, as noted by Willmot (1988), they miss the possibility of $\gamma = 1$.

Al-Khasawneh, Mohanad F. 2010. Estimating the negative binomial dispersion parameter. *Asian Journal of Mathematics & Statistics* 3: 1–15. doi:10.3923/ajms.2010.1.15. [CrossRef]

Bardwell, George E., and Edwin L. Crow. 1964. A two-parameter family of hyper-Poisson distributions. *Journal of the American Statistical Association* 59: 133–41. [CrossRef]

Boswell, M. T., and Ganapati P. Patil. 1970. Chance mechanisms generating the negative binomial distributions. In *Random Counts in Models and Structures*. Edited by G. P. Patil. London: University Press, vol. 1, chp. 1, pp. 3–22.

Cameron, A. Colin, and Pravin K. Trivedi. 1986. Econometric models based on count data: Comparisons and applications of some estimators and tests. *Journal of Applied Econometrics* 1: 29–53. [CrossRef]

Cameron, A. Colin, and Pravin K. Trivedi. 2013. *Regression Analysis of Count Data*, 2nd ed. Econometric Society Monographs No. 53. Cambridge: Cambridge University Press.

Consul, Prem C. 1989. *Generalized Poisson Distribution: Properties and Applications*. Statistics: Textbooks and Monographs 99. New York: Marcel Dekker Inc.

Crow, Edwin L., and George E. Bardwell. 1965. Estimation of the parameters of the hyper-Poisson distributions. In *Classical and Contagious Discrete Distributions. Proceedings of the International Symposium held at McGill University, Montreal, Canada, August 15–August 20, 1963*. Edited by G. P. Patil. Calcutta: Statistical Publishing Society; Oxford: Pergamon Press, pp. 127–40.

Dacey, Michael F. 1972. A family of discrete probability distributions defined by the generalized hypergeometric series. *Sankhyā: The Indian Journal of Statistics, Series B* 34: 243–50.

Davidson, Russell, and James G. MacKinnon. 1987. Implicit alternatives and the local power of test statistics. *Econometrica* 55: 1305–29. [CrossRef]

Dean, C. B. 1992. Testing for overdispersion in Poisson and binomial regression models. *Journal of the American Statistical Association* 87: 451–57. doi:10.2307/2290276. [CrossRef]

Dean, C. B., and J. F. Lawless. 1989. Tests for detecting overdispersion in Poisson regression models. *Journal of the American Statistical Association* 84: 467–72. doi:10.2307/2289931. [CrossRef]

Fang, Yue. 2003. GMM tests for the Katz family of distributions. *Journal of Statistical Planning and Inference* 110: 55–73. [CrossRef]

Frome, Edward L., Michael H. Kutner, and John J. Beauchamp. 1973. Regression analysis of Poisson-distributed data. *Journal of the American Statistical Association* 68: 935–40. doi:10.2307/2284525. [CrossRef]

Gart, John J. 1964. The analysis of Poisson regression with an application in virology. *Biometrika* 51: 517–21. [CrossRef]

Ghahfarokhi, Mohammad Ali Baradaran, Hosseiyn Iravani, and M. R. Sepehri. 2008. Application of Katz family of distributions for detecting and testing overdispersion in Poisson regression models. *World Academy of Science, Engineering and Technology 42*: 514–19.

Gilbert, Christopher L. 1979. Econometric models for discrete economic processes. Paper presented at Econometric Society European Meeting. Athens, Greece, September 3.

Gilbert, Christopher L. 1982. Economic models for discrete (integer valued) economic processes. In *Selected Papers on Contemporary Econometric Problems*. Edited by E. G. Charatsis. Athens: Athens School of Economics and Business Science, pp. 255–83.

Greene, William H. 2007. Functional form and heterogeneity in models for count data. *Foundations and Trends ® in Econometrics* 1: 113–218. doi:10.1561/0800000008. [CrossRef]

Greene, William H. 2008. Functional forms for the negative binomial model for count data. *Economics Letters* 99: 585–90. doi:10.1016/j.econlet.2007.10.015. [CrossRef]

Greenwood, M., and G. U. Yule. 1920. An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks or of repeated accidents. *The Journal of the Royal Statistical Society, Series A* 83: 255–79. [CrossRef]

Gurland, John. 2006. Katz system of distributions. In *Encyclopedia of Statistical Sciences*. Edited by S. Kotz, N. Balakrishnan, C. B. Read and B. Vidakovic. New York: John Wiley & Sons, Inc., vol. 6, pp. 3824–25. doi:10.1002/0471667196.ess1334.pub2. [CrossRef]

Haight, Frank A. 1967. *Handbook of the Poisson Distribution*. New York: John Wiley & Sons, Inc.

Hausman, Jerry, Bronwyn H. Hall, and Zvi Griliches. 1984. Econometric models for count data with an application to the patents-r & d relationship. *Econometrica* 52: 909–38.

Hellinger, Ernst. 1909. Neue begründung der theorie quadratischer formen von unendlichvielen veänderlichen. *Journal für die reine und angewandte Mathematik* 136: 210–71. [CrossRef]

Hess, Klaus Th., Anett Liewald, and Klaus D. Schmidt. 2002. An extension of Panjer's recursion. *ASTIN Bulletin* 32: 283–97. doi:10.2143/AST.32.2.1030. [CrossRef]

Hilbe, Joseph M. 2011. *Negative Binomial Regression*, 2nd ed. Cambridge: Cambridge University Press.

Hilbe, Joseph M. 2014. *Modeling Count Data*. New York: Cambridge University Press.

Joe, Harry, and Rong Zhu. 2005. Generalized Poisson distribution: The property of mixture of Poisson and comparison with negative binomial distribution. *Biometrical Journal* 47: 219–29. doi:10.1002/bimj.200410102. [CrossRef]

Johnson, Norman L., and Samuel Kotz. 1969. *Discrete Distributions*. New York: John Wiley & Sons, Inc.

Johnson, Norman L., Samuel Kotz, and Adrienne W. Kemp. 1993. *Univariate Discrete Distributions*, 2nd ed. New York: John Wiley & Sons, Inc.

Jorgenson, Dale W. 1961. Multiple regression analysis of a Poisson process. *Journal of the American Statistical Association* 56: 235–45. [CrossRef]

Katz, Leo. 1945. *Characteristics of Frequency Functions Defined by First Order Difference Equations*. Ph. D. thesis, University of Michigan, Ann Arbor, MI, USA.

Katz, Leo. 1946. On the class of functions defined by the difference equation $(x+1)f(x+1) = (a+bx)f(x)$ (Abstract). *Annals of Mathematical Statistics* 17: 501.

Katz, Leo. 1948. Frequency functions defined by the Pearson difference equation (Abstract). *Annals of Mathematical Statistics* 19: 120.

Katz, Leo. 1965. Unified treatment of a broad class of discrete distributions. In *Classical and Contagious Discrete Distributions. Proceedings of the International Symposium held at McGill University, Montreal, Canada, August 15–August 20, 1963*. Edited by G. P. Patil. Calcutta: Statistical Publishing Society; Oxford: Pergamon Press, pp. 175–82.

Kemp, Adrienne W. 1968. A wide class of discrete distributions and the associated differential equations. *Sankhyā: The Indian Journal of Statistics, Series A (1961–2002)* 30: 401–10.

King, Gary. 1989. Variance specification in event count models: From restrictive assumptions to a generalized estimator. *American Journal of Political Science* 33: 762–84. [CrossRef]

King, Maxwell L. 1987. Towards a theory of point optimal testing. *Econometric Reviews* 6: 169–218. [CrossRef]

King, Maxwell L., and Sivagowry Srianantha kumar. 2015. Point optimal testing: A survey of the post 1987 literature. *Model Assisted Statistics and Applications* 10: 179–96. doi:10.3233/MAS-150323. [CrossRef]

Lawless, Jerald F. 1987a. Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics* 15: 209–25. [CrossRef]

Lawless, J. F. 1987b. Regression methods for Poisson process data. *Journal of the American Statistical Association* 82: 808–15. doi:10.2307/2288790. [CrossRef]

Lee, Lung-Fei. 1986. Specification test for Poisson regression models. *International Economic Review* 27: 689–706. [CrossRef]

McCullagh, Peter, and John A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. Monographs On Statistics and Applied Probability 37. London: Chapman & Hall\CRC. doi:10.1007/978-1-4899-3242-6. [CrossRef]

Miller, David W. 1998. *Fitting Frequency Distributions Philosophy and Practice. Part 1: Discrete Distributions*, 2nd ed. Self-published.

Nelder, John A., and Robert W. M. Wedderburn. 1972. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)* 135: 370–84. [CrossRef]

Ord, J. Keith. 1967a. On a system of discrete distributions. *Biometrika* 54: 649–56. [CrossRef]

Ord, J. Keith. 1967b. *On Families of Discrete Distributions*. Ph. D. thesis, University of London, London, UK.

Ord, J. Keith. 1972. *Families of Frequency Distributions*. London: Griffin.

Panjer, Harry H. 1981. Recursive evaluation of a family of compound distributions. *ASTIN Bulletin* 12: 22–26. doi:10.1017/S0515036100006796. [CrossRef]

Pearson, Karl. 1894. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 185: 71–110 (+ 5 plates). doi:10.1098/rsta.1894.0003. [CrossRef]

Pearson, Karl. 1895. Contributions to the mathematical theory of evolution. — II. Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 186: 343–414 (+ 10 plates). doi:10.1098/rsta.1895.0010. [CrossRef]

Pestana, Dinis D., and Sílvio F. Velosa. 2004. Extensions of Katz-Panjer families of discrete distributions. *REVSTAT Statisical Journal* 2: 145–62.

Qu, Yinsheng, G. J. Beck, and G. W. Williams. 1990. Polya-Eggenberger distribution: Parameter estimation and hypothesis tests. *Biometrical Journal* 32: 229–42. doi:10.1002/bimj.4710320215. [CrossRef]

Raschke, Christian, and William H. Greene. 2010. Corrigendum to "functional forms for the negative binomial model for count data". *Economics Letters* 107: 313. doi:10.1016/j.econlet.2007.10.015. [CrossRef]

Slater, Lucy J. 1966. *Generalized Hypergeometric Functions*. Cambridge: Cambridge University Press.

Staff, P. J. 1964. The displaced Poisson distribution. *Australian Journal of Statistics* 6: 12–20. doi:10.1111/j.1467-842X.1964.tb00146.x. [CrossRef]

Staff, P. J. 1967. The displaced Poisson distribution. Region B. *Journal of the American Statistical Association* 62: 643–54.

Sundt, Bjørn, and William S. Jewell. 1981. Further results on recursive evaluation of compound distributions. *ASTIN Bulletin* 12: 27–39. doi:10.1017/S0515036100006802. [CrossRef]

Weisstein, Eric W. 2019. Gauss's Test. From MathWorld — A Wolfram Web Resource. Available online: http://mathworld.wolfram.com/GausssTest.html (accessed on 21 December 2019).

Willmot, Gordon. 1988. Sundt and Jewell's family of discrete distributions. *ASTIN Bulletin* 18: 17–29. doi:10.2143/AST.18.1.2014957. [CrossRef]

Winkelmann, Rainer. 2008. *Econometric Analysis of Count Data*, 5th ed. Berlin: Springer. doi:10.1007/978-3-540-78389-3. [CrossRef]

Yang, Zhao, James W. Hardin, Cheryl L. Addy, and Quang H. Vuong. 2007. Testing approaches for overdispersion in Poisson regression versus the generalized Poisson model. *Biometrical Journal* 49: 565–84. doi:10.1002/bimj.200610340. [CrossRef] [PubMed]

Yang, Zhao, James W. Hardin, and Cheryl L. Addy. 2009. A score test for overdispersion in Poisson regression based on the generalized Poisson-2 model. *Journal of Statistical Planning and Inference* 139: 1514–21. doi:10.1016/j.jspi.2008.08.018. [CrossRef]