Barakat, Bilal Fouad

Working Paper

Generalised Poisson Distributions for Modelling Parity

Vienna Institute of Demography Working Papers, No. 7/2016

Provided in Cooperation with:
Vienna Institute of Demography (VID), Austrian Academy of Sciences

Suggested Citation: Barakat, Bilal Fouad (2016) : Generalised Poisson Distributions for Modelling Parity, Vienna Institute of Demography Working Papers, No. 7/2016, Austrian Academy of Sciences (ÖAW), Vienna Institute of Demography (VID), Vienna, https://doi.org/10.1553/0x003cd01c

This Version is available at:
https://hdl.handle.net/10419/156315

VIENNA INSTITUTE OF DEMOGRAPHY
Working Papers
7/2016

Bilal Barakat

Generalised Poisson Distributions for Modelling Parity

Vienna Institute of Demography
Austrian Academy of Sciences

Welthandelsplatz 2 / Level 2
1020 Vienna · Austria

E-Mail: vid@oeaw.ac.at
Website: www.oeaw.ac.at/vid

Abstract

Conventional parametric count distributions, namely the Poisson and Negative-Binomial models, do not offer satisfactory descriptions of empirical distributions of completed cohort parity. One reason is that they cannot model variance-to-mean ratios below unity, that is, underdispersion, which is typical of low-fertility parity distributions. Statisticians have relatively recently revived two generalised count distributions that can model both overdispersion and underdispersion, but that have to date not attracted the attention of demographers. The objective of this note is to assess the utility of these distributions, the Conway-Maxwell-Poisson and Gamma Count models, for the modelling of parity distributions, using both simulations and maximum-likelihood fitting to empirical data from the Human Fertility Database (HFD). The results show that these generalised count distributions offer a greatly improved fit compared to customary Poisson and Negative-Binomial models in the presence of underdispersion, without loss of performance in the presence of equi- or overdispersion.

Keywords

Fertility, parity, discrete probability distributions, underdispersion.

Author

Bilal Barakat, Wittgenstein Centre for Demography and Global Human Capital (IIASA, VID/OAW, WU), Vienna, Austria. Email: bilal.barakat@oeaw.ac.at

Generalised Poisson Distributions for Modelling Parity

Bilal Barakat

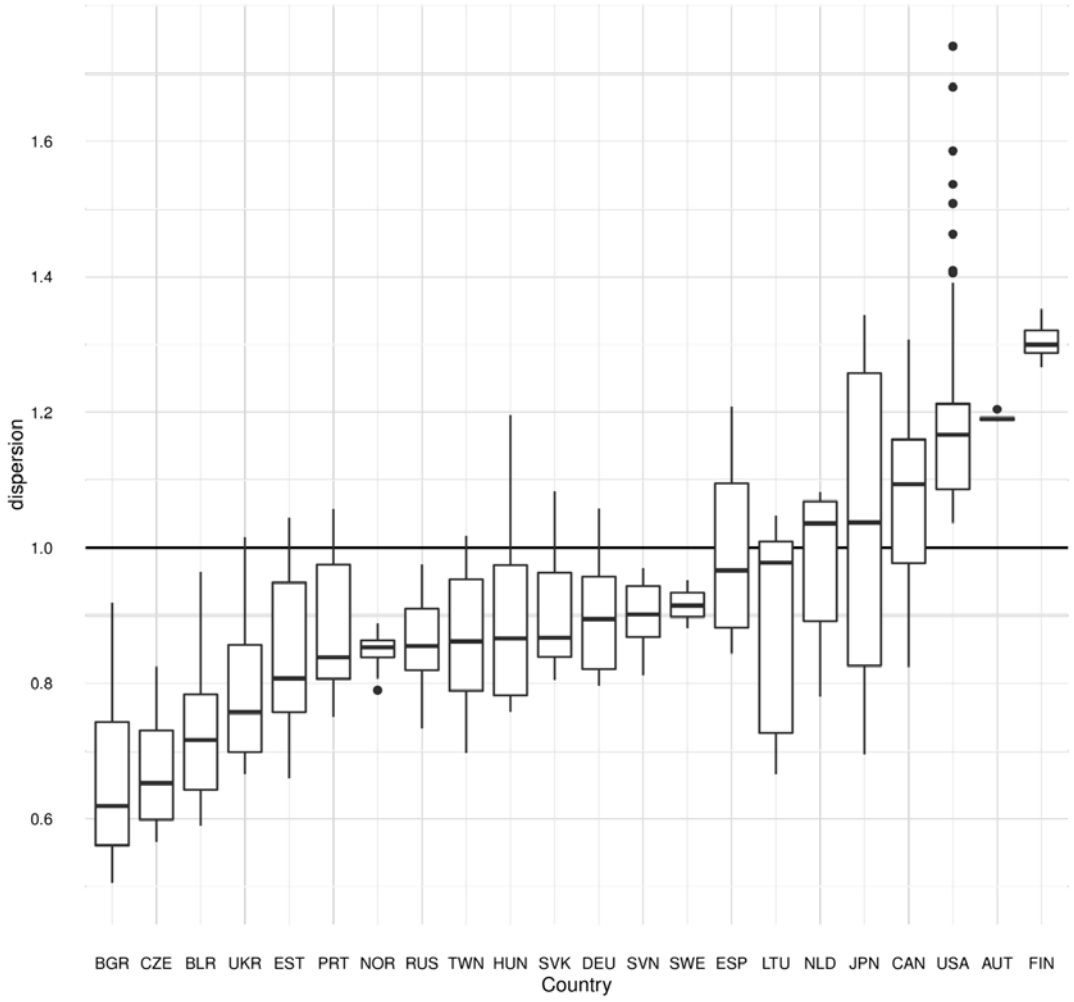
1. Introduction

The number of live births a women experiences, her parity, is an integer count. The statistical analysis of parities therefore requires the use of discrete count distributions. (Fully non-parametric approaches are an alternative in some, but by no means all applications, and suffer serious disadvantages of their own, notably a lack of analytic parsimony.) Unfortunately, while there is a vast number of distributions for *continuous* outcomes, the analysis of discrete count outcomes, including in demography, has until recently been limited to a choice between two: the Poisson distribution, and the Negative-Binomial distribution.

It is well-known that, by construction, the mean of a Poisson distribution equals its variance. Equivalently, its variance-to-mean ratio equals one, a measure also known as statistical dispersion. Accordingly, a Poisson distribution can be expected to fit only poorly to an empirical distribution whose dispersion differs considerably from unity. In applied statistics generally, attention has largely focused on the need to account for *overdispersion*, that is, distributions with dispersion considerably larger than one. This is unsurprising, given that mixtures of Poisson distributions are always overdispersed, and heterogeneity is one of the most common ways in which we expect reality to diverge from simple statistical models. Indeed, both the Negative-Binomial distribution, as well as the 'zero-inflated' Poisson, where a certain share of structurally-zero outcomes are assumed, can be mathematically interpreted as special cases of Poisson mixtures, and like the regular Poisson distribution, are structurally unable to model *underdispersion*.

The need for alternative modelling options in the context of human birth parities arises from the fact that in low-fertility settings *underdispersion* is actually the norm, in other words: parity distributions whose mean considerably exceeds their variances. Even severe underdispersion is far more common than overdispersion. Figure 2 demonstrates this using data from the Human Fertility Database (HFD) (additional details regarding the data are provided in Section 3.1).

Figure 1: Statistical dispersion (variance-to-mean ratio) in completed cohort parity distributions of the HFD



Statistically speaking, underdispersion could arise as a consequence of any of the following: a) a positive inter-personal correlation in terms of the child count ('If others have more/fewer children, so will I.'); or of mechanisms that diminish the occurrance of 'runaway' parity, where some women tend towards extremely high birth counts, while others are 'stuck' at low levels, namely b) a parity progression rate that is negatively related to the parity already achieved ('The more children I already have, the less likely I am to have more.'); or c) a parity progression rate that is positively correlated with the waiting time since the last birth ('The longer it's been since the last birth, the more likely I am to have another child.').

Only during the last decade have two instances of 'generalised Poisson distributions' that formalise the latter two effects and thereby allow for parametric modeling of *underdispersed* counts seen a modest 'revival' in applied statistics. They do not, however, appear to have been exploited in demographic analysis yet, despite the underdispersion of parity counts. The

present aim is to begin to fill this gap by providing an introduction and first assessment of their utility to demographers.

2. Generalized Count Distributions

2.1.The Conway-Maxwell-Poisson Distribution (COM-Poisson)

This distribution was originally proposed by Conway and Maxwell (1962) and more recently revived by Shmueli et al. (2005). It generalises the standard Poisson distribution by allowing the probabilities to decay more rapidly or more slowly as the distance from the mean increases. As such, it formalises mechanism b) above. Formally, its probability function takes the form:

$$P(Y = n) = \frac{\lambda^n}{(n!)^\nu} \frac{1}{Z(\lambda, \nu)},$$

for $n = 0,1,2, \ldots$, where the normalising constant is $Z(\lambda, \nu) = \sum_{i=0}^{inf} \frac{\lambda^i}{(i!)^\nu}$, and the parameters must satisfy the constraints $\lambda > 0, \nu \geq 0$. This specification implies that the ratios of successive probabilities can be expressed as:

$$\frac{Y = n - 1}{Y = n} = \frac{n^\nu}{\lambda}.$$

For $\nu = 1$ this reduces to the regular Poisson case, while $\nu \leq 1$ and $\nu \geq 1$ result in overdispersion respectively underdispersion.

2.2.The Gamma Count Distribution

The Gamma count distribution arises from the assumption that the waiting times between births follow a Gamma distribution (rather than an exponential distribution, as in the Poisson model). In other words, it formalises mechanism c) mentioned above. Depending on the parameters of this Gamma distribution, the hazard can be modelled to increase or decrease as a function of the waiting time, corresponding to underdispersion and overdispersion respectively. Accessible derivations for the Gamma Count model are provided by Winkelmann (2008), including asymptotics.

Specifically, the Gamma count model takes the following form:

$$P(Y = n) = G(\alpha n, \beta T) - G(\alpha(n + 1), \beta T)$$

for $n = 0,1,2, \ldots$, where $G(\alpha k, \beta T)$ is the regularized lower incomplete Gamma function

$$G(\alpha n, \beta T) = \frac{1}{\Gamma(n\alpha)} \int_0^{\beta T} u^{n\alpha-1} exp^{-u} du,$$

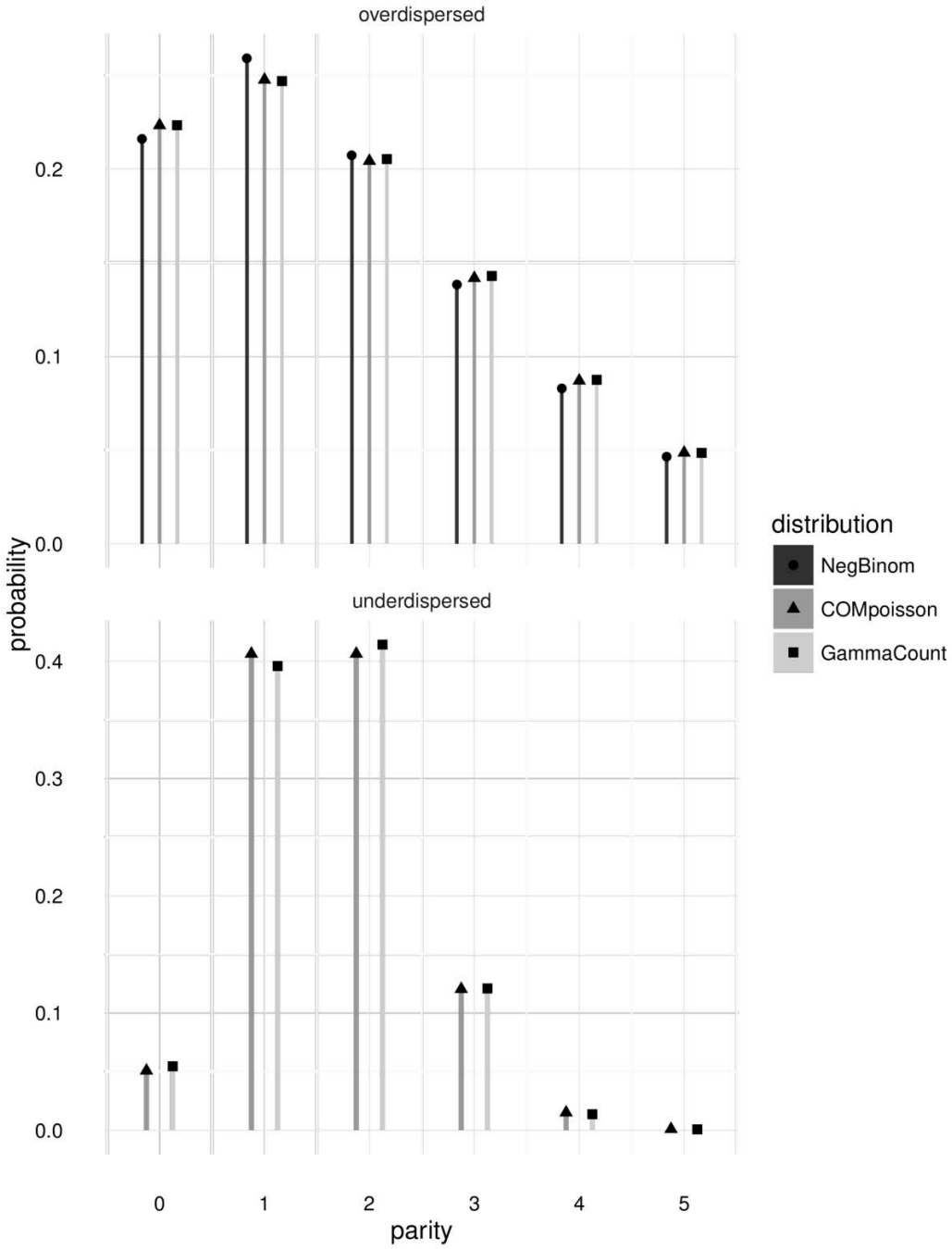
and T is the scale of the overall exposure period. We have $\alpha, \beta > 0$ and $G(0, \beta T) \equiv 1$ by assumption. In our setting, $T = 1$ may be assumed without loss of generality. Then α is the dispersion factor, $\frac{\alpha}{\beta}$ is the mean waiting time between births, and—asymptotically—$\frac{\beta}{\alpha}$ is the expected count—although this approximation can be poor in the parameter range of interest for fertility applications and should not be relied on. For $\alpha = 1$, the model reduces to the special case of Poisson counts. The regression is for the waiting times, so that we may equate the linear predictor with $\lambda_i = \frac{\alpha}{\beta_i}$

,A well-known property of the standard Poisson model is that the number of events per unit of exposure is a sufficient statistic for the mean. One implication is that two individuals experiencing ten years of exposure and one event each, and one individual experiencing twenty years of exposure and two events, result in the same estimate. In practice, this is frequently exploited to allow for the aggregation of data without loss of information: for any given values of possible covariates, only the total amount of exposure and total number of events needs to be recorded. It is important to note that this operational shortcut is not possible with the generalised Poisson models! Because the hazard is a non-constant function of the waiting time or parity already attained, the way the eventless episodes are distributed between individuals does matter.

2.3.Gamma Count as a Swiss Army Knife

Fortunately, it turns out that what initially appears as a potentially confusing proliferation of options for modelling count data ultimately leads to a simplification. Figure 3 shows both the COM-Poisson and Gamma Count model fitted to an overdispersed target drawn from a Negative Binomial distribution, and the Gamma Count model fitted to an underdispersed target drawn from a COM-Poisson distribution. This illustrates two points. Firstly, that the availability of fully generalised count distributions makes the Negative Binomial model redundant for practical purposes, because it can be well-approximated when the COM-Poisson or Gamma Count models are set to be overdispersed. Secondly, that their unique ability to model underdispersion sets the COM-Poisson and Gamma Count distributions apart from other count models, but not from each other, because they can mimic each other very closely. This was tested for this study across the entire parameter range of interest in fertility applications. As a matter of fact, the case displayed in Figure 2 displays the maximal discrepancy found, with the typical error being an order of magnitude smaller than the one shown here.

Figure 2: Simulated parity distributions resulting from Maximum-Likelihood fits of: COM-Poisson and Gamma count models to an overdispersed target distribution sampled from a Negative Binomial distribution (top panel), and a Gamma count model to an underdispersed target sampled from a COM-Poisson distribution (bottom panel)



The effective equivalence of these two distributions is striking because their theoretical derivations bear no obvious relationship to each other, and no formal mathematical (asymptotic?) equivalence appears to have been established in the literature. Indeed, Winkelmann (2008) does not mention the COM-Poisson model in his derivation of the Gamma count model, or indeed at all in his book on count models. Conversely, neither do Shmueli et al. (2005), who established the statistical properties of COM-Poisson, cite Winkelmann or mention the Gamma count model. While it seems unlikely that the almost perfect match between the two distributions is coincidental, the question of their formal mathematical relationship is not pursued further here.

The marginal effective difference between the COM-Poisson and Gamma Count distributions does not make them entirely redundant, however. Firstly, the conceptual derivations are different, so depending on the argument being made, in particular whether it focuses on parity progression or on waiting times, either the COM-Poisson model or Gamma count model may be more appropriate. Secondly, and following from the first point, for regression analysis the choice of model is dictated by whether the dependent variable is mean waiting time or mean birth count directly. Thirdly, the *purpose* of modelling may be decisive, because the two distributions differ in their practical properties. On the one hand, the statistical properties of the COM-Poisson model have been investigated more fully (Sellers and Shmueli 2010), and asymptotic significance tests are available. For simulations, on the other hand, the Gamma Count model has the advantage that it appears to be vastly more efficient computationally, by one or two orders of magnitude. This is no doubt due to the fact that the underlying Gamma function benefits from being a common mathematical function for which highly optimised algorithms are standard. For illustration, on the system used to generate this report, fitting the Gamma count distribution to a Negative Binomial target one hundred times took 0 seconds of computation time, compared to 3 for the COM-Poisson distribution. Since it is also much faster to sample from, the Gamma count model is also preferable for inferential approaches involving frequent (re-)sampling from the distribution, namely both bootstrapping and Bayesian inference.

So while there is a role for the COM-Poisson distribution for certain applications, a case can be made to consider the Gamma Count distribution as a general-purpose default for modelling both over- and underdispersed distributions of human birth parities.

3. Empirical Analysis

3.1.Completed Cohort Parity from the Human Fertility Database

The Human Fertility Database, at least with respect to time series of completed parity, is focused on industrialised high-income countries. Cumulative fertility rates by birth order from the HFD were extracted for all countries for which these were available at the time of writing. The strength of these data for present purposes is the fact that they carefully account for exposure rates and mortality, and that they provide a consistent longitudinal perspective. The

limitations (for present purposes) are, firstly, that, even for the earliest cohorts, only a relatively limited range of average fertility levels are represented, namely the low fertility end of the spectrum, and secondly, that high parities are aggregated at 5+. As the aggregation occurs at the level of cumulative fertility rates (CCFR), the calculated share at parity 4 is also affected, corresponding to the difference between CCFR4 and CCFR at exactly 5 (rather than 5+). Three different synthetic assumptions about the true spread of the reported CCFR5+ over parities 5-10 were tried, namely a uniform distribution, and a linear or exponential decline. All presented analyses are based on the exponential model, but, unless noted otherwise, the conclusions are qualitatively robust in the sense of not being sensitive to the choice of imputation.

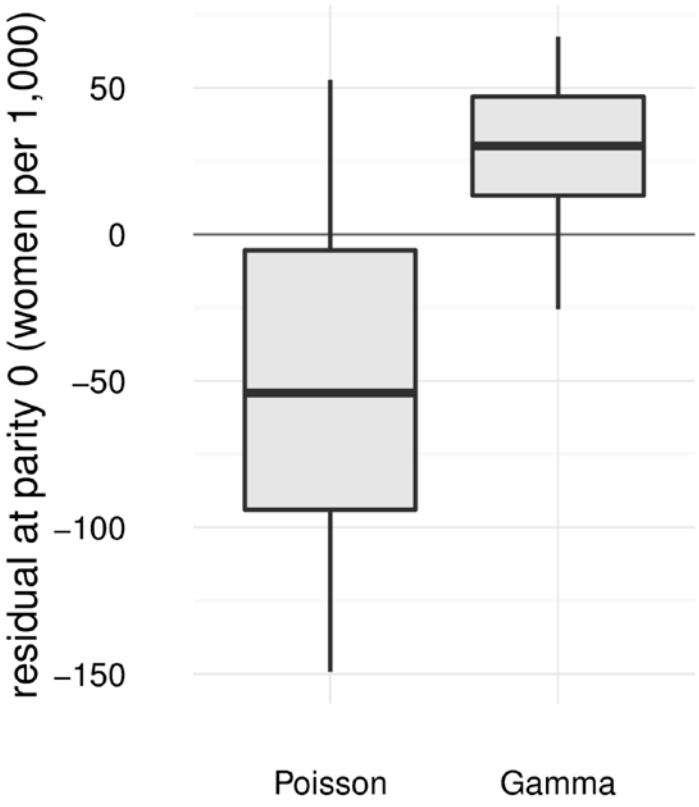
3.2.Zero-Inflation

Before comparing the *overall* fit of generalised and regular Poisson distributions to the empirical HFD data, it is worthwhile examining the residual at parity 0 separately. It is common in count data models that zero counts have a special status vis-à-vis higher counts.

In the following, I therefore restrict attention to the case of zero-inflation. Recall the key difference between a zero-inflation model and a hurdle model, namely that zero-inflation assumes that cases of parity 0 are contributed by two sources, a fixed zero group and some of the observations sampled from the basic distribution, whereas in the hurdle model, cases of parity 0 are contributed *only* by those not crossing the initial hurdle. In the context of modelling birth parities, the former is more plausible than the latter, since even women who do cross the hurdle of wanting to have children and being able to have them may nevertheless end up childless by chance.

To gain some insight into the relationship between zero-inflation on the one hand and the regular Poisson and Gamma count distributions on the other (the Negative Binomial and COM-Poisson models add no information since their fits are each practically identical to one shown), Figure 3 displays the absolute residual at parity 0, in other words, the share of “excess” or “missing” zeroes in the data.

Figure 3: Absolute residuals (observed – fitted) at parity 0 of Maximum-Likelihood fits of different models to empirical HFD distributions of completed cohort parity in terms of women per 1,000



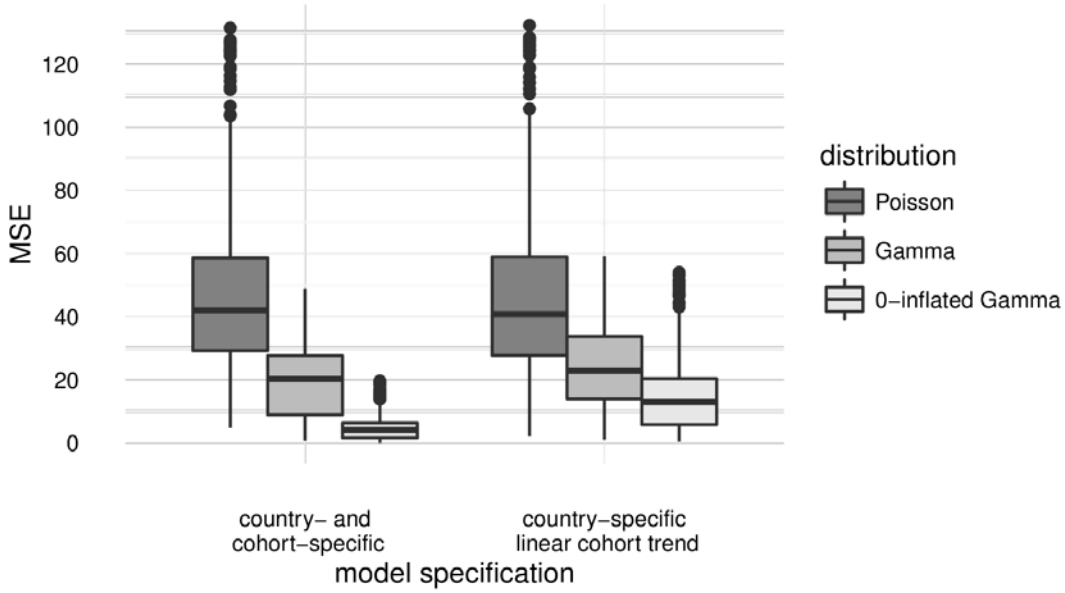
It is evident that the regular Poisson model actually predicts *too many* zeroes in general, rather than too few. Actually, this is unsurprising given we know the vast majority of HFD parity distributions to be underdispersed. By observing that a probability point mass at zero can actually be interpreted as a Poisson distribution with mean and variance zero, it becomes clear that the zero-inflated Poisson model is actually a special case of a mixture of Poisson distributions. Accordingly, it always results in *overdispersion*. An underdispersed distribution is therefore unlikely to exhibit excess zeroes relative to a regular Poisson distribution.

The presence of excess zeroes relative to the underdispersed Gamma count baseline is positive for two reasons. Substantively, we know that the true data generating process, namely human fertility, does in fact involve a small proportion of women whose probability of giving birth is close to zero. From this point of view, the presence of moderate zero-inflation relative to the Gamma count distribution actually raises its plausibility. Moreover, in practical terms, it allows for improving the fit to the data by explicitly taking zero-inflation into account—an option not available to the Poisson (or indeed, the Negative Binomial) model, that is already overestimating the proportion at parity 0.

3.3.Fits to Empirical Parity Distributions

With this in mind, Figure 4 compares the fits of the regular Poisson, Gamma Count, and zero-inflated Gamma Count models to the empirical HFD completed cohort birth parity distributions, in terms of Mean Squared Error based on the original scale of women per 1,000. To simplify the presentation, the redundant Negative Binomial and COM-Poisson fits are omitted. The former is redundant because most of the observed distributions are underdispersed, and so the Negative Binomial would reduce to the Poisson case. The latter is redundant because we already established that the COM-Poisson and Gamma Count distributions closely mimic each other in the relevant parameter range, and therefore perform approximately equally well in fitting the empirical data.

Figure 4: Mean Squared Error of Maximum-Likelihood fits of different models to empirical HFD distributions of completed cohort parity in terms of women per 1,000



The left set of boxplots shows the fits to each country-and-cohort-specific parity distribution individually. Of course, the mere fact that the Gamma Count distribution fits the data better than the Poisson distribution, and that the zero-inflated Gamma Count distribution fits better still, is to be expected, given that the number of independent parameters (and degrees of freedom) increases from one to two to three as we move through these models. However, even the three-parameter zero-inflated Gamma Count distribution cannot be said to be overfitted. Technically the fit here is to eleven data points for each distribution, namely parities 0 through 10, but even restricting attention to those with meaningful frequencies, 0 to 5, say, still leaves the data with six degrees of freedom. So even the most 'complex' of the three models is still parsimonious, with the additional parameters all enjoying meaningful substantive interpretations and achieving a fit as close to perfect as one can hope for in modelling natural phenomena. With a typical Mean Squared Error of less than nine in the vast

majority of cases, and typically closer to four, the zero-inflated Gamma Count model is generally within two or three women per 1,000 of the true parity share. Moreover, the right set of boxplots demonstrates that the great improvement in fit over the Poisson distribution is certainly not due to approximating a saturated model: this specification assumes linear (over cohorts) country-specific trends in each parameter, and therefore uses two (Poisson), four (Gamma Count), respectively six (zero-inflated Gamma Count) parameters to fit all parities 0 through 10 for between 6 (Austria, Finland) and 57 (USA) cohorts at once. This is not proposed as an appropriate, much less optimal, regression specification, merely to illustrate that the advantage of generalised count distributions in the presence of underdispersion remains considerable even in applications more typical of real-life research.

References

Conway, Richard W, and William L Maxwell. 1962. “A Queuing Model with State Dependent Service Rates.” *Journal of Industrial Engineering* 12 (2): 132–36.

Sellers, K.F., and G. Shmueli. 2010. “A Flexible Regression Model for Count Data.” *The Annals of Applied Statistics* 4 (2). Institute of Mathematical Statistics: 943–61.

Shmueli, G., T.P. Minka, J.B. Kadane, S. Borle, and P. Boatwright. 2005. “A Useful Distribution for Fitting Discrete Data: Revival of the Conway–Maxwell–Poisson Distribution.” *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 54 (1). Wiley Online Library: 127–42.

Winkelmann, R. 2008. *Econometric Analysis of Count Data*. Springer Verlag.

VIENNA INSTITUTE OF DEMOGRAPHY

Working Papers

Kohlenberger, Judith, Isabella Buber-Ennser, Bernhard Rengs and Zakarya Al Zalak, *A Social Survey on Asylum Seekers in and around Vienna in Fall 2015: Methodological Approach and Field Observations*, VID Working Paper 6/2016.

Barakat, Bilal and Robin Shields, *Just Another level? Comparing Quantitative Patterns of Global School and Higher Education Expansion*, VID Working Paper 5/2016.

Bloom, David E., Michael Kuhn and Klaus Prettner, *Africa's Prospects for Enjoying a Demographic Dividend*, VID Working Paper 4/2016.

Frankovic, Ivan, Michael Kuhn and Stefan Wrzaczek, *Medical Care within an OLG Economy with Realistic Demography*, VID Working Paper 3/2016.

Abel, Guy J., *Estimates of Global Bilateral Migration Flows by Gender between 1960 and 2015*, VID Working Paper 2/2016.

Testa, Maria Rita, Valeria Bordone, Beata Osiewalska and Vegard Skirbekk, *The Relation between Mother's Socio-Economic Status and Daughter's Fertility Intentions in Austria, Italy, Bulgaria, and Norway*, VID Working Paper 1/2016.

Hoffmann, Roman and Raya Muttarak, *A Tale of Disaster Experience in Two Countries: Does Education Promote Disaster Preparedness in the Philippines and Thailand*, VID Working Paper 9/2015.

Klotz, Johannes and Richard Gisser, *Mortality Differentials by Religious Denomination in Vienna 1981-2002*, VID Working Paper 8/2015.

Steiber, Nadia and Barbara Haas, *Overworked or Underemployed? Actual and Preferred Household Employment Patterns in the Context of the Economic Crisis*, VID Working Paper 7/2015.

Beaujouan, Eva, Zuzanna Brzozowska and Krystof Zeman, *Childlessness Trends in Twentieth-Century Europe: Limited Link to Growing Educational Attainment*, VID Working Paper 6/2015.

Abel, Guy, *Estimates of Global Bilateral Migration Flows by Gender between 1960 and 2010*, VID Working Paper 5/2015.

Spijker, Jeroen, *Alternative Indicators of Population Ageing: An Inventory*, VID Working Paper 4/2015.
