DEMOGRAPHIC RESEARCH
A peer-reviewed, open-access journal of population sciences

DEMOGRAPHIC RESEARCH

VOLUME 36, ARTICLE 26, PAGES 745–758
PUBLISHED 14 MARCH 2017
http://www.demographic-research.org/Volumes/Vol36/26/
DOI: 10.4054/DemRes.2017.36.26

Descriptive Finding

Generalised count distributions for modelling parity

Bilal Barakat

Contents

1	Introduction	746
2	Generalised count distributions	748
2.1	The Conway-Maxwell-Poisson distribution (COM-Poisson)	748
2.2	The gamma count distribution	749
2.3	General comments	749
2.4	Gamma count as a Swiss army knife	750
3	Empirical analysis	752
3.1	Completed cohort parity from the Human Fertility Database	752
3.2	Zero inflation	753
3.3	Fits to empirical parity distributions	755
4	Conclusion	756
	References	757

Demographic Research: Volume 36, Article 26
Descriptive Finding

Generalised count distributions for modelling parity

Bilal Barakat[1]

Abstract

BACKGROUND
Parametric count distributions customarily used in demography – 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, i.e., underdispersion, which is typical of low-fertility parity distributions. Statisticians have recently revived two generalised count distributions that can model both over- and underdispersion, but they have not attracted demographers' attention to date.

OBJECTIVE
The objective of this paper is to assess the utility of these alternative general count distributions, namely the Conway-Maxwell-Poisson and gamma count models, for the modeling of distributions of completed parity.

METHODS
Simulations and maximum-likelihood estimation are used to assess their fit to empirical data from the Human Fertility Database (HFD).

RESULTS
The results show that the generalised count distributions offer a dramatically improved fit compared to customary Poisson and negative binomial models in the presence of underdispersion, without performance loss in the case of equidispersion or overdispersion.

CONCLUSIONS
This gain in accuracy suggests generalised count distributions should be used as a matter of course in studies of fertility that examine completed parity as an outcome.

CONTRIBUTION
This note performs a transfer of the state of the art in count data modelling and regression in the more technical statistical literature to the field of demography, allowing demographers to benefit from more accurate estimation in fertility research.

[1] Österreichische Akademie der Wissenschaften, Vienna Institute of Demography, Austria. E-Mail: bilal.barakat@oeaw.ac.at.

1. Introduction

The number of live births a woman experiences, her parity, is an integer count. The statistical analysis of parities therefore requires the use of discrete count distributions. Fully nonparametric 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, the choice between parametric count distribution has until recently effectively been limited to the Poisson distribution and the negative binomial distribution, including in demographic analysis (e.g., Parrado and Morgan 2008; Nisén et al. 2014).

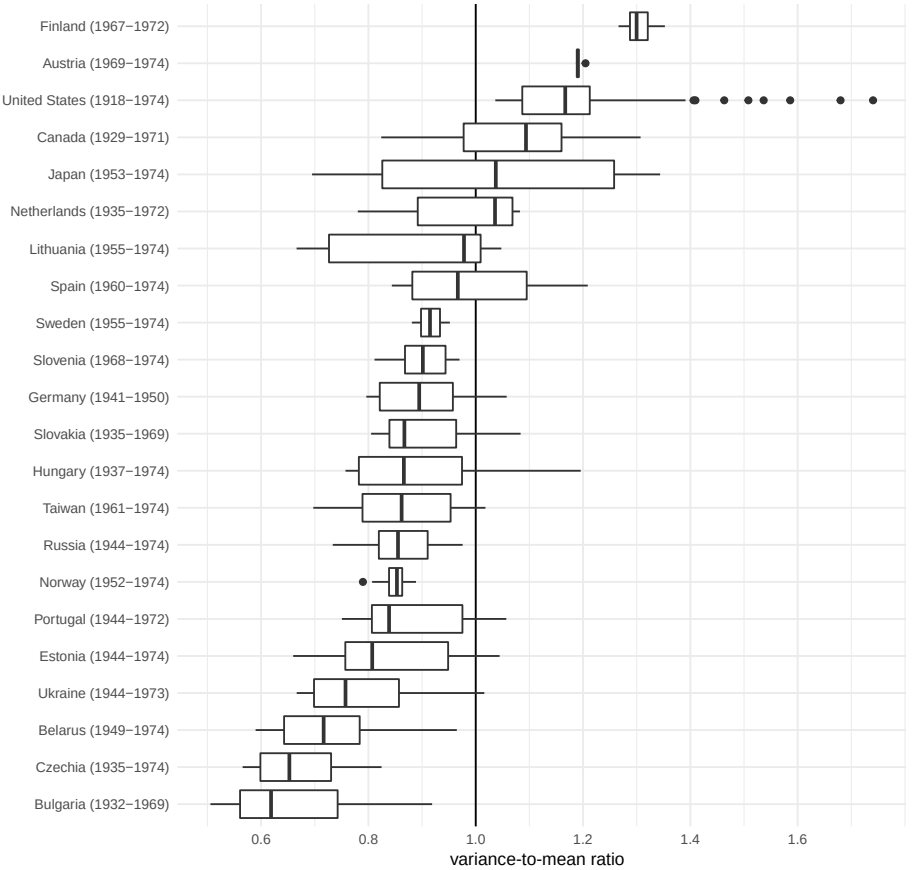
It is well known that the variance-to-mean ratio of a Poisson distribution, its dispersion, equals one by construction. Accordingly, Poisson distributions fit poorly to empirical distributions whose dispersion differs considerably from unity. In applied statistics generally, attention has largely focused on the need to account for overdispersion, that is, 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 and the 'zero-inflated' Poisson, where a certain share of structural-zero outcomes are assumed, can be mathematically interpreted as special cases of Poisson mixtures, and, like the regular Poisson distribution, they are structurally unable to model underdispersion.

The need for alternative modeling 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. Figure 1 demonstrates this using data from the Human Fertility Database (HFD).

Statistically speaking, underdispersion could arise as a consequence of: a) a positive interpersonal correlation in terms of the child count, b) 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 a parity progression rate that is negatively related to the parity already achieved, or c) a parity progression rate that is positively correlated with the waiting time since the last birth.

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 common 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.

Figure 1: Statistical dispersion (variance-to-mean ratio) in completed cohort parity distributions of the HFD, by country, across birth cohorts (in parentheses)



Note: Thick horizontal line: median across cohorts; box: inter-quartile range (IQR); whiskers: observations within 1.5 times IQR; points: outliers.

2. Generalised 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 $Z(\lambda, \nu) = \sum_{i=0}^{\infty}, \frac{\lambda^i}{(i!)^\nu}$ and the parameters must satisfy the constraints $\lambda > 0, \nu \geq 0$. Parameters λ and ν may be interpreted as representing the rate and dispersion of the distribution in the general sense that for fixed ν the mean of the distribution increases with λ, and that for fixed λ, the distribution becomes more spread out for decreasing ν. Unfortunately, no closed-form expression exists to relate these parameters to the distribution's moments directly. Our intuitive interpretation of the parameters must therefore rest on the fact that the ratios of successive probabilities can be expressed simply as:

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

This means that λ determines the overall relationship between the probabilities at successive parities while ν determines how these relationships change with increasing parity. For $\nu = 1$ this reduces to the regular Poisson case, while $\nu < 1$ and $\nu > 1$ result in overdispersion or underdispersion, respectively. In the HFD data analysed here, the estimated values of λ or range from 1.4 to 27.8, and ν ranges from 0.8 to 3.8. This excludes 17 country-cohort dyads (out of 579) with extremely low or high variance-to-mean ratios for which the estimates suffered from convergence problems.

2.2 The gamma count distribution

The gamma count distribution formalises mechanism c) mentioned above and assumes that the waiting times between births follow a gamma distribution (rather than an exponential distribution, as in the Poisson model). 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:

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

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

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

and T is the scale of the overall exposure period. We have $\alpha, \beta \in \mathbf{R}^+$ and $G(0, \beta T) \equiv 1$ by assumption. In our setting $T = 1$ may be assumed without loss of generality. Asymptotically, α is the dispersion factor, $\frac{\alpha}{\beta}$ is the mean waiting time between births, and $\frac{\beta}{\alpha}$ approaches mean parity. Note, however, that these asymptotic approximations 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, and $\alpha < 1$ and $\alpha > 1$ result in overdispersion or underdispersion. In the HFD data analysed here, the estimated values of α range from 0.8 to 4.5, and β ranges from 1.4 to 10.8.

2.3 General comments

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. 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 neither the COM-Poisson nor the gamma count model allow for this operational shortcut. Because the hazard is a nonconstant function of the waiting time or parity already attained, the way the eventless episodes are distributed between individuals does matter.

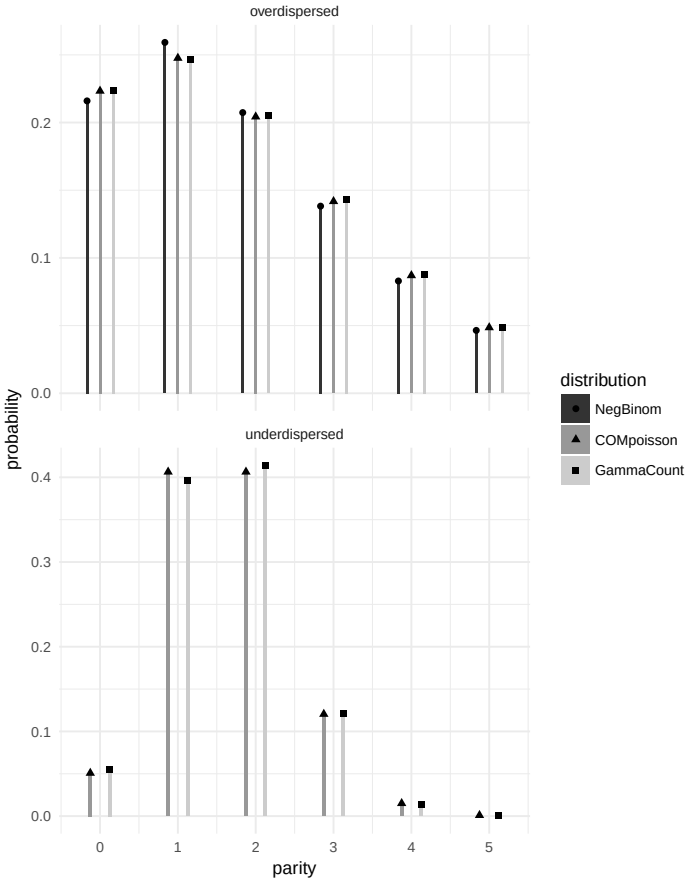
Existing implementations of the COM-Poisson model are available in the `compoisson`, `CompGLM`, and `ComPoissonReg` packages for `R`, for instance, as well as for `SAS` (in the `COUNTREG` procedure) and `MATLAB`, but not – to the best of my knowledge – for `STATA`. The gamma count model does not currently appear to be available off the shelf on any platform, but its definition (Eq. 1) is straightforward to implement in terms of the gamma function. This approach was followed here. The code underlying the present analysis is available as an `R` package alongside this article and was executed by the author under `R version 3.3.2`.

2.4 Gamma count as a Swiss army knife

Fortunately, what initially appears as a potentially confusing proliferation of options for modelling count data ultimately leads to a simplification. Figure 2 shows both the COM-Poisson and gamma count models 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, it shows that the availability of fully generalised count distributions makes the negative binomial model practically redundant for human parity modelling, because it can be well approximated by the COM-Poisson or gamma count models of overdispersion within the relevant range of mean parity and dispersion. Secondly, it shows that their unique ability to model underdispersion sets the latter two 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.

The effective equivalence of these two distributions is striking because 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. Conversely, neither do Shmueli et al. (2005), who established the statistical properties of COM-Poisson, 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.

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)**



Their numerical similarity does not make the COM-Poisson and gamma count distributions entirely redundant, however. Firstly, the conceptual derivations are different, so depending on whether the argument being made focuses on parity progression or on

waiting times, either the COM-Poisson model or gamma count model may be more appropriate. Similarly, 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. While the statistical properties of the COM-Poisson model have been investigated more fully (Sellers and Shmueli 2010) and asymptotic significance tests are available, 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. The gamma count model may therefore be preferable for simulation and 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 for the gamma count distribution as a general-purpose default for modeling 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 parity completed by age 40, is focused on industrialised high-income countries.

Cumulative fertility rates by birth order were extracted from the HFD for all countries for which these were available at the time of writing. The countries included are evident from Figure 2. The range of cohorts included differs by country according to availability, with the earliest being birth cohort 1918 (USA), the 1940s being more typical for other countries, and the latest being birth cohort 1974 for most countries. The strength of this data for present purposes is the fact that it carefully accounts for exposure rates and mortality, and that it provides 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: uniform distribution, linear decline, and 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.

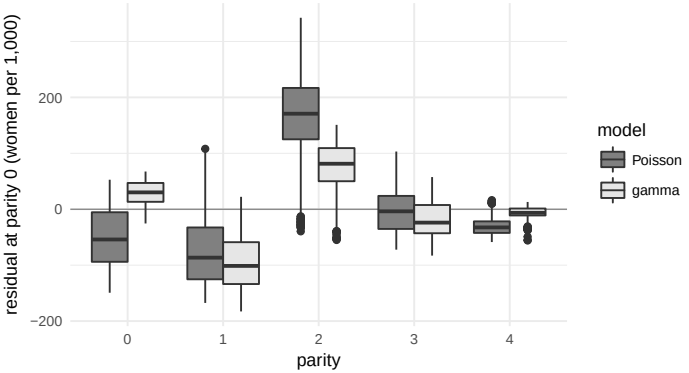
3.2 Zero inflation

Figure 3 displays the average absolute residual by parity. The pattern of these residuals for the Poisson model is dominated by the relatively large share of women with two children. This means that both parity 0 and positive parities contribute to the observed underdispersion.

While the gamma model evidently represents a significant improvement over the Poisson (and equivalent NegBinom) fit in the underdispersed case, the question whether its fit is already good enough in absolute terms needs to be reviewed critically depending on the application. In particular, examples can be found across a wide range of parity patterns where the empirical pattern is not modelled satisfactorily, including some or most cohorts in Canada, Hungary, or Japan, among others (for assessing the fits to individual country-cohort-specific parity distributions, 'rootograms' [Kleiber and Zeileis 2016] can be generated within the `R` package accompanying this article). This suggests that the way these empirical parity distributions differ from the Poisson idealisation is not limited to the presence of underdispersion. Crucially, as can be seen in Figure 3, the residual deviation remaining after allowing for underdispersion is systematic, in that a large residual remains at parity 1 specifically. Nonetheless, as shown in the following, while the simple gamma model by itself cannot match this pattern, it does pave the way for significantly reducing the residual in a way the Poisson model does not. The key lies with parity 0, rather than parity 1, however.

It is common in count data models for zero counts to have a special status vis-à-vis higher counts. The two most common specifications are zero inflation and hurdle models. Zero inflation assumes that cases of parity 0 are contributed by two sources – a structurally-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 appears more relevant than the latter. In terms of theories of the underlying data-generating process, where demographic fertility models stipulate a two-stage process they tend to consider (soft) requirements for being exposed to the risk of childbearing (fecundity, marriage in traditional societies, the decision to have children), but even healthy married women who desire children may of course remain childless by chance. In the following, I therefore limit my attention to the case of zero inflation.

Figure 3: Absolute residuals (observed – fitted) by parity of maximum-likelihood fits of different models to empirical HFD distributions of completed cohort parity, across countries and cohorts, in terms of women per 1,000



Note: Thick horizontal line: median across cohorts; box: inter-quartile range (IQR); whiskers: observations within 1.5 times IQR; points: outliers.

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 the one shown), focus on parity 0 in Figure 3. It is evident that, in general, the regular Poisson model predicts too many zeroes rather than too few. This is unsurprising; we know the vast majority of HFD parity distributions to be underdispersed. By observing that a probability point mass at zero can 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.

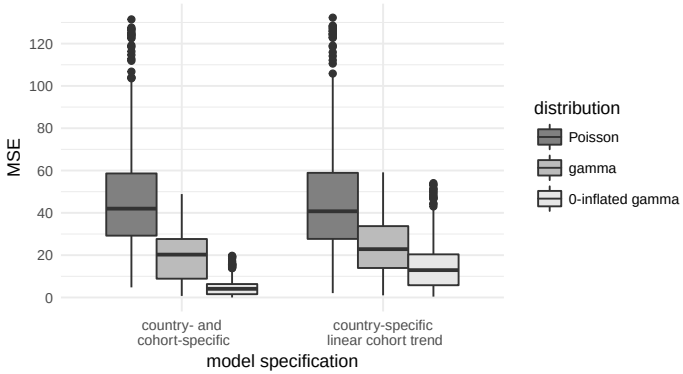
The presence of excess zeroes relative to the underdispersed gamma count baseline is to be welcomed for two reasons. Substantively, we know that the true data-generating process, namely human fertility, does in fact involve a small proportion of women whose probability of giving birth is close to zero. By reflecting this fact, the presence of moderate zero inflation relative to the Gamma count distribution makes this distribution more plausible as a model for birth parity. Moreover, in practical terms, it allows for improving the fit to the data by explicitly taking zero inflation into account—an option not avail-

able to the Poisson model (or indeed, the negative binomial model), which 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 again. 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 (MSE) of maximum-likelihood fits of different models to empirical HFD distributions of completed cohort parity, across countries and cohorts, in terms of women per 1,000



Note: Thick horizontal line: median across cohorts; box: inter-quartile range (IQR); whiskers: observations within 1.5 times IQR; points: outliers.

The left set of box plots 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 11 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: The mean squared error of the zero-inflated gamma count model is less than nine in the vast majority of cases, and typically closer to four, which means its predicted values generally differ from the observed values by only two or three women per 1,000. Moreover, the right set of box plots 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), or six (zero-inflated gamma count) parameters, respectively, to fit all parities 0 through 10 for between 6 (Austria, Finland) and 57 (USA) cohorts at once. While merely illustrative, this specification is close to applications in real-life research in terms of its general structure.

4. Conclusion

Generalised count distributions, specifically the Conway-Maxwell-Poisson and gamma count distributions, and especially their zero-inflated variants, offer a clear advantage over the Poisson and negative-binomial models still dominant in demographic research in the widespread presence of underdispersion in distributions of completed birth parity. The associated disadvantage is largely one of convenience rather than statistical (and would disappear once standard software packages implement these models); it is more than outweighed by the substantial gain in modelling accuracy. A case can therefore be made for using the more general distributions as a matter of course in analyses of parity.

References

Conway, R.W. and Maxwell, W.L. (1962). A queuing model with state dependent service rates. *Journal of Industrial Engineering* 12(2): 132–136.

Human Fertility Database (HFD). Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria). Available at www.humanfertility.org (data downloaded on [2016-10-21]).

Kleiber, C. and Zeileis, A. (2016). Visualizing count data regressions using rootograms. *The American Statistician* 70(3): 296–303. doi:10.1080/00031305.2016.1173590.

Nisén, J., Myrskylaä, M., Silventoinen, K., and Martikainen, P. (2014). Effect of family background on the educational gradient in lifetime fertility of Finnish women born 1940–50. *Population Studies* 68(3): 321–337. doi:10.1080/00324728.2014.913807.

Parrado, E.A. and Morgan, S.P. (2008). Intergenerational fertility among hispanic women: New evidence of immigrant assimilation. *Demography* 45(3): 651–671.

R Core Team (2016). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/

Sellers, K.F. and Shmueli, G. (2010). A flexible regression model for count data. *The Annals of Applied Statistics* 4(2): 943–961. doi:10.1214/09-AOAS306.

Shmueli, G., Minka, T.P., Kadane, J.B., Borle, S., and Boatwright, P. (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): 127–142. doi:10.1111/j.1467-9876.2005.00474.x.

Winkelmann, R. (2008). *Econometric analysis of count data*. 5th edition. Berlin: Springer.