A new concept for count distributions

Per-Erik Hagmark

A new concept for count distributions

Per-Erik Hagmark*

Machine Design and Operation Laboratory, Tampere University of Technology
Korkeakoulunkatu 6, Box 589, FIN-33101 Tampere, Finland

Abstract. A new concept, called silhouette, and the related parameterization are introduced and studied. Applications show how to extend maximally the mean-variance domain of a count distribution, and how to construct a single variable for any mean-variance and any requirements on distribution shape.

Key words. Computation, count data model, distribution shape, matrix, mean-variance domain, parameterization, under- and overdispersion.

MSC. 60E05, 60-08, 15A09, 70C20

1. Introduction

We consider *count variables*, i.e. nonnegative integer-valued random variables, whose mean μ and deviation σ are finite. A $\mu\sigma$-*domain*, i.e. the set of all pairs (μ, σ) of a set of count variables, will be called *maximal* if it contains all pairs (μ, σ) satisfying

$$(\mu - [\mu])([\mu] + 1 - \mu) < \sigma^2 < \mu(N - \mu), \tag{1}$$

where N is the supremum of the variables (possibly $N = \infty$), and $[\mu]$ is the largest integer not exceeding μ. On the other hand, no $\mu\sigma$-domain exceeds the closure of (1).

In commonly used count distributions, the $\mu\sigma$-domain is very seldom maximal, and good general shape flexibility is practically non-existent. In count data modeling this can mean that the $\mu\sigma$-domain of the planned model does not contain the mean-deviation pair estimated from the data, or that no distribution shape offered by the model matches the

* *Correspondence*: Ryssjegränden 2, FIN-02260 Esbo (Finland).
E-mail: Per-Erik.Hagmark@tut.fi. *Tel.*: +358 (0) 98021696

data. The famous distribution of Consul and Jain (1973) is a typical example with seriously incomplete underdispersion ability, i.e. σ stays on a nonzero distance from the left side of (1). The so-called 'generalized Poisson law' again does have maximal μσ-domain, but the shape is always unimodal, and besides, the relations between the parameters and the μσ-pair are laborious (Morlat 1952, Winkelmann 1995). For theory and practice of count models, see e.g. Johnson et al. (1992), Ridout and Besbeas (2004), Castillo & Perez-Casany (2005), and Hagmark (2008).

This study develops a new general approach. We introduce a new concept, the 'silhouette', and a related one-parameter extension for non-binary bounded count variables. Basic theoretical results with examples are presented in Sec.2-5, applications follow in Sec.6-8, and a summary in Sec.9.

2. Basic concepts and formulas

Let $F_0, F_1, F_2, \ldots$ be a non-decreasing sequence with $0 \le F_n \le 1$ and $\lim_{n\to\infty} F_n = 1$. In other words, F_n is the (cumulative) distribution of a count variable (Cv). The related sequence $Y_0 = 0$, $Y_{n+1} = Y_n + F_n$ will be called *integral distribution* (Id). Clearly, the sequence $n - Y_n$ is non-decreasing, and the mean is given by $\mu = \lim_{n\to\infty}(n - Y_n)$. If $\mu < \infty$, it follows that $n - Y_n \le \mu$, and combining with $Y_n \ge 0$ one has the basic inequality

$$Y_n \ge y(\mu)_n, \quad y(\mu)_n = \begin{cases} 0 & \text{if } n \le [\mu] \\ n - \mu & \text{if } n > [\mu] \end{cases}. \tag{2}$$

After further calculation, one obtains the variance formula

$$\sigma^2 = (\mu - [\mu])([\mu] + 1 - \mu) + 2\sum_{n=1}^{\infty}(Y_n - y(\mu)_n). \tag{3}$$

The sequence $y(\mu)_n$ defined in (2) is the Id of the binary Cv with mean μ and values $[\mu]$ & $[\mu]+1$. These Cvs will be called *minimal.* Among all Cvs with the same mean, the minimal Cv has the least variance $(\mu - [\mu])([\mu] + 1 - \mu)$, which follows from (2) and (3), and is the left side of (1). For an example, see Fig.1.

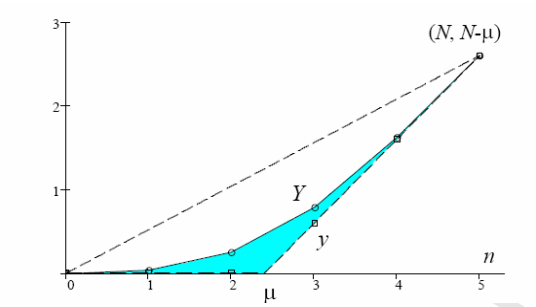

Figure 1. The integral distribution of a binomial Cv and the corresponding minimal Cv. Mean μ = 2.4, max. N = 5, Y = (0, 0.038, 0.2515, 0.789, 1.6255, 2.6), y = (0, 0, 0, 0.6, 1.6, 2.6).

Remark. Two general geometric features can also be pointed out in Fig.1. The variance is twice the area of the shadowed polygon, and the Id is always inside the dash triangle. This follows easily from (2), (3) and the convexity of an Id.

3. Definition of the silhouette

The basic definitions and results in Section 2 lead immediately to a characterization of the class of Cvs to be studied hereafter.

Theorem 1. *The sequence* $Y_0, Y_1, Y_2, \ldots$ *is the Id of a Cv with minimum* 0 (*probability* > 0), *maximum* $N \geq 2$ (*probability* > 0) *and mean* $\mu \in (0, N)$, *if and only if*

(i) $Y_0 = 0$ & $Y_n = n - \mu$, $n \geq N$

(ii) $Y_1 > 0$ & $Y_{N-1} > N - \mu - 1$

(iii) $Y_{n+1} - 2Y_n + Y_{n-1} \geq 0$, $n = 1 \ldots N\text{-}1$.

Moreover, (i)*,* (ii) *and* (iii) *imply* $Y_n > y(\mu)_n$, $n = 1...N\text{-}1$. ■

Definition. The *silhouette* of a Cv with minimum 0, maximum $N \geq 2$ and Id Y is defined as the nonnegative N-1-tuple

$$\alpha_n = \frac{Y_{n+1} - 2Y_n + Y_{n-1}}{Y_n - y(\mu)_n}, \quad n = 1...N\text{-}1, \tag{4}$$

where the mean $\mu = N - Y_N$. For example, the silhouette of the binomial Cv with $N = 5$ and $\mu = 2.4$ is $\alpha \approx (4.615, 1.288, 1.582, 5.417)$, Fig.1.

The basic properties of the silhouette concept will be derived in the next two sections. For this purpose we write (4) as a matrix equation:

$$\boldsymbol{A}(\alpha)\begin{bmatrix} Y_1 \\ Y_2 \\ ... \\ Y_{N-2} \\ Y_{N-1} \end{bmatrix} = \begin{bmatrix} \alpha_1 y(\mu)_1 \\ \alpha_2 y(\mu)_2 \\ ... \\ \alpha_{N-2} y(\mu)_{N-2} \\ \alpha_{N-1} y(\mu)_{N-1} + N - \mu \end{bmatrix}, \tag{5}$$

where

$$\boldsymbol{A}(\alpha) = \begin{bmatrix} 2+\alpha_1 & -1 & 0 & .. & 0 & 0 & 0 \\ -1 & 2+\alpha_2 & -1 & .. & 0 & 0 & 0 \\ .. & .. & .. & .. & .. & .. & .. \\ 0 & 0 & 0 & .. & -1 & 2+\alpha_{N-2} & -1 \\ 0 & 0 & 0 & .. & 0 & -1 & 2+\alpha_{N-1} \end{bmatrix}. \tag{6}$$

4. Construction and parameterization

Every mean-silhouette pair $(\mu, \alpha_1...\alpha_{N\text{-}1})$ is of course an element of the set $(0, N) \times [0, \infty)^{N\text{-}1}$. Next we show that the converse is true. Therefore, let (5) and (6) be a formal starting point with *any* $\mu \in (0, N)$ and *any* nonnegative N-1-tuple α.

The inverse of the 'symmetric Jacobi' matrix $\boldsymbol{A}(\alpha)$ is a 'one-pair' matrix that can be constructed as follows:

$$U_1 = 1 \qquad U_{n+1} = (2+\alpha_n)U_n - U_{n-1} \qquad (U_0 = 0) \tag{7}$$

$$V_{N-1} = 1 \qquad V_{N-n-1} = (2+\alpha_{N-n})V_{N-n} - V_{N-n+1} \qquad (V_N = 0) \tag{8}$$

$$A^{-1}(\alpha)_{n,j} = \frac{1}{U_N}\begin{cases} U_n V_j & if \quad n \le j \\ U_j V_n & if \quad n > j \end{cases}, \quad n = 1...N\text{-}1, j = 1...N\text{-}1. \tag{9}$$

For more details, see Vandebril et al. (2008), Sec. 3.1.3 and 7.2.3. Further, define

$$B(\mu)_n = \begin{cases} [\mu]+1-\mu & if \ n = [\mu] \\ \mu - [\mu] & if \ n = [\mu]+1, \\ 0 & otherwise \end{cases} \quad n = 1...N\text{-}1. \tag{10}$$

Now equation (5) and its solution adopt the form

$$\boldsymbol{A}(\alpha)\big(\boldsymbol{Y} - \boldsymbol{y}(\mu)\big) = \boldsymbol{B}(\mu) \tag{11}$$

$$\boldsymbol{Y} = \boldsymbol{y}(\mu) + \boldsymbol{A}^{-1}(\alpha)\,\boldsymbol{B}(\mu). \tag{12}$$

We complete the components $Y_1 \ldots Y_{N\text{-}1}$ according to Theorem 1i, $Y_0 = 0$, $Y_N = N\text{-}\mu$ …, and we show that this sequence Y is the Id of a unique Cv.

It is easy to see that every element of $\boldsymbol{U}$ (7), $\boldsymbol{V}$ (8) and $\boldsymbol{A}^{-1}$ (9) is positive, and that at least one element of $\boldsymbol{B}$ (10) is positive. Hence, (12) implies $Y_n > y(\mu)_n$, $n = 1...N\text{-}1$. Especially, we get $Y_1 > y(\mu)_1 \ge 0$, and $Y_{N-1} > y(\mu)_{N-1} \ge N - 1 - \mu = Y_N - 1$, i.e. condition (ii) in Theorem 1. Then we have from (5) and (2) $Y_{n+1} - 2Y_n + Y_{n\text{-}1} = \alpha_n\big(Y_n - y(\mu)_n\big) \ge 0$, $n = 1...N\text{-}1$, i.e. condition (iii) in Theorem 1 and (4). Now Theorem 1 yields the next result.

Theorem 2. *Every element* $(\mu, \alpha) \in (0, N) \times [0, \infty)^{N-1}$, $N \geq 2$, *is the mean-silhouette pair of a unique Cv. No other Cvs with minimum* 0 *and maximum N exist.* ■

In other words, (μ, α) is a complete and one-to-one parameterization with *no* other restrictions or interrelations than $0 < \mu < N$ and $\alpha_n \geq 0$ ($n = 1 \ldots N\text{-}1$). The corresponding Cvs will be denoted $Z(\mu, \alpha)$.

5. Relations between the silhouette and the variance

We study the parameterization defined above. The following properties are quite comfortable in theory and applications.

Theorem 3. *The Id components* Y_n ($n = 1 \ldots N\text{-}1$) *and the variance* σ^2 *of* $Z(\mu, \alpha)$ *are functions of the silhouette components* α_j ($j = 1 \ldots N\text{-}1$) *with the following properties:*

(i) *Monotony:* $\dfrac{\partial Y_n}{\partial \alpha_j} < 0$ *and* $\dfrac{\partial \sigma^2}{\partial \alpha_j} < 0$.

(ii) *If* $\alpha_j \to 0$ *for each j, then* $Y_n \to n(1 - \mu/N)$ *and* $\sigma^2 \to \mu(N - \mu)$ *from below.*

(iii) *Assume that* ($\alpha_{[\mu]} \to \infty$ *if* $1 \leq [\mu] \leq N\text{-}1$) *&* ($\alpha_{[\mu]+1} \to \infty$ *if* $0 \leq [\mu] < \mu < N\text{-}1$).

Then $Y_n \to y(\mu)_n$ *and* $\sigma^2 \to (\mu - [\mu])([\mu] + 1 - \mu)$ *from above.* ■

Proof. (i) Writing $\boldsymbol{\Delta} = \boldsymbol{Y} - \boldsymbol{y}$ in (11) and applying partial differentiation one has

$$\boldsymbol{A}\frac{\partial \boldsymbol{\Delta}}{\partial \alpha_j} + \frac{\partial \boldsymbol{A}}{\partial \alpha_j}\boldsymbol{\Delta} = \frac{\partial \boldsymbol{B}}{\partial \alpha_j}$$

where $\dfrac{\partial A_{n,k}}{\partial \alpha_j} = \begin{cases} 1 & \text{if } n = k = j \\ 0 & \text{otherwise} \end{cases}$ and $\dfrac{\partial B_n}{\partial \alpha_j} = 0$. Hence,

$$\boldsymbol{A}\frac{\partial \boldsymbol{\Delta}}{\partial \alpha_j} + \Delta_j \boldsymbol{I}^{\langle j \rangle} = \boldsymbol{0}, \quad \frac{\partial \boldsymbol{\Delta}}{\partial \alpha_j} = -\Delta_j \boldsymbol{A}^{-1\langle j \rangle}, \quad \frac{\partial \Delta_n}{\partial \alpha_j} = -\Delta_j A^{-1}{}_{n,j}$$

where $\boldsymbol{I}$ is the identity matrix and $\boldsymbol{B}^{\langle j \rangle}$ is the j^{th} column of $\boldsymbol{B}$. Since all elements of $\boldsymbol{\Delta}$ and $\boldsymbol{A}^{-1}$ are positive, the first inequality follows, and so the second one follows from (3).

(ii) If $\alpha_n \to 0$ for each $n = 1...N\text{-}1$, it follows from (4) and $Y_n - y(\mu)_n \le (1-\mu/N)\mu$ that the probability $f_n = Y_{n+1} - 2Y_n + Y_{n-1} \to 0$. Hence, the points (n, Y_n) approach the straight line (0, 0)...(N, N-μ) and the limit statements follow (Fig.1 and the remark).

(iii) First, (4) adopts the form $Y_n = y_n + f_n/\alpha_n$ if $\alpha_n > 0$ ($0 \le f_n \le 1$). If $\alpha_{[\mu]} \to \infty$ then $Y_{[\mu]} \to y_{[\mu]}$. Using the convexity of Ids and $Y_0 = y_0$, one has $Y_k \to y_k$ (= 0) for $k \le [\mu]$. In case $[\mu] = \mu$, more is true: The convexity and $Y_N = y_N$ (= N-μ) imply also $Y_k \to y_k$ (= k-μ) for $k \ge [\mu]+1$. One is left with the case $\alpha_{[\mu]+1} \to \infty$ & $[\mu] < \mu$. Now $Y_{[\mu]+1} \to y_{[\mu]+1}$, and a similar argumentation yields $Y_k \to y_k$ for all $k \ge [\mu]+1$. The variance follows from (3).

Hereby the proof is complete. ■

Remark. The extreme Ids in Theorem 3 belong to two binary Cvs with mean μ. In the case (ii) one has the $\{0, N\}$-valued Cv $Z(\mu, 0\ldots0)$, and in the case (iii) one has the $\{[\mu], [\mu]+1\}$-valued minimal Cv (Sec.2). Compare with inequality (1) and Fig.1.

6. Extension for full under- and overdispersion ability

For applications and systematic computation we draw a useful variance formula from (3) and (12), and formulate a condition for pairs $(\mu, \alpha) \in (0, N) \times [0, \infty)^{N-1}$:

$$\mathrm{Var}(Z(\mu, \alpha)) = \big(\mu - [\mu]\big)\big([\mu] + 1 - \mu\big) + 2\sum \boldsymbol{A}^{-1}(\alpha)\,\boldsymbol{B}(\mu) \tag{13}$$

$$(\alpha_{[\mu]} > 0 \textit{ if } 1 \le [\mu] \le N\text{-}1) \;\&\; (\alpha_{[\mu]+1} > 0 \textit{ if } 0 \le [\mu] < \mu < N\text{–}1). \tag{14}$$

For example, every pair $(\mu, \alpha) \in (0, N) \times (0, \infty)^{N-1}$ satisfies (14). The following statement is an immediate consequence of earlier results.

Corollary. *If* μ and α *satisfy condition* (14), *then* $\phi > 0 \mapsto \mathrm{Var}(Z(\mu, \alpha\phi))$, $(\alpha\phi)_n = \phi\alpha_n$, *is a decreasing continuous bijection from* $(0, \infty)$ *to* $((\mu-[\mu])([\mu]+1-\mu), \mu(N-\mu))$. ■

For fixed μ and α, the one-parameter Cv $Z(\mu, \alpha\phi)$, $\phi > 0$, will be called the *extension* of $Z(\mu, \alpha)$. According to the corollary, if μ and α satisfy (14), then the extension has *full under- and overdispersion ability*, i.e. for any theoretically possible σ (1), there is a unique $\phi > 0$ such that $\mathrm{Var}(Z(\mu, \alpha\phi)) = \sigma^2$. The "extra dispersion parameter" ϕ can be computed with (15) to any accuracy. More exactly, denoting $\phi(\varepsilon) = \Phi(\sigma, \mu, \alpha, \varepsilon)$, $\varepsilon > 0$, one has $\phi = \lim_{\varepsilon \to 0} \phi(\varepsilon)$ and $|\mathrm{Var}(Z(\mu, \alpha\phi(\varepsilon))) - \sigma^2| < \varepsilon$.

$$\Phi(\sigma, \mu, \alpha, \varepsilon) = \left| \begin{array}{l} v = \sigma^2 - (\mu - [\mu])([\mu] + 1 - \mu) \\ h = 1 \\ \phi = 0 \\ dv = 2\sum \boldsymbol{A}^{-1}(\alpha\phi)\,\boldsymbol{B}(\mu) - v \\ \textit{while } |dv| \geq \varepsilon \\ \quad h = (-1)^{(dv>0)}|h|/2 \\ \quad \phi = \phi - \ln(1 + h\,e^{\phi}) \\ \quad dv = 2\sum \boldsymbol{A}^{-1}(\alpha\phi)\,\boldsymbol{B}(\mu) - v \\ \phi\,(\textit{output}) \end{array} \right. \tag{15}$$

Example. Suppose maximum $N = 5$, mean $\mu = 2.4$ and variance $\sigma^2 = 0.9$ have been assessed from count data. The binomial model satisfying N and μ is $Z(\mu, \alpha)$ with $\alpha \approx (4.615, 1.288, 1.582, 5.417)$, Sec.2-3, but the variance is $\mu - \mu^2/N = 1.248 > \sigma^2$ (underdispersion). Since σ satisfies (1) and α satisfies (14), there is a ϕ such that $Z(\mu, \alpha\phi)$ has the desired variance σ^2. From (15), $\phi \approx 1.714$, and (12) yields related distributions:

n	0	1	2	3	4	5	σ^2	ϕ
$F_n(\alpha)$	0.038	0.2135	0.5375	0.8365	0.9745	1	1.248	1
$F_n(\alpha\phi)$	0.0178	0.1587	0.5482	0.8864	0.989	1	0.9	1.714

7. Extending for maximal mean-deviation domain

Every parameterized Cv with minimum 0 and maximum $N \geq 2$ is (in principle) of the form $Z(\mu(\theta), \alpha(\theta))$, where θ is the parameter (Theorem 2). If the mean domain is $(0, N)$, i.e. maximal, and all pairs $(\mu(\theta), \alpha(\theta))$ satisfy (14), then it follows from Sec.6 that the extensions $Z(\mu(\theta), \alpha(\theta)\phi)$ form a parameterized Cv with maximal $\mu\sigma$-domain (Sec.1).

Example. Consider all binomial Cvs with the same maximum $N \geq 2$. The point probabilities are strictly positive for all means $\mu \in (0, N)$, so (4) implies that the silhouettes $\alpha(\mu)$ satisfy condition (14). Hence, the two-parameter $Z(\mu, \alpha(\mu)\phi)$ defines a generalized binomial Cv with maximal $\mu\sigma$-domain. Some of the ϕ-level curves $\mathrm{Var}(Z(\mu, \alpha(\mu)\phi)) = \sigma^2$ in the $\mu\sigma$-plane are depicted in Fig.2 ($N = 5$). The curve $\phi = 1$ is the $\mu\sigma$-domain of the binomial Cv (equidispersion), and the curves $\phi = 0.2$ and $\phi = 5$ are examples of overdispersion and underdispersion, respectively. The single point is the case $(\mu, \sigma) = (2.4, \sqrt{0.9})$ from the example in Sec.6.

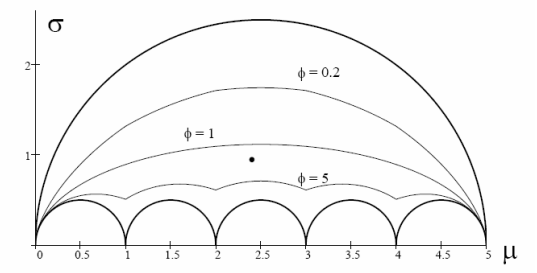

Figure 2. Maximal $\mu\sigma$-domain for $N = 5$ (bounded by half circles). The single point and the ϕ-level curves concern the extended binomial distribution explained in the main text.

8. Design of single count variables (an outline)

Suppose a non-binary Cv is to be designed for minimum 0, maximum $N \geq 2$, mean μ, and deviation σ. The required triplet (N, μ, σ) must be theoretically possible, i.e. satisfy (1). In addition, the final distribution F should also resemble a given distribution shape model G, i.e. the Euclidean norm $||F - G||$ must be minimized, at least nearly.

Consider the parameterized Cv $T(\alpha) = Z(\mu, \alpha \lim_{\varepsilon \to 0} \Phi(\sigma, \mu, \alpha, \varepsilon))$, $\alpha \in (0,\infty)^{N-1}$, with related distribution $F(\alpha)$. According to Sec.6, every $T(\alpha)$ has automatically the correct triplet (N, μ, σ). It can also be shown that, among all Cvs with this triplet, the subset $\{T(\alpha)\}$ is dense in distribution! Thus, a suitable minimization procedure applied to $||F(\alpha) - G||$ produces a satisfactory α. Also, if G is not very contradictory to μ and σ, then a good initial α for the search is given by the silhouette of G. (Compare with the example in Sec.6.)

9. Summary

We have considered integer-valued count variables with minimum 0 and maximum $N \geq 2$. Our presentation is self-contained, and only elementary concepts are needed. Instead, a new concept, the silhouette, and a related complete parameterization, are introduced. Several theoretically and computationally advantageous properties are found, e.g. the monotonous effect of the silhouette on the integral distribution and the variance.

The theory (Sec.1-5) is followed by applications (Sec.6-8). The one-parameter extension of any single count variable satisfying (14) possesses full under- and overdispersion ability. A procedure that enlarges the mean-deviation domain of a parameterized count variable to maximal size is described, and as an example, a new

generalized binomial distribution is defined. Finally, an outline is given for how a single count variable can be constructed for any theoretically appropriate maximum-mean-deviation triplet, and with a distribution shape resembling a desired model.

References

Castillo, J., Perez-Casany, M., 2005. Overdispersed and underdispersed Poisson generalizations. Journal of Statistical Planning and Inference 134 (2), 486-500.

Consul, P.C., Jain, G.C., 1973. A generalization of the Poisson distribution. Technometrics 15, 791-799.

Hagmark, P.-E., 2008. On construction and simulation of count data models. Mathematics and Computers in Simulation 77, 72-80.

Johnson, N.L., Kotz, S., Kemp, A.W., 1992. Univariate Discrete Distributions (2nd ed.). John Wiley & Sons, New York.

Morlat, G., 1952. Sur une generalisation de la loi de Poisson. Comptes Rendus. Academie des Sciences, Paris, Series A 235, 933-935.

Ridout, M.S., Besbeas, P., 2004. An empirical model for underdispersed count data. Statistical Modelling 4, 77-89.

Vandebril, R., van Barel, M., Mastronardi, N., 2008. Matrix Computations and Semiseparable Matrices – Vol. 1. John Hopkins University Press, Baltimore.

Winkelmann, R., 1995. Duration dependence and dispersion in count data models, Journal of Business and Economic Statistics 13, 467-474.