Dispersion Index, Panjer Stopped Sums, and Variance of Mixtures

Malva, Madalena
Instituto Politécnico de Viseu, DM (ESTV), and CEAUL
Campus Politécnico de Repeses
3504–510 Viseu, Portugal
malva@mat.estv.ipv.pt

Mendonça, Sandra
U. Madeira, DME, and CEAUL
Campus Universitário da Penteada
9000–390 Funchal,Portugal
smendonca@uma.pt

Pestana, Dinis D.
U. Lisboa, DEIO (FCUL), and CEAUL
Campo Grande
1749–016 Lisboa, Portugal
dinis.pestana@fc.ul.pt

Sequeira, Fernando
U. Lisboa, DEIO(FCUL), and CEAUL
Campo Grande
1749–016 Lisboa, Portugal
fjsequeira@fc.ul.pt

1 Introduction

Let $S_n = \sum_{k=1}^{n} X_k$, with independent summands $X_k \frown Bernoulli(p_k)$, under the restriction $\mathbb{E}\,[S_n] = \sum_{k=1}^{n} p_k = np$. Then $\text{var}[S_n] = \sum_{k=1}^{n} p_k\,(1-p_k)$ is maximum when all the $p_k = p$. This result, that Nedelman and Wallenius (1986) considered surprising, may be established with very elementary arguments; in fact,

$$\text{var}(S_n) = \sum_{k=1}^{n} p_k - \sum_{k=1}^{n} p_k^2 = np - \sum_{k=1}^{n} p_k^2$$

and rewriting $\sum_{k=1}^{n} p_k^2 = \sum_{k=1}^{n-1} p_k^2 + \left(np - \sum_{k=1}^{n-1} p_k\right)^2$ it is immediate to establish that $\sum_{k=1}^{n} p_k^2$ is minimum when $2\,p_i = 2\left(np - \sum_{k=1}^{n-1} p_k\right) = 2p_n$, $i = 1, 2\ldots, n-1$, i.e. $p_i = p$, $i = 1, \ldots, n$. Thus $\text{var}(S_n)$ is maximum when the summands are i.i.d. $Bernoulli(p)$ random variables.

This elementary proof can immediately be extended to establish that $S_\nu = \sum_{k=1}^{\nu} X_k$, with independent $X_k \frown Binomial(n_k, p_k)$ summands, has maximum variance, under the constraints $\sum_{k=1}^{\nu} n_k = N$ and $\mathbb{E}\,[S_\nu] = \sum_{k=1}^{\nu} n_k p_k = Np$, when all the $p_k = p$, since in $\text{var}(S_\nu) = Np - \sum_{k=1}^{\nu} n_k p_k^2$ we get minimum $\sum_{k=1}^{\nu} n_k p_k^2$ when $p_i = p_\nu$, $i = 1, \ldots, \nu - 1$, as it is easily seen rewriting

$\sum_{k=1}^{\nu} n_k p_k^2 = \sum_{k=1}^{\nu-1} n_k p_k^2 + \frac{\left(Np - \sum_{k=1}^{\nu-1} n_k\, p_k\right)^2}{N - \sum_{k=1}^{\nu-1} n_k}$. Therefore, the $Binomial(N,p)$ random variable has greater variance than any other sum with the same expected value Np of independent binomials.

Morris (1982) investigated the class of members of the natural exponential family for which the variance is at most a quadratic function of the mean; 'he used this property to obtain unified results and gain insight concerning limit laws' (Johnson *et al.*, 2005, p. 113). Morris class contains only six families (Binomial, Poisson, NegativeBinomial, Gamma, Gaussian, and GHS — Generalized Hyperbolic Secant). In section 2 we investigate the variance of sums $S_\nu = \sum_{k=1}^{\nu} X_k$ of independent summands whose variance $\sigma_k^2 = P(\mu_k)$ (with the usual notation $\mu_k = \mathbb{E}(X_k)$), where P is a polynomial of degree at most 2.

On the other hand, all sums of n independent $Poisson(\mu_k)$ random variables with the same expected value $n\mu$ has the same variance, and the sum of n i.i.d. *Geometric*(p) random variables has smaller variance than the sum of n independent *Geometric*(p_k) random variables, under the constraint that it has the same expected value $\frac{n(1-p)}{p}$. The Binomial, Poisson and Negative Binomial (of which the Geometric is a special case) random variables are the non-degenerate members of the Panjer class, for which the recursive relation $p_{n+1} = p_n\left(a + \frac{b}{n+1}\right)$, $n = 0, 1, \ldots$ is valid. Observe that the subfamilies of Panjer class are the discrete Morris families.

Considering the dispersion index $\frac{\text{var}(X)}{\mathbb{E}(X)}$, it is immediate that the Binomial random variables are underdispersed, the Negative Binomials are overdispersed, and that the Poisson is a yardstick, having dispersion index 1. In section 3 we look into the dispersion parameter getting some further results on the extremal properties of the variance of sums of i.i.d. variables, and we put forward some comments on the variance of mixtures.

2 Summands whose variance is quadratic in the mean

Let $S_\nu = \sum_{k=1}^{\nu} X_k$, where the summands X_k are members of a family of random variables with variance σ_k^2, and assume that $\sigma_k^2 = \alpha\mu_k^2 + \beta\mu_k + \gamma$. Morris (1982) demonstrated that there exist exactly six univariate natural exponential families whose variance is a quadratic function of the mean,

Family	α	β	γ
$Gaussian(\mu, \sigma^2)$	0	0	σ^2
$Poisson(\mu)$	0	1	0
$Binomial(n, p)$	$-\frac{1}{n}$	1	0
$NegativeBinomial(\nu, p)$	$\frac{1}{\nu}$	1	0
$Gamma(\nu, \delta)$	$\frac{1}{\nu}$	0	0
$GHS(\alpha)$	$\frac{1}{\alpha}$	0	α

We shall focus on those.

1. With independent summands $X_k \frown Poisson(\mu_k)$, under the constraint $\sum_{k=1}^{\nu} \mu_k = \nu\,\mu$, the variance of the sum $\text{var}(S_\nu) = \sum_{k=1}^{\nu} \mu_k$ is constant.
2. For Binomial, Negative Binomial, Gamma and GHS random variables, using standard lagrangian constrained optimization, it is obvious that the variance of S_ν subject to $\sum_{k=1}^{\nu} \mu_k = \nu\,\mu$, has an extremum if $\mu_k = \mu,\ k = 1, 2, \ldots, \nu$.

(a) If $X_k \frown Binomial(n, p_k)$, that extremum is a maximum (observe that we may consider, more generally, different parameters n_k, as seen in section 1).

(b) If $X_k \frown NegativeBinomial(n, p_k)$, that extremum is a minimum.

(c) If $X_k \frown Gama(\nu, \delta_k)$, that extremum is a minimum.

To enliven the intuition: the parabola $y = x^2$ is below $y = x$ if $x \in (0, 1)$, and above if $x > 1$, and its derivative $2x$ is increasing for all positive x.

3 Variance and the dispersion index

The results in section 2 settles the problem in what concerns the families of Panjer random variables: subject to the restriction of having the same expected value, the variance of the sum of independent (underdispersed) Binomial random variables is maximum if they all have the same parameter p, is always the same if the summands are Poisson, and is minimum if the summands are (overdispersed) Negative Binomials with the same parameter p. The Panjer counting variables play an important role in the theory randomly stopped sums, since from the above recursion it is possible to construct an algorithm to compute iteratively the density of stopped sums with Panjer subordinator, for instance of the aggregate risk (Panjer, 1981; Rólski *et al.*, 1999), and provide an elementary characterization of infinitely divisible and of geo-infinitely divisible random variables as limit of Poisson stopped and of Geometric stopped sums, respectively, as observed by Pestana and Velosa (2004).

Observing that dispersion index of $X \frown Binomial(p)$ is $1 - p < 1$, the dispersion index of $X \frown Poisson(\mu)$ is 1, and the dispersion index of $X \frown NegativeBinomial(\nu,\, p)$ is $\frac{1}{p} > 1$, it seemed worthwhile to investigate further the behaviour of the variance of $S_\nu = \sum_{k=1}^{\nu} X_k$ taking into account the whether the summands are underdispersed or overdispersed; observe that the sum of independent *Hipergeometric*(N, n, p_i) random variables, that are underdispersed, has maximum variance when the summands are i.i.d., exactly as it happens with binomial summands. To carry this out, and assuming that $X > 0$ (since the dispersion index $\frac{\text{var}(X)}{\mathbb{E}(X)}$ is difficult to interpret unless this is so), we write $\frac{\sigma_i^2}{\mu_i} = 1 - g(\mu_i)$ with $g(\mu_i) \in (0, 1)$ for underdispersion and $\frac{\sigma_i^2}{\mu_i} = 1 + g(\mu_i)$ with $g(\mu_i) > 0$ for overdispersion to establish some more general results on the optimality of the variance of sums of i.i.d. summands.

Using standard lagrangian optimization techniques, it is straightforward to establish that the condition $g(\mu_k) + \mu_k g'(\mu_k) = g(\mu_\nu) + \mu_\nu g'(\mu_\nu),\ k = 1, ..., \nu - 1$, which is trivially true whenever $\mu_k = \mu$, is sufficient to get extremal variance. Hence, once again, the case of i.i.d. summands leads to maximum variance if $g(\mu_i) \in (0, 1),\ i = 1, \ldots, \nu$, or to minimum variance, otherwise.

Gurland's (1957) theorem characterizes pairs of random variables for which mixing and compounding has the same result; using Johnson *et al.* (2005) notations: consider two distributions F_1 and F_2 with probability generating functions $\mathcal{G}_1$ and $\mathcal{G}_2$, respectively, where $\mathcal{G}_2$ depends on a parameter θ in such way that $\mathcal{G}_2(z \mid k\theta) = (\mathcal{G}_2(z \mid \theta))^k$; then the mixed distribution $F_2(K\theta) \bigwedge_K F_1$ has probability generating function $\mathcal{G}_1\,(\mathcal{G}_2(z \mid \theta))$, i.e. it corresponds the compound (stopped sum) $F_1 \bigvee F_2$.

The fact that the Panjer counting variables are natural subordinators of randomly stopped sums, and Gurland's (1957) theorem relating stopped sums and mixtures, is a promising path to investigate variances of mixtures.

REFERENCES (RÉFERENCES)

Gurland, J. (1957). Some interrelation among compound and generalized distributions. *Biometrika* **44**, 265–268.

Johnson, N. L., Kotz, S., and Kemp, A. W. (2005). *Univariate Discrete Distributions*, 3rd ed., Wiley, New York.

Morris, C. N. (1982). Natural exponential families with quadratic variance functions, *Annals of Statistics* **11**, 65–80.

Nedelman, J. And Wallenius, T. (1986). Bernoulli trials, Poisson trials, surprising variances and Jensens inequality. *The American Statistician* **40**, 286–289.

Panjer, H. H. (1981). Recursive evaluation of a family of compound distributions, *ASTIN Bulletin* **12**, 22–26.

Pestana, D. D., and Velosa, S. F. (2005). Extensions of Katz-Panjer families of discrete distributions, *REVSTAT* **2**, 145–162.

Rólski, T., Schmidli, H., Schmidt, V., and Teugels, J. (1999). *Stochastic Processes for Insurance and Finance*, Wiley, New York.

RÉSUMÉ (ABSTRACT)

Using Jensen's inequality, Nedelman and Wallenius (1986) established that the sum of n *independent Bernoulli variables with the same parameter* p *has maximum variance among all sums, with the same mean* np*, of independent Bernoulli random variables. We provide a simple proof of this result, and discuss similar results for other families of counting random variables, namely those from Panjer class and Morris class of Natural Exponential Families whose variance is at most quadratic in the mean, restricting our attention to those members which have positive support, so that the dispersion index makes sense.*