Stat Papers (2011) 52:971–977
DOI 10.1007/s00362-009-0304-y

REGULAR ARTICLE

Expectation of a uniform random variable with uniform observation errors after selection of the highest observations

Cornelis A. Van Bochove

Received: 2 March 2009 / Revised: 17 December 2009 / Published online: 14 January 2010

Abstract In selection processes of a random variable with random observation errors, the crucial variable is the conditional expectation of the target variable given the sum of the observations. An example is the selection of the most talented young researchers for tenure track. This paper derives an explicit expression for this conditional expectation for the case that both the target variable and the observation errors have a uniform, but different, distribution.

Keywords Uniform distribution · Selection · Conditional expectation · Observation error · Research career

Mathematics Subject Classification (2000) 62E15

1 Introduction

This paper derives the conditional expectation of a random variable with a uniform distribution, given the sum of this variable and of an arbitrary number of independent variables with a second uniform distribution. The need to obtain this conditional expectation arises in the computation of the expected value of a uniform random variable with independent uniform observation errors, given the sum of a number of observations. The problem was inspired by an analysis of the selection of the most talented young researchers for tenure track. Research talent (the target variable) cannot be observed directly, but has to be inferred from actual research output over a number of years. An individual's output can be considered as the sum of his or her research talent

C. A. Van Bochove (✉)
Centre for Science and Technology Studies (CWTS), Leiden University, Leiden, The Netherlands
e-mail: cbochove@cwts.nl
URL: http://www.socialsciences.leidenuniv.nl/cwts/

(appropriately defined) and an annual random component (the ‘observation error’). If the s percent researchers with the highest observed output are selected for tenure track, the expected value of the target variable within this group increases with the length of the selection period. This increased accuracy has to be balanced against the cost of extending the selection period. This requires the computation of the expected value of the target variable in the upper s percent of the distribution of actual output after a number of years, which is equivalent to obtaining the conditional expectation of the target variable given the sum of the observations over the selection period.

For some purposes[1] a numerical computation of the conditional expectation is insufficient and an analytical expression is needed. Obtaining it requires an analytical expression for the distribution function of the sum of the observations. For most distributions this is not available. However, if both the target variable and the observation errors have uniform distributions, an analytical expression of the distribution of the sum is a special case of a result recently published in *Statistical Papers* (Sadooghi-Alvandi et al. 2009). We will show that in this case it also possible to derive an analytical expression for the conditional expectation.[2]

2 Implicit expression for the conditional expectation

Formally, we consider the case of a random variable V with a uniform distribution on $(0, b)$. Let there be $n-1$ observations of V, each having a random error with another uniform distribution. We will denote the error of observation i by X_{i+1}. This allows us to define the sum S of $n-1$ observations as $S = \sum_{i=1}^{n} X_i$, where $X_1 = (n-1)V$. Then S is the sum of n independent random variables X_i, with X_1 uniform on $(0, a_1)$, $a_1 = (n-1)b$, and $X_i (i \geq 2)$ uniform on $(0, a_2)$; consequently $0 \leq S \leq a_1 + (n-1)a_2$. We are interested in the expectation of X_1 given the value of $S : E\left(X_1|S > s\right) = \int_0^{a_1} x_1 f_{X_1|S}\left(x_1|s\right) \mathrm{d}x_1$, where $f_{X_1|S}$ is the conditional probability density function (pdf) of X_1 given $S > s$. The corresponding conditional distribution function is:

$$\begin{aligned} F_{X_1|S}\left(x_1|S\right) = P\left[X_1 \leq x_1|S > s\right] &= \frac{P\left[X_1 \leq x_1 \wedge S > s\right]}{P\left[S > s\right]} \\ &= \frac{P\left[X_1 \leq x_1\right] - P\left[X_1 \leq x_1 \wedge S < s\right]}{1 - P\left[S < s\right]} \\ &= \frac{F_{X_1}\left(x_1\right) - P\left[X_1 \leq x_1 \wedge S_2 + X_1 < s\right]}{1 - F\left(s\right)}, \end{aligned} \tag{1}$$

where $S_2 = \sum_{i=2}^{n} X_i$ is the sum of the observation errors, $F_{X_1}\left(x_1\right)$ is the marginal distribution of x_1 and $F(s)$ is the distribution function of S. For $x_1 > s$, the probability in the extreme right hand side of (1) is:

[1] In our case this purpose was the development of a simulation model of tenure track policy in a national research system. Such a model requires a specification of the relation between average research productivity, and the rate and speed of selection.

[2] Though this work has been inspired by the specific case of the design of tenure track systems, it is also applicable to the design of other selection processes, provided uniform distributions are at least a reasonable first approximation for the target variable and the observation errors.

$$\begin{aligned} P\left[X_1 \leq x_1 \wedge S_2 + X_1 < s\right] &= P\left[s < X_1 \leq x_1 \wedge S_2 + X_1 < s\right] \\ &\quad + P\left[X_1 < s \wedge S_2 + X_1 < s\right] \\ &= P\left[X_1 < s \wedge S_2 + X_1 < s\right], \end{aligned} \tag{2}$$

where the last equality follows from the fact that the first of the two probabilities in the second expression is zero if $x_1 > s$. Defining $f_2(s_2)$ as the pdf of S_2, the independence of X_1 and S_2 implies:

$$P\left[X_1 < x_1 \wedge S_2 + X_1 < s\right] = \int_0^{\min\{x_1,s\}} \int_0^{s-y_1} f_{X_1}\left(y_1\right) f_2\left(y_2\right) \mathrm{d}y_2 \mathrm{d}y_1 \tag{3}$$

The conditional pdf is now given by:

$$\begin{aligned} f_{X_1|S}\left(x_1|s\right) &= \frac{dF_{X_1|S}\left(x_1|s\right)}{dx_1} = \frac{1}{1 - F\left(s\right)} \\ &\quad \times \left\{ f_{X_1}\left(x_1\right) - \frac{\mathrm{d} \int_0^{\min\{x_1,s\}} \int_0^{s-y_1} f x_1\left(y_1\right) f_2\left(y_2\right) \mathrm{d}y_2 \mathrm{d}y_1}{\mathrm{d}x_1} \right\} \end{aligned} \tag{4}$$

Boundary conditions of the second integral in (4) stem from $f_2(y_2) = 0$ for $y_2 < 0$ and for $y_2 > (n-1)a_2$. For notational compactness it is convenient to define $[x]_+ = \max(x, 0)$; $[x]_+^i = [\max(x, 0)]^i$; the Dirac operator $\delta(z) = 0, 1 \Leftrightarrow z \leq 0, > 0$; and the function $z_x(i) = x - ia_2$. The integral in (3) and (4) can now be written as:

$$\begin{aligned} &\delta\left(z_s\left(n-1\right)\right) \int_0^{\min\{x_1, z_s(n-1)\}} f_{X_1}\left(y_1\right) \mathrm{d}y_1 \\ &\quad + \delta\left(x_1 - z_s\left(n-1\right)\right) \int_{[z_s(n-1)]_+}^{\min\{x_1,s\}} f_{X_1}\left(y_1\right) F_2\left(s - y_1\right) \mathrm{d}y_1 \end{aligned} \tag{5}$$

Evaluation of the integrals as far as possible and substitution of $f_{X_1}\left(x_1\right) = \frac{1}{a_1}$ yields:

$$\begin{aligned} \int_0^{\min\{x_1,s\}} \int_0^{s-y_1} f_{X_1}\left(y_1\right) f_2\left(y_2\right) \mathrm{d}y_2 \mathrm{d}y_1 &= \delta\left(z_s\left(n-1\right)\right) \frac{\min\left\{x_1, z_s\left(n-1\right)\right\}}{a_1} \\ + \delta\left(x_1 - z_s\left(n-1\right)\right) \int_{[z_s(n-1)]_+}^{\min\{x_1,s\}} \frac{F_2\left(s - y_1\right)}{a_1} \mathrm{d}y_1 \end{aligned} \tag{6}$$

Differentiating the right hand side of (6) with respect to x_1 and inserting the result in (4) we find:

$$f_{X_1|S}\left(x_1|s\right) = \delta\left(x_1 - z_s\left(n-1\right)\right) \frac{1 - \delta\left(s - x_1\right) F_2\left(s - x_1\right)}{a_1\left(1 - F\left(s\right)\right)} \tag{7}$$

From which the conditional expectation is seen to be:

$$E\left(X_1|S > s\right) = \frac{a_1^2 - \left[z_s\left(n-1\right)\right]_+^2}{2a_1\left(1 - F\left(s\right)\right)} - \int\limits_{\left[z_s(n-1)\right]_+}^{\min\{s,a_1\}} \left\{\frac{x_1 F_2\left(s - x_1\right)}{a_1\left(1 - F\left(s\right)\right)}\right\} \mathrm{d}x_1 \tag{8}$$

3 Explicit expression for the conditional expectation

Sadooghi-Alvandi et al. (2009) provide a general expression for the pdf of the sum S of n independent uniform variables X_i with $0 < X_i < b_i$:

$$f\left(s\right) = \frac{1}{(n-1)!\Pi_1^n b_i} \left\{s^{n-1} + \sum_{k=1}^{n} (-1)^k \sum_{J_k} \left[s - \sum_{l=1}^{k} b_{j_l}\right]_+^{n-1}\right\},$$

where $J_k = \{(j_1, j_2, \ldots, j_k); 1 \leq j_1 < j_2 < \cdots < j_k \leq n\}$, the j_i's being integer. In our case the X_i are identical for $i \geq 2$, implying $b_{j_l} = a_2$ for $j_l > 1$. There are $\binom{n-1}{k-1}$ k-tuples with $j_1 = 1$ and $\binom{n-1}{k}$ k-tuples with all j_l's > 1. Let δ_{nk} denote the Kronecker delta. The density function of S now is:

$$\begin{aligned} f\left(s\right) = \frac{1}{a_1 a_2^{n-1}\left(n-1\right)!} \sum_{k=0}^{n} (-1)^k &\left\{\binom{n-1}{k-1}\left(1 - \delta_{k0}\right)\left[z_s\left(k-1\right) - a_1\right]_+^{n-1}\right. \\ &\left. + \left(1 - \delta_{nk}\right)\binom{n-1}{k}\left[z_s\left(k\right)\right]_+^{n-1}\right\} \end{aligned} \tag{9}$$

The conditional expectation in (8) contains both the distribution of S, $F(s)$, and the distribution of the sum of the observation errors, $F_2(s_2)$. To obtain the pdf of the sum of the observation errors, $f_2(s_2)$, from (9), we first set $a_1 = a_2$:

$$f\left(s\right) = \frac{1}{a_2^n\left(n-1\right)!} \sum_{k=0}^{n} (-1)^k \binom{n}{k} \left[z_s\left(k\right)\right]_+^{n-1}$$

and next substitute $n - 1$ for n:

$$f_2\left(s_2\right) = \frac{1}{a_2^{n-1}\left(n-2\right)!} \sum_{k=0}^{n-1} (-1)^k \binom{n-1}{k} \left[z_{s_2}\left(k\right)\right]_+^{n-2} \tag{10}$$

To derive $F(s)$ and $F_2(s_2)$ from (9) and (10), respectively, note that

$$\int\limits_{0}^{s} \left[z_y(i)\right]_+^n \mathrm{d}y = \delta(z_s(i)) \int\limits_{-z_0(i)}^{s} \left(z_y(i)\right)^n \mathrm{d}y = \frac{[z_s(i)]_+^{n+1}}{n+1} \tag{11}$$

Consequently:

$$\begin{aligned} F(s) = \int\limits_{0}^{s} f(y)\,\mathrm{d}y = \frac{1}{n!a_1a_2^{n-1}} \sum_{k=0}^{n} (-1)^k \Bigg\{ (1-\delta_{nk}) \binom{n-1}{k} [z_s(k)]_+^n \\ + (1-\delta_{k0}) \binom{n-1}{k-1} [z_s(k-1)-a_1]_+^n \Bigg\} \end{aligned} \tag{12}$$

$$F_2(s_2) = \delta(s_2) \frac{1}{(n-1)!a_2^{n-1}} \sum_{k=0}^{n-1} (-1)^k \binom{n-1}{k} \left[z_{s_2}(k)\right]_+^{n-1} \tag{13}$$

Inserting (12) and (13), the core of the integral in (8) is $(p > q > 0)$:

$$\int\limits_{q}^{p} x_1 F_2(s-x_1)\,\mathrm{d}x_1 = \int\limits_{q}^{p} \frac{x_1}{(n-1)!a_2^{n-1}} \sum_{k=0}^{n-1} (-1)^k \binom{n-1}{k} [z_s(k)-a_1]_+^{n-1} \mathrm{d}x_1 \tag{14}$$

The right hand side of (14) contains integrals of the type $\int_q^p y\,(c-y)_+^n\,\mathrm{d}y$, where integration yields:

$$\int\limits_{q}^{p} y\,(c-y)_+^{n-1}\,\mathrm{d}y = \frac{q\,(c-q)_+^n}{n} + \frac{(c-q)_+^{n+1}}{n\,(n+1)} - \frac{p\,(c-p)_+^n}{n} - \frac{(c-p)_+^{n+1}}{n\,(n+1)},$$

implying:

$$\begin{aligned} \int\limits_{q}^{p} x_1 F_2(s-x_1)\,\mathrm{d}x_1 = \frac{1}{n!a_2^{n-1}} \sum_{k=0}^{n-1} (-1)^k \binom{n-1}{k} \Bigg[q\,[z_s(k)-q]_+^n \\ + \frac{[z_s(k)-q]_+^{n+1}}{n+1} - p\,[z_s(k)-p]_+^n - \frac{[z_s(k)-p]_+^{n+1}}{n+1} \Bigg] \end{aligned} \tag{15}$$

We now obtain the desired explicit expression for the conditional expectation:

$$\begin{aligned}E\left(X_1|S>s\right)=&\frac{a_1^2-\left[z_s\left(n-1\right)\right]_+^2}{2a_1\left(1-F\left(s\right)\right)}-\frac{1}{a_1a_2^{n-1}\left(1-F\left(s\right)\right)n!}\sum_{k=0}^{n-1}(-1)^k\binom{n-1}{k}\\&\times\left\{\left[z_s\left(n-1\right)\right]_+\left[z_s\left(k\right)-\left[z_s\left(n-1\right)\right]_+\right]_+^n\right.\\&+\frac{\left[z_s\left(k\right)-\left[z_s\left(n-1\right)\right]_+\right]_+^{n+1}}{n+1}-a_1\left[z_s\left(k\right)-a_1\right]_+^n\\&\left.-\frac{\left[z_s\left(k\right)-a_1\right]_+^{n+1}}{n+1}\right\}\end{aligned}\qquad(16)$$

4 Illustration: expected error equal to expected target value, best half selected

To illustrate the values of $E(X_1|S>s)$ as n increases, we consider the case $a_1=(n-1)a_2$, implying that in every period the variances (and the expected values) of the

Table 1 Expected value of the target variable in the upper half of the observations, as fraction of the upper boundary of the target variable, variances of observation errors and target variable equal

$n-1$	$\frac{E(X_1\|S>s)}{a_1}$
1	0.667
2	0.708
3	0.722
4	0.729
5	0.733
6	0.736
7	0.738
8	0.740
9	0.741
10	0.742
11	0.742
12	0.743
13	0.744
14	0.744
15	0.744
16	0.745
17	0.745
18	0.745
19	0.746
20	0.746

observation error and the target variable are equal. We specifically consider $s = a_1$, implying that the conditional expectation gives the expected value of x_1 in the upper half of the actual observations. In a selection process this would mean that the best half is selected. In this case, (16) simplifies to:

$$E\left(X_1|S > s\right) = a_1 - \frac{2a_1\left(n-1\right)^{n-1}}{(n+1)!}\sum_{k=0}^{n-1}(-1)^k\binom{n-1}{k}\left(\frac{n-k-1}{n-1}\right)^{n+1}$$

The value of $E(X_1|S > s)$ as a fraction of a_1 is given in Table 1. Note that this fraction converges very rapidly to 0.75, the expected value in the upper half of the distribution if there are no observation errors. This rapid convergence occurs even though we consider the case where the variance of the observation errors is relatively large: equal to that of the target variable.

Though this finding admittedly depends to some degree on the assumption of uniform distributions, it is likely that it has important implications for the case of tenure track selection that inspired this paper. In the case of tenure track, the selection period in many countries has become very long, often 15 years or more. This is in stark contrast to the employment conditions in other sectors of society, where selection periods of a few years before appointment to a permanent post are common. Our result, obtained through elementary but rigorous statistical reasoning, indicates that it is unlikely that the far longer selection periods in the research system lead to a substantially greater accuracy in the selection of the most talented researchers.

Acknowledgements I would like to thank Ludo Waltman of CWTS and an anonymous referee for comments that improved the presentation.

Reference

Sadooghi-Alvandi SM, Nematollahi AR, Habibi R (2009) On the distribution of the sum of independent uniform random variables. Stat Pap 50:171–175