Saralees Nadarajah; Samuel Kotz

Exact and approximate distributions for the product of Dirichlet components

Kybernetika, Vol. 40 (2004), No. 6, [735]--744

Persistent URL: http://dml.cz/dmlcz/135630

KYBERNETIKA — VOLUME 40 (2004), NUMBER 6, PAGES 735–744

EXACT AND APPROXIMATE DISTRIBUTIONS FOR THE PRODUCT OF DIRICHLET COMPONENTS

SARALEES NADARAJAH AND SAMUEL KOTZ

It is well known that $X/(X+Y)$ has the beta distribution when X and Y follow the Dirichlet distribution. Linear combinations of the form $\alpha X + \beta Y$ have also been studied in Provost and Cheong [24]. In this paper, we derive the exact distribution of the product $P = XY$ (involving the Gauss hypergeometric function) and the corresponding moment properties. We also propose an approximation and show evidence to prove its robustness. This approximation will be useful especially to the practitioners of the Dirichlet distribution.

Keywords: approximation, Dirichlet distribution, Gauss hypergeometric function

AMS Subject Classification: 33C90, 62E17, 62E99

1. INTRODUCTION

Since the 1930s, the statistics literature has seen many developments in the theory and applications of linear combinations and ratios of random variables. Some of these include:

— Ratios of normal random variables appear as sampling distributions in single equation models, in simultaneous equations models, as posterior distributions for parameters of regression models and as modeling distributions, especially in economics when demand models involve the indirect utility function (details in [32]).

— Weighted sums of uniform random variables – in addition to the well known application to the generation of random variables – have applications in stochastic processes which in many cases can be modeled by these weighted sums. In computer vision algorithms these weighted sums play a pivotal role ([10]). An earlier application of the linear combinations of uniform random variables is given in connection with the distribution of errors in nth tabular differences Δ^n ([15]).

— Ratio of linear combinations of chi-squared random variables are part of von Neumann's [31] test statistics (mean square successive difference divided by the variance). These ratios appear in various two-stage tests ([30]). They are also used in tests on structural coefficients of a multivariate linear functional relationship model (details in [2, 25]).

— Sums of independent gamma random variables have applications in queuing theory problems such as determination of the total waiting time and in civil engineering problems such as determination of the total excess water flow into a dam. They also appear in test statistics used to determine the confidence limits for the coefficient of variation of fiber diameters ([8, 14]) and in connection with the inference about the mean of the two-parameter gamma distribution ([6]).

— Linear combinations of inverted gamma random variables are used for testing hypotheses and interval estimation based on generalized p-values, specifically for the Behrens–Fisher problem and variance components in balanced mixed linear models ([32]).

— As to the Beta distributions their linear combinations occur in calculations of the power of a number of tests in ANOVA ([18]) among other applications. More generally, the linear combinations are used for detecting changes in the location of the distribution of a sequence of observations in quality control problems ([13]). [20] – [23] and [19] provided applications of sums and ratios to availability, Bayesian quality control and reliability.

— Linear combinations of the form $T = a_1 t_{f_1} + a_2 t_{f_2}$, where t_f denotes the Student t random variable based on f degrees of freedom, represents the Behrens–Fisher statistic and – as early as the middle of the twentieth century – Stein [29] and Chapman [1] developed a two-stage sampling procedure involving the T to test whether the ratio of two normal random variables is equal to a specified constant.

— Weighted sums of the Poisson parameters are used in medical applications for directly standardized mortality rates ([3]).

In this paper, we consider the distribution of $P = XY$ when X and Y are distributed according to the joint pdf

$$f(x, y) = \frac{\Gamma(a+b+c) x^{a-1} y^{b-1} (1-x-y)^{c-1}}{\Gamma(a)\Gamma(b)\Gamma(c)} \tag{1}$$

for $x > 0$, $y > 0$, $x + y < 1$, $a > 0$, $b > 0$ and $c > 0$. This is known as the Dirichlet distribution (see, for example, [12]). It has received applications in many areas, including Bayesian statistics, contingency tables, correspondence analysis, environmental sciences, forensic science, geochemistry, image analysis, life testing, misclassification, molecular biology, neural networks, non-parametric statistics, PERT, and statistical decision theory (see, for example, [7]) for illustrations of some of these application areas).

The paper is organized as follows. In Sections 2 and 3, we derive exact expressions for the pdf and moments of $P = XY$, involving the Gauss hypergeometric function defined by

$${}_2F_1(\alpha, \beta; \gamma; x) = \sum_{k=0}^{\infty} \frac{(\alpha)_k (\beta)_k}{(\gamma)_k} \frac{x^k}{k!}$$

(where $(c)_k = c(c+1)\cdots(c+k-1)$ denotes the ascending factorial), the properties of which can be found in [26] and [5]. In Section 4, we propose an approximation for the distribution of P and show evidence to prove that the it is quite robust. This approximation will be useful especially to the practitioners of the Dirichlet distribution.

2. PDFS

Theorem 1 derives the pdf of $P = XY$ when X and Y are distributed according to (1).

Theorem 1. If X and Y are jointly distributed according to (1) then

$$\begin{aligned} f_P(p) &= \frac{\Gamma(a+b+c)\Gamma(c)}{2^{a-b-c}\Gamma(a)\Gamma(b)\Gamma(2c)} p^{b-1}(1-4p)^{c-1/2}\left(1-\sqrt{1-4p}\right)^{a-b-c} \\ &\quad \times {}_2F_1\left(c, b+c-a; 2c; 2-\frac{1+\sqrt{1-4p}}{2p}\right) \end{aligned} \tag{2}$$

for $0 < p < 1/4$.

Proof. From (1), the joint pdf of $(X, P) = (X, XY)$ becomes

$$\begin{aligned} f(x,p) &= \frac{\Gamma(a+b+c)}{\Gamma(a)\Gamma(b)\Gamma(c)} x^{a-2}\left(\frac{p}{x}\right)^{b-1}\left(1-x-\frac{p}{x}\right)^{c-1} \\ &= \frac{\Gamma(a+b+c)}{\Gamma(a)\Gamma(b)\Gamma(c)} p^{b-1}x^{a-b-c}\left(x-p_1\right)^{c-1}\left(p_2-x\right)^{c-1}, \end{aligned}$$

where $p_1 = (1-\sqrt{1-4p})/2$ and $p_2 = (1+\sqrt{1-4p})/2$. Thus, the pdf of P can be written as

$$f_P(p) = \frac{\Gamma(a+b+c)}{\Gamma(a)\Gamma(b)\Gamma(c)} p^{b-1}\int_{p_1}^{p_2} x^{a-b-c}\left(x-p_1\right)^{c-1}\left(p_2-x\right)^{c-1}\,\mathrm{d}x. \tag{3}$$

By equation (2.2.6.1) in Prudnikov [26, Vol. 1], the integral in (3) can be calculated as

$$\begin{aligned} &\int_{p_1}^{p_2} x^{a-b-c}\left(x-p_1\right)^{c-1}\left(p_2-x\right)^{c-1}\,\mathrm{d}x \\ &= B(c,c)p_1^{a-b-c}\left(p_2-p_1\right)^{2c-1}\ {}_2F_1\left(c, b+c-a; 2c; 1-\frac{p_2}{p_1}\right). \end{aligned} \tag{4}$$

The result in (2) follows by combining (3) and (4). □

The following corollary notes two special cases where (2) reduces to elementary forms.

Corollary 1. If $c = 1$ then (2) reduces to

$$f_P(p) = \frac{2^{b-a+1}\Gamma(a+b+1)}{\Gamma(a)\Gamma(b)} p^{b-1}\sqrt{1-4p}\left(1-\sqrt{1-4p}\right)^{a-b-1}$$

for $0 < p < 1/4$. If $b = a + c$ then (2) reduces to

$$f_P(p) = \frac{4^{a+c}\Gamma\left(a+c+1/2\right)}{\Gamma(a)\Gamma\left(c+1/2\right)} \frac{p^{a+c-1}(1-4p)^{c-1/2}}{\left(1-\sqrt{1-4p}\right)^{2c}}$$

for $0 < p < 1/4$.

Proof. The proof follows by standard properties of the Gauss hypergeometric function, see [26] and [5]. □

3. MOMENTS

Here, we derive the moments of $P = XY$ when X and Y are distributed according to (1).

Theorem 2. If X and Y are jointly distributed according to (1) then

$$E\left(P^n\right) = \frac{\Gamma(a+b+c)\Gamma(a+n)\Gamma(b+n)}{\Gamma(a+b+c+2n)\Gamma(a)\Gamma(b)} \tag{5}$$

for $n \geq 1$. Using properties of the gamma function, (5) can be rewritten as

$$E\left(P^n\right) = \frac{a(a+1)\cdots(a+n-1)b(b+1)\cdots(b+n-1)}{(a+b+c)(a+b+c+1)\cdots(a+b+c+2n-1)}$$

for $n \geq 1$. In particular, the first two moments of P are

$$E(P) = \frac{ab}{(a+b+c)(a+b+c+1)} \tag{6}$$

and

$$E\left(P^2\right) = \frac{a(a+1)b(b+1)}{(a+b+c)(a+b+c+1)(a+b+c+2)(a+b+c+3)}. \tag{7}$$

Proof. Note that $E(P^n) = E(X^nY^n)$ and this is the product moment of the Dirichlet distribution, which is well known (see, for example, [12]). □

4. APPROXIMATION

In view of the fact that $4P$ has support in the interval $[0,1]$, we are motivated to approximate its distribution by a suitable member of the two-parameter beta family of distributions:

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \tag{8}$$

for $0 < x < 1$, $\alpha > 0$ and $\beta > 0$. The choice of the beta parameters (α and β) is made using the method of moments. Equating the first two moments of $4P$ with those of the beta distribution, we have

$$4E(P) = \frac{\alpha}{\alpha+\beta}$$

and

$$16E\left(P^2\right) = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}$$

which we must solve simultaneously to find the beta parameters α and β. After some algebraic manipulation, we find the solutions as

$$\alpha = E(P)\frac{E(P) - 4E\left(P^2\right)}{E\left(P^2\right) - E^2(P)} \tag{9}$$

and

$$\beta = \left\{\frac{1}{4} - E(P)\right\}\frac{E(P) - 4E\left(P^2\right)}{E\left(P^2\right) - E^2(P)}. \tag{10}$$

The two moments $E(P)$ and $E(P^2)$ can be computed using (6) and (7), respectively, for given values of the parameters a, b and c.

Table 1. Estimates of (α,β) for selected (a,b,c).

a	b	c	α	β
0.5	0.5	0.5	0.375	1.031
0.5	0.5	3	0.239	4.543
0.5	3	0.5	0.474	1.105
0.5	3	3	0.497	3.539
3	0.5	0.5	0.474	1.105
3	0.5	3	0.450	3.539
3	3	0.5	2.831	1.003
3	3	3	2.429	3.643
1	3	3	0.978	3.584
1	1	0.5	0.854	1.014
1	3	1	1	1.5
1	1	1	0.778	1.556

Approximations of the above kind have been proposed before; see, for example, [4, 28] and [8]. But this is the first time it has been proposed for correlated beta random variables. In order to show robustness of the approximation, we selected twelve values for the parameters (a, b, c) and computed the corresponding estimates for (α, β) using (9) and (10). The selected parameters (a, b, c) and the estimates are shown in the table above. We checked robustness by comparing the exact and approximated pdfs of $4P$ as given by (2) and (8), respectively. These comparisons are illustrated in Figures 1, 2 and 3. It is evident that the approximation is quite robust. We hope that this approximation will be useful – especially to the practitioners of the Dirichlet distribution – since it avoids the use of the Gauss hypergeometric function and since the beta distribution is widely accessible in standard statistical packages.

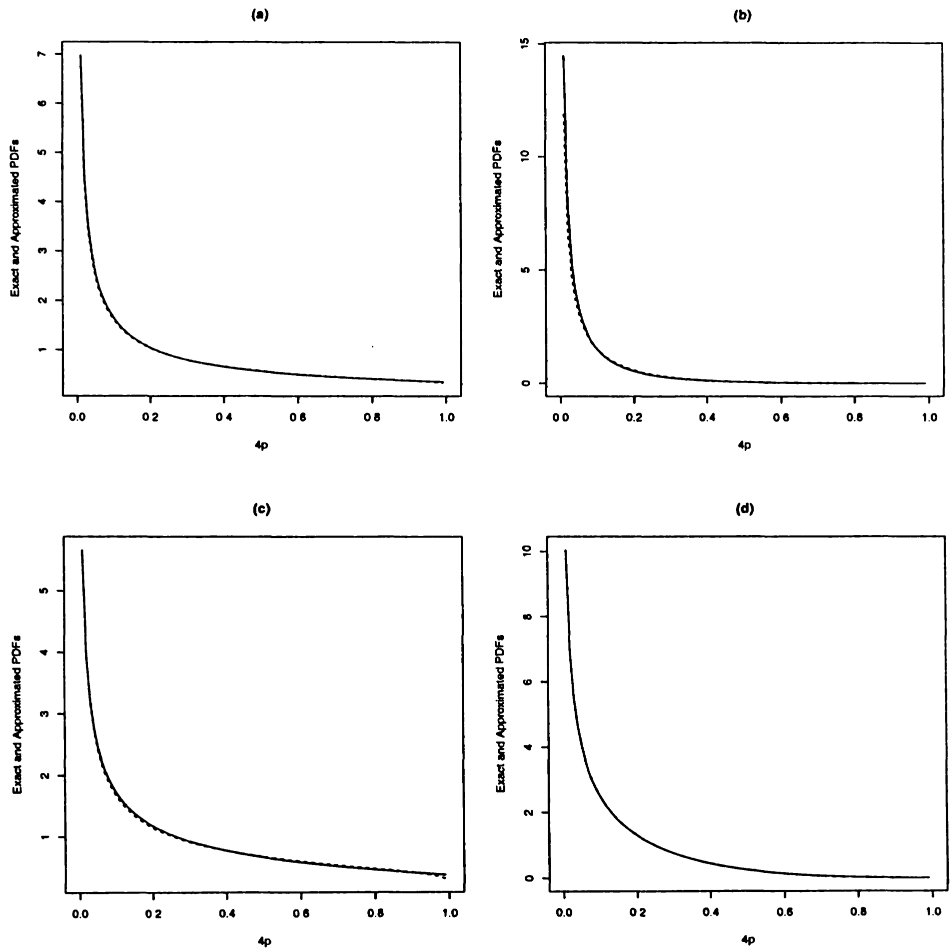

Fig. 1. The exact pdf (solid curve) and the approximated pdf (broken curve) of $P = XY$ for (a): $(a, b, c) = (0.5, 0.5, 0.5)$; (b): $(a, b, c) = (0.5, 0.5, 3)$; (c): $(a, b, c) = (0.5, 3, 0.5)$; and, (d): $(a, b, c) = (0.5, 3, 3)$.

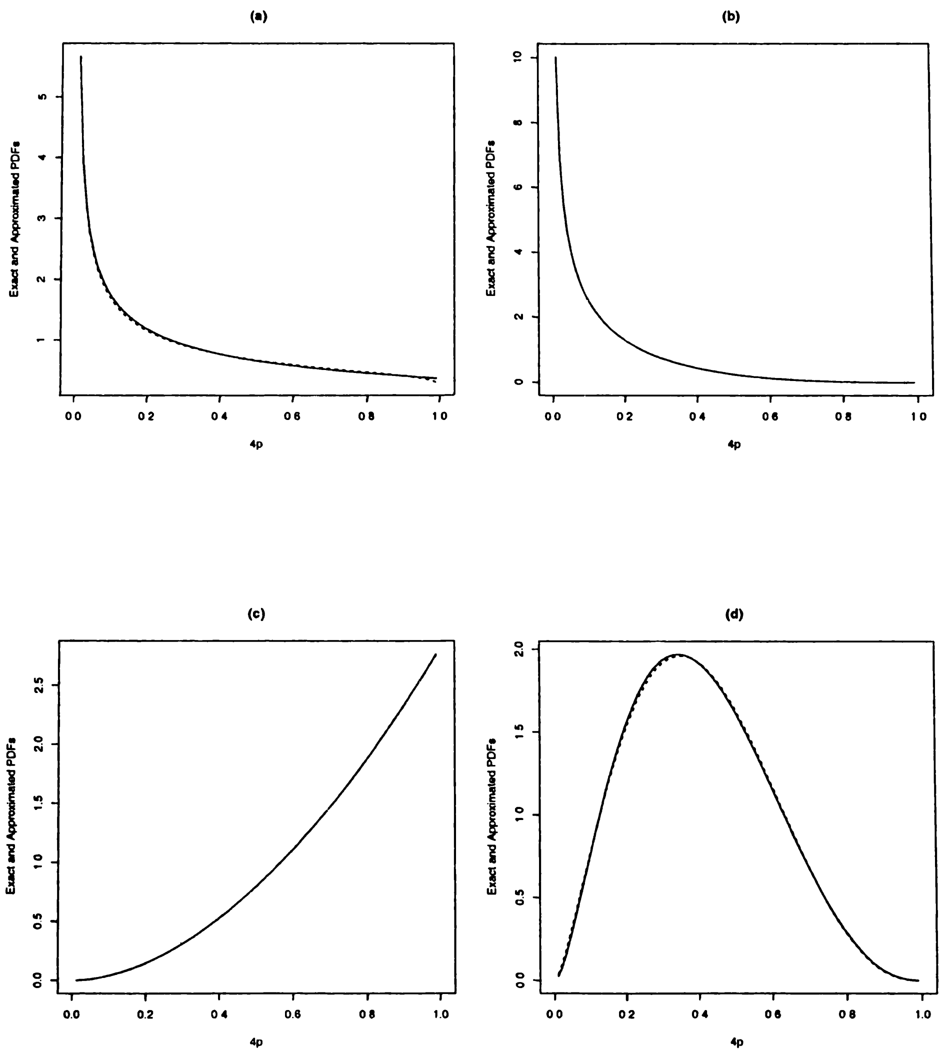

Fig. 2. The exact pdf (solid curve) and the approximated pdf (broken curve) of $P = XY$ for (a): $(a, b, c) = (3, 0.5, 0.5)$; (b): $(a, b, c) = (3, 0.5, 3)$; (c): $(a, b, c) = (3, 3, 0.5)$; and, (d): $(a, b, c) = (3, 3, 3)$.

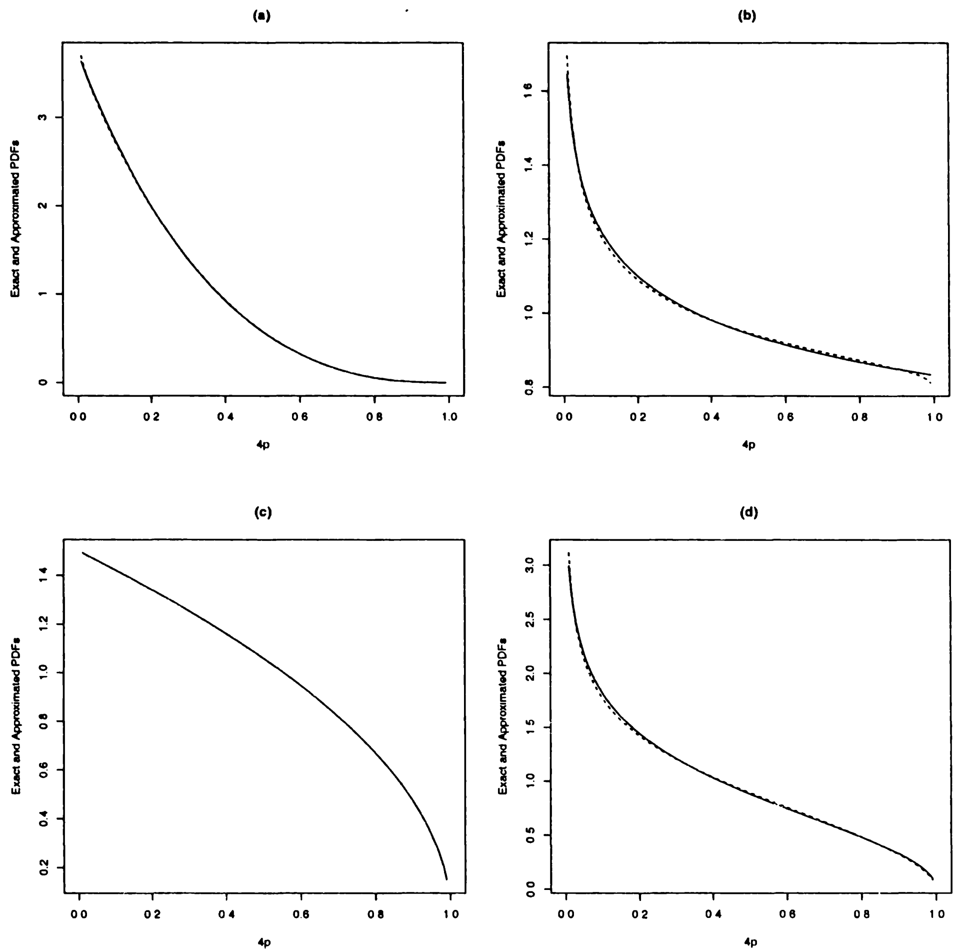

Fig. 3. The exact pdf (solid curve) and the approximated pdf (broken curve) of $P = XY$ for (a): $(a, b, c) = (1, 3, 3)$; (b): $(a, b, c) = (1, 1, 0.5)$; (c): $(a, b, c) = (1, 3, 1)$; and, (d): $(a, b, c) = (1, 1, 1)$.

(Received April 30, 2004.)

REFERENCES

[1] D. G. Chapman: Some two-sample tests. Ann. Math. Statist. *21* (1950), 601–606.

[2] Y. P. Chaubey, A. B. M. Talukder, and Nur Enayet:. Exact moments of a ratio of two positive quadratic forms in normal variables. Comm. Statist. – Theory Methods *12* (1983), 675–679.

[3] A. J. Dobson, K. Kulasmaa, and J. Scherer: Confidence intervals for weighted sums of Poisson parameters. Statist. Medicine *10* (1991), 457–462.

[4] D.-Y. Fan: The distribution of the product of independent beta variables. Comm. Statist. - Theory Methods *20* (1991), 4043–4052.
[5] I. S. Gradshteyn and I. M. Ryzhik: Table of Integrals, Series, and Products. Sixth edition. Academic Press, San Diego 2000.
[6] J. V. Grice and L. J. Bain: Inferences concerning the mean of the gamma distribution. J. Amer. Statist. Assoc. *75* (1980), 929–933.
[7] A. K. Gupta and S. Nadarajah: Handbook of Beta Distribution and Its Applications. Marcel Dekker, New York 2004.
[8] O. A. Y. Jackson: Fitting a gamma or log-normal distribution to fibre-diameter measurements on wool tops. Appl. Statist. *18* (1969), 70–75.
[9] N. Johannesson and N. Giri: On approximations involving the beta distribution. Comm. Statist. - Simulation Computation *24* (1995), 489–503.
[10] B. Kamgar-Parsi, B. Kamgar-Parsi, and M. Brosh: Distribution and moments of weighted sum of uniform random variables with applications in reducing Monte Carlo simulations. J. Statist. Comput. Simulation *52* (1995), 399–414.
[11] C. V. Kimball and D. J. Scheibner: Error bars for sonic slowness measurements. Geophysics *63* (1998), 345–353.
[12] S. Kotz, N. Balakrishnan, and N. L. Johnson: Continuous Multivariate Distributions. Volume 1: Models and Applications. Second edition. Wiley, New York 2000.
[13] T. L. Lai: Control charts based on weighted sums. Ann. Statist. *2* (1974), 134–147.
[14] H. Linhart: Approximate confidence limits for the coefficient of variation of gamma distributions. Biometrics *21* (1965), 733–738.
[15] A. N. Lowan and J. Laderman: On the distribution of errors in nth tabular differences. Ann. Math. Statist. *10* (1939), 360–364.
[16] H. J. Malik: The distribution of the product of two noncentral beta variates. Naval Res. Logist. Quart. *17* (1970), 327–330.
[17] N. N. Mikhail and D. S. Tracy: The exact non-null distribution of Wilk's Λ criterion in the bivariate collinear case. Canad. Math. Bull. *17* (1975), 757–758.
[18] K. L. Monti and P. K. Sen: The locally optimal combination of independent test statistics. J. Amer. Statist. Assoc. *71* (1976), 903–911.
[19] T. Pham-Gia: Distributions of the ratios of independent beta variables and applications. Comm. Statist. - Theory Methods *29* (2000), 2693–2715.
[20] T. Pham-Gia and N. Turkkan: Bayesian analysis of the difference of two proportions. Comm. Statist. - Theory Methods *22* (1993), 1755–1771.
[21] T. Pham-Gia and N. Turkkan: Reliability of a standby system with beta component lifelength. IEEE Trans. Reliability (1994), 71–75.
[22] T. Pham-Gia, and N. Turkkan: Distribution of the linear combination of two general beta variables and applications. Comm. Statist. - Theory Methods *27* (1998), 1851–1869.
[23] T. Pham-Gia and N. Turkkan: The product and quotient of general beta distributions. Statist. Papers *43* (2002) , 537–550.
[24] S. B. Provost and Y.-H. Cheong: On the distribution of linear combinations of the components of a Dirichlet random vector. Canad. J. Statist. *28* (2000), 417–425.
[25] S. B. Provost and E. M. Rudiuk: The exact density function of the ratio of two dependent linear combinations of chi-square variables. Ann. Inst. Statist. Math. *46* (1994), 557–571.
[26] A. P. Prudnikov, Y. A. Brychkov, and O. I. Marichev: Integrals and Series. Volumes 1, 2 and 3. Gordon and Breach Science Publishers, Amsterdam 1986.
[27] B. Rousseau and D. M. Ennis: A Thurstonian model for the dual pair (4IAX) discrimination method. Perception & Psychophysics *63* (2001), 1083–1090.
[28] D. Sculli and K. L. Wong: The maximum and sum of two beta variables in the analysis of PERT networks. Omega Internat. J. Manag. Sci. *13* (1985), 233–240.

[29] C. Stein: A two-sample test for a linear hypothesis whose power is independent of the variance. Ann. Math. Statist. *16* (1945), 243–258.
[30] T. Toyoda and K. Ohtani: Testing equality between sets of coefficients after a preliminary test for equality of disturbance variances in two linear regressions. J. Econometrics *31* (1986), 67–80.
[31] J. von Neumann: Distribution of the ratio of the mean square successive difference to the variance. Ann. Math. Statist. *12* (1941), 367–395.
[32] V. Witkovský: Computing the distribution of a linear combination of inverted gamma variables. Kybernetika *37* (2001), 79–90.
[33] A. J. Yatchew: Multivariate distributions involving ratios of normal variables. Comm. Statist. - Theory Methods *15* (1986), 1905–1926.

Saralees Nadarajah, Department of Mathematics, University of South Florida, Tampa, Florida 33620. U. S. A.
e-mail: snadaraj@math.iupui.edu

Samuel Kotz, Department of Engineering Management and Systems Engineering, The George Washington University, Washington, D.C. 20052. U. S. A.
e-mail: kotz@gwu.edu