APPLICATIONES MATHEMATICAE
34,1 (2007), pp. 15–27

Saralees Nadarajah (Manchester)

A MODEL FOR PROPORTIONS WITH MEDICAL APPLICATIONS

Abstract. Data that are proportions arise most frequently in biomedical research. In this paper, the exact distributions of $R = X + Y$ and $W = X/(X + Y)$ and the corresponding moment properties are derived when X and Y are proportions and arise from the most flexible bivariate beta distribution known to date. The associated estimation procedures are developed. Finally, two medical data sets are used to illustrate possible applications.

1. Introduction. Data comprising of proportions arise most frequently in biomedical research. Enumeration of all of the published work on proportions as related to bio medical research will be an exhaustive exercise. We refer the readers to the excellent books by Aitchison (1986), Diggle *et al.* (1994, 2002) and Fleiss *et al.* (2003).

In this paper, we consider the important problem of sums and ratios of proportions. This problem always arises with respect to data on proportions. For example, suppose that one wishes to test the effectiveness of two treatments, say A and B. The two treatments are tested on a group of patients. Let X denote the success proportion of treatment A and Y the success proportion of treatment B. Clearly, X and Y are random variables. The ratio $W = X/(X + Y)$ will represent the effectiveness of treatment A over treatment B.

In the above example, the joint probability density function (pdf) of X and Y will belong to the class of bivariate beta distributions. The most generalized form of the bivariate beta distribution known to date (due to

2000 *Mathematics Subject Classification*: 33C90, 62E99.

Key words and phrases: bivariate beta distribution, estimation, moments, ratio of random variables, sum of random variables.

Connor and Mosimann (1969)) is given by the joint pdf

$$f(x, y) = K x^{a-1} y^{b-1} (1-x)^{c-b-d} (1-x-y)^{d-1} \tag{1}$$

for $x > 0$, $y > 0$, $x + y < 1$, $a > 0$, $b > 0$, $c > 0$ and $d > 0$, where K is the normalizing constant given by

$$K = \frac{\Gamma(a+c)\Gamma(b+d)}{\Gamma(a)\Gamma(b)\Gamma(c)\Gamma(d)}. \tag{2}$$

The usual bivariate beta distribution arises as the particular case of (1) for $c = b + d$. The univariate marginals of (1) are given by

$$\begin{aligned} f_X(x) &= \frac{x^{a-1}(1-x)^{c-1}}{B(a,c)} \\ f_Y(y) &= K B(a,d) y^{b-1} (1-y)^{a+d-1} \ {}_2F_1(a, b+d-c; a+d; 1-y), \\ E(X^m) &= \frac{B(m+a,c)}{B(a,c)}, \\ E(Y^n) &= \frac{\Gamma(n+b)\Gamma(n+c)\Gamma(a+c)\Gamma(b+d)}{\Gamma(b)\Gamma(c)\Gamma(n+a+c)\Gamma(n+b+d)}, \end{aligned}$$

where the ${}_2F_1$ hypergeometric function (also known as the Gauss hypergeometric function) is defined by

$${}_2F_1(a, b; c; x) = \sum_{k=0}^{\infty} \frac{(a)_k (b)_k}{(c)_k} \frac{x^k}{k!},$$

where $(f)_k = f(f+1) \cdots (f+k-1)$ denotes the ascending factorial. The conditional distributions of (1) are given by the pdfs

$$\begin{aligned} f_{X|Y}(x \,|\, y) &= \frac{x^{a-1}(1-x)^{c-b-d} \left(1 - \dfrac{x}{1-y}\right)^{d-1}}{B(a,d)(1-y)^a \ {}_2F_1(a, b+d-c; a+d; 1-y)}, \\ f_{Y|X}(y \,|\, x) &= K B(a,c) (1-x)^{-b} y^{b-1} \left(1 - \frac{y}{1-x}\right)^{d-1}. \end{aligned}$$

Note that the conditional distribution of $X/(1-y)$ given $Y = y$ belongs to Libby and Novick (1982)'s generalized beta family with the parameters a, d, $1-y$ and $b+d-c$. The conditional distribution of $Y/(1-x)$ given $X = x$ belongs to the standard beta family with the parameters b and d. The conditional moments are:

$$\begin{aligned} E(X^m \,|\, y) &= \frac{(1-y)^m \ {}_2F_1(m+a, b+d-c; m+a+d; 1-y)}{B(a,d) \ {}_2F_1(a, b+d-c; a+d; 1-y)}, \\ E(Y^n \,|\, x) &= \frac{B(b+n,d)}{B(b,d)} (1-x)^n. \end{aligned}$$

In this paper, we derive the exact distributions of $R = X + Y$ and $W = X/(X + Y)$ (Section 2) and the corresponding moment properties (Section 3) when X and Y are correlated beta random variables with the joint pdf given by (1). We also derive the associated estimation procedures (Section 4) and provide applications using two medical data sets (Sections 5 and 6).

The calculations of this paper involve several special functions, including the incomplete beta function defined by

$$B_x(a,b) = \int_0^x t^{a-1}(1-t)^{b-1}\, dt,$$

the Appell function of the first kind defined by

$$F_1(a,b,c;d;x,y) = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} \frac{(a)_{m+n}(b)_m(c)_n x^m y^n}{(d)_{m+n} m!n!},$$

and the ${}_3F_2$ hypergeometric function defined by

$${}_3F_2\,(a,b,c;d,e;x) = \sum_{k=0}^{\infty} \frac{(a)_k(b)_k(c)_k}{(d)_k(e)_k} \frac{x^k}{k!}.$$

The properties of the above special functions can be found in Prudnikov *et al.* (1986) and Gradshteyn and Ryzhik (2000).

2. Probability density functions. Theorems 1 and 2 derive the pdfs of $R = X+Y$ and $W = X/(X+Y)$ when X and Y are distributed according to (1).

Theorem 1. *If X and Y are jointly distributed according to* (1) *then the pdf $f_R(r)$ of R can be expressed in one of the equivalent forms*:

$$KB(a,b)r^{a+b-1}(1-r)^{d-1}\ {}_2F_1\,(a,b+d-c;a+b;r), \tag{3}$$

$$Kr^{b-1}(1-r)^{d-1}\sum_{i=0}^{\infty}\binom{b-1}{i}(-r)^{-i}B_r(i+a,c-b-d+1), \tag{4}$$

$$Kr^{a+b-1}(1-r)^{d-1}\sum_{i=0}^{\infty}\binom{c-b-d}{i}(-r)^{i}B(i+a,b), \tag{5}$$

$$Kr^{a+b-1}(1-r)^{d-1}\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\binom{b-1}{i}\binom{c-b-d}{j}\frac{(-1)^{i+j}r^j}{i+j+a} \tag{6}$$

for $0 < r < 1$.

Proof. From (1), the joint pdf of $(R,W) = (X+Y, X/R)$ becomes

$$f(r,w) = Kr^{a+b-1}(1-r)^{d-1}(1-rw)^{c-b-d}w^{a-1}(1-w)^{b-1}. \tag{7}$$

Thus, the pdf of R can be written as

$$f_R(r) = K r^{a+b-1} (1-r)^{d-1} \int_0^1 (1-rw)^{c-b-d} w^{a-1} (1-w)^{b-1} \, dw. \tag{8}$$

The result in (3) follows by applying equation (2.2.6.1) in Prudnikov *et al.* (1986, Volume 1) to calculate the integral in (8). The result in (4) follows by using the series expansion

$$(1-w)^{b-1} = \sum_{i=0}^{\infty} \binom{b-1}{i} (-w)^i \tag{9}$$

to rewrite (8) as

$$\begin{aligned} f_R(r) &= K r^{a+b-1} (1-r)^{d-1} \int_0^1 (1-rw)^{c-b-d} w^{a-1} \sum_{i=0}^{\infty} \binom{b-1}{i} (-w)^i \, dw \\ &= K r^{a+b-1} (1-r)^{d-1} \sum_{i=0}^{\infty} \binom{b-1}{i} (-1)^i \int_0^1 (1-rw)^{c-b-d} w^{i+a-1} \, dw \\ &= K r^{a+b-1} (1-r)^{d-1} \sum_{i=0}^{\infty} \binom{b-1}{i} (-1)^i r^{-i-a} B_r(i+a, c-b-d+1). \end{aligned}$$

The result in (5) follows by using the series expansion

$$(1-rw)^{c-b-d} = \sum_{i=0}^{\infty} \binom{c-b-d}{i} (-rw)^i \tag{10}$$

to rewrite (8) as

$$\begin{aligned} f_R(r) &= K r^{a+b-1} (1-r)^{d-1} \int_0^1 w^{a-1} (1-w)^{b-1} \sum_{i=0}^{\infty} \binom{c-b-d}{i} (-rw)^i \, dw \\ &= K r^{a+b-1} (1-r)^{d-1} \sum_{i=0}^{\infty} \binom{c-b-d}{i} (-r)^i \int_0^1 w^{i+a-1} (1-w)^{b-1} \, dw \\ &= K r^{a+b-1} (1-r)^{d-1} \sum_{i=0}^{\infty} \binom{c-b-d}{i} (-r)^i B\left(i+a, b\right). \end{aligned}$$

The result in (6) follows by using both the series expansions (9) and (10). ■

THEOREM 2. *If X and Y are jointly distributed according to* (1) *then the pdf $f_W(w)$ of W can be expressed in one of the equivalent forms*:

$$K B(a+b, d) w^{a-1} (1-w)^{b-1} \; {}_2F_1\left(a+b, b+d-c; a+b+d; w\right), \tag{11}$$

$$Kw^{-b-1}(1-w)^{b-1}\sum_{i=0}^{\infty}\binom{d-1}{i}(-w)^{-i}B_w(i+a+b,c-b-d+1), \tag{12}$$

$$Kw^{a-1}(1-w)^{b-1}\sum_{i=0}^{\infty}\binom{c-b-d}{i}(-w)^{i}B(i+a+b,d), \tag{13}$$

$$Kw^{a-1}(1-w)^{b-1}\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\binom{d-1}{i}\binom{c-b-d}{j}\frac{(-1)^{i+j}w^j}{i+j+a+b} \tag{14}$$

for $0<w<1$.

Proof. Using (7), the pdf of W can be written as

$$f_W(w)=Kw^{a-1}(1-w)^{b-1}\int_0^1 r^{a+b-1}(1-r)^{d-1}(1-rw)^{c-b-d}dr. \tag{15}$$

The result in (11) follows by applying equation (2.2.6.1) in Prudnikov *et al.* (1986, Volume 1) to calculate the integral in (15). The result in (12) follows by using the series expansion

$$(1-r)^{d-1}=\sum_{i=0}^{\infty}\binom{d-1}{i}(-r)^i \tag{16}$$

to rewrite (15) as

$$\begin{aligned} f_W(w) &= Kw^{a-1}(1-w)^{b-1}\int_0^1 r^{a+b-1}\left\{\sum_{i=0}^{\infty}\binom{d-1}{i}(-r)^i\right\}(1-rw)^{c-b-d}\,dr \\ &= Kw^{a-1}(1-w)^{b-1}\sum_{i=0}^{\infty}\binom{d-1}{i}(-1)^i\int_0^1 r^{i+a+b-1}(1-rw)^{c-b-d}\,dr \\ &= Kw^{a-1}(1-w)^{b-1} \\ &\quad\times\sum_{i=0}^{\infty}\binom{d-1}{i}(-1)^i w^{-i-a-b}B_w(a+b+i,c-b-d+1). \end{aligned}$$

The result in (13) follows by using the series expansion (10) to rewrite (15) as

$$\begin{aligned} f_W(w) &= Kw^{a-1}(1-w)^{b-1}\int_0^1 r^{a+b-1}(1-r)^{d-1}\left\{\sum_{i=0}^{\infty}\binom{c-b-d}{i}(-rw)^i\right\}dr \\ &= Kw^{a-1}(1-w)^{b-1}\sum_{i=0}^{\infty}\binom{c-b-d}{i}(-w)^i\int_0^1 r^{i+a+b-1}(1-r)^{d-1}dr \\ &= Kw^{a-1}(1-w)^{b-1}\sum_{i=0}^{\infty}\binom{c-b-d}{i}(-w)^iB\left(i+a+b,d\right). \end{aligned}$$

The result in (14) follows by using both the series expansions (16) and (10). ■

Note that if $b-1$, $d-1$ and $c-b-d$ are integers then the infinite sums in (4)–(6) and (12)–(14) reduce to finite sums. Thus, these representations provide a convenient way of computing the pdfs of R and W in the case $b-1$, $d-1$ and $c-b-d$ are integers.

3. Moments. Here, we derive the moments of $R = X + Y$ and $W = X/(X+Y)$ when X and Y are distributed according to (1). We need the following lemma.

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

$$E(X^m Y^n) = \frac{K\Gamma(m+a)\Gamma(n+b)\Gamma(n+c)\Gamma(d)}{\Gamma(m+n+a+c)\Gamma(n+b+d)}$$

for $m \geq 1$ and $n \geq 1$. In particular,

$$\operatorname{Cov}(X,Y) = -\frac{abc}{(a+c)^2(a+c+1)(b+d)}. \tag{17}$$

Proof. One can write

$$\begin{aligned} E(X^m Y^n) &= K \int_0^1 \int_0^{1-x} x^{m+a-1} y^{n+b-1} (1-x)^{c-b-d} (1-x-y)^{d-1}\, dy\, dx \\ &= K \int_0^1 \int_0^{1-x} x^{(m+a)-1} y^{(n+b)-1} (1-x)^{(n+c)-(n+b)-d} (1-x-y)^{d-1}\, dy\, dx. \end{aligned}$$

Thus, the result follows from (2). ∎

The moments of $R = X + Y$ are now simple consequences of this lemma as illustrated in Theorem 3. The moments of $W = X/(X+Y)$ require a separate treatment as shown by Theorem 4.

Theorem 3. *If X and Y are jointly distributed according to* (1) *then*

$$E(R^n) = K \sum_{l=0}^{n} \binom{n}{l} \frac{K\Gamma(n-l+a)\Gamma(l+b)\Gamma(l+c)\Gamma(d)}{\Gamma(n+a+c)\Gamma(l+b+d)}$$

for $n \geq 1$.

Proof. The result follows by writing

$$E((X+Y)^n) = \sum_{l=0}^{n} \binom{n}{l} E(X^{n-l} Y^l)$$

and applying Lemma 1 to each expectation in the sum. ∎

Theorem 4. *If X and Y are jointly distributed according to* (1) *then*

$$\begin{aligned} E(W^n) &= KB(a+b,d)B(a+n,b) \\ &\quad \times {}_3F_2\,(a+b, b+d-c, a+n; a+b+d, a+b+n; 1) \end{aligned} \tag{18}$$

for $n \geq 1$.

Proof. Using (11), one can write

$$E(W^n) = KB(a+b,d)\int_0^1 w^{n+a-1}(1-w)^{b-1} \times {}_2F_1\,(a+b, b+d-c; a+b+d; w)\,dw. \tag{19}$$

The result of the theorem follows by applying equation (2.21.1.5) in Prudnikov *et al.* (1986, Volume 3) to calculate the integral in (19). ∎

4. Estimation. Here, we derive procedures for the maximum likelihood estimation of the parameters of (1). If $\{(x_i, y_i),\ i = 1, \ldots, n\}$ is a random sample from (1) then the log-likelihood function can be written as

$$\begin{aligned}\log L(a,b,c,d)& \\ = (a-1)\sum_{i=1}^n \log x_i + (b-1)\sum_{i=1}^n \log y_i + (c-b-d)\sum_{i=1}^n \log(1-x_i)& \\ + (d-1)\sum_{i=1}^n \log(1-x_i-y_i) + n\log\Gamma(a+c) + n\log\Gamma(b+d)& \\ - n\log\Gamma(a) - n\log\Gamma(b) - n\log\Gamma(c) - n\log\Gamma(d).&\end{aligned}$$

The first-order derivatives of $\log L$ with respect to a, b, c and d are:

$$\begin{aligned}\frac{\partial \log L}{\partial a} &= \sum_{i=1}^n \log x_i + n\Psi(a+c) - n\Psi(a), \\ \frac{\partial \log L}{\partial b} &= \sum_{i=1}^n \log y_i - \sum_{i=1}^n \log(1-x_i) + n\Psi(b+d) - n\Psi(b), \\ \frac{\partial \log L}{\partial c} &= \sum_{i=1}^n \log(1-x_i) + n\Psi(a+c) - n\Psi(c), \\ \frac{\partial \log L}{\partial d} &= \sum_{i=1}^n \log(1-x_i-y_i) - \sum_{i=1}^n \log(1-x_i) + n\Psi(b+d) - n\Psi(d),\end{aligned}$$

where $\Psi(x) = d\log\Gamma(x)/dx$ denotes the digamma function. Thus, the maximum likelihood estimators of (a,b,c,d) are the simultaneous solutions of the equations

$$\sum_{i=1}^n \log x_i = -n\Psi(a+c) + n\Psi(a), \tag{20}$$

$$\sum_{i=1}^n \log y_i - \sum_{i=1}^n \log(1-x_i) = -n\Psi(b+d) + n\Psi(b), \tag{21}$$

$$\sum_{i=1}^{n} \log{(1-x_i)} = -n\Psi(a+c) + n\Psi(c), \tag{22}$$

$$\sum_{i=1}^{n} \log{(1-x_i-y_i)} - \sum_{i=1}^{n} \log(1-x_i) = -n\Psi(b+d) + n\Psi(d). \tag{23}$$

The associated Fisher information matrix requires the second-order derivatives of $\log L$ which can be calculated as:

$$\begin{aligned}
\frac{\partial^2 \log L}{\partial a^2} &= n\Psi'(a+c) - n\Psi'(a),\\
\frac{\partial^2 \log L}{\partial a \partial c} &= n\Psi'(a+c),\\
\frac{\partial^2 \log L}{\partial b^2} &= n\Psi'(b+d) - n\Psi'(b),\\
\frac{\partial^2 \log L}{\partial b \partial d} &= n\Psi'(b+d),\\
\frac{\partial^2 \log L}{\partial c^2} &= n\Psi'(a+c) - n\Psi'(c),\\
\frac{\partial^2 \log L}{\partial d^2} &= n\Psi'(b+d) - \Psi'(d).
\end{aligned}$$

The remaining second-order derivatives are zero, which shows that the mles of a and b are independent, so are the mles of a and d, and the mles of b and c. Since the non-zero second-order derivatives above are all constants the corresponding elements of the Fisher information matrix are:

$$E\left(-\frac{\partial^2 \log L}{\partial a^2}\right) = -n\Psi'(a+c) + n\Psi'(a), \tag{24}$$

$$E\left(-\frac{\partial^2 \log L}{\partial a \partial c}\right) = -n\Psi'(a+c), \tag{25}$$

$$E\left(-\frac{\partial^2 \log L}{\partial b^2}\right) = -n\Psi'(b+d) + n\Psi'(b), \tag{26}$$

$$E\left(-\frac{\partial^2 \log L}{\partial b \partial d}\right) = -n\Psi'(b+d), \tag{27}$$

$$E\left(-\frac{\partial^2 \log L}{\partial c^2}\right) = -n\Psi'(a+c) + n\Psi'(c), \tag{28}$$

$$E\left(-\frac{\partial^2 \log L}{\partial d^2}\right) = -n\Psi'(b+d) + \Psi'(d). \tag{29}$$

5. Data set 1. Here, we consider a data set on white-cell compositions of 30 blood cells by two different methods, see Table 1 (data set 11 of

Aitchison (1986)). The methods used are: microscopic inspection and image analysis. The composition variables are: G = granulocytes, L = lymphocytes and M = monocytes.

Table 1. White cell compositions by two different methods

	Microscopic inspection			Image analysis		
Sample	G	L	M	G	L	M
1	0.732	0.256	0.012	0.763	0.223	0.014
2	0.664	0.280	0.056	0.681	0.262	0.057
3	0.725	0.214	0.067	0.748	0.198	0.054
4	0.806	0.175	0.019	0.867	0.116	0.017
5	0.620	0.351	0.029	0.565	0.408	0.026
6	0.856	0.113	0.031	0.885	0.086	0.029
7	0.957	0.030	0.013	0.966	0.023	0.011
8	0.927	0.053	0.020	0.935	0.047	0.018
9	0.903	0.072	0.025	0.921	0.055	0.024
10	0.936	0.055	0.009	0.943	0.048	0.009
11	0.871	0.114	0.015	0.888	0.097	0.015
12	0.445	0.523	0.032	0.569	0.401	0.038
13	0.240	0.736	0.024	0.225	0.747	0.028
14	0.475	0.472	0.053	0.577	0.370	0.054
15	0.318	0.663	0.019	0.272	0.706	0.022
16	0.462	0.516	0.022	0.544	0.432	0.025
17	0.376	0.252	0.372	0.364	0.245	0.391
18	0.440	0.240	0.320	0.496	0.180	0.324
19	0.583	0.142	0.275	0.629	0.099	0.272
20	0.399	0.169	0.432	0.500	0.115	0.384
21	0.804	0.152	0.044	0.805	0.155	0.04
22	0.655	0.263	0.082	0.659	0.247	0.094
23	0.725	0.218	0.057	0.769	0.179	0.052
24	0.650	0.298	0.052	0.665	0.283	0.052
25	0.370	0.166	0.464	0.388	0.159	0.452
26	0.175	0.802	0.023	0.262	0.709	0.028
27	0.328	0.627	0.045	0.395	0.561	0.043
28	0.427	0.511	0.062	0.388	0.542	0.07
29	0.943	0.046	0.011	0.948	0.041	0.011
30	0.860	0.108	0.032	0.886	0.086	0.027

The interest is in knowing whether the two different methods lead to different compositional results, i.e. are the proportions of G and L the same for microscopic inspection and image analysis? An obvious model in this situation would be the bivariate beta distribution given by (1). We fitted (1) to the three bivariate data sets on proportions: data set 1 containing the values (G, L) obtained by microscopic inspection, data set 2 containing the values (G, L) obtained by image analysis, and data set 3 containing the values (G, L) obtained by both microscopic inspection and image analysis. The maximum likelihood estimates $(\widehat{a}, \widehat{b}, \widehat{c}, \widehat{d})$ obtained by solving (20)–(23) are shown in Table 2. The last column of the table gives the negative logarithm of the maximized likelihood (NLLH).

Table 2. Parameter estimates of (1)

Data set	$\widehat{a}$	$\widehat{b}$	$\widehat{c}$	$\widehat{d}$	NLLH
1	2.277	3.180	1.358	1.004	−56.6
2	2.406	2.880	1.276	1.017	−58.8
3	2.332	3.016	1.312	1.008	−115.2

We use the standard likelihood ratio test to test for homogeneity. Evidently, the two methods are not significantly different in terms of the white cell compositions. The variance covariance matrices of $(\widehat{a}, \widehat{b}, \widehat{c}, \widehat{d})$ for the three data sets obtained by inverting the matrices given by (24)–(29) are:

$$\begin{pmatrix} 0.334 & 0.000 & 0.141 & 0.000 \\ 0.000 & 0.727 & 0.000 & 0.144 \\ 0.141 & 0.000 & 0.103 & 0.000 \\ 0.000 & 0.144 & 0.000 & 0.053 \end{pmatrix},$$

$$\begin{pmatrix} 0.380 & 0.000 & 0.139 & 0.000 \\ 0.000 & 0.586 & 0.000 & 0.130 \\ 0.139 & 0.000 & 0.090 & 0.000 \\ 0.000 & 0.130 & 0.000 & 0.054 \end{pmatrix},$$

$$\begin{pmatrix} 0.177 & 0.000 & 0.069 & 0.000 \\ 0.000 & 0.324 & 0.000 & 0.068 \\ 0.069 & 0.000 & 0.048 & 0.000 \\ 0.000 & 0.068 & 0.000 & 0.027 \end{pmatrix}.$$

The corresponding estimates of $\mathrm{Corr}(X, Y)$ obtained using (17) are provided by Table 3.

Table 3. Correlation coefficient

Data set	$\mathrm{Corr}(X, Y)$
1	−0.902
2	−0.894
3	−0.898

Hence, one can conclude that the compositional variables are strongly correlated and that the two methods are homogeneous.

6. Data set 2. Here, we consider a data set on serum protein compositions of blood samples of 30 patients with two disease types, see Table 4 (data set 16 of Aitchison (1986)). The composition variables are: A = albumin, P = pre-albumin and G = globulin. The first 14 patients have disease type A and the remaining have disease type B.

Table 4. Serum protein compositions of blood samples

	Serum protein		
Patient No	A	P	G
A1	0.348	0.197	0.455
A2	0.386	0.239	0.383
A3	0.471	0.240	0.289
A4	0.427	0.245	0.328
A5	0.346	0.230	0.423
A6	0.485	0.231	0.284
A7	0.398	0.217	0.384
A8	0.537	0.219	0.244
A9	0.316	0.213	0.472
A10	0.543	0.251	0.206
A11	0.409	0.228	0.362
A12	0.322	0.236	0.442
A13	0.372	0.229	0.399
A14	0.348	0.197	0.455
B15	0.452	0.234	0.314
B16	0.490	0.189	0.321
B17	0.512	0.219	0.270
B18	0.429	0.180	0.383
B19	0.424	0.177	0.399
B20	0.377	0.175	0.448
B21	0.556	0.223	0.222
B22	0.264	0.206	0.530
B23	0.311	0.179	0.509
B24	0.338	0.194	0.467
B25	0.396	0.285	0.320
B26	0.438	0.244	0.318
B27	0.347	0.224	0.429
B28	0.376	0.181	0.442
B29	0.278	0.156	0.566
B30	0.333	0.190	0.477

The interest is in knowing whether the two different disease types yield different serum protein compositions, i.e. are the proportions of A and P the same for the two different disease types? We model this situation by the bivariate beta distribution given by (1). We fitted (1) to the three bivariate data sets on proportions: data set 1 containing the values (A, P) for the first 14 patients, data set 2 containing the values (A, P) for the last 16 patients, and data set 3 containing the values (A, P) for all the 30 patients. The maximum likelihood estimates $(\widehat{a}, \widehat{b}, \widehat{c}, \widehat{d})$ as well as the NLLH values are shown in Table 5.

Table 5. Parameter estimates of (1)

Data set	$\widehat{a}$	$\widehat{b}$	$\widehat{c}$	$\widehat{d}$	NLLH
1	19.736	22.439	27.786	33.709	−43.3
2	14.778	11.765	22.993	22.451	−43.9
3	16.370	13.442	24.306	22.914	−84.2

Comparison of the NLLH values shows that the disease types are not significantly different in terms of the serum protein compositions. The variance covariance matrices of $(\widehat{a}, \widehat{b}, \widehat{c}, \widehat{d})$ for the three data sets are:

$$\begin{pmatrix} 54.864 & 0.000 & 75.864 & 0.000 \\ 0.000 & 71.025 & 0.000 & 105.124 \\ 75.864 & 0.000 & 109.547 & 0.000 \\ 0.000 & 105.124 & 0.000 & 161.478 \end{pmatrix},$$

$$\begin{pmatrix} 26.775 & 0.000 & 40.763 & 0.000 \\ 0.000 & 16.869 & 0.000 & 31.484 \\ 40.763 & 0.000 & 65.606 & 0.000 \\ 0.000 & 31.484 & 0.000 & 62.693 \end{pmatrix},$$

$$\begin{pmatrix} 17.558 & 0.000 & 25.539 & 0.000 \\ 0.000 & 11.786 & 0.000 & 19.659 \\ 25.539 & 0.000 & 39.098 & 0.000 \\ 0.000 & 19.659 & 0.000 & 34.785 \end{pmatrix}.$$

For the combined data set (data set 3), the correlation coefficient between the serum proteins A and P is −0.508.

References

J. Aitchison (1986), *The Statistical Analysis of Compositional Data*, Chapman and Hall, London.

R. J. Connor and J. E. Mosimann (1969), *Concepts of independence for proportions with a generalization of the Dirichlet distribution*, J. Amer. Statist. Assoc. 64, 194–206.

P. J. Diggle, K.-Y. Liang and S. L. Zeger (1994), *Analysis of Longitudinal Data*, 1st ed., Oxford Univ. Press, Oxford.

P. J. Diggle, P. J. Heagerty, K.-Y. Liang and S. L. Zeger (2002), *Analysis of Longitudinal Data*, 2nd ed., Oxford Univ. Press, Oxford.

J. L. Fleiss, B. Levin and M. C. Paik (2003), *Statistical Methods for Rates and Proportions*, 3rd ed., Wiley, New York.

I. S. Gradshteyn and I. M. Ryzhik (2000), *Tables of Integrals*, *Series*, *and Products*, 6th ed., Academic Press, San Diego.

D. L. Libby and M. R. Novick (1982), *Multivariate generalized beta-distributions with applications to utility assessment* J. Educational Statist. 7, 271–294.

A. P. Prudnikov, Y. A. Brychkov and O. I. Marichev (1986), *Integrals and Series*, Vols. 1, 2 and 3, Gordon and Breach, Amsterdam.

School of Mathematics
University of Manchester
Manchester M60 1QD, UK
E-mail: saralees.nadarajah@manchester.ac.uk

Received on 14.12.2006;
revised version on 14.3.2007 (1850)