PLISKA

STUDIA MATHEMATICA
BULGARICA

ПЛИСКА

БЪЛГАРСКИ
МАТЕМАТИЧЕСКИ
СТУДИИ

Pliska Stud. Math. Bulgar. **16** (2004), 159-170

PLISKA
STUDIA MATHEMATICA
BULGARICA

AN ESTIMATE OF THE PROBABILITY $\Pr(X < Y)$

Saralees Nadarajah, Georgi K. Mitov, Kosto V. Mitov[1]

In the area of stress-strength models there has been a large amount of work as regards estimation of the probability $R = \Pr(X < Y)$ when X and Y are independent random variables belonging to the same univariate family of distributions. In this paper we propose an estimate of this quantity based on a simple property of the uniform distribution. We illustrate the use of the estimate with bootstrap confidence intervals for four commonly known distributions (normal, exponential, gamma and beta).

1. Introduction

Let X and Y be independent random variables with probability density functions (pdfs) f_X, f_Y and cumulative distribution functions (cdfs) F_X, F_Y. With this notation, one can write

$$\begin{aligned} R = \Pr\left(X < Y\right) &= \int_{-\infty}^{\infty}\int_{-\infty}^{x} f_X(u) f_Y(x) du dx = \\ &= \int_{-\infty}^{\infty} F_X(x) f_Y(x) dx = \int_{-\infty}^{\infty} F_X(x) dF_Y(x). \end{aligned} \tag{1}$$

In an earlier paper of Birnboum and MaCarty [1] the authors studied a distribution-free upper confidence bound for this probability, based on independent

[1] The third author is supported by by NFSI-Bulgaria, Grant No. MM-1101/2001.

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

Key words: Beta Distribution, Bootstrap Confidence Intervals, Exponential Distribution, Gamma Distribution, Normal Distribution, Stress-Strength, Uniform Distribution.

samples of X and Y. They submitted an estimate obtained by substituting in (1), the corresponding empirical CDF $\hat{F}_X(x)$ and $\hat{F}_Y(y)$ obtained from the independent samples $x_1 \leq x_2 \leq \ldots \leq x_m$ and $y_1 \leq y_2 \leq \ldots \leq y_n$, i.e.

$$\bar{R} = \int_{-\infty}^{\infty} \hat{F}_X(x) d\hat{F}_Y(x) = \frac{U}{mn}, \tag{2}$$

where

$$U = \text{ number of pairs } (x_i, y_j) \text{ such that } x_i < y_j, \tag{3}$$

is the Wilcoxon-Mann-Whitney statistics.

The above statistics is applicable in the case when the samples consist of non grouped observations, but in case of grouped observations it must be changed in an appropriate way.

In the present note we provide another distribution-free estimate of R which is applicable for grouped and non grouped observations.

It is based on the following simple observation for the uniform distribution.

2. Uniform Distribution

Let $X \in U(a,b)$ and $Y \in U(c,d)$ be independent uniform random variables. The expression for $\Pr(X < Y)$ will depend on the relative positions of a, b, c and d.

So, in the case when $a < b \leq c < d$, i.e. the support of X is to the left of the support of Y, $\Pr(X < Y) = 1$. If $c < d \leq a < b$, it is clear that $\Pr(X < Y) = 1$.

Suppose now that the supports have non empty intersection. We consider one possible case only. Assume that $a < c < b < d$, i.e. $[a,b] \bigcap [c,d] = [c,b] \neq \emptyset$. In this case, it is easily seen that

$$\begin{aligned} \Pr\left(X < Y | X \in [a,c], Y \in [c,b]\right) &= 1, \\ \Pr\left(X < Y | X \in [a,c], Y \in [b,d]\right) &= 1, \\ \Pr\left(X < Y | X \in [c,b], Y \in [b,d]\right) &= 1 \end{aligned}$$

and

$$\Pr\left(X < Y | X \in [c,b], Y \in [c,b]\right) = \frac{1}{2}.$$

The last relation follows from the fact that

$$\begin{aligned}
& \Pr\left(X < Y | X \in [c,b], Y \in [c,b]\right) \\
= & \frac{\Pr\left(X < Y, X \in [c,b], Y \in [c,b]\right)}{\Pr\left(X \in [c,b], Y \in [c,b]\right)} \\
= & \frac{\displaystyle\int_{-\infty}^{\infty} \Pr\left(X < Y, X \in [c,b] | Y = y \in [c,b]\right) d\Pr(Y \le y)}{\Pr(X \in [c,b]) \Pr(Y \in [c,b])} \\
= & \frac{\displaystyle\int_{c}^{b} \Pr\left(X < y, X \in [c,b]\right) d\Pr(Y \le y)}{\Pr(X \in [c,b]) \Pr(Y \in [c,b])} \\
= & \frac{\displaystyle\int_{c}^{b} \Pr(c < X \le y) d\Pr(Y \le y)}{\Pr(X \in [c,b]) \Pr(Y \in [c,b])}
\end{aligned}$$

and since

$$\begin{aligned}
& \Pr(X \in [c,b]) = \frac{b-c}{b-a}, \quad \Pr(Y \in [c,b]) = \frac{b-c}{d-c}, \\
& \Pr(c < X \le y) = \frac{y-c}{b-a}, \quad d\Pr(Y \le y) = \frac{dy}{d-c}.
\end{aligned}$$

Thus, by the total probability formula, one obtains

$$\begin{aligned}
\Pr(X < Y) &= \Pr(X \in [a,c]) \Pr(Y \in [c,b]) + \Pr(X \in [a,c]) \Pr(Y \in [b,d]) \\
&+ \Pr(X \in [c,b]) \Pr(Y \in [b,d]) + \frac{1}{2} \Pr(X \in [c,b]) \Pr(Y \in [c,b]).
\end{aligned}$$

In the other cases of intersection of supports the calculations are similar.

The fact that $\Pr(X < Y | X, Y \in I) = \frac{1}{2}$, where I is a given interval can be generalized easily for densities which are simple functions. Let us note that a histogram we build on the base of a sample is a function of this type.

This property of the uniform distribution is used in the unpublished paper [6] for ordering of stochastic numbers in the stochastic arithmetic.

3. Simple Densities

Here, we extend the above formulas to the case when the pdfs of X and Y are simple functions, i.e.

$$f_X(x) \quad = \quad p_n \ge 0, \text{ for all } x \in (x_{n-1}, x_n]$$

and

$$f_Y(y) \quad = \quad q_n \geq 0, \text{ for all } y \in (x_{n-1}, x_n]$$

for $n \in \mathbb{Z} = \{-\infty, \ldots, -2, -1, 0, 1, 2, \ldots, \infty\}$.

For densities of this kind it is not difficult to check that for any two given intervals $(x_{n-1}, x_n]$ and $(x_{m-1}, x_m]$:

1. If $n < m$, i.e. $x_n \leq x_{m-1}$ then

$$\begin{aligned} & \Pr\left(X < Y, X \in (x_{n-1}, x_n], Y \in (x_{m-1}, x_m]\right) \\ = \; & p_n q_m \left(x_n - x_{n-1}\right)\left(x_m - x_{m-1}\right). \end{aligned}$$

2. If $n = m$, i.e. $(x_{n-1}, x_n] \equiv (x_{m-1}, x_m]$ then

$$\Pr\left(X < Y, X \in (x_{n-1}, x_n], Y \in [x_{n-1}, x_n]\right) \quad = \quad \frac{1}{2} p_n q_n \left(x_n - x_{n-1}\right)^2.$$

3. If $n > m$, i.e. $x_{n-1} \geq x_m$ then

$$\Pr\left(X < Y, X \in (x_{n-1}, x_n], Y \in (x_{m-1}, x_m]\right) \quad = \quad 0.$$

Thus, by the total probability formula, one obtains the exact formula:

$$\begin{aligned} \Pr(X < Y) \quad = \quad & \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{m-1} \Pr\left(X \in (x_{n-1}, x_n]\right) \Pr\left(Y \in (x_{m-1}, x_m]\right) \\ & + \frac{1}{2} \sum_{n=-\infty}^{\infty} \Pr\left(X \in (x_{n-1}, x_n]\right) \Pr\left(Y \in (x_{n-1}, x_n]\right) \\ = \quad & \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{m-1} p_n q_m \left(x_n - x_{n-1}\right)\left(x_m - x_{m-1}\right) \\ & + \frac{1}{2} \sum_{n=-\infty}^{\infty} p_n q_n \left(x_n - x_{n-1}\right)^2. \end{aligned} \tag{4}$$

4. Continuous Densities

Now we turn to the continuous case. Suppose that X and Y are independent continuous random variables with pdfs $f_X(x)$ and $f_Y(y)$. Then for every $m, n \in \mathbb{Z}$

we have

$$\begin{aligned} & \Pr\left(X \in (x_{n-1}, x_n], Y \in (x_{m-1}, x_m]\right) \\ = & \Pr\left(X \in (x_{n-1}, x_n]\right) \Pr\left(Y \in (x_{m-1}, x_m]\right) \\ = & \left(\int_{x_{n-1}}^{x_n} f_X(x) dx\right) \left(\int_{x_{m-1}}^{x_m} f_Y(y) dy\right) \\ = & f_X\left(\xi_n\right) f_Y\left(\eta_m\right) \left(x_n - x_{n-1}\right) \left(x_m - x_{m-1}\right), \end{aligned}$$

where the last step follows by the mean value theorem for definite integrals for $\xi_n \in (x_{n-1}, x_n)$ and $\eta_m \in (x_{m-1}, x_m)$. Note that one can estimate

$$f_X(x) \quad \approx \quad f_X\left(\xi_n\right), \text{ for all } x \in (x_{n-1}, x_n]$$

and

$$f_Y(y) \quad \approx \quad f_Y\left(\eta_n\right), \text{ for all } y \in (x_{n-1}, x_n]$$

for $n \in \mathbb{Z}$. Using these relations and (4), one obtains the estimate

$$\begin{aligned} \Pr(X < Y) \quad \approx \quad & \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{n-1} f_X\left(\xi_m\right) f_Y\left(\eta_n\right) \left(x_n - x_{n-1}\right) \left(x_m - x_{m-1}\right) \\ & + \frac{1}{2} \sum_{n=-\infty}^{\infty} f_X\left(\xi_n\right) f_Y\left(\eta_n\right) \left(x_n - x_{n-1}\right)^2 . \end{aligned} \tag{5}$$

It is not difficult to see that the first summand on the right hand side of (5) tends to $\Pr(X \in (x_{n-1}, x_n]) \Pr(Y \in (x_{m-1}, x_m])$ when the maximum length of intervals $[x_{n-1}, x_n)$ approaches 0. The second summand vanishes to 0 under the same operation. Thus, if we consider a family of sequences $\{x_n^k, n \in \mathbb{Z}\}_{k=1}^{\infty}$ such that

$$d_k \quad = \quad \sup\left(x_n^{(k)} - x_{n-1}^{(k)}\right) \to 0, \quad k \to \infty$$

then the right hand side of (5) would tend to the exact value of the probability R given by (1).

5. An Algorithm

Here, we provide an algorithm for estimating R applicable for both simple and continuous pdfs and for grouped or non grouped data too.

1. Case of non grouped data.

- Suppose one has independent samples of X and Y:

  $$x_1, x_2, \ldots, x_{n_1}$$

  and

  $$y_1, y_2, \ldots, y_{n_2}.$$

  Let $z_{[1]} \leq z_{[2]} \leq \cdots \leq z_{[n]}$ denote the order statistics of the pooled sample, where $n = n_1 + n_2$.

- Let m be an integer $m < \min(n_1, n_2)$. Partition the interval $[z_{[1]}, z_{[n]}]$ into m segments $(t_{i-1}, t_i]$, $i = 1, 2, \ldots, m$ of equal length $t_i - t_{i-1} = (z_{[n]} - z_{[1]})/m$.

- Estimate the probabilities

  $$\Pr\left(X \in (t_{i-1}, t_i]\right)$$

  and

  $$\Pr\left(Y \in (t_{i-1}, t_i]\right)$$

  by

  $$\widehat{p}_i = \frac{\#\{x_1, \ldots, x_{n_1}\} \in (t_{i-1}, t_i]}{n_1}$$

  and

  $$\widehat{q}_i = \frac{\#\{y_1, \ldots, y_{n_2}\} \in (t_{i-1}, t_i]}{n_2},$$

  respectively, for $i = 1, 2, \ldots, m$.

- By (5), the estimate for R is:

  $$\widehat{R} = \sum_{i=2}^{m} \sum_{j=1}^{i-1} \widehat{q}_i \widehat{p}_j + \frac{1}{2} \sum_{i=1}^{m} \widehat{p}_i \widehat{q}_i. \tag{6}$$

2. Case of grouped data.
Suppose that the observations of X are grouped into M intervals

$$[x_1, x_2], (x_2, x_3], \ldots, (x_M, x_{M+1}]$$

and that the observations of Y are grouped into N intervals

$$[y_1, y_2], (y_2, y_3], \dots, (y_N, y_{N+1}].$$

Since there is no any information about the distribution of the observations inside the intervals, i.e. the observations are uniformly distributed in each interval we can construct a joint set of intervals

$$[t_0, t_1], (t_1, t_2], \dots, (t_{m-1}, t_m]$$

for both samples as follows

$$\min\{M, N\} \le m \le \max\{M, N\},$$

$$t_0 = \min\{x_1, y_1\}, t_m = \max\{x_M, y_N\},$$

$$h = (t_m - t_0)/m, \quad t_l = t_0 + lh, \quad l = 1, \dots, m.$$

Now the observations can be redistributed from the initial intervals into the new ones proportionally on their lengths.

This estimate has the appeal of being simple and does not depend on the parameters of the distributions. Moreover, it is not necessary the random variables X and Y belong to the same family of distributions.

It would be of interest to investigate the sampling properties of (6) such as unbiasedness, consistency and the asymptotic distribution. This would be the subject of a follow-up paper.

6. Examples

In this section, we illustrate the use of (6) with bootstrap confidence intervals for the four most commonly known statistics distributions: normal, exponential, gamma and the beta distributions. From each of these distributions, we simulated two samples with $n_1 = 100$ and $n_2 = 100$. We took $m = 40$. Bootstrap confidence intervals (CIs) were constructed by producing 1000 replications of the estimate given by (6). All the calculations were performed by using the R language (Ihaka and Gentleman, 1996).

6.1. Normal Distribution

For normally distributed random variables, we simulated $X \in N(0, 1)$ and $Y \in N(\mu, 1)$ for $\mu = -3.0, -2.9, \dots, 2.9, 3.0$. The exact value of $R = \Pr(X < Y)$ is easily calculated by the simple formula (see Downton (1973)):

$$R = \Phi\left(\frac{\mu}{\sqrt{2}}\right), \tag{7}$$

where $\Phi(\cdot)$ denotes the cdf of the standard normal distribution. The results of the simulations are illustrated graphically in Figure 1.

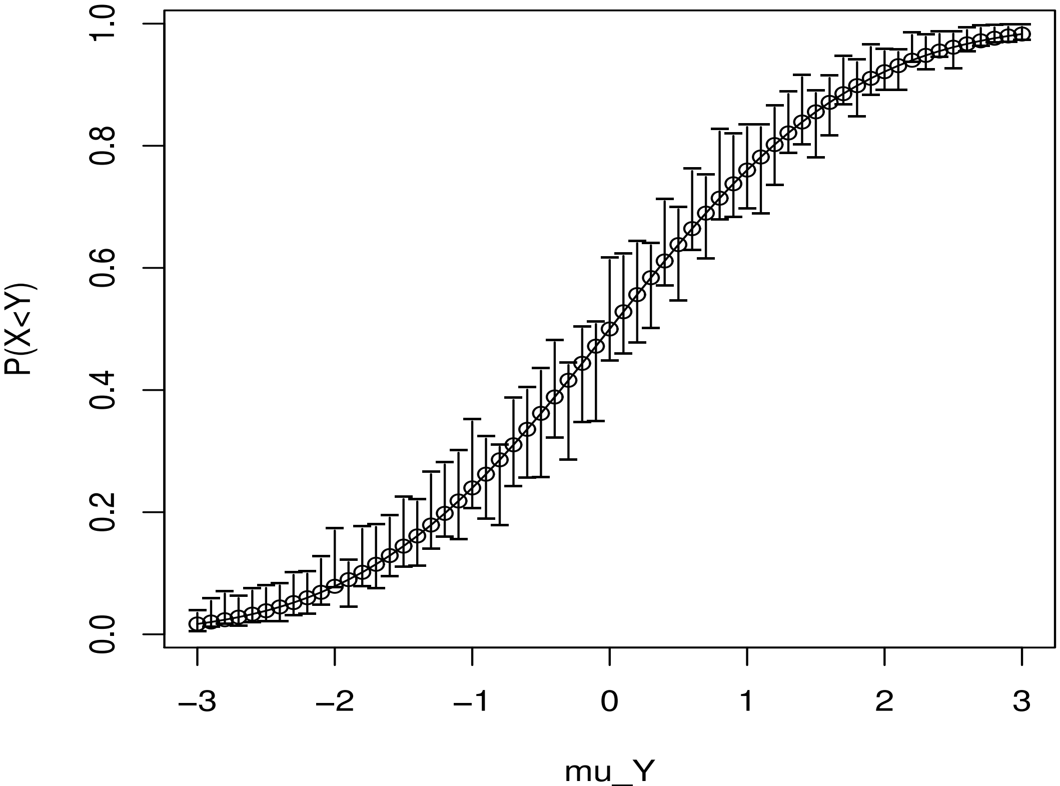

Figure 1: Estimates of $\Pr(X < Y)$ with 95% bootstrap confidence intervals for $X \in N(0,1)$, $Y \in N(\mu,1)$ and $\mu = -3.0, -2.9, \ldots, 2.9, 3.0$. The solid curve gives the exact values of $\Pr(X < Y)$ computed using (7).

6.2. Exponential Distribution

For exponentially distributed random variables, we simulated $X \in \mathit{Exp}(1)$ and $Y \in \mathit{Exp}(\lambda)$ for $\lambda = 1.0, 1.1, \ldots, 5.9, 6.0$. The exact value of $R = \Pr(X < Y)$ is easily calculated by the simple formula:

$$R = \frac{1}{1+\lambda}; \tag{8}$$

see Nadarajah (2003). The results of the simulations are shown graphically in Figure 2.

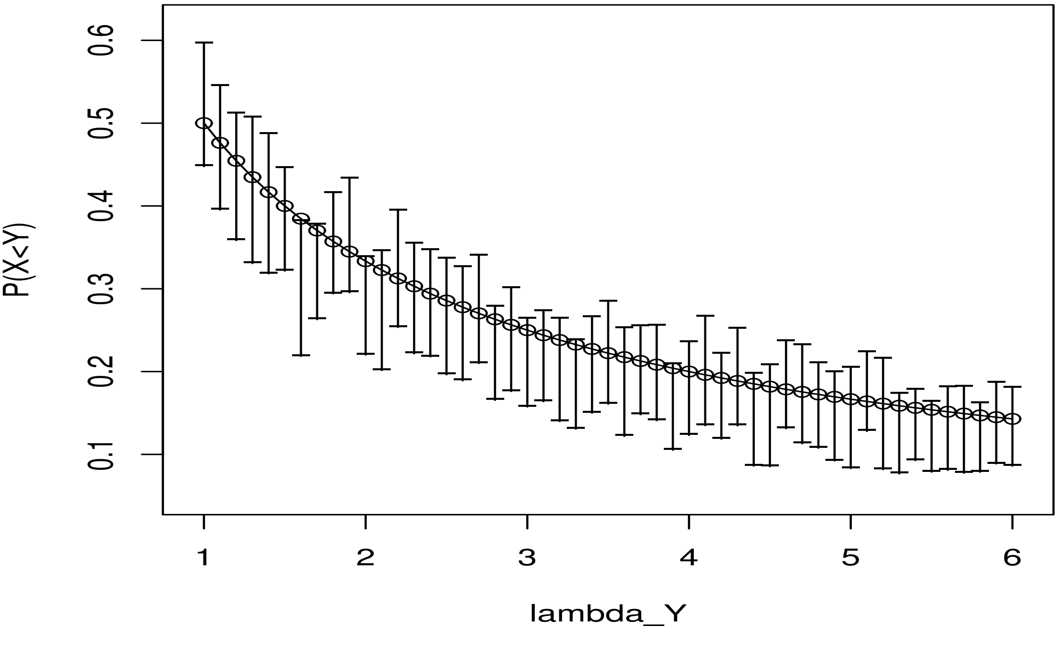

Figure 2: Estimates of $\Pr(X < Y)$ with 95% bootstrap confidence intervals for $X \in \mathit{Exp}(1)$, $Y \in \mathit{Exp}(\lambda)$ and $\lambda = 1.0, 1.1, \ldots, 5.9, 6.0$. The solid curve gives the exact values of $\Pr(X < Y)$ computed using (8).

6.3. Gamma Distribution

For gamma distributed random variables, we simulated $X \in \mathit{Gamma}\ (1, 1)$ and $Y \in \mathit{Gamma}\ (\lambda, 1)$ for $\lambda = 1.0, 1.1, \ldots, 3.9, 4.0$. The exact value of $R = \Pr(X <$

Y) is calculated by the formula (see Nadarajah (2003)):

$$R = 2^{-(1+\lambda)} \, {}_2F_1\left(1, 1+\lambda; 2; \frac{1}{2}\right), \tag{9}$$

where ${}_2F_1$ denotes the Gauss hypergeometric function defined by

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

The results of the simulations are shown graphically in Figure 3.

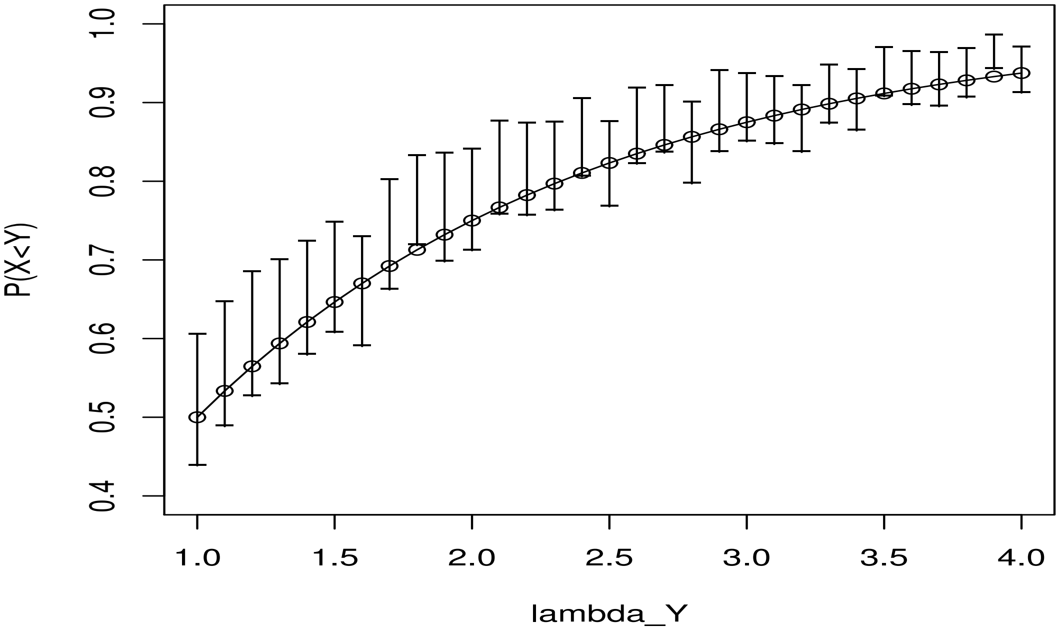

Figure 3: Estimates of $\Pr(X < Y)$ with 95% bootstrap confidence intervals for $X \in Gamma(1, 1)$, $Y \in Gamma(\lambda, 1)$ and $\lambda = 1.0, 1.1, \ldots, 3.9, 4.0$. The solid curve gives the exact values of $\Pr(X < Y)$ computed using (9).

6.4. Beta Distribution

In this example, we simulated $X \in Beta(a_1, b_1)$ and $Y \in Beta(a_2, b_2)$. The exact value of $R = \Pr(X < Y)$ is (see Nadarajah (2002)):

$$(10)\quad R = \frac{B\left(a_1 + a_2, b_2\right)\ {}_3F_2\left(a_1, 1 - b_1, a_1 + a_2; 1 + a_1, a_1 + a_2 + b_2; 1\right)}{a_1 B\left(a_1, b_1\right) B\left(a_2, b_2\right)},$$

where $B(\cdot, \cdot)$ denotes the Beta function defined by

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

and ${}_3F_2$ denotes the hypergeometric function defined by

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

The results of the simulations are given in Table 1.

Table 1: $X \in Beta(a_1, b_1)$, $Y \in Beta(a_2, b_2)$

a_1	b_1	a_2	b_2	Exact Value	95% CI
0.8	0.5	0.8	0.5	0.499	(0.404, 0.563)
0.8	0.5	0.5	0.8	0.300	(0.207, 0.338)
0.9	0.5	0.7	0.4	0.505	(0.457, 0.620)
0.9	0.5	0.4	0.7	0.265	(0.187, 0.319)
0.2	1.5	0.9	2.0	0.827	(0.710, 0.846)
1.5	0.2	0.9	2.0	0.047	(0.025, 0.091)
2.0	0.5	0.5	2.0	0.043	(0.034, 0.095)
0.5	2.0	0.5	2.0	0.563	(0.407, 0.578)
2.0	3.0	2.0	3.0	0.500	(0.451, 0.621)
2.0	3.0	3.0	2.0	0.243	(0.203, 0.344)
2.0	3.0	4.0	5.0	0.576	(0.448, 0.606)
2.0	3.0	5.0	4.0	0.727	(0.612, 0.766)

The exact values were computed using (10).

REFERENCES

[1] Z. W. Birnbaum, R. C. McCarty, A distribution-free upper confidence bound for $\Pr\{Y < X\}$, based on independent samples of X and Y. *The Annals of Mathematical Statistics*, **29** (1958), 558–562.

[2] F. Downton, On the estimation of Pr($Y < X$) in the normal case. *Technometrics*, **15** (1973), 551–558.

[3] R. Ihaka, R. Gentleman, R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5** (1996), 299–314.

[4] S. Nadarajah, Reliability for beta models. *Serdica Mathematical Journal*, **28** (2002), 1001–1016.

[5] S. Nadarajah, Reliability for lifetime distributions. *Mathematical and Computer Modelling*, **37** (2003), 683–688.

[6] R. Alt, S. Markov, G. K. Mitov, On an order relation between distributions, (unpublished)

Saralees Nadarajah
Department of Mathematics
University of South Florida
Tampa, Florida 33620, USA
e-mail: snadaraj@chuma1.cas.usf.edu

Georgi K. Mitov
Faculty of Mathematics and Informatics
Sofia University
5 J. Boucher Str., 1407 Sofia, Bulgaria
e-mail: gkmitov@yahoo.com

Kosto V. Mitov
Faculty of Aviation
National Military University
5856 Dolna Mitropolia, Bulgaria
e-mail: kmitov@af-acad.bg