Exact Calculation of Beta Inequalities

John Cook
Department of Biostatistics, Box 447
The University of Texas, M. D. Anderson Cancer Center
1515 Holcombe Blvd., Houston, Texas 77030, USA
cook@mdanderson.org

November 2, 2005*

Abstract

This paper addresses the problem of evaluating $P(X > Y)$ where X and Y are independent beta random variables. We cast the problem in terms of a hypergeometric function and use hypergeometric identities to calculate the probability in closed form for certain values of the distribution parameters.

Keywords: beta distribution, stochastic inequalities, hypergeometric functions, closed form

1 Introduction

Define $g(a, b, c, d)$ to be the probability of a sample from a beta(a, b) random variable being larger than an independent sample from a beta(c, d) random variable. Thus for positive parameters a, b, c, and d,

$$g(a, b, c, d) = \int_0^1 \frac{x^{a-1}(1-x)^{b-1}}{B(a, b)} I_x(c, d)\, dx \tag{1}$$

where $I_x(c, d)$ is the incomplete beta function, the CDF of a beta(c, d) random variable.

*Revised February 21, 2006

The function g plays a central role in the calculation of randomization probabilities for many outcome-adaptive clinical trials with binary end points. Suppose the prior probabilities of response on each of two arms in a randomized trial have beta priors. Because the beta prior is conjugate for binomial distributions, the posterior distributions on the probability of response are also beta distributions. A simple adaptive randomization scheme is to assign patients to an arm with probability equal to the probability that that arm is superior. If the posterior distribution on one arm is beta(a, b) and the other is beta(c, d), then the probability of assigning the first arm is $g(a, b, c, d)$. More generally, one often assigns the first arm with probability

$$\frac{g(a, b, c, d)^\lambda}{g(a, b, c, d)^\lambda + g(c, d, a, b)^\lambda}$$

for some $\lambda > 0$. Values of $\lambda < 1$ dampen the the effect of the data on the randomization probabilities, and values $\lambda > 1$ amplify the effect of the data. See [3], [4], and [5] for theoretical background. See [11] for software to simulate adaptively randomized trials.

The function g is also also at the heart of the safety monitoring methods of Thall, Simon, and Estey [9]. As above, the probabilities of response on each of two arms (either two active control arms or an active arm and a historical standard) are distributed as beta random variables. Accrual to an arm stops if the probability of that arm being superior falls below some threshold. That is, if $g(a, b, c, d)$ is too small, the first arm is closed. See [8] for software implementing this method.

In general, the function g cannot be evaluated in closed form and must be approximated numerically, as for example in [10]. In [6], we gave general methods for computing g for several distribution families, including beta. Here we focus on special cases that can be evaluated in terms of gamma functions. These special cases may be directly useful in applications. Also, the special cases given here provide test cases for numerical software used to evaluate g for general arguments.

The symmetries

$$g(a, b, c, d) = g(d, c, b, a) = g(d, b, c, a) = 1 - g(c, d, a, b)$$

are developed in [6]. For the 24 permutations of the 4 arguments, there are at most six different values of g. These are $g(a, b, c, d)$, $g(a, b, d, c)$, $g(a, c, d, b)$ and their complementary probabilities.

If we define

$$h(a,b,c,d) = \frac{B(a+c,b+d)}{B(a,b)B(c,d)} \quad (2)$$

$$= \frac{\Gamma(a+c)\Gamma(b+d)\Gamma(a+b)\Gamma(c+d)}{\Gamma(a)\Gamma(b)\Gamma(c)\Gamma(d)\Gamma(a+b+c+d)}. \quad (3)$$

then [6] shows that the following recurrence relations hold:

$$\begin{aligned} g(a+1,b,c,d) &= g(a,b,c,d) + h(a,b,c,d)/a \\ g(a,b+1,c,d) &= g(a,b,c,d) - h(a,b,c,d)/b \\ g(a,b,c+1,d) &= g(a,b,c,d) - h(a,b,c,d)/c \\ g(a,b,c,d+1) &= g(a,b,c,d) + h(a,b,c,d)/d \end{aligned}$$

2 Case: one integer argument

First we consider the special case $d = 1$. We have

$$\begin{aligned} g(a,b,c,1) &= \int_0^1 \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)} \int_0^x ct^{c-1}\,dt\;dx \\ &= \int_0^1 \frac{x^{a+c-1}(1-x)^{b-1}}{B(a,b)} \\ &= \frac{B(a+c,b)}{B(a,b)} \\ &= \frac{\Gamma(a+b)\,\Gamma(a+c)}{\Gamma(a+b+c)\,\Gamma(a)}. \end{aligned}$$

Using the recurrence relationship for the fourth argument, we can reduce the case of d being any positive integer to the case $d = 1$.

Using symmetries of g we can permute any parameter into the third argument, and so we can compute $g(a,b,c,d)$ exactly if any one of the arguments is an integer.

3 Derivation of Hypergeometric Form

In [6] we showed that

$$g(a,b,c,d) = \sum_{k=0}^{\infty} \frac{h(a,b,c+k,d)}{c+k}.$$

If we define

$$t_k = \frac{h(a,b,c+k,d)}{c+k}$$

then a little algebra shows that

$$\frac{t_{k+1}}{t_k} = \frac{(a+c+k)(c+d+k)}{(c+k+1)(a+b+c+d+k)}.$$

Since t_{k+1}/t_k is a rational function of k, it follows that the infinite sum for g is a constant times a hypergeometric function. (See [2] or [7] for background on hypergeometric functions.) Furthermore, one can read the hypergeometric parameters from the factorization of t_{k+1}/t_k. It follows that

$$g(a,b,c,d) = \frac{h(a,b,c,d)}{c}\, {}_3F_2\left(\begin{matrix} a+c,\ c+d,\ 1 \\ c+1,\ a+b+c+d \end{matrix} \middle| 1 \right) \tag{4}$$

By representing our function g in terms of a hypergeometric function, we open the possibility of taking advantage of cataloged identities known for such functions. See [12] for thousands of hypergeometric identities.

While the original definition of g given in (1) requires all parameters to be positive, equation (4) does not. The function h requires that none the beta function arguments in its definition are non-positive integers. The hypergeometric function ${}_3F_2$ requires the real part of $s_\ell - s_u$ be positive in order for the series to converge, where s_ℓ is the sum of the lower parameters and s_u is the sum of the upper parameters. In our case $s_\ell - s_u$ is simply b. By using the parameter symmetries, we can make any parameter the "b" parameter.

Even though only positive parameters in g corresponds inequality probabilities, it will be useful below to use negative arguments along the way to compute g for certain positive arguments.

4 Case $a + b + c + d = 1$

In this section, we consider the case $a+b+c+d = 1$. Then the ${}_3F_2$ function has 1 as both an upper and lower parameter and reduces to the simpler hypergeometric function ${}_2F_1$ and so

$$g(a,b,c,d) = \frac{h(a,b,c,d)}{c}\, {}_2F_1\left(\begin{array}{c} a+c,\ c+d \\ c+1 \end{array} \middle| 1 \right).$$

A theorem of Gauss ([1], equation 15.1.20) allows us to evaluate the ${}_2F_1$ term in closed form:

$${}_2F_1\left(\begin{array}{c} a+c,\ c+d \\ c+1 \end{array} \middle| 1 \right) = \frac{\Gamma(c+1)\,\Gamma(1-a-c-d)}{\Gamma(1-a)\,\Gamma(1-d)}.$$

Using the reflection formula ([1], equation 6.1.17)

$$\Gamma(z)\Gamma(1-z) = \pi \csc(\pi z)$$

we find that for $a + b + c + d = 1$,

$$g(a,b,c,d) = \frac{\sin(\pi a)\sin(\pi d)}{\sin(\pi(a+b))\sin(\pi(b+d))}.$$

We can repeatedly apply the recurrence relations given above to increase the parameters by any integer amount. Therefore if we can compute $g(a,b,c,d)$, we can compute $g(a+i, b+j, c+k, d+\ell)$ for any positive integers i, j, k, and ℓ.

Furthermore, we can often compute $g(a,b,c,d)$ when $a + b + c + d$ is any integer. Consider, for example, $g(.7,.8,.1,.4)$. The parameters sum to 2 and given the definition (1) there would be nothing we could do. However, equation (4) allows negative arguments, and so we apply the recurrence relationship for the first argument and evaluate

$$g(.7,.8,.1,.4) = g(-.3,.8,.1,.4) + h(-.3,.8,.1,.4)/.3.$$

This strategy will not work if $h(a,b,c,d)$ is undefined, *i.e.* we cannot have $a+b$, $c+d$, $a+c$, or $b+d$ equal to zero. For example, we could not use the approach in this section to evaluate $g(.7,.3,.4,.6)$ because decreasing any parameter by 1 would cause h to be undefined. However, we give a method below that can be used for this case.

5 Case $a + b = c + d = 1$

Very often in practice, statisticians choose prior parameters for the beta distribution that sum to an integer. This is because the sum of the parameters can be interpreted as the number of "observations" contained in the prior and it seems natural to choose priors that make this value an integer.

If $a + b = c + d = 1$, we have

$$\begin{aligned} h(a,b,c,d) &= \frac{B(a+c,b+d)}{B(a,b)B(c,d)} \\ &= \frac{\Gamma(a+c)\Gamma(b+d)}{\Gamma(a)\Gamma(b)\Gamma(c)\Gamma(d)} \\ &= \Gamma(a+c)\Gamma(b+d)\sin(\pi a)\sin(\pi d)/\pi^2. \end{aligned}$$

Also,

$$\begin{aligned} {}_3F_2\left(\begin{array}{c} a+c,\ c+d,\ 1 \\ c+1,\ a+b+c+d \end{array} \middle| 1 \right) &= {}_3F_2\left(\begin{array}{c} a+c,\ 1,\ 1 \\ c+1,\ 2 \end{array} \middle| 1 \right) \\ &= \frac{c(\psi(c) - \psi(1-a))}{a+c-1} \end{aligned}$$

by equation 07.27.03.0053.01 of [12], provided $a + c \neq 1$ and $1 - a > 0$. Here $\psi(z)$ is the logarithmic of the Γ function, $\Gamma'(z)/\Gamma(z)$.

Therefore equation (4) can be evaluated analytically for $a+b = c+d = 1$, provided $a+c \neq 1$ and $a > 0$. The recurrence relations can be used to parlay these results to the case of $a + b$ and $c + d$ being positive integers.

6 Conclusion

In this note we defined the function $g(a, b, c, d)$ and briefly outlined some of its uses in conducting clinical trials. We expressed this function as a hypergeometric function, extending the function's domain and making it possible to apply well known identities. We showed that the function can be evaluated in closed form provided one of the arguments is a positive integer, and can often be evaluated in closed form if the parameters sum to a positive integer.

References

[1] Milton Abramowitz and Irene Stegun. *Handbook of Mathematical Functions*, Dover (1972)

[2] George Andrews, Richard Askey, and Ranjan Roy. *Special Functions*, Cambridge. (2000)

[3] Donald A. Berry and G. E. Eick. Adaptive assignment versus balanced randomization in clinical trials: a decision analysis. *Statistics in Medicine*, vol 14, 231-246 (1995).

[4] Donald A. Berry. Statistical Innovations in Cancer Research. In *Cancer Medicine e.6.* Chapter 33. London: BC Decker. (Ed: Holland J, Frei T et al.) (2003)

[5] Donald A. Berry. Bayesian statistics and the ethics of clinical trials. *Statistical Science* Volume 19(1):175-187. (2004)

[6] John D. Cook. Numerical Computation of Stochastic Inequality Probabilities. MDACC technical report UTMDABTR-008-03, (2003), revised 2005.

Available at http://www.mdanderson.org/pdf/biostats_utmdabtr00803.pdf.

[7] Ronald Graham, Donald Knuth, and Oren Patashnik *Concrete Mathematics*, Addison Wesley, 2nd edition. (1994)

[8] Hoang Nguyen and John D. Cook. Multc Lean [Computer software]. (2005)

Available at http://biostatistics.mdanderson.org/SoftwareDownload

[9] Peter Thall, Richard Simon, and Elihu Estey. Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes, *Statistics in Medicine*, vol 14, 357-379 (1995).

[10] J. Kyle Wathen, Hoang Nguyen, and John D. Cook. Inequality Calculator [Computer software] (2005)

Available at http://biostatistics.mdanderson.org/SoftwareDownload

[11] J. Kyle Wathen and Odis Wooten. Adaptive Randomization 3.0 [Computer software]. (2005)

Available at http://biostatistics.mdanderson.org/SoftwareDownload

[12] http://functions.wolfram.com/HypergeometricFunctions/