应用概率统计 第十八卷
第 一期 2002年2月

Chinese Journal of Applied Probability
and Statistics Vol. 18 No. 1 Feb. 2002

Saddlepoint Approximations for Function of Means from Several Populations*

BING-YI JING
(Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong)
ZE-HUI LI
(Department of Mathematics, Lanzhou University, Lanzhou, 730000)

Abstract

Saddlepoint approximations for marginal tail probabilities for a real-valued function of vector means from *several* populations are developed. The approximations are then shown to give great numerical accuracy, which are demonstrated in some numerical examples. Application to the bootstrap is also considered in order to avoid intensive Monte Carlo simulations.

Keywords: Bootstrap, Edgeworth expansion, saddlepoint approximation.
AMS Subject Classification: 62E20, 62G09.

§1. Introduction

Saddlepoint approximations of the Lugannani-Rice type for marginal (and conditional) tail probabilities for a real-valued function of a random vector have recently been considered by different authors, for examples, Skovgaard (1987), Daniels and Young (1991), DiCiccio and Martin (1991), Wang (1993), Jing and Robinson (1994) and others. The method used by these authors is a common one: first find a 1–1 transformation from the random vector to a new random vector, and then integrate out the unwanted variables in the joint density of the new random vector by Laplace approximation, finally apply Temme's method to obtain the desired tail area probabilities. In this paper, we shall employ the same techniques to obtain saddlepoint approximations to a function of vector means from several populations. The approach taken here closely follows that of Jing and Robinson (1994) for the case of a function of vector means from one population. Several examples will be considered to illustrate the accuracy of the saddlepoint approximations. We shall also investigate the performance of these approximations in the bootstrap context.

§2. The saddlepoint approximations for several populations

For simplicity, we shall only consider saddlepoint approximations to a function of vector means from *two* populations; the generalization to the case of more than two populations is straightforward.

Suppose we have two sets of observations $\{X_1, \cdots, X_m\}$ and $\{Y_1, \cdots, Y_n\}$, where X_i's and Y_j's are d_1-dimensional and d_2-dimensional random vectors R^{d_1} and R^{d_2} (both d_1 and d_1 are positive integers) from two populations with distribution functions F and G, respectively. Without loss of generality, we assume that $m \geq n$. In many case, we are interested in comparing some quantities from the two populations. For instance, we may be interested in the difference of the means $\mathsf{E}X_1 - \mathsf{E}Y_1$, the ratio of the means $\mathsf{E}X_1/\mathsf{E}Y_1$, or the ratio of the variances $\mathsf{Var}(X_1)/\mathsf{Var}(Y_1)$, etc. Note that all these quantities can be expressed as functions of means from the two population, and can be treated together. First let us formulate the problem.

*Research supported by National Nature Science Foundation of China 19871035 and Nature Science Foundation of Gansu Province B4.
Received 1999.10.12. Revised 2001.1.3.

Let $a_1 = g_1(x, y)$ be a real-valued function from R^d to R^1, where $x \in R^{d_1}$ and $y \in R^{d_2}$ and $d = d_1 + d_2$. Let $\overline{X} = \sum X_i/m$ and $\overline{Y} = \sum Y_i/n$. We wish to approximate

$$\mathsf{P}(A_1 \geq a_1) \equiv \mathsf{P}(g_1(\overline{X}, \overline{Y}) \geq a_1).$$

Clearly, the difference of two means $\mathsf{E}X_1 - \mathsf{E}Y_1$, the ratio of two means $\mathsf{E}X_1/\mathsf{E}Y_1$, and the ratio of two variances $\mathsf{Var}(X_1)/\mathsf{Var}(Y_1)$ can all be put into the forms of $g_1(\mu_x, \mu_y)$ which can be estimated by $g_1(\overline{X}, \overline{Y})$, where $\mu_x = \mathsf{E}X_1$ and $\mu_y = \mathsf{E}Y_1$. The purpose of this paper is to study the distribution function of $g_1(\overline{X}, \overline{Y})$. Note that $g_1(\overline{X}, \overline{Y})$ can not be put into the usual smooth function of vector means from one single population, as X_i's and Y_j's are from different distributions. However, as can be seen from Theorem 1 below, the saddlepoint approximations for both cases are strikingly similar. The development here will closely follow that of Jing and Robinson (1994), the outline of which is sketched below.

For any two vectors a and b, we shall use $\langle a, b\rangle$ to denote their inner product. Let $K_x(\theta) = \log(\mathsf{E}e^{\langle \theta, X_1\rangle})$ and $K_y(\tau) = \log(\mathsf{E}e^{\langle \tau, Y_1\rangle})$, where $\theta \in R^{d_1}$ and $\tau \in R^{d_2}$, be the cumulant generating functions of X_1 and Y_1, respectively. By saddlepoint approximations for the density of $\overline{X}$ and $\overline{Y}$ (see Daniels (1987), for instance), we have

$$\begin{aligned} f_{\overline{X},\overline{Y}}(x, y) =& \frac{e^{-m[\langle \theta, x\rangle - K_x(\theta)]}}{(2\pi/m)^{d_1/2}[\det(K_x''(\theta))]^{1/2}} \frac{e^{-n[\langle \tau, y\rangle - K_y(\tau)]}}{(2\pi/n)^{d_2/2}[\det(K_y''(\tau))]^{1/2}}(1 + O(n^{-1})) \\ =& \frac{e^{-N\Lambda(x,y)}}{(2\pi/N)^{d/2}\Delta^{1/2}}(1 + O(n^{-1})), \end{aligned}$$

where the saddlepoints θ and τ are solutions of $K_x'(\theta) = x$ and $K_y'(\tau) = y$, respectively, and

$$\begin{aligned} \Lambda(x, y) =& \frac{m}{N}(\langle \theta, x\rangle - K_x(\theta)) + \frac{n}{N}(\langle \tau, y\rangle - K_y(\tau)), \\ \Delta =& \det\Big[\frac{N}{m}K_x''(\theta)\Big] \det\Big[\frac{N}{n}K_y''(\tau)\Big]. \end{aligned}$$

Now let us construct a 1–1 transformation

$$\begin{cases} a_1 = g_1(x_1, \cdots, x_{d_1}, y_1, \cdots, y_{d_2}) \\ a_2 = g_2(x_1, \cdots, x_{d_1}, y_1, \cdots, y_{d_2}) \\ \cdots\cdots\cdots \\ a_d = g_d(x_1, \cdots, x_{d_1}, y_1, \cdots, y_{d_2}), \end{cases}$$

that is, $a = g(x, y) \equiv g(z)$. Write the inverse transformation as $z = g^{-1}(a)$. Denote the absolute value of the jacobian of the transformation as $J(a) = |\det(\partial z/\partial a)|$. So the joint density of $A \equiv g(\overline{X}, \overline{Y}) \equiv (g_1(\overline{X}, \overline{Y}), \cdots, g_d(\overline{X}, \overline{Y}))$ is $f_A(a) = f_{\overline{X},\overline{Y}}(x(a), y(a))J(a)$. Hence the marginal density for $A_1 = g_1(\overline{X}, \overline{Y})$ is

$$f_{A_1}(a) = \int_{R^{d-1}} f_A(a)da_2 \cdots da_d = \frac{e^{-NH(a_1)}\widetilde{J}}{\widetilde{\Delta}^{1/2}[\det(\widetilde{L}_{22})]^{1/2}}(1 + O(n^{-1})), \tag{1}$$

where

$$\begin{aligned} L(a) =& \Lambda(z(a)), \\ L_{22} =& \partial L(a)/\partial(a_2, \cdots, a_d), \\ H(a_1) =& \inf_{a_2, \cdots, a_d} L(a) \equiv L(a_1, \widetilde{a}_2, \cdots, \widetilde{a}_d), \end{aligned}$$

and write $\widetilde{a} = (a_1, \widetilde{a}_2, \cdots, \widetilde{a}_d)$, (i.e. $\widetilde{a}$ minimizes $L(a)$ for fixed a_1) and also $\widetilde{J}$, $\widetilde{L}_{22}$ and $\widetilde{\Delta}$ are evaluated at $\widetilde{a}$.

To get a saddlepoint approximation for the tail probability $\mathsf{P}(A_1 \geq a_1)$, we can integrate $f_{A_1}(a)$ in (1). By applying the Temme's method as in Jing and Robinson (1994), we get the following theorem.

Theorem 1 Under some regularity conditions, we have

$$\mathsf{P}(A_1 \geq a_1) = 1 - \Phi(\widehat{w}\sqrt{N}) - \frac{\phi(\widehat{w}\sqrt{N})}{\sqrt{N}}\Big(\frac{1}{\widehat{w}} - \frac{1}{\widehat{u}} + O(n^{-3/2})\Big), \tag{2}$$

where

$$\widehat{w} = \sqrt{2H(a_1)}\,\mathrm{sign}(a_1 - \alpha_1),$$
$$\widehat{u} = \sqrt{\widetilde{\Delta}\ \det(\widetilde{L}_{22})}\ \widetilde{L}_1/\widetilde{J},$$

where $\widetilde{L}_1 = \partial L(\widetilde{a})/\partial a_1$ and $\alpha_1 = \{a_1 : \inf_{a_1} H(a_1)\}$.

To apply Theorem 1, one can take the following easy steps.

Step 1: For fixed a_1, one can solve $(\widetilde{\theta}, \widetilde{\tau}, \widetilde{a}_2, \cdots, \widetilde{a}_d)$ from

$$\begin{cases} K_x'(\theta) = x(a) \\ K_y'(\tau) = y(a) \\ L_2(a) = \dfrac{m}{N}\Big(\dfrac{\partial x(a)}{\partial(a_2, \cdots, a_d)}\Big)^T\theta + \dfrac{n}{N}\Big(\dfrac{\partial y(a)}{\partial(a_2, \cdots, a_d)}\Big)^T\tau = 0. \end{cases}$$

Note here we have $(2d-1)$ unknowns and $(2d-1)$ equations.

Step 2: Write $\widetilde{a} = (a_1, \widetilde{a}_2, \cdots, \widetilde{a}_d)$, then calculate

$$\begin{aligned} H(a_1) =& L(\widetilde{a}) = \frac{m}{N}[\langle\widetilde{\theta}, x(\widetilde{a})\rangle - K_x(\widetilde{\theta})] + \frac{n}{N}[\langle\widetilde{\tau}, y(\widetilde{a})\rangle - K_y(\widetilde{\tau})], \\ \widetilde{J}(a_1) =& \Big|\det\Big(\frac{\partial z(\widetilde{a})}{\partial a}\Big)\Big|, \\ \widetilde{L}_1 =& \frac{m}{N}\Big(\frac{\partial x(\widetilde{a})}{\partial a_1}\Big)^T\widetilde{\theta} + \frac{n}{N}\Big(\frac{\partial y(\widetilde{a})}{\partial a_1}\Big)^T\widetilde{\tau}, \\ \widetilde{L}_{22} =& \frac{m}{N}\widetilde{L}_{x22} + \frac{n}{N}\widetilde{L}_{y22}, \end{aligned}$$

where

$$\widetilde{L}_{x22} = \Big(\frac{\partial x(\widetilde{a})}{\partial(a_2, \cdots, a_d)}\Big)^T(K_x''(\widetilde{\theta}))^{-1}\Big(\frac{\partial x(\widetilde{a})}{\partial(a_2, \cdots, a_d)}\Big) + M_x(\widetilde{a}, \widetilde{\theta}).$$

$$M_x(a, \theta) = \begin{pmatrix} \langle\theta, \dfrac{\partial^2 x}{\partial a_2 \partial a_2}\rangle & \cdots & \langle\theta, \dfrac{\partial^2 x}{\partial a_2 \partial a_d}\rangle \\ \cdots & \cdots & \cdots \\ \langle\theta, \dfrac{\partial^2 x}{\partial a_d \partial a_2}\rangle & \cdots & \langle\theta, \dfrac{\partial^2 x}{\partial a_d \partial a_d}\rangle \end{pmatrix}_{(d-1)\times(d-1)},$$

and $\widetilde{L}_{y22}$ and $M_y(\widetilde{a}, \widetilde{\tau})$ are similarly defined.

Step 3: Calculate $\widehat{w}$ and $\widehat{u}$, and apply (2) in Theorem 1.

An alternative saddlepoint approximation can also be obtained, which does not depend on the transformation $a = g(z)$, see Jing and Robinson (1994), DiCiccio and Martin (1991) for more details in related cases. We shall not pursue this here. But it should be pointed out that it is often easier to construct simple 1-1 transformation $a = g(z)$ (which is often easy to find) and then apply Theorem 1 than using the alternative formula.

§3. Some numerical examples

3.1 An exact case

Let $F = N(0,1)$ and $G = N(0,1)$. To find out the distribution for $\mathsf{P}(\overline{X} + \overline{Y} \geq a_1)$, we choose $d = 2$, $a_1 = x + y$ and $a_2 = y$. Applying Theorem 1, we get

$$\mathsf{P}(\overline{X} + \overline{Y} \geq a_1) = 1 - \Phi(a_1/\sqrt{1/m + 1/n}).$$

In this case, the saddlepoint approximation is exact.

3.2 Two-sample studentized-t statistic

Let $X_1, \cdots, X_m \sim N(0,1)$ and $Y_1, \cdots, Y_n \sim N(0,1)$. Define the two-sample studentized t statistic as

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{{S_x}^2/m + {S_y}^2/n}},$$

where $S_x^2 = \sum(X_i - \overline{X})^2/(m-1)$, $S_y^2 = \sum(Y_j - \overline{Y})^2/(n-1)$. It is known that when $m = n$, T follows Student-t distribution with $(2n-2)$ degrees of freedom. So in the following, we shall apply the saddlepoint approximation (2) for $\mathsf{P}(T \geq a_1)$ and compare the results with the exact probabilities.

To apply Theorem 1, we need to find out a 1–1 transformation and also the cumulant generating functions. For the first part, we choose $d = 4$ and the transformation to be

$$\begin{cases} a_1 = (x_1 - y_1)/\sqrt{(x_2 - x_1^2)/(m-1) + (y_2 - y_1^2)/(n-1)} \\ a_2 = x_2 - x_1^2 \\ a_3 = y_1 \\ a_4 = y_2 - y_1^2. \end{cases}$$

Its inverse transformation is

$$\begin{cases} x_1 = a_3 + a_1\sqrt{a_2/(m-1) + a_4/(n-1)} \\ x_2 = a_2 + x_1^2 \\ y_1 = a_3 \\ y_2 = a_4 + a_3^2. \end{cases}$$

It is easy to see that the Jacobian of this transformation is $\sqrt{a_2/(m-1) + a_4/(n-1)}$. The cumulant generating functions of both distributions are

$$K_x(\theta_1, \theta_2) = \frac{\theta_1^2}{2(1-2\theta_2)} - \log\Big(\frac{1-2\theta_2}{2}\Big),$$
$$K_y(\tau_1, \tau_2) = \frac{\tau_1^2}{2(1-2\tau_2)} - \log\Big(\frac{1-2\tau_2}{2}\Big).$$

Some numerical results are presented in Table 1 to illustrate the performance of the saddlepoint approximations. For the case $m = n$, we know that T follows Student-t distribution with $(2n-2)$ degrees of freedom, so the exact percentiles can be found from the t table. To calculate approximate percentiles by saddlepoint approximations, we can invert $\mathsf{P}(A_1 \geq a_1) = \alpha$ for various values of α, using saddlepoint approximations (2) in Theorem 1. For comparison purposes, we also include percentiles obtained from the Edgeworth expansions for T, which have been derived in Hall and Martin (1988). It takes the following form,

$$\mathsf{P}(T \leq x) = \Phi(x) + p_1(x)\phi(x) + p_2(x)\phi(x) + O(n^{-3/2}), \tag{3}$$

where

$$p_1(x) = \frac{1}{6}\gamma(2x^2+1),$$
$$p_2(x) = x\Big[\frac{1}{12}\kappa(x^2-3) - \frac{1}{18}\gamma^2(x^4+2x^2-3) - \frac{1}{4}(ax^2+3b)\Big],$$

where σ_x^2, γ_x, κ_x are the variance, skewness and kurtosis for X with similar quantities defined for Y, and $\sigma_{pool}^2 = \sigma_x^2/m + \sigma_y^2/n$ and

$$\gamma = \Big(\frac{\gamma_x}{m^2} - \frac{\gamma_y}{n^2}\Big)\Big/\sigma_{pool}^3, \qquad \kappa = \Big(\frac{\kappa_x}{m^3} + \frac{\kappa_y}{n^3}\Big)\Big/\sigma_{pool}^4,$$
$$a = \Big(\frac{\sigma_x^4}{m^3} + \frac{\sigma_y^4}{n^3}\Big)\Big/\sigma_{pool}^3, \qquad b = 2\sigma_x^2\sigma_y^2\Big(\frac{m+n}{m^2n^2}\Big)\Big/\sigma_{pool}^3.$$

As can be seen from Table 1, the percentiles corresponding to saddlepoint approximations are extremely close to the "exact" ones, for both $m = n = 5$ and $m = n = 20$ cases. On the other hand, the percentiles

obtained from the Edgeworth expansions perform very poorly by comparison, particularly for the case small sample size $m = n = 5$.

Table 1 Comparisons of the percentiles of T under normal distributions

	$m = n = 5$			$m = n = 20$		
Probability	"Exact"	Saddlepoint	Edgeworth	"Exact"	Saddlepoint	Edgeworth
0.60	0.262	0.262	0.403	0.255	0.255	0.291
0.70	0.546	0.541	0.678	0.528	0.529	0.563
0.80	0.889	0.881	1.004	0.851	0.852	0.883
0.90	1.397	1.377	1.460	1.304	1.303	1.329
0.95	1.860	1.843	1.841	1.686	1.685	1.698
0.99	2.897	2.869	2.558	2.429	2.428	2.394
0.995	3.355	3.323	2.821	2.712	2.711	2.650
0.999	4.501	4.457	3.360	3.319	3.319	3.178
0.9995	5.041	4.993	3.569	3.566	3.565	3.383
0.9999	6.442	6.379	4.013	4.116	4.114	4.385

§4. Application to the bootstrap

Suppose we have two independent data sets $\mathcal{X} = \{X_1, \cdots, X_m\}$ and $\mathcal{Y} = \{Y_1, \cdots, Y_n\}$ from two unknown distributions F and G, respectively. Define T as in Example 2 except that $(m-1)$ and $(n-1)$ in the definitions of S_x^2 and S_y^2 will now be replaced by m and n. Then the bootstrap estimate of $\mathsf{P}(T \geq a_1)$ is $\mathsf{P}^*(T^* \geq a_1)$, where

$$T^* = \frac{(\overline{X}^* - \overline{Y}^*) - (\overline{X} - \overline{Y})}{\sqrt{S_x^{*2}/m + S_y^{*2}/n}}$$

and $\mathcal{X}^* = \{X_1^*, \cdots, X_m^*\}$ and $\mathcal{Y}^* = \{Y_1^*, \cdots, Y_n^*\}$ are drawn with replacement from $\mathcal{X}$ and $\mathcal{Y}$, respectively, and $\overline{X}^*$, $\overline{Y}^*$ are the bootstrap means, S_x^{*2} and S_y^{*2} are the bootstrap variances of $\mathcal{X}^*$ and $\mathcal{Y}^*$, respectively. Also P^* indicates the probability conditional on $\mathcal{X}$ and $\mathcal{Y}$.

Generally, $\mathsf{P}^*(T^* \geq a_1)$ can be estimated by Monte Carlo simulation. But now we apply the saddlepoint approximation (2) to avoid the intensive simulation, in a similar way to Davison and Hinkley (1988) and Daniels and Young (1991). Other related work includes DiCiccio and Martin (1991), Wang (1993), Jing and Robinson (1994) and so on. To apply the saddlepoint approximation (2), we choose the same transformation as in Example 2, except that we now replace $m-1$ and $n-1$ there to m and n, respectively. The empirical cumulant generating functions for $\mathcal{X}$ and $\mathcal{Y}$ are

$$\widehat{K}_x(\theta_1, \theta_2) = \log\Big(\frac{1}{m}\sum_{i=1}^{m} \exp(\theta_1(x_i - \overline{x}) + \theta_2(x_i - \overline{x})^2)\Big),$$
$$\widehat{K}_y(\tau_1, \tau_2) = \log\Big(\frac{1}{n}\sum_{j=1}^{n} \exp(\tau_1(y_j - \overline{y}) + \tau_2(y_j - \overline{y})^2)\Big).$$

For comparison purposes, we generated two random samples of sizes $m = n = 10$ from two independent normal distributions. The data are

$$X_i : -1.35, -0.81, -0.24, 0.28, -1.55, -0.59, 0.50, 0.82, 0.96, -1.82;$$
$$Y_j : -0.61, -0.82, 1.32, -0.27, 0.93, -0.99, -0.11, 1.27, 0.87, 0.04.$$

We calculate the tail probability $\mathsf{P}^*(T^* \geq a_1)$ for various values of a_1 by the following three methods: "exact" probability obtained by 100,000 simulations; saddlepoint approximations; and Edgeworth expansions, which is the bootstrap version of (3) obtained by replacing all the population quantities by their sample counterparts. The results are presented in Table 2.

As can be seen from the table, the saddlepoint approximations give very accurate results, in particular, at the tails of the distributions, as one might have expected. On the other hand, Edgeworth expansions perform rather poorly at the tails. It is interesting to note, though, that Edgeworth expansions give better approximations at the center of the distribution here.

Table 2 Comparisons of the bootstrap probabilities $\mathsf{P}^*(T^* \geq a_1)$ $(m = n = 10)$

	$\mathsf{P}^*(T^* \geq a_1)$		
a_1	"Exact"	Saddlepoint	Edgeworth
3.267	0.01	0.012	0.006
2.713	0.025	0.020	0.016
2.294	0.05	0.051	0.044
1.835	0.10	0.114	0.112
1.313	0.20	0.185	0.188
0.964	0.30	0.265	0.274
0.670	0.40	0.353	0.368
0.400	0.50	0.451	0.471
0.132	0.60	0.559	0.584
-0.1529	0.70	0.679	0.707
-0.4842	0.80	0.819	0.853
-0.9655	0.90	0.901	0.933
-1.3813	0.95	0.946	0.972
-1.7493	0.975	0.976	0.993

References

[1] Daniels, H.E., Uniform approximations for tail probabilities, *International Statistical Review*, **55**(1987). 37–48.
[2] Daniels, H.E. and Young, G.A., Saddlepoint approximation for the studentized mean, with an application to the bootstrap, *Biometrika*, **78**(1991), 169–179.
[3] Davison, A.C. and Hinkley, D.V., Saddlepoint approximations in resampling methods, *Biometrika*, **75**(1988), 417–431.
[4] DiCiccio, T.J. and Martin, M.A., Approximations of marginal tail probabilities for a class of smooth functions with applications to Bayesian and conditional inference, *Biometrika*, **78**(1991), 891–902.
[5] Hall and Martin, On the bootstrap and two-sample problems, *Austal. J. Statist.*, **30A**(1988), 179–192.
[6] Jing, B.Y. and Robinson, J., Saddlepoint approximations for marginal and conditional probabilities of transformed variables. *Annals of Statistics*, **22(3)**(1994), 1115–1132.
[7] Skovgaard, Ib.M., Saddlepoint expansions for conditional distributions, *J. Appl. Prob.*, **24**(1987). 875–887.
[8] Wang, S., Saddlepoint approximations in conditional inference, *J. Appl. Probab.*, **30**(1993), 397–404.

几个总体均值函数的鞍点逼近

邢 炳 义
(香港科学技术大学, 香港)
李 泽 慧
(兰州大学, 兰州, 730000)

本文发展了几个总体的均值向量的函数的尾部概率的鞍点逼近方法, 并应用它于Bootstrap估计中, 以代替Monte Carlo模拟, 一些数值例子说明了其近似精度.

关键词: Bootstrap, Edgeworth展开, 鞍点逼近.

学科分类号: 0212.