温利民,张良超,章溢,等. 基于贝塔分布的最优置信区间研究[J]. 江西师范大学学报(自然科学版),2022,46(4):342-348.
WEN Limin,ZHANG Liangchao,ZHANG Yi,et al. The study on optimal confidence interval based on beta distribution[J]. Journal of Jiangxi Normal University(Natural Science),2022,46(4):342-348.

文章编号: 1000-5862(2022)04-0342-07

基于贝塔分布的最优置信区间研究

温利民[1],张良超[1],章 溢[2],刘 蔚[1]
(1. 江西师范大学数学与统计学院,江西 南昌 330022;2. 江西师范大学财政金融学院,江西 南昌 330022)

摘要: 基于贝塔分布的概率特征性质,该文研究了一类特殊的贝塔分布的最优区间估计;进而,将得到的区间估计与等尾置信区间进行了比较. 结果表明:使用最短置信区间作为未知参数的区间估计,估计的精度得到显著提高. 最后,利用数值模拟的方法给出了贝塔分布的最短区间估计用表.

关键词: 贝塔分布;等尾置信区间;最短置信区间
中图分类号: O 211.9 **文献标志码:** A **DOI:** 10.16357/j.cnki.issn1000-5862.2022.04.03

0 引言

在概率论与数理统计中,区间估计是参数估计的重要内容. 设θ是总体的一个未知参数,在总体为连续型分布的情况下,通过枢轴量法可得到参数的置信区间. 然而当枢轴量的分布为单峰非对称时,利用传统方法构造的区间是等尾置信区间,而不是最优置信区间.

关于最优置信区间的定义,常见的有2种:一种是在给定置信水平的区间估计下要求平均区间长度最短,另一种是在给定平均区间长度下要求置信度尽可能大或精确度尽可能高. 本文主要考虑第1种定义,即在给定置信度水平下求解平均值区间长度最短的区间估计. 在数理统计中,关于置信区间的最优性的研究较多. 夏乐天等[1]讨论了指数分布参数的最短区间估计;袁长迎等[2]在伽玛分布形状参数已知时研究了尺度参数的最短区间估计;徐美萍等[3]研究了在威布尔分布中尺度参数的最短区间估计;王秀丽[4]研究了均匀分布参数的最短置信区间;薛峰等[5]利用粒子群优化算法研究了贝塔分布参数的最短置信区间. 在区间估计问题中,枢轴量G的分布通常是单峰分布,如正态分布、t分布、卡方分布、F分布等. 由于正态分布和t分布是单峰对称分布,且未知参数θ在枢轴量G的分子上,所以用传统方法构造的区间就是最短置信区间. 然而由于卡方分布和F分布为单峰非对称分布,所以传统方法构造的区间不是最短置信区间. 孙鹏哲等[6]研究了卡方分布的最短置信区间,得出在各种自由度和常用的置信水平下最优左侧尾概率分配统计表. 李广正[7]运用拉格朗日乘数法和利用Mathematica软件给出了F分布的最短区间估计用表. 上述文献对于F分布和卡方分布的最短置信区间讨论较多,而对于贝塔分布的研究较少.

在已有的研究基础上,本文给出在一类形状参数下基于贝塔分布的最短置信区间,这可以适用于当枢轴量服从贝塔分布时求解最短置信区间. 利用具有单峰的贝塔分布的密度函数,可得到待估参数的最短置信区间. 此外,在贝叶斯框架下,当参数的后验分布服从贝塔分布时,可得到参数的最短后验区间估计. 通过模拟分析,验证贝塔分布的最短置信区间的优越性.

1 最短置信区间的概念界定

1.1 区间估计的定义及说明

定义1 设总体X具有概率函数$f(x;\theta)$,θ为未

收稿日期: 2021-11-16
基金项目: 国家自然科学基金(72263019)和江西省学位与研究生教育教学改革课题(JXYJG-2021-066)资助项目.
作者简介: 温利民(1979—),男,江西石城人,教授,博士,博士生导师,主要从事精算学与金融统计推断的研究. E-mail: wlmjxnu@163.com

知参数. $X_1,X_2,\cdots,X_n$ 是取自总体 X 的一个样本,若对于事先给定的 $\alpha(0<\alpha<1)$,存在 2 个统计量 $T_1=T_1(X_1,X_2,\cdots,X_n)$ 与 $T_2=T_2(X_1,X_2,\cdots,X_n)$ 使得 $P(T_1\leqslant\theta\leqslant T_2)=1-\alpha$ 成立,则称区间 $[T_1,T_2]$ 为参数 θ 的置信水平为 $1-\alpha$ 的置信区间,其中 T_1 和 T_2 分别被称为置信水平 $1-\alpha$ 的置信下限和置信上限.

由定义可以看出,T_1 和 T_2 都是不依赖于未知参数的随机变量,因此置信区间 $[T_1,T_2]$ 是随机区间. $P_\theta(T_1\leqslant\theta\leqslant T_2)=1-\alpha$ 表示: 对样本 $X_1,X_2,\cdots,X_n$ 观测多次,得到许多不同的区间 $[T_1,T_2]$,在这些确定的区间中,大约有 $(1-\alpha)\times100\%$ 的比例包含了未知参数 θ 的真值,而约有 $\alpha\times100\%$ 的比例不包含其真值. 需特别注意,对于一次抽样所得到的一个区间,决不能理解为"不等式 $T_1\leqslant\theta\leqslant T_2$ 成立的概率为 $1-\alpha$". 因为在给定样本下的 T_1 和 T_2 是 2 个确定的数,从而只有 2 种可能: 要么这个区间包含 θ; 要么这个区间不包含 θ. 因此,定义说明区间 $[T_1(X_1,X_2,\cdots,X_n),T_2(X_1,X_2,\cdots,X_n)]$ 属于包含未知参数 θ 的区间类的置信水平是 $1-\alpha$,这也说明置信水平与概率是有所区别的,不可混淆.

1.2　区间估计的可靠度和精确度及其关系

当参数真值为 θ 时,自然希望随机区间 $[T_1,T_2]$ 包含 θ 的概率 $P_\theta(T_1\leqslant\theta\leqslant T_2)$ 要大. 因此,一个好的区间估计应该对所有属于参数空间 Θ 的 θ,概率 $P_\theta(T_1\leqslant\theta\leqslant T_2)$ 都相当大.

定义 2　设 $[T_1(X_1,X_2,\cdots,X_n),T_2(X_1,X_2,\cdots,X_n)]$ 为参数 θ 的一个区间估计,则称区间包含 θ 的概率在参数空间 Θ 上的下确界 $\inf\limits_{\theta\in\Theta}P_\theta(T_1\leqslant\theta\leqslant T_2)$ 为该区间估计的置信系数.

若一个区间估计的置信系数越大,则该区间估计的可靠度越高. 但是,构造一个置信系数很大的区间估计并不是一件难事. 如将明天中午 12 点的气温估计在 $-10\sim50$ ℃ 之间,这个估计的可靠度很高,但由于它的范围太大,很不精确,所以一个好的区间估计还有一个精确度的要求.

1.3　最短置信区间的确定

区间估计的精确度的标准不止一个,常用的标准有 2 个:

1) 区间 $[T_1(X_1,X_2,\cdots,X_n),T_2(X_1,X_2,\cdots,X_n)]$ 的平均长度 $E_\theta[T_2-T_1]$ 要短,即区间的范围不能太大,这是符合实际的;

2) 设参数真值为 θ,在 $\theta^*\neq\theta$ 时,自然希望区间 $[T_1,T_2]$ 包含 θ^* 的概率要小,即区间 $[T_1,T_2]$ 包含非真值的情况出现越少越好.

在给定样本容量 n 后,可靠度与精确度是相互制约着的. 为了提高可靠度,可以通过增大区间范围来实现,但是会降低精确度. 反过来,为了提高精确度,可通过减小区间范围来实现,但是会降低可靠度. 为此本文采用 J. Neyman 建议的某种折中方案: 在使得置信系数达到一定要求的前提下,寻找精确度尽可能高的区间估计,也就是要求区间平均长度尽可能短,或者区间包含非真值的概率尽可能小,这 2 个要求可能同时达到,也可能不同时达到.

下面介绍构造置信区间的常用方法,即枢轴量法. 可按下列 3 个步骤构造 $\eta=g(\theta)$ 的置信区间.

1) 构造样本 $(X_1,X_2,\cdots,X_n)$ 和未知参数 η 的一个函数 $G=G(X_1,X_2,\cdots,X_n;\eta)$,要求 G 的分布与未知参数 η 无关,称具有这种性质的函数为枢轴量.

2) 对给定的 $\alpha(0<\alpha<1)$,选取 2 个常数 c 和 $d(c<d)$,使得

$$P_\theta(c\leqslant G(X_1,X_2,\cdots,X_n;\eta)\leqslant d)=1-\alpha,\forall\theta\in\Theta.$$

3) 若不等式 $c\leqslant G(X_1,X_2,\cdots,X_n;\eta)\leqslant d$ 可等价地变换为

$$T_1(X_1,X_2,\cdots,X_n)\leqslant\eta\leqslant T_2(X_1,X_2,\cdots,X_n),$$

则

$$P_\theta(T_1(X_1,X_2,\cdots,X_n)\leqslant\eta\leqslant T_2(X_1,X_2,\cdots,X_n))=1-\alpha,\forall\theta\in\Theta,$$

从而 $[T_1(X_1,X_2,\cdots,X_n),T_2(X_1,X_2,\cdots,X_n)]$ 是 η 的一个置信水平为 $1-\alpha$ 的置信区间. 当 $G(X_1,X_2,\cdots,X_n;\eta)$ 是 η 的连续严格单调函数时,这 2 个不等式的等价关系总是可以做到的. 当 $g(\theta)=\theta$ 时,$[T_1(X_1,X_2,\cdots,X_n),T_2(X_1,X_2,\cdots,X_n)]$ 是 θ 的一个置信水平为 $1-\alpha$ 的置信区间.

一般来讲,满足要求的 c 和 d 是不唯一的,若有可能,应选在平均区间长度 $E_\theta[T_2-T_1]$ 达到最短时的 c 与 d,则此时所求得的置信区间被称为置信水平 $1-\alpha$ 的最短置信区间. 由于区间平均长度与所构造的枢轴量密切相关,所以接下来考虑在 2 类枢轴量形式下的最短置信区间估计. 以下讨论都是在枢轴量服从单峰分布情况下进行的.

1) 枢轴量 G 具有如下形式:

$$G=T(X_1,X_2,\cdots,X_n)(\theta+U(X_1,X_2,\cdots,X_n)),$$

其中 $T(X_1,X_2,\cdots,X_n)>0$.

由 $P_\theta(c\leqslant G\leqslant d)=1-\alpha$(即 $P_\theta(c\leqslant T(\theta+U)\leqslant d)=1-\alpha$)得到参数 θ 的置信区间为 $[cT^{-1}-U,dT^{-1}-U]$,平均区间长度为 $(d-c)E_\theta(T^{-1})$. 考虑在平均区间长度最短下的区间估计,即为求如下条件极值问题:

$$\begin{cases}\min\ d-c,\\ P_\theta(c\leqslant G\leqslant d)=1-\alpha.\end{cases}\tag{1}$$

运用拉格朗日乘数法,令 $L=d-c+\lambda(F(d)-F(c)-1+\alpha)$,对 L 关于 c、d 分别求偏导并令其为 0 得

$$\begin{cases}\partial L/\partial c=-1-\lambda f(c)=0,\\ \partial L/\partial d=1+\lambda f(d)=0,\end{cases}\tag{2}$$

其中 $F(\cdot)$、$f(\cdot)$ 分别表示 G 的分布函数与密度函数. 由式(2)可知,$f(c)=f(d)$. 所以条件极值问题(1)可转化为如下所示的 2 元方程组求解问题:

$$\begin{cases}f(c)=f(d),\\ F(d)-F(c)=1-\alpha.\end{cases}\tag{3}$$

当 $f(x)$ 为单峰对称密度函数(比如正态分布、t 分布的密度函数)时,由式(3)容易看出,此时等尾置信区间即为最短置信区间. 当 $f(x)$ 为单峰非对称密度函数(如卡方分布和 F 分布的密度函数)时,只要求解式(3)就可得最短置信区间,文献[8]证明了式(3)有唯一解,具体求解可利用求根法或黄金分割法[9]得到.

2)枢轴量 G 具有如下形式:

$G=T(X_1,X_2,\cdots,X_n)(\theta+U(X_1,X_2,\cdots,X_n))^{-1}$,其中 $T(X_1,X_2,\cdots,X_n)\geqslant 0$.

由 $P_\theta(c\leqslant G\leqslant d)=1-\alpha$(即 $P_\theta(c\leqslant T(\theta+U)^{-1}\leqslant d)=1-\alpha$,其中 c、d 同号)得到参数 θ 的置信区间为 $[d^{-1}T-U,c^{-1}T-U]$,平均区间长度为 $(c^{-1}-d^{-1})E_\theta(T)$. 考虑在平均区间长度最短下的区间估计,即为求如下条件极值问题:

$$\begin{cases}\min\ c^{-1}-d^{-1},\\ P_\theta(c\leqslant G\leqslant d)=1-\alpha.\end{cases}\tag{4}$$

同理运用拉格朗日乘数法,把条件极值问题(4)可转化为如下所示的 2 元方程组求解问题:

$$\begin{cases}c^2f(c)=d^2f(d),\\ F(d)-F(c)=1-\alpha.\end{cases}\tag{5}$$

当 $f(x)$ 为单峰对称密度函数(如正态分布、t 分布的密度函数)时,由式(5)容易看出,此时等尾置信区间对应的 c、d 满足 $f(c)=f(d)$,然而 $c^2f(c)\neq d^2f(d)$,从而等尾置信区间不是最短置信区间,且文献[8]证明了式(5)有唯一解.

2 贝塔分布的最短置信区间

现有文献对于 F 分布和卡方分布的最短置信区间讨论较多,而对于贝塔分布的研究较少. 事实上,贝塔分布在一类形状参数下也是单峰非对称分布. 接下来研究在枢轴量服从贝塔分布时的最短置信区间,首先给出一个引理.

引理 1[10] 设总体 X 的密度函数为 $p(x)$,分布函数为 $F(x)$,$X_1,X_2,\cdots,X_n$ 为样本,则样本极差 $Z=X_{(n)}-X_{(1)}$ 的分布函数为

$$F_Z(x)=n\int_{-\infty}^{+\infty}(F(t+x)-F(t))^{n-1}p(t)\,\mathrm{d}t.$$

2.1 单峰贝塔分布的最短置信区间

定理 1 设 $X_1,X_2,\cdots,X_n$ 是来自均匀分布 $U(\theta_1,\theta_2)$ 的一个样本,则 $\theta_2-\theta_1$ 的一个无偏估计为

$$\hat{\theta}=(n+1)(X_{(n)}-X_{(1)})/(n-1).$$

证 令 $Y_i=(X_i-\theta_1)/(\theta_2-\theta_1)$,$i=1,2,\cdots,n$,则 Y_i 独立同分布于 $U(0,1)$,由引理 1 可知

$$Y_{(n)}-Y_{(1)}\sim\mathrm{Beta}(n-1,2).$$

由贝塔分布的数学期望公式有

$$E(Y_{(n)}-Y_{(1)})=E((X_{(n)}-X_{(1)})/(\theta_2-\theta_1))=(n-1)/(n+1),$$

所以 $E(\hat{\theta})=\theta_2-\theta_1$. 定理 1 得证.

由定理 1 的证明过程知,事实上可构造枢轴量:

$$G=(X_{(n)}-X_{(1)})/(\theta_2-\theta_1)\sim\mathrm{Beta}(n-1,2).$$

对给定的置信水平 $1-\alpha$,若 c、d 满足

$$P(c\leqslant G\leqslant d)=1-\alpha,$$

可得 $\theta_2-\theta_1$ 的置信区间为

$$[(X_{(n)}-X_{(1)})d^{-1},(X_{(n)}-X_{(1)})c^{-1}].\tag{6}$$

注意到,要使得式(6)的平均区间长度最短等价于求解式(5),其中 $F(\cdot)$ 为 $\mathrm{Beta}(n-1,2)$ 的分布函数,$f(\cdot)$ 为 $\mathrm{Beta}(n-1,2)$ 的密度函数.

2.2 与等尾置信区间的比较

接下来讨论给定置信水平 0.90 和 0.95 以及样本容量 n,比较 $\theta_2-\theta_1$ 的最短置信区间与等尾置信区间. 首先由式(6)知,平均区间长度正比于 $c^{-1}-d^{-1}$,因此记等尾置信区间长度为

$$L_1=(\mathrm{Beta}_{\alpha/2}(n-1,2))^{-1}-(\mathrm{Beta}_{1-\alpha/2}(n-1,2))^{-1}.$$

其中 c、d 是在式(6)中当平均区间长度达到最短时的取值,具体可通过 Matlab 软件中的解方程组命令 fsolve 求解,因此记最短置信区间长度为

$$L_2=c^{-1}-d^{-1}.$$

2 者的相对差异记为 $e(n)=(L_1-L_2)/L_2$. 对不同的样本容量,得到如表 1 和表 2 所示的结果.

在表 1、表 2 中的 c、d 是在式(6)中当平均区间长度达到最短时的取值. 从表 1、表 2 可以看出:在给定置信水平情况下,2 种方法求得的置信区间长度都随着样本容量 n 的增加而变短,而且 c、d 也随着样本容量增加而变大. 这是由于贝塔分布的形状参数 $n-1$ 增加,其密度函数越呈现“尖峰左偏”形状,样本的集中趋势越来越明显.

表 1　在置信水平 0.90 下最短置信区间与等尾置信区间的比较

n	L_1	L_2	c	d	$e(n)$ /%
4	2.914 3	2.116 8	0.320 08	0.992 65	37.67
6	1.324 2	1.038 5	0.488 79	0.992 72	27.51
8	0.840 6	0.681 0	0.592 58	0.993 47	23.44
10	0.612 4	0.505 1	0.661 86	0.994 22	21.25
12	0.480 7	0.401 0	0.711 17	0.994 86	19.89
14	0.395 2	0.332 2	0.748 01	0.995 38	18.96
16	0.335 4	0.283 6	0.776 54	0.995 82	18.29
18	0.291 2	0.247 3	0.799 29	0.996 19	17.77
20	0.257 3	0.219 2	0.817 85	0.996 50	17.37
22	0.230 4	0.196 8	0.833 27	0.996 76	17.05
24	0.208 6	0.178 6	0.846 28	0.996 99	16.78
26	0.190 5	0.163 5	0.857 42	0.997 19	16.56

表 2　在置信水平 0.95 下最短置信区间与等尾置信区间的比较

n	L_1	L_2	c	d	$e(n)$ /%
4	4.079 0	3.021 0	0.248 51	0.997 10	35.02
6	1.742 1	1.389 6	0.417 92	0.996 84	25.36
8	1.079 1	0.887 7	0.528 90	0.997 05	21.55
10	0.776 0	0.649 3	0.605 34	0.997 32	19.52
12	0.604 2	0.510 9	0.660 79	0.997 58	18.25
14	0.494 0	0.420 8	0.702 72	0.997 80	17.39
16	0.417 6	0.357 6	0.735 50	0.998 00	16.76
18	0.361 5	0.310 8	0.761 80	0.998 16	16.29
20	0.318 6	0.274 8	0.783 37	0.998 30	15.92
22	0.284 8	0.246 3	0.801 37	0.998 42	15.62
24	0.257 4	0.223 1	0.816 61	0.998 53	15.37
26	0.234 8	0.203 9	0.829 69	0.998 62	15.16

通过对比可以看出: 在每一个置信水平和样本容量的组合下,本文计算的最短置信区间要优于等尾置信区间,随着样本容量的增大,2 种置信区间的相对差异逐渐减小,但即使样本容量取到 26,在置信水平 0.95 的情况下相对差异仍有 15.16%. 在样本容量 $n \leqslant 10$ 时,2 者的相对差异更大,大多数达到 20% 以上. 由于给定样本容量,在置信水平 0.95 情况下的 2 种置信区间的相对差异比在置信水平 0.90 的情况下的更小,所以,在小样本情况下最短置信区间优势明显.

事实上,当参数 $m > 1$、$n > 1$ 时,贝塔分布 Beta(m,n) 都是单峰分布. 如 $m > n > 1$ 是左偏单峰分布,$n > m > 1$ 是右偏单峰分布,$n = m > 1$ 是单峰对称分布. 当枢轴量 G 具有如下形式:

$$G = T(X_1, X_2, \cdots, X_n)\,(\theta + U(X_1, X_2, \cdots, X_n))^{-1}$$

且服从贝塔分布时,本文构造了在置信水平 0.90 和 0.95 下基于 Beta(m,n) 的最短区间估计用表,整理在表 3 ~ 表 6 中.

表 3　在置信水平 0.90 下最短置信区间的左侧端点值

n	m							
	5	6	7	8	9	10	11	12
5	0.297 15	0.349 41	0.394 93	0.434 77	0.469 86	0.500 97	0.528 70	0.553 56
6	0.263 47	0.312 81	0.356 47	0.395 22	0.429 77	0.460 72	0.488 57	0.513 74
7	0.236 78	0.283 31	0.325 04	0.362 52	0.396 26	0.426 76	0.454 43	0.479 62
8	0.215 07	0.259 00	0.298 84	0.334 96	0.367 78	0.397 66	0.424 97	0.449 98
9	0.197 05	0.238 59	0.276 62	0.311 39	0.343 22	0.372 41	0.399 23	0.423 94
10	0.181 84	0.221 20	0.257 53	0.290 99	0.321 82	0.350 25	0.376 53	0.400 85
11	0.168 83	0.206 20	0.240 94	0.273 14	0.302 98	0.330 65	0.356 34	0.380 23
12	0.157 57	0.193 13	0.226 38	0.257 38	0.286 26	0.313 16	0.338 25	0.361 67

表 4 在置信水平 0.90 下最短置信区间的右侧端点值

n	m							
	5	6	7	8	9	10	11	12
5	0.856 33	0.867 52	0.877 42	0.886 12	0.893 77	0.900 52	0.906 50	0.911 83
6	0.808 07	0.822 45	0.835 21	0.846 46	0.856 39	0.865 19	0.873 01	0.880 01
7	0.763 21	0.780 17	0.795 27	0.808 65	0.820 52	0.831 09	0.840 53	0.849 00
8	0.722 03	0.741 00	0.757 98	0.773 11	0.786 59	0.798 65	0.809 47	0.819 23
9	0.684 39	0.704 91	0.723 38	0.739 91	0.754 72	0.768 03	0.780 02	0.790 87
10	0.650 05	0.671 74	0.691 35	0.709 01	0.724 90	0.739 23	0.752 20	0.763 98
11	0.618 68	0.641 24	0.661 73	0.680 27	0.697 02	0.712 19	0.725 98	0.738 54
12	0.589 99	0.613 17	0.634 32	0.653 54	0.670 98	0.686 84	0.701 29	0.714 51

表 5 在置信水平 0.95 下最短置信区间的左侧端点值

n	m							
	5	6	7	8	9	10	11	12
5	0.249 47	0.301 00	0.346 71	0.387 29	0.423 42	0.455 73	0.484 73	0.510 90
6	0.220 49	0.268 61	0.311 96	0.350 98	0.386 14	0.417 90	0.446 68	0.472 86
7	0.197 66	0.242 66	0.283 74	0.321 13	0.355 15	0.386 17	0.414 50	0.440 45
8	0.179 17	0.221 37	0.260 31	0.296 09	0.328 94	0.359 10	0.386 85	0.412 42
9	0.163 87	0.203 57	0.240 53	0.274 77	0.306 43	0.335 71	0.362 80	0.387 90
10	0.151 01	0.188 45	0.223 59	0.256 37	0.286 88	0.315 26	0.341 65	0.366 23
11	0.140 04	0.175 45	0.208 91	0.240 33	0.269 73	0.297 21	0.322 90	0.346 93
12	0.130 56	0.164 14	0.196 06	0.226 20	0.254 54	0.281 16	0.306 15	0.329 62

表 6 在置信水平 0.95 下最短置信区间的右侧端点值

n	m							
	5	6	7	8	9	10	11	12
5	0.888 29	0.896 48	0.903 87	0.910 46	0.916 30	0.921 49	0.926 12	0.930 25
6	0.844 38	0.855 52	0.865 57	0.874 51	0.882 47	0.889 54	0.895 86	0.901 53
7	0.802 33	0.816 00	0.828 33	0.839 34	0.849 17	0.857 94	0.865 80	0.872 88
8	0.762 89	0.778 62	0.792 87	0.805 64	0.817 08	0.827 33	0.836 56	0.844 89
9	0.726 25	0.743 65	0.759 46	0.773 70	0.786 51	0.798 03	0.808 44	0.817 87
10	0.692 37	0.711 09	0.728 17	0.743 62	0.757 57	0.770 17	0.781 60	0.791 98
11	0.661 09	0.680 85	0.698 94	0.715 37	0.730 27	0.743 78	0.756 07	0.767 27
12	0.632 22	0.652 78	0.671 67	0.688 89	0.704 56	0.718 83	0.731 85	0.743 76

从表 3 ~ 表 6 可以看出: 在给定置信水平情况下,n 固定,m 越大,最短置信区间的左侧端点值越大,最短置信区间的右侧端点值也越大,但增大幅度比左侧端点值更小. 在给定置信水平情况下,m 固定,n 越大,最短置信区间的左侧端点值越小,最短置信区间的右侧端点值也越小,但减小幅度比左侧端点值更大. m 和 n 都固定,在置信水平更高情况下的左侧端点值更小以及右侧端点值更大,这导致置信区间长度增大,这体现了可靠度与精确度其实是相互制约的.

2.3 Beta(a,1) 型的最短置信区间

前面的分析讨论都是在基于枢轴量的分布为单峰分布的假设下得到的. 但当 $a > 1$ 时,Beta(a,1) 分布的密度函数是单调递增的. 接下来讨论当枢轴量服从 Beta(a,1) 分布时在 2 类枢轴量形式下的最短置信区间估计.

1) 枢轴量 G 具有如下形式:

$G = T(X_1, X_2, \cdots, X_n)(\theta + U(X_1, X_2, \cdots, X_n))$,其中 $T(X_1, X_2, \cdots, X_n) > 0$.

由 $P_\theta(c \leqslant G \leqslant d) = 1 - \alpha$(即 $P_\theta(c \leqslant T(\theta + U) \leqslant d) = 1 - \alpha$) 得到参数 θ 的置信区间为 $[cT^{-1} - U, dT^{-1} - U]$,平均区间长度为 $(d - c)E_\theta(T^{-1})$. 此外,G 的密度函数为

$$g(x) = \begin{cases} ax^{\alpha-1}, 0 < x < 1, \\ 0, \quad \text{其他}, \end{cases}$$

则寻求 c、d 使得

$$\int_c^d g(x)\,\mathrm{d}x = 1 - \alpha. \tag{7}$$

根据密度函数单调递增的特点,对于给定的置信水平,在平均区间长度最短下,应当选取 $g(x)$ 取值较大的部分,即应选取 c_0,使得

$$P(c_0 \leqslant G \leqslant 1) = \int_{c_0}^1 g(x)\,\mathrm{d}x = 1 - \alpha, \tag{8}$$

其中 $c_0 = \alpha^{1/a}$,得到参数 θ 的置信区间为 $[c_0T^{-1} - U, T^{-1} - U]$,平均区间长度为 $(1 - c_0)E_\theta(T^{-1})$. 由于密度函数严格单增,所以观察式(7)和式(8),显然有 $d - c \geqslant 1 - c_0$. 综上所述,此时的最短置信区间为 $[c_0T^{-1} - U, T^{-1} - U]$.

2) 枢轴量 G 具有如下形式:

$G = T(X_1, X_2, \cdots, X_n)(\theta + U(X_1, X_2, \cdots, X_n))^{-1}$,

其中 $T(X_1, X_2, \cdots, X_n) \geqslant 0$.

若 $P_\theta(c \leqslant G \leqslant d) = 1 - \alpha$(即 $P_\theta(c \leqslant T(\theta + U)^{-1} \leqslant d) = 1 - \alpha$)得到参数 θ 的置信区间为 $[d^{-1}T - U, c^{-1}T - U]$,平均区间长度为 $(c^{-1} - d^{-1})E_\theta(T)$. 令

$$\int_0^c g(x)\,\mathrm{d}x = \alpha_1, \int_d^1 g(x)\,\mathrm{d}x = \alpha_2,$$

则 $\alpha_1 + \alpha_2 = \alpha$,且 $c = \alpha_1^{1/a}$, $d = (1 - \alpha_2)^{1/a}$.

若 $P_\theta(c_0 \leqslant G \leqslant 1) = \int_{c_0}^1 g(x)\,\mathrm{d}x = 1 - \alpha$,则 $c_0 = \alpha^{1/a}$,得到参数 θ 的置信区间为 $[T - U, c_0^{-1}T - U]$,平均区间长度为 $(c_0^{-1} - 1)E_\theta(T)$. 注意到

$$d - c \geqslant 1 - c_0, cd = (\alpha_1(1 - \alpha_2))^{1/a} \leqslant \alpha_1^{1/a} \leqslant \alpha^{1/a} = c_0,$$

所以 $(d - c)/(cd) \geqslant (1 - c_0)/c_0$.

综上所述,此时的最短置信区间为 $[T - U, c_0^{-1}T - U]$.

2.4　最短后验置信区间

对于区间估计问题,在上述讨论中把参数看成一个常数,在求置信区间时要构造一个枢轴量,这一点技巧性较强,有时是比较困难的. 并且在理解置信水平和置信区间时也会产生困难,而贝叶斯方法具有处理方便和含义清晰的优点. 贝叶斯统计方法是英国统计学家托马斯·贝叶斯(Thomas Bayes)提出的一种方法,其主要的核心思想是将未知参数看成随机变量. 这使得统计学的区间估计得到了更好的解释. 贝叶斯统计方法已成为现代统计学不可或缺的重要内容,在数理统计、生物统计、医学统计、环境统计、金融统计与精算等领域[11-15]中都有广泛的应用.

若参数 θ 的先验分布为 $\pi(\theta)$,样本分布函数为 $F(x;\theta)$,由贝叶斯定理可得参数 θ 的后验分布 $\pi_*(\theta \mid x)$. 若给定概率 $1 - \alpha$,找到一个区间 $[c,d]$,使得 $P(c \leqslant \theta \leqslant d \mid x) = 1 - \alpha$ 成立,这样求得的区间就是参数 θ 的贝叶斯置信区间,称 $1 - \alpha$ 为置信水平. 注意到,在贝叶斯统计中,把参数 θ 看成是随机变量,直接从后验分布中推导得出置信区间,并且把置信水平 $1 - \alpha$ 很自然地解释为参数落入这一区间的概率.

置信水平和平均区间长度是评价贝叶斯区间估计的 2 个标准,在置信水平给定的情况下,希望平均区间长度越短越好. 接下来,考虑在参数 θ 的后验分布为贝塔分布时的最短后验区间估计.

设 $X_1, X_2, \cdots, X_n$ 是来自负二项分布 $NB(m,\theta)$ 的样本,其分布函数为

$$F_X(x) = C_{m+x-1}^x \theta^m (1 - \theta)^x, x = 0, 1, 2, \cdots,$$

且假设 θ 的先验分布为 $\mathrm{Beta}(a,b)$. 根据贝叶斯定理,简单计算可得 θ 的后验分布为 $\mathrm{Beta}(a^*, b^*)$,其中 $a^* = mn + a$, $b^* = \sum_{i=1}^n X_i + b$.

给定置信水平 $1 - \alpha$,若 $P(c \leqslant \theta \leqslant d \mid x) = F(d) - F(c) = 1 - \alpha$,则得到 θ 的贝叶斯置信区间为 $[c,d]$,区间长度为 $d - c$. 欲求最短置信区间,即为求如下条件极值问题:

$$\begin{cases} \min\ d - c, \\ F(d) - F(c) = 1 - \alpha. \end{cases} \tag{9}$$

运用拉格朗日乘数法,把条件极值问题(9)可转化为如下所示的 2 元方程组求解问题:

$$\begin{cases} f(c) = f(d), \\ F(d) - F(c) = 1 - \alpha, \end{cases}$$

其中 $F(\cdot)$、$f(\cdot)$ 分别表示 $\mathrm{Beta}(a^*, b^*)$ 的分布函数与密度函数.

3　数值例子

假设某企业生产的圆盘直径服从均匀分布 $U(c,d)$,实际统计 14 个该种圆盘直径,数据如下: 8.022 2, 7.965 0, 8.016 0, 8.001 9, 8.047 3, 8.014 9, 8.030 0, 7.995 4, 7.993 2, 8.032 5, 7.958 3, 7.963 3, 7.967 3, 7.989 1.

由定理 1 得到 $d - c$ 的估计值为 0.102 7. 当置信

水平为 0.90 时,经过简单计算,等尾置信区间为 [0.091 4,0.126 5];查表 1 可知,最短置信区间为 [0.089 4,0.119 0]. 同理计算,当置信水平为 0.95 时,等尾置信区间为 [0.090 6,0.134 6],最短置信区间为 [0.089 2,0.126 7]. 以上结果表明:当置信水平为0.90和0.95时,本文方法得到的置信区间长度比等尾置信区间长度更短,且它们的长度之比分别为 0.843 3 和 0.852 3,由此可见这种误差是不可忽略的.

4 总结

通过上述分析可以看出:研究在枢轴量服从贝塔分布时的最短置信区间是十分有必要且有意义的,尤其是当样本容量较小时. 本文介绍的方法在理论上可以计算在更广泛的参数组合下的贝塔分布的最短置信区间.

5 参考文献

[1] 夏乐天,郭宝才,肖艳文. 指数分布参数置信区间的最短化研究 [J]. 河海大学学报(自然科学版),2003,31(3):355-357.

[2] 袁长迎,徐明民. 伽玛分布参数的最短置信区间 [J]. 数理统计与管理,2006,25(4):435-437.

[3] 徐美萍,于健,王若. 威布尔分布中尺度参数的最短区间估计 [J]. 江西师范大学学报(自然科学版),2014,38(3):226-228.

[4] 王秀丽. 均匀分布参数的最短置信区间 [J]. 数学的实践与认识,2008,38(9):57-60.

[5] 薛峰,高尚. 贝塔分布的最短置信区间的粒子群优化算法 [J]. 科学技术与工程,2012,12(17):4061-4064.

[6] 孙鹏哲,闫在在,董洪芹,等. χ^2 分布的最短置信区间 [J]. 内蒙古统计,2009(5):44-45.

[7] 李广正. 基于 F 分布的最短置信区间研究 [J]. 统计与决策,2018(12):18-20.

[8] 刘瑞香. 枢轴量为单峰分布的最短区间估计 [J]. 统计与决策,2011(17):164-165.

[9] 李丽颖,宋立新. 关于参数的最短置信区间的数值计算 [J]. 统计与决策,2016(12):68-69.

[10] 茆诗松,王静龙,濮晓龙. 高等数理统计 [M]. 北京:高等教育出版社,2006.

[11] 徐登可,田瑞琴. 函数型空间自回归模型的贝叶斯估计 [J]. 高校应用数学学报,2022,37(3):323-336.

[12] 章溢,周金亮. 索赔次数的贝叶斯预测与信度近似 [J]. 江西师范大学学报(自然科学版),2021,45(4):353-361.

[13] 王纯杰,罗琳琳,李纯净,等. 删失数据下部分线性模型的贝叶斯 P-样条估计 [J]. 东北师大学报(自然科学版),2020,52(4):25-32.

[14] 张良超,周金亮,温利民. 零膨胀泊松模型中风险参数的贝叶斯估计 [J]. 江西师范大学学报(自然科学版),2020,44(3):269-274.

[15] 刘思杨,蔡艳. 应用 Stan 软件包实现 IRT 模型的贝叶斯参数估计 [J]. 江西师范大学学报(自然科学版),2020,44(3):282-291.

The Study on Optimal Confidence Interval Based on Beta Distribution

WEN Limin[1],ZHANG Liangchao[1],ZHANG Yi[2],LIU Wei[1]

(1. School of Mathematics and Statistics,Jiangxi Normal University,Nanchang Jiangxi 330022,China;
2. School of Finance,Jiangxi Normal University,Nanchang Jiangxi 330022,China)

Abstract: Based on the probability properties of the beta distribution,the optimal interval estimation of a special kind of Beta distribution is considered. Furthermore,the obtained interval estimates are compared with equal-tailed confidence intervals. The results show that using the shortest confidence interval as the interval estimation of unknown parameters,the accuracy of the estimation is significantly improved. Finally,a table for estimating the shortest interval of the beta distribution is given by means of numerical simulation.

Key words: Beta distribution; equal-tail confidence interval; shortest confidence interval

(责任编辑:曾剑锋)