

指数分布参数置信区间的最短化研究

夏乐天,郭宝才,肖艳文

(河海大学理学院,江苏 南京 210098)

摘要 从置信区间的本质意义出发,通过数值计算的方法,对于给定的置信度 0.90,0.95 和 0.99,在样本容量从 2 到 22 的范围内,求得指数分布参数的最短置信区间,并对通常方法求得的置信区间的长度与最短置信区间的长度进行了对比分析.结果表明 在小样本(≤ 11)的情形下,用最短置信区间来作未知参数的区间估计,将会使估计精度得到显著的提高.

关键词 指数分布 参数 置信区间 最短化

中图分类号 O212 **文献标识码** A **文章编号** 1000-1980(2003)03-0355-03

1 问题的提出

设简单随机样本 $\xi_1, \dots, \xi_n$ 来自总体 ξ , ξ 的概率函数为 $f(x; \theta)$, 其中 θ 未知, 若有统计量 $\theta_i = \theta_i(\xi_1, \dots, \xi_n)$, $i = 1, 2$, $\theta_1 < \theta_2$, 使得 $P\{\theta_1 < \theta < \theta_2\} = 1 - \alpha = \gamma$, 则称 (θ_1, θ_2) 为 θ 的置信区间, θ_1 为置信下限, θ_2 为置信上限, γ 为置信度. (θ_1, θ_2) 也称为 θ 的区间估计. 显然, 对于给定的置信度, 置信区间的长度 $L = \theta_2 - \theta_1$ 是越短越好, 它代表区间估计的精度. 现设简单随机样本 $\xi_1, \dots, \xi_n$ 来自参数为 λ 的指数分布总体, 总体的概率密度函数

$$f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

其中 $\lambda > 0$ 未知. 易知, $\eta = 2n\lambda\bar{\xi}$ 服从自由度为 $2n$ 的 χ^2 分布 $\chi^2(2n)$. 对于给定的置信度 $\gamma = 1 - \alpha$, 用通常教科书^[1,2]上介绍的方法可求得 λ 的置信区间为

$$\left(\frac{\chi_{1-\alpha/2}^2(2n)}{2n\bar{\xi}}, \frac{\chi_{\alpha/2}^2(2n)}{2n\bar{\xi}} \right) \tag{1}$$
$$\bar{\xi} = \frac{1}{n} \sum_{i=1}^n \xi_i$$

其中 $\chi_{\alpha/2}^2(2n)$ 和 $\chi_{1-\alpha/2}^2(2n)$ 分别是 $\chi^2(2n)$ 的上、下侧 $\alpha/2$ 分位点.

由于 χ^2 分布的概率密度是偏态的函数, 因此上述按概率对称求得的 λ 的置信区间的长度绝不是最短的, 从而也不是最佳的. 那么, 如何在置信度给定的前提下, 寻找 λ 的最短置信区间呢? 文献[3]曾对正态总体 $N(\mu, \sigma^2)$ 中 σ 的最短置信区间作过简单的讨论, 但是很遗憾, 指数分布总体中 λ 的最短置信区间问题的讨论, 迄今为止一直未在有关文献中发现.

2 λ 的最短置信区间的推求

当 $n \geq 2$ 时, $\chi^2(2n)$ 的概率密度是

$$f(y) = \begin{cases} \frac{1}{2^n \Gamma(n)} y^{n-1} e^{-y/2} & y > 0 \\ 0 & y \leq 0 \end{cases} \tag{2}$$

其概率密度曲线如图 1 所示.

已知, 对于给定的置信度 $\gamma = 1 - \alpha$, 用通常教科书上介绍的方法可求得的 λ 的置信区间如(1)所示. 当

样本观测值 $x_1, \cdots, x_n$ 得到之后,置信区间的长度

$$L = \frac{1}{2n\bar{x}}[\chi^2_{\alpha/2}(2n) - \chi^2_{1-\alpha/2}(2n)]$$
$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$
这样推求 λ 的置信区间的理由是考虑操作比较方便,它的长度完全取决于 $\chi^2(2n)$ 两个按概率对称的 $\alpha/2$ 分位点的差 $\chi^2_{\alpha/2}(2n) - \chi^2_{1-\alpha/2}(2n)$. 从密度曲线的图形可以看出, λ 的长度最短的置信区间应该是 $(\frac{a}{2n\bar{\xi}}, \frac{b}{2n\bar{\xi}})$. 其中 (a, b) 的位置比 $(\chi^2_{1-\alpha/2}(2n), \chi^2_{\alpha/2}(2n))$ 可能要稍微往左偏一些, 因为密度曲线的峰值偏向原点. 由 $\eta = 2n\lambda \bar{\xi} \sim \chi^2(2n)$, 令 $P\{a < \eta < b\} = 1 - \alpha$, 确实可解得 λ 的置信区间为 $(\frac{a}{2n\bar{\xi}}, \frac{b}{2n\bar{\xi}})$, 此时置信区间的长度 $L = \frac{1}{2n\bar{\xi}}(b - a)$.

定理 设简单随机样本 $\xi_1, \cdots, \xi_n$ 来自参数为 λ 的指数分布总体, 则 λ 的长度最短的置信区间 $(\frac{a}{2n\bar{\xi}}, \frac{b}{2n\bar{\xi}})$ 中的 a, b 应满足方程:

$$a^{n-1}e^{-a/2} - b^{n-1}e^{-b/2} = 0 \tag{3}$$

证 因为 $P\{a < \eta < b\} = \int_a^b f(y) dy = 1 - \alpha$ (4)
所以 b 可看作是 a 的可微函数. 对式(4)的两端关于 a 求导, 得

$$f(b) \frac{db}{da} - f(a) = 0$$
$$\frac{db}{da} = \frac{f(a)}{f(b)} = \frac{a^{n-1}e^{-a/2}}{b^{n-1}e^{-b/2}}$$

从而有 $\frac{dL}{da} = 0$ 即得式(3).

3 计算结果

联立方程(3)和(4),用数值计算中的矩形法,对样本容量 $n \in \{2, 3, \cdots, 22\}$,从初值 $b_0(n) = \chi^2_{\alpha/2}(2n)$ 开始,以步长 10^{-5} 向左搜索至 $\chi^2_{\alpha}(2n)$, 可求得 λ 的最短置信区间中的 a, b , 见表 1(精度控制在 10^{-3} 之内).

表 1 指数分布参数 λ 的最短置信区间中的 a, b

Table 1 a and b for minimum confidence interval of exponentially distributed parameter λ

γ				γ				γ			
n	0.90	0.95	0.99	n	0.90	0.95	0.99	n	0.90	0.95	0.99
2	0.168	0.085	0.017	9	8.343	7.245	5.404	16	18.943	17.206	14.138
	7.864	9.530	13.286		27.326	29.955	35.541		44.697	47.958	54.763
3	0.883	0.607	0.264	10	9.786	8.584	6.544	17	20.529	18.713	15.493
	10.958	12.803	16.902		29.876	32.607	38.390		47.109	50.448	57.404
4	0.875	1.425	0.786	11	11.259	9.958	7.729	18	22.127	20.235	16.868
	13.892	15.897	20.296		32.398	35.227	41.197		49.509	52.924	60.026
5	3.017	2.414	1.498	12	12.756	11.361	8.951	19	23.737	21.771	18.261
	16.711	18.861	23.533		34.895	37.818	43.968		51.898	55.386	62.631
6	4.258	3.516	2.344	13	14.276	12.791	10.207	20	25.357	23.319	19.670
	19.446	21.729	26.654		37.372	40.384	46.706		54.276	57.836	65.220
7	5.570	4.700	3.291	14	15.816	14.243	11.492	21	26.987	24.879	21.093
	22.119	24.525	29.684		39.830	42.927	49.416		56.64	60.275	67.794
8	6.935	5.948	4.316	15	17.372	15.715	12.803	22	28.626	26.449	22.531
	24.742	27.263	32.642		42.271	45.451	52.101		59.005	62.703	70.354

说明 对于 n 的每一个值,第一行为 a ,第二行为 b .

比较 λ 的最短置信区间的长度与按概率对称求得的置信区间的长度,可得到这两个长度的绝对误差和相对误差见表 2. 表 2 中 $L_0 = b - a$ 表示 λ 的最短置信区间长度的 $2n\bar{x}$ 倍, $L = \chi^2_{\alpha/2}(2n) - \chi^2_{1-\alpha/2}(2n)$ 表示 λ 的按概率对称求得的置信区间长度的 $2n\bar{x}$ 倍,它们分别代表对应置信区间的长度; $e = (L - L_0)/L_0$ 代表后者的相对误差,表示用 L_0 代替 L 可使 λ 的置信区间的精度提高的百分数.

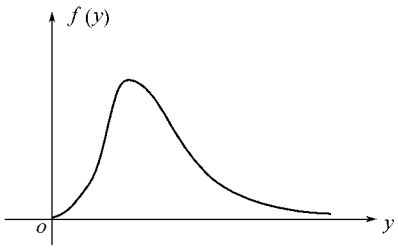


图 1 $\chi^2(2n)$ 的概率密度曲线

Fig.1 Probability density function of $\chi^2(2n)$

表 2 λ 的最短置信区间的精度分析

Table 2 Precision analysis of the minimum confidence interval of parameter λ

n	γ = 0.90				γ = 0.95				γ = 0.99			
	L ₀	L	L - L ₀	e/%	L ₀	L	L - L ₀	e/%	L ₀	L	L - L ₀	e/%
2	7.697	8.755	1.078	14.01	9.446	10.659	1.213	12.85	13.268	14.653	1.385	10.44
3	10.076	10.957	0.881	8.75	12.196	13.212	1.016	8.33	16.638	17.872	1.234	7.42
4	12.081	12.774	0.756	6.29	14.472	15.355	0.883	6.10	19.510	20.611	1.101	5.64
5	13.694	14.367	0.673	4.92	16.447	17.236	0.789	4.80	22.035	23.032	0.997	4.52
6	15.188	15.800	0.612	4.03	18.213	18.933	0.720	3.95	24.309	25.225	0.916	3.77
7	16.549	17.114	0.565	3.41	19.824	20.490	0.666	3.36	26.393	27.244	0.851	3.23
8	17.808	18.334	0.526	2.96	21.315	21.937	0.622	2.92	28.326	29.125	0.799	2.82
9	18.983	19.479	0.496	2.61	22.709	23.295	0.586	2.58	30.137	30.891	0.754	2.50
10	20.090	20.559	0.469	2.33	24.023	24.579	0.556	2.31	31.846	32.563	0.717	2.25
11	21.139	21.586	0.447	2.11	25.269	25.799	0.530	2.10	33.468	34.153	0.685	2.05
12	22.139	22.567	0.428	1.93	26.456	26.963	0.507	1.92	35.016	35.673	0.657	1.88
13	23.096	23.506	0.410	1.78	27.593	28.079	0.486	1.76	36.499	37.130	0.631	1.73
14	24.014	24.409	0.359	1.64	28.684	29.153	0.469	1.63	37.924	38.532	0.608	1.60
15	24.899	25.280	0.381	1.53	29.736	30.188	0.452	1.52	39.298	39.885	0.587	1.49
16	25.753	26.122	0.369	1.43	30.752	31.189	0.437	1.42	40.625	41.194	0.569	1.40
17	26.580	26.938	0.358	1.35	31.735	32.160	0.425	1.34	41.910	42.463	0.553	1.32
18	27.382	27.729	0.347	1.27	32.689	33.101	0.412	1.26	43.157	43.694	0.537	1.24
19	28.161	28.500	0.339	1.20	33.616	34.081	0.402	1.20	44.370	44.892	0.552	1.18
20	28.920	29.249	0.329	1.14	34.517	34.909	0.392	1.13	45.550	46.059	0.509	1.12
21	29.659	29.980	0.321	1.08	35.396	35.778	0.382	1.08	46.700	47.198	0.498	1.07
22	30.380	30.694	1.04	1.04	36.254	36.626	0.372	1.03	47.823	48.309	0.486	1.02

4 结 语

表 1 中的数据可供实际查用,具有非常现实的意义.表 2 表明:用表 1 中的数据作指数分布总体参数 λ 的区间估计,当容量 n 不超过 11 时,置信区间的长度有较明显的缩短,幅度在 2% ~ 10% ;容量 n 从 12 ~ 22 时,置信区间长度缩短的幅度在 1% ~ 2% .置信区间缩短的幅度随着容量 n 增大呈递减趋势,当容量 n > 22 时,研究 λ 的置信区间的最短化已意义不大.事实上,根据 Fisher 定理,对于 $\eta = 2n\lambda\bar{\xi} \sim \chi^2(2n)$, $\sqrt{2\eta}$ 的极限分布是正态分布 $N(\sqrt{4n-1},1)^{[4]}$,它是对称分布.所以,当 n 增大时,两种方法算得的置信区间的长度的差异不断地缩小是自然的.在区间估计的小样本理论中,凡是抽样分布的总体不是单峰对称的,则都可以进行未知参数的最短置信区间的研究.

参考文献:

[1] 陈希孺.概率论与数理统计[M].合肥:中国科技大学出版社,1992.195—204.
[2] BLOM Cunnar.Probability and statistics theory and applications[M].NY:Springer,1989.235—241.
[3] ROBERT V H,ELLIOT A T.Probability and statistical inference[M].3rd ed.NY:Macmillan Pub,1989.209—213.
[4] 印凡成,夏乐天.概率论与数理统计[M].南京:河海大学出版社,2000.156—157,190—191.

Study on minimum confidence interval for exponentially distributed parameters

XIA Le-tian, GUO Bao-cai, XIAO Yan-wen

(College of Sciences, Hohai Univ., Nanjing 210098, China)

Abstract: Based on the definition of the confidence interval and the numerical calculation method, the minimum confidence interval for exponentially distributed parameters, with a sample size from 2 to 22, is obtained at the degree of confidence $\gamma = 0.90, 0.95$ and 0.99 . A comparison of the length of the minimum confidence interval with that of the confidence interval calculated with conventional methods shows that in the case of small sample size ($n \leq 11$), the adoption of the minimum confidence interval can obviously improve the precision of interval estimation for unknown parameters.

Key words: exponential distribution; parameter; confidence interval; minimum length