

文章编号: 1000-8055(2007)07-1025-05

不完全区间数据最佳线性无偏估计方法

傅惠民, 敖 亮

(北京航空航天大学 小样本技术研究中心, 北京 100083)

摘 要: 建立了不完全区间数据秩统计量的概率密度函数和联合概率密度函数, 给出了其平均秩、中位秩和百分位秩以及不完全区间数据统计量的均值、方差和协方差计算公式. 在此基础上, 针对位置-尺度分布族提出一种不完全区间数据最佳线性无偏估计方法, 得到了其百分位值的置信限估计, 并对工程中常见的极值分布、两参数 Weibull 分布、正态分布进行了详细讨论.

关 键 词: 航空、航天推进系统; 不完全区间数据; 秩统计量; 概率密度函数; 最佳线性无偏估计; 均值; 方差; 协方差

中图分类号: TB301; O211 文献标识码: A

Best linear unbiased estimate method for incomplete interval censored data

FU Huimin, AO Liang

(Research Center of Small Sample Technology,

Beijing University of Aeronautics and Astronautics, Beijing 100083, China)

Abstract: The probability density function and joint probability density function of rank statistics for incomplete interval censored data are established, including the formulas of mean ranks, median ranks and percent ranks of rank statistics; furthermore, the means, variances and covariances of statistics for incomplete interval censored data are given in the paper. According to this, the best linear unbiased estimate method of location-scale distribution for incomplete interval censored data is presented, the confidence limits of percentiles for the incomplete interval censored data are obtained, and extreme value distribution, two-parameter Weibull distribution, normal distribution are discussed in detail.

Key words: aerospace propulsion system; incomplete interval censored data; rank statistic; probability density function; best linear unbiased estimate; mean; variance; covariance

最佳线性无偏估计方法不仅计算简单, 而且得到的估计量性质优良, 在工程中有着广泛的应用, 然而传统的最佳线性无偏估计方法只适用于完全数据和截尾数据, 对区间数据、不完全数据和不完全区间数据却无法进行处理. 为此, 文献[1]讨论了区间统计量及其分布理论, 建立了区间数据最佳线性无偏估计方法; 文献[2]建立了不完全

数据的秩分布, 给出平均秩和百分位秩公式; 文献[3]给出了不完全数据秩统计量的联合概率密度函数, 建立了不完全数据最佳线性无偏估计方法. 在上述工作的基础上, 本文建立了不完全区间数据统计量的概率密度函数和联合概率密度函数, 给出了其平均秩、中位秩和百分位秩公式, 提出了不完全区间数据最佳线性无偏估计方法.

收稿日期: 2007-01-04; 修订日期: 2007-07-02

基金项目: 国家自然科学基金项目(10472006); 北京市教育委员会共建项目(XK100060418)

作者简介: 傅惠民(1956-), 男, 浙江遂昌人, 北京航空航天大学航空科学与工程学院“长江学者”特聘教授, 博士, 从事小样本信息技术、软校准技术、数据融合方法及可靠性研究.

1 不完全区间数据秩统计量的分布

1.1 不完全区间数据秩统计量

设来自分布 $F(t)$ 的一组不完全区间数据具有如下形式

$$(t_1, n_1', m_1'), (t_2, n_2', m_2'), \dots, (t_q, n_q', m_q') \quad (1)$$

式(1)表示位于 (t_{j-1}, t_j) 内有 n_j' 个完全数据, t_j 处有 m_j' 个中止数据, $j = 1, 2, \dots, q$, 其中 $t_0 = -\infty$, $t_1 < t_2 < \dots < t_q$, 数据总量为 n . 则有

$$\sum_{j=1}^q (n_j' + m_j') = n \quad (2)$$

令 $n_i = \sum_{j=1}^i n_j'$, $m_i = \sum_{j=1}^i m_j'$ ($i = 1, 2, \dots, q$), $n_{-1} = n_0 =$

$$h_{ir} = \frac{(n - n_i - r)!}{(n - n_i - m_{i-1})!} \prod_{s=1}^{i-1} \frac{(n + 1 - n_s - m_s)!}{(n + 1 - n_s - m_{s-1})!} \times$$

$$\sum_{r_1+r_2+\dots+r_{i-1}=r} \left[\prod_{s=1}^{i-1} \frac{(m_s - m_{s-1})! (n_i - n_s + \sum_{l=s}^{i-1} r_l)!}{r_s! (m_s - m_{s-1} - r_s)! (n_i - n_s + \sum_{l=s}^{i-1} r_l - r_s)!} \right] \quad (5)$$

式中 h_{ir} 表示如果 t_i 前的中止数据继续试验的话, 有且仅有 r 个中止数据在 t_i 前失效这一事件 A_{ir} 发生的概率, 其中 $h_{i0} = 1$.

1.3 不完全区间数据秩统计量的联合概率密度函数

根据条件概率公式可以证明, 第 i_1 个秩统计

$$g_{r_1 r_2}(P_{i_1}^*, P_{i_2}^*) = \frac{(n + 2)!}{(n_{i_1} + r_1)! (n_{i_2} + r_2 - n_{i_1})! (n - n_{i_2} - r_2 - r_1)!} \times (P_{i_1}^*)^{n_{i_1} + r_1} (P_{i_2}^* - P_{i_1}^*)^{n_{i_2} + r_2 - n_{i_1}} (1 - P_{i_2}^*)^{n - n_{i_2} - r_2 - r_1} \quad (7)$$

式中 $h_{i_1 r_1}$ 是事件 $A_{i_1 r_1}$ 发生的概率, 可由式(5)计算; $h_{i_2 r_2}^*$ 是在事件 $A_{i_1 r_1}$ 发生的情况下, 事件 $A_{i_2 r_2}^*$ 发生的条件概率, 它相当于计算不完全区间数据

$$\begin{cases} n_i^* = m_i^* = 0 & i < i_1 \\ n_i^* = n_i + r_1 & i \geq i_1 \\ m_i^* = m_i - r_1 & i \geq i_1 \end{cases} \quad (8)$$

的事件 A_{ir}^* 发生的概率, 仍可由式(5)求得.

2 不完全区间数据秩统计量的平均秩、中位秩和百分位秩

2.1 不完全区间数据秩统计量的平均秩

根据式(3), 可以得到第 i 个秩统计量 P_i^* 的均值(平均秩)为

$$\bar{P}_i^* = E(P_i^*) = \sum_{r=0}^{m_{i-1}} h_{ir} \int_0^{P_i^*} P_i^* g_r(P_i^*) dP_i^* \quad (9)$$

$m_{-1} = m_0 = 0$. 由文献[1]知, 可将 t_i 看作第 i 个不完全区间数据秩统计量 T_i^* 的一个实现值. 设 $P_i^* = F(T_i^*)$, 则称 P_i^* 为第 i 个不完全区间数据秩统计量.

1.2 不完全区间数据秩统计量的概率密度函数

由文献[2]可以证明第 i ($i = 1, 2, \dots, q$) 个不完全区间数据秩统计量的概率密度函数为

$$g(P_i^*) = \sum_{r=0}^{m_{i-1}} h_{ir} g_r(P_i^*) \quad (3)$$

式中

$$g_r(P_i^*) = \frac{(n + 1)!}{(n_i + r)! (n - n_i - r)!} \times (P_i^*)^{n_i + r} (1 - P_i^*)^{n - n_i - r} \quad (4)$$

量 $P_{i_1}^*$ 和第 i_2 个秩统计量 $P_{i_2}^*$ 的联合概率密度函数可由下式给出^[3]

$$g(P_{i_1}^*, P_{i_2}^*) = \sum_{r_1=0}^{m_{i_1-1}} h_{i_1 r_1} \sum_{r_2=0}^{m_{i_2-1} - r_1} h_{i_2 r_2}^* g_{r_1 r_2}(P_{i_1}^*, P_{i_2}^*) \quad 1 \leq i_1 < i_2 \leq q \quad (6)$$

式中

$$\bar{P}_i^* = \frac{n_i + 1}{n + 2} + \frac{1}{n + 2} \sum_{r=0}^{m_{i-1}} r h_{ir} \quad (10)$$

2.2 不完全区间数据秩统计量的中位秩和百分位秩

第 i 个秩统计量 P_i^* 的中位秩 P_{50i}^* 由下式计算

$$\sum_{r=0}^{m_{i-1}} h_{ir} \int_0^{P_{50i}^*} g_r(P_i^*) dP_i^* = \frac{1}{2} \quad (11)$$

第 i 个秩统计量 P_i^* 的置信度为 γ 的百分位秩 P_{γ}^* 由下式计算

$$\sum_{r=0}^{m_{i-1}} h_{ir} \int_0^{P_{\gamma}^*} g_r(P_i^*) dP_i^* = \gamma \quad (12)$$

3 不完全区间数据最佳线性无偏估计

3.1 位置-尺度分布族

设 T 为位置-尺度分布族随机变量, 即

$$T \sim F\left(\frac{t-\mu}{\sigma}\right), \quad -\infty < t < +\infty, \\ -\infty < \mu < +\infty, \quad \sigma > 0 \quad (13)$$

式中 μ 和 σ 为位置参数和尺度参数. 令 $T^{\circ} = (T - \mu) \backslash \sigma$, 则 T° 被称为标准位置-尺度分布族随机变量, 其分布函数 $F(t^{\circ})$ 和概率密度函数 $f(t^{\circ})$ 不含任何未知参数. 正态分布、两参数 Weibull 分布和极值分布等均属于位置-尺度分布族.

3.2 不完全区间数据统计量的均值、方差和协方差

设 $(t_1, n_1, m_1), (t_2, n_2, m_2), \dots, (t_q, n_q, m_q)$ 是来自位置-尺度分布族 $F\left(\frac{t-\mu}{\sigma}\right)$ 的一组不完全区间数据, 则可将 t_i 看作第 i 个不完全区间数据统计量 T_i^* 的一个实现值. 通过 $T_i^{\circ} = (T_i^* - \mu) \backslash \sigma$ 可以将 T_i^* 变换为来自标准位置-尺度分布 $F(t^{\circ})$ 的不完全区间数据统计量 T_i° .

根据式(3)可以得到第 i 个标准不完全区间数据统计量 T_i° 的均值为

$$\mu = E(T_i^{\circ}) = \sum_{r=0}^{m_i-1} h_r \int_0^1 F^{-1}(P_i^*) g_r(P_i^*) dP_i^* \quad (14)$$

第 i 个标准不完全区间数据统计量 T_i° 的方差由下式计算

$$v_{ii} = \text{Var}(T_i^{\circ}) = E[(T_i^{\circ})^2] - \mu^2 \quad (15)$$

式中

$$E[(T_i^{\circ})^2] = \sum_{r=0}^{m_i-1} h_r \int_0^1 [F^{-1}(P_i^*)]^2 g_r(P_i^*) dP_i^* \quad (16)$$

第 i_1 个标准不完全区间数据统计量 $T_{i_1}^{\circ}$ 和第 i_2 个标准不完全区间数据统计量 $T_{i_2}^{\circ}$ 的协方差为

$$v_{i_1 i_2} = \text{Cov}(T_{i_1}^{\circ}, T_{i_2}^{\circ}) = E(T_{i_1}^{\circ} T_{i_2}^{\circ}) - \mu_{i_1} \mu_{i_2} \\ i_1, i_2 = 1, 2, \dots, q \text{ 且 } i_1 < i_2 \quad (17)$$

式中

$$E(T_{i_1}^{\circ} T_{i_2}^{\circ}) = \sum_{r_1=0}^{m_{i_1}-1} h_{i_1 r_1} \sum_{r_2=0}^{m_{i_2}-1} h_{i_2 r_2}^* \times \\ \int_0^1 \int_0^1 F^{-1}(P_{i_1}^*) F^{-1}(P_{i_2}^*) g_{r_1 r_2}(P_{i_1}^*, P_{i_2}^*) dP_{i_1}^* dP_{i_2}^* \quad (18)$$

3.3 位置-尺度分布最佳线性无偏估计

由于

$$t_i = \mu + \sigma u_i^{\circ} \quad i = 1, 2, \dots, q \quad (19)$$

所以

$$t_i = \mu + \sigma u_i + \varepsilon \quad i = 1, 2, \dots, q \quad (20)$$

式中 $\varepsilon = \sigma(u_i^{\circ} - u_i)$, 并且 $E(\varepsilon) = 0$ 和 $\text{Cov}(\varepsilon_1, \varepsilon_2) = \sigma^2 v_{i_1 i_2}$. 设 $V = (v_{i_1 i_2})_{q \times q}$ 和 $V^{-1} = (v^{i_1 i_2})_{q \times q}$, 根据加权最小二乘估计法, 令

$$Q = \sum_{i_1, i_2=1}^q (t_{i_1} - \mu - \sigma u_{i_1}) v^{i_1 i_2} (t_{i_2} - \mu - \sigma u_{i_2}) \quad (21)$$

由 $\frac{\partial Q}{\partial \mu} = \frac{\partial Q}{\partial \sigma} = 0$ 求得位置参数 μ 和尺度参数 σ 的估计量为

$$\hat{\mu} = \sum_{i=1}^q D_i t_i \quad (22)$$

$$\hat{\sigma} = \sum_{i=1}^q C_i t_i \quad (23)$$

式中

$$D_i = \sum_{i_2=1}^q A v^{i i_2} + \sum_{i_2=1}^q B u_{i_2} v^{i i_2} \quad (24)$$

$$C_i = \sum_{i_2=1}^q B v^{i i_2} + \sum_{i_2=1}^q l u_{i_2} v^{i i_2} \quad (25)$$

式中

$$A = \frac{1}{\Delta} \sum_{i_2=1}^q \sum_{i_1=1}^q u_{i_1} u_{i_2} v^{i_1 i_2} \quad (26)$$

$$B = -\frac{1}{\Delta} \sum_{i_2=1}^q \sum_{i_1=1}^q u_{i_1} v^{i_1 i_2} \quad (27)$$

$$l = \frac{1}{\Delta} \sum_{i_2=1}^q \sum_{i_1=1}^q v^{i_1 i_2} \quad (28)$$

$$\Delta = \left(\sum_{i_2=1}^q \sum_{i_1=1}^q v^{i_1 i_2} \right) \left(\sum_{i_2=1}^q \sum_{i_1=1}^q u_{i_1} u_{i_2} v^{i_1 i_2} \right) - \\ \left(\sum_{i_2=1}^q \sum_{i_1=1}^q u_{i_1} v^{i_1 i_2} \right)^2 \quad (29)$$

根据高斯-马尔科夫定理可知, 估计量 $\hat{\mu}$ 和 $\hat{\sigma}$ 具有无偏且方差最小的性质, 即最佳线性无偏估计. $\hat{\mu}$ 和 $\hat{\sigma}$ 的方差和协方差为

$$\text{Var}(\hat{\mu}) = A \sigma^2 \quad (30)$$

$$\text{Var}(\hat{\sigma}) = l \sigma^2 \quad (31)$$

$$\text{Cov}(\hat{\mu}, \hat{\sigma}) = B \sigma^2 \quad (32)$$

3.4 位置-尺度分布置信限估计

式(22)和式(23)给出的 $\hat{\mu}$ 和 $\hat{\sigma}$ 具有渐近正态分布的性质, 大量 Monte Carlo 模拟也证明了这一点. 所以, 当 q 稍大时, $\hat{\mu}$ 和 $\hat{\sigma}$ 近似服从正态分布. 此时, 位置参数 μ 和尺度参数 σ 置信度为 γ 的单侧置信下限 μ_L, σ_L 和单侧置信上限 μ_U, σ_U 分别由下面各式给出

$$\mu_L = \hat{\mu} + \frac{\sigma}{1 - u_L^2} [u_L^2 B - u_L \sqrt{u_L^2 (B^2 - A l) + A}] \quad (33)$$

$$\mu_U = \mu + \frac{\sigma}{1 - u_{\sqrt{l}}^2} [u_{\sqrt{l}}^2 B + u_{\sqrt{l}} \sqrt{u_{\sqrt{l}}^2 (B^2 - Al) + A}] \quad (34)$$

$$\sigma_L = \frac{\sigma}{1 + u_{\sqrt{l}} \sqrt{l}} \quad (35)$$

$$\alpha_U = \frac{\sigma}{1 - u_{\sqrt{l}} \sqrt{l}} \quad (36)$$

式中 μ 和 σ 分别为位置参数 μ 和尺度参数 σ 的最佳线性无偏估计量, $u_{\sqrt{l}}$ 为标准正态偏量, A 、 B 、 l 由式(26)~(28)给出。

随机变量 T 概率为 P 的百分位值 $t_P = \mu + F^{-1}(P)\sigma$ 的无偏估计量由下式给出

$$t_P = \mu + F^{-1}(P)\sigma \quad (37)$$

其置信水平为 γ 的单侧置信下、上限分别为

$$t_{PL} = \mu + \frac{\sigma}{1 - u_{\sqrt{l}}^2} [F^{-1}(P) + u_{\sqrt{l}}^2 B - k_P] \quad (38)$$

$$t_{PU} = \mu + \frac{\sigma}{1 - u_{\sqrt{l}}^2} [F^{-1}(P) + u_{\sqrt{l}}^2 B + k_P] \quad (39)$$

式中

$$k_P = u_{\sqrt{l}} \left\{ u_{\sqrt{l}}^2 (B^2 - Al) + A + [F^{-1}(P)]^2 l + 2F^{-1}(P)B \right\}^{1/2} \quad (40)$$

$$v_{i_1 i_2} = \text{Cov}(T_{i_1}^o, T_{i_2}^o) = \sum_{r_1=0}^{m_{i_1}-1} h_{i_1 r_1} \sum_{r_2=0}^{m_{i_2}-1} h_{i_2 r_2} \int_0^{P_{i_1}^*} \int_0^{P_{i_2}^*} \ln \ln \frac{1}{1 - P_{i_1}^*} \ln \ln \frac{1}{1 - P_{i_2}^*} g_{r_1 r_2}(P_{i_1}^*, P_{i_2}^*) dP_{i_1}^* dP_{i_2}^* - u_{i_1} u_{i_2} \quad (41)$$

$i_1, i_2 = 1, 2, \dots, q \text{ 且 } i_1 < i_2$

首先, 根据式(43)~(45) 计算求得 u_i 和 $v_{i_1 i_2}$, $i_1, i_2 = 1, 2, \dots, q$, 然后, 由式(24)~(25) 计算 D_i 和 C_i , 并由式(22)和(23) 得到极值分布参数 μ 和 σ 的最佳线性无偏估计 $\hat{\mu}$ 和 $\hat{\sigma}$ 。

4.2 Weibull 分布不完全区间数据最佳线性无偏估计

对于两参数 Weibull 分布

$$F(t) = 1 - \exp\left[-\left(\frac{t}{\beta}\right)^{\alpha}\right] \quad (46)$$

令 $x = \ln t$, 则式(46) 变换为极值分布

$$G(x) = 1 - \exp\left[-\exp\left(\frac{x - \mu}{\sigma}\right)\right] \quad (47)$$

式中 $\mu = \ln \beta$ 和 $\sigma = 1/\alpha$ 所以, 可以首先采用 4.1 节

$$u_i = E(T_i^o) = \sum_{r=0}^{m_i-1} h_r \int_0^1 \Phi^{-1}(P_i^*) g_r(P_i^*) dP_i^* \quad i = 1, 2, \dots, q \quad (50)$$

$$v_{ii} = \text{Var}(T_i^o) = \sum_{r=0}^{m_i-1} h_r \int_0^1 [\Phi^{-1}(P_i^*)]^2 g_r(P_i^*) dP_i^* - u_i^2 \quad i = 1, 2, \dots, q \quad (51)$$

$$v_{i_1 i_2} = \text{Cov}(T_{i_1}^o, T_{i_2}^o) = \sum_{r_1=0}^{m_{i_1}-1} h_{i_1 r_1} \sum_{r_2=0}^{m_{i_2}-1} h_{i_2 r_2} \int_0^{P_{i_1}^*} \int_0^{P_{i_2}^*} \Phi^{-1}(P_{i_1}^*) \Phi^{-1}(P_{i_2}^*) g_{r_1 r_2}(P_{i_1}^*, P_{i_2}^*) dP_{i_1}^* dP_{i_2}^* - u_{i_1} u_{i_2} \quad (52)$$

$i_1, i_2 = 1, 2, \dots, q \text{ 且 } i_1 < i_2$

首先, 根据式(50)~(52) 计算求得 u_i 和 $v_{i_1 i_2}$, $i_1, i_2 = 1, 2, \dots, q$, 然后, 由式(24)~(25) 计算 D_i 和

4 常用分布不完全区间数据最佳线性无偏估计

4.1 极值分布不完全区间数据最佳线性无偏估计

设极值分布的分布函数为

$$F\left(\frac{t - \mu}{\sigma}\right) = 1 - \exp\left[-\exp\left(\frac{t - \mu}{\sigma}\right)\right] \quad (41)$$

令 $t^o = \frac{t - \mu}{\sigma}$, 则可以得到标准极值分布函数

$$F(t^o) = 1 - \exp[-\exp(t^o)] \quad (42)$$

由式(14)~(18) 可知, 标准极值分布不完全区间数据统计量的均值、方差和协方差为

$$u_i = E(T_i^o) = \sum_{r=0}^{m_i-1} h_{ir} \int_0^1 \ln \ln \frac{1}{1 - P_i^*} g_r(P_i^*) dP_i^* \quad i = 1, 2, \dots, q \quad (43)$$

$$v_{ii} = \text{Var}(T_i^o) = \sum_{r=0}^{m_i-1} h_{ir} \int_0^1 \left[\ln \ln \frac{1}{1 - P_i^*} \right]^2 g_r(P_i^*) dP_i^* - u_i^2 \quad i = 1, 2, \dots, q \quad (44)$$

方法得到 μ 和 σ 的最佳线性无偏估计 $\hat{\mu}$ 和 $\hat{\sigma}$, 然后再计算 α 和 β 的估计值 $\hat{\alpha} = 1/\hat{\sigma}$ 和 $\hat{\beta} = \exp(\hat{\mu})$ 。

4.3 正态分布不完全区间数据最佳线性无偏估计

对于正态分布

$$F(t) = \Phi\left(\frac{t - \mu}{\sigma}\right) \quad (48)$$

令 $t^o = \frac{t - \mu}{\sigma}$, 则得标准正态分布函数

$$\Phi(t^o) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^0 \exp\left[-\frac{u^2}{2}\right] du \quad (49)$$

由式(14)~(18) 可知, 标准正态分布不完全区间数据统计量的均值、方差和协方差为

C_i , 并由式(22)和(23)得到正态分布均值 μ 和标准差 σ 的最佳线性无偏估计 $\hat{\mu}$ 和 $\hat{\sigma}$.

5 结 论

- (1) 建立了不完全区间数据秩统计量的概率密度函数和联合概率密度函数.
- (2) 给出了不完全区间数据秩统计量的平均秩、中位秩和百分位秩公式.
- (3) 建立了不完全区间数据统计量的均值、方差和协方差计算公式.
- (4) 提出了不完全区间数据最佳线性无偏估计方法, 在此基础上给出了位置- 尺度分布族百分位值的置信限估计公式.
- (5) 对工程中常用的极值分布、两参数 Weibull 分布、正态分布不完全区间数据最佳线性

无偏估计进行了详细讨论.

参考文献:

[1] 傅惠民. 区间统计量及其分布[J]. 机械强度, 2005, 27(6): 752-757.
FU Huimin. Interval statistic and its distribution[J]. Journal of Mechanical Strength, 2005, 27(6): 752-757.

[2] 傅惠民. 不完全数据秩分布理论[J]. 航空学报, 1993, 14(11): A578-584.
FU Huimin. Theory of incomplete data rank distributions [J]. Journal of Acta Aeronautica et Astronautica Sinica, 1993, 14(11): A578-584.

[3] 傅惠民, 王华胜. 不完全数据最佳线性无偏估计方法[J]. 航空动力学报, 2004, 19(2): 169-173.
FU Huimin, WANG Huasheng. Best linear unbiased estimate method for incomplete data [J]. Journal of Aerospace Power, 2004, 19(2): 169-173.