Statistics and Application 统计学与应用, 2024, 13(2), 259-264
Published Online April 2024 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2024.132026

Hans汉斯

一类不完全数据下指数分布的参数估计

宋　翌，文　婷，何家洪

北部湾大学理学院，广西　钦州

收稿日期：2024年1月23日；录用日期：2024年3月27日；发布日期：2024年4月2日

摘　要

本文针对指数分布，依据极大似然估计方法和贝叶斯估计方法，研究了在右删失数据下指数分布密度函数中参数λ的估计问题，并通过实验数据比较了两种方法的估计效果。研究表明在样本容量相同的情况下，贝叶斯估计方法要比极大似然估计方法偏差更小，可信度更高。

关键词

右删失数据，参数估计，指数分布

Parameter Estimation of Exponential Distribution under a Class of Incomplete Data

Yi Song, Ting Wen, Jiahong He

School of Science, Beibu Gulf University, Qinzhou Guangxi

Received: Jan. 23rd, 2024; accepted: Mar. 27th, 2024; published: Apr. 2nd, 2024

Abstract

This paper studies the estimation problem of parameter λ in the exponential distribution density function under right censored data based on the maximum likelihood estimation method and Bayes estimation method and compares the estimation efficiency of the two methods through experimental data. The results show that under the same sample size, the Bayes estimation method has a smaller deviation and higher reliability than the maximum likelihood estimation method.

文章引用: 宋翌, 文婷, 何家洪. 一类不完全数据下指数分布的参数估计[J]. 统计学与应用, 2024, 13(2): 259-264.
DOI: 10.12677/sa.2024.132026

Keywords

Right Censored Data, Parameter Estimation, Exponential Distribution

1. 引言

在实际问题当中，我们所获取的数据不一定是完整的，它会由于在收集或是保存过程中部分因素导致所获得的数据不完整。比如说对于生存分析，在我们的研究过程当中，相关人员若可以明确的记录每个研究对象发生某特定终点事件的具体时间，将这种数据称之为完全数据，与完全数据相反的是如果科研人员不能观测记录到研究对象发生终点事件的具体时间，则把这种数据称为缺失数据。这种缺失数据在一定意义上来说是不可避免会发生的，像一些数据收集方法不合理的行业，如医药行业，它缺失的数据高达百分之六十以上，缺失数据的存在意味着我们所获得信息就不完整。这就使得科学研究人员需要寻找新的方法来解决统计分析中的参数估计问题。

我国学者对于右删失数据下参数估计问题有着广泛的研究。周旭等探讨了在截断右删失数据下伽马分析的参数推断问题[1]；侯兰宝等人研究了定时区间删失下指数分布的参数估计[2]；董小刚等研究了部分区间删失数据下广义指数分布的参数估计及应用[3]；刘长林等进一步研究了无失效数据场合指数分布可靠度的 Bayes 估计[4]；刘慧馨研究了右删失数据下多响应 ATF 模型的两阶段估计[5]等。

通过查找大量的文献资料发现对处理右删失数据学者们主要采用的方法为随机加权法及多重插补法等，本文主要研究了在右删失数据下用极大似然估计法和贝叶斯估计法去估计指数分布的参数，并结合实证数据讨论两种估计方法的优劣性。

2. 右删失数据

对于不完全数据分为右删失数据，左删失数据以及区间删失数据，其中右删失数据是指研究对象只知道起始事件发生的时间，终点事件发生的时间未知，比如说研究冰箱寿命，冰箱无法制冷为终点时间，但这个事件并不知道什么时候会发生，这种生存时间的类型就叫做右删失数据。在一些寿命问题的调查上，我们也不可能说终点事件没有发生就不结束，根据观察 f 时间结束的不同，右删失数据还可以进一步的分为三种类型，分别是 I 型删失，II 型删失，III 型删失。

I 型删失是指在调查研究过程当中，所有的研究对象观察的起始时间是一样的，并且观察对象的结束时间也固定，但已经发生终点事件的观察对象除外。

如图 1 所示，A，B，E，G，I 表示终点事件已发生，C，D，F，H，J 为没有发生终点时间，到了研究所固定的结束时间，这样 C，D，F，H，J 的数据就是缺失的，但是它们的生存时间必定超过研究时间。

II 型删失是指所有的研究对象的起始观察时间是相同的，一直要观察到有足够的研究对象发生终点事件，其他研究对象的生存时间并不可知，这种情况造成的数据缺失就是 II 型缺失，这种缺失可以说是在调查研究开始之前就已经想好了缺失数据的比例。

如图 2 所示，研究对象共有 10 个，有 6 个发生的终点时间，则研究结束，而剩下的 4 个的生存时间并不可知。

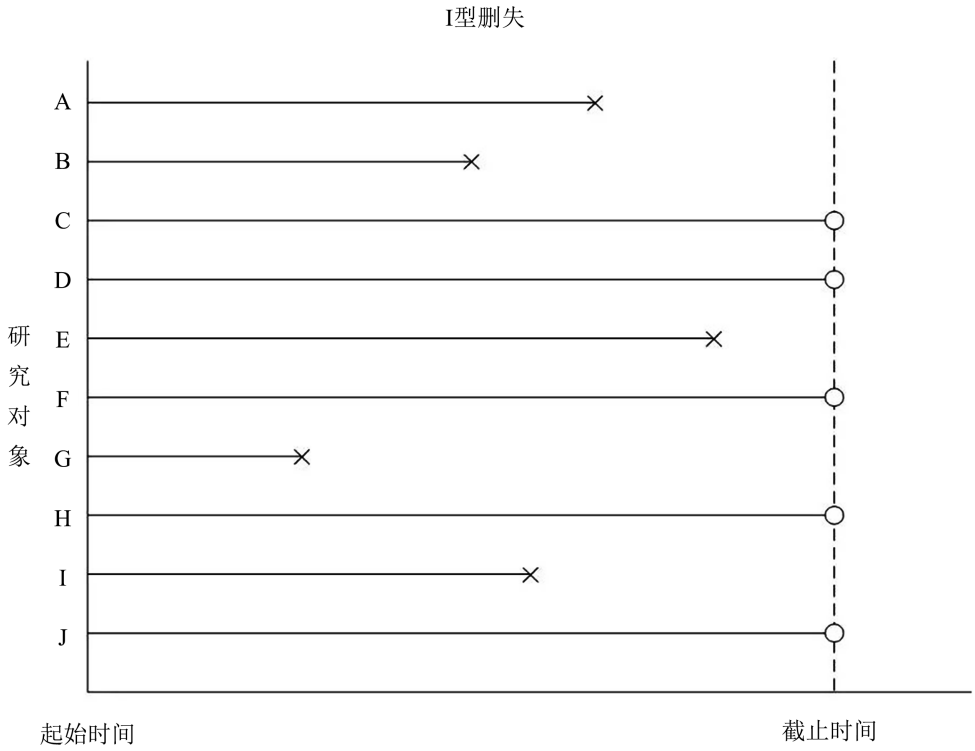

Figure 1. I type censoring data
图 1. I 型删失数据

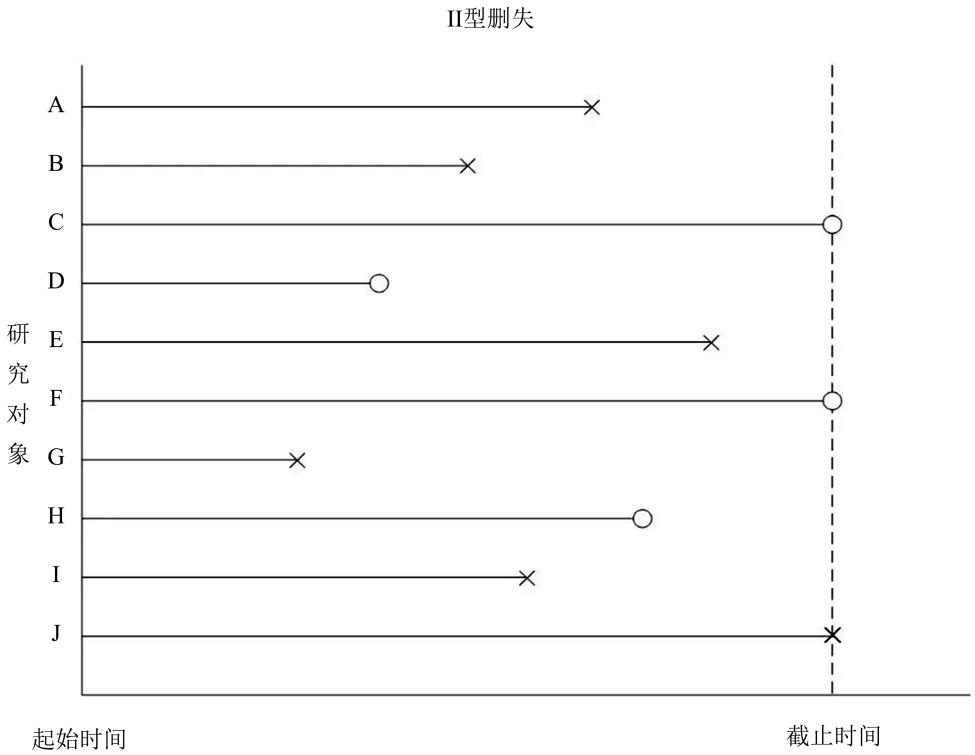

Figure 2. II type censoring data
图 2. II 型删失数据

III 型缺失是指所有研究对象的起始时间不确定，与此同时，研究对象发生终点事件的时间也不可知，不在本文研究范围内。

3. 右删失数据下指数分布参数估计及其性质

3.1. 右删失数据下指数分布参数 λ 的极大似然估计

选取不考虑日常损耗的节能灯的寿命时间作为实验研究观测对象，已知这个观测总体符合指数分布，指数分布的分布函数为：

$$F(x)=\begin{cases}1-\mathrm{e}^{-\lambda x}, x\succ 0\\ 0, x\le 0\end{cases}$$

其概率密度为：

$$f(x)=\begin{cases}\lambda\mathrm{e}^{-\lambda x}, x\succ 0\\ 0, x\le 0\end{cases}$$

其中参数 λ 为失效率，在这里可以解释为在单位时间内节能灯报废的个数。我们从选定时间开始观测节能灯寿命时间，把节能灯发生故障不能正常使用作为终点事件，以一个月为限定，在这一个月内，已发生终点事件的节能灯所获得寿命数据为正常的，而没有发生终点事件，但到了观测时间，这类节能灯获得的数据为右删失数据。设观测总体为 200 个，其观测结果记做 $(x_1,x_2,x_3,\cdots,x_{200})$，其中 x_i 以概率 p 缺失 $(0\prec p\prec 1)$，则实际观测所获得的寿命数据为 (x_i,ξ_i)，$(i=1,2,3,\cdots,200)$，记 ξ_i，$\xi_i=\begin{cases}1, \text{在一个月内}\\ 0, x_i \text{超过一个月}\end{cases}$。这里的 $(x_1,x_2,x_3,\cdots,x_{200})$ 与 $(\xi_1,\xi_2,\xi_3,\cdots,\xi_i)$ 是相互独立的，其中 $p(\xi_i=1)=1-p(\xi_i=0)=1-p$，可以知道 ξ_i 服从两点分布。若记 $n=\sum_{i=1}^{200}\xi_i$，表示 $\xi_i=1$ 的次数，则可以理解为 200 个节能灯中得到观测值的节能灯个数，n 是一个二项分布的随机变量。记观测值为 $(y_1,y_2,y_3,\cdots,y_n)$，这里的 $\sum_{j=1}^{n}y_j=\sum_{i=1}^{200}x_i\xi_i$。

由上述我们可得极大似然函数为：

$$L(\lambda)=\prod_{j=1}^{n}f(y_j,\lambda)=\prod_{j=1}^{n}\lambda\mathrm{e}^{-\lambda y_j}=\lambda^n\mathrm{e}^{-\lambda\sum_{j=1}^{n}y_j}。$$

由上式得 λ 的极大似然估计为：$\lambda=\dfrac{n}{\sum_{j=1}^{n}y_j}$。

上面所采用的方法是没处理缺失数据所做的估计，这样估计误差会很大。如果对于缺失数据直接以 30 天作为替代值，则可以得到 λ 的极大似然估计为 $\lambda=\dfrac{200}{\sum_{i=1}^{200}x_i}$。

现通过我们获得节能灯的数据，由所得观测数据计算得出不处理缺失数据时的参数 $\lambda=0.0378$。若采取插补法，以 30 代替缺失数据的观测值，则计算得出参数 $\lambda=0.0287$。从两个数据我们可以看出不对缺失数据进行处理所估计出的参数值相比于插补法处理缺失数据所估计出来的参数值偏差更大。

3.2. 右删失数据下指数分布参数 λ 的贝叶斯估计

我们仍然选取上述节能灯作为观测对象，采取贝叶斯估计方法对参数进行估计。总体服从指数分布，则选取倒伽马分布 $\Pi(\alpha,\beta)$ 作为 λ 的先验分布，其中 $\alpha\succ 0,\beta\succ 0$ 为超参数。λ 的先验分布为：

$$\Pi(\lambda)=\frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{-(\alpha+1)}\mathrm{e}^{-\frac{\beta}{\lambda}}。$$

要估计参数 λ，我们要先求后验分布的核，后验分布的核等于先验分布 $\Pi(\lambda)$ 乘以样本分布 $f(x)$，则 $\Pi(\lambda)f(x)=\frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{-(\alpha+1+n)}\mathrm{e}^{-\frac{1}{\lambda}\left(\sum_{i=1}^{n}x_i+\beta\right)}$，若取平方损失函数，则 λ 的贝叶斯估计为：$\lambda=\frac{\sum_{i=1}^{200}x_i+\beta}{\alpha+200-1}$ (以插补法处理数据后的参数估计值)，若对缺失数据不做处理，则 λ 的贝叶斯估计为：$\lambda=\frac{\sum_{j=1}^{n}y_j+\beta}{\alpha+n-1}$。

我们根据所获得的节能灯观测数据进行估计，取 $\alpha=3$，$\beta=0.05$，在对缺失数据不做处理时，λ 的参数估计值为 0.0384。用插补法处理缺失数据时得到 λ 的参数估计值为 0.0292。同样我们可以看出用贝叶斯方法对参数进行估计，依旧是对缺失数据进行处理所得到的估计值偏差更小。

3.3. 两种估计方法的比较

以上两种参数估计的方法，对于在右删失数据下的指数总体而言，我们通过上述实例计算可以知道，极大似然估计方法所得参数估计值的偏差更小。但是贝叶斯估计方法所得参数估计值与极大似然估计方法所得估计值相差不大，这可能跟主观的对先验概率的取值以及样本容量有一定的关系。为此，我们进行进一步的模拟比较，假设参数 λ 的真值等于 20，所选择的先验分布中的参 $\alpha=0.8,\beta=60$，样本容量分别取 10 和 20，在右删失数据的情况下比较两种估计方法的优劣性。对样本容量为 10 和样本容量为 20 的情况下都产生 25 组模拟样本，分别计算两种估计方法下估计量的相对误差和相对平方误差，其中 λ_1 为用极大似然估计方法计算得出的估计量，λ_2 为用贝叶斯方法计算得出的估计量。若缺失数据个数为 3 个，所计算出的结果比较见下表 1。

Table 1. Comparison of errors under different missing data
表 1. 不同缺失数据下误差比较

缺失数据个数	$\frac{Bias(\lambda_1)}{\lambda}$	$\frac{Bias(\lambda_2)}{\lambda}$	$\frac{MSE(\lambda_1)}{\lambda^2}$	$\frac{MSE(\lambda_2)}{\lambda^2}$
10 (3)	0.11	0.08	0.012	0.007
20 (3)	0.08	0.07	0.005	0.004

表中，$\frac{Bias(\lambda_i)}{\lambda}$ 表示相对误差，$\frac{MSE(\lambda_i)}{\lambda^2}$ 表示相对均方误差。10(3)表示样本容量为 10，缺失数据为 3，20(3)为样本容量为 20，缺失数据为 3。

4. 结论

由以上讨论可知，对指数分布参数 λ 用极大似然和贝叶斯两种方法估计，其在样本容量相同的情况下，对比目前文献中常见的随机加权法和多重插补法估计效果更好。同时用极大似然估计方法所得的相对误差和相对均方误差都比用贝叶斯估计方法所得结果要高，这就说明贝叶斯估计参数的方法要比极大似然估计方法偏差更小，可信度更高。

基金项目

2020 广西高校中青年教师科研基础能力提升项目“几类不完全数据的若干统计推断问题研究”(编号：2020KY10012)；2019 钦州学院人才引进科研启动项目“基于缺失数据及随机截断数据的若干统计推断”(编号：19KYQD45)。

参考文献

[1] 周旭, 田茂再. 左截断右删失数据下伽马分布的参数推断[J]. 山东理工大学学报(自然科学版), 2023, 37(6): 1-6.

[2] 侯兰宝. 定时区间删失下指数分布的参数估计[J]. 统计与决策, 2020, 36(5): 20-24.

[3] 董小刚, 彭小草, 蒋京京, 等. 部分区间删失数据下广义指数分布的参数估计及应用[J]. 吉林大学学报(理学版), 2022, 60(3): 557-567.

[4] 刘长林, 李云飞. 无失效数据场合指数分布可靠度的Bayes估计[J]. 西华师范大学学报(自然科学版), 2020, 41(1): 59-64.

[5] 刘慧馨. 右删失数据下多响应AFT模型的两阶段估计[J]. 应用概率统计, 2023, 39(1): 10-26.