第 24 卷第 5 期　　四川理工学院学报(自然科学版)　　Vol. 24　No. 5
2011 年 10 月　　Journal of Sichuan University of Science & Engineering(Natural Science Edition)　　Oct. 2011
文章编号:1673-1549(2011)05-0604-03

泊松分布参数估计的比较研究

王丙参[1], 魏艳华[1], 孙春晓[2]
(1. 天水师范学院数学与统计学院, 甘肃　天水　741001;2. 西北农林科技大学理学院, 陕西　杨凌　712100)

摘　要:研究了泊松分布点估计及区间估计,并证明了样本均值是参数 λ 的优良估计量。利用贝叶斯统计分析方法,在取先验分布为共轭分布的情形下,给出了最大后验密度可信区间,即最短可信区间,并通过实例与经典区间估计进行了比较。

关键词:泊松分布;点估计;可信区间
中图分类号:O212　　**文献标识码**:A

泊松分布是一种极其重要的离散型分布,常用来描述稀有事件,如保险精算中,经常用来描述索赔次数,但在实际问题中往往不知参数,因此如何根据抽样结果对参数进行准确估计具有重要意义[1-3]。未知参数 θ 的区间估计比点估计有明显的优势,它即给出参数真值所在的范围,又给出该范围包含真值的可信程度。显然在置信水平 $1-\alpha$ 确定的前提下,区间的长度越短越好。若枢轴量的密度函数(*pdf*)是对称的单峰函数,当两侧各取 $\frac{\alpha}{2}$ 时,区间的长度为最短。如果枢轴量的 *pdf* 非对称,仍按对称情况确定的区间并非最短[4]。签于此,本文研究了泊松分布点估计及区间估计,在取先验分布为共轭分布伽玛分布的情形下,给出了最大后验密度可信区间,即最短可信区间,最后通过实例与经典区间估计进行了比较。

1　泊松分布

定义 1　泊松分布 X 以全体自然数为一切可能值,分布列为 $P(X=k)=e^{-\lambda}\frac{\lambda^k}{k!},k=0,1,\cdots$, 记为 $X\sim P(\lambda)$。

满足下面三个条件的 *r. vX* 服从泊松分布[5]:

(1)普通性:在充分小的观察单位上, X 的取值最多为 1。

(2)平稳性: X 的取值只与单位时间 t 有关,而与观察单位的位置无关。

(3)独立增量性:在某个观察单位上 X 的取值与前面各不同观察单位上 X 的取值均独立。

为了描述稀有事件,只含有一个参数的泊松分布往往是第一选择。由于 Poisson 分布的均值等于方差,所以当风险集体同质时,理赔次数服从泊松分布,但实际中,同质性保单组合的索赔次数并不完全满足此规律。此外,保单组合中索赔次数为零保单数相对较多,主要是保险公司采用了风险回避机制,如免赔额、无赔款优待费率体系等,使得投保人在发生事故时,会权衡利益得失而决定是否索赔。基于索赔次数的这些特点,引出了调零的复合泊松分布类。

2　点估计的比较研究

设 $X_1,\cdots,X_n$ 是来自泊松分布 $P(\lambda)$ 的样本,样本观测值为 $x_1,\cdots,x_n$, $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}X_i$。由格里纹科定理,利用经验分布去替换总体分布是数理统计的理论基础。矩估计简单易行,又具有良好性质,故此方法经久不衰,其实最简单最直接的方法往往也是最有效的方法。显然 λ 的矩估计为 $\bar{x}$。最大似然估计的本质是样本来自使样本出现可能性最大的那个总体,而似然函数可以衡量样本出现概率的大小,因此需找未知参数的估计值使得似然函数达到最大。

$$L(\lambda)=\prod_{i=1}^{n}p(x_i;\lambda)=e^{-n\lambda}\lambda^{n\bar{x}}\prod_{i=1}^{n}\frac{1}{x_i!}$$

收稿日期:2011-09-19
基金项目:甘肃省自然科学研究基金(096RJZE106);甘肃省教育厅项目(0908-07);天水师范学院科研基金(TSA0931)
作者简介:王丙参(1983-),男,河南南阳人,讲师,硕士,主要从事随机分析与保险精算方面的研究,(E-mail)wangbingcan2004@163.com

令偏导等于 0,可得 $\frac{\partial LnL(\lambda)}{\partial\lambda} = -n + \frac{n\bar{x}}{\lambda} = 0$, 即 $\hat{\lambda} = \bar{x}$。可见矩估计与最大似然估计是一样的,人类追求的终极目标就是公平与效率,同理我们追求的最终目标是找到一个最佳统计量,尽管方法不一样,但如果最佳结果存在则一样。

引理 1[6]　设总体 X 的 $pdff(x;\theta)$ 为指数型分布族,即样本的联合 pdf 具有如下形式:

$$\prod_{i=1}^{n} f(x_i;\theta) = C(\lambda)\exp\{\sum_{j=1}^{m} b_j(\lambda) T_j(x_1,\cdots,x_n)\} h(x_1,\cdots,x_n)$$

其中 $\theta = (\theta_1,\cdots,\theta_n)^T \in \Theta$。如果 Θ 中包含一个 m 维矩形,而且 $B = (b_1(\theta),\cdots,b_n(\theta))^T$ 的值域包含一个 m 维开集,则 $T = (T_1(x_1,\cdots,x_n),\cdots,T_n(x_1,\cdots,x_n))^T$ 是参数 θ 的一个充分完备统计量。

定理 1　$\bar{x}$ 是 λ 的一致无偏有效估计量、一致最小方差无偏估计量、相合估计量、充分统计量及完备统计量。

证明　(1)由于 $E\bar{x} = \lambda$, 则 $\bar{x}$ 为一致无偏估计量。

(2) $I(\lambda) = E[\frac{\partial Lnp(X;\lambda)}{\partial\lambda}]^2 = E(\frac{X}{\lambda} - 1)^2 = \frac{E(X-\lambda)^2}{\lambda^2} = \frac{1}{\lambda}$, λ 的任一无偏估计量 $\hat{\lambda}$ 都满足 $D\hat{\lambda} \geqslant \frac{1}{nI(\lambda)} = \frac{\lambda}{n} = D\bar{x}$, 即 $\bar{x}$ 为有效无偏估计量。

(3)有效估计必是一致最小方差无偏估计。

(4) $P(|\bar{x} - \lambda| \geqslant \varepsilon) \leqslant \frac{1}{\varepsilon^2}E(\bar{x}-\lambda)^2 = \frac{\lambda}{n\varepsilon^2} \to 0$, 所以 $\bar{x}$ 为相合估计量。

(5) $L(\lambda) = \prod_{i=1}^{n} p(x_i;\lambda) = e^{-n\lambda}\lambda^{n\bar{x}}\prod_{i=1}^{n}\frac{1}{x_i!}$ 令 $T(x_1,\cdots,x_n) = \frac{\sum_{i=1}^{n} x_i}{n}$, $h(x_1,\cdots,x_n) = \frac{1}{\prod_{i=1}^{n} x_i!}$, $g(T(x_1,\cdots,x_n);\lambda) = \lambda^{nT}e^{-n\lambda}$, 则 $L(\lambda) = h(x_1,\cdots,x_n)g(T(x_1,\cdots,x_n);\lambda)$, 由因子分解定理知 $\bar{x}$ 为充分统计量。

(6)由泊松分布的可加性得 $n\bar{x} \sim P(n\lambda)$, 设 $g(\bar{x})$ 使得 $E_\lambda(g(\bar{x}))\sum_{k=0}^{\infty} g(\frac{k}{n})\frac{(n\lambda)^k}{k!}e^{-n\lambda} = 0, \forall\lambda > 0$, 即 $\sum_{k=0}^{\infty} g(\frac{k}{n})\frac{(n\lambda)^k}{k!}e^{-n\lambda} = 0, \forall\lambda > 0$, 上述幂级数对一切 $\lambda > 0$ 恒收敛 0,只能系数全为 0,可见 $P_\lambda(g(\bar{x})) = 0, \forall\lambda > 0$, 故 $\bar{x}$ 为充分统计量。

也可证明如下:

$$L(\lambda) = \prod_{i=1}^{n} p(x_i;\lambda) = e^{-n\lambda}\lambda^{n\bar{x}}\prod_{i=1}^{n}\frac{1}{x_i!} = e^{-n\lambda}\exp\{n\bar{x}\ln\lambda\}\frac{1}{\prod_{i=1}^{n} x_i!} = C(\lambda)\exp\{\sum_{j=1}^{m} b_j(\lambda)T_j(x_1,\cdots,x_n)\} h(x_1,\cdots,x_n)$$

其中 $C(\lambda) = e^{-n\lambda}$, $h(x_1,\cdots,x_n) = \frac{1}{\prod_{i=1}^{n} x_i!}$, $T(x_1,\cdots,x_n) = \bar{x}$, $b(\lambda) = n\ln\lambda$。

由引理 1 可知 $\bar{x}$ 为充分完备统计量。

由于泊松分布的均值与方差相等,故样本均值也是 σ^2 的一个估计。由于样本的信息比较分散,为了便于分析,引入统计量对样本进行加工,而加工有好、坏之分,为评价好坏,引入了上述统计量指标,如无偏估计保证了没有系统误差,有效估计意味着加工的效率比较高,没有过多的信息遗漏。估计量是一种估计的方法,而估计值是用此方法的一次实现,二者不可混淆。

3　区间估计的比较研究

引理 2[6]　若 $X \sim Be(a,b)$, 则 $Y = X/(1-X) \sim Z(a,b)$; 若 $X \sim Z(n_1/2,n_2/2)$, 则 $Y = (n_2/n_1)X \sim F(n_1,n_2)$。

设 $X_1,\cdots,X_n$ 是来自泊松分布 $P(\lambda)$ 的样本,样本观测值为 $x_1,\cdots,x_n$, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$, 令 $k = \sum_{i=1}^{n} x_i$ 为样本总计数且 $k \sim P(n\lambda)$。当 $k \geqslant 1$ 时,由方程 $\sum_{i=0}^{k-1}\frac{c^i}{i!}e^{-c} = 1 - \alpha + \beta$, $\sum_{i=0}^{k}\frac{d^i}{i!}e^{-d} = \beta$ 可解出置信下限 $c = n\lambda_1$ 和置信上限 $d = n\lambda_2$, 其中 k 为样本总计数, $1-\alpha$ 为置信度, $0 \leqslant \beta \leqslant \alpha, 0 < \alpha < 1$, 则区间长度 $L = d - c$。一般我们采用等概率法 $\beta = \frac{\alpha}{2}$ 求得 c,d, 可此时 L 不一定最短。我们如何在适当的 β 下,使得 L 最短? 当 α 给定时, c,d,L 都是 β 的函数。令 $f(x) = \sum_{i=0}^{m}\frac{x^i}{i!}e^{-x}$, 则 $f'(x) = -\frac{m^i}{m!}e^{-x} < 0$, 故 β 可唯一解出,从而 c,d,L 唯一确定。当 $k = 0$ 时,显然 $c = 0$, 从而 $\beta = \alpha$ 可得 $d = -\ln\alpha$。当 $k = 1$ 时, $\frac{dL}{d\beta} = \frac{d(d-c)}{d\beta} < 0$, 于是当 $\beta = \alpha$ 时, L 有最小值,此时 $c = 0, e^{-d}(1+d) = \alpha$。当 $k \geqslant 2, \beta \to 0 \Rightarrow d \to \infty$, 从而 $\frac{dL}{d\beta} < 0$; 当 $\beta \to \alpha$ 时, $c \to 0$, 从而 $\frac{dL}{d\beta} > 0$; 当 $c \leqslant k-1, d > k$ 时, $\frac{d^2L}{d\beta^2} \geqslant 0$, 故此时在 $(0,\alpha)$ 上, L 的最小值存在且唯一。泊松分布的 pdf 在 k 处有极大值,当 $1-\alpha$ 给定时, L 的最短区间 (c,d) 应包含 k 且在附近,当 α 比较小时, $c \leqslant k-1, d > k$ 可以满足。由上所知我们可以采取 $\beta = \alpha$ 以 $h = 0.00001$ 为步长向左搜索[3], L 先由大变小,再由小变大,在最小处求得 c,d。

由于贝叶斯统计推断利用了先验知识,往往收到较

好的效果,尤其对于小样本。若取 λ 的先验分布为其共轭分布 $\Gamma(a,b)$,由于 $\pi(\lambda)=\frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}$,则 $\pi(\lambda\mid x)\propto\lambda^{n\bar{x}+a-1}e^{-(n+b)\lambda}$,即 λ 的后验分布为 $\Gamma(n\bar{x}+a,b+n\bar{x})$。又 $2(b+n\bar{x})\lambda\sim\chi^2(2(n\bar{x}+a))$,令 $Y=2(n\bar{x}+b)\lambda$,$P(c<Y<d)=P(\frac{c}{2(n\bar{x}+b)}<\lambda<\frac{d}{2(n\bar{x}+b)})$,从而 λ 的可信水平在 $1-\alpha$ 的可信区间长度为 $\frac{d-c}{2(n\bar{x}+b)}$,要使可信区间长度最短,即 $d-c$ 最小,于是:令

$$L=d-c+\mu[F(d,2(n\bar{x}+a))-F(c,2(n\bar{x}+a))-1+\alpha]$$

对 c,d 求偏导并等于 0 可得:

$$\frac{\partial L}{\partial c}=-1-\mu f(c,2(n\bar{x}+a))=0$$

$$\frac{\partial L}{\partial d}=1+\mu f(d,2(n\bar{x}+a))=0$$

从而有 $f(c,2(n\bar{x}+a))=f(d,2(n\bar{x}+a))$。令 $f'(x,2(n\bar{x}+a))=0$ 求得极大值点 $x_0=2(n\bar{x}+a-1)$,当 $x<x_0$ 时,$f(x,2(n\bar{x}+a))$ 严格单调递增;当 $x>x_0$ 时,$f(x,2(n\bar{x}+a))$ 严格单调递减;又因为 $\lim\limits_{x\to\infty}f(x,2(n\bar{x}+a))=\lim\limits_{x\to0}f(x,2(n\bar{x}+a))=0$,故对 $f(c,2(n\bar{x}+a))=f(d,2(n\bar{x}+a))$ 可唯一解出 $c=g(d)$。再由 $F(d,2(n\bar{x}+a))-F(c,2(n\bar{x}+a))=1-\alpha$ 可得 $d=\tilde{d}$,从而可得 $c=\tilde{c}$,求得可信水平为 $1-\alpha$ 的最优区间估计为 $(\frac{\tilde{c}}{2(n\bar{x}+b)},\frac{\tilde{d}}{2(n\bar{x}+b)})$。

例 1 设在 1200 m 长的磁带上缺陷数服从 $P(\lambda)$,λ 的先验分布是 $\Gamma(3,1)$,如今对三盘磁带做检查,分别发现 2、0、6 个缺陷,求 λ 在可信度为 95% 最短可信区间[7]。

解 由题意可知 $a=3,b=1$。又因为 $\bar{x}=\frac{8}{3}$,$n=3$,所以 λ 的后验分布为 $\Gamma(11,9)$,即 $Y\triangleq18\lambda\sim\chi^2(22)$,从而求 λ 在可信度为 95% 最短可信区间为 $\frac{(d-c)}{18}$,由前面推导可得 $c=9.95787$,$d=35.22668$,最短可信区间为 (0.55321, 1.95704),区间长度为 1.403823。若采用等尾区间估计可得(2.0433728936, 0.6101289297),区间长度为 1.43324。显然可见最大后验密度可信区间比一般的等尾区间要短。运行程序如下:

```
function y = myfun100(x)
y(1) = chi2pdf(x(1),22) - chi2pdf(x(2),22);
y(2) = chi2cdf(x(2),22) - chi2cdf(x(1),22) -95;
format long; x = [40,70];
x = fsolve('myfun100',x) x/18
data; d = cinv(0.975,22)/18;put d = ;
c = cinv(0.025,22)/18;put c = ;
a = d - c;put a = ; run;
```

综上所述,实际上对于统计量的 $pdff(x)$ 为单峰函数,只要取 $f(d)=f(c)$ 且 $F(d)-F(c)=1-\alpha$ 就可使得区间估计最短。

参 考 文 献:

[1] 王丙参,魏艳华.保费收取次数为负二项随机过程的风险模型[J].江西师范大学学报,2010,34(6):604-608.

[2] 魏艳华,王丙参,宋立新.均匀分布的优良特性及其应用[J].四川理工学院学报:自然科学版,2010,23(4):385-387.

[3] 匡荣彭,胜 光.Poisson 分布参数估计的几点注记[J]宿州教育学院学报,2007,10(6):132-133.

[4] 岑 忠,丁 勇.泊松分布参数的最短置信区间[J].中国卫生统计,2010,27(2):133-134.

[5] 马先莹,孙红卫,相 静.Poisson 分布易被忽视的重要性质[J]大学数学,2009,25(5):184-186.

[6] 赵选民,徐 伟.数理统计[M].北京:科学出版社,2002.

[7] 茆诗松.贝叶斯统计[M].北京:中国统计出版社,1999.

Comparative Research of Parameter Estimation of Poisson Distribution

WANG Bing-can[1], *WEI Yan-hua*[1], *SUN Chun-xiao*[2]

(1. School of Mathematics and Statistics, Tianshui Normal University, Tianshui 741001, China;
2. College of Science, Northwest University of Agriculture and Forestry, Yangling 712100, China)

Abstract: Point estimation and interval estimation of Poisson distribution are discussed. It proves that the sample mean is a good estimator of the parameters λ. When prior distributions are conjugate distributions, it gives their confidence interval of the highest posterior density by using the method of Bayes statistical analysis, furthermore, the shortest confidence interval and classical interval estimation are compared by an example.

Key words: Poisson distribution; point estimation; confidence interval