DOI:10.13546/j.cnki.tjyjc.2020.01.006



# 零浮动Poisson项目计数法在敏感数据抽样调查中的应用

刘　寅[1]，吴　琴[2]

（1.中南财经政法大学 统计与数学学院，武汉 430074；2.华南师范大学 数学科学学院，广州 510631）

**摘　要**：文章将Poisson-Poisson项目计数法进行推广，提出零浮动Poisson项目计数法，其中，非敏感辅助变量来自于一个参数已知的零浮动Poisson分布。并给出了该模型下敏感参数极大似然估计的EM算法以及构造其置信区间的bootstrap方法。此外，还对该模型保护受访者隐私的能力加以讨论，发现该模型的隐私保护要优于Poisson-Poisson项目计数法。最后，从随机模拟的结果表明在该模型下利用本文所介绍的分析方法可以得到敏感参数的较为准确的估计。

**关键词**：零浮动Poisson项目计数法；敏感数据抽样调查；EM算法；隐私保护度

**中图分类号**：C81　　**文献标识码**：A　　**文章编号**：1002-6487(2020)01-0029-03

## 0　引言

针对敏感数据收集的抽样调查技术与分析方法近几十年来得到广泛关注。如果直接询问受访者有关敏感信息的问题，由于涉及到个人隐私而导致难以获得理想的数据。Warner(1965)[1]最早提出的一个Warner模型通过引入一种随机化装置来辅助调查人员间接地获得有关敏感信息的数据。在随后的几十年中，关于敏感数据收集的抽样调查技术不断发展，主要形成了三大分支技术。其中，以Miller(1984)[2]提出的设计方法为代表的项目技术方法在社会学、心理学领域有着广泛的应用。该方法要求受访者被随机的分配到两个组，分别回答若干个非敏感、不相关的且答案为“是”或“否”的问题或者同样的非敏感、不相关问题加上一个研究人员所关心的、答案也为“是”或“否”的敏感问题。在调查过程中，受访者无需针对每一个问题分别给出答案而只需报告答案为“是”的问题的个数即可。Tian 等 (2017)[3]创造性地提出一个Poisson项目计数法，该方法将项目计数法中若干个非敏感、不相关的且答案为“是”或“否”的问题用一个答案可为0、1、2、…中的任一数字的非敏感问题记性替代，该模型需要两组样本对敏感参数进行估计，其中一组受访者仅报告非敏感问题的答案，另一组受访者则需报告非敏感问题与敏感问题之和。

然而上述方法仅适用于对某个特定群体中受访者具有某一敏感特征的比率进行研究，但是某些情况下，我们所关心的敏感特征是离散型定量数据，而我们希望对这一特征的平均水平加以了解。因此，Liu 等(2018)[4]提出Poisson-Poisson项目计数法来对诸如此类的问题进行研究。在调查过程中，受访者需要对一个敏感问题 $Q_X$ (例如：你在过去的一个月中吸毒的次数？)作出回答。由于受访者关于敏感问题 $Q_X$ 的答案可能为0、1、2、…中的任一数字，因此，可以假定与敏感问题 $Q_X$ 相应的随机变量 $X$ 来自Poisson分布，即 $X\sim Poisson(\theta)$，$\theta>0$。为了保护受访者的隐私不被泄露，调查人员需借助电脑软件产生来自参数已知的Poisson分布的随机数，受访者无需报告关于敏感问题 $Q_X$ 的答案是多少，只需报告所产生的随机数与该问题的答案之和即可。对于调查者来说，由于他并不知道受访者关于敏感问题的答案是什么，因此受访者的隐私被有效保护起来。记与所产生的随机数相对应的随机变量为 $U$，与敏感问题 $Q_X$ 相对应的随机变量为 $X$，与受访者答案相对应的随机变量为 $Y$，故 $U$ 与 $X$ 独立且 $Y=U+X$，其中，$U\sim Poisson(\lambda_0)$，$\lambda_0>0$ 且 $\lambda_0$ 已知。

但是在实际调查中，假定 $U\sim Poisson(\lambda_0)$ 并不一定总是合适。对于某些受访者来说，他们可能并不理解随机数的产生原理因此并不愿意配合调查。在这种情况下，调查人员可以用一个答案为0、1、2、…中的、与敏感问题 $Q_X$ 不相关的、非敏感的问题 $Q_U$ 进行替换，例如，“你在过去一年中出国旅游的次数？”。注意到受访者对于问题 $Q_U$ 的答案可能并非服从一个标准的Poisson分布，因为大多数受访者可能并没有出国旅游的经历，从而受访者关于非敏感问题 $Q_U$ 的真实回答中可能包含过多的0(超过正常的Poisson分布的0所出现的频数)。这种情况下，考虑使用零浮动Poisson分布来作为 $U$ 的分布更为恰当。因此，本文在Liu 等 (2018)[4]提出的Poisson-Poisson项目计数法的基础上做进一步推广，考虑零浮动Poisson项目计数法来

**基金项目**：国家自然科学基金资助项目(11601524；11401226)；中南财经政法大学青年教师资助项目(31721811206)

**作者简介**：刘　寅(1986—)，女，湖北咸宁人，博士研究生，讲师，研究方向：敏感性抽样调查。

吴　琴(1985—)，女，河北保定人，博士研究生，讲师，研究方向：敏感性抽样调查。

对离散型定量敏感数据的均值特征进行分析，给出相应的参数估计方法，并对所提出的模型保护受访者隐私的能力进行讨论。

## 1 零浮动Poisson项目计数法调查设计

在实际调查中，受访者同时面对一个敏感问题 $Q_X$（例如：你在过去的一个月中吸毒的次数？）和一个非敏感、不相关问题 $Q_U$（例如：你在过去一年中出国旅游的次数？）。受访者无需分别报告两个问题的答案，只需根据真实情况报告两个问题的答案之和即可。令随机变量 $U$ 和 $X$ 分别表示问题 $Q_U$ 和 $Q_X$ 的答案，考虑下述零浮动Poisson项目计数法：

$$Y=U+X \tag{1}$$

其中，非敏感变量 $U$ 来自参数已知的零浮动Poisson分布，即 $U\sim ZIP(\phi_0, \lambda_0)$，$0\le\phi_0\le 1$，$\lambda>0$；敏感变量 $X$ 来自Poisson分布，即 $X\sim Poisson(\theta)$，$\theta$ 为感兴趣的未知参数；$U$ 与 $X$ 相互独立。

由公式(1)可知，随机变量 $Y$ 的概率质量函数为：

$$f(Y=y)=\phi_0\cdot\frac{\theta^y e^{-\theta}}{y!}+(1-\phi_0)\cdot\frac{(\lambda_0+\theta)^y e^{-(\lambda_0+\theta)}}{y!} \tag{2}$$

即：

$$f(Y=y)=\phi_0\cdot Poisson(y|\theta)+(1-\phi_0)\cdot Poisson(y|\lambda_0+\theta) \tag{3}$$

因此，随机变量 $Y$ 可看作为两个Poisson分布——$Poisson(\theta)$ 与 $Poisson(\lambda_0+\theta)$——随机变量的加权组合。

## 2 极大似然估计的EM算法

假设在一次调查中，调查人员共收集到 $n$ 个有效的回答，记 $Y_{obs}=\{y_1,\dots,y_n\}$，则 $\theta$ 的观测数据对数似然函数为：

$$\ell(\theta|Y_{obs})=\sum_{i=1}^{n}\log\left[\phi_0\cdot\frac{\theta^{y_i}e^{-\theta}}{y_i!}+(1-\phi_0)\cdot\frac{(\lambda_0+\theta)^{y_i}e^{-(\lambda_0+\theta)}}{y_i!}\right]$$

由于在上述对数似然函数中，log函数的里层是一个求和的形式，因此，直接令 $d\ell(\theta|Y_{obs})/d\theta=0$ 无法直接求得未知参数 $\theta$ 的极大似然估计的显示表达。因此，本文考虑用EM算法[5]来导出 $\theta$ 的极大似然估计。

为了达到这一目的，此处需要引入两层潜在变量。首先，引入潜在变量序列 $\{Z_1,\dots,Z_n\}$，使得对第 $i$ 个受访者，1被拆分成 $Z_i$ 和 $1-Z_i$ 两部分，其中，$Z_i$ 对应于 $\phi\cdot Poisson(\theta)$ 部分而 $1-Z_i$ 对应于 $(1-\phi)\cdot Poisson(\lambda+\theta)$ 部分。因此，$\{Z_1,\dots,Z_n\}$ 的条件预测分布为：

$$Z_i|(Y_{obs},\theta)=Bernoulli\left(\phi_0\cdot\frac{\theta^{y_i}e^{-\theta}}{y_i!}\Big/f(y_i)\right),\quad i=1,\dots,n \tag{4}$$

记 $\{z_1,\dots,z_n\}$ 为 $\{Z_1,\dots,Z_n\}$ 的实现值。在此基础之上，引入潜在变量 $Z$ 使得 $\sum_{i=1}^{n}y_i(1-z_i)$ 被分解为 $Z$ 和 $\sum_{i=1}^{n}y_i(1-z_i)-Z$ 两部分。故，$Z$ 的条件预测分布为：

$$Z\sim Binomial\left(\sum_{i=1}^{n}y_i(1-z_i),\ \frac{\lambda_0}{\lambda_0+\theta}\right) \tag{5}$$

记 $z$ 为 $Z$ 的实现值，完全数据为 $Y_{com}=\{y_1,\dots,y_n,z_1,\dots,z_n,z\}$。则，$\theta$ 的完全数据对数似然函数为：

$$\ell(\theta|Y_{com})\propto\left(\sum_{i=1}^{n}y_i-z\right)\log(\theta)-n\theta$$

EM算法的M步计算 $\theta$ 的完全数据极大似然估计：

$$\hat{\theta}=\left(\sum_{i=1}^{n}y_i-z\right)\Big/n \tag{6}$$

E步将公式(6)中的 $z$ 由其条件期望进行替代。结合公式(4)和公式(5)，$z$ 的条件期望为：

$$E(Z|Y_{obs},\theta)=\left[\sum_{i=1}^{n}y_i\left(1-\phi_0\cdot\frac{\theta^{y_i}e^{-\theta}}{y_i!}\Big/f(y_i)\right)\right]\cdot\frac{\lambda_0}{\lambda_0+\theta} \tag{7}$$

## 3 Bootstrap置信区间

由于敏感参数 $\theta$ 的极大似然估计没有显示表达，只能通过由公式(6)和公式(7)定义的EM迭代算法来进行求解，因而关于 $\hat{\theta}$ 的方差也难以得到其精确表达。因此，本文将采用bootstrap自助抽样的方法来构造关于敏感参数 $\theta$ 的置信区间。首先，从分布 $ZIP(\phi,\lambda)$ 中产生独立的样本 $\{u_1^*,\dots u_n^*\}$；其次，基于由公式(6)和公式(7)定义的EM算法得到的 $\hat{\theta}$，独立地从分布 $Poisson(\hat{\theta})$ 中产生样本 $\{x_1^*,\dots x_n^*\}$；然后，令 $y_i^*=u_i^*+x_i^*$，$i=1,\dots,n$，得到一组bootstrap样本 $Y_{obs}^*=\{y_1^*,\dots,y_n^*\}$。依据公式(6)和公式(7)计算极大似然估计 $\hat{\theta}^*$。将此过程重复 $G$ 次，得到 $G$ 个bootstrap估计 $\{\hat{\theta}^*(1),\dots,\hat{\theta}^*(G)\}$。因此，$\hat{\theta}$ 的标准误差，$se(\hat{\theta})$，可由 $G$ 个bootstrap估计的样本标准差来估计，即：

$$sd(\hat{\theta})=\left\{\frac{1}{G-1}\sum_{g=1}^{G}\left[\hat{\theta}^*(g)-\left(\hat{\theta}^*(1)+\cdots+\hat{\theta}^*(G)\right)\Big/G\right]^2\right\}^{1/2} \tag{8}$$

如果 $\{\hat{\theta}^*(1),\dots,\hat{\theta}^*(G)\}$ 近似正态分布，则 $\theta$ 的一个置信水平为95%的bootstrap置信区间为：

$$\left[\hat{\theta}-z_{\alpha/2}\cdot sd(\hat{\theta}),\ \hat{\theta}+z_{\alpha/2}\cdot sd(\hat{\theta})\right] \tag{9}$$

其中，$z_{\alpha/2}$ 表示标准正态分布的 $\alpha/2$ 上分位点。如果 $\{\hat{\theta}^*(1),\dots,\hat{\theta}^*(G)\}$ 并非近似正态分布，则 $\theta$ 的一个置信水平为95%的bootstrap置信区间为：

$$\left[\hat{\theta}_L,\ \hat{\theta}_U\right] \tag{10}$$

其中，$\hat{\theta}_L$ 和 $\hat{\theta}_U$ 分别为 $\{\hat{\theta}^*(1),\dots,\hat{\theta}^*(G)\}$ 的 $\alpha/2$ 和 $1-\alpha/2$ 分位点。

## 4 隐私保护度

对于敏感数据调查设计来说，该设计的有效与否在一定程度上取决于该设计能否较好地保护受访者隐私不被

泄露。为了衡量在零浮动Poisson项目计数法下受访者隐私受到保护的程度，本文定义该设计的隐私保护度为：

$$DPP_y=\Pr(X\geq 1|Y=y)=1-\Pr(X=0|Y=y)$$
$$=1-\Pr(U=y,\ X=0|Y=y)$$
$$=\begin{cases}0, & y=0\\ 1-\dfrac{(1-\phi_0)\lambda_0{}^y e^{-(\lambda_0+\theta)}}{\phi_0\theta^y e^{-\theta}+(1-\phi_0)(\lambda_0+\theta)^y e^{-(\lambda_0+\theta)}}, & y\geq 1\end{cases} \quad (11)$$

从公式(11)中可以看出，$DPP_y\in[0,1]$。由公式(11)所定义的$DPP_y$反应的是在受访者的回答中暴露其具有明暗属性特征的概率，故，$DPP_y$越小越接近于0，受访者的隐私保护得越好。当$y=0$时，由于此时受访者不具备敏感特征，故暴露其隐私的概率为0。当$y>0$时，由于非敏感变量$U$的干扰，受访者本身可能具有敏感特征也可能不具有敏感特征，此时，由公式(11)易验证：

$$DPP_y\leq 1-\frac{(1-\phi_0)\lambda_0{}^y e^{-(\lambda_0+\theta)}}{(1-\phi_0)(\lambda_0+\theta)^y e^{-(\lambda_0+\theta)}}=1-\left(\frac{\lambda_0}{\lambda_0+\theta}\right)^y=DPP_y^{PPICT}$$

换言之，当受访者报告的答案大于0时，零浮动Poisson项目计数法比Liu等(2018)[4]提出的Poisson-Poisson项目计数法在受访者隐私保护上的表现更好。

## 5 随机模拟

采用随机模拟的方式来考察不同$(\phi_0,\lambda_0)$组合下利用本文所介绍的分析方法来对敏感参数$\theta$进行估计的效果。此处，独立地产生$L=1000$组观测样本，每一组样本容量均为$n=100$。对于每一组随机样本，利用公式(6)和公式(7)所定义的EM算法来计算$\theta$的估计，并将这1000个估计的平均作为$\theta$的极大似然估计，如表1中第3列所示。依据所得到的$\theta$的极大似然估计，再利用上文所介绍的bootstrap方法产生$G=1000$组随机样本，并计算相应的bootstrap均值、bootstrap标准差以及两种bootstrap置信区间，如表1中第4—7列所示。

表1 不同$(\phi_0,\lambda_0)$组合下$\theta$的极大似然估计、标准差及bootstrap置信区间

| 参数 | | MLE | Mean① | SD② | 95% Bootstrap CI③ | 95% Bootstrap CI④ |
|---|---|---|---|---|---|---|
| $\theta=2$ | $\phi_0=0.2,\lambda_0=1$ | 2.0024 | 1.9988 | 0.1709 | [1.6674, 2.3374] | [1.6739, 2.3467] |
| | $\phi_0=0.2,\lambda_0=2$ | 2.0078 | 2.0115 | 0.2015 | [1.6129, 2.4028] | [1.5876, 2.4183] |
| | $\phi_0=0.2,\lambda_0=3$ | 2.0150 | 2.0278 | 0.2403 | [1.5441, 2.4859] | [1.5305, 2.5255] |
| | $\phi_0=0.2,\lambda_0=4$ | 1.9978 | 1.9960 | 0.2715 | [1.4657, 2.5298] | [1.4653, 2.5159] |
| | $\phi_0=0.2,\lambda_0=5$ | 2.0077 | 2.0211 | 0.2942 | [1.4310, 2.5843] | [1.4952, 2.6116] |
| $\theta=5$ | $\phi_0=0.4,\lambda_0=1$ | 4.9891 | 4.9855 | 0.2398 | [4.5191, 5.4592] | [4.5070, 5.4640] |
| | $\phi_0=0.4,\lambda_0=2$ | 5.0027 | 4.9966 | 0.2724 | [4.4687, 5.5367] | [4.4702, 5.5435] |
| | $\phi_0=0.4,\lambda_0=3$ | 4.9991 | 4.9961 | 0.2963 | [4.4183, 5.5799] | [4.4212, 5.5983] |
| | $\phi_0=0.4,\lambda_0=4$ | 4.9859 | 4.9990 | 0.3326 | [4.3340, 5.6378] | [4.3511, 5.6223] |
| | $\phi_0=0.4,\lambda_0=5$ | 5.0085 | 5.0104 | 0.3517 | [4.3191, 5.6978] | [4.3164, 5.7191] |

注：①Mean = bootstrap估计的均值；②SD = bootstrap估计的标准差，由公式(8)给出；③由公式(9)定义的Bootstrap置信区间；④由公式(10)定义的Bootstrap置信区间。

由表1的结果可以看出，对于不同的$(\phi_0,\lambda_0)$组合，由公式(6)和公式(7)所定义的EM算法计算得到的未知敏感参数$\theta$的极大似然估计以及由bootstrap方法所得到的未知敏感参数$\theta$的均值都非常接近于其真实值。当$\theta$的真实值为2时，相应的估计的标准差在区间[0.17, 0.30]之间；当$\theta$的真实值为5时，相应的估计的标准差在区间[0.23, 0.36]之间，估计精度较好。此外随着辅助非敏感零浮动Poisson随机变量中参数$\lambda$取值的增加，关于未知敏感参数$\theta$的估计的标准差随之增加，并且其相应的两种bootstrap置信区间的宽度增加，精确度有所下降。造成这一现象的可能原因在于随着$\lambda$的增大，对于敏感信息收集所造成的干扰越多，因此估计精度受到影响。但总体上来说，估计效果不错，较为准确。

## 6 结论

本文将Liu等(2018)[4]提出的Poisson-Poisson项目计数法进行推广，提出零浮动Poisson项目计数法，其中，非敏感辅助变量来自于参数已知的零浮动Poisson分布，即$ZIP(\phi_0,\lambda_0)$。特别地，当$\phi_0=0$时，零浮动Poisson项目计数法退化成Poisson-Poisson项目计数模型。换言之，Poisson-Poisson项目计数模型为零浮动Poisson项目计数法的一个特殊情形，因此，零浮动Poisson加法模型的应用范围更广。此外，从受访者隐私受到保护的方面来看，本文所讨论的零浮动Poisson项目计数法比Poisson-Poisson项目计数法的表现更好。在本文中，仅考虑了非敏感辅助零浮动Poisson变量的参数$\phi_0$和$\lambda_0$已知的情形，对于未知的情况，将在未来的研究中进行讨论。

**参考文献：**


[1]Warner S L. Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias[J]. Journal of the American Statistical Association, 1965, (60).

[2]Miller J D. A New Survey Technique for Studying Deviant Behavior [D]. Washington: George Washington University, 1984.

[3]Tian G L, Tang M L, Wu Q,et al. Poisson and Negative Binomial Item Count Techniques for Surveys With Sensitive Questions[J]. Statistical Methods in Medical Research, 2017, 26(2).

[4]Liu Y, Tian G L, Tang M L,et al.Poisson-Poisson Item Count Techniques for Survey With Sensitive Discrete Quantitative Data [J]. Statistical Papers, 2018.

[5]Dempster A P, Larid N M,Rubin D B.Maximum Likelihood From Incomplete Data via the EM Algorithm (with discussion) [J]. Journal of the Royal Statistical Society, Series B, 1977, 39(1)


（责任编辑/刘柳青）

DOI:10.13546/j.cnki.tjyjc.2020.01.007



# 基于贝叶斯的Bootstrap置信区间

张芃哓[1]，田茂再[1, 2a,2b]

（1.兰州财经大学 统计学院，兰州 730020；2.中国人民大学 a.应用统计科学研究中心；b.统计学院，北京 100872）

**摘 要**：在许多领域中，Bootstrap成为一种数据处理的有效方法。很多情况下，模型中感兴趣的参数的置信区间难以构建，为了解决这一问题，文章提出了一个新的贝叶斯Bootstrap置信区间的估计量，并做了蒙特卡洛模拟比较，结果比经典区间估计方法和经典Bootstrap方法更优，并进行了实例分析。

**关键词**：Bootstrap；贝叶斯；置信区间

**中图分类号**：O21　　**文献标识码**：A　　**文章编号**：1002-6487(2020)01-0032-04

## 0 引言

统计学是研究数据分析的算法，它研究的性质和算法的性质都随时间而变化。在20世纪上半叶，统计学家可用的主要工具是数学、逻辑和早期计算器。自Efron(1979)[1]发表Bootstrap后，很快就获得了广泛关注，其原因是它对列出的每个统计子组提出了即时的内部测试问题。一些统计学家认识到，主要基于分析操作的统计程序并没有为处理计算机时代出现的数据集提供有效的技术。蒙特卡罗算法是用一种非常直观的方法来估计依赖于未知参数的复杂抽样分布。这种蒙特卡罗方法本身就很有趣，它反映了算法和计算实验在统计学中日益重要的作用。

Bootstrap估计成为统计误差估计和假设检验中常用的工具之一。近年来Bootstrap理论在国外引起了广泛关注，同时有一系列的版本和扩展[2]。Arlot等(2010)利用广义加权Bootstrap重采样分位数，构造了高维独立随机分布的高斯向量(或对称有界分布)的样本均值的非渐近置信区间；Spokoiny(2015)[3]证明了Bootstrap估计在参数维度不断增加、样本容量有限以及模型误设情况下的有效性。

近年来大数据越来越普遍，数据越来越多，传统方式处理数据较为复杂，而贝叶斯理论能很好地处理复杂数据，所以本文引用Rubin(1981)[4]的贝叶斯理论，他们为贝叶斯推理提供了Bootstrap方法，也提出了适当的渐近分析。Newton和Raftery(1994)[5]又进一步提出了基于似然的Bootstrap方法。

当今各行各业都涉及数据，重新研究数据分析相关理

**基金项目**：国家自然科学基金资助项目(11861042)

**作者简介**：张芃哓(1993—)，男，甘肃兰州人，硕士研究生，研究方向：统计模型。

（通讯作者）田茂再(1969—)，男，湖南凤凰人，教授，博士生导师，研究方向：复杂数据分析。

# Application of Zero-inflated Poisson Item Count Technique for Sample Surveys With Sensitive Data

Liu Yin[1], Wu Qin[2]

(1. School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan 430074, China; 2. School of Mathematics, South China Normal University, Guangzhou 510631, China)

**Abstract**: This paper extends the Poisson-Poisson item count technique to the zero-inflated Poisson item count technique, where the non-sensitive auxiliary variable comes from a zero-inflated Poisson distribution with known parameters. The paper also provides the EM algorithm to calculate the maximum likelihood estimate for the sensitive parameter as well as the bootstrap method in constructing confidence intervals. Furthermore, the paper discusses the ability of privacy protection of the proposed model and finds out that it performs better than that of the Poisson-Poisson item count approach. Finally, the results of the stochastic simulation show that a more accurate estimate of the sensitive parameters can be obtained by using the analytical method described in the paper.

**Key words**: zero-inflated Poisson item count technique; sample surveys with sensitive characteristics; EM algorithm; degree of privacy protection