Statistics and Application 统计学与应用, 2023, 12(5), 1352-1360
Published Online October 2023 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2023.125139

Hans汉斯

MAR机制下泊松项目计数技术对敏感问题的研究分析

杨　丹

重庆理工大学理学院，重庆

收稿日期：2023年9月20日；录用日期：2023年10月11日；发布日期：2023年10月25日

摘　要

在对敏感问题进行研究时，出于对隐私的保护，受访者往往拒绝回答或是给出不真实的答案，故现如今敏感问题的研究增益精进，目前十分有效又广泛使用的一种模型为泊松项目计数技术模型，该模型从实验设计上就很好地保护了受访者的隐私以及匿名性，从而引导受访者给出真实有效的答案。然而尽管实验设计有效，在真实调查中，仍然会出现数据缺失的情况，数据缺失分为几种不同情况，分为完全随机缺失，随机缺失和不可忽略的缺失；而由于泊松项目计数技术的实验设计能够保护受访者隐私，因此，假设数据的缺失不是由于受访者害怕泄露隐私而拒绝回答；故本文研究泊松项目计数技术在随机缺失(MAR)下的统计理论推导。经理论计算得，MAR条件下的泊松项目计数技术中，其中的随机缺失数据只依赖于可观测到的数据；故在随机缺失数据下的泊松项目计数技术模型，可以只通过可观测数据进行计算。

关键词

随机缺失，泊松项目计数技术，敏感问题

Analysis of the Sensitivity of the Counting Technology for the Poisson Project under the MAR Mechanism

Dan Yang

College of Science, Chongqing University of Technology, Chongqing

Received: Sep. 20th, 2023; accepted: Oct. 11th, 2023; published: Oct. 25th, 2023

Abstract

When researching sensitive issues, respondents often refuse to answer or give untrue answers due to the protection of privacy, so nowadays, the research on sensitive issues has been improved,

文章引用: 杨丹. MAR 机制下泊松项目计数技术对敏感问题的研究分析[J]. 统计学与应用, 2023, 12(5): 1352-1360.
DOI: 10.12677/sa.2023.125139

and a very effective and widely used model is the Poisson item counting technique model, which protects the privacy and anonymity of the respondents from the experimental design and guides the respondents to give true and effective answers. However, despite the validity of the experimental design, in real surveys, there are still cases of missing data, which are categorized into several different situations, including completely random missing, random missing, and non-negligible missing; since the experimental design of Poisson's item-counting technique protects the privacy of the respondents, it is assumed that the missing data is not due to the fact that respondents are afraid of revealing their privacy and refusing to answer; therefore, this paper investigates the effect of Poisson's item-counting technique on the random missing (MMS) and the anonymity of the respondents counting technique under missing at random (MAR). After theoretical calculations, the Poisson item counting technique under MAR conditions, in which the random missing data only depends on the observable data; therefore, the model of Poisson item counting technique under random missing data can be calculated only by observable data.

Keywords

Stochastic Absence, Poisson Item Counting Technique, Sensitivity Issue

1. 引言

1.1. 研究背景

目前敏感问题的研究越来越重要，关于敏感问题的研究也越来越多，在涉及到敏感问题的时候，人们往往会为了保护个人隐私在涉及到敏感问题，拒绝回答或给出不真实的答案都会导致敏感比例的估计造成偏差，使研究结果不准确。

故 Warner (1965) [1]首先提出了一种利用随机器械的随机响应机制以研究敏感问题，以潜在的数理统计方法来保护受访者的隐私，由随机作答模式进行参数估计以推测敏感性群体比□；Greenberg *et al*. (1969) [2]提出的不相关问题的随机作答模式，改进了 Warner 模型中两个相关问题可能导致回答不真实的情况；Christofifides (2003) [3]在随机作答设计下将此研究中的二分法推广到多分法，提出了多分法的敏感比例估计；Miller (1984) [4]提出了项目技术技术(ICT)不再需要随机器械，而是通过统计分布来估计敏感比例；Guo-Liang Tian *et al*. (2017) [5]简化了 ICT 考虑了伯努利分布和泊松分布(负二项分布)下的 PICT 和 NICT 模型以估计敏感比例。

由于敏感问题的实验设计上主要存在两个问题：

1) 由于隐私泄露，受访者不给予真实答案，从而影响参数估计。

2) 由于实验设计缺陷，受访者不给予答案，造成数据缺失。

在 PICT 的提出后，此项关于敏感问题的计数备受关注，改善了 ICT 的部分机制缺陷，不仅更好的保护了受访者的隐私，并且实验设计方法又非常的简单实用；现在关于改善受访者不真实回答的问题已经有了非常多改进的实验设计，故本文将在 PICT 的基础上，利用 MAR 机制研究数据缺失下的敏感问题。

1.2. 研究意义

现在已经有基于的随机响应机制的缺失数据下的敏感问题研究，目前对于敏感问题发展出了更多的

非随机响应模型，不仅简化了随机响应模型，不再依赖一个随机器械，调查也更加的方便且节约成本，故泊松项目计数技术已经被广泛使用，可仍存在数据缺失的现象，由于敏感问题的实验设计都在很好地保护受访者的隐私，所以在受访者均给出真实答案的强假设下，本文假设数据的缺失不是由于受访者想要保护自己隐私而拒绝回答问题，而认为缺失数据与数据本身无关，而仅与观测到的数据有关。

为进一步解决敏感问题中的缺失数据问题，本文主要研究在随机缺失模型(MAR)下，PICT 中缺失数据下的敏感问题分析。进一步改进泊松项目计数技术，解决数据缺失问题，推进敏感问题的研究。

1.3. 研究现状

在数据研究中，时常都会出现数据缺失的现象，而在敏感问题的研究中，受访者拒绝回答涉及隐私的信息更是常见，也会出现数据缺失的现象。针对 Warn (1965)、Greenberg *et al*. (1969)以及 Christofides (2003)的随机作答涉及，部分解释变量缺失的问题目前已有一些研究。Lee *et al*. (2011) [6]提出了逻辑回归的半参数估计方法来处理部分数据缺失；Wang *et al*. (2002) [7]又提出了联合条件似然的方法解决部分数据缺失问题；Wu and Tian *et al*. (2020) [8]提出了非服从情况下(存在受访者具备敏感特征却给出假的 0)的 PICT 模型，以解决受访者为保护个人隐私而故意给出不真实答案的情况并给出相应的统计推断，指出不考虑非服从情况会导致验证的偏差。

2. 模型理论概述

2.1. 泊松项目计数技术(PICT)

一种项目计数技术模型(ICT)被 Miller 等人提出，主要通过对随机作答问题的合理设计，得到真实有效的答案，这种模型不再需要随机器械，从而减少了成本简化了设计。ICT 的实验设计由 K 个非敏感问题和 1 个敏感问题组成，实验人员被随机分为控制组和实验组，控制组对仅回答 K 个非敏感问题，实验组不仅要回答 K 个相同的非敏感问题，还要回答 1 个敏感问题，所以问题均为二项式问题，仅需要回答“Yes”或者“No”，最终受访者只需要提交“Yes”的数量。这项研究模型简化了实验设计，被广泛应用，但是 ICT 模型存在一个致命的缺陷，当实验组提交的结果是 K + 1 个“Yes”的时候，受访者无疑具有敏感属性，在这种情况下，ICT 的实验设计不能够很好的隐藏受访者的隐私，从而得到不正确(故意错填或者不填)的答案，实验结果的置信区间和参数估计均会受到影响，这会使 ICT 的可应用性和可信度降低。

为了克服这一问题，Guo-Liang Tian *et al*.对 ICT 模型进行改进并提出了两个新的模型，分别是 Poisson ICT (PICT) and Negative Binomial ICT (NICT)，ICT 虽然简化了实验设计，但是不能很好的保护受访者的隐私，而 PICT 和 NICT 的实验设计，不仅简单且可行性高，并且能够很好的保护受访者的隐私，从而使实验结果可信度更好。PICT/NICT 的实验设计由仅 1 个非敏感问题和 1 个敏感问题组成，其中，非敏感问题服从泊松分布($X = 0,1,2,\cdots$)，而敏感问题服从伯努利分布($Z = 1$ 具有敏感属性，相反没有)，受访者同样是被随机的分配到控制组和实验组，最终受访者只需提交两个问题的结果总和数据($Y = X + Z$)，如见图 1；当数据结果过离散的时候，将泊松分布替换为负二项分布。

2.2. 随机缺失机制

在数据集中，我们往往将不含缺失数据的变量称为完全变量，存在缺失值的变量称为不完全变量。

Rubin (1976) [9]提出三种不同的缺失机制：

1) 完全随机化缺失(MCAR)：完全随机化缺失是指缺失的观测值与其他可观测到的变量及此观测值是没有关系的。即数据的丢失是随机的，不依赖于任何不完全变量和完全变量。

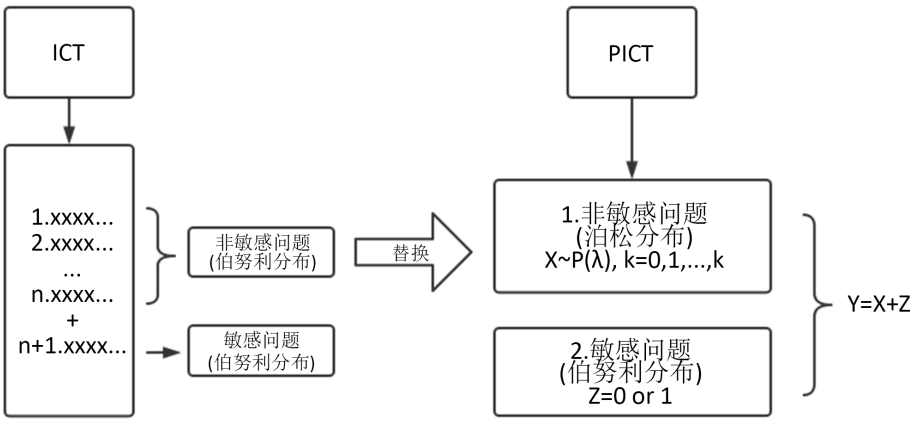

Figure 1. ICT model and PICT model
图 1. ICT 模型及 PICT 模型

2) 随机缺失(MAR)：随机化缺失则是指数据的确缺失值与其他可观测到的变量有关，但与缺失的数据值本身没有关系。

3) 不可忽略的缺失(MNAR)：不可忽略的缺失则是缺失的数据与其本身及其他观测的变量有关。

本文主要研究 MAR 下数据缺失处理，随机缺失意味着缺失的概率与缺失数据本身无关，但仅与部分已观测到的数据有关，数据的缺失不是完全随机的，该类数据的缺失依赖于其他完全变量，用数学公式表示为：

$$P(\delta \mid y) = P(\delta \mid y_{obs}) \tag{1}$$

其中 y_{obs} 为可观测部分数据，δ 为示性函数：

$$\delta = \begin{cases} 1 & y \text{可观测} \\ 0 & y \text{观测不到} \end{cases} \tag{2}$$

3. 模型建立

3.1. 随机缺失下模型参数估计理论

令 $y = (y_1, y_2, \cdots, y_n)$ 是来自密度为 $f(y;\theta)$ 的无限总体的随机样本的实现，θ 为感兴趣的参数。在式(2)的示性函数下，假设响应概率模型为：

$$\Pr(\delta \mid y) = \Pr(\delta \mid y;\phi) \tag{3}$$

其中 ϕ 是这个模型的参数。

令 $(y_{i,obd}, y_{i,mis})$ 分别为 y_i 的观测部分和缺失部分，因此，与其去观测 (δ_i, y_i)，而是去观测 $(\delta_i, y_{i,obs})$。

$$y_{i,obs} = \begin{cases} y_i & \delta_i = 1 \\ * & \delta_i = 0 \end{cases}$$

在给出的模型(3)下，$(\delta_i, y_{i,obs})$ 的边缘密度函数为：

$$\tilde{f}(y_{i,obs}, \delta_i; \theta, \phi) = \int f(y_i;\theta) P(\delta_i \mid y_i;\phi) \mathrm{d}\mu(y_{i,mis}) \tag{4}$$

那么在独立同分布条件下，通过对式(4)进行累积求和，可以得到 y 和 δ 的联合概率密度函数为：

$$\tilde{f}(y_{obs}, \delta_i; \theta, \phi) = \prod_{i=1}^{n} \tilde{f}(y_{i,obs}, \delta_i; \theta, \phi) \tag{5}$$

式(5)可被称为观测似然，给出观测域为

$$\Re\left(y_{obs,\delta}\right)=\left\{y;y_{obs}\left(y_i,\delta_i\right)=y_{i,obs},i=1,\cdots,n\right\}$$

则在观测域$\Re\left(y_{obs,\delta}\right)$下，观测似然函数定义为

$$\begin{aligned}L_{obs}\left(\theta,\phi\right)&=\int_{\Re\left(y_{obs,\delta}\right)}f\left(y;\theta\right)P\left(\delta\mid y;\phi\right)\mathrm{d}\mu\left(y\right)\\&=\prod_{i=1}^{n}\left[\int f\left(y_i;\theta\right)P\left(\delta\mid y_i;\phi\right)\mathrm{d}\mu\left(y_{i,mis}\right)\right]\\&=\int f\left(y;\theta\right)P\left(\delta\mid y;\phi\right)\mathrm{d}\mu\left(y_{mis}\right)\\&=\int f\left(y,\delta;\theta,\phi\right)\mathrm{d}\mu\left(y_{mis}\right)\end{aligned}\tag{6}$$

$f\left(y\right)$为y的密度函数，$f\left(\delta\mid y\right)$为$\delta$在$y$条件下的条件密度函数，且

$$\begin{cases}f\left(\delta_i\mid y_i\right)=\left\{\pi\left(y_i;\phi\right)\right\}^{\delta}\left\{1-\pi\left(y_i;\phi\right)\right\}^{1-\delta}\\\pi\left(y_i;\phi\right)=\Pr\left(\delta_i=1\mid y_i;\phi\right)\end{cases}$$

参数ϕ可以被看作是讨厌参数，并不是直接感兴趣的参数，但依旧要进行估计。

在标量y的特殊情况下，观测似然为：

$$L_{obs}\left(\theta,\phi\right)=\prod_{\delta_i=1}\left[f\left(y_i;\theta\right)\pi\left(y_i;\phi\right)\right]\times\prod_{\delta_i=0}\left[\int f\left(y_i;\theta\right)\left\{1-\pi\left(y_i;\phi\right)\right\}\mathrm{d}y_i\right]\tag{7}$$

定理 3.1 (Rubin,1976) $P_{\phi}\left(\delta\mid y\right)$为给定$y$下$\delta$的联合密度函数，$f_{\theta}\left(y\right)$是$y$的联合密度函数，在给定条件下：1) 参数$\theta$和$\phi$是不同的；2) MAR 的条件下，观测似然可以写为：

$$L_{obs}\left(\theta,\phi\right)=L_1\left(\theta\right)L_2\left(\phi\right)\tag{8}$$

则θ的 MLE 就可以通过最大化$L_1\left(\theta\right)$来获得。

在(7)式中，MAR下的似然函数由两部分组成，对于前半部分可观测得到的数据可以直接得到观测似然，而对于后半部分的缺失数据，需要通过积分的形式求得，为当似然函数以积分的形式存在时，计算上是很困难的，所以我们引入得分函数和平均得分函数，通过平均得分函数的特有性质，来求得参数的极大似然估计。

1) 得分函数$S\left(\theta;y\right)$定义为：

$$S\left(\theta;y\right)=\frac{\partial}{\partial\theta}\ln f\left(y,\theta\right)\tag{9}$$

本质上，得分函数就是联合密度函数求对数的一阶偏导，而在独立同分布的条件下时，似然函数的本质就是联合密度函数，故得分函数又可以理解为是对数似然函数的一阶偏导。

在正则条件下，允许交换积分和微分的顺序，存在关于期望和方差的性质：

$$\begin{aligned}E_{\theta}\left\{S\left(\theta;y\right)\right\}&=0\\V_{\theta}\left\{S\left(\theta;y\right)\right\}&=E_{\theta}\left\{I\left(\theta;y\right)\right\}\end{aligned}\tag{10}$$

2) 平均得分函数$\overline{S}\left(\eta\right)$定义为：($\eta=\left(\theta;\phi\right)$，$\eta$为位置参数的集合)

$$\overline{S}\left(\eta\right)=E\left\{S_{com}\left(\eta\right)\mid y_{obs},\delta\right\}$$

其中完整数据集下的得分函数$S_{com}\left(\eta\right)$就是通过(9)式进行计算。

$$S_{com}\left(\eta\right)=\frac{\partial}{\partial\eta}\log f\left(y,\delta;\eta\right)$$

平均得分函数具有与得分函数相同的期望和方差基本性质。

对式(7)的似然函数求一阶偏导得到函数。在 MAR 下，平均得分函数为：

$$\bar{S}(\eta)=\sum_{\delta_i=1}S(y;\eta)+\sum_{\delta_i=0}E\{S(y;\eta)\}$$

由于式(10)得分函数的期望为 0，故平均得分函数的另外一个有效性质为：在正则条件下，观测得分函数等于平均得分函数(Fisher, 1922) [10]。

$$S_{obs}(\eta)=\bar{S}(\eta)$$

这时在令平均得分函数等于 0 ($\bar{S}(\eta)=0$)得到唯一解来求解参数就容易很多。

3.2. 完整数据下 PICT 参数估计

令 $Y=X+Z$，且 $X\sim Poisson(\lambda)$，$Z\sim Bernoulli(\pi)$，X 与 Z 相互独立，λ 为待估计的未知参数，另外还需要估计感兴趣的参数 $\pi=\Pr(Z=1)$ ($Z=1$ 具有敏感特征，否则不具备敏感特征)；令 $\{x_i\}_{i=1}^{n_1}$ 和 $\{y_i\}_{i=1}^{n_2}$ 分别为控制组(记为 $Y^{(1)}$)和实验组(记为 $Y^{(2)}$)的观测数据。

通过 EM 算法计算得到 π 的极大似然估计，引入潜在数据 $Y_{latent}=\{z_1,\cdots,z_{n_2}\}$，也就是在实验组中敏感问题的潜在数据。因此完整数据为 $Y_{com}=\{Y_{obs},Y_{latent}\}$，因此完整数据下的极大似然函数为：

$$L(\pi,\lambda\mid Y_{obs},Y_{latent})=\pi^{\sum_{j=1}^{n_2}z_j}(1-\pi)^{n_2-\sum_{j=1}^{n_2}z_j}\times\left(\prod_{i=1}^{n_1}\frac{\lambda^{x_i}e^{-\lambda}}{x!}\right)\left(\prod_{i=1}^{n_2}\frac{\lambda^{y_j-z_j}e^{-\lambda}}{(y_j-z_j)!}\right)\tag{11}$$

由于(9)式子存在潜在数据，则没有明确的解析解，故采用 EM 算法进行迭代计算。

M 步：通过对数似然函数求偏到令其为 0，可以得到完整数据下 π 和 λ 的极大似然估计。

$$\begin{aligned}\hat{\pi}&=\frac{1}{n_2}\sum_{j=1}^{n_2}z_j\\ \hat{\lambda}&=\frac{\sum_{i=1}^{n_1}x_i+\sum_{j=1}^{n_2}(y_j-z_j)}{n_1+n_2}\end{aligned}\tag{12}$$

E 步：对缺失变量 Z_j 求期望

$$E(Z_j\mid Y_{obs},\pi,\lambda)=\frac{\pi y_j}{\pi y_j+\lambda(1-\pi)}\tag{13}$$

(13)替换(12)中的 Z_j，E 步和 M 步不断迭代之至收敛。

3.3. 随机缺失下 PICT 的模型参数估计

3.3.1. 模型建立

在前两节的 MAR 和 PICT 的理论基础下，本节将 IPCT 运用到 MAR 模型中，研究数据随机缺失下 PICT 的模型建立。

在控制组 $Y^{(1)}$ 中前 r_1 个数据为可以观测到的数据，n_1-r_1 个数据是没有观测到的数据，记为 $y_{mis}^{(1)}$；在实验组 $Y^{(2)}$ 中前 r_2 个数据为可以观测到的数据，而后 n_2-r_2 个数据是没有观测到的数据，记为 $y_{mis}^{(2)}$；$y_{mis}=(y_{mis}^{(1)},y_{mis}^{(2)})$，完整数据则为 $y_{com}=\{y_{obs},y_{mis},y_{latent}\}$，其中实验组中的潜在数据为 $y_{latent}=\{z_1,\cdots,z_{n_2}\}$，那么根据式(11)可以构建 MAR 下的似然函数为：

$$L_{obs}(\lambda,\pi)=\left\{\prod_{i=1}^{r_1}\frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\times\prod_{j=1}^{r_2}\frac{e^{-\lambda}\lambda^{y_j-z_j}}{(y_j-z_j)!}\pi^{z_j}(1-\pi)^{1-z_j}\right\}^{I\{\delta=1\}}\times\left\{\prod_{i=r_1+1}^{n_1}\frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\times\prod_{j=r_2+1}^{n_2}\frac{e^{-\lambda}\lambda^{y_j-z_j}}{(y_j-z_j)!}\pi^{z_j}(1-\pi)^{1-z_j}\right\}^{I\{\delta=0\}}$$

由定理 3.1 可知，MAR 随机缺失机制下，数据只依赖与可以观测到的数据，由式 8，对数似然函数就可以取观测部分的对数似然函数替代完全数据下的对数似然函数。

取对数，可得到观测对数似然函数为：

$$\ln L_{obs}(\lambda,\pi)=\alpha-r_1\lambda+\sum_{i=1}^{r_1}x_i\ln\lambda-r_2\lambda+\sum_{j=1}^{r_2}(y_j-z_j)\ln\lambda+\sum_{j=1}^{r_2}z_j\ln\pi+\sum_{j=1}^{r_2}(1-z_j)(1-\pi) \tag{14}$$

其中，$\alpha=-\sum_{i=1}^{r_1}\ln x_i!-\sum_{j=1}^{r_2}\ln(y_j-z_j)!$。

由式(14)再分别对 λ 和 π 求偏导得到平均得分函数，再令其为 0 求得解，得到：

$$\overline{S}_1(\lambda,\pi)=\frac{\sum_{i=1}^{r_1}x_i+\sum_{j=1}^{r_2}(y_i-z_i)}{\lambda}-(r_1+r_2)$$

$$\overline{S}_2(\lambda,\pi)=\frac{\sum_{i}^{r_2}z_i}{\pi}-\frac{\sum_{i}^{r_2}(1-z_i)}{1-\pi}$$

再令其为 0 求得解，得到：

$$\begin{aligned}\hat{\pi}'&=\frac{1}{r_2}\sum_{j=1}^{r_2}z_j\\ \hat{\lambda}'&=\frac{\sum_{i=1}^{r_1}x_i+\sum_{j=1}^{r_2}(y_j-z_j)}{r_1+r_2}\end{aligned} \tag{15}$$

所以在 MAR 下，所得到的估计量也可以通过简单地忽略样本的缺失部分来获得。

在上式(15)中，z_i 作为潜在变量，依旧是观测不到的，所以采用 EM 算法，由于 $z_i\sim Bernoulli(\pi)$，其概率为 $\Pr(\pi=1)=\frac{\pi y_j}{\pi y_j+\lambda(1-\pi)}$，故 z_i 的条件期望为，

$$E(Z_j\mid Y_{obs},\pi,\lambda)=\frac{\pi y_j}{\pi y_j+\lambda(1-\pi)} \tag{16}$$

再将式 16 代入式 15 中替换掉 z_i，不断循环之至收敛。

3.3.2. Bootstrap 置信区间估计

Bootstrap 方法是一种重采样技术，用来估计标准误差、置信区间和偏差。根据 Bootstrap 抽样，每次抽样都可以计算得一个均数，若设置重抽样次数为 $G=1000$，则可以得到 1000 个均数，以这些均属为原始数据，再求出这 1000 个均数的均数，得到的均数值就是利用 Bootstrap 方法得到的点估计，而对于 95% 的置信区间，则需要分别计算出第 2.5%和 97.5%的分位数。

那么将 Bootstrap 方法运用到 MAR 下的 PICT 模拟中，感兴趣的参数为敏感属性 π，通过 EM 算法可以计算得到 $\hat{\pi}'$ (如式 15)，产生独立同分布的数据，通过 Bootstrap 方法计算产生 $\hat{\pi}'^*$，迭代 G 次，得到 $\{\hat{\pi}'(g)^*\}_{g=1}^{G}$；那么 π 的 $(1-\alpha)100\%$ 的分位数置信区间为：

$$\left[\hat{\pi}'_{100(\alpha/2)},\hat{\pi}'_{100(1-\alpha/2)}\right]$$

4. 模拟研究

对模型进行模拟实验，产生随机数 X 为均值为 λ 的泊松分布，Z 服从伯努利分布。控制组中 $Y^{(1)}=X$，实验组中 $Y^{(2)}=X+Z$；由于样本量的大小会对实验结果产生影响，故本文选取 3 中不同的样本量大小进行模拟实验，分别为小样本 $N=100$，中样本 $N=500$，和大样本 $N=1000$，其中随机缺失 20%的数据；参数的初值选取为 $\pi=[0.05,0.10,0.15,0.20,0.25]$，$\lambda=[2,3,4]$，置信水平设置为 $\alpha=0.05$。

1) 完整数据下 $\hat{\pi}$ 的估计和置信宽度如下：

Table 1. Parameter π estimation under complete data
表 1. 完整数据下参数 π 估计

		$\lambda=2$	$\lambda=3$	$\lambda=4$
	π	$\hat{\pi}$	$\hat{\pi}$	$\hat{\pi}$
$N=100$	0.05	0.0508	0.0507	0.0501
	0.10	0.1009	0.0995	0.1008
	0.15	0.1501	0.1505	0.1497
	0.20	0.2085	0.1996	0.1996
	0.25	0.2503	0.2570	0.2594
$N=500$	0.05	0.0509	0.0500	0.0509
	0.10	0.1009	0.1001	0.1008
	0.15	0.1508	0.1503	0.1507
	0.20	0.2008	0.2001	0.2006
	0.25	0.2509	0.2509	0.2508
$N=1000$	0.05	0.0500	0.0500	0.0500
	0.10	0.0999	0.1000	0.1000
	0.15	0.1500	0.1499	0.1500
	0.20	0.1999	0.2000	0.1999
	0.25	0.2499	0.2501	0.2499

Table 2. Parameter π estimation under missing data
表 2. 缺失数据下参数 π 估计

		$\lambda=2$	$\lambda=3$	$\lambda=4$
	π	$\hat{\pi}$	$\hat{\pi}$	$\hat{\pi}$
$N=100$	0.05	0.0594	0.0507	0.0503
	0.10	0.1075	0.1083	0.1086
	0.15	0.1574	0.1494	0.1576
	0.20	0.2052	0.2007	0.2091
	0.25	0.2570	0.2580	0.2442
$N=500$	0.05	0.0500	0.0500	0.0499
	0.10	0.1000	0.0999	0.1007
	0.15	0.1504	0.1507	0.1507
	0.20	0.2005	0.2009	0.2009
	0.25	0.2500	0.2506	0.2505

Continued

	0.05	0.0500	0.0509	0.0509
	0.10	0.1000	0.1009	0.0999
$N = 1000$	0.15	0.1500	0.1509	0.1499
	0.20	0.2006	0.2008	0.1999
	0.25	0.2507	0.2509	0.2507

根据表 1~2，实验结果表明，当样本量越大的时候参数的估计效果越好；采用 EM 算法，会得到更加可靠的参数估计结果，模拟数据表明参数估计结果较好，偏差较小；实验的模拟结果也可以看出，在完整数据下和随机缺失下的参数估计结果差距非常小，证实了在 MAR 随机缺失时，数据只依赖于观测数据。

5. 结论

在随机缺失模型 MAR 下，数据不完全缺失，而缺失部分的数据只依赖于可观测到的数据，对于缺失部分数据可以直接忽略不计，所得到的估计量也可以通过简单地忽略样本的缺失部分来获得，所以对于 MAR 模型下数据缺失处理方法有删除，插补，不处理等。

而本文研究有效且广泛应用的泊松项目计数技术模型，在数据缺失下的情况；由于泊松项目计数技术模型的实验设计，主要就是为了通过统计推断的方法，保护受访者的隐私，从而得到真实可靠的答案，故本文假设数据的缺失不再是由于受访者害怕泄露隐私而拒绝回答而导致，而是由一些随机影响因素(如受访者由于时间冲突没有进行调查)，而导致的随机缺失，即研究在 MAR 模型下泊松项目计数技术的计算原理实现。经理论计算得到，MAR 条件下的泊松项目计数技术中，其中的随机缺失数据只依赖于可观测到的数据；故在随机缺失数据下的泊松项目计数技术模型，可以只通过可观测数据进行计算。

参考文献

[1] Warner, S.L. (1965) Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association*, **60**, 63-69. https://doi.org/10.1080/01621459.1965.10480775

[2] Greenberg, B.G., Abul-Ela, E.L.A., Simmons, W.R. and Horvitz, D.G. (1969) The Unrelated Question Randomized Response Model: Theoretical Framework. *Journal of the American Statistical Association*, **64**, 520-539. https://doi.org/10.1080/01621459.1969.10500991

[3] Christofides, T.C. (2003) A Generalized Randomized Response Technique. *Metrika*, **57**, 195-200. https://doi.org/10.1007/s001840200216

[4] Miller, J.D. (1984) A New Survey Technique for Studying Deviant Behavior. Doctoral Dissertation, George Washington University, Washington, DC.

[5] Tian, G.L., Tang, M.L., Wu, Q., *et al*. (2017) Poisson and Negative Binomial Item Count Techniques for Surveys with Sensitive Question. *Statistical Methods in Medical Research*, **26**, 931-947. https://doi.org/10.1177/0962280214563345

[6] Lee, S.M., Li, C.S., Hsieh, S.H., *et al*. (2011) Semiparametric Estimation of Logistic Regression Model with Missing Covariates and Outcome. *Metrika*, **75**, 621-653. https://doi.org/10.1007/s00184-011-0345-9

[7] Wang, C.Y., Chen, J.C., Lee, S.M. and Ou, S.T. (2002) Joint Conditional Likelihood Estimator in Logistic Regression with Missing Covariate Data. *Statistica Sinica*, **12**, 555-574.

[8] Wu, Q., Tang, M.L., Fung, W.H., *et al*. (2020) Poisson Item Count Techniques with Noncompliance. *Statistics in Medicine*, **39**, 4480-4498. https://doi.org/10.1002/sim.8736

[9] Rubin, D.B. (1976) Inference and Missing Data. *Biometrika*, **63**, 581-592. https://doi.org/10.1002/sim.8736

[10] Fisher, R.A. (1922) On the Mathematical Foundation of Theoretical Statistics. *Philosophical Transactions of the Royal Society*, 222, 309-368. https://doi.org/10.1098/rsta.1922.0009