基于社会敏感问题的问卷调查方案设计与分析

张立卓, 郭 伟

(对外经济贸易大学 统计学院, 北京 100029)

摘 要: 利用概率论与数理统计方法, 以对酒后驾车现象为例, 就社会敏感问题的问卷调查方案设计, 建立数学模型, 并对模型结果进行分析.

关键词: 调查方案; 模型; 估计量; 数学期望; 标准差

1 引言

在社会问卷调查中, 会遇到因涉及个人隐私或利害关系而不受调查对象欢迎, 感到尴尬的所谓敏感问题, 敏感性问题是指与个人或单位的隐私或私人利益有关而不便向外界透露的问题. 例如, 你是否有过酒后驾车、赌博、偷税、漏税、考试作弊等行为? 这时, 即使做了无记名的问卷调查, 也很难消除被调查者的顾虑与担扰, 极有可能拒绝或故意做出错误的回答, 从而难以保证调查数据的真实性, 导致调查结果存在很大的偏差. 因此寻求解决敏感性问题调查的有效方法至关重要. 文献 [1] 以学生考试作弊现象的调查与估计为例, 分别介绍了 Warnner 模型 [4]、Simmons 模型 [5] 和 Christofides 模型 [6], 分析了三种模型的特点与优缺点. 本文受文献 [1] 中 Christofides 模型的启发, 以酒后驾车现象为例, 依据两种不同的选答方式, 建立模型 I 与 II, 并对模型结果进行比较与分析. 本文之所以选择以酒后驾车现象为例进行调查与估计, 一方面是因为酒后驾车所导致的交通事故给人民的生命财产带来严重的危害, 有效治理酒后驾车刻不容缓. 通过问卷调查来高精度地估计有过酒后驾车行为的司机所占的比例, 有利于为交通管理部门减少并预防交通事故、治理酒后驾车提供决策参考. 另一方面, 通过对 Christofides 模型一定程度的推广, 提高了模型的适用范围, 并对以后的数学建模案例教学提供更多的范例和应用示范.

2 模型 I

2.1 问卷调查方案设计与选答规则

调查者制作两套外形完全相同的卡片, 其中第一套卡片上写有: A_1: 你有过酒后驾车行为吗? 如果有过, 请回答数字 "1", 如果从未有过, 请回答数字 "0". 第二套卡片上写有: 请直接回答数字 "1". 将两套卡片充分混合后放在一个盒子中, 被调查者随机抽取一张, 看后放回, 根据卡片上的问题做出回答.

2.2 假设

1) 被调查司机共 n 位, 每位均独立作答;

2) 被调查者中有过酒后驾车行为的司机真实回答问题的概率为 $\alpha(0 < \alpha \leq 1$, 且可由调

收稿日期: 2015-10-17

查者事先估计), 其他情形均真实作答;

3) 盒子中第一套卡片所占的比例为 k, 且 $k \neq 0$;

4) 被调查者中回答 "1" 的人数共 m_1 位.

2.3 模型建立与求解

设每位被调查者有过酒后驾车行为的概率为 p_{A_1}, 我们要估计的是有过酒后驾车行为的司机的比例, 该比例可用 p_{A_1} 来估计, 为此下面求 p_{A_1} 的估计量.

引入随机变量 X_i[1],

$$X_i = \begin{cases} 1, & \text{若第 } i \text{ 位被调查者回答 "1"}, \\ 0, & \text{若第 } i \text{ 位被调查者回答 "0"}, \end{cases} \quad i = 1, 2, \ldots, n,$$

X_i	0	1
P	1-p	p

其中 p 为每位被调查者回答 "1" 的概率, 显然 $X_1, X_2, \cdots, X_n$ 独立同分布, 可视为来自总体 X 的简单随机样本, 其数学期望与方差分别为 $E(X_i) = p$, $D(X_i) = p(1-p)$, 由全概率公式[2],

$$P\{X_i = 1\} = p = k\alpha p_{A_1} + (1-k) \tag{1}$$

于是,

$$P\{X_i = 0\} = 1 - p = k - k\alpha p_{A_1} \tag{2}$$

由题设, 在 $X_1, X_2, \cdots, X_n$ 的观察值 $x_1, x_2, \cdots, x_n$ 中, 有 m_1 个取 1, $n - m_1$ 个取 0, 因此 $X_1 + X_2 + \cdots + X_n = m_1$, 建立关于 p_{A_1} 的似然函数[3],

$$L(x_1, x_2, \cdots, x_n; p_{A_1}) = \prod_{i=1}^{n} P\{X_i = x_i\} = p^{m_1}(1-p)^{n-m_1}$$

上式两边取对数, 并将 (1),(2) 式代入得,

$$\ln L(p_{A_1}) = m_1 \ln\left(k\alpha p_{A_1} + (1-k)\right) + (n - m_1)\ln\left(k - k\alpha p_{A_1}\right)$$

上式两边关于 p_{A_1} 求导, 并令导数为零,

$$\frac{d}{dp_{A_1}} \ln L(p_{A_1}) = m_1 \frac{k\alpha}{k\alpha p_{A_1} + (1-k)} + (n - m_1)\frac{-k\alpha}{k - k\alpha p_{A_1}} = 0$$

解之得 p_{A_1} 的最大似然估计值为

$$\widehat{p_{A_1}} = \frac{1}{\alpha k}\left(\frac{m_1}{n} - (1-k)\right) = \frac{1}{\alpha k}\left(\frac{1}{n}\sum_{i=1}^{n} x_i - (1-k)\right), \ k \neq 0, \ \alpha \neq 0 \tag{3}$$

进而得 p_{A_1} 的最大似然估计量为

$$\widehat{p_{A_1}} = \frac{1}{\alpha k}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - (1-k)\right), \ k \neq 0, \ \alpha \neq 0 \tag{4}$$

易知, 上式也是 p_{A_1} 的矩估计量.

2.4 模型性质与分析

由 (1) 式知, 该估计量的数学期望为

$$\begin{aligned} E(\widehat{p_{A_1}}) &= \frac{1}{\alpha k} E\left(\frac{1}{n}\sum_{i=1}^{n} X_i - (1-k)\right) = \frac{1}{\alpha k}\left(\frac{1}{n}\sum_{i=1}^{n} E(X_i) - (1-k)\right) \\ &= \frac{1}{\alpha k}\left(p - (1-k)\right) = p_{A_1} \end{aligned}$$

该估计量是无偏的, 其方差为

$$D(\widehat{p_{A_1}}) = \frac{1}{\alpha^2 k^2} D\left(\frac{1}{n}\sum_{i=1}^{n} X_i - (1-k)\right) = \frac{1}{\alpha^2 k^2}\left(\frac{1}{n^2}\sum_{i=1}^{n} D(X_i)\right) = \frac{p(1-p)}{\alpha^2 n k^2}$$

为了对方差作进一步的分析, 将上式的右端分解为

$$D(\widehat{p_{A_1}}) = \frac{p_{A1}(1-\alpha p_{A_1})}{\alpha n} + \frac{(1-k)(1-\alpha p_{A_1})}{\alpha^2 k n}. \tag{5}$$

从 (5) 式可以看出, 估计量的方差 $D(\widehat{p_{A_1}})$ 由两部分组成, 第一部分是直接调查 (即 $k=1$) 情况下的方差, 而第二部分是由于随机选答机制所带来的方差 [1].

2.5 模型评价

首先, 较之于回答 "是" 与 "否", 模型 I 中所要求回答的是数字, 可以减轻被调查者的顾虑, 实现更大可能的真实作答. 其次, 对部分被调查者不真实回答问题的因素已作考虑, 可使模型反映的数据更客观.

3 模型 II

3.1 问卷调查方案设计与选答规则

问题 A_2 : 你有过酒后驾车行为吗?

调查者准备一套外形完全相同的卡片, 每张卡片上写有 "0" 或 "1" 中某一数字, 被调查者随机地抽取一张, 看后放回. 若被调查者有过酒后驾车行为, 则回答 "1" 与他抽取的数字之差, 反之, 则回答他抽取的数字. 对部分有过酒后驾车行为的被调查者有可能不真实回答问题的因素应加以考虑.

3.2 假设

1) 被调查司机共 n 位, 每位均独立作答;

2) 被调查者中有过酒后驾车行为的真实回答问题 A_2 的概率为 $\alpha(0<\alpha\leq 1$, 且可由调查者事先估计), 其他情形均真实作答;

3) 盒子中写有数字 "1" 的卡片所占的比例为 k, 且 $k\neq\frac{1}{2}$;

4) 被调查者中回答数字 "1" 的人数为 m_2 位.

3.3 模型建立与求解

引入随机变量 X_i[1], 表示第 i 位被调查者抽到的数字, 则 X_i 的概率分布为

X_i	0	1
P	1-k	k

$i=1,2,\cdots,n,$

显然 $X_1, X_2, \cdots, X_n$ 独立同分布, 且其数学期望与方差分别为

$$E(X_i)=k,\ D(X_i)=k(1-k),$$

又引入随机变量 Y_i[1],

$$Y_i=\begin{cases}1, & \text{若第 } i \text{ 位被调查者有过酒后驾车行为,}\\ 0, & \text{若第 } i \text{ 位被调查者从未有过酒后驾车行为,}\end{cases} \quad i=1,2,\ldots,n,$$

显然 $Y_1, Y_2, \ldots, Y_n$ 独立同分布, 设每位被调查者有过酒后驾车行为的概率为 p_{A_2}, 则

Y_i	0	1
P	1-p_{A_2}	p_{A_2}

易知 $X_1, X_2, \cdots, X_n$ 与 $Y_1, Y_2, \cdots, Y_n$ 相互独立,

再引入随机变量 Z_i[1], 表示第 i 位被调查者所回答的数字, 即

$$Z_i = |X_i - Y_i|, \quad i = 1, 2, \cdots, n$$

显然 $Z_1, Z_2, \cdots, Z_n$ 独立同分布, 由全概率公式 [2],

$$P\{Z_i = 1\} = p = k(1 - p_{A_2}) + k(1-\alpha)p_{A_2} + (1-k)\alpha p_{A_2} = k - (2k-1)\alpha p_{A_2} \tag{6}$$

于是,

$$P\{Z_i = 0\} = 1 - p = 1 - k + (2k-1)\alpha p_{A_2} \tag{7}$$

其中 p 为每位被调查者回答数字 "1" 的概率, Z_i 的数学期望与方差分别为

$$E(Z_i) = k - (2k-1)\alpha p_{A_2}, \ D(Z_i) = k - k^2 + (2k-1)^2 \alpha p_{A_2}(1 - \alpha p_{A_2})$$

依假设, 在 $Z_1, Z_2, \cdots, Z_n$ 的观察值 $z_1, z_2, \cdots, z_n$ 中有 m_2 个取 1, $n - m_2$ 个取 0, 因此 $Z_1 + Z_2 + \cdots + Z_n = m_2$, 建立关于 p_{A_2} 的似然函数 [3],

$$L(z_1, z_2, \cdots, z_n; p_{A_2}) = \prod_{i=1}^{n} P\{Z_i = z_i\} = p^{m_2}(1-p)^{n-m_2}$$

上式两边取对数, 并将 (6), (7) 式代入得,

$$\ln L(p_{A_2}) = m_2 \ln\left(k - (2k-1)\alpha p_{A_2}\right) + (n - m_2)\ln\left(1 - k + (2k-1)\alpha p_{A_2}\right)$$

上式两边关于 p_{A_2} 求导, 并令导数为零,

$$\frac{d}{dp_{A_2}} \ln L(p_{A_2}) = m_2 \frac{-(2k-1)\alpha}{k - (2k-1)\alpha p_{A_2}} + (n - m_2)\frac{(2k-1)\alpha}{1 - k + (2k-1)\alpha p_{A_2}} = 0$$

注意 $k \neq \frac{1}{2}$, 解之得 p_{A_2} 的最大似然估计值为

$$\widehat{p_{A_2}} = \frac{1}{\alpha(2k-1)}\left(k - \frac{m_2}{n}\right) = \frac{1}{\alpha(2k-1)}\left(k - \frac{1}{n}\sum_{i=1}^{n} z_i\right), k \neq \frac{1}{2}, \alpha \neq 0 \tag{8}$$

进而得 p_{A_2} 的最大似然估计量为

$$\widehat{p_{A_2}} = \frac{1}{\alpha(2k-1)}\left(k - \frac{1}{n}\sum_{i=1}^{n} Z_i\right), k \neq \frac{1}{2}, \alpha \neq 0, \tag{9}$$

易知, 上式也是 p_{A_2} 的矩估计量.

3.4 模型性质与分析

由 (6) 式知, 该估计量的数学期望为

$$E(\widehat{p_{A_2}}) = \frac{1}{\alpha(2k-1)}\left(k - E\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\right)\right) = \frac{1}{\alpha(2k-1)}\left(k - E(Z_i)\right) = p_{A_2}$$

该估计量是无偏的, 其方差为

$$D(\widehat{p_{A_2}}) = \frac{1}{\alpha^2(2k-1)^2} D\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\right) = \frac{1}{\alpha^2(2k-1)^2 n} D(Z_i) \tag{10}$$

$$= \frac{p_{A_2}(1 - \alpha p_{A_2})}{\alpha n} + \frac{k(1-k)}{\alpha^2(2k-1)^2 n} \tag{11}$$

从上式可以看出, 估计量的方差 $D(\widehat{p_{A_2}})$ 由两部分组成, 第一部分是直接调查 (即 $k = 1$) 情况下的方差, 而第二部分是由于随机选答机制所带来的方差 [1].

3.5 模型评价

由 (5) 式, (11) 式知, 模型 I 与模型 II 估计量的方差都是两部分之和,

$$D(\widehat{p_{A_1}}) = \frac{p_{A1}(1 - \alpha p_{A_1})}{\alpha n} + \frac{(1-k)(1 - \alpha p_{A_1})}{\alpha^2 k n} = \frac{p_{A1}(1 - \alpha p_{A_1})}{\alpha n} + \frac{1-k}{\alpha^2 n} \cdot \frac{1 - \alpha p_{A_1}}{k}, \tag{12}$$

$$D(\widehat{p_{A_2}})=\frac{p_{A_2}(1-\alpha p_{A_2})}{\alpha n}+\frac{k(1-k)}{\alpha^2(2k-1)^2n},=\frac{p_{A_2}(1-\alpha p_{A_2})}{\alpha n}+\frac{1-k}{\alpha^2 n}\cdot\frac{k}{(2k-1)^2} \tag{13.}$$

第一部分都是对敏感问题直接调查下所产生的方差, 且相同, 下面来比较第二部分. 不妨假设模型 I 中第一套卡片所占的比例 k 与模型 II 中数字 "1" 所占的比例 k 相同,

1) 因为 $0<\alpha\le 1, 0<p_{A_1}<1$, 所以

$$\frac{1-\alpha p_{A_1}}{k}<\frac{1}{k}$$

如果 $\frac{1}{k}\le\frac{k}{(2k-1)^2}$, 即 $k^2\ge(2k-1)^2$, $(3k-1)(k-1)\le 0$, 解之 $\frac{1}{3}\le k\le 1$, 又依假设, 当 $\frac{1}{3}\le k\le 1$, 且 $k\ne\frac{1}{2}$ 时,

$$\frac{(1-k)(1-\alpha p_{A_1})}{\alpha^2 kn}<\frac{1-k}{\alpha^2 n}\cdot\frac{1}{k}\le\frac{1-k}{\alpha^2 n}\cdot\frac{k}{(2k-1)^2}$$

即

$$D(\widehat{p_{A_1}})<D(\widehat{p_{A_2}}). \tag{14}$$

上式表明, 此时模型 I 优于模型 II.

2) 因为 $0<p_{A_1}<1$, 所以

$$\frac{1-\alpha p_{A_1}}{k}>\frac{1-\alpha}{k}.$$

假设 $\alpha\ne 1$, 如果 $\frac{1-\alpha}{k}\ge\frac{k}{(2k-1)^2}$, 即 $1-\alpha\ge\frac{k^2}{(2k-1)^2}$, 也即

$$\sqrt{1-\alpha}\ge\frac{k}{|2k-1|}. \tag{15}$$

当 $\frac{1}{2}<k\le 1$ 时, 由上式知, $\sqrt{1-\alpha}\ge\frac{k}{2k-1}$, 解之,

$$k\ge\frac{\sqrt{1-\alpha}}{2\sqrt{1-\alpha}-1}.$$

而因为 $\frac{\sqrt{1-\alpha}}{2\sqrt{1-\alpha}-1}>\frac{\sqrt{1-\alpha}}{2\sqrt{1-\alpha}-\sqrt{1-\alpha}}=1$, 即 $k>1$(舍), 当 $0<k<\frac{1}{2}$ 时, 由 (15) 式知, $\sqrt{1-\alpha}\ge\frac{k}{1-2k}$, 解之,

$$k\le\frac{\sqrt{1-\alpha}}{2\sqrt{1-\alpha}+1}.$$

令 $f(x)=\frac{\sqrt{1-x}}{2\sqrt{1-x}+1}(0\le x<1)$, 则

$$f'(x)=\frac{d}{dx}\left(\frac{\sqrt{1-x}}{2\sqrt{1-x}+1}\right)=-\frac{1}{2\sqrt{1-x}\left(2\sqrt{1-x}+1\right)^2}<0$$

所以 $f(x)\downarrow$ 单调递减, 于是当 $0<\alpha<1$ 时, $f(\alpha)<f(0)=\frac{1}{3}$, 因此, 当 $0<k\le\frac{\sqrt{1-\alpha}}{2\sqrt{1-\alpha}+1}<\frac{1}{3}$ 时,

$$\frac{1-k}{\alpha^2 n}\cdot\frac{1-\alpha p_{A_1}}{k}>\frac{1-k}{\alpha^2 n}\cdot\frac{1-\alpha}{k}\ge\frac{1-k}{\alpha^2 n}\cdot\frac{k}{(2k-1)^2}$$

即

$$D(\widehat{p_{A_1}})>D(\widehat{p_{A_2}}). \tag{16}$$

上式表明, 此时模型 II 优于模型 I.

综上所述, 当 $0<\alpha\le 1$, $\frac{1}{3}\le k\le 1$, 且 $k\ne\frac{1}{2}$ 时, $D(\widehat{p_{A_1}})<D(\widehat{p_{A_2}})$.

当 $0<\alpha<1, 0<k<\frac{1}{3}$ 时, $D(\widehat{p_{A_1}})>D(\widehat{p_{A_2}})$.

参考文献

[1] 姜启源, 谢金星, 叶俊. 数学模型 (第四版)[M]. 北京, 高等教育出版社, 2011,317-324.
[2] 苏淳. 概率论 (第二版)[M]. 北京, 科学出版社, 2010,101-106.
[3] 李贤平, 沈崇圣, 陈子毅. 概率论与数理统计 [M]. 上海, 复旦大学出版社, 2003,322-324.
[4] Warner S L, Randomized response: a survey technique for eliminating evasive answer bias[J]. J American Stat Assoc, 1965, 60: 63-69.
[5] Greenberg, Bernard G, Abul-Ela, Abdel-Latif A, Simmons, Walt R, Horvitz, Daniel G. The unrelated question randomized response model: Theoretical framework[J]. J Amer Statist Assoc, (1969, 64: 520-539.
[6] Christofides, Tasos C. A generalized randomized response technique[J]. Metrika, 2003, 57: 195-200.

Design and Analysis for Questionnaires Investigation about Sensitive Social Problem

ZHANG Li-Zhuo, GUO Wei

(School of Statistics, University of International Business and Economics, Beijing 100029, China)

Abstract: This paper is concerned with the design for questionnaires investigation about sensitive problem by using probability theory and statistics methods. We construct two different mathematical models by taking drunker driving as an example. The comparison and analysis among them are presented.

Keywords: survey program; model; estimators; mathematical expectation; standard deviation