应用概率统计　第 35 卷
第 3 期　2019 年 6 月

Chinese Journal of Applied Probability and Statistics
Jun., 2019, Vol. 35, No. 3, pp. 221-232
doi: 10.3969/j.issn.1001-4268.2019.03.001

基于超总体伪设计与组合样本的候选者数据库网络调查的推断研究*

刘　展*　　潘莹丽

(湖北大学数学与统计学学院, 应用数学湖北省重点实验室, 武汉, 430062)

摘　要: 候选者数据库网络调查的推断问题是网络调查发展中迫切需要解决的问题. 基于此, 提出基于超总体伪设计与组合样本的非概率抽样推断方法: 对网络候选者数据库的调查样本建立超总体模型来构造伪权数, 并根据网络候选者数据库的调查样本和概率样本的组合样本计算总体均值的估计, 最后根据超总体模型的方差估计理论推导出目标总体均值估计的方差估计式, 同时采用 Bootstrap 与 Jackknife 方法来估计总体均值估计的方差, 并比较不同方差估计方法的效果. 研究结果表明: 基于超总体伪设计与组合样本的总体均值估计效率高于仅使用概率样本的估计和仅使用网络候选者数据库的调查样本加权的估计, 估计效果较好; 方差估计方面, 采用 VM1、VM2 与 VM3 方法计算的方差估计相比而言更好.

关键词: 超总体; 伪设计; 组合样本; 网络候选者数据库; 非概率抽样

中图分类号: C811

英文引用格式: LIU Z, PAN Y L. Research on inference of candidate database web surveys based on superpopulation pseudo design and the combined sample [J]. Chinese J Appl Probab Statist, 2019, 35(3): 221–232. (in Chinese)

§1. 引　　言

候选者数据库网络调查, 作为网络调查的一种方法, 具有成本低廉、方便快捷等特点, 已经被广泛的应用于市场调查、政府调研、民意调查以及学术研究之中, 受到人们广泛的关注与重视. 网络调查的候选者数据库, 简称网络候选者数据库, 也称为网络访问固定样本, 就是愿意完成网络调查的网络访问 (上网) 人群, 这就意味着存在一个潜在的受访者的样本数据库, 在未来的数据收集中, 如果他们被选择为调查对象, 他们将愿意配合完成调查[1]. 候选者数据库网络调查就是从网络候选者数据库中抽取样本进行网络调查, 其获得的调查样本称为网络候选者数据库的调查样本. 根据候选者数据库网络调查的涵义, 可知候选者数据库网络调查为非概率抽样调查, 网络候选者数据库的调查样本为非概率样本, 其入样概率未知、权数未知, 无法使用传统的抽样推断理论进行统计推断, 由此阻碍了数

*国家社会科学基金项目 (批准号: 18BTJ022) 资助.

*通讯作者, E-mail: eleen_20040109@163.com.

本文 2017 年 7 月 28 日收到, 2018 年 2 月 12 日收到修改稿.

据的进一步开发和利用, 不利于网络调查的发展. 因此, 如何解决非概率抽样的统计推断问题, 特别是候选者数据库网络调查的统计推断问题, 是网络调查发展的迫切需求.

目前国外已有一些研究者对候选者数据库网络调查的非概率抽样统计推断进行了研究, 而国内关于这方面的研究很少. Lee[2] 将倾向得分调整与校准调整相结合, 试图替代传统的事后调整, 减少非随机样本选择和候选者数据库网络调查覆盖不全所导致的偏差. Lee 与 Valliant[3]提出首先建立倾向得分模型进行倾向得分调整, 然后进行校准调整获得最终的调整权数, 最后利用该调整权数对网络候选者数据库的调查样本进行加权调整来估计总体. Valliant 和 Dever[4] 结合网络候选者数据库与参考样本, 建立倾向模型来估计成为一名网络调查志愿者的概率即倾向得分, 最后利用估计的倾向得分进行加权调整来估计总体. 以上研究只是在估计过程中使用了相关的概率样本, 最终的估计只使用了候选者数据库网络调查的数据, 概率样本的信息利用不足. Elliott 和 Haviland[5] 将概率样本与非概率样本结合构造出一个组合估计, 其正好是两个样本均值的加权平均, 并使用两个样本的数据估计总体. 然而, 他们所提出的组合估计与剩余标准偏差有关, 而要保证剩余标准偏差估计的精度就需要概率样本的样本量比较大, 这就意味着当概率样本的样本量比较小时, 他们所提出的估计精度不高. Elliott[6] 进一步提出将非概率样本与概率样本结合, 根据贝叶斯定理对非概率样本构造伪权数, 并利用两个样本数据共同估计总体, 他所提出的方法相关统计量的偏差与均方误差有所减少, 估计精度有所提高. 刘展与金勇进[7] 提出将网络访问固定样本的调查样本与概率样本结合, 利用倾向得分逆加权和加权组调整构造伪权数, 并利用两个样本数据来估计总体, 研究结果表明无论概率样本的样本量大小, 他们所提出的总体均值估计方法效果较好, 估计精度较高. 然而, 以上所提出的伪权数构造均受到概率样本抽取的影响. 基于此, 本研究提出基于超总体伪设计与组合样本的候选者数据库网络调查的推断方法, 即从超总体的角度出发, 对网络候选者数据库的调查样本 (非概率样本) 建立超总体模型来构造伪权数, 并利用非概率样本与概率样本的数据共同估计总体, 这也是国内外相关研究文献未曾涉及的一个方面. 本文所提出的伪权数构造不会受到概率样本抽取的影响, 构造过程简单且可操作性较强.

§2. 基于超总体的伪权数构造

在超总体模型方法中, 感兴趣的变量被认为是一个服从某一分布的随机变量, 即总体取值是随机的, 并假定有限总体为某一个超总体的一次随机实现, 即假定总体本身为超总体模型中的一个样本. 在总体取值既定的情况下抽取样本 (不一定是随机抽取的), 并根据抽取的样本建立一个统计模型拟合目标变量 Y, 最后根据 Y 的分布即 $f(Y \mid X;\Theta)$, 其中 X 为协向量, Θ 为参数, 由样本推断总体. 超总体模型方法的随机性来源于变量本身而非抽样过程, 推断并不要求抽取单元的随机性, 但需要对模型进行假定. 在一定的模型假定下, 根据联合分布探索抽中的样本单元与未被抽中的样本单元之间的关系, 建立超总体模型, 再通过收集的样本数据来估计 (预测) 未被抽中的样本单元, 从而得到总体的估计. 一般情况

下, 保证所建立的超总体模型绝对正确是很难把握的, 需要根据获得的样本数据对模型进行检验, 只有当模型检验通过的情况下, 才可以使用该模型.

超总体模型方法的样本选择 (抽样) 机制类似于缺失数据的缺失机制, 也可分为可忽略的选择机制和不可忽略的选择机制. 如果 $f(\delta_S \mid Y, X; \Phi) = f(\delta_S \mid X; \Phi)$, 即总体单元进入样本 S 的概率不依赖于 Y, 则为可忽略的选择机制, 否则为不可忽略的选择机制. 其中, δ_S 为总体单元是否在样本 S 中的示性变量, Φ 为参数. 如果是可忽略的样本选择机制, 可以根据样本数据直接建立 Y 与 X 的模型, 并根据该模型来推断总体. 如果是不可忽略的样本选择机制, 则样本的选择不仅与 X 有关还可能与要估计的 Y 有关, 此时建立模型相对来说较为困难. 对于概率样本, 由于可进行抽样设计来获得, 其随机分布受抽样者的控制, 所以往往可以通过抽样设计达到可忽略的样本选择机制. 对于非概率抽样, 有一些目的性的非概率样本是可忽略的样本选择机制. 比如, 美国能源信息管理局抽取了具有较大 x 值的 n 个单元组成的非概率样本, 以及 Royall 提出的基于协变量总体矩 (如均值、方差) 的平衡抽样, 都属于可忽略的样本选择机制. 然而, 样本单元选择并没有受到很好控制的一些非概率样本可能就不是可忽略的样本选择机制, 此时可以将基于倾向得分伪设计的方法和超总体方法结合. 本文假定非概率样本的选择为可忽略的样本选择机制.

超总体模型方法中样本单元与非样本单元 (总体中不在样本中的单元) 的模型形式可能相同也可能不同. 目标变量 Y 可以根据样本单元和非样本单元划分为两个部分, 即 $Y = (Y_S, Y_{\overline{S}})$, 其中 $\overline{S}$ 表示不在样本中的单元集合, 则有 $f(Y \mid X; \Theta) = f(Y_S \mid Y_{\overline{S}}, X; \Theta) \cdot f(Y_{\overline{S}} \mid X; \Theta)$. 如果 $f(Y_S \mid Y_{\overline{S}}, X; \Theta) = f(Y_S \mid X; \Theta)$, 则 Y_S 与 $Y_{\overline{S}}$ 在协向量 X 的条件下是独立的, 此时有

$$f(Y \mid X; \Theta) = f(Y_S \mid X; \Theta) f(Y_{\overline{S}} \mid X; \Theta). \tag{1}$$

如果基于模型的推断目标是参数 Θ, 可仅基于 $f(Y_S \mid X; \Theta)$ 即基于样本单元建立模型来估计参数. 然而, 如果是对整个的总体 Y 进行推断, 则需要估计 $f(Y_{\overline{S}} \mid X; \Theta)$, 即根据非样本单元建立模型来估计 $Y_{\overline{S}}$, 从而得到总体 Y 的估计 $\widehat{Y} = Y_S + \widehat{Y}_{\overline{S}}$. 如果 $f(Y_{\overline{S}} \mid X; \Theta)$ 的模型形式与 $f(Y_S \mid X; \Theta)$ 的模型形式相同, 则根据 S 样本单元建立的模型就可以直接用于预测非样本单元上的 $Y_{\overline{S}}$. 如果 $f(Y_{\overline{S}} \mid X; \Theta)$ 的模型形式与 $f(Y_S \mid X; \Theta)$ 的模型形式不相同, 则还需要根据非样本单元建立模型来估计 $Y_{\overline{S}}$, 然而非样本单元是不在样本中的单元, 是未知的, 难以构建模型估计, 也就难以推断整个总体了. 本文假定 $f(Y_{\overline{S}} \mid X; \Theta)$ 的模型形式与 $f(Y_S \mid X; \Theta)$ 相同. 在这种假定下, 样本单元与非样本单元都服从同样的模型, 模型的参数可由样本单元数据来估计, 然后利用该模型对非样本单元进行预测, 最终实现总体的估计.

在以上假定下, 如果有一个网络候选者数据库的调查样本 S_1 和一个相关概率样本 S_2, 其中 S_1 为从网络候选者数据库中随机抽取的样本, X 为两个样本共同的协向量. S_1 本质上为非概率样本, 其单元的选择概率未知、权数未知, 需要对 S_1 单元构造权数, 因为该权数为构造的, 不是真实的, 不妨称其为伪权数. 现在考虑如何建立超总体模型来为网络候选者数据库的调查样本 S_1 (样本量为 n_1) 单元构造伪权数. 假设总体第 i 个单元的目标变量为 Y_i, 其 p 维的协向量为 $X_i = (X_{i1}, X_{i2}, \ldots, X_{ip})$, $i = 1, 2, \ldots, N$, N 为总体单元个数. 协

向量的总体总量为 $T_U=(T_{X1},T_{X2},\ldots,T_{Xp})'$, 其中 $T_{Xj}=\sum\limits_{i=1}^{N}X_{ij}$, $j=1,2,\ldots,p$. 可对 Y_i 的均值建立一个关于 X_i 的线性回归模型如下:

$$E_M(Y_i)=X_i\beta, \tag{2}$$

$$V_M(Y_i)=v_i, \tag{3}$$

其中 $\beta=(\beta_1,\beta_2,\ldots,\beta_p)'$ 为回归系数, v_i 为方差参数, 下标 M 意味着关于模型的期望. 对于网络候选者数据库的调查样本 S_1, 样本量为 n_1, 假设其协向量的总体总量 T_U 已知, 可建立上述超总体回归模型, 回归系数的估计为 $\widehat{\beta}=(X'_{S_1}X_{S_1})^{-1}X'_{S_1}Y_{S_1}$, X_{S_1} 为样本 S_1 单元的 $n_1\times p$ 维的协变量矩阵, Y_{S_1} 为样本 S_1 中的 n_1 维的 Y 向量. 由超总体回归模型即可预测不在样本 S_1 中单元的 Y 值, 即 $\widehat{Y}_i=X_i\widehat{\beta}$, 从而可得到总体总量的估计为

$$\begin{aligned}\widehat{T}&=\sum_{i\in S_1}Y_i+\sum_{i\notin S_1}\widehat{Y}_i\\&=\sum_{i\in S_1}Y_i+\sum_{i\notin S_1}(X_i\widehat{\beta})=\sum_{i\in S_1}Y_i+(T_U-T_{S_1})'\widehat{\beta}\\&=\sum_{i\in S_1}Y_i+(T_U-T_{S_1})'(X'_{S_1}X_{S_1})^{-1}X'_{S_1}Y_{S_1}\\&=\sum_{i\in S_1}Y_i+(T_U-T_{S_1})'(X'_{S_1}X_{S_1})^{-1}\sum_{i\in S_1}(X_iY_i)\\&=\sum_{i\in S_1}[1+(T_U-T_{S_1})'(X'_{S_1}X_{S_1})^{-1}X_i]Y_i=\sum_{i\in S_1}w_{1i}Y_i,\end{aligned} \tag{4}$$

其中 $w_{1i}=1+(T_U-T_{S_1})'(X'_{S_1}X_{S_1})^{-1}X_i$, T_{S_1} 为样本 S_1 中 X 的总量 (和). (4) 式中的总体总量可视为 Y_i 的加权和, 单元 i 的权数为 w_{1i}, 即构造出了样本 S_1 单元的伪权数. 在基于超总体方法构造出网络候选者数据库的调查样本 S_1 单元的伪权数之后, 将网络候选者数据库的调查样本 S_1 与概率样本 S_2 结合, 对网络候选者数据库的调查样本的伪权数和概率样本的基础权数进行标准化[7], 得到最终的组合样本单元的权数. 以上伪权数是通过超总体的方法构造的, 可称之为超总体伪设计方法. 需要注意的是, 超总体线性回归模型拟合的越好, X 对 Y 的解释程度越高, 最后的估计效果也越好.

§3. 总体均值估计与方差估计

在基于超总体方法构造的网络候选者数据库的调查样本伪权数 $w_{1i}=1+(T_U-T_{S_1})'\cdot(X'_{S_1}X_{S_1})^{-1}X_i$ 的基础上, 考虑总体规模 N 已知的情况. 根据非概率样本 S_1 的伪权数 w_{1i} 与另一个概率样本 S_2 的基础权数 w_{2j}, 再结合组合样本数据, 最终可得到总体均值的估计为 $\widehat{\overline{Y}}=n_1\widehat{\overline{Y}}_1/(n_1+n_2)+n_2\widehat{\overline{Y}}_2/(n_1+n_2)$ (推导方式见文献 [7]), 其中 $\widehat{\overline{Y}}_1=\sum\limits_{i=1}^{n_1}(w_{1i}Y_{1i})/N$, $\widehat{\overline{Y}}_2=\sum\limits_{j=1}^{n_2}(w_{2j}Y_{2j})/N$, Y_{1i}、Y_{2j} 分别为 S_1 与 S_2 的目标变量值, n_1、n_2 分别为 S_1 与 S_2 的样

本量. 因为总体规模 N 已知, 此处 $\widehat{\overline{Y}}_1$ 与 $\widehat{\overline{Y}}_2$ 的形式实际上是用 N 替换总体规模未知时的分母 $\sum\limits_{j=1}^{n_2} w_{2j}$ 或 $\sum\limits_{i=1}^{n_1} w_{1i}$ 而得到.

对于总体均值估计的方差估计, 由于非概率样本 S_1 与概率样本 S_2 是分别独立抽取的, 所以有

$$\mathsf{Var}\,(\widehat{\overline{Y}}) = \Big(\frac{n_1}{n_1+n_2}\Big)^2 \cdot \mathsf{Var}\,(\widehat{\overline{Y}}_1) + \Big(\frac{n_2}{n_1+n_2}\Big)^2 \cdot \mathsf{Var}\,(\widehat{\overline{Y}}_2). \tag{5}$$

$\mathsf{Var}(\widehat{\overline{Y}}_1)$ 为基于超总体模型估计的方差, 可根据超总体模型推导求出. 具体地, 对于 $\mathsf{Var}(\widehat{\overline{Y}}_1)$, 根据 Valliant 等人 ([8; Chap. 5]) 描述的方法, 可得

$$\begin{aligned}\widehat{T} - T &= \sum_{i\in S_1} w_{1i}Y_i - \sum_{i\in U} Y_i = \sum_{i\in S_1} w_{1i}Y_i - \Big(\sum_{i\in S_1} Y_i + \sum_{i\in U-S_1} Y_i\Big) \\ &= \sum_{i\in S_1} (w_{1i}-1)Y_i - \sum_{i\in U-S_1} Y_i = \sum_{i\in S_1} a_iY_i - \sum_{i\in \overline{S}_1} Y_i,\end{aligned} \tag{6}$$

其中 T 为 Y 的总体总量, $a_i = w_{1i} - 1$, $\overline{S}_1 = U - S_1$. 由此可得总体总量估计 $\widehat{T}$ 的预测方差为

$$V_M(\widehat{T} - T) = \sum_{i\in S_1} a_i^2 v_i + \sum_{i\in \overline{S}_1} v_i. \tag{7}$$

由于 $a_i = O(N/n_1)$, $\sum\limits_{i\in S_1} a_i^2 v_i = n_1 \cdot O(N^2/n_1^2) = O(N^2/n_1)$, $\sum\limits_{i\in \overline{S}_1} v_i = O(N - n_1)$, 而 $N^2/n_1 \cdot [1/(N-n_1)] \approx N/n_1$, 所以当总体的抽样比较小时, $\sum\limits_{i\in \overline{S}_1} v_i$ 相对于 $\sum\limits_{i\in S_1} a_i^2 v_i$ 就比较小, 可以忽略不计, 即 $V_M(\widehat{T} - T) \approx \sum\limits_{i\in S_1} a_i^2 v_i$, 从而可得到总体总量估计的方差估计为 $\widehat{V}_M(\widehat{T}) \approx \sum\limits_{i\in S_1} a_i^2 \widehat{v}_i$, 进一步可得到总体均值估计的方差估计为 $\widehat{V}_M(\widehat{\overline{Y}}_1) \approx N^{-2} \sum\limits_{i\in S_1} a_i^2 \widehat{v}_i$. 其中 $\widehat{v}_i$ 可根据残差 $e_i = Y_i - X_i\widehat{\beta}$ 来估计, 主要有三种方法: (i) $\widehat{v}_i = e_i^2$; (ii) $\widehat{v}_i = e_i^2/(1-h_{ii})$; (iii) $\widehat{v}_i = [e_i/(1-h_{ii})]^2$, 其中 h_{ii} 为单元 i 的杆杠值, 为帽子矩阵 $H = X_{S_1}(X_{S_1}'X_{S_1})^{-1}X_{S_1}'$ 的对角线上的元素. 由上可计算出 $\mathsf{Var}\,(\widehat{\overline{Y}}_1)$ 的估计, 即 $\widehat{V}_M(\widehat{\overline{Y}}_1)$.

对于 $\mathsf{Var}\,(\widehat{\overline{Y}}_2)$, 由于 S_2 为概率样本, $\mathsf{Var}\,(\widehat{\overline{Y}}_2)$ 的估计易求. 在计算出 $\mathsf{Var}\,(\widehat{\overline{Y}}_1)$ 的估计 $\widehat{V}_M(\widehat{\overline{Y}}_1)$ 和 $\mathsf{Var}\,(\widehat{\overline{Y}}_2)$ 的估计 $\widehat{V}(\widehat{\overline{Y}}_2)$ 之后就可得到总体均值估计的方差估计为

$$\widehat{V}(\widehat{\overline{Y}}) = \Big(\frac{n_1}{n_1+n_2}\Big)^2 \cdot \widehat{V}_M(\widehat{\overline{Y}}_1) + \Big(\frac{n_2}{n_1+n_2}\Big)^2 \cdot \widehat{V}(\widehat{\overline{Y}}_2). \tag{8}$$

如果样本 S_2 为不放回的简单随机样本, 则 $\widehat{\overline{Y}}_2$ 的方差估计为 $\widehat{V}(\widehat{\overline{Y}}_2) = (1-f)s^2/n_2$, 其中 $f = n_2/N$ 为样本 S_2 的抽样比, s^2 为样本 S_2 的样本方差. 将 $\widehat{V}_M(\widehat{\overline{Y}}_1)$ 与 $\widehat{V}(\widehat{\overline{Y}}_2)$ 的具体表达式带入 (8) 式, 得到总体均值估计的方差估计为

$$\begin{aligned}\widehat{V}(\widehat{\overline{Y}}) &= \Big(\frac{n_1}{n_1+n_2}\Big)^2 \cdot \widehat{V}_M(\widehat{\overline{Y}}_1) + \Big(\frac{n_2}{n_1+n_2}\Big)^2 \cdot \widehat{V}(\widehat{\overline{Y}}_2) \\ &\approx \Big(\frac{n_1}{n_1+n_2}\Big)^2 \cdot \frac{\sum\limits_{i\in S_1} a_i^2 \widehat{v}_i}{N^2} + \Big(\frac{n_2}{n_1+n_2}\Big)^2 \cdot \frac{1-n_2/N}{n_2} s^2\end{aligned}$$

$$= \frac{1}{(n_1+n_2)^2N^2}\Big[n_1^2\sum_{i\in S_1}a_i^2\widehat{v}_i + Nn_2(N-n_2)s^2\Big]. \tag{9}$$

在这里, 将采用 $\widehat{v}_i = e_i^2$、$\widehat{v}_i = e_i^2/(1-h_{ii})$、$\widehat{v}_i = [e_i/(1-h_{ii})]^2$, 并根据 (9) 式计算方差估计 $\widehat{V}(\widehat{\overline{Y}})$ 的方法分别简记为 VM1、VM2 与 VM3 方法.

至此已由总体均值估计 $\widehat{\overline{Y}}$ 的表达式 $\widehat{\overline{Y}} = n_1\widehat{\overline{Y}}_1/(n_1+n_2) + n_2\widehat{\overline{Y}}_2/(n_1+n_2)$ 直接推导出了方差估计式. 此时可能会产生一个疑问, 倾向得分伪设计[7]与超总体伪设计方法得到的总体均值估计的表达式基本相同, 只是 $\widehat{\overline{Y}}_1$ 中的伪权数不同, 那么倾向得分伪设计方法中是否也可通过 $\widehat{\overline{Y}}$ 的表达式直接推导出方差估计式? 可以看到, 若通过 $\widehat{\overline{Y}}$ 的表达式推导方差估计, 需要计算 $\widehat{V}(\widehat{\overline{Y}}_1)$ 和 $\widehat{V}(\widehat{\overline{Y}}_2)$, 由于 S_2 为概率样本, 所以两种方法中 $\widehat{V}(\widehat{\overline{Y}}_2)$ 都容易计算, 关键是 $\widehat{V}(\widehat{\overline{Y}}_1)$ 的计算. 一方面, 由于倾向得分伪设计中伪权数的构造过程相对于超总体方法来说更为复杂, 伪权数的组成成分更多更杂, 简单地根据随机抽样推断理论计算出的方差估计 $\widehat{V}(\widehat{\overline{Y}}_1)$ 可能并不能反映出伪权数构造的复杂过程, 也不能解释非概率样本中伪权数估计的抽样变动性, 估计效果可能不好. 另一方面, 超总体伪设计方法中伪权数的构造是在非概率样本 S_1 上建立超总体模型来构造的, 并不依赖基于抽样设计抽取的概率样本 S_2 (只是最终总体估计时使用到 S_2 的数据), 对抽样的变动不会很敏感, 而且基于模型的推断已有一定的理论基础, 可以直接根据模型推断理论计算出 $\widehat{\overline{Y}}_1$ 的方差估计 $\widehat{V}(\widehat{\overline{Y}}_1)$. 因此, 倾向得分伪设计并不适合通过 $\widehat{\overline{Y}}$ 的表达式直接推导出方差估计式. 此外, 为了进一步探索方差估计的效果, 在模拟和实证研究中同时采用 Bootstrap 和 Jackknife 方法来计算方差估计.

§4. 模拟研究

为了研究所提出的估计方法的效果, 现进行模拟研究. 在模拟中, 生成一个规模为 $N = 10\,000$ 的有限总体, 这个有限总体是从一个无限总体中产生的, 该无限总体的协变量与目标变量的分布分别为

$$X_i \sim \exp(5.5); \qquad Y_i = 4.5X_i + e_i; \qquad e_i \sim \mathrm{N}(1.5, 4.5),$$

协变量 X_i 被用于超总体模型的建立. 从有限总体中不放回简单随机抽取 $M = 5\,000$ 的样本, 视其为网络候选者数据库 V. 在总体和网络候选者数据库固定的情况下, 分别从总体和网络候选者数据库中不放回简单随机抽取 $n_2 = 100, 200, 500, 1\,000, 1\,500$ 的一个概率样本 S_2 和对应的 $n_1 = 150, 300, 750, 1\,500, 1\,800$ 的一个样本 S_1 (网络候选者数据库的调查样本), 重复抽取 S_1 和 S_2 $1\,000$ 次 ($B = 1\,000$), 得到 $1\,000$ 组蒙特卡罗样本. 在每组蒙特卡罗样本上, 对网络候选者数据库的调查样本 S_1 建立超总体回归模型, 并计算出 S_1 样本单元的权数 $w_{1i} = 1 + (T_U - T_{S_1})'(X'_{S_1}X_{S_1})^{-1}X_i$. 由于此处建立的超总体模型为一元线性回归模型, 故 S_1 中第 i 个单元的权数为 $w_{1i} = 1 + \big(\sum_{i=1}^{N}X_i - \sum_{i=1}^{n_1}X_i\big)\big(\sum_{i=1}^{n_1}X_i^2\big)^{-1}X_i$. 最后, 将

S_1 与 S_2 结合, 并利用组合样本数据估计总体均值 $\widehat{\overline{Y}} = n_1\widehat{\overline{Y}}_1/(n_1+n_2) + n_2\widehat{\overline{Y}}_2/(n_1+n_2)$, 进一步计算 1 000 组蒙特卡罗样本上总体均值估计 $\widehat{\overline{Y}}$ 的均值、方差、相对偏差 (relative bias, 简记为 RB) 与均方误差. 相对偏差为: $\mathrm{RB}(\%) = 100 \times (\overline{\widehat{\theta}} - \theta)/\theta$, 其中 θ 为要估计的真实参数, $\widehat{\theta}$ 为 θ 的一个估计, $\overline{\widehat{\theta}}$ 为 1 000 个样本上 $\widehat{\theta}$ 的均值, 相对偏差越小, 估计的效果越好. 为了便于比较, 同时报告 $\widehat{\overline{Y}}_1$ 与 $\widehat{\overline{Y}}_2$ 的结果, 见表 1.

表 1　基于超总体伪设计与组合样本的总体均值估计的模拟结果

		均值	方差	相对偏差 (%)	均方误差
$n_1 = 150$, $n_2 = 100$	$\widehat{\overline{Y}}$	0.80706	0.01165	0.24671	0.01164
	$\widehat{\overline{Y}}_1$	0.81406	0.01360	1.11581	0.01366
	$\widehat{\overline{Y}}_2$	0.79657	0.04466	-1.05694	0.04468
$n_1 = 300$, $n_2 = 200$	$\widehat{\overline{Y}}$	0.80959	0.00531	0.56006	0.00532
	$\widehat{\overline{Y}}_1$	0.81127	0.00645	0.76882	0.00648
	$\widehat{\overline{Y}}_2$	0.80706	0.02031	0.24692	0.02029
$n_1 = 750$, $n_2 = 500$	$\widehat{\overline{Y}}$	0.80942	0.00223	0.53917	0.00224
	$\widehat{\overline{Y}}_1$	0.81227	0.00220	0.89400	0.00225
	$\widehat{\overline{Y}}_2$	0.80513	0.00884	0.00694	0.00883
$n_1 = 1\,500$, $n_2 = 1\,000$	$\widehat{\overline{Y}}$	0.80761	0.00104	0.31468	0.00104
	$\widehat{\overline{Y}}_1$	0.80930	0.00097	0.52451	0.00099
	$\widehat{\overline{Y}}_2$	0.80508	0.00422	-0.00007	0.00422
$n_1 = 1\,800$, $n_2 = 1\,500$	$\widehat{\overline{Y}}$	0.80691	0.00075	0.22714	0.00075
	$\widehat{\overline{Y}}_1$	0.81078	0.00070	0.70889	0.00073
	$\widehat{\overline{Y}}_2$	0.80225	0.00247	-0.35096	0.00248

由表 1 可以看到无论样本量大小, 仅利用非概率样本 S_1 或概率样本 S_2 以及结合 S_1 与 S_2 得到的总体均值估计均在 0.8 附近, 相差不大. 在方差上, 仅利用 S_2 的总体均值估计的方差在不同的样本量情况下都是最大的, 并且 $\widehat{\overline{Y}}$、$\widehat{\overline{Y}}_1$、$\widehat{\overline{Y}}_2$ 的方差均随着样本量的增大而减小. 从相对偏差来看, 当 $n_1 = 150$, $n_2 = 100$ 和 $n_1 = 1\,800$, $n_2 = 1\,500$ 时, $\widehat{\overline{Y}}$ 的相对偏差绝对值均最小, $\widehat{\overline{Y}}_1$ 的相对偏差绝对值均最大; 其他样本量情况下, $\widehat{\overline{Y}}$ 的相对偏差绝对值均处于 $\widehat{\overline{Y}}_1$ 和 $\widehat{\overline{Y}}_2$ 之间, $\widehat{\overline{Y}}_1$ 的相对偏差绝对值都是最大的. 总的来看, 无论样本量大小, $\widehat{\overline{Y}}$、$\widehat{\overline{Y}}_1$、$\widehat{\overline{Y}}_2$ 的相对偏差绝对值均小于 2%, 偏差较小, 并且随着样本量的增大, $\widehat{\overline{Y}}$ 的相对偏差绝对值基本呈下降趋势. 在均方误差上, 当 $n_1 = 150$, $n_2 = 100$; $n_1 = 300$, $n_2 = 200$; $n_1 = 750$, $n_2 = 500$ 时, $\widehat{\overline{Y}}$ 的均方误差都是最小的; 当$n_1 = 1\,500$, $n_2 = 1\,000$ 和 $n_1 = 1\,800$, $n_2 = 1\,500$ 时, $\widehat{\overline{Y}}_1$ 的均方误差均最小, 次小的是 $\widehat{\overline{Y}}$, 其与 $\widehat{\overline{Y}}_1$ 的均方误差相差不到 0.0001, 几乎相同; 无论样本量大小, $\widehat{\overline{Y}}_2$ 的均方误差都是最大的, 说明基于超总体伪设计和组合样本的总体均值估计的效率较高, 仅利用概率样本的总体均值估计的效率最低. 此外, 随着样

本量增大, $\widehat{\overline{Y}}$、$\widehat{\overline{Y}}_1$、$\widehat{\overline{Y}}_2$ 的均方误差均逐渐递减, 表明随着样本量增大三个估计量的估计效率均逐渐提高. 从相对偏差和均方误差综合来看, 基于超总体伪设计和组合样本的总体均值估计效果较好.

除了点估计, 同时考虑方差估计. 在每组蒙特卡罗样本上, 根据 (9) 式计算出总体均值估计的方差估计, 不妨记为 $\widehat{V}_M(\widehat{\overline{Y}})$, 其中 (9) 式中的 $\widehat{v}_i$ 采用三种方法: $\widehat{v}_i = e_i^2$、$\widehat{v}_i = e_i^2/(1-h_{ii})$ 和 $\widehat{v}_i = [e_i/(1-h_{ii})]^2$ 来计算, 这里 $h_{ii} = X_i^2 \big/ \sum\limits_{i=1}^{n_1} X_i^2$. 同时, 采用 Bootstrap 和 Jackknife 方法来计算方差估计. 利用 Bootstrap 方法估计方差时, 为了避免重复抽样时只抽取到 S_1 或 S_2 的样本, 分别在 S_1、S_2 中不放回简单随机抽取样本量为 $n_1/2$ 和 $n_2/2$ 的两个样本组合为一个 Bootstrap 样本, 然后放回去之后再使用同样的方法重新抽取一个 Bootstrap 样本, 共抽取 100 次, 得到 100 个 Bootstrap 样本, 在这 100 个 Bootstrap 样本上计算方差估计. 利用 Jackknife 方法估计方差时, 将样本 $S = S_1 \cup S_2$ 划分成 5 个随机组. Bootstrap 和 Jackknife 方法计算得到的方差估计分别记为 $\widehat{V}_{bs}(\widehat{\overline{Y}})$ 和 $\widehat{V}_{jk}(\widehat{\overline{Y}})$. 最后, 在 1 000 组蒙特卡罗样本上计算方差估计的标准误差 (standard error, 简记为 SE), 相对偏差与 95% 置信区间覆盖率 (coverage rates of 95 percent confidence interval, 简记为 CR). 方差估计的标准误差是蒙特卡罗样本得到的方差估计的标准差除以蒙特卡罗样本个数的平方根, 即 $\mathrm{SE} = S/\sqrt{B}$, 其中 $S = \sqrt{\sum\limits_{b=1}^{B}[\widehat{V}_b(\widehat{\theta}) - \overline{\widehat{V}}(\widehat{\theta})]^2/(B-1)}$, $\overline{\widehat{V}}(\widehat{\theta}) = \sum\limits_{b=1}^{B}\widehat{V}_b(\widehat{\theta})/B$ 分别为 1 000 个样本上方差估计的标准差与均值, 标准误差越小, 方差估计的可靠度越大. 方差估计的相对偏差为: $\mathrm{RB}(\%) = 100 \times [\widehat{V}(\widehat{\theta}) - \mathrm{MSE}(\widehat{\theta})]/\mathrm{MSE}(\widehat{\theta})$, 其中 $\mathrm{MSE}(\widehat{\theta})$ 为 1 000 个样本上目标总体均值估计的均方误差, 相对偏差越小, 方差估计的效果越好. 方差估计的 95% 置信区间覆盖率为 1 000 个样本中满足 $|\widehat{\theta}-\theta|\big/\sqrt{\widehat{V}(\widehat{\theta})} \leqslant 1.96$ 的样本个数所占的百分比, 覆盖率越接近于 95%, 方差估计的效果越好. 最终计算结果见表 2, 这里 $\widehat{V}_{M1}(\widehat{\overline{Y}})$、$\widehat{V}_{M2}(\widehat{\overline{Y}})$、$\widehat{V}_{M3}(\widehat{\overline{Y}})$ 分别表示采用 $\widehat{v}_i = e_i^2$、$\widehat{v}_i = e_i^2/(1-h_{ii})$ 和 $\widehat{v}_i = [e_i/(1-h_{ii})]^2$ 计算的方差估计.

由表 2 可见, 无论样本量大小, 标准误差最低的是采用 VM1 方法得到的方差估计, 最高的是 Jackknife 方差估计, 总的来看五个方差估计的标准误差均非常接近于 0, 可靠度均比较高, 且均随着样本量的增大而减小. 从相对偏差来看, 当 $n_1 = 150$, $n_2 = 100$ 时, 五个方差估计的相对偏差绝对值均小于 10%, 其中 $\widehat{V}_{M1}(\widehat{\overline{Y}})$ 的相对偏差绝对值最小, $\widehat{V}_{jk}(\widehat{\overline{Y}})$ 的相对偏差绝对值最大; 当 $n_1 = 300$, $n_2 = 200$ 时, 仅有 $\widehat{V}_{M1}(\widehat{\overline{Y}})$ 的相对偏差绝对值低于 10%, 其他方差估计的相对偏差绝对值均高于 10%; 当 $n_1 = 750$, $n_2 = 500$ 时, $\widehat{V}_{jk}(\widehat{\overline{Y}})$ 的相对偏差绝对值大于 10%, 其他方差估计的相对偏差绝对值均小于 10%, 且 $\widehat{V}_{M1}(\widehat{\overline{Y}})$、$\widehat{V}_{M2}(\widehat{\overline{Y}})$ 和 $\widehat{V}_{M3}(\widehat{\overline{Y}})$ 的相对偏差绝对值不到 2%; 当 $n_1 = 1\,500$, $n_2 = 1\,000$ 和 $n_1 = 1\,800$, $n_2 = 1\,500$ 时, $\widehat{V}_{M1}(\widehat{\overline{Y}})$、$\widehat{V}_{M2}(\widehat{\overline{Y}})$ 和 $\widehat{V}_{M3}(\widehat{\overline{Y}})$ 的相对偏差绝对值均小于 4%, $\widehat{V}_{bs}(\widehat{\overline{Y}})$ 和 $\widehat{V}_{jk}(\widehat{\overline{Y}})$ 的相对偏差绝对值均大于 10%, 偏差较大. 总的来看, 无论样本量大小, $\widehat{V}_{M1}(\widehat{\overline{Y}})$ 的相对偏差绝对值均低于 10%, 偏差较小. 在 95% 置信区间覆盖率上, 当 $n_1 = 150$, $n_2 = 100$; $n_1 = 750$, $n_2 = 500$ 和 $n_1 = 1\,500$, $n_2 = 1\,000$ 时, 覆盖率最高的均为 Bootstrap 方差估计, 且比较接近 95%, 近似于正态覆盖, 其次是 $\widehat{V}_{M1}(\widehat{\overline{Y}})$、$\widehat{V}_{M2}(\widehat{\overline{Y}})$ 和 $\widehat{V}_{M3}(\widehat{\overline{Y}})$, 覆盖率最低的都是 Jackknife

表 2　基于超总体伪设计与组合样本的方差估计的模拟结果

		$\widehat{V}_{M1}(\widehat{\overline{Y}})$	$\widehat{V}_{M2}(\widehat{\overline{Y}})$	$\widehat{V}_{M3}(\widehat{\overline{Y}})$	$\widehat{V}_{bs}(\widehat{\overline{Y}})$	$\widehat{V}_{jk}(\widehat{\overline{Y}})$
	SE	5.93603e-05	6.17108e-05	6.44271e-05	8.65833e-05	2.98546e-04
$n_1 = 150$, $n_2 = 100$	RB (%)	1.39009	2.88442	4.49552	7.01908	7.25730
	CR (%)	94.8	94.90	95.00	95.00	87.80
	SE	1.95905e-05	1.99543e-05	2.03449e-05	3.52039e-05	0.00014
$n_1 = 300$, $n_2 = 200$	RB (%)	9.43243	10.23632	11.07073	15.58569	17.12736
	CR (%)	96.00	96.40	96.50	96.30	90.50
	SE	4.54482e-06	4.58134e-06	4.61897e-06	1.16925e-05	5.73502e-05
$n_1 = 750$, $n_2 = 500$	RB (%)	-1.42167	-1.13965	-0.85331	9.34029	10.23809
	CR (%)	94.10	94.10	94.10	95.30	88.50
	SE	1.32106e-06	1.32570e-06	1.33042e-06	5.77307e-06	2.82666e-05
$n_1 = 1500$, $n_2 = 1\,000$	RB (%)	-3.48277	-3.35363	-3.22351	16.85805	20.00666
	CR (%)	93.90	93.90	93.90	95.50	89.50
	SE	8.34094e-07	8.36128e-07	8.38187e-07	4.42997e-06	2.28913e-05
$n_1 = 1\,800$, $n_2 = 1\,500$	RB (%)	0.22991	0.32442	0.41952	27.61355	31.53672
	CR (%)	95.60	95.60	95.60	97.60	91.30

方差估计; 其他样本量情况下覆盖率最接近 95% 的均为 $\widehat{V}_{M1}(\widehat{\overline{Y}})$, 覆盖率最低的仍然是 Jackknife 方差估计. 总体上除了 Jackknife 方差估计外其他方差估计的覆盖率均在 93% 与 98% 之间, 具有较好的置信区间覆盖. 从相对偏差和 95% 置信区间覆盖率两个方面来看, 优先考虑 $\widehat{V}_{M1}(\widehat{\overline{Y}})$, 其次是 $\widehat{V}_{M2}(\widehat{\overline{Y}})$ 和 $\widehat{V}_{M3}(\widehat{\overline{Y}})$, $\widehat{V}_{jk}(\widehat{\overline{Y}})$的估计效果最差.

§5.　实证分析

采用 2014 年美国行为风险因素监测调查数据 (下载网址 http://www.cdc.gov/BRFSS), 对基于超总体伪设计与组合样本的非概率抽样推断方法进行实证分析. 选取变量为每天平均睡眠时间 (SLEPTIM1)、过去是否有抑郁情绪 (ADDEPEV2), 以及网络使用变量 (INTERNET) 等 3 个变量, 在去掉缺失、“不知道” 或 “拒绝” 等答案的单元后, 剩下 445 744 个单元. 其中, 每天平均睡眠时间取值为 1–24 小时, 过去是否有抑郁情绪取值为 1 或 2, 将其转换成 0 或 1, 0 表示没有, 1 表示有. 假设该 445 744 个单元构成目标总体 ($N = 445\,744$), 其中过去是否有抑郁情绪为目标变量, 每天平均睡眠时间为协变量, 并且过去是否有抑郁情绪的均值, 即有抑郁情绪的人所占的比例为 0.19109, 方差为 0.15458. 首先从目标总体中不放回简单随机抽取一个样本量为 200 的概率样本 S_2 ($n_2 = 200$), 其样本单元的基础权数为 $w_{2i} = N/n_2$, $i = 1, 2, \ldots, n_2$. 将 445 744 个单元中使用网络的单元视为网

络候选者数据库, 并从网络候选者数据库中不放回简单随机抽取 300 个单元, 得到网络候选者数据库的调查样本 S_1 ($n_1 = 300$). 由于网络候选者数据库是使用网络的总体单元, 再从中随机抽取单元作为网络候选者数据库的调查样本 S_1, S_1 即为非概率样本, 从整个抽取的过程可以看到样本的选择与目标变量并无关系, 并不依赖于目标变量, 该样本选择机制属于可忽略的选择机制.

在实证研究中选取过去是否有抑郁情绪为目标变量 Y, 每天平均睡眠时间为协变量 X, 按照一般的统计规律, 是否有抑郁情绪与每天平均睡眠时间的关系应该是一定的, 因此对于样本单元与非样本单元来说, 是否有抑郁情绪与每天平均睡眠时间之间关系的模型应该是相同的, 与本文前面的假定相符. 由此可根据网络候选者数据库的调查样本 S_1 建立是否有抑郁情绪与每天平均睡眠时间的超总体回归模型, 从而构造出 S_1 的伪权数, 最后将 S_1 与 S_2 结合, 利用组合样本数据计算得到总体均值的估计 $\widehat{\overline{Y}} = n_1\widehat{\overline{Y}}_1/(n_1+n_2) + n_2\widehat{\overline{Y}}_2/(n_1+n_2)$. 重复上述过程 1 000 次, 得到 1 000 个总体均值估计 $\widehat{\overline{Y}}$, 最后计算 1 000 个总体均值估计 $\widehat{\overline{Y}}$ 的均值、方差、相对偏差与均方误差. 为了便于比较, 同时计算仅利用非概率样本 S_1 或概率样本 S_2 得到的总体均值估计 $\widehat{\overline{Y}}_1$、$\widehat{\overline{Y}}_2$ 的均值、方差、相对偏差与均方误差.

表 3 基于超总体伪设计与组合样本的总体均值估计的实证结果

	均值	方差	相对偏差 (%)	均方误差
$\widehat{\overline{Y}}$	0.18460	0.00030	-3.39580	0.00034
$\widehat{\overline{Y}}_1$	0.17967	0.00051	-5.97923	0.00064
$\widehat{\overline{Y}}_2$	0.19201	0.00077	0.47935	0.00077

表 3 显示仅利用非概率样本 S_1 与伪权数计算的总体均值估计 $\widehat{\overline{Y}}_1$ 以及基于超总体伪设计和组合样本的总体均值估计 $\widehat{\overline{Y}}$ 都在 0.18 附近, 而仅利用概率样本 S_2 得到的总体均值估计接近 0.19, 前两者与后者差距不大. 从方差来看, $\widehat{\overline{Y}}_2$ 的离散程度最大, $\widehat{\overline{Y}}$ 的离散程度最小. 从相对偏差来看, 基于超总体伪设计和组合样本的总体均值估计 $\widehat{\overline{Y}}$, 以及仅利用 S_1 得到的总体均值估计 $\widehat{\overline{Y}}_1$ 均低估了目标总体均值, 而仅利用 S_2 得到的总体均值估计 $\widehat{\overline{Y}}_2$ 高估了目标总体均值, 无论是高估还是低估, 偏离程度最大的是 $\widehat{\overline{Y}}_1$, 最小的是 $\widehat{\overline{Y}}_2$, 并且总体上三个估计的相对偏差绝对值均低于 6%. 最后, $\widehat{\overline{Y}}$、$\widehat{\overline{Y}}_1$、$\widehat{\overline{Y}}_2$ 在均方误差上依次递增, 表明基于超总体伪设计和组合样本的总体均值估计的效率是最高的, 仅利用概率样本 S_2 估计总体均值的效率最低.

采用直接推导的方差估计式、Bootstrap 和 Jackknife 三种方法来计算方差估计 $\widehat{V}_{M1}(\widehat{\overline{Y}})$、$\widehat{V}_{M2}(\widehat{\overline{Y}})$、$\widehat{V}_{M3}(\widehat{\overline{Y}})$、$\widehat{V}_{bs}(\widehat{\overline{Y}})$ 与 $\widehat{V}_{jk}(\widehat{\overline{Y}})$, 前三者分别为采用 $\widehat{v}_i = e_i^2$、$\widehat{v}_i = e_i^2/(1-h_{ii})$ 和 $\widehat{v}_i = [e_i/(1-h_{ii})]^2$ 计算的方差估计. 利用 Bootstrap 方法估计方差时, 分别在 S_1 与 S_2 中不放回简单随机抽取样本大小为 150 ($n_{1bs} = 150$) 和 100 ($n_{2bs} = 100$) 的两个样本, 将其组合为一个 Bootstrap 样本; 利用 Jackknife 方法估计方差时, 将样本 $S = S_1 \cup S_2$ 随机

分成 5 个组. 最终计算的方差估计的标准误差 (SE)、相对偏差 (RB) 与 95% 置信区间覆盖率 (CR) 见表 4.

表 4　基于超总体伪设计与组合样本的方差估计的实证结果

	$\widehat{V}_{M1}(\widehat{\overline{Y}})$	$\widehat{V}_{M2}(\widehat{\overline{Y}})$	$\widehat{V}_{M3}(\widehat{\overline{Y}})$	$\widehat{V}_{bs}(\widehat{\overline{Y}})$	$\widehat{V}_{jk}(\widehat{\overline{Y}})$
SE	7.18803e-07	7.21272e-07	7.23779e-07	1.46868e-06	6.55531e-06
RB (%)	0.40786	0.65137	0.89622	-0.27199	-1.29472
CR (%)	93.70	93.70	93.70	93.00	84.80

由表 4 可以看到, 五个方差估计中标准误差最低的是 VM1 方法的方差估计, 最高的是 Jackknife 方差估计, 并且 $\widehat{V}_{M1}(\widehat{\overline{Y}})$、$\widehat{V}_{M2}(\widehat{\overline{Y}})$、$\widehat{V}_{M3}(\widehat{\overline{Y}})$ 的标准误差依次递增, 总的来说五个方差估计的标准误差都较小. 从相对偏差来看, 五个方差估计的相对偏差绝对值均小于 2%, 其中相对偏差绝对值最小的是 Bootstrap 方差估计, 其次是采用 VM1 方法计算的方差估计 $\widehat{V}_{M1}(\widehat{\overline{Y}})$, 最大的是 Jackknife 方差估计, 表明 Bootstrap 方差估计与采用 VM1 方法计算的方差估计偏差最小. 在 95% 置信区间覆盖率方面, $\widehat{V}_{M1}(\widehat{\overline{Y}})$、$\widehat{V}_{M2}(\widehat{\overline{Y}})$、$\widehat{V}_{M3}(\widehat{\overline{Y}})$ 的覆盖率相同且是最高的, 最接近 95%, 方差估计效果较好, Jackknife 方差估计的覆盖率不到 85%, 估计效果最差, 这与模拟结果基本一致. 综合考虑三个指标, $\widehat{V}_{M1}(\widehat{\overline{Y}})$ 的估计效果最好, Jackknife 方差估计的效果最差.

§6. 结　　论

本文针对候选者数据库网络调查的统计推断问题, 提出了基于超总体伪设计与组合样本的非概率抽样推断方法. 首先对非概率样本即网络候选者数据库的调查样本建立超总体模型来构造伪权数, 然后将非概率样本与概率样本结合利用组合样本数据来估计目标总体. 进一步根据超总体模型的方差估计理论直接推导出总体均值估计的方差估计, 并比较了该方差估计与 Bootstrap 方差估计、Jackknife 方差估计的效果. 模拟研究与实证分析表明基于超总体伪设计与组合样本的总体均值估计在估计效率上好于仅使用非概率样本或概率样本的总体均值估计, 并且利用 VM1、VM2 和 VM3 方法计算的方差估计效果均比较好.

本文所提出的方法将基于模型的方法与基于伪设计的方法较为巧妙的融合, 获得了精度较高的总体估计, 充分利用了基于模型的特点和优势, 为非概率抽样的统计推断提供了一条新的思路.

参 考 文 献

[1] SVENSSON J. Web panel surveys – can they be designed and used in a scientifically sound way? [C] // *Proceedings 59th ISI World Statistics Congress, 25-30 August 2013, Hong Kong, China*. Hague, Netherlands: International Statistical Institute, 2013: 567–572.

[2] LEE S. Propensity score adjustment as a weighting scheme for volunteer panel web surveys [J]. *J Off Stat*, 2006, **22(2)**: 329–349.

[3] LEE S, VALLIANT R. Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment [J]. *Sociol Methods Res*, 2009, **37(3)**: 319–343.

[4] VALLIANT R, DEVER J A. Estimating propensity adjustments for volunteer web surveys [J]. *Sociol Methods Res*, 2011, **40(1)**: 105–137.

[5] ELLIOTT M N, HAVILAND A. Use of a web-based convenience sample to supplement a probability sample [J]. *Surv Methodol*, 2007, **33(2)**: 211–215.

[6] ELLIOTT M R. Combining data from probability and non-probability samples using pseudo-weights [J]. *Survey Practice*, 2009, **2(6)**: 1–7.

[7] 刘展, 金勇进. 网络访问固定样本调查的统计推断研究 [J]. 统计与信息论坛, 2017, **32(2)**: 3–10.

[8] VALLIANT R, DORFMAN A H, ROYALL R M. *Finite Population Sampling and Inference: A Prediction Approach* [M]. New York: Wiley, 2000.

Research on Inference of Candidate Database Web Surveys Based on Superpopulation Pseudo Design and the Combined Sample

LIU Zhan PAN Yingli

(*Faculty of Mathematics and Statistics, Hubei Key Laboratory of Applied Mathematics, Hubei University, Wuhan, 430062, China*)

Abstract: How to solve the inference problem of candidate database web surveys is an urgent problem to be solved in the development of web survey. In order to solve this problem, the inference method of non-probability sampling based on superpopulation pseudo design and the combined sample is proposed. A superpopulation model is firstly built up to construct pseudo weights for a survey sample of the web candidate database. The estimator of the population mean is then computed according to the combined sample composed of the survey sample of the web candidate database and a probability sample. The variance estimator of the population mean estimator is lastly derived according to the variance estimation theory of the superpopulation model. The Bootstrap and Jackknife methods are also used to compute the variance estimator. And all these variance estimation methods are compared. The research results show that the population mean estimator based on superpopulation pseudo design and the combined sample is better, and has higher efficiency than the estimator only using the probability sample and the weighted estimator only using the survey sample of the web candidate database. The variance estimator computed by using the VM1, VM2 and VM3 method are relatively better.

Keywords: superpopulation; pseudo design; combined sample; web candidate database; non-probability sampling

2010 Mathematics Subject Classification: 62D05