【统计理论与方法】

DOI:10.20207/j.cnki.1007-3116.2024.0003

概率-非概率调查样本的整合推断问题研究：核匹配方法

王 俊[1a,1b]，金勇进[2]

（1. 中国社会科学院 a. 人口与劳动经济研究所，b. 人力资源研究中心，北京 100006；
2. 中国人民大学 应用统计科学研究中心，北京 100872）

摘要：基于网络便利调查、大数据平台的数据收集方式，在实践中得到了广泛的发展，但获得的样本本质上均是非概率样本。利用非概率样本推断目标总体特征面临着潜在的偏差，如涵盖偏差、自我选择偏差等。近年来，对融合概率调查和非概率调查的数据资源，以估计有限总体特征问题的讨论较多，但依然存在较多问题。在已有研究的基础上，对非概率样本和概率样本均测量了辅助变量，但只有非概率样本测量了研究变量的背景下，介绍基于倾向得分框架的权数构造方法；在倾向得分核匹配方法的基础上，提出了基于融合概率和非概率样本协变量平衡的核函数带宽选择方法，为非概率样本构造倾向得分核匹配权数。模拟结果显示基于倾向得分核匹配的方法能够显著降低非概率样本的偏差，提出的融合样本协变量平衡的带宽方法能够有效减少估计量的相对偏差、绝对相对偏差和标准差。

关键词：非概率样本；融合数据；带宽选择；统计推断；核匹配

中图分类号：C811 **文献标识码**：A **文章编号**：1007－3116(2024)10－0003－10

引用格式：王俊，金勇进. 概率-非概率调查样本的整合推断问题研究：核匹配方法[J]. 统计与信息论坛，2024，39(10)：3-12.

Citation Form：WANG Jun，JIN Yongjin. Statistical inference with integrated probability and non-probability samples：kernel matching method[J]. Journal of statistics and information，2024，39(10)：3-12.

一、引 言

随着互联网、大数据技术的发展，基于互联网的非概率抽样调查、网络爬虫的数据收集方式在社会科学研究中得到了广泛的应用与发展。然而这些方式收集到的样本本质上是非概率样本。与概率抽样调查相比，研究人员虽可以快速、低成本的方式获得样本量较大的非概率样本，但由于非概率样本的入样机制未知，利用其对目标总体进行统计推断，将面临潜在的涵盖偏差、自我选择偏差等[1-2]。传统的权数调整方法，如事后分层、校准、迭代调整等，无法有效减少非概率网络样本的偏差，当调整变量选择不当时，甚至会显著增加加权估计的方差[3]。近些年来，国内外学术界基于整合概率样本和非概率样本以对总体特征进行统计推断的探讨逐渐增加[3-8]。

整合概率样本和非概率样本对目标总体特征推断时，概率样本可是现存的、高质量的概率样本，也可是与非概率网络调查同期执行的小规模概率调查获取的样本。非概率样本主要是网络候选者数据库调查、自我选择调查、应答者驱动调查等网络便利调查获取的调查样本[8]，或者是大数据平台上的数据资源。目前，

收稿日期：2023－12－30

基金项目：国家社会科学基金项目“复杂抽样数据的统计推断方法及其应用研究”(19BTJ012)；中国社会科学院重大经济社会调查项目“中国城市劳动力调查”(GQDC2023022)；中国社会科学院研究所实验室综合资助项目“人口仿真与就业监测实验室”(2024SYZH008)；全国统计科学研究项目“噪声惩罚下的多视角聚类研究”(2023LY093)

作者简介：王 俊，男，安徽六安人，博士，助理研究员，研究方向：抽样调查，应用统计；
金勇进(通讯作者)，男，北京人，博士，教授，博士生导师，研究方向：抽样技术，应用统计。

对概率-非概率样本的融合推断的研究主要可以分为两种情形：一是概率样本和非概率样本同时测量辅助变量和研究变量，如金勇进和刘晓宇讨论的线上、线下调查样本的融合推断[8]，王俊和刘展讨论的概率-非概率混合样本的推断[9]；二是概率样本仅测量辅助变量，非概率样本同时测量辅助变量和研究变量，如 Valliant 和 Dever、金勇进和刘展等讨论的网络志愿者样本库调查[10-11]。实践中，由于研究经费的限制，多数研究人员或机构无法同时开展非概率调查和概率抽样调查。此时，能使用的概率样本通常是仅包含辅助变量的外部大型概率调查样本，或者非常小规模的概率调查样本。因此，第二类情形的应用相对较为普遍。本文以该情形为背景，讨论如何基于概率样本为非概率样本构造权数，以减少估计总体特征时的潜在偏差。

二、基于倾向得分框架的权数构造方法

在概率样本仅测量辅助变量、非概率样本同时测量辅助变量和研究变量的情形下，基于倾向得分框架的推断方法应用较为广泛。该框架将概率样本作为参照样本，为非概率样本构造一套权数，从而减少非概率样本的代表性偏差，其中讨论较多的方法为逆倾向得分权数、倾向得分分组加权权数、倾向得分匹配权数。为方便讨论，令参照样本记为 s^{ref}，样本量为 n^{ref}，样本单元的权重为 $\{d_i^r, i\in s^{\mathrm{ref}}\}$。非概率样本记为 s^{np}，样本量为 n^{np}。参照样本和非概率样本合并样本表示为 $s=\{s^{\mathrm{ref}}, s^{\mathrm{np}}\}$，$I_i=1$ 表示合并后样本 s 中的单元 i 为非概率样本 s^{np} 中的单元，$I_i=0$ 表示合并后样本中的单元 i 为参照样本中的单元。合并后的样本权重为 $\{d^r, d\}$，其中 d^r 为概率样本的权重，d 为非概率样本的权重，假定 $d=1$。

(一)逆倾向得分权数

逆倾向得分权数加权法在国内外文献中被称为伪随机推断法、伪设计推断法[8,11-12]。在该方法中，倾向得分解释为控制协变量后，总体单元参与非概率调查的概率[13]。由于非概率样本的入样概率未知，通常将概率样本作为参照样本，使用 Valliant 和 Dever 提出的加权逻辑回归法估计倾向得分[10]。假定估计出的非概率样本 s^{np} 的倾向得分为 $\{\hat{e}_i, i\in s^{\mathrm{np}}\}$，则对于任一单元 $i\in s^{\mathrm{np}}$，其逆倾向得分权数为 $w_i=1/\hat{e}_i$。总体的逆倾向得分加权(IPW)估计量表示为：

$$\bar{y}_{\mathrm{IPW}}=\frac{1}{\sum_{i\in s^{\mathrm{np}}} w_i}\sum_{i\in s^{\mathrm{np}}} w_i y_i \tag{1}$$

(二)倾向得分分组权数

为了降低 IPW 估计量方差较大的情况，Lee 和 Valliant 提出了基于倾向得分的分组逆加权估计量[14]。该方法首先根据估计出的倾向得分大小将所有非概率样本单元进行排序，并根据分位数将非概率样本分为 K 组，实践中通常分为 5 组[11]。然后计算各组内倾向得分的平均值 $\bar{e}_k$ $(k=1,2,\cdots,K)$。此时，各组内的样本单元权数为 $w_{ki}=1/\bar{e}_k$，$(k=1,2,\cdots,K)$。分组加权估计量则表示为：

$$\bar{y}_{\mathrm{GIW}}=\frac{1}{\sum_{k=1}^{K}\sum_{i\in s^k} w_{ki}}\sum_{k=1}^{K}\sum_{i\in s^k} w_{ki} y_{ki} \tag{2}$$

(三)最近邻倾向得分匹配权数

倾向得分最近邻匹配权数法使得非概率样本和概率样本在样本结构上尽可能保持一致，本质上是基于最近邻方法利用非概率样本为概率样本的研究变量进行插补，进而利用概率样本的插补值对总体特征进行统计推断。目前非概率样本的统计推断中对基于倾向得分的最近邻匹配权数法讨论较少，金勇进和刘晓宇、Liu 和 Valliant 对该方法进行了讨论[8,15]。假定基于合并后样本估计出的倾向得分为 $\{\hat{e}_i, i\in s^{\mathrm{np}}\cup s^{\mathrm{ref}}\}$，则参照样本中单元 i 的未观测值 y_i 的 k 最近邻倾向得分匹配(kNNM)估计为：

$$\hat{y}_i=\frac{1}{k}\sum_{j\in \mathrm{kNN}(d(\hat{e}_i,\hat{e}_u),u\in s^{\mathrm{np}},u\neq i)} y_j \tag{3}$$

其中，$d(\hat{e}_i,\hat{e}_u)$ 为以倾向得分衡量的参照样本中单元 i 和非概率样本中单元 u 之间的距离，距离函数 $d(\cdot)$ 可以选择绝对距离、平方距离等；$\mathrm{kNN}(d(\hat{e}_i,\hat{e}_u),u\in s^{\mathrm{np}},u\neq i)$ 表示以估计出的倾向得分为衡量对象，离参照样本单元 i 最近的 k 个非概率样本单元的集合。

将离参照样本中单元 i 最近的 k 个非概率样本单元观测值的平均值作为参照样本中单元 i 的插补，则基于倾向得分的 k 最近邻匹配估计量 $\hat{\bar{Y}}_{\mathrm{kNNM}}$ 表示为：

$$\hat{\bar{Y}}_{\mathrm{kNNM}}=\frac{1}{\sum_{i\in s^{\mathrm{ref}}}w_i}\sum_{i\in s^{\mathrm{ref}}}d_i^r\hat{y}_i=\frac{1}{\sum_{i\in s^{\mathrm{ref}}}w_i}\sum_{i\in s^{\mathrm{ref}}}d_i^r\frac{1}{k}\sum_{j\in \mathrm{kNN}(d(\hat{e}_i,\hat{e}_u),u\in s^{\mathrm{np}})}y_j$$

$$=\frac{1}{\sum_{i\in s^{\mathrm{ref}}}w_i}\sum_{j\in s^{\mathrm{np}}}\sum_{i\in s^{\mathrm{ref}}}\frac{1}{k}d_i^rI(j\in \mathrm{kNN}(d(\hat{e}_i,\hat{e}_u),u\in s^{\mathrm{np}}))y_j \tag{4}$$

令 $w_j=\sum_{i\in s^{\mathrm{ref}}}d_i^r\frac{1}{k}I(j\in \mathrm{kNN}(d(\hat{e}_i,\hat{e}_u),u\in s^{\mathrm{np}}))$，则有：

$$\hat{\bar{Y}}_{\mathrm{kNNM}}=\frac{1}{\sum_{i\in s^{\mathrm{np}}}w_j}\sum_{i\in s^{\mathrm{np}}}w_jy_j \tag{5}$$

从式(5)可见，$\hat{\bar{Y}}_{\mathrm{kNNM}}$可表示为非概率样本单元的线性加权的形式，$w_j$ 为非概率样本中单元 j 的 k 最近邻匹配权数。易证 $\sum_{j\in s^{\mathrm{np}}}w_j=\sum_{i\in s^{\mathrm{ref}}}d_i^r$。

(四)方法评述

使用 IPW 时，如何解释估计的倾向得分十分重要，直接决定了 IPW 估计量外推的总体范围。Valliant 和 Dever 提出的加权逻辑回归，实质是将参照样本作为目标总体抽样框的估计，估计的倾向得分解释为参照样本估计的总体中单元参加非概率调查的概率[10]。若基于无加权逻辑回归估计倾向得分，估计的倾向得分解释为合并后样本中的单元参加非概率调查的概率 $P(i\in s^{\mathrm{np}}\mid x,s=s^{\mathrm{np}}\cup s^{\mathrm{ref}})$。此时，若合并后样本 s 不能看成来自于总体的简单随机样本，逆倾向得分权数只能将非概率样本还原至合并后样本 s，而不能外推至更大的总体[10]。国内外学者讨论了利用倾向得分为非概率样本构造逆倾向得分权数的方法，模拟研究均显示，当模型指定正确的时候，逆倾向得分权重能够显著减少非概率网络样本的偏差[16-18]。然而，Austin 通过模拟研究表明，当总体规模较大，且非概率网络样本单元抽样比较小时，基于加权逻辑回归估计出的倾向得分容易出现较小，甚至几乎为零的情况，进而导致逆倾向得分加权(IPW)估计量的方差较大，甚至无法使用的情形[19]。此外，逆倾向得分加权估计量的效果对倾向得分模型的设定误差较为敏感，当倾向得分模型设定不当时，逆倾向得分加权估计量无法有效减少非概率样本的偏差。虽然基于倾向得分分组的方法能够降低极端逆倾向得分权数的影响和模型设定误差的影响[20]，但对减少估计量的偏差是有限度的。虽该方法估计量的方差较小，但其 95%置信区间包含真值的比例较低[10]，本文不再列出模拟结果讨论。

相比较于 IPW 估计量和倾向得分分组加权估计量，倾向得分匹配方法受模型指定误差的影响相对较小。在倾向得分最近邻匹配权数中，最近邻个数 k 的选择对匹配权数加权估计量的结果影响较大，若 k 的选择过小将会导致匹配结果受异常值影响较大，也可能会引起非概率样本单元未完全匹配的情况。Wang 等提出了基于倾向得分的核匹配权数方法，该方法在基于最近邻匹配的基础上引入了核函数，使得在一对多匹配过程中，不同距离的非概率样本单元的权重不一样，其本质是利用非概率样本训练非参数回归模型，并利用该模型对概率样本中未观测的研究变量进行预测，进一步基于预测的结果和概率样本的权数得到总体特征的加权估计[20]。在 Wang 等提出的方法基础上，Kern 等考察了各倾向得分估计方法对倾向得分核匹配估计量效率的影响，比如逻辑回归、随机树等[21]。Wang 和 Kern 等在提出的核匹配权数法中未对核函数中的带宽选择问题进行详细的讨论[20-21]。通常情况下，非概率样本和概率样本之间的样本结构存在较大差异，在非概率样本上训练得到的最优带宽，未必在概率样本上得到较好的预测效果。此外，倾向得分匹配的本质是使得协变量在概率样本和非概率样本上的分布尽可能保持一致，这为带宽的选择提供一个新的思路。

基于此，本文在倾向得分核匹配权数法的基础上，提出融合样本协变量平衡法估计核函数中带宽，并讨论该方法的估计量理论性质。进一步，通过模拟研究验证本文提出融合样本协变量平衡法得到的匹配估计量的效果。最后，对本文的结果进行讨论，并指出实践中需要注意的问题。

三、融合样本协变量平衡的倾向得分核匹配权数

(一)倾向得分核匹配权数

Wang 等提出了倾向得分核匹配方法为非概率样本构造核匹配权数，实质是基于训练出的核回归模型对概率样本的未观测值进行预测，并基于概率样本的预测值形成总体特征的加权估计[20]。该方法首先将估

计的倾向得分作为协变量，并利用非概率样本估计核平滑回归模型 $\hat{m}(\hat{e}(x);h)$。假定参照样本中单元 i 的 y_i 的倾向得分核匹配估计表示为 $\hat{m}(\hat{e}(x_i))$：

$$\hat{m}(\pi(x_i)) = \sum_{j\in s^{np}} \frac{K\left(\frac{d(\hat{e}_i-\hat{e}_j)}{h}\right)}{\sum_{j\in s^{np}} K\left(\frac{d(\hat{e}_i-\hat{e}_j)}{h}\right)} y_j \tag{6}$$

则总体均值 $\bar{Y}$ 的倾向得分核匹配(KM)估计量表示为：

$$\hat{\bar{y}}_{KM} = \frac{1}{\sum_{i\in s^{ref}} w_i} \sum_{i\in s^{ref}} d_i^r \hat{m}(\pi(x_i)) \tag{7}$$

由于模型 $\hat{m}(\hat{e}(x);h)$ 可表示为非概率样本单元研究变量的线性加权和，即 $\hat{m}(\hat{e}(x);h) = \sum_{j\in s^{np}} \cdot w^m(h) y_j$。则 $\hat{\bar{y}}_{KM}$ 可表示为非概率样本研究变量的线性加权和形式：

$$\begin{aligned}\hat{\bar{y}}_{KM} &= \frac{1}{\sum_{i\in s^{ref}} w_i} \sum_{i\in s^{ref}} d_i^r \sum_{j\in s^{np}} \frac{K\left(\frac{d(\hat{e}_i-\hat{e}_j)}{h}\right)}{\sum_{j\in s^{np}} K\left(\frac{d(\hat{e}_i-\hat{e}_j)}{h}\right)} y_j \\ &= \frac{1}{\sum_{i\in s^{ref}} w_i} \sum_{j\in s^{np}} \sum_{i\in s^{ref}} d_i^r \frac{K\left(\frac{d(\hat{e}_i-\hat{e}_j)}{h}\right)}{\sum_{j\in s^{np}} K\left(\frac{d(\hat{e}_i-\hat{e}_j)}{h}\right)} y_j = \frac{1}{N}\sum_{j\in s^{np}} w_j^{KM} y_j\end{aligned} \tag{8}$$

其中 $w_j^{KM} = \sum_{i\in s^{ref}} d_i^r \frac{K\left(\frac{d(\hat{e}_i-\hat{e}_j)}{h}\right)}{\sum_{k\in s^{np}} K\left(\frac{d(\hat{e}_i-\hat{e}_k)}{h}\right)}$，为非概率样本中第 j 个单元的核匹配权数；$\hat{N} = \sum_{j\in s^{np}} \cdot w_j^{KM}$；$h$ 为核函数的带宽，可以通过经验原则(Rule of thumb)或者交叉验证选取[20]。

倾向得分核匹配中的带宽选择方法包括经验原则法、交叉验证法。交叉验证法基于非概率样本训练和评估模型，并选择使得交叉验证误差最小的带宽，即在带宽选择时仅使用了非概率样本的信息，造成了概率样本信息的浪费。Wang 和 Kern 等在讨论倾向得分核匹配方法时，仅基于交叉验证、经验原则等确定带宽[20-21]，即仅使用了非概率样本训练和评估模型。通常来说，非概率样本和概率样本之间的样本结构、变量分布存在较大差异，使用非概率样本训练得到的最优带宽，虽然能使得基于非概率样本计算出的损失函数值最小，但无法保证其在概率样本中最小。虽然能够保证选择的带宽使得估计量的偏差和方差在非概率样本中达到相对最优，但在概率样本中不一定能够达到偏差和方差的最优。因此，需要进一步讨论带宽的选择问题。

(二)融合样本协变量平衡的带宽选择方法

在因果推断中，倾向得分匹配的一个重要目的是使得控制组和干预组之间的协变量分布尽可能一致。若选择的协变量和研究变量高度相关，且在控制了这些协变量后总体单元是否参与非概率调查和研究变量不相关时，该方法能够获得较好的结果。本部分提出基于融合概率-非概率样本协变量平衡的带宽选择方法。该方法选择使参照样本和加权后非概率样本在协变量上的分布差异尽可能小的带宽，进而得到基于融合概率-非概率样本协变量平衡的核匹配(ICBKM)权数。本文提出的融合概率-非概率样本协变量平衡的带宽选择法中，带宽的选择并非使得非概率样本的预测误差达到最小，而是使非概率样本协变量的核权数加权平均和概率样本的协变量的加权均值的差异尽可能最小，即式(9)：

$$\min_h (\bar{x}_{prob} - \bar{x}_{np}^{KM})(\bar{x}_{prob} - \bar{x}_{np}^{KM})^T \tag{9}$$

其中，$\bar{x}_{np}^{KM} = \frac{1}{N}\sum_{j\in s^{np}} w_j^{ICBKM} x_j$；$w_j^{ICBKM}$ 为非概率样本中单元 j 的基于融合样本协变量平衡的带宽选择方法得到的核匹配权数，表示为 $w_j^{ICBKM} = \sum_{i\in s^{ref}} d_i^r \frac{K\left(\frac{d(\hat{e}_i-\hat{e}_j)}{h}\right)}{\sum_{j\in s^{np}} K\left(\frac{d(\hat{e}_i-\hat{e}_j)}{h}\right)}$；核函数 $K(\cdot)$ 可以选择高斯核等；h

为带宽，根据式(9)进行选择。易证，$\sum_{j\in s^{np}} w_j^{ICBKM} = \sum_{i\in s^{ref}} d_i^r$，即非概率样本的匹配权数和等于参照样本的权数和。

此时，基于融合样本协变量平衡的带宽选择方法得到的核匹配权数加权估计量(ICBKM)表示为：

$$\hat{\bar{y}}_{ICBKM} = \frac{1}{N}\sum_{j\in s^{np}} w_j^{ICBKM} y_j \tag{10}$$

(三)核匹配权数加权估计量的性质

假定概率样本的权数已校准至和总体规模 N 一致，即式(10)中 $\hat{N}=N$。假定有限总体 $U=\{(Y_1,X_1,e(x_1)),(Y_2,X_2,e(x_2)),\cdots,(Y_N,X_N,e(x_N))\}$，从超总体 $F(y,x,e(x))$ 独立生成，有限总体的均值为 $\bar{Y}$，超总体模型中 $E(y)=\mu$。当控制了协变量 X 后，合并样本中的任意一单元 i 无论属于非概率样本还是概率样本均与研究变量 y 不相关，且满足条件(*I*)、$0<e(x)^{-1}<+\infty$、$e(x)$在概率样本和非概率样本之间有足够的重叠时，可以证明：

$$E(y\mid e(x),s^{ref}) = E(y\mid e(x),s^{np}) = E(y\mid e(x)) \tag{11}$$

此时，当满足如下(Ⅱ)～(Ⅴ)的条件下，易证，当总体规模 $N\to\infty$，$n^{ref}\to\infty$，$n^{np}\to\infty$时，$E(\hat{\bar{y}}_{ICBKM})\xrightarrow{D^{ref}}\bar{Y}$，$D^{ref}$为参照样本的抽样设计。

(Ⅰ)对任一样本 $i\in\{s^{ref}\cup s^{np}\}$，满足

$P(i\in s^{np}\mid x,y,s^{ref}\cup s^{np}) = P(i\in s^{np}\mid x,s^{ref}\cup s^{np})$

(Ⅱ)核函数 $K(\mu)$满足，$\int K(\mu)\mathrm{d}\mu=1$，$\sup_{\mu}\mid K(\mu)\mid<\infty$，以及 $\lim_{\mid\mu\mid<\infty}\mid\mu\mid\cdot\mid K(\mu)\mid=0$；

(Ⅲ)带宽 $h=h(n^{np})$，$h\to 0$，当 $n^{np}\to\infty$时，$h\cdot n^{np}\to\infty$；

(Ⅳ)研究变量 Y 的二阶矩存在，即 $E(y^2<\infty)$；

(Ⅴ)概率样本中任一单元的权重 $d_i<\alpha,\alpha\in\mathbb{R}_+$。

(四)基于 bootstrap 的方差估计

由于方差估计相对复杂，本文采取 bootstrap 方法估计各估计量的方差。此时不仅能考虑复杂概率样本的抽样误差，还能将由于抽样的随机性带来的模型的估计误差考虑进来。假设原参照样本为 s^{ref}，样本单元的权重为$\{d_i^r,i\in s^{ref}\}$。非概率样本为 s^{np}，样本单元的权重为 1。具体估计过程如下：

第一步，从参照样本中根据有放回式简单随机抽样抽取一个样本 s_b^{ref}，假定再抽样样本量为 n_b^{ref}；

第二步，从非概率样本中根据有放回式简单随机抽样方法抽取一个简单随机样本 s_b^{np}，再抽样样本量为 n_b^{np}；

第三步，将概率样本的第 b 个 bootstrap 样本 s_b^{ref} 和非概率样本的第 b 个 bootstrap 样本 s_b^{np} 合并成一个样本估计倾向得分 $e=P(i\in s_b^{np}\mid X;s_b^{ref}\cup s_b^{np})$。则 s_b^{ref} 中第 i 个单元 y_i 的核匹配估计为：

$$\hat{m}(\hat{e}(x_i)) = \frac{\sum_{j\in s_b^{np}} K\left(\frac{d(\hat{e}_i-\hat{e}_j)}{h}\right)y_j}{\sum_{j\in s_b^{np}} K\left(\frac{d(\hat{e}_i-\hat{e}_j)}{h}\right)} \tag{12}$$

则第 b 个 bootstrap 样本的核匹配估计量为：

$$\begin{aligned}\bar{Y}_{KM}^{(b)} &= \frac{1}{n_p^{rep}}\sum_{i\in s_b^{ref}}\hat{m}(\hat{e}(x_i)) = \frac{1}{n_p^{rep}}\sum_{i\in s_b^{ref}}\frac{\sum_{j\in s_b^{np}} K\left(\frac{d(\hat{e}_i-\hat{e}_j)}{h}\right)y_j}{\sum_{j\in s_b^{np}} K\left(\frac{d(\hat{e}_i-\hat{e}_j)}{h}\right)}\\ &= \frac{1}{n_{np}^{rep}}\frac{n_{np}^{rep}}{n_p^{rep}}\sum_{j\in s_b^{np}}\sum_{i\in s_b^{ref}}\frac{K\left(\frac{d(\hat{e}_i-\hat{e}_j)}{h}\right)}{\sum_{j\in s_b^{np}} K\left(\frac{d(\hat{e}_i-\hat{e}_j)}{h}\right)}y_j = \frac{1}{n_{np}^{rep}}\sum_{j\in s_b^{np}} w_j^b y_j\end{aligned} \tag{13}$$

其中，$w_j^b = \sum_{i\in s_b^{ref}}\frac{K\left(\frac{d(\hat{e}_i-\hat{e}_j)}{h}\right)}{\sum_{j\in s_b^{np}} K\left(\frac{d(\hat{e}_i-\hat{e}_j)}{h}\right)}$。

第四步，重复第一步至第三部 B 次，得到 B 次匹配估计量的均值 $\hat{y}_{\text{ker}}$ 和估计量的方差 $v(\hat{y}_{\text{KM}})$ 分别为：

$$\bar{y}_{\text{ker}} = \frac{1}{B}\sum_{b=1}^{B}\bar{Y}_{\text{KM}}^{(b)} \tag{14}$$

$$v(\bar{y}_{\text{KM}}) = \frac{1}{B}\sum_{b=1}^{B}(\bar{Y}_{\text{KM}}^{(b)} - \bar{y}_{\text{ker}})^2 \tag{15}$$

四、模拟研究

由于调查组织者对非概率网络调查理解不够深入，当前已公开的网络调查数据和外部大型社会调查数据重叠的变量较少，重叠的变量间也不具有可比性，无法直接使用。本部分通过模拟研究，在理想的情况下，对提出的方法进行评估。

（一）模拟数据说明

模拟研究中的总体 U 包括辅助变量 $X=(X_1,X_2,X_3,X_4)$、用于模拟概率抽样设计的设计变量 Z 以及研究变量 Y，总体规模为 1 000 000。X 由式(16)生成、设计变量 Z 由式(17)生成、研究变量 Y 由式(18)生成。

$$X\sim N(\mu,\Sigma),\mu=(4,0,4),\Sigma=\text{diag}(4,1,2) \tag{16}$$

$$Z\sim\text{uniform}(100,10\ 000) \tag{17}$$

$$y=-0.06+0.95x_1+0.7x_2+0.25x_3+0.65x_4+\varepsilon,\varepsilon\sim N(0,1) \tag{18}$$

概率抽样设计为根据设计变量 Z 的大小成比例的 PPS 抽样，总体任一单元 i 的包含概率 $\pi s_i = n_p \cdot \frac{Z_i}{\sum_{j\in U}Z_j}$，其中 n_p 为概率样本的样本量。非概率样本的样本选择机制假定为式(19)的 logit 线性模型：

$$\text{Logit}(\pi_i^{\text{np}})=0.1+x_1+x_2+x_3+x_4 \tag{19}$$

其中，总体中任一单元参加非概率调查的包含概率为 $\pi_i^{*} = n_{\text{np}}\frac{\pi_i^{\text{np}}}{\sum_{j\in U}\pi_j^{\text{np}}}$，其中 n_{np} 为非概率调查的样本量。

（二）模拟过程

概率调查的组织实施费用通常较高，如果概率样本的样本量较大将会导致研究经费的上升，这对大多数经费有限的研究机构或者个人来说是不可承受的。本部分设置概率样本的样本量 $n_p=500$，抽样比为 0.5‰。非概率样本的样本量为 $n_{\text{np}}=3\ 000$，抽样为 3.3‰。模拟过程具体如下：

第一步，根据设计变量 Z 大小成比例的 PPS 抽样，从模拟总体中抽取样本量为 $n_p=500$ 的概率样本 s_p；根据式(19)指定的线性 logit 形式选取样本量为 $n_{\text{np}}=3\ 000$ 的非概率样本 s_{np}；

第二步，根据概率样本得到总体特征的 HT 估计 $\hat{y}_{\text{HT}}^{\text{prob}}$。根据非概率样本得到总体特征的简单均值估计 $\hat{y}_{\text{simp}}^{\text{np}}$；

第三步，将非概率样本 s_{np} 和概率样本 s_p 合并，利用加权和未加权的 logit 回归分别估计样本单元的倾向得分，分别基于估计出的倾向得分计算总体的 IPW 估计 $\hat{y}_{\text{IPW}}^{\text{wted}}$ 和 $\hat{y}_{\text{IPW}}^{\text{unwted}}$、1 最近邻倾向得分匹配权数加权估计 $\hat{y}_{\text{1NNM}}^{\text{wted}}$ 和 $\hat{y}_{\text{1NNM}}^{\text{unwted}}$、10 最近邻倾向得分匹配权数加权估计 $\hat{y}_{\text{10NNM}}^{\text{wted}}$ 和 $\hat{y}_{\text{10NNM}}^{\text{unwted}}$、基于经验原则选择带宽的核平滑权数加权估计 $\hat{y}_{\text{KM\_siv}}^{\text{wted}}$ 和 $\hat{y}_{\text{KM\_siv}}^{\text{unwted}}$、基于非概率样本训练误差最小选择带宽的核平滑权数加权估计 $\hat{y}_{\text{KM\_np}}^{\text{wted}}$ 和 $\hat{y}_{\text{KM\_np}}^{\text{unwted}}$、本文提出的基于融合样本协变量平衡带宽选择方法的核平滑权数加权估计 $\hat{y}_{\text{ICBKM}}^{\text{wted}}$ 和 $\hat{y}_{\text{ICBKM}}^{\text{unwted}}$，并基于 bootstrap 方法计算估计量的标准差。最后，重复上面的两个过程 300 次，计算 95%置信区间包含真值比例。

（三）模拟结果分析

通过估计结果的相对偏差、绝对相对偏差、标准差、95%置信区间包含真值的比例四个指标对基于融合样本协变量平衡的带宽选择方法的核匹配权数（ICBKM）加权估计量的效果进行评估，并和 IPW、1 最近邻匹配权数（1NNM）加权估计、10 最近邻匹配权数（10NNM）加权估计、Wang 等讨论的基于经验原则和非概率样本训练误差最小原则选择的带宽得到的核匹配权数加权估计进行比较[20]。相对偏差（R. bias）、绝对相对偏差（A. R. bias）、标准差（S. E）、95%置信区间包含真值比例（C. R.）分别定义为：

$$\text{R. bias}=\frac{\hat{\bar{y}}-\bar{Y}}{\bar{Y}}\times 100\% \tag{20}$$

$$\text{A. R. bias}=\left|\frac{\hat{\bar{y}}-\bar{Y}}{\bar{Y}}\right|\times 100\% \tag{21}$$

$$\text{S. E.}=\sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\bar{y}}_b-\frac{1}{B}\sum_{b=1}^{B}\hat{\bar{y}}_b)^2} \tag{22}$$

$$\text{C. R.}=\frac{1}{B}\sum_{b=1}^{B}I(\bar{Y}\in[\hat{\bar{y}}_b\pm\text{S. E.}(\hat{\bar{y}}_b)])\times 100\% \tag{23}$$

模拟结果(表1)中，概率样本HT估计的平均相对偏差为0.03%，非概率样本简单均值的平均偏差为-10.09%。Valliant和Dever提出的基于加权逻辑回归的IPW估计量[10]、kNNM加权估计量和KM加权估计量均能够明显减少非概率样本中的偏差。从IPW估计的表现上看，基于无加权逻辑回归、加权逻辑回归得到的IPW估计量的平均相对偏差较大，分别为-8.85%和-3.83%，加权逻辑回归的IPW估计量优于未加权逻辑回归的IPW估计量，因此使用IPW估计量时需要注意。各倾向得分匹配权数加权估计量均能够显著减少非概率样本的偏差，基于无加权逻辑回归估计的倾向得分的匹配权数加权估计量的相对偏差均小于基于加权逻辑回归估计的倾向得分得到的结果。具体来看，基于加权逻辑回归估计的倾向得分得到的1NNM加权估计量和10NNM加权估计量的相对偏差分别为-1.02%和-1.08%，基于无加权逻辑回归估计的倾向得分得到的1NNM加权估计量和10NNM加权估计量的相对偏差均在1%以内，分别为-0.60%和-0.56%。基于加权逻辑回归得到的倾向得分匹配权数加权估计量的相对偏差的绝对值均大于1%，但小于2.60%，优于IPW估计量的结果，其中基于经验原则选择带宽的核匹配加权估计量的相对偏差最大，为-2.59%，本文提出的融合样本协变量平衡方法选择带宽和基于非概率样本训练误差最小方法选择带宽得到的核匹配权数加权估计相对偏差分别为-1.69%和-1.20%。基于无加权逻辑回归估计的倾向得分得到的核匹配权数加权估计量相对偏差较小，绝对值均小于1.05%，其中基于非概率样本训练误差最小选择带宽方法的偏差为-1.03%，本文提出的融合样本协变量平衡的带宽选择方法的相对偏差为-0.99%，基于经验法则选择的带宽得到的相对偏差最小为-0.60%。

表1　模拟结果

估计方法	相对偏差(%)	绝对相对偏差(%)	标准差	95%置信区间包含真值比例(%)
$\hat{y}_{\text{simp}}^{\text{non-prob}}$	-10.09	10.08	0.032 8	0.00
$\hat{y}_{\text{HT}}^{\text{prob}}$	0.03	2.06	0.082 2	94.67
$\hat{y}_{\text{IPW}}^{\text{wted}}$	-3.83	4.07	0.077 5	64.00
$\hat{y}_{\text{IPW}}^{\text{unwted}}$	-8.85	8.84	0.031 2	0.00
$\hat{y}_{\text{1NNM}}^{\text{wted}}$	-1.02	5.30	0.190 9	94.00
$\hat{y}_{\text{1NNM}}^{\text{unwted}}$	-0.56	4.25	0.172 1	97.00
$\hat{y}_{\text{10NNM}}^{\text{wted}}$	-1.08	4.08	0.141 5	93.00
$\hat{y}_{\text{10NNM}}^{\text{unwted}}$	-0.60	2.95	0.111 6	95.00
$\hat{y}_{\text{KM\_np}}^{\text{wted}}$	-1.20	3.77	0.140 0	93.33
$\hat{y}_{\text{KM\_np}}^{\text{unwted}}$	-1.03	3.44	0.118 5	88.00
$\hat{y}_{\text{KM\_siv}}^{\text{wted}}$	-2.59	3.34	0.097 2	85.00
$\hat{y}_{\text{KM\_siv}}^{\text{unwted}}$	-0.60	3.52	0.117 9	92.00
$\hat{y}_{\text{ICBKM}}^{\text{wted}}$	-1.69	3.49	0.111 3	87.33
$\hat{y}_{\text{ICBKM}}^{\text{unwted}}$	-0.99	2.89	0.095 9	91.67

从绝对相对偏差(表1和图1)看，概率样本HT估计的平均绝对相对偏差为2.06%，非概率样本简单均值的平均绝对相对偏差为10.08%。IPW估计量的平均绝对相对偏差均大于4%，无加权和加权逻辑回归情形下，IPW估计量的平均绝对相对偏差分别为8.84%和4.07%，实际应用IPW估计量时需要谨慎。除了1NNM加权估计量外，倾向得分核匹配权数加权估计量的绝对相对偏差整体较小，均低于4%，且无加权逻辑回归得到的匹配权数加权估计量的表现均优于加权逻辑回归得到的结果。无加权逻辑回归估计的倾向得分情形下，本文提出的融合样本协变量平衡的带宽选择方法得到的核匹配权数加权估计量的平均绝对相对偏差最小，为2.89%；其次分别为10NNM加权估计量(2.95%)和经验原则选择带宽得到的倾向得分匹配

权数加权估计量(3.34%)。1NNM 加权估计量和基于经验原则选择带宽得到的核匹配权数估计量的平均绝对相对偏差较大,分别为 4.25%和 3.52%。加权逻辑回归倾向得分的情形下,1NNM 和 10NNM 加权估计量的绝对相对偏差较高,分别为 5.30%和 4.08%。经验原则选择的带宽得到的匹配估计量、本文提出的融合样本协变量平衡选择的带宽得到的 ICBKM 加权估计量的绝对相对偏差较小,分别为 3.34%和 3.49%。整体上看,加权逻辑回归估计的倾向得分构造的 IPW 估计量的表现优于无加权逻辑回归时的 IPW 估计量。各倾向得分匹配权数加权估计量中,无加权逻辑回归估计出的倾向得分优于加权逻辑回归估计出的倾向得分,本文提出的 ICBKM 加权估计量的效果在绝对相对偏差上表现相对较好。

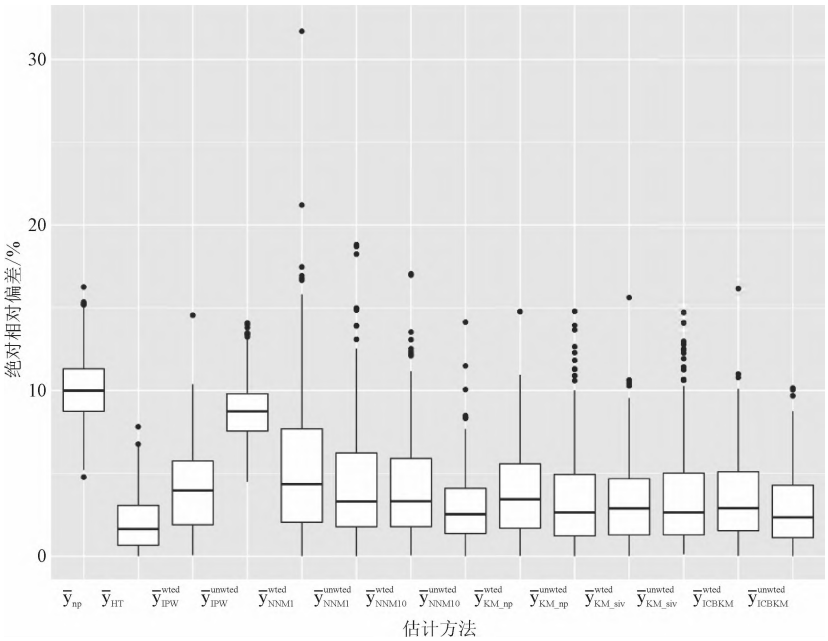

图 1　绝对相对偏差(%)

从估计量的标准差来看,在基于倾向得分框架的加权估计量中,本文提出的基于无加权逻辑回归估计出的倾向得分构建的 ICBKM 权数加权估计量的标准差最小,为 0.095 9,其他基于倾向得分匹配权数加权估计量的标准差均在 0.11 左右,基于 1NNM 权数加权估计量的标准差最大,均大于 0.17。从 95%置信区间包含真值的比例上看,基于 kNNM 权数加权估计量的 95%置信区间包含比例均在 93%～97%之间,其他倾向得分核匹配估计量的置信区间包含真值的比例均在 90%上下,本文提出的基于无加权逻辑回归估计出的倾向得分构建的 ICBKM 估计量的比例为 91.67%。值得注意的是基于加权和无加权逻辑回归估计出的倾向得分构建的 IPW 估计量的包含真值的比例相对较小,分别为 64%和 0%,可能的原因是在抽样的过程中,本文对真实的倾向得分进行了标准化,使得倾向得分的和为期望样本量,进而导致估计倾向得分时面临潜在的模型设定误差,这也从另外一个方面说明,倾向得分模型的设定对 IPW 估计量的效果有着较大的影响,而基于倾向得分匹配方法构建的加权估计量受模型设定误差的影响相对较小,因此实际应用过程中,对 IPW 估计量的使用需要谨慎。

五、结论与建议

本文在非概率样本和概率样本均测量了辅助变量,但仅有非概率样本测量了研究变量的背景下,讨论了利用概率样本的代表性优势为非概率样本构造匹配权数的方法,包括基于最近邻的倾向得分匹配权数和倾向得分核匹配权数,并在 Wang 讨论的核匹配权数加权估计量的基础上,提出了基于融合概率样本和非概率样本协变量平衡的带宽选择方法。模拟研究结果表明,相比较于 IPW 估计量,基于倾向得分的最近邻匹配权数加权估计量、基于倾向得分的核匹配估计量受模型指定误差影响相对较小,且能够显著降低非概率样本估计总体特征的相对偏差、绝对相对偏差。基于最近邻的倾向得分匹配权数加权估计方法虽能够有效降低

非概率样本的相对偏差、提高95%置信区间包含真值的比例，但当k较小时，其绝对相对偏差和估计量的标准差相对较大，实际应用过程中，可以利用本文提出的融合样本协变量平衡的方法通过交叉验证选取最优的k。基于倾向得分核匹配权数加权估计量能够显著降低非概率样本估计总体时的相对偏差和绝对相对偏差，基于无加权逻辑回归估计出的倾向得分构建的倾向得分匹配权数加权估计量的效果一致优于加权逻辑回归得到的结果，其中本文提出基于融合样本协变量平衡的带宽选择方法得到的匹配权数加权估计量的绝对相对偏差和标准差在核匹配估计量中最小。因此，在倾向得分匹配权数的框架下，基于概率样本为非概率样本构造匹配权数的方法有较大的实用价值。然而，由于非概率调查设计人员、大数据平台的规划人员对非概率样本认识的不足，在最初设计调查问卷或者收集的数据信息时，未考虑到与概率样本的变量重叠和可比性的问题，使得当前多数非概率样本和现存的大型概率调查重合变量较少、重合变量的问题缺乏可比性等，导致无法有效融合概率和非概率样本对总体特征进行统计推断。因此，增加非概率调查和概率调查之间问卷设计的协同性，形成非概率调查实践与总体推断的行业规范是未来充分利用和开发非概率样本的基础。

参考文献：

[1] 王莹，万舒晨. 大数据时代抽样调查面临的挑战与机遇[J]. 统计与信息论坛，2016，31(6)：33-36.

[2] BAKER R，BRICK J，BATES N，et al. Summary report of the AAPOR task force on non-probability sampling[J/OL]. 2013. https：//doi. org/10. 1093/JSSAM/SMT008.

[3] YEAGER D S，KROSNICK J A，CHANG L，et al. Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples[J]. The public opinion quarterly，2011，75(4)：709-747.

[4] KIM J K，PARK S，CHEN Y，et al. Combining non-probability and probability survey samples through mass imputation [J]. Journal of the royal statistical society：series A (statistics in society)，2021，184(3)：941-963.

[5] WISNIOWSKI A，SAKSHAUG J W，PEREZ RUIZ D A，et al. Integrating probability and nonprobability samples for survey inference[J]. Journal of survey statistics and methodology，2020，8(1)：120-147.

[6] YANG S，KIM J K. Statistical data integration in survey sampling：a review[J]. Japanese journal of statistics and data science，2020，3(2)：625-650.

[7] ROBBINS M W，GHOSH-DASTIDAR B，RAMCHAND R. Blending probability and nonprobability samples with applications to a survey of military caregivers[J]. Journal of survey statistics and methodology，2021，9(5)：1114-1145.

[8] 金勇进，刘晓宇. 线上、线下调查数据的融合问题——以基于伪设计的校准为例[J]. 统计研究，2022，39(1)：122-131.

[9] 王俊，刘展. 概率-非概率混合样本的统计推断问题研究[J]. 数理统计与管理，2021，40(6)：1069-1079.

[10] VALLIANT R，DEVER J A. Estimating propensity adjustments for volunteer web surveys[J]. Sociological methods & research，2011，40(1)：105-137.

[11] 金勇进，刘展. 大数据背景下非概率抽样的统计推断问题[J]. 统计研究，2016，33(3)：11-17.

[12] ELLIOTT M R，VALLIANT R. Inference for nonprobability samples[J]. Statistical science，2017，32(2)：249-264.

[13] 金勇进，刘晓宇. 权数对基于模型推断的影响分析[J]. 统计与信息论坛，2022，37(3)：3-13.

[14] LEE S，VALLIANT R. Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment[J]. Sociological methods & research，2009，37(3)：319-343.

[15] LIU Z，VALLIANT R. Investigating an alternative for estimation from a nonprobability sample：matching plus calibration [J]. Journal of official statistics，2023，39(1)：45-78.

[16] 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.

[17] 王俊，金勇进，司亚娟. 网络便利样本的推断问题研究[J]. 统计与决策，2018，34(18)：10-15.

[18] VALLIANT R. Comparing alternatives for estimation from nonprobability samples[J]. Journal of survey statistics and methodology，2020，8(2)：231-263.

[19] AUSTIN P C. Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis [J]. Statistics in medicine，2016，35(30)：5642-5655.

[20] WANG L，GRAUBARD B I，KATKI H A，et al. Improving external validity of epidemiologic cohort analyses：a kernel weighting approach[J]. Journal of the royal statistical society：series A (statistics in society)，2020，183(3)：1293-1311.

[21] KERN C，LI Y，WANG L. Boosted kernel weighting — Using statistical learning to improve inference from nonprobability samples[J]. Journal of survey statistics and methodology，2021，9(5)：1088-1113.

Statistical Inference with Integrated Probability and Non-probability Samples: Kernel Matching Method

WANG Jun[1a,1b], JIN Yongjin[2]

(a. Institute of Population and Labor Economics, b. Human Resources Research Center, 1. Chinese Academy of Social Sciences, Beijing 100006, China; 2. Center for Applied Statistics, Renmin University of China, Beijing 100872, China)

Abstract: The data collection methods based on web convenient survey and big data platform have been widely adopted in social science research. However, the obtained samples are essentially non-probability samples. Finite population quantile estimates using those non-probability samples face potential biases, such as coverage bias, self-selection bias. In recent years, there have been considerable discussions on integrating data sources from probability surveys and non-probability surveys to estimate finite population quantiles, however, there are still many issues remain. This paper first introduces methods for constructing weights for non-probability sample based on propensity score frameworks, including inverse propensity score weighting, grouped inverse propensity score weighting and propensity score matching, when only covariates are measured for probability sample, but both covariates and study variables are measured for non-probability sample. Building on propensity score kernel matching methods, a kernel bandwidth selection method is selected, that balances covariates between probability sample and non-probability sample, to construct propensity score kernel matching weights for non-probability sample. Simulation results indicate that weighting non-probability samples using propensity score matching methods provides superior performance compared to inverse propensity score weighting, particularly when the propensity score model is estimated using unweighted logistic regression. Although inverse propensity score weighting can mitigate biases in non-probability samples when propensity score model is estimated by weighted logistic regression, it tends to have a larger standard error and a lower coverage rate compared to propensity score matching methods. However, inverse propensity score weighting method can not reduce bias of non-probability sample when propensity score model is estimated by unweighted logistic regression For k-nearest neighbors (kNN) propensity score matching estimators, both the standard error and coverage rate are adversely affected when k is small. In contrast, kernel matching methods based on propensity scores substantially reduce biases and standard errors while improving the coverage rate for non-probability samples. Furthermore, the kernel bandwidth selection method proposed effectively decreases both relative and absolute biases as well as the standard error of the estimates.

Key words: non-probability sample; integrated sample; bandwidth selection; statistical inference; kernel matching

(责任编辑:张爱婷)