2023-2024 年度

山西省统计科学研究课题

优秀成果

项目编号 2023LD008

项目类别 重大课题

项目名称 山西市、县（区）总体个体户增加值核算抽样方法设计

项目负责人 张 帅

承担单位 山西财经大学

课题组成员 王子涵 曹苗苗 任欣颖 李倩 孙含笑

项目编号	2023LD008

山西省统计科学研究项目结项评审活页

（活页文字表述中不得直接或间接透露个人相关背景材料）

课题名称：山西市、县（区）总体个体户增加值核算抽样方法设计

内容简介（本课题解决的主要问题，重点和难点，学术价值，创新之处）：

一、本课题主要解决的问题

我国的GDP公布数据分为省、市、县（区）三个层次。个体户增加值是GDP的构成部分之一。在第五次全国经济普查（以下简称“五经普”）中，个体户抽样方案的实施对于全面、准确地了解我国个体户的经济状况至关重要。国家五经普平台抽取的个体户样本是对省有代表性的样本，对地市和县不具有代表性，这使个体户增加值在市和县（区）之间分配缺乏准确的依据。国家五经普平台的抽样结果显示，每个县（区）都有被抽到的样本，但这些样本对县（区）和市都不具有代表性。因此，要想准确分配个体户增加值，研究个体户的市、县（区）代表性样本抽样方法是十分必要的。

（一）如何获取对省、市、县（区）均具有代表性的样本?

省级样本、市级样本和县（区）级样本兼容才是最节约样本容量的方案，这样注定不能省、市、县（区）独立抽样，独立抽样会造成样本的巨大浪费。由此想到，自下而上的抽样思路，这种思路只需设计县（区）级抽样方案，获取县（区）级代表性样本，得到县（区）级个体户增加值推断结果，然后逐层汇总，得到市级和省级数据。本项目的研究关键是县（区）级个体户抽样方法设计和增加值推断。

（二）如何充分利用国家抽取的对省有代表性的样本?

国家已经抽取了对省有代表性的个体户样本，在不能改变国家抽样方案的情况下，只有充分利用国家的抽样信息，以达到节约成本的目的。一种思路是在国家抽取的样本中填加新样本，直至推断误差控制在要求范围内，一种思路是将国家抽取的样本，视为一个子总体，在总体中扣除，只对剩余总体进行抽样。经过多种方法尝试，最终选取了第二种思路。

（三）如何对总体进行行业分层？

为服务于GDP分行业数据的获取，个体户增加值有必要分行业推断。国民经济行业分类有国家划分标准，门类、大类、中类和小类，以往国家对经济普查个体户行业分类采用的是工业、建筑业、批发零售业、住宿餐饮业、服务业等五个行业分类。本项目为保持与省级抽样推断结果具有可比性也采用了国家的五行业分类。

（四）如何确定全面调查层和抽样调查层？

国家五经普和四经普的个体户抽样方案设计，均将个体户进行了全面调查层和抽样调查层划分。国家的抽样方案往往是基于大量的模拟，最终采用的一种较为节约成本的抽样方法。故本项目也借鉴了这种全面调查层和抽样调查层划分思路，但是具体如何划分是一个难题。以清查数据中，数据质量较高的就业人数，作为全面调查层和抽样调查层的划分标志。通过分行业大量模拟，从不划分全面调查层，到可能的各种划分门槛的模拟，直至找到最优门槛值。实践表明，全面调查层和抽样调查层的划分会减少大量样本容量，是非常必要的。同时，各地区各行业的最优划分门槛应该是不同的。

（五）如何推断增加值？

根据国家普查登记表612的登记数据，设计以下县（区）级个体户增加值推断公式，该公式适用于五个行业：

个体户增加值 = 全面调查层的增加值 + 非全面调查层的增加值 + 上报平台的个体户增加值。

其中，非全面调查层增加值 =(非全面调查层抽样增加值 / 非全面调查层样本个数) * 非全面调查层总体数；

非全面调查层抽样增加值=非全面调查层抽样的雇佣员工总支出+非全面调查层抽样的全年的营业收入-非全面调查层抽样的全年与经营相关的总支出+非全面调查层抽样的总支出的全年缴纳的各种税费；

全面调查层的增加值=全面调查层的雇佣员工总支出+全面调查层的全年的营业收入-全面调查层的全年与经营相关的总支出+全面调查层的总支出的全年缴纳的各种税费；

上报平台的个体户增加值=上报平台的个体户雇佣员工总支出 + 上报平台的个体户全年的营业收入 - 上报平台的个体户全年与经营相关的总支出 + 上报平台的个体户全年缴纳的各种税费。

二、重点和难点

（一）重点

研究重点有如下三点：

第一，县（区）级个体户代表性样本抽样方法设计。对于山西的县（区）级个体户代表性样本，无需向国家五经普抽样方案中那样，进行整群抽样。县域面积并不是很大，整群抽样会导致相同推断精度条件下，样本容量大大增加，并且分行业整群抽样，也基本覆盖了全县，不会带来明显的成本节约。也就是说县（区）级代表性样本抽样方法设计可在国家五经普抽样方案中进行改进。

第二，全面调查层和抽样调查层的划分界限确定。国家五经普抽样方案中，依据个体户是否达到线上规模来划分全面调查层和抽样调查层，达到线上规模的作为全面调查层，剩余作为抽样调查层。这种确定没有以节约成本为前提，本研究以清查数据质量较高的从业人数作为划分标志，通过模拟，找到最优划分门槛值，研究发现，不同地区，不同行业，划分门槛不同，有必要针对具体地区和行业划分不同门槛值。最终模拟发现，某县代表性样本容量并不必国家五经普平台抽到的省级代表性样本多，但却达到了对该县有代表性的目的。

第三，抽样精度的判断依据确定。国家五经普方案，以估计普查小区内个体户数量作为抽样精度的判断依据，这种估计略显粗糙。以就业人数作为估计内容，判断抽样精度明显更加精细。

（二）难点

在进行省、市、县（区）三级样本的抽样调查时，确保每一级样本的代表性，同时避免资源浪费，是一项极具挑战性的任务。如果处理不当，可能导致各级样本独立，无法有效反映总体情况，从而造成资源和信息的浪费。目前，关于如何在不浪费样本信息的前提下，兼顾三级样本代表性的文献和研究相对缺乏，这为本项目的实施增加了难度。

本项目的主要内容包括以下几个方面：

样本代表性的兼顾策略：研究如何在确保省级样本代表性的基础上，使市级和县级样本能够充分反映省级总体的特征。同时，探索如何使县级样本在代表市级总体的同时，也能够反映省级总体的情况。

抽样方法的探索与实施：开发一种高效的抽样方法，用于核算山西省市、县（区）的个体户增加值。这包括利用经济普查清查数据作为抽样框，结合辅助信息，通过模拟训练来筛选出易于实施且成本效益高的抽样策略。

多级抽样的优化：研究如何在多级抽样中优化样本分配，确保每一级样本的代表性，同时减少不必要的重复调查，提高抽样效率。

辅助信息的利用：深入分析清查辅助信息，如行业分布、经济活动类型等，以提高抽样的精度和代表性。

模拟训练与策略筛选：通过模拟训练，评估不同抽样策略的效果，筛选出最优的抽样方法。这包括对样本容量、抽样误差、成本效益等进行综合评估。

资源节约与效率提升：探索如何在保证样本代表性的前提下，减少资源浪费，提高调查效率。这可能涉及到创新的抽样设计、数据收集方法和分析技术。

理论与实践的结合：将理论研究与实际应用相结合，通过实证研究验证所提出抽样方法的有效

性，并根据实际调查结果不断调整和优化抽样策略。

三、学术价值

（一）研究方法特色

本研究在个体户增加值核算抽样方法上的特色显著，主要表现在以下几个方面：

1. 自下而上的分层抽样策略

采用了自下而上的分层抽样方法，这种方法从县（区）个体户出发，依据其所属的行业、经营规模等特征进行分层，确保样本能够全面捕捉县（区）级经济的内在多样性。与传统的自上而下抽样方法相比，自下而上的策略更能反映市、县（区）级的实际情况，提高了样本选择的效率和统计估计的精确度。

2. 双重调查层的划分

基于从业人员期末人数的门槛值，将样本分为全面调查层和抽样调查层。全面调查层包含所有从业人员数量超过一定门槛值的个体户，确保这些对经济总量影响较大的单位被全面覆盖；而抽样调查层则针对从业人员数量较少的个体户进行抽样调查，以优化样本选择过程，确保样本能够准确反映各行业的实际经营状况。这种双重调查层的设计不仅提高了调查效率，也降低了调查成本。

3. 门槛值的引入与优化

通过设置门槛值来区分不同规模的经营实体，这一创新性的设计减少了不必要的样本数量，降低了调查成本。同时，门槛值的优化确保了调查结果的准确性和可靠性。根据山西地区的特点和经济发展水平，经过多次模拟和实地调研，确定了合理的门槛值，确保了样本的代表性和调查结果的准确性。

4. 模拟优化技术与机器学习技术的结合

在抽样调查的抽样方式选择和样本容量的确定等关键问题上，本研究引入了模拟优化技术和机器学习技术。通过模拟训练，获得了大量样本数据，并利用机器学习算法对这些数据进行分析和处理，以获取既定相对误差约束下的最优调查方案。这种技术的结合不仅提高了抽样调查的科学性和系统性，也确保了调查结果的准确性和可靠性。

（二）研究方法突破

本研究在个体户增加值核算抽样方法上的突破主要体现在以下几个方面：

1. 针对县（区）级层面的抽样方案设计

考虑到市、县（区）总体与省总体在经济结构和个体户特点上的差异，本研究提出了适用于县级层面的抽样方案。这一方案充分考虑了县级经济的实际情况和个体户的特点，确保了调查结果的针对性和实用性。这是对现有抽样方法的重要补充和突破。

2. 模拟训练与实际应用的结合

本研究不仅在理论上进行了模拟训练，还将优化后的抽样方案应用于实际的抽样调查中。通过实际调查数据的反馈和验证，不断对抽样方案进行修正和完善，确保了研究成果的实用性和有效性。

这种模拟训练与实际应用的结合方式不仅提高了研究的科学性，也增强了研究成果的可信度。

3. 数据质量控制的强化

本研究通过一系列严格的质量控制措施，结合现代信息技术，如大数据分析和人工智能算法，提高了数据推算的精确度和效率。对调查数据进行了严格的审核和校验，确保了数据的准确性和可靠性。同时，利用大数据分析和人工智能算法对数据进行深入挖掘和分析，提高了数据推算的精确度和效率。

（三）应用价值

本研究在应用价值方面也具有重要意义，主要表现在以下几个方面：

1. 政策制定支持

本研究成果能够为国统局和省统计局优化个体户抽样方案提供借鉴。为政府部门制定经济政策和普查政策，尤其是针对个体户的政策提供数据支持和决策依据。政府部门可以根据研究结果了解个体户的发展状况和对经济的影响，从而制定更加精准有效的政策来推动个体户经济的发展和繁荣。

2. 促进地方经济健康发展

通过优化抽样方法并准确评估不同行业、不同规模个体户的发展状况，本研究为地方经济的健康发展提供了数据保障。政府部门可以根据研究结果制定更加精准有效的政策措施来推动不同行业的发展和升级，从而促进地方经济的整体繁荣和进步。

3. 提升调查效率和成本效益

本研究提出的抽样设计思路和数据处理方法有助于提高调查的科学性和系统性，确保调查结果的准确性和可靠性。同时，通过优化抽样方法和数据处理流程，可以降低调查成本并提高调查效率，为政府部门和研究机构提供更加高效、准确的数据支持，同时也为经济普查工作的开展提供了一些新的思路。

四、创新之处

（一）抽样方法的创新设计

项目在抽样方法上进行了创新，特别是在剔除国家平台样本后，重新设计了抽样方案，这种设计考虑了县级地区的经济特点，在节约成本前提下，获取了县（区）级样本的代表性。以自下而上的思路设计县（区）级样本的抽样方法，进而获得市和省级代表性样本，从而到由县（区）数据逐层向上推断市、省乃至国家的数据。这种自下而上的思路，比国统局五经普和四经普倡导的抽取省级代表性本，省级利用这些样本推断市和县（区）个体户增加值的做法要准确，并且成本也不会明显提高。

（二）自下而上的分层抽样策略

采用了自下而上的分层抽样方法，这种方法从县（区）个体户出发，依据其所属的行业、经营规模等特征进行分层，确保样本能够全面捕捉县（区）级经济的内在多样性。与传统的自上而下抽样方法相比，自下而上的策略更能反映市、县（区）级的实际情况，提高了样本选择的效率和统计估计的精确度。

（三）双重调查层的划分

基于从业人员期末人数的门槛值，将样本分为全面调查层和抽样调查层。全面调查层包含所有从业人员数量超过一定门槛值的个体户，确保这些对经济总量影响较大的单位被全面覆盖；而抽样调查层则针对从业人员数量较少的个体户进行抽样调查，以优化样本选择过程，确保样本能够准确反映各行业的实际经营状况。这种双重调查层的设计不仅提高了调查效率，也降低了调查成本。

（四）门槛值的优化求解

通过设置门槛值来区分不同规模的经营实体，这一创新性的设计减少了不必要的样本数量，降低了调查成本。同时，门槛值的优化确保了调查结果的准确性和可靠性。根据山西地区的特点和经济发展水平，经过多次模拟和实地调研，确定了合理的门槛值，确保了样本的代表性和调查结果的准确性。

（五）模拟优化技术与机器学习技术的结合

在抽样调查的抽样方式选择和样本容量的确定等关键问题上，引入了模拟优化技术和机器学习技术。通过模拟训练，获得了大量样本数据，并利用机器学习算法对这些数据进行分析和处理，以获取既定相对误差约束下的最优调查方案。这种技术的结合不仅提高了抽样调查的科学性和系统性，也确保了调查结果的准确性和可靠性。