Statistics and Application 统计学与应用, 2023, 12(4), 895-900
Published Online August 2023 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2023.124094

Hans汉斯

论统计调查中的误差控制

孙艺嘉[1]，张自然[1]，罗灵茜[1]，尤七七[1]，张丽蓉[1]，涂现峰[2]

[1]嘉兴南湖学院商贸管理学院，浙江 嘉兴
[2]嘉兴南湖学院信息工程学院，浙江 嘉兴

收稿日期：2023年7月10日；录用日期：2023年7月31日；发布日期：2023年8月14日

摘　要

统计调查被应用在生产生活的方方面面，随着当代信息技术的发展，日益成为一项重要的数据分析方法，其全面性、准确性和普遍性使其成为实验的普适选择。而其中，数据质量的控制对于后续的统计分析起着关键性作用。本文将社会调查研究工作与数据分析具体方法相融合，以实际调查“浙江省大学生状态转变的探究”为例，从问卷数据回收分析的流程出发，分别探寻数据筛选、信效度检验、内在逻辑检验、随机性检验四个角度中误差控制的实时体现，通过数据、控制流程以及结论的共同体现，论证统计调查中的误差控制的重要性及其实际应用。

关键词

误差控制，统计调查，数据筛选，随机性检验

Error Control in Statistical Investigation

Yijia Sun[1], Ziran Zhang[1], Lingxi Luo[1], Qiqi You[1], Lirong Zhang[1], Xianfeng Tu[2]

[1]School of Business Management, Jiaxing Nanhu University, Jiaxing Zhejiang
[2]School of Information Engineering, Jiaxing Nanhu University, Jiaxing Zhejiang

Received: Jul. 10th, 2023; accepted: Jul. 31st, 2023; published: Aug. 14th, 2023

Abstract

Statistical survey is used in all aspects of production and life. With the development of modern information technology, it has increasingly become an important data analysis method. Its comprehensiveness, accuracy and universality make it a popular choice for experiments. Among them, the control of data quality plays a key role in the subsequent statistical analysis. This paper combines social investigation and research work with specific methods of data analysis, takes the actual investigation “Investigation on the state transformation of college students in Zhejiang Province” as an example, and starts from the process of questionnaire data recovery and analysis, respectively

文章引用: 孙艺嘉, 张自然, 罗灵茜, 尤七七, 张丽蓉, 涂现峰. 论统计调查中的误差控制[J]. 统计学与应用, 2023, 12(4): 895-900. DOI: 10.12677/sa.2023.124094

to explore the real-time manifestation of error control in four aspects: data screening, reliability and validity test, internal logic test and randomness test. The importance of error control in statistical investigation and its practical application are demonstrated through data, control process and conclusion.

Keywords

Error Control, Statistical Survey, Data Screening, Test for Randomness

1. 引言

统计调查是一项有针对性的，根据调研目的严慎搜集并科学处理调查所需相关资料的统计方法，是一种非常常见的数据分析手段。随着当代新机技术与网络新兴技术的飞速发展，计算机等电子计算器的不断更新迭代，统计调查呈现出了其新的优势和极强的普适性，对于数据的处理更加便捷和直观，并且可消化的数据库的能力越来越强，当下，越来越多的人通过统计调查这一方法整理和解决生活中各方面的数据分析问题，在可查的文献中，与统计调查应用相关的期刊就可达七万多篇，并且不断涉及了日常生活及学术调查的方方面面，为人类的社会实践和社会调查不断提供着强劲辅助，借予其良好的适应性和普适性，不断地发展统计调查自身在人文社科方面的拓展性应用。

此外，误差控制是经由将测量所得量值与参考值之间进行不断的比对处理后得到新的数据库的数据处理过程，在问卷数据回收之后的统计工作中占据着重要位置。统计工作五个阶段中的前四个阶段都是极易产生统计误差的阶段，但由于统计分析工作在统计调查中的关键性，其运行的良好程度是确保统计工作成果准确性的重要保证。而由于统计误差和随机误差的存在，问卷回收得到的数据往往存在较大的波动性，存在无效数据的可能性。因此问卷在回收之后还要进行数据统计分析，并应用误差控制将整个数据进行重新规整和排布，从而更加确保问卷分析所得结果的有效性和准确性。因此，误差控制在问卷分析中占据着关键性位置。

2. 国内外研究评述

国内目前已有的关于统计调查中误差控制的研究成果主要关注误差控制难题，探索提升调查数据质量的方法。在社会调查质量面临误差侵蚀与方法误用的困境下，臧雷振、王栋(2022) [1]认为构建一个以减少整体性调查误差为目标，以增强调查结果适用性为宗旨，以加强对社会调查全流程质量控制为手段，融合多元路径的全面社会调查质量保证框架非常必要。网络调查因其特有的便捷性、经济性、多样性被广泛应用，但在实际应用过程中，公众仍认为网络调查数据可信度较低。钱思汐(2022) [2]研究发现，在网络调查的实际实施过程中，仍然存在样本代表性不足、拒收率高、理解问卷困难等问题，认为可以从“友好”设计问卷、加强被调查者信息保护、科学运用激励方式、多平台邀请应答、加强收回数据质量审核等五方面提升调查数据的可信度，为统计分析提供数据质量保证。张华(2022) [3]关注于数据收集之前的数据质量控制，针对网络调查中的非抽样误差控制提出了界定网络调查的适用范围，科学设计问卷，有效运用随机化问答技术等建议。刘明宇(2022) [4]将“计算机辅助电话调查云系统”应用在社情民意调查环节中，通过严谨而智能的调查方法，减少调查过程中产生的数据误差。在数据质量控制阶段，根据

预设比例提取审核前后数据的综合误差率，定量评估数据质量，同时利用缺失率、信度检验、效度检验等相关方法计算调查数据的可靠程度。

如今多国、多区域和多文化背景下的目的调查(或“3MC”调查)对于全球和区域决策和理论建设变得越来越重要，Lars Lyberg (2018) [5]等学者提出管理和评估 3MC 调查质量的一个关键方法是总调查误差(TSE)框架和相关的调查过程质量，调查举例说明了 PIAAC 可能造成错误的原因，以及通过有效的质量保证(QA)和质量控制(QC)方法解决这些问题的方法。Meyer Bruce D (2020) [6]等学者认为调查误差是普遍存在的，为总结和分析测量误差的程度、来源和后果，他们估计总的调查误差，将其分解为三个高水平的误差源：广义覆盖误差、项目无响应误差和测量误差。定义了可以利用数据组合进行估计的总测量误差框架关键组成部分的经验对应关系。

3. 实例

以具体调查“浙江省大学生状态转变的探究”为例，该调查主要通过问卷数据分析大学生状态转变指数。问卷共有 50 道题目，其中 1~7 题为基本情况调查，8~50 题为量表题目，是模型分析的主要部分，包含大学生学习、心理、生活各方面的测量。通过样框和入样人数的确定均参考统计学的相关理论，严格遵守抽样调查的相关原则，对浙江省各大高校的大学生进行线下和线上的问卷调查，总共收集到原始样本 497 份，经数据质量控制后得到有效样本 391 份。下面从四个方面：数据筛选、内在逻辑检验、随机性检验与量表部分的信效度检验清洗问卷，确保分析前的数据真实有效。

3.1. 数据筛选

数据筛选主要指对问卷数据进行预处理，检查数据中是否含有存在无效问卷、异常值等，从而剔除无效样本数据。通过查阅资料以及阅读文献[7]，将筛选的方法总结为以下几个方面：

1) 时长问题。从全部问卷的做题时长来看，剔除时长过短问卷或过长问卷，时长阈值问题是此步的关键，本研究时长区间设置为预调查成员认真做问卷时长的最大值与最小值，此部分删除样本 72 份。

2) 重复问题。a) 单个样本中，被调查者所选选项重复过多，比如说 90%选 *C*；b) 所有样本中剔除重复样本，此部分删除样本 11 份。

3) 逻辑错误或前后明显矛盾。如学历为专科的选择大四年级，此部分删除样本 1 份。

4) 漏选、多选问卷。根据分析需要剔除，慎重因某条选项缺失直接剔除，如排序题漏选，只需要分析排序题题目存在漏选时，剔除该样本，此部分删除样本 0 份。

在数据正式录入之前，本小组先对收集到的间在信息及其数据进行审核，将含有不完整、不一致(矛盾)、重复(规律)、极端化数据的问卷均归为无效问卷剔除。最终剔除无效样本 84 份，得到 391 份有效问卷。

数据筛选通过对信息的筛查和审核，以最基础的筛查手段帮助研究保留最重要的原始数据部分，从根本上去除了会导致研究结论出现偏离现实和极端值的某种可能，对于后续的信息处理和数据模型的运用都具有积极的筑基作用。

3.2. 信效度检验

1) 良好的信度检验结果

信度(Reliability)即可靠性，是指使用相同指标或测量工具重复测量相同事物时，得到相同结果的一致性程度，调查问卷的信度是为了进一步考察量表所测结果的稳定性以及一致性。我们采用 Cronbach 信度系数来检测问卷的可靠性，对问卷量表题目利用 Cronbach 法进行信度检验。具体结果见表 1：

Table 1. Reliability statistics
表 1. 可靠性统计

克隆巴赫 Alpha	基于标准化项的克隆巴赫 Alpha	项数
0.879	0.880	43

计算各维度评分及总体评价的克隆巴赫系数(Cronbach's α)，一般认为该系数高于 0.8 为信度高；介于 0.7~0.8 之间，说明信度较好。从上表可知：信度系数值为 0.879，大于 0.8，因而说明研究数据信度质量很高。

2) 良好的效度检验结果

效度(Validity)即有效性，指测量工具或手段能够准确测出所需测量的事物的程度。进行效度检验是为了要确定设计的量表题目是否合理，是否能有效反应研究人员的研究目标，测量结果与要考察的内容越吻合，则效度越高。本小组采用常用的 KMO 法对问卷中的量表题目进行效度检验。具体效度检验结果见表 2：

整体效度：

Table 2. KMO and Bartlett tests
表 2. KMO 和巴特利特检验

KMO 取样适切性量数		0.851
巴特利特球形度检验	近似卡方	7488.810
	自由度	903
	显著性	0.000

计算评价 KMO 值，一般认为该值大于 0.8，则说明研究数据非常适合提取。从上表可知，KMO 值为 0.851，Bartlett 球形检验显著性为 0.000，表明问卷具有良好的结构效度，很适合做因子分析。

综上所述，本调查在问卷的设计过程中，各项指标的测量问题和答案基本都是在已有文献的基础上设计的，因此具有较好的理论基础，同时问卷总体的信度和效度均处于较高的可接受范围。此外，本问卷的各个条目由小组成员经过了细致的讨论分析，保证了较好的内容效度和准则效度。

通过信度效度检验，来对矫调查问卷选题的可信度和有效性，并以良好的信度效度检验结果进一步说明调查问卷设计的可行性，推证后续调研分析的可进行性，以及所得结果的可信度和实践意义。

3.3. 内在逻辑检验

逻辑检验主要是运用样本选项中的某些相关性，通过统计数据相互之间的逻辑关系，检验其是否存在误差的方法。与数据筛选中逻辑错误不同，逻辑检验主要通过特定的统计方法以确定两个或多个、两组或两组以上变量间关系，通常用相关性分析、回归性分析等。而逻辑错误是筛选无效问卷的第一道程序，主要判断前后题项的选择是否存在简单的逻辑错误和明显冲突。

通过网上查阅资料以及阅读文献[8]，得到大学生月支出费用与大学生年级有正相关的关系，即认为随着年级的增大月支出费用也会增大，基于此检验样本的内在逻辑是否正确，样本月支出区间从小到大分别赋值 1~4，同样年级从小到大同样分别赋值 1~4，调用 SPSS 运算得到相关性为 0.103，且在显著性水平 0.05 下显著相关，具体结果见表 3，即样本的内在逻辑性成立。

Table 3. Analysis results of correlation between monthly living expenses and grades
表 3. 月生活费支出与年级的相关性分析结果

		月支出	年级
月支出	Pearson 相关性	1	0.103*
	N	391	391
年级	Pearson 相关性	0.103*	1
	N	391	391

通过对问卷的内在逻辑检验，经由对变量间的逻辑检验分析，找到不同控制因素间的关系，以此定夺前后题项的选择是否存在简单的逻辑错误和明显冲突，从而保证问卷数据的逻辑通顺，筛除问卷中逻辑意义上对冲的无效问卷，使得误差控制后的数据具有可采纳性，促进数据的有效性，推动后续数据分析的顺利进行。

3.4. 样本的随机性检验

对样本进行随机性检验，目的是为了检验所得样本是否为特定抽样方法下的随机样本[9]，即在总体中每一个单位被抽取的机会是均等的，不至于出现倾向性误差，是控制抽样调查中数据质量的一种有效手段。如果抽样不满足随机性要求，可能会导致对总体的推断错误，我们对辅助变量生活费支出进行随机性检验。

检验统计量为大学生月生活费用支出的均值，样本量较大时，根据中心极限定理，均值近似服从正态分布。方法：样本均值大于等于某值，这种情况的概率很小(小于等于显著性水平 0.01)，小概率事件不可能发生，故提出假设：

H_0：问卷获得的样本为随机样本；H_1：不是。

Table 4. Calculation table of average and variance of monthly living expenses of sample college students
表 4. 样本大学生月生活支出平均数与方差计算表

月支出(元)	750	1250	1750	2250
频数	14	167	133	77

总体均值为 1601 元，即浙江省大学月平均支出为 1601 元。计算样本均值[10]与样本方差，因月支出费用统计为组距式数列，采用平均值代替对应组，处理结果见表 4，采用加权平均数以及加权方差计算公式，可得样本均值为 1599.10 元，方差为 169046.51 元^2，在原假设下，

$$P(\bar{x}>1599.10)=1-\Phi\left(\frac{1599.10-1601}{\sqrt{169046.51}}\right)=\Phi(0.0046)=0.502>0.01$$

落入接受域中，即在 99%的置信度下接受原假设，即样本满足随机性。

随机性检验的良好结果肯定了样本抽样的机会均等性，肯定了每一个样本被抽取出来的概率相等，由此保障了经由这些样本所得数据结论的普适性，剔除了由于样本抽取个体概率不一致带来的数据结论存在偏向性的可能，对于促进后续数据分析结论的可适用性具有积极意义。

4. 结语

数据质量的控制对于后续的统计分析起到重要作用，目前已有的关于统计调查中误差控制的研究成果主要关注误差控制难题。本文以“浙江省大学生状态转变的探究”为例，从数据筛选、信效度检验、内在逻辑检验三个角度进行误差控制：首先对收集到的初步信息及其数据进行数据筛选，将含有不完整、不一致、重复、极端化数据的问卷均归为无效问卷剔除，筛选后得到进一步分析样本，在将样本通过随机性检验，论证样本是否受到了某些非随机因素的干扰，来检验样本的合理性。随后，再将问卷进至信度效度检验阶段，通过信度检验来监测问卷数据的可靠性与有效性，通过效度检验来确定问卷设计题项的合理性，确定其能有效反映研究人员的研究目标。最终，再通过内在逻辑性检验，检验问卷题设的内在逻辑性。通过数据、控制流程以及结论的共同体现，论证统计调查中的误差控制的重要性及其实际应用。

数据质量控制是整个统计调查的基石，数据质量控制的质量决定了数据分析的有效性和准确性。本文经过对统计数据的层层误差控制，通过不断的试错、调整，得到合理的数据结果，论证误差控制在数据控制阶段的关键性，并认为通过此方法能够助使后续的数据分析阶段更加的顺利，且更易借此实现分析结果的多维度解读。

基金项目

嘉兴南湖学院 2022 年大学生科研训练计划(SRT)，项目代码：8517223215。

参考文献

[1] 臧雷振, 王栋. 社会调查质量: 调查误差、结果适用性及质量控制[J]. 江海学刊, 2022, 340(4): 137-145.

[2] 钱思汐. 网络问卷调查数据质量控制研究[J]. 中国统计, 2022(1): 73-76.

[3] 张华. 网络调查中的非抽样误差——以开放式 Web 调查为例[J]. 统计学报, 2022, 3(5): 70-82.

[4] 刘明宇. 民意调查数据质量控制和方法研究——以“黑龙江省基本公共服务满意度调查”为例[J]. 统计与咨询, 2022(1): 8-13.

[5] Lyberg, L., Hibben, K.C. and Pennell, B.-E. (2018) Applying the Total Survey Error Framework to PIAAC. *Quality Assurance in Education*, **26**, 153-168. https://doi.org/10.1108/QAE-07-2017-0035

[6] Meyer, B.D. and Mittag, N. (2020) An Empirical Total Survey Error Decomposition Using Data Combination. *Journal of Econometrics*, **224**, 286-305. https://doi.org/10.3386/w25737

[7] 李培凯. 九种方法轻松筛选无效作答——及对问卷研究设计的启示[EB/OL]. https://zhuanlan.zhihu.com/p/103352963?from=singlemessage&utmid=0, 2023-03-14.

[8] 秦菊香. 大学生不同年级与月生活费关系的调查研究[J]. 大众投资指南, 2018(17): 262.

[9] 艾小青, 金勇进. 抽样调查下样本随机性的检验[J]. 统计研究, 2009, 26(9): 28-31.

[10] 猎学网. 大一新生每月多少生活费才够花?全国各地大学生生活费排行出炉! [EB/OL]. https://baijiahao.baidu.com/s?id=1739943613451243501&wfr=spider&for=pc, 2023-03-14.