



经济与社会研究丛书

非随机化回答模型 研究

梁敏 著

FEI SUIJIHUA HUIDA MOXING YANJIU



暨南大学出版社
JINAN UNIVERSITY PRESS



经济与社会研究丛书

非随机化回答模型 研究

梁敏 著

FEI SUIJIHUA HUIDA MOXING YANJIU



暨南大学出版社
JINAN UNIVERSITY PRESS

中国·广州

图书在版编目 (CIP) 数据

非随机化回答模型研究/梁敏著. —广州:暨南大学出版社, 2019. 3
(经济与社会研究丛书)
ISBN 978 - 7 - 5668 - 2581 - 0

I. ①非… II. ①梁… III. ①统计调查—调查方法—研究 IV. ①C811

中国版本图书馆 CIP 数据核字 (2019) 第 040735 号

非随机化回答模型研究

FEI SUIJIHUA HUIDA MOXING YANJIU

著 者: 梁 敏

出 版 人: 徐义雄

责任编辑: 曾鑫华 高 婷

责任校对: 林 琼

责任印制: 汤慧君 周一丹

出版发行: 暨南大学出版社 (510630)

电 话: 总编室 (8620) 85221601

营销部 (8620) 85225284 85228291 85228292 (邮购)

传 真: (8620) 85221583 (办公室) 85223774 (营销部)

网 址: <http://www.jnupress.com>

排 版: 广州市天河星辰文化发展部照排中心

印 刷: 广州市穗彩印务有限公司

开 本: 787mm × 960mm 1/16

印 张: 8.5

字 数: 140 千

版 次: 2019 年 3 月第 1 版

印 次: 2019 年 3 月第 1 次

定 价: 32.00 元

(暨大版图书如有印装质量问题, 请与出版社总编室联系调换)

前 言

采用什么样的方法对敏感性问题实施调查，才能做到既能保护被调查者隐私，又能获得所需要的真实资料，是统计调查必须面对和需要解决的问题。随机化回答技术在解决这一问题上取得了丰硕的研究成果。使用随机化回答技术进行调查需要使用随机化装置在现场回答，这些装置的制作较为复杂，试验方法也有较高技术要求，还需要解释说明技术的原理以消除被调查者的疑虑，这种技术不仅操作不方便，而且实施成本高，更重要的是不易推广应用，极大地限制了敏感性问题调查的范围。为了解决以上问题，本书试图从以下三方面对随机化回答技术进行改进。

首先，在已有随机化回答模型的基础上提出四个改进的模型。关于定性问题，提出改进的 Simmons 模型和改进的 Simmons 模型的简化模型。前者介于直接提问模型和 Simmons 模型之间，其精度比 Simmons 模型高；后者是前者的简化，其操作更简便。关于定量问题，提出改进的加法模型和改进的乘法模型。前者是对加法模型的改进，其精度比加法模型高；后者是对乘法模型的改进，其精度比乘法模型高。

其次，对随机化回答技术的经典模型和新提出模型进行改造利用，用无关问题或敏感性问题替代随机化装置，或在问卷中使用独立的非敏感性问题来间接获得所需信息，实现了敏感性问题 的问卷调查。关于定性敏感性问题 的问卷调查设计了九种方法，其中八种用于二分类敏感性问题，一种用于多分类敏感性问题。关于定量敏感性问题 的问卷调查设计了六种方法。

最后，对改进的四个随机化回答模型和设计的去除随机化装置的方法分别讨论了放回简单随机抽样和不放回简单随机抽样两种方式，给出了相应的无偏

估计量、估计量的方差和方差的估计量，并以两个典型模型为代表说明了它们的应用。

由于水平和时间、经费的限制，虽然本书对敏感性问题的调查方法进行了系统的研究，但对于一些重要的问题，比如如何对不同调查方式进行科学比较和如何科学定义隐私的保护度等问题本书尚未进行研究，这些都是我们今后要继续完成的工作。由于笔者的知识和水平有限，书中难免有错误和不足之处，恳请读者批评指正！

梁 敏

2018 年 12 月于北京师范大学珠海分校

目 录

前 言	1
1 绪 论	1
1.1 研究背景和研究意义	1
1.2 研究现状及述评	3
1.3 研究内容与创新点	7
2 经典随机化回答技术介绍	9
2.1 用于定性敏感性问题的经典随机化回答技术	9
2.2 用于定量敏感性问题的经典随机化回答技术	24
3 改进的随机化回答技术	28
3.1 用于定性敏感性问题的改进随机化回答技术	28
3.2 用于定量敏感性问题的改进随机化回答技术	35
4 用于定性敏感性问题的非随机化回答技术	44
4.1 同时询问多个问题以去除随机化装置的模型	45
4.2 用问题替代随机化装置的模型	54
5 用于定量敏感性问题的非随机化回答技术	74
5.1 使用独立非敏感性问题以去除随机化装置的模型	74

5.2 用无关问题替代随机化装置的模型 84

6 应用研究 103

6.1 定性敏感性问题替代装置简化模型的分层整群抽样 104

6.2 应用研究 106

6.3 定量敏感性问题倍数回答模型的分层整群抽样 110

6.4 应用研究 112

6.5 其他模型问题设计 114

7 总结与展望 117

7.1 总结 117

7.2 研究展望 119

附 录 121

参考文献 123

后 记 132

1 绪 论

1.1 研究背景和研究意义

1.1.1 研究背景

在调查中，经常会遇到敏感性问题，即所调查的内容涉及商业机密或个人隐私而不便向外界透露的问题，如考生考试作弊、企业偷税漏税、个人的不良嗜好或行为等。敏感性问题按问题指标特征可分为定性敏感性和定量敏感性问题两类。定性敏感性问题反映敏感性问题的性质和类别，其调查目标是得到具有各种敏感特征的人或事物的比例。早期的敏感性问题调查研究多局限于此类敏感性问题，并将敏感性问题特征分为两类（二项选择）或者多类（多项选择）。在实际生活中，很多敏感性问题我们仅知道其属于哪一类是远远不够的，我们还需了解特定人群或事物中敏感性问题的数量特征，即需要了解定量敏感性问题。定量敏感性问题反映了敏感特性的数值大小，其统计指标主要是均值。

对于敏感性问题，调查中若采用直接调查的方式，很难取得被调查者的信任和配合，往往会引起被调查者的抵触而拒绝回答或故意做出错误的回答，从而大大增加了调查中的非抽样误差。因此对这一类调查必须采用经过特别设计的调查方法，以消除被调查者的顾虑，使他们能够如实回答问题。这种特别设计的非直接提问方式能够使被调查者提供的信息看上去不敏感，从而保护被调查者的隐私，但调查者可以从被调查者的回答中提取出所需信息。

1965 年，Warner 首次提出调查敏感性问题的特殊处理方法，即随机化回答技术。此后，人们通过各种努力来不断修改、完善随机化回答模型，逐步形成用于定性敏感性问题与定量敏感性问题的两大类随机化回答模型。

运用随机化回答技术，虽然可能使估计量的方差增大，导致估计精度下降，但能够得到接近真实的回答，大大消除虚报和瞒报的现象，从而减少非抽样误差。这种以牺牲一定的估计精度换取真实回答从而减少非抽样误差的做法是值得尝试的。为此，本书在前人研究的基础上，又提出几种改进的模型。这些模型在一定条件下比经典模型的精度更高。

但是，使用随机化回答技术和模型方法对敏感性问题进行调查需要在现场使用随机化装置，这不仅增加了调查的成本和操作的难度，而且不适合在大的范围内实施。

随着计算机技术、互联网技术的飞速发展，家用电脑和手机的普及，互联网用户数量的快速攀升，网络在人们的工作、生活中占据着越来越重要的地位，利用互联网进行网上统计调查的需求变得越来越高。如果不使用随机化装置，就可以利用微信等对敏感性问题进行网络调查，那么关于敏感性问题的调查范围就可以大大拓宽，且这种调查方式顺应了大数据时代的要求。敏感性问题网络调查组织方式相对简单，调查数据能够实时上传到数据中心，不仅可以缩短数据录入时间、有效降低数据汇总过程中的人为登记误差，而且可以基于调查系统产生的并行数据开展调查误差分析等多方面的研究与应用。

1.1.2 研究意义

基于以上背景，本书选择去除随机化装置的敏感性问题调查为研究对象，使随机化回答技术可以用于敏感性问题的非现场调查，并对经典的和改进的随机化回答模型进行改造，提出用无关问题或敏感性问题替代随机化装置，或在问卷中使用独立的非敏感性问题来间接获得所需信息。这是一种去除随机化装置的敏感性问题调查方法。

本书的理论意义在于提供了一个提高经典随机化回答模型精度的思路，使常用的随机化回答模型更加有效。这拓宽了随机化回答技术的使用范畴，适用

于不同类型的调查,适应大数据时代统计调查的特点,使随机化回答技术变得更加实用。尤其在敏感性问题非现场调查过程中,调查者只需要运用一定的方法对被调查者进行间接指导,对调查过程做间接控制,从而为敏感性问题调查开辟一条更为便捷的途径。

本书的实践意义在于节省人力、物力,以较小样本量就能达到同样的精度,拓展了敏感性问题的调查范围和领域。在当今大数据时代,数据获取不再局限于常规渠道,数据采集的模式大大扩展,一些传统模式中无法获得数据的渠道在当今大数据时代都可以进行有效的数据提取。通过实施去除随机化装置的敏感性问题调查,可降低调查成本,减轻调查者和被调查者的负担,达到缩短调查时间、提高调查的质量和效率的目的。

1.2 研究现状及述评

考察国内外关于敏感性问题抽样调查研究方面的文献资料可知,最近 50 多年来很多学者针对不同类型的问题,对随机化回答技术和非随机化回答技术进行了广泛而深入的研究,前者需要使用随机化装置而后者不需要。下文对这两方面的研究成果进行简要的回顾和评述。

1.2.1 随机化回答技术研究现状及述评

Warner 提出的随机化回答技术是向被调查者展示两个与敏感性问题有关但完全对立的问题,在他设计的模型中,由于两个调查问题都是敏感性问题,只是提法相反,被调查者可能仍有疑虑而不愿合作。1967 年,Simmons 等人对 Warner 模型进行了改进,提出了随机化回答技术的无关问题法。此后,人们通过各种努力,主要从提高精度和保护度(二者变化方向相反)两方面来不断修改、完善上述两个随机化回答模型,逐步形成用于定性敏感性问题 and 定量敏感性问题的两大类模型。前者具有代表性的模型有 Greenberg 模型、Kuk 模型、Mangat 模型、Christofides 模型等,不仅形成了多种随机化回答模型,而且研究领域也从具有两项选择的随机化回答模型扩展到具有多项选择的随机化回

答模型。此外，有的学者将贝叶斯方法应用于随机化回答技术中，还有学者研究了同时估计多个敏感性特征以及个体不完全如实回答的情况。定量敏感性问题是定性敏感性问题的直接推广，现今很多定量敏感性问题的结果是在已有的定性问题的研究基础上得到的。具有代表性的用于定量敏感性问题的模型有 Greenberg 模型、加法模型、乘法模型、Gupta 模型等。

为了推广应用随机化回答技术，不少学者还将这种技术从简单随机抽样扩展到分层抽样、整群抽样等。

纵观国内外在随机化回答技术研究方面的情况，可以总结出值得继续深入研究的一些发展趋势：

第一，Warner 模型和 Simmons 模型是定性随机化回答模型的基石，其他模型都是在这两个模型的基础上改进得到的。而 Simmons 模型克服了 Warner 模型中的两个缺点。所以，在定性敏感性问题的调查中，Simmons 模型应用较为广泛，但是该模型通过增加无关的非敏感信息来降低敏感信息的比例，虽然可以改善被调查者的合作态度，但最终回答在理论上有 $(1 - p)$ 的概率回答与被调查者无关的问题，这会牺牲这部分样本中关于敏感性问题的信息，从而增大估计量的方差。因此为了使估计量达到同样的精度，必须增大样本量，这样就会耗费更多的人力、物力。如何使最后回答包含更多与敏感性问题有关的信息、减少无用信息是我们需要考虑和解决的问题。

第二，对于定量随机化回答模型，加法模型优于 Greenberg 模型。乘法模型和加法模型实质上是一样的，都是用一些无关的随机数据来干扰真实的数据，使调查者不能从被调查者的回答中确切地推断出其敏感数值的大小，但调查者可以从被调查者所回答的数字中提取出所需信息。在定量敏感性问题的调查中，加法模型和乘法模型应用比较广泛，如何进一步提高这二者的精度是一个值得研究的问题。

1.2.2 非随机化回答技术研究现状及述评

随机化回答技术在敏感性问题抽样调查中的应用十分广泛，例如用于调查学生是否学术不端、运动员是否服用兴奋剂等。无论是哪种随机化回答技术，

其中必有被调查者必须直接面对的敏感性问题,而且回答之前调查者要进行繁杂的规则说明,这样不利于被调查者的合作。为了使随机化回答技术的实用性得到进一步扩展,有少数学者研究了去除随机化装置的敏感性问题调查方法,如条款计数技术 (Item Count Technique)、提名技术、反面问题法以及投影技法等,并将其应用于实践当中。

1. 条款计数技术

1979 年, Raghavarao 和 Federers 首次提出条款计数技术:从总体中有放回地随机抽取两个独立样本,第一个样本面对 G 个无关的非敏感性问题,第二个样本在 G 个无关的非敏感性问题基础上再添加一个敏感性问题,即第二个样本面对的是 $G+1$ 个问题。对于定性敏感性问题,被调查者给出回答“是”的问题数目。具有敏感性特征的人在总体中的比例的无偏估计量为第二个样本与第一个样本的回答“是”的问题数目均值之差。

可以看出,若第二个样本中有人回答“ $G+1$ ”时,则该被调查者肯定具有敏感性特征,被调查者的隐私没有得到保护,因而 2007 年 Chaudhuri 和 Christofides 对该方法进行了改进,对第一个样本增加了一个问题“我具有敏感性特征 A 或不相关的非敏感性特征 F ”,第二个样本的最后一个问题改成了“我不具有敏感性特征 A 或不具有不相关的非敏感性特征 F ”。

对于定量敏感性问题,被调查者给出所有问题答案的和。敏感特性的数值均值的无偏估计量为第二个样本与第一个样本的样本均值之差。

条款计数技术容易被理解,可以用于大规模调查,但其公式推导比较复杂,参数 G 要比较大才能较好地保护被调查者隐私。但是当 G 很大时,要找 G 个无关的非敏感性问题不是件容易的事情。非敏感性问题太多会延长调查时间,使被调查者产生厌烦心理,给出回答“是”的问题数目时容易出错,而且估计量的方差随着 G 的增大而增大。

2. 提名技术

提名技术由 Miller 于 1985 年首次提出,后由 Thompson (1992)、Chaudhuri (2000、2008) 进一步改进。在该方法中,被调查者回答的不是其本人的情况,而是其认识的人中具有敏感性特征的人数,即

$$x_j = A_j / (1 + B_j)$$

其中 A_j 为第 j 个人回答的其认识的人中具有敏感性特征的人数, B_j 为第 j 个人回答的具有敏感性特征的人的朋友数, 则具有敏感性特征的人在总体中的比

例的无偏估计量为:
$$\frac{\sum_{j=1}^n x_j}{n}。$$

这种方法需要很大的样本量, 导致工作量和调查费用增多。

3. 反面问题法

2006 年, Esponda 首次提出反面问题法, 2009 年 Esponda 和 Guerrero 又进一步改进了该模型。反面提问“我不属于”而非“我属于”的好处在于选项中有多个答案符合被调查者的情况, 而被调查者只选择其中一个答案, 因而能够保护被调查者的隐私。例如:

- 我的月收入不是 ()。
- A. 低于 5 000 元 B. 5 000 ~ 10 000 元 C. 高于 10 000 元

若被调查者属于选项 C, 则其会选择 A 或 B, 假设概率都是 0.5, 令 X 为真实选项变量, Y 为被调查者填写的选项变量, 则

$$\begin{aligned} P(Y = A) &= P(Y = A | X = A)P(X = A) + \\ &\quad P(Y = A | X = B)P(X = B) + \\ &\quad P(Y = A | X = C)P(X = C) \\ &= 0 + 0.5P(X = B) + 0.5P(X = C) \end{aligned}$$

同理可得另外两个方程, 于是可求出 $P(X = A)$ 、 $P(X = B)$ 和 $P(X = C)$ 。

这种方法中被调查者选择每个选项的概率未必是相等的, 比如收入属于选项 B 的人可能更倾向于选择 C 而不是 A, 如何确定人们选择每个选项的概率是一个具有挑战性的问题。

4. 投影技法

2004 年, 史锋莘和刘建平从心理学角度出发, 对调查问卷精心设计、巧妙安排, 通过抓住被调查者的心理来获取有效信息。这种方法主要适用于敏感性较低的问题, 通常需要对调查者进行专门的培训。

以上都是不含随机化装置的可以用于非面对面的敏感性问题的调查方法,能由被调查者自行填写,保密性和时效性强,有效地克服了被调查者对敏感性问题的心理障碍。但由上述分析可知,这些方法都各有缺点,而且真正能用于定量敏感性问题的非常少。

1.3 研究内容与创新点

1.3.1 研究内容

本书的研究内容主要包括:

第1章概述全书基本脉络,介绍敏感性问题的特殊调查方法,以及国内外关于随机化回答技术的研究情况,并进行评述,指出存在的问题和值得深入研究的领域,并给出本书研究的主要内容,归纳出可能的创新点。

第2章是随机化回答技术的介绍。分别详细介绍用于定性和定量敏感性问题的经典模型。

第3章提出四个精度比经典模型更高的随机化回答模型。

第4章对经典模型和提出的模型进行改造利用,提出了用于定性敏感性问题的去除随机化装置的调查方法,实现了敏感性问题的问卷调查。该方法不需要调查者亲临现场,不受调查范围、调查规模及调查单位聚散的限制,给出了相应的无偏估计量,推算出估计量的方差和方差的估计量。

第5章提出了用于定量敏感性问题的去除随机化装置的调查方法。

第6章分别用一个实例对用于定性和定量敏感性问题的非随机化回答技术(根据提出模型改编)进行分析。

第7章对本书进行了总结,对问卷调查技术的理论体系及应用前景进行了展望,指出本书研究的不足和待改进之处。

1.3.2 创新点

本书对国内外已有的和笔者改进的随机化回答模型进行改编,对定性敏感

性问题（包括二项选择和多项选择问题）和定量敏感性问题都提出了去除随机化装置的敏感性问题调查方法。

本书的创新点在于：

第一，提出四个精度更高的关于定性和定量敏感性问题的随机化回答模型，给出了相应的无偏估计量，推算出估计量的方差和方差的估计量。

第二，对随机化回答技术的经典模型和笔者提出的精度更高的随机化回答模型进行改造利用，提出用无关问题或敏感性问题替代随机化装置，或在问卷中使用非敏感性问题来间接获得所需信息的九种关于定性敏感性问题的问卷模型，分别讨论了放回简单随机抽样和不放回简单随机抽样两种方式。该方法去除了制作复杂的随机化装置，因而成本低、易于推广应用，适用于不同类型的调查。

第三，提出六种关于定量敏感性问题的问卷调查方法，分别讨论了放回简单随机抽样和不放回简单随机抽样两种方式。该技术不需要随机化装置，和普通问卷调查一样，具有操作简便、调查成本低、效率高、适合于大规模调查等优点，在敏感性问题调查中具有广阔的应用前景。

2 经典随机化回答技术介绍

2.1 用于定性敏感性问题的经典随机化回答技术

2.1.1 Warner 模型

美国统计学家 Warner 于 1965 年首次提出随机化回答模型（即 Warner 模型），开创了敏感性问题调查的先河。该模型的调查设计是向被调查者展示两个与敏感性问题（具有特征 A ）有关但完全对立的问题：一个问题是“你具有特征 A 吗？”另一个问题是“你不具有特征 A 吗？”这两个问题的答案都只有“是”与“否”。实施随机化回答模型需要设计一种随机化装置，使被调查者以已知概率 p 回答第一个问题，而以概率 $1 - p$ 回答第二个问题。重要的是调查者不知道被调查者回答的是哪一个问题，从而保护了被调查者的隐私。

设总体中具有特征 A 的比例为 π_A ，则被调查者回答“是”的概率为：

$$P(\text{是}) = p\pi_A + (1 - p)(1 - \pi_A)$$

考虑一个放回简单随机样本，若 n 个被调查者中有 m 个回答“是”，则 π_A 的一个估计 $\hat{\pi}_A$ 满足：

$$\frac{m}{n} = p\hat{\pi}_A + (1 - p)(1 - \hat{\pi}_A)$$

当 $p \neq \frac{1}{2}$ 时，有：

$$\hat{\pi}_A = \frac{1}{2p - 1} \left[\frac{m}{n} - (1 - p) \right]$$

其方差为：

$$V(\hat{\pi}_A) = \frac{\pi_A(1 - \pi_A)}{n} + \frac{p(1 - p)}{n(2p - 1)^2} \tag{2.1}$$

它的一个无偏估计为：

$$v(\hat{\pi}_A) = \frac{\hat{\pi}_A(1 - \hat{\pi}_A)}{n - 1} + \frac{p(1 - p)}{n(2p - 1)^2}$$

2.1.2 Simmons 模型（无关问题模型）

由于 Warner 模型中两个调查问题都是敏感性问题，只是提法相反，对此被调查者可能仍有疑虑而不愿合作。另外，Warner 模型要求 $p \neq \frac{1}{2}$ ，且当 p 接近 $\frac{1}{2}$ 时估计量的方差增大，而当 p 离 $\frac{1}{2}$ 太远，即接近 0 或 1 时，又会增加被调查者的顾虑。为克服 Warner 模型的上述缺点，Simmons（1967）将敏感性问题的对立问题改成了与调查问题毫无关联的非敏感性问题，总体中具有非敏感性特征 B 的比例 π_B 在设计时已知。

样本中回答“是”的概率为：

$$P(\text{是}) = p\pi_A + (1 - p)\pi_B$$

则：

$$\pi_A = \frac{1}{p}[P(\text{是}) - (1 - p)\pi_B]$$

π_A 的一个估计 $\hat{\pi}_A$ 满足：

$$\frac{m}{n} = p\hat{\pi}_A + (1 - p)\pi_B$$

则：

$$\hat{\pi}_A = \frac{1}{p}\left[\frac{m}{n} - (1 - p)\pi_B\right]$$

其方差为：

$$V(\hat{\pi}_A) = \frac{\pi_A(1 - \pi_A)}{n} + \frac{(1 - p)^2\pi_B(1 - \pi_B)}{np^2} + \frac{(1 - p)(\pi_A + \pi_B - 2\pi_A\pi_B)}{np} \tag{2.2}$$