美国政治研究中的抽样调查方法①

任莉颖

〔内容提要〕抽样调查是研究美国政治的重要观测工具。自 19 世纪末至今，美国政治研究中的抽样调查从起源、发展，到 21 世纪以来遇到挑战。面对概率抽样调查覆盖误差增大、应答率下降和成本上升，非概率抽样调查的兴起，以及来自大数据的竞争等问题，抽样调查研究者们不断创新，正在探索响应式调查设计，非概率样本的统计推断，以及与大数据结合应用等方法。本文采用总调查误差的框架，从测量误差、覆盖误差、无应答误差和调整误差四个方面分析了 2016 年美国总统大选前民调失灵的原因。概率抽样调查、非概率抽样调查和大数据各有自己的主要应用场域，未来的发展中三种数据采集手段会相互校验、融合使用，而高质量的概率抽样调查是衡量非概率抽样调查或大数据质量的参照基准。

关键词：美国政治 研究方法 抽样调查 非概率抽样调查 大数据 响应式调查设计 总调查误差

抽样调查是研究美国政治的重要观测工具，其作用堪比望远镜之于天文学，显微镜之于生物学。② 抽样调查有两种应用形式，一是设计相对简单，题量较少，时效性强的民意调查（polling），常被媒体和政党用来搜集民众想法，反映民众意愿；另一是设计相对复杂，题量较大，实施时间也较长的学术或政策调查，被政治学者们用在挖

① 本文是国家社会科学基金“计算社会科学背景下的政治学研究方法变革研究”（项目号：19BZZ010）的阶段成果。感谢中国社会科学院美国研究所赵梅、王欢两位老师的动议和指导，感谢《美国研究》匿名审稿专家的修改建议，笔者文责自负。

② Henry E. Brady, “Contributions of Survey Research to Political Science,” *PS: Political Science and Politics*, Vol. 33, No.1 (2000), p.47.

掘政治现象的影响机制，比较各国之间的异同，分析随时间变化的趋势等深入的研究上。

当然，抽样调查方法不仅仅应用于政治学，在任何一个学科研究中，只要涉及观察人群特征或个人组成单位（如家庭、机构、企业等）的群体特征，抽样调查都可以成为合适的工具。抽样调查的优势在于不必调查群体中的所有个体，而是概率选取部分个体构成调查样本，通过对这些调查样本特征的分析获知群体的特征。①

本文聚焦抽样调查在美国的发展历程及其所面临的挑战，对抽样调查自19世纪末至今的发展历程进行梳理，借助总调查误差框架分析2016年美国总统大选前民调失灵现象的原因，并对抽样调查近年来遇到的挑战、对策及未来发展进行探讨。

一 抽样调查的发展历程

《公共舆论季刊》（*Public Opinion Quarterly*，又译《民意季刊》）是美国抽样调查研究方面最有影响的期刊，在其创刊50周年（1987）和75周年（2011）之际，分别出版过两期专刊，请抽样调查领域的资深学者撰写文章，从多个方面回顾抽样调查的历史、审视其现状，并预测其将来。1987年，第一本关于抽样调查发展史的专著出版，作者琼·康弗斯（Jean Converse）是密歇根大学社会学系教授，时任密歇根大学社会抽样调查研究中心（Survey Research Center，SRC）底特律地区研究（Detroit Area Study）②项目负责人。她以抽样调查实践者的内部视角和亲身经历记述了抽样调查从1890年到1960年的扎根、成形的发展历程。③ 也许是受到这本书的影响，2011年，美国著名抽样调查专家、时任美国人口普查局主任的罗伯特·格罗夫斯（Robert Groves）教授在文章中将抽样调查发展历史的第一个分界点界定在1960年，而把第二个分界点界定在1990年。④ 本文按照格罗夫斯的分段方法，把抽样调查的发展历史划为发轫期、发展期和迷失期三个发展阶段，并进行评析。

（一）1890年至1960年，抽样调查发轫期

① 抽样调查有广义和狭义之分。广义的抽样调查包括概率抽样和非概率抽样两种形式，而狭义的抽样调查仅指概率抽样的调查。本文采用狭义的定义，对于非概率抽样调查将做出特别说明。

② “底特律地区研究”是美国密歇根大学社会研究院1951～2004年间在大底特律地区进行的年度抽样社会调查。该调查不仅收集了大底特律地区丰富的社会变迁数据，也是密歇根大学社会研究院抽样调查教育的实践田野。

③ Jean M. Converse，*Survey Research in the United States: Roots and Emergence 1890～1960*（University of California Press，Berkeley，1987）.

④ Robert M. Groves，“Three Eras of Survey Research，” *Public Opinion Quarterly*，Vol.75，Issue 5，Special Issue（2011），pp.861～871.

抽样调查是一种相对于普查的方法创新。普查的历史悠久，最早可以追溯到六千多年前的巴比伦，主要目的是清点管辖区域内的人口并掌握他们的基本信息。普查长期以来在美国具有重要的地位，因为美国的国会由参议院和众议院组成。参议院的议员议席按照每州两名的方式分配，而众议院的议员议席则是根据各州的人口数分配。

然而，普查并非易事，耗时、耗资、耗力，各国要成立专业的统计部门来实施这项工作，于是承担这项繁重工作的人想出“偷懒”的方法也是自然而然。这些人中就有挪威中央统计局主任安德斯·凯(Anders Kaier)，他早在1897年出版了一份报告，首次提出了用代表性样本代替普查所有人口的做法。他提出的做法类似于配额抽样，是依据一些辅助信息有目的地选取一个“平衡”的样本，而该样本的各方面特征可以反映出普查的人群特征。这在当时是一个革命的想法，[①]安德斯·凯不屈不挠地到处宣传，却四处碰壁。尽管政府部门对这一方法持谨慎态度，商业调查公司却乐于接受省时、省力、省钱的创新。于是1936年美国大选年发生了抽样调查史上一个里程碑式的事件。大选年预测总统竞选结果是美国政治的一个热点，当时久负盛名的《读者文摘》(*Reader's Digest*)杂志邀请其大量读者参与调查，曾经在1916年到1932年间成功预测当选的总统。在1936年的调查中，《读者文摘》收到了240万名读者的应答，统计结果显示阿尔夫·兰登将胜出。盖洛普公司(Gallup)采用配额抽样的方法，仅根据很少样本的调查，得出了相反的结论。当年大选结果如盖洛普公司预测，富兰克林·罗斯福当选。这个事件在公众中产生了强烈的反响，也影响到政府和学术界，代表性样本的思想开始生根发芽。

抽样调查史上另一个里程碑式的事件，是1934年著名统计学家耶日·内曼(Jerzy Neyman)关于概率抽样论文的发表。这篇文章论证并提供了从样本推断到总体的方法及基于大样本的置信区间估计，为概率抽选代表性样本奠定了理论基础。学者们在这一理论的基础上不断探索，完善概率抽样的步骤，并测试方法的有效性。在美国测试这一方法最佳的环境就是总统竞选，终于在1948年的总统选举中，概率抽样的方法击败盖洛普的配额抽样方法，成功预测杜鲁门获胜。这一事件引起了美国社会科学研究委员会(Social Science Research Council)的重视，提出民意调查应采

① Xiaoli Meng, “Statistical Paradises and Paradoxes in Big Data (I): Law of Large Populations, Big Data Paradox, and the 2016 US Presidential Election,” *The Annals of Applied Statistics*, Vol.12, Issue 2 (2018), pp.685~726. 孟晓犁形象地用搅拌浓汤来类比这种思想，浓汤在充分搅拌后就能够使每一勺都能尝到同样的味道，而与容器的大小无关。在抽样调查中，从总体中选取代表性样本的方法就是起到类似搅拌浓汤的作用。

用更佳技术提高准确性的建议。① 这一建议结束了一段时期以来配额抽样与概率抽样之争，概率抽样成为美国公认的最优调查方法。

与此同时，问卷标准化提问方面也取得了很大进展，社会学和心理学研究者们为此做出巨大贡献。调查是社会学者常用的研究手段，特别是在 19 世纪末的社会改良运动中，诸多社会学者走入伦敦的贫民窟，匹兹堡的工人区等地通过访谈收集了丰富的信息。这些学者对访谈方法既有信心又有经验。然而，当一些商业或私营调查机构想做大量的访问时，聘用的往往是缺乏训练的新手，而研究者发现提问用语和方式对于态度性问题影响尤大，于是他们设计了统一的标准化问题，要求访员严格按照问卷文字提问。在态度性问题的设计上，研究者们开始借鉴心理学上的赋值方法，但又觉得那种赋值方法过于烦琐。1929 年，伦西斯·利克特(Rensis Likert) 在他的博士论文中使用了一种单个问题加上分程度答案的形式，简化了态度性问题的测量，这种方法一直沿用至今。

20 世纪 40 年代到 60 年代是美国抽样调查史上的“黄金时代”。当时的数据采集手段以访员面对面访问和邮寄问卷自填为主，应答率普遍在 70%以上，而且无应答的主要原因是接触不到受访者，而非被拒绝访问。那时访员通常由退休妇女或照顾孩子的专职妈妈承担，她们既有很好的资历，又有对调查的热情。同时，这一时期抽样调查数量较少，人们对此还有较大的新鲜感。②

这一阶段也见证了知名抽样调查专业机构和研究协会的诞生。1941 年，国家民意研究中心(National Opinion Research Center，NORC) 在丹佛大学创办，后来由于创办者哈里·菲尔德(Harry Field) 的意外去世，该中心由克莱德·哈特(Clyde Hart) 接任而在 1947 年转到芝加哥大学。1946 年，密歇根大学抽样调查中心在利克特的领导下组建。两个机构后来分别承担了美国两个重要的抽样调查项目，一个是社会学的综合社会调查(General Social Survey，GSS)，另一个是政治学的美国全国选举调查(American National Election Studies，ANES)。1947 年，在哈里·菲尔德的倡议下，一些致力于民意研究的抽样调查先锋创办了美国民意研究协会(American Association for Public Opinion Research，AAPOR)，并于次年出版发行了至今仍有重要影响力的专业杂志《公共舆论季刊》。

回顾这一阶段，从凯到内曼，从乔治·盖洛普(George Gallup) 到利克特，这些人物对抽样调查从无到有、从被拒绝到被接受，发挥了重要的作用。他们对抽样调查的

① Martin R. Frankel and Lester R. Frankel, “Fifty Years of Survey Sampling in the United States,” *Public Opinion Quarterly*, Vol.51, Issue 4 (1987), pp.S127~S138.

② Eleanor Singer, “Reflections on Surveys' Past and Future,” *Journal of Survey Statistics and Methodology*, Vol.4, Issue 4 (2016), pp.463~475.

满腔热情也源于社会关怀，相信自己在为营造更好的社会创造有用的工具。如盖洛普在《民主的脉搏》一书中提出通过抽样调查来反映人民的声音，而利克特也曾针对当时新政集权化的形势下政府官员远离民众，提出通过抽样调查了解民意的解决方案。①

（二）1960 年至 1990 年，抽样调查发展期

1987 年，"综合社会调查"的创始人詹姆斯·戴维斯（James Davis）在评论康弗斯的《美国抽样调查：扎根发芽，1890～1960》一书时，提出他期望的下一部著作题为《抽样调查 1960～1990：大寒潮》。②戴维斯这一论断有些过早，在 1960 年至 1990 年的 30 年里，抽样调查虽然度过了蜜月期的火热，但相比 1990 年以后，还只是渐生寒意。

在这一阶段，技术促进了抽样调查的突飞猛进。为了保证样本的代表性，抽样调查最重要的是建构一个定义明确、无遗漏无重复的抽样框。对于家庭调查来说，抽样框就是一个完备的家庭列表。如果出现遗漏，则会有人或家庭没有机会被访问到（称之为覆盖误差），这会严重影响到抽样调查的代表性。美国早期抽样调查的家庭列表主要有两种来源：一是地理区域（如行政单位、普查地区划分等）内的住户名单或住址列表；二是商业公司编辑的电话号码簿中的住宅电话号码。在住宅电话未能完全普及时，第二种来源的抽样框显然会有严重的覆盖误差。因此，"黄金时代"的高质量抽样调查仅采用第一种方法建构抽样框，只有一些商业或私营调查机构使用住宅电话号码组织访问。

20 世纪 60 年代末，美国住宅电话得到了普及。由于使用住宅电话号码建构抽样框效率高，费用低而得到越来越多调查机构的青睐。在没有统计理论支撑的情况下，政府和学术调查往往采取观望的态度，而调查统计学家们则有巨大的压力为这一实践赋以合理性。一种建立在概率抽样理论上的电话号码随机抽样的方式诞生了。这种方法考虑美国电话号码的构成，前六位数字对应特定的地理区域，于是随机生成后四位号码，实现电话号码的随机抽选。同时计算机技术也有了进步，并开始应用在电话访问上，被称为计算机辅助电话访问（Computer-Assisted Telephone Interviewing，CATI）。

这一时期调查研究者将认知心理学理论和方法应用到问题用语和问卷结构等方面的测试中，这方面研究成果最为丰硕的是密歇根大学霍华德·舒曼（Howard Schuman）教授，他利用在抽样调查中嵌入随机分组的实验方法（称为调查实验），研究了

① Robert M.Groves，p.864.

② James A. Davis，"Book Review：Survey Research in the United States：Roots and Emergence，1890～1960，" *Public Opinion Quarterly*，Vol.53，Issue 1（1989），pp.136～138.

问题顺序①、开放或封闭的答案选项设计②、态度问题中设立中间选项③等做法对受访者应答的影响。

这一时期美国联邦政府和研究基金会投入了大量资金用来支持抽样调查，多个全国大规模纵贯追踪调查都是在这一时期启动的。如 1968 年的“收入动态追踪调查”、1969 年“全国教育进展评估”、1972 年“综合社会调查”，以及 1973 年“全国刑事犯罪调查”（后改名为“全国刑事犯罪受害者调查”）等。政治学领域的大型调查也不胜枚举，如“五国公民文化研究”（1969）、“欧洲晴雨表调查”（1970）、“国际社会调查项目”（1985）等。有学者统计，美国联邦政府在 1984 年批准资助 131 个家庭或个人调查项目，涉及受访者 264.7 万人；1989 年批准资助 162 个家庭或个人调查项目，涉及受访者 264.8 万人。④

抽样调查为了解大规模民众或群体的态度及变化打开了大门，成为用政治事实阐释政治科学的重要工具。⑤ 与早期盖洛普对于抽样调查与民主关系的乐观态度不同，这一时期政治研究者们深深被抽样调查中的一些发现所困扰。美国式民主制度标榜民治、民有、民享的政府，公众是如何参与政治及通过投票选择执政方成为政治学研究者们最迫切想要了解的问题。在发轫时期末，已经有学者指出选民在投票时对竞选各方的政策观点并不了解，⑥而美国著名政治学家菲利普·康弗斯（Philip Converse）在 1964 年发表的文章《公众信仰体系的本质》（The Nature of Belief System in Mass Publics）所引发的争论延续至今。⑦ 这篇文章记录了公众在政治议题上态度的分化：一些人受信仰体系所限，观点清晰，且很难改变；而另一些人则信仰体系不清，态度不稳定，且随机变化。早期有研究认为这个发现仅体现了 20 世纪 50 年代的公众特征，20 世纪 60 年代末 70 年代初，特别是在反越南战争的时期内，美国公众对

① Howard Schuman, Stanley Presser, and Jacob Ludwig, “Context Effects on Survey Responses to Questions About Abortion,” *Public Opinion Quarterly* Vol.45, Issue 2 (1981), pp.216~223.

② Howard Schuman and Stanley Presser, “The Open and Closed Question,” *American Sociological Review*, Vol.44, Issue 4 (1979), pp.692~712.

③ Stanley Presser and Howard Schuman, “The Measurement of a Middle Position in Attitude Surveys,” *Public Opinion Quarterly*, Vol.44, Issue 1 (1980), pp.70~85.

④ Stanley Presser and Susan Mcculloch, “The growth of survey research in the United States: Government-sponsored surveys, 1984~2004,” *Social Science Research*, Vol.40, Issue 4 (2011), pp.1019~1024.

⑤ Henry E. Brady, p.48.

⑥ 参见 Samuel A. Stouffer, *Communism, Conformity & Civil Liberties: A Cross-Section of the Nation Speaks Its Mind* (New York: Doubleday, 1955)。

⑦ Philip E. Converse, “The Nature of Belief Systems in Mass Publics,” Reprinted in *Critic Review*, Vol.18, Issue 1~3 (2006), pp.1~74.

于政治的知情程度很高。① 然而，到 20 世纪 80 年代，大多数学者已经认同了康弗斯的观点，进而考虑这种“无态度”(non-attitude) 的情形对于民主质量的影响。争论主要围绕两条主线，一条是针对“无知的理性”(rational ignorance) ，认为吸收并储存信息的成本大于收益时，选民的“无知”是理性选择的后果，但是选民们在做投票决定时用认知捷径来弥补其信息的不足;②另一条主线接受个人层面的无知，但是所有选民作为一个整体却显示出“聚合理性”③。2006 年，《评论家评论》(*Critic Review*) 杂志组织论坛并出版专刊，十几位资深政治学者著文评论康弗斯 1964 年的这篇文章，康弗斯本人也做了回应。由此可见，康弗斯的早期发现对美国政治研究影响深远。④

当美国政治学者们在为丰富的研究数据欢欣鼓舞时，一个现象引起了抽样调查者们的注意，就是受访者的应答率在下降。概率抽样理论的假定是所有样本都被调查到，早期抽样调查的高应答率使得调查统计学家们没有为此担忧。于是，此时调查者们最想证实的是: 无应答现象是否已在抽样调查中普遍出现，以及是什么原因导致这一现象。无应答研究成为抽样调查下一个时期的重点。

总的来看，抽样调查在 20 世纪 60 年代至 90 年代的 30 年里巩固了其在社会科学研究领域中的地位，虽然已出现了一些令人不安的迹象，如应答率的下降和调查成本的上升，但忧患主要来自内部。然而，进入 20 世纪 90 年代以后，科技的进步带来了外部的挑战，抽样调查一时迷失了方向。

(三) 1990 年至今，抽样调查迷失期

不幸的是，调查者们的研究发现全球的抽样调查都不同程度地显示出应答率下降的趋势，直接原因是受访者拒绝接受访问的比例增加，⑤而社会资本下降是造成这一现象的重要社会原因⑥。

---

① 参见 Norman H. Nie and Kristi Andersen, “Mass Belief Systems Revisited: Political Change and Attitude Structure,” *The Journal of Politics*, Vol.36, Issue 3 (1974) , pp.540~591。

② 参见 Henry E. Brady and Paul M. Sniderman, “Attitude Attribution: A Group Basis for Political Reasoning,” *American Political Science Review*, Vol.79, Issue 4 (1985) , pp.1061~1078; Michael X. Delli Carpini and Scott Keeter, *What Americans Know about Politics and Why It Matters* (New Haven, CT: Yale) 。

③ 参见 Philip E. Converse, “Popular Representation and the Distribution of Information,” in John A. Ferejohn and James H. Kuklinski eds., *Information and Democratic Processes*(Chicago: University of Illinois Press, 1990) ; Benjamin I. Page and Robert Y. Shapiro, *The Rational Public Fifty Years of Trends in Americans´ Policy Preferences* (Chicago: University of Chicago Press, 1992) ; Larry M. Bartels, “Uninformed Votes: Information Effects in Presidential Elections,” *American Journal of Political Science*. Vol.40, Issue 1 (1996) , pp.194~230。

④ *Critic Review*, Vol.18, Issue 1~3 (2006) .

⑤ Edith De Leeuw and Wim de Heer, “Trends in Household Survey Nonresponse: A Longitudinal and International Comparison,” in Robert M. Groves, Don A. Dillman, John L. Eltinge, and Roderick J. A. Little eds., *Survey Nonresponse* (New York: Wiley, 2002) , pp.41~54.

⑥ J. Michael Brick and Douglas Williams, “Explaining Rising Nonresponse Rates in Cross-Sectional Surveys,” *Annals of The American Academy of Political and Social Science*, Vol.645, Issue 1 (2013) , pp.36~59.

当时应答率的计算方式多样，常常会造成不同调查项目之间无法比较。因此美国民意调查研究协会在 1998 年出版了《标准化定义：抽样调查案例代码的最终配置与结果率》，[①]制定了统一的应答率计算标准。调查者们努力寻找办法来提高应答率，他们尝试了培训访员转化拒访的技巧、调配不同特征的访员，提升受访者的酬金等方法，然而这些措施不但没有改变应答率下降的趋势，反而使调查成本越来越高。

技术进步加剧了抽样调查的困境，甚至导致了抽样调查的退步。移动电话普及后，一些家庭不再安装使用住宅固定电话，造成原有的基于住宅电话的抽样框出现严重的覆盖误差。然而对于移动电话，一个家庭可能不仅只有一个号码，甚至一个人也可能拥有多个号码，而且移动电话和居住区域之间不是完全对应，因此，仅依靠移动电话号码建构抽样框会产生更为严重的问题。这时，严谨的抽样调查不得不退回到基于邮递系统的住址列表来选取代表性调查样本，无力承担昂贵调查费用的商业调查公司在寻找新的替代方案。

同时互联网技术的发展催生了网络问卷调查，这种调查属于最为传统的受访者自填调查模式，类似于早期邮寄问卷调查。与纸版问卷相比，网络问卷问题形式更为丰富，不仅包括文字性问题，也可以插入音频、图片和视频等多媒体信息；网页上可以设定问题的跳问路径、弹出问题的帮助信息，以降低在没有访员引导情况下可能出现的填答错误；网络问卷还可以随机设定问题或选项的顺序，避免顺序效应带来的测量误差。然而，这种调查模式的致命弱点是无法确定抽样框。政府或学术调查会从邮政地址列表中抽取代表性样本，将网络调查的链接通过邮件发送给选中的家庭，然后采用电话或真人到访的方式进行补访。一些商业调查公司则走上了当年《读者文摘》的老路，在网站上推送链接，网民自愿参与调查。还有一些调查公司建设网络调查样本库，主动招募网络调查的志愿者，登记他们的基本社会人口信息，然后采用配额抽样的方法发送调查链接。2010 年，美国民意研究协会宣布尽管这样的样本库有一些用处，“当研究目标是为了精确地估计总体参数值时，研究者应该避免使用在线非概率样本库”。此外，在网络调查自制（DIY）工具的辅助下，似乎人人可以做调查，抽样调查的专业化被漠视，抽样设计被忽略，测量设计上也鱼目混珠，抽样调查被“游戏化”或“娱乐化”。[②] 抽样调查统计学家们又遇到和当年电话调查普及时同样的压力，就是如何为这种非概率抽样的网络调查提供理论支持，于是非概率抽样调查的统计推断问题成为这一阶段的研究热点。

---

① 该标准之后不断更新，最新发布的定义是 2016 年第 9 版。

② Mick P. Couper, “Is the sky falling? New Technology, Changing Media, and the Future of Surveys,” *Survey Research Methods*, Vol.7, Issue 3 (2013), pp.145~156.

这一时期出现的另一个“复兴”是调查实验，就是通过把调查样本随机分配到实验组和对照组，将实验设计嵌入抽样调查中。① 如前所述，这一方法早就被应用在抽样调查方法的比较研究中。与传统的实验室实验相比，调查实验的被试（样本）是概率抽选，在概率论的支持下可以将实验结果推论到更大的总体（称之为外部效度）；而且被试（样本）数量大，同质性低，也提高了实验结果的有效性（称之为内部效度）。当经济学的触角伸入抽样调查中，用调查数据分析因果机制成为重要的需求。政治学者们在这方面深受影响，当计算机辅助调查的技术得到应用，对样本的随机化分配成为易事，他们便马上利用这一技术优势，将调查实验嵌入大型抽样调查中。他们还发明了一种测量敏感问题态度的实验，称之为列举实验（list experiment）。这种方法是将问卷分为两个版本，随机分配给受访者。一个版本的问卷中包括一组有关态度或行为的常规问题，另一个版本的问卷中同样包括这些问题，但多出一条关于态度或行为的敏感问题。通过比较两组问题的均值，就可以得出敏感态度或行为的发生比例。美国政治学家保罗·辛德曼（Paul Sniderman）是推动将调查实验运用到美国政治研究中的核心人物。在他自己关于种族偏见与歧视的研究中，调查实验是重要的研究方法。② 他还申请到美国自然科学基金（NSF）的资助，创建了社会科学分时实验室（Time-sharing Experiments in the Social Sciences，TESS）。社会科学分时实验室采用一个调查项目搭载多个调查实验的方法，公开征集调查实验的研究计划，并搭建了拥有全国代表性样本的网络调查平台采集调查实验数据。

互联网、物联网、社交媒体的普及，开辟了数据采集的新阶段。任何电脑、移动设备或传感器上的操作都可以被机器自动记录、存储或传输，产生了巨大数量的数据，被称之为“大数据”。大数据给人的感觉是可以记录下任何人所做的任何事。这种情况下既无须抽样，也无须调查，数据已经在那里了。一时间，抽样调查仿佛遇到了“灭顶之灾”。然而，大数据这个巨人对于社会科学研究也有诸多羁绊。一是大数据并非是理想中的总体数据，总是有一些人会被有意无意地排除在这些设备或网络之外，不同的人被机器捕捉到数据的概率不同且不知。与非概率抽样的网络调查数据相似，大数据虽在规模上取胜，但同样不能推论总体；二是大数据是有机产生的，或称“有机数据”，③数据量虽大，但信息含量低，噪音干扰多，数据处理不易。对于研究者来说属于“二手数据”，如果不清楚数据产生的机制，很容易得出错误的结论；三是大

① 任莉颖《用问卷做实验：调查-实验法的概论与操作》，重庆大学出版社，2018 年版。

② 参见 Paul M. Sniderman and Thomas Piazza，*The Scar of Race* Cambridge（MA：Harvard Unversity Press，1993）；Paul M. Sniderman and Edward G. Carmines，*Reaching beyond Race*（Cambridge，Mass：Harvard University Press），1997.

③ Robert M. Groves，p.866.

数据并非公共资源，大多掌握在商业公司或私营机构中，在很大程度上成为谋利的私有财产，而无意于帮助理解社会。虽然如此，对于研究者来说，大数据获取相对容易，成本也低，具有很强的吸引力。

在这一阶段，抽样调查遇到了严峻的挑战，分别是：第一，应答率的下降及调查成本的上升；第二，非概率抽样调查的死灰复燃；第三，大数据的横空出世。抽样调查研究者们没有姑息待命，下部分将重点介绍他们在这方面的努力。

二 抽样调查的新探索

最糟糕的是，我们越是哀叹我们的营养食谱日益被忽视，快餐就越被生产、消费，甚至被誉为未来时代的美食。事实上，我们一些最有经验的厨师正在不知疲倦地工作，以保持我们悠久的烹饪技能，而其他人在为快速烹饪的比赛做准备。

——孟晓犁①

美国著名华裔统计学家孟晓犁借用这个比喻来形容抽样调查所处的境地。的确，一批献身于抽样调查的研究者们正在致力维护抽样调查的声誉，开发适应现时代特征的应用策略，以保证抽样调查数据的质量。与此同时，他们积极面对新的形势，顺应历史潮流，探索非概率抽样调查统计推断的可能性以及与大数据的互利互惠。

（一）概率抽样调查的自救：响应式调查设计

虽然可以简单地认为，抽样调查应答率低，概率抽样就不能保证提供对于总体特征参数的无偏估计，但是应答率低到何种程度才能破坏推断的有效性却一直没有明确答案。抽样调查应答率的降低促使调查研究者们不得不认真考察无应答率与无应答偏差的关系。格罗夫斯在 2006 年和 2008 年两次发表论文证明无应答率和无应答偏差没有直接联系。如他和同事利用 59 项研究中的 959 个估计值进行分析，发现无应答率与这些估计值偏差相关系数仅在 0.20 左右，只有在调查变量与应答倾向高度相关的情况下，无应答率才会影响到无应答误差。因此，同一个调查内的不同变量的无应答误差是不同的。② 美国抽样调查专家迈克尔·布里克（Michael Brick）和罗杰·图兰吉（Roger Tourangeau）利用同样的数据进一步分析发现，当把这些估计值偏差按照所属研究进行汇总时，可以发现无应答率与无应答偏差在调查之间存在较强的

① Xiaoli Meng，p.686.

② Robert M. Groves and Emilia Peytcheva，"The Impact of Nonresponse Rates on Nonresponse Bias：A Meta-Analysis，" *Public Opinion Quarterly*，Vol.72，Issue 2（2008），pp.167～189.

相关性。也就是说，较高应答率的调查，研究变量的总体偏差相对较低[①]。

针对越来越多的家庭或个人不愿意参与抽样调查，格罗夫斯等利用计算机辅助调查能够获取并及时提供关于调查过程的数据（称之为并行数据）的便利，在2006年提出了“响应式调查设计”（responsive survey design）的思路，其基本框架包括以下四个方面[②]：

预先确定一组可能会影响到调查成本和误差的设计特性；

针对设计特性，确定一套测量成本和误差属性的指标，并在数据采集的最初阶段监测这些指标；

在权衡成本和误差得失的基础上，在后续阶段改变设计特性；

将不同阶段的数据组合成最终的数据集。

研究者们在这个框架的基础上进行拓展。一个研究取向是在调查开始，根据抽样框或其他关于样本的辅助数据，对不同的人群总体分派不同的调查操作指示，这种方法也被称为“适应式调查设计”（adaptive survey design），有别于在调查开始后基于前一阶段的情况进行修改的响应式设计。另一个研究取向是不再划分为独立的阶段，而是在全过程中根据需要进行调整，这种做法被称作“动态调查设计”（dynamic survey design）。[③]

这些设计共同关注的四个元素是：设计特性、辅助数据、质量和成本的测量指标，以及质量-成本的优化。[④] 辅助数据是测量指标设计的基础。按照所利用的辅助数据，这些指标可分为三大类：第一类是应答率，仅依据受访者是否应答即可计算；第二类指标除了应答率，还加入了抽样框数据和并行数据，如R指标（R indicator）和分组应答率的变异系数；第三类指标比第二类指标又增添了调查数据，如缺失信息率等。[⑤] 其中，R指标尝试采用模型的方法来预测受访者的应答倾向，[⑥]对辅助数据的

---

① J. Michael Brick and Roger Tourangeau, “Responsive Survey Designs for Reducing Nonresponse Bias,” *Journal of Official Statistics* Vol.33, Issue 3 (2017), pp.735~752.

② Robert M. Groves and Steven G. Heeringa, “Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs,” *Journal of The Royal Statistical Society Series A-statistics in Society*, Vol.169, Issue 3 (2006), pp.439~457.

③ Roger Tourangeau et al. “Adaptive and Responsive Survey Designs: A Review and Assessment,” *Journal of The Royal Statistical Society Series A-statistics in Society*, Vol.180, Issue 1 (2017), pp.203~223.

④ Asaph Young Chun, Barry Schouten, and James Wagner, “JOS Special Issue on Responsive and Adaptive Survey Design: Looking Back to See Forward - Editorial: In Memory of Professor Stephen E. Fienberg, 1942~2016,” *Journal of Official Statistics*, Vol.33, Issue 3 (2017), pp.571~577.

⑤ James Wagner, “A Comparison of Alternative Indicators for the Risk of Nonresponse Bias,” *Public Opinion Quarterly*, Vol.76, Issue 3 (2012), pp.555~575.

⑥ 任莉颖、邱泽奇、丁华、严洁：《问卷调查质量研究：应答代表性评估》，载《社会》，2014年第1期，第196~214页。

来源和质量要求更高。如联系记录和访员观察等并行数据极易产生测量误差，会减弱他们与调查变量或受访者应答倾向间可能存在的关系。

抽样调查研究者们通过真实的项目、实验或仿真模拟对响应式调查设计的效果进行评估。总览这些研究，图兰吉等人得出以下结论①：

第一，调查方案的重大变化（如更短的问卷、更大的激励措施或转为面访），与简单坚持一贯的数据采集方案相比，更有可能减少无应答偏差。但是，在调查预算不断缩减的时代，所有调查都很难减少无应答偏差。

第二，尽管许多尝试使用倾向模型来提高数据收集的效率，但是无论是提高应答率还是对降低应答倾向的变异，收获甚微。究其原因有几个方面：（1）受当时的调查环境所限，取得显著进展的难度较大；（2）辅助变量对于应答倾向的预测力不足，导致倾向模型对于数据采集没有起到有效的指导作用；（3）即使模型准确地预测了应答倾向，实地执行时不一定选择了有效的干预措施；（4）即使选择了有效的干预措施，却不能有效地监控访员忠实地执行指令。

第三，相比依据同样的辅助变量对调查数据进行事后加权，在数据采集时利用响应式调查设计实现样本的平衡不仅有助于降低偏差，还可以减少加权对调查估计值方差的影响。

响应式调查设计被认为是现代抽样调查的核心技术。为此，密歇根大学抽样调查中心设置了专门的暑期培训课程，用以推广这个技术在美国及全球抽样调查实践中的应用。

（二）非概率抽样调查的希望：统计推断

研究者们一边努力寻找挽救概率抽样调查的良方，一边重新审视死灰复燃的非概率抽样调查。2011 年，美国民意研究协会任命了一个特别工作组，由美国国内知名抽样调查专家组成，“研究在何种情况下，不使用概率样本的各种调查设计仍可用于推断更大的总体”。②

这里的“推断”指的是“统计推断”，用工作组给出的定义是：对总体特征进行估计，并且对这些估计的可靠性提供某种度量的一组程序。这组程序要基于理论和明确的假设，那些没有理论基础而收集数据并做出估计的方法不能用作统计推断。例如街角拦访、网上自愿参与的方便抽样（convenience sampling）调查在进行估计时如果没有任何基于理论的调整，是与统计推断无份的。

---

① Roger Tourangeau et al., pp.219~223.

② Reg Baker et al., “Summary Report of the AAPOR Task Force on Non-probability Sampling,” *Journal of Survey Statistics and Methodology*, Vol.1, Issue 2 (2013), pp.90~143.

非概率抽样被孟晓犁比喻为“快餐”，能快速满足人的需要，但含有对人体有害的成分。这些“有害的成分”表现为：(1) 有部分人被排除在调查之外，导致严重的覆盖误差；(2) 受访者自愿参与，导致自我选择的偏差；(3) 高水平的无应答率。虽然在网络和大数据时代，非概率抽样调查可以在短时间内采集到大量的数据，然而也无法降低对总体估计值的偏差。孟晓犁提出，估计值的偏差是三个部分数值的乘积：第一部分是数据质量测量，表现为研究变量 X 与样本应答指标 R 的相关系数；第二部分是数据数量测量，表现为(N-n)/n 的平方根，其中 N 是总体规模，n 是样本规模；第三部分是问题难度测量，采用 X 的标准差。由此可见，首先，估计值偏差并非是样本规模的函数，而是样本规模与总体规模比值的函数；其次，样本相对规模对估计值偏差的影响会同时受到质量测量和难度测量的制约。因此，在没有考虑数据质量的情况下，样本量的大小不能决定估计值的准确程度，反而会出现“大数据悖论”，即“数据越多，我们越容易欺骗自己”(the more the data, the surer we fool ourselves)。①

提升非概率抽样调查的数据质量是让这一方法获得新生的唯一希望。在抽样调查中，利用样本估计进行统计推断有两种不同的思路，分别是基于设计的估计和基于模型的估计。概率抽样调查属于基于设计的估计方法，随机化抽样设计保证每个样本的入选概率是可知的，入选概率的倒数就是样本的权重，在估计时通过权重把样本还原为总体，从而实现统计推断的功能。基于模型的推断将抽样调查中的有限总体视为特定形式的超总体的一次随机实现，数据产生的机制可以通过超总体模型加以刻画，利用抽样调查获取的样本观测数据进行拟合，对没有观测到的变量值进行预测，从而实现对总体的统计推断。②

对于非概率抽样调查，没有一个严格的随机化抽样设计，但是可以通过一些干预实现“准随机化”(quasi-randomization)。第一个方法是计算出样本的伪包含概率(pseudo-inclusion probability)，转化为权重，用来纠正选择偏差。具体做法是选取一个供参考的调查(reference survey)，可以是质量上可信的可公开获取的概率抽样调查数据集，也可以是调查机构并行实施的概率抽样调查，要求是作为参考的调查要与非概率抽样调查都含有与研究变量高度相关的协变量。将参考数据集里的样本和自愿参与调查的样本混合在一起，根据共同的协变量拟合模型来预测作为非概率样本的概率，转换为伪权重。如果只需要对非概率样本进行分析，则使用这个伪权重；如果概率样本和非概率样本合并使用，还需对伪权重和概率样本的权重进行标准化，确

① Xiaoli Meng, p.686.

② 金勇进、郝一炜《非概率样本的模型推断》，载《数学的实践与认识》，2019 年第 5 期，第 246~255 页。

保合并后的权重之和接近总体规模。①

第二个方法是样本匹配(sampling match)。样本匹配的重点也是选择参考数据源,将非概率样本的背景特征与目标总体进行匹配。参考数据源可以是普查数据,也可以是推断目标总体的高质量的概率抽样调查数据。传统的配额抽样就是一种简单的样本匹配。这种方法从普查数据中选取一些社会人口属性,如性别、年龄、受教育程度等变量作为协变量,然后根据这些协变量的交互分层来分配样本,实现样本在这些协变量上的构成与总体相似。样本匹配方法的关键是要找到和研究变量相关的协变量,然而不同的研究主题,相关的协变量不尽一致,而且协变量的数量也可能是多个。于是,研究者开发出用倾向值进行匹配的方法。具体做法是从参考数据源中抽取一个随机样本,这个样本可以看作推断目标总体的概率样本,这个概率样本需包含和研究变量相关的重要的协变量信息。然后,根据这些协变量,通过倾向值匹配的方法,从非概率样本中选取匹配样本。最后,利用匹配样本的调查数据实现对总体的估计。也就是说,通过匹配的方法,使匹配样本与概率选取的目标样本有相似的性质,因此可以根据匹配样本对目标总体进行推断。②

第三个方法是链接跟踪网络抽样方法(link-tracing network sampling),适用于有社会联系的没有可得抽样框的特殊人群的抽样。如应答者驱动抽样(Respondent-Driven Sampling)就是这样一种方法。具体做法类似滚雪球抽样,也是通过前一个应答者来招募下一个受访者,不同的是这种方法对于招募的路径及每个应答者招募的人数有所限定,并且利用统计方法进行评估,直至达到某种"均衡"即可结束调查。这种方法在满足一些假定的情况下可以获取接近概率抽样的样本。但在实际中这些假定很难得到满足,即使得到满足了,估计值的方差也可能相当高。③

上述三种方法都是基于设计的估计思路,是从总体选择样本,通过样本来反射总体;另一种思路是基于模型的估计,不考虑样本的选择机制,而是用样本来预测总体。模型估计的假定类似抽样调查中数据的随机缺失(Missing at Random, MAR)机制,认为在控制住一系列协变量的情况下,样本与非样本在研究变量的特征上是相似的,因此通过利用样本数据,纳入这些协变量拟合模型,模型的参数可以用来预测非样本或总体的特征。常见的例子就是事后的校准权重,如采用普查数据中的性别、年龄、受教育程度等变量构建的分层(poststratification)或倾斜(raking)权重。最新的方法则是通过建构回归模型、倾向值模型,或多层次回归模型,以及采用贝叶斯分析方法

① Michael R. Elliott and Richard Valliant, "Inference for Nonprobability Samples," *Statistical Science*, Vol.32, Issue 2 (2017), pp.249~264.

② Reg Baker et al., pp.94~95.

③ Reg Baker et al., p.96.

(Bayesian Analysis)来估计总体参数。这些方法也可以应用在概率抽样调查中,用于处理覆盖误差或无应答误差导致的估计偏差。

那么,如何判断哪种方法更好呢?美国皮尤研究中心的资深研究方法专家安德鲁·默瑟(Andrew Mercer)等认为抽样调查的估计偏差取决于三个要素:一是互换性(exchangeability),含义是观测的样本与没有观测的样本是可以互换的,或者是有条件的互换,也就是说可以实现二者在研究变量上的表现无差异;二是正概性(positivity),意思是每一个观测的样本都是正概率入选,不存在总体中的某一个群体从观测的样本中完全缺失;三是组成性(Composition),就是观测的样本分布与目标总体相匹配,或者通过调整后匹配①。如配额抽样或事后倾斜权重仅在组成性上有所改进,对于其他两个方面没有任何助益。而样本匹配的方法可以在一定程度上保证互换性、正概性,辅以事后调整权重,也可以改善组成性,因此具有一定的优势。

无论哪种方法,最关键的是要获取和研究变量高度相关、测量误差小的协变量,模型的方法还要求在模型的设定上减少误差。然而,实现这些并非易事,因此非概率抽样的统计推断具有相当大程度的不确定性。

(三)抽样调查与大数据的互补与互助

大数据成为社会热点后,2015 年,美国民意研究协会又成立了一个特别工作组,来调研大数据的特性及对抽样调查的影响。② 工作组的专家们认为,大数据属于"发现"的数据,是先出现数据,而后研究者根据自己的研究需要去"收割"。而调查数据则属于"制造"的数据,是研究者根据研究需要先设计,然后按照设计来有控制地采集数据。由此,大数据的出现带来了研究范式的改变。传统研究范式是从理论到假设再到数据,最后通过统计检验来验证假设,提出新理论,或修正、扩展原理论。大数据则在一定程度上脱离了理论驱动的研究范式,转向数据驱动,利用数据量大、数据颗粒精细的优势来挖掘细节和变量间的相关性。

大数据在对专业人员技能上的要求也与抽样调查有所不同。抽样调查的专业训练注重抽样和测量的设计,以及在数据采集过程的质量控制,以最大限度地降低总调查误差为目标。收集到的数据采用结构化方式存储,数据清洗主要包括逻辑性检验、数据值合理性的查验,以及元数据的修订。后期数据处理包括对变量缺失值的插补、覆盖误差和无应答误差的调整等。总的来看,前期投入大,技能要求高,后期工作主要是对前期工作中出现问题的弥补。大数据的采集属于"直接收割",最需要的是计

① Andrew W. Mercer et al, "Theory and Practice in Nonprobability Surveys: Parallels between Causal Inference and Survey Inference," *Public Opinion Quarterly*, Vol.81, Issue S1 (2017), pp.250~271.

② Lilli Japec et al., "Big Data in Survey Research AAPOR Task Force Report," *Public Opinion Quarterly*, Vol.79, Issue 4 (2015), pp.839~880.

算机数据管理技能，在不同的时间点从不同的数据源聚合并形成数据集。收割上来的数据良莠不分，格式多样，没有统一的结构。这时需要专业人员对数据进行清洗，去粗取精，去伪存真，统一测度，并形成可供分析的数据库格式。因此，大数据采集的前期成本低，速度快，但后期数据清洗和加工的工作量巨大。而且由于大数据可以在网上轻松获取，没有经过专业训练的业余数据分析人员数量增长，可能会导致大数据处理和分析质量的下降，基于数据的结论不可靠。

大数据自身除了具有大量(volume)、快速(velocity)和多样(variety)的特征外，还具有易变(variability)、存真(veracity)和复杂(complexity)的特征。美国政治学家大卫·拉泽(David Lazer)教授等将“谷歌流感趋势”(Google Flu Trend，GFT)预测失误归结为两个原因：一个是大数据的“狂妄”，认为大数据可以替代传统数据收集和分析，忽视了基本的测量、建构效度和信度以及数据间的依赖性等问题；另一是搜索引擎算法的变动，提出搜索行为不仅是由外部因素决定的，也是由服务提供者培育的。当谷歌公司为了支持其业务而改变算法，向用户推荐其他内容的搜索，实际上就改变了数据的生成机制，导致错误的估计。拉泽等还提出用户也有可能改变数据生成机制，如政治竞选团队和商业公司意识到新闻媒体正在监控社交媒体，他们会使用一些策略以造成他们的候选人或产品正在流行的假象。还因为数据产权、个人隐私等问题，大数据很难支撑科学研究的复制(replication)检验①。相比之下，抽样调查的数据生成机制稳定、透明，数据可通过共享的方式供其他研究者复制，但调查数据的精细程度和时效性较弱，在时空动态分析，以及检测复杂的相互作用方面也有较大局限。

因此，抽样调查数据和大数据是两种各具优缺点的研究工具，二者可以在研究内容上互补，在研究方法上互助。

首先在研究内容上，抽样调查和大数据的发现可以互相激发。如抽样调查中常会有一些重要的现象或人群，由于数据量小而无法使用常规的统计手段分析，大数据则可以扩大对这些现象或人群数据的采集，使研究内容上更为全面。大数据也可以提供新的视角和方法，如研究中运用空间分析的方法，考察州内县级收入分配的聚合情况，可深入探讨抽样调查数据中所发现的收入不平等与健康的关系。② 另外，大数据有助于发现正在发生的事件，以及发展的趋势，却常常无法解释这个事件为什么会发生，或者为什么会偏离某种趋势，这时则需要借助抽样调查的精心设计来探究。随着大数据在社会科学研究上的应用增加，为更深入理解大数据发现的问题而进行抽

① David Lazer et al，“The Parable of Google Flu：Traps in Big Data Analysis，”*Science*，Vol.343，Issue 6176 (2014)，pp.1203~1205.

② Timothy L. Haithcoat et al，“Income Inequality and Health：Expanding Our Understanding of State-Level Effects by Using a Geospatial Big Data Approach，”*Social Science Computer Review*，doi：10.1177/0894439319872991.

样调查的需求可能也会随之增长。

在研究方法上，如前所述，基于普查或行政管理的大数据早已应用在抽样框的设计以及事后的权重调整上，有助于降低抽样调查由于覆盖误差或无应答误差导致的估计偏差。对于非概率抽样调查，这些数据可以作为重要的协变量，用以准随机化设计、伪权重的计算和模型估计。此外，一些个体可识别的大数据可以直接和调查数据链接，如将收入登记数据与关于选举的调查数据相关联，探讨个人财政状况对于选举决定的影响。① 这样做一方面可以丰富研究数据，另一方面也可以避免自报数据的测量误差，还可以减少调查数据采集的负担。对于个体不可识别但可以分类汇总的大数据，则可以通过统计值，与调查数据联合建构多层次模型，以满足特定的研究目的。

抽样调查也将大数据的技术用于提高调查质量和降低调查成本。如利用地理信息系统(GIS)建立抽样框，并采用卫星定位系统(GPS)进行住址抽样，②或基于计算机辅助调查系统记录的键盘痕迹数据计算单题访问时长，用于纠正访员不合规范的访问行为，③或将机器学习技术应用到职业应答的文本编码④等。在响应式调查设计的执行中，更是需要依靠计算机记录的各种并行数据及大数据的可视化手段来控制整个数据采集过程。

美国著名抽样调查专家米克·库珀(Mick Couper)甚至认为，大数据有可能解放抽样调查。他认为抽样调查的过量和商业化是导致抽样调查应答率下降，拒访率上升的重要原因。如果大数据可以带来抽样调查的减少，可能意味着完成的调查质量更高，也会提高抽样调查在受访者心目中的价值⑤。

三 抽样调查失灵了吗?:以2016年美国大选民意调查为例

从前文的介绍可以看出，抽样调查进入现时代被各种各样的实践问题所困扰，虽然有强大的概率抽样理论支撑，在现实面前却显得无力回天。那么，抽样调查作为研

① Andrew Healy, Mikael Persson, and Erik Snowberg, "Digging into the Pocketbook: Evidence on Economic Voting from Income Registry Data Matched to a Voter Survey," *American Political Science Review*, Vol. 111, Issue 4 (2017), pp.771~785.

② Pierre F. Landry and Mingming Shen, "Reaching Migrants in Survey Research: The Use of the Global Positioning System to Reduce Coverage Bias in China," *Political Analysis*, Vol.13, Issue 1 (2005), pp.1~22.

③ 严洁、邱泽奇、任莉颖、丁华、孙妍《社会调查质量研究:访员臆答与干预效果》，载《社会学研究》，2012年第2期，第168~181页。

④ 吴琼、戴利红、张婧申《机器学习在社会调查职业编码中的应用》，载《调研世界》，2019年第9期，第56~60页。

⑤ Mick P. Couper, p.156.

究工具还值得信赖吗?

美国大选一直是抽样调查的“试金石”。1936 年和 1948 年两次美国大选为抽样调查确立几十年来的“霸主”地位提供了机遇。2016 年美国大选中再次爆出冷门,民意调查中一直被看好的希拉里·克林顿败给了唐纳德·特朗普,人们在被选举结果震惊之余,也对民意调查的准确性提出了质疑。

美国民意研究协会一直对抽样调查的表现保持高度的关注,早在 2016 年春季就成立了一个委员会,任务是总结当年大选前民调的准确性,审查不同民调方法的差异,并从历史的角度进行评估。大选结束后,这个委员会对在大选前 13 天内进行的 22 个全国民意调查和 422 个州内民意调查,以及在其他调查数据或实验数据的辅助下进行了严谨充分的论证,发现有明确证据支持的解释是:(1)部分选民在临近选举日时改变了之前的选举决定,或从之前的不确定到转向特朗普;(2)在民意调查的样本中拥有大学学历的选民被过度代表,而低学历的选民代表性不足;(3)与 2012 年美国总统选举相比,投票的选民结构也发生了变化。部分证据显示,一些民调机构利用模型预测选民投票的可能性上存在失误。虽然当时最为普遍的说法是一些支持特朗普的选民没有在民意调查中如实报告,但委员会的多方取证没有支持这一说法。①

对于抽样调查质量的评估,总调查误差(Total Survey Error,TSE)框架是一个有效的工具。② 这个框架下,抽样调查的生命历程有两条主线,一条是测量,路径是构建—测量—应答—修订后的数据;另一条是代表性,路径是目标总体—抽样框—样本—受访者—事后权重调整。两条路径汇合,生成调查统计值。在这两条路径上,每一个阶段或环节都有产生误差的风险。如第一条路径就分别对应着建构效度(测量在多大程度上构建了要研究的概念)、测量误差(理想的测量和实际的测量之间的差异)和过程误差(对实际测量结果加工成研究数据时造成的偏差);第二条路径则对应着覆盖误差(目标总体与抽样框对应的总体之间的差异)、抽样误差(从抽样框中选取部分样本时的统计误差)、无应答误差(受访者完全应答的估值与实际不完全应答的估值之间的差异)和调整误差(对样本估值进行事后调整时造成的误差)。用这个框架来分析 2016 年美国大选前的民调预测失误,发现主要问题在于抽样调查过程中的测量误差、覆盖误差、无应答误差和调整误差。

(一)测量误差

委员会的报告(以下简称报告)中检验了四个可能的解释,其中两个解释属于测

① Courtney Kennedy et al, “An Evaluation of the 2016 Election Polls in the United States,” *Public Opinion Quarterly*, Vol.82, Issue 1 (2018), pp.1~33.

② Robert M. Groves and Lars E. Lyberg, “Total Survey Error: Past, Present, and Future,” *Public Opinion Quarterly*, Vol.74, Issue 5 (2010), pp.849~879.

量误差方面的原因。[①] 一个是选民投票前的临时决定。用于预测的民意调查要在选举前进行，一般认为，在调查方法同样严谨的情况下，民调的日期离选举日越近，预测的结果就越准确。这种看法的根据在于选民在接受调查后到真正投票时这段时期内可能会由于某些事件而改变他们的想法。也就是说，民意调查采集到的只是应答者最终投票决定的近似测量。报告引用了一个选举日当天的出口民调(exit polling)的研究结果，发现在竞选的最后一周，在选民中出现了明显的有利于特朗普的情形，特别是在特朗普以微弱优势胜出的那四个州。皮尤研究中心的回访民调也发现有11%的受访者承认他们在投票箱前做出了和选举前不一样的决定。这种临时改变决定的做法并非是2016年大选所独有，但之前一般改变想法的人会在民主党和共和党候选人之间平均分配，而这个回访调查却发现，在这些改变投票决定的受访者中，转而选择特朗普的比例比转向克林顿的比例多出16个百分点。对于竞争如此激烈的大选，这个测量误差可能就会决定预测的准确性。

另一个和测量误差相关的解释被称为“害羞的特朗普”(shy Trump)，指的是支持特朗普的受访者在民意调查中没有坦诚自己真实的投票决定，从而造成民意调查获取的是错误的信息。在美国，种族和性别通常是两个具有政治正确色彩的话题，而在2016年选举中希拉里·克林顿是美国历史上第一个女性总统候选人，特朗普则被控有种族和性别上的歧视，所以支持特朗普的受访者出于社会期许或政治正确的原因不愿意透露真实想法似乎是一个非常合理的解释。报告重点从访员效应角度来证实(伪)这个解释。以往的研究发现，受访者对于一些敏感问题的应答可能会因为对访员的不信任，或访员的某些特征(如性别和种族)而隐藏自己的真实想法。然而，专家们基于对调查模式的比较和一些调查实验的研究，没有发现支持的证据。他们也假设如果这个解释成立，同一州内特朗普与共和党参议员在民调预测与实际得票的差异上会表现不同，这一间接的假设也没有被证实。因此，由于社会期许或政治正确而导致的测量误差至少是不严重的。

(二)覆盖误差

美国大选民调的总体界定上有些复杂，可以分为符合资格的选民总体、登记注册的选民总体，以及实际投票的选民总体。这三个总体按此顺序排列，后者都是前者的一个子集，或是一个大数据样本。大选民调的理论目标总体是符合资格的选民总体，操作化的目标总体是登记注册的选民总体，而统计推断的目标总体则是实际投票的选民总体。

报告中分析的民意调查主要有六种设计：(1)网络自愿式调查，样本来自调查公

① Courtney Kennedy et al., pp.1~33.

司建设的网络调查样本库;(2)电话调查,样本选取基于固定电话和手机号码的随机数字拨号(RDD);(3)电话调查,样本框是各州的选民登记文件;(4)交互式语音系统(IVR)调查,样本框是各州的选民登记文件;(5)交互式语音系统和电话调查的混合模式;(6)交互式语音系统和网络调查的混合模式。

根据以往的研究发现,上网和不上网的人之间有系统的差异,网络自愿式调查很容易将老人、低学历或蓝领劳动者排除在外;使用固定电话和手机的人群也不一样,后者一般为年轻人,在种族和民族上较为分化。基于选民登记文件的抽样框要比电话号码建构的抽样框质量更好,但前者更适用于州内民调,对于全国民调则不易获取。按照美国联邦法规的规定,交互式语音系统功能只能用于固定电话,而美国大约一半的成年人没有固定电话。因此,仅采用交互式语音系统方式的州内民调,即使有选民登记文件作为抽样框,仍会存在严重的覆盖误差,这也是为什么一些民调采用交互式语音系统和电话或网络相结合的方式。

报告中一个意外的发现是,这些仅采用交互式语音系统方式的州内民调在预测的准确度上最高。推测其原因是由于这些漏掉的手机用户有大部分是非裔美国人或年轻选民,他们对实际投票的参与率较低,因此漏掉这些特征人群的样本结构反而与实际投票总体的结构更接近。那些为了弥补这一覆盖误差而补充了网络或电话调查的民调,反而表现较差。

(三)无应答误差

根据皮尤研究中心2012年的报告,当时电话访问的应答率已经低于10%。2016年美国大选前的民意调查主要采用访员主导或交互式语音系统形式的电话访问,虽然应答率不知,但已有许多研究证明在这些民调中低学历的选民代表性不足,而拥有大学及以上学历的选民被过度代表。

如果明确知道无应答样本的特征,一般通过权重调整就可以基本避免估计偏差。然而,报告中发现大多数的州内民调都没有对教育结构进行事后调整,而在全国性民调中约有一半做了调整。究其原因,是因为州内民调大多采用州内的选民登记文件作为抽样框,这些文件中包括了登记选民的年龄、性别、地域分布、党派注册和过往的投票历史等信息,却唯独漏掉了受教育程度。这些民调在访问时也未能补充受访者教育程度的信息,致使无法对样本的无应答误差进行纠偏。

令人疑惑的是,这些州内的民调以前也没有依据教育程度进行权重调整,却没有发现大的预测失误。报告发现,2016年的选民特征与投票选择之间的关系和2012年大选时有所不同。出口民调数据显示,在2012年无论是全美还是威斯康星、宾夕法尼亚和密歇根三个"摇摆州"(swing state),选民的受教育程度与对民主党派候选人的支持呈现U型关系,即受教育程度低和受教育程度高的选民都更为支持民主党

派候选人。但在 2016 年的大选，选民受教育程度与对民主党派候选人的支持几乎呈直线上升关系，即选民的受教育程度越高，越支持民主党候选人。这样，在 2012 年的民调中如果不对过度代表的高学历样本和代表性不足的低学历样本进行调整，不会造成麻烦，因为两个人群在支持方向上比较一致。但在 2016 年的民调中，如果不做调整，就会带来较大的偏差，这时低学历样本的代表性不足将造成过高地估计民主党候选人的支持率。

报告还查证了另外一种无应答误差的可能，即坚定支持特朗普的地区的选民是否在民调中代表性不足。逻辑是，如果人口普查数据显示有 13% 的美国人生活在坚定支持特朗普的地区，但民调估计只有 9% 的美国人生活在这些地区，这就证明民调确实系统性地遗漏了特朗普的支持者。受数据所限，报告中仅对电话调查进行了分析，没有发现有明显的证据支持这个假设。然而，孟晓犁用“数据缺陷指标”(data defect index) 来分析这次美国大选前的民调，确实发现在特朗普的支持者中无应答的概率更高。①

(四) 调整误差

如前所述，美国大选前的民调统计推断的目标总体是实际投票的选民总体，然而这一总体也是一个大数据样本，每次大选时不同特征的选民的投票意愿不同，将造成实际投票的选民总体与登记注册的选民总体有结构性的差异。所以，解释 2016 年美国大选民调失灵的一个可能的理由就是不同民调在预测当年选民投票的可能性，以及对调查估计所做的相应调整上出了错误。报告发现不同的民调在可能的选民(likely voter) 的预测模型设定上各有千秋，对于估计结果的影响也不尽相同。但有证据表明，在几个摇摆州，降低大学及以上学历样本的权重会提升预测的准确性，而调高非西班牙裔黑人的权重则会降低准确性。虽然大家都意识到需要对民调估计值进行调整以降低覆盖误差、无应答误差，以及选民投票行为的自选择误差，然而这些调整在多大程度上发挥了降低偏差的作用却很难判断。

尽管 2016 年美国大选前的民调存在各样误差的隐患，但委员会的报告证实，以历史标准来衡量，至少全国性的民调整体上看是准确的。2016 年的误差水平不到自 1936 年现代民调出现以来全国民调平均误差的一半，也低于 1992 年以来的平均误差。州内的民调则问题严重些，过高估计了对希拉里·克林顿的支持。但总的来看，美国民调不存在对某个党派候选人的系统性偏差。全国和州内民调的趋势线都显示，在任何一次选举中，民调在党派倾向上是随机的。

① Meng Xiaoli, p.712.

四 结语: 抽样调查的未来

《公共舆论季刊》在 2017 年出版特刊,主题为“抽样调查: 今天和明天”。在开篇文章的正文前引用了两位资深调查专家的观点。①

我想我对抽样调查不抱太大希望。你们(的调查)降到了 9%的应答率,或者类似水平? 我只是在我们的工具箱里找不到任何东西能胜过产生这种行为的巨大的社会力量。

——罗伯特·格罗夫斯②

如果我们认为就一个特定的主题询问一个有代表性的样本,对于其他方法收集的信息是有价值的补充,我们就必须投入金钱,并做出必要的努力来实现这些目标。

——艾利诺·辛格③

从格罗夫斯和辛格的观点,可以清楚地看到抽样调查的发展到了一个十字路口: 抽样调查是该坚守原先的方向,还是应该让步于非概率抽样调查或大数据?

一部分调查研究者还在不懈地努力,想尽各种办法维护抽样调查的质量; 同时,无论是社会环境的原因造成应答率下降,还是政府对抽样调查经费的缩减,都使得抽样调查不具有持续性。相比之下,非概率抽样调查或大数据虽然成本低、时效快、数据量大,但数据质量不尽如人意。尽管大多数抽样调查领域的资深学者都认为数据采集进入一个多工具的时代,但似乎每个工具都不够完美,这会使基于数据分析的量化研究陷入困境: 再完美的模型,如果建立在糟糕的数据上,也没有用处,甚至有害。

笔者认为,研究数据采集进入了一个新的生态环境,抽样调查、非概率抽样调查和大数据是这个生态环境中互动的三个主体。

首先,三种数据采集手段短期内会各据一方。与之前一样,在没有扎实的理论支持下,政府仍然会继续依靠传统的概率抽样调查手段,根据对有代表性的调查数据的分析,为政策制定提供信息支持。因此,主要承接政府项目的维斯塔特(Westat)等专业调查机构还在致力于概率抽样调查的研究与实施。④ 非概率抽样调查数据和大数据的主要用户是商业公司或媒体,主要是服务于自己的业务模型,或用于采集时效性

---

① Peter V. Miller, “Is There a Future for Surveys?” *Public Opinion Quarterly*, Vol.81, Issue S1 (2017), pp.205~212.

② Hermann Habermann, Courtney Kennedy, and Partha Lahiri, “A Conversation with Robert Groves,” *Statistical Science*, Vol.32, Issue 1 (2017), p.131.

③ Eleanor Singer, pp.473~474.

④ 根据作者与 Westat 资深抽样统计专家郝虹生 2019 年 11 月的谈话。

强、成本低的社会数据。在学术研究上,知名的纵贯调查项目仍是重要的数据资源,然而也有一些研究会采用相对认可的新的技术手段。如美国政治学期刊《美国政治学评论》(*American Political Science Review*)在2019年10月网上首发的一篇学术论文中,研究者使用了两种数据来源。两个数据源都是非概率样本,样本量也不大,但是对这项研究很适用。同时按照要求作者也把所用数据及分析程序公布在哈佛大学的研究数据存储(dataverse)网站上,供同行复制或检验研究结果。

其次,三种数据采集手段会互相校验,概率抽样调查仍是衡量非概率抽样调查或大数据质量的参照基准。因此,培植高质量的抽样调查仍然必不可少。调查研究者们在执行过程中采用响应式调查设计降低调查误差的同时,也要利用或开发不同的统计工具,加强对缺失数据的处理等方面的研究。同时,调查研究者们也需要建立一种数据质量的度量,供用户来区分不同类型的调查,或不同类型的估计值的质量,并且教育用户该怎样选择数据。如大数据工作组的专家们基于抽样调查的总调查误差提出了大数据总体误差(Big Data Total Error)框架①,孟晓犁提出的"数据缺陷指数"都是潜在的评估工具②。

最后,三种数据采集手段融合使用,将促进不同数据源的组合。美国国家统计委员会(Committee on National Statistics)的一个小组已经开展研究,以"促进联邦统计项目的范式转变,即使用来自政府和私营部门的不同数据源的组合,而不是单一的普查、调查或行政记录"。他们认为,抽样调查范式已经衰落,而新的统计需要基于抽样调查数据和非抽样调查数据的组合。③ 然而,多数据源的组合会遇到诸多障碍,但这是一个不可避免的趋势,并且是值得努力的方向。

塞翁失马,焉知非福。对于抽样调查来说,也许正如库珀所希望的,多种数据采集工具并存可以减少对抽样调查数量的需求,进而减少对受访者的过度搅扰,转变人们对于抽样调查的态度,从而将抽样调查做到"少而精",回归到1960年以前的"黄金时代"。

任莉颖:中国社会科学院社会学研究所副研究员

(本文责任编辑:李墨)

① Lilli Japec et al., pp.853~855.
② Meng Xiaoli, p.693.
③ Peter V. Miller, p.211.

tions of the multilateral trading system have been influenced and promoted by international institutional leadership. The United States has played an important leading role in the post-war world trading system, that is, the international institutional leadership. The multilateral trading system has the attributes of public goods. The international institutional leadership provided by the United States depends on both its international and domestic bases. According to the international and domestic bases, the international institutional leadership can be grouped into four types: structural leadership, conservative leadership, radical leadership, and leadership deficit. Different types of international institutional leadership have promoted different transformations of the multilateral trading system.

Focal Topic
Approaches and Developments in American Political Studies

Survey Research Methods in American Political Studies

Survey research is an important observational instrument for American political studies. It has quickly developed since it originated in the late 19th century, but it confronts serious challenges in the 21st century. To deal with the problems such as growing coverage errors, decreasing response rates and increasing costs, non-probability sampling survey's rise, and big data's competition, survey researchers are innovating, proposing responsive survey design, inventing statistical inferential techniques for non-probability sampling surveys, and exploring possible joint usage with big data. Using the total survey error frame, this article analyzes the reasons of the polling's predication failures in the 2016 presidential election from the aspects of measurement error, coverage error, non-response error, and adjustment error. It concludes that probability sampling surveys, non-probability sampling survey, and big data have their own domains of application.

Statistical Analysis Methods and American Political Science Research

Statistical analysis methods play an important role in American political science research. From the beginning of the 20th century, when the statistical analysis method was germinated in American political science research, to the situation where statistical analysis is commonly used in political science research, to the current era of big data influencing the research paradigm, its emergence and development have gone through different sta-