【统计理论与方法】

域的样本量分配方法研究

吕 萍[1]，郭淡泊[2]

（1. 北京大学 中国社会科学调查中心，北京 100871；2. 常熟理工学院 数学与统计学院，江苏 常熟 215500）

摘要：域估计是抽样调查的研究热点之一，其核心问题是样本量问题。在实践中，域包含计划域和非计划域，若为计划域，兼顾总体和域的估计精度的抽样设计方案更有效，其中关键是域样本量分配问题。域样本量的研究方法主要有 5 种，包含传统的比例、最优、内曼、等量分配方法和 Power 分配方法。通过使用中国家庭追踪调查数据对 5 种方法进行比较研究。研究表明，比例、最优、内曼分配方法有助于提高总体的估计精度，当总样本量较小时，无法兼顾域的估计精度；等量分配方法能提高域的估计精度，便于域之间的比较，但降低了总体的估计精度；Power 分配方法虽然略微降低了总体的估计精度，却显著提高了域的估计精度，是兼顾总体和域的估计精度的样本分配方法。

关键词：域估计；比例分配；内曼分配；等量分配；Power 分配

中图分类号：C811；O212.2 文献标志码：A 文章编号：1007－3116(2018)05－0003－05

一、引 言

域估计是抽样调查的研究热点问题，受到社会各界越来越多的关注，冯士雍将域估计问题列为抽样调查的三个重要前沿方向之一[1－2]。Rao 曾说过“再怎么强调域估计研究的重要性也不为过。”[1]因此，域估计问题受到越来越多的研究者的重视。

在实践中，域包含计划域和非计划域。计划域是在抽样设计之前确定的域，一般情况下，通过调整抽样设计兼顾总体和域的估计精度，多以层、阶段和子总体的形式存在。域总体已知，域样本量固定，可以避免域样本量极小或为零的情况，估计方法简单。非计划域是在抽样设计或调查之后确定的，域样本量是随机的，可能非常小甚至为零，估计方法复杂。在实践中，尤其是大型的抽样调查中，大部分域是计划域。

兼顾域和总体的估计精度，最常用的方法是将每个域视为子总体，分别设计抽样方案和计算域样本量，但会导致抽样设计复杂且总样本量异常大，违背抽样设计的初衷。因此，对于计划域，最理想的方法是找到兼顾总体和域的“最优”或“次优”的抽样设计方法，尤其大型抽样调查，调查经费和执行能力等有保障。这样的抽样设计方法主要有两类：第一类，通过改进抽样设计，例如细分层、减少整群、多重抽样框、连续性抽样调查方法、整合调查数据和混合模式等方法。第二类，通过层层抽样的方法、ABC 三级一套样本法和样本追加策略等方法。这些方法都是调整样本量，增加域内有效样本量，提高域的估计精度。因此，域样本量的分配是非常重要的。

本文主要对计划域的样本量分配进行深入研究，并使用实际调查案例进行实证研究，得出 Power 样本量方法是较好的兼顾域和总体的估计精度的方法。

二、样本量分配方法

在抽样设计中，样本量的计算需要考虑总体大小、总体和各层的变异程度、置信度、调查经费、抽样方法、应答率和执行能力等。若为分层抽样还需要

收稿日期：2017－09－15；修复日期：2017－11－07

基金项目：国家社会科学基金项目《追踪调查中小域估计的方法及其应用研究》(11CTJ005)；全国统计科学研究重点项目《小区域估计方法在政府统计中的应用》(2012LZ055)

作者简介：吕 萍，女，山东泰安人，副研究员，研究方向：统计调查技术和统计数据分析；

郭淡泊，女，福建厦门人，讲师，研究方向：经济统计分析。

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

考虑总样本量在各层的分配[3-4]。同样地，若为域估计，且是计划域，当总样本量固定时，同样需要考虑域的样本量分配，兼顾总体和域的估计精度。

假设目标变量是总体的均值 $\overline{Y}=\sum_d W_d\overline{Y}_d$，得到总体的均值估计量是 $\overline{y}_{st}=\sum_d W_d\overline{y}_d$，其中 $W_d=N_d/N$ 表示域 $d(=1,2,\cdots,L)$ 的域权(相对大小)，N_d 是域 d 的总量，N 是总体的总量，$\overline{Y}_d$ 表示域的总体均值，$\overline{y}_d$ 表示域的样本均值。

总体的方差估计量是：$V(\overline{y_{st}})=\sum_{d=1}^{L}W_d{}^2V(\overline{y_d})=\sum_{d=1}^{L}W_d{}^2\frac{1-f_d}{n_d}s_d^2=\sum_{d=1}^{L}\frac{W_d{}^2s_d^2}{n_d}-\sum_{d=1}^{L}\frac{W_ds_d^2}{N}$，其中 s_d^2 是域 $d(=1,2,\cdots,L)$ 的样本方差，f_d 是域 d 的抽样比例，n_d 是域 d 的样本量，$V(\overline{y_d})$ 是域的方差估计量。

假设总样本量是 n，与分层抽样设计相同，域样本量与各域的规模、域内方差有关，若考虑调查费用，则还与调查费用有关。因此，域估计的样本量分配，同样可以考虑分层抽样中的比例分配、最优分配和内曼分配并结合域的估计精度。

(一)比例分配

比例分配是域的样本量与域规模成比例，这种方法仅涉及域规模，域规模越大，域内样本量越大，即 $n_d^{pro}=\frac{n}{N}N_d$，当各域的费用和方差差异不大，域规模差异较大时多采用比例分配。该方法是等概率的，因此总量和均值估计量都是自加权的，数据分析相对简单，是最常用的样本量分配方法。

(二)最优分配

最优分配是在给定费用下使估计量的方差最小，或给定方差下使总费用最小。假设总费用是线性函数，即 $C=c_0+\sum_d c_dn_d$，其中 c_0 是固定费用(包含编制抽样框、组织协调费用和抽样费用等)，c_d 是域 d 中样本单元的平均费用(包含调查员的工资和差旅费等)，$S_d(y)$ 是域 d 的总体方差。得到 $n_d^{opt}=n\frac{N_dS_d(y)/\sqrt{c_d}}{\sum_d N_dS_d(y)/\sqrt{c_d}}$。因此，域规模越大、变异程度越大和费用越小，域样本量越大。当所有域的方差和费用都相同时，最优分配与比例分配相同。

(三)内曼分配

内曼分配是在最优分配的基础上，假定各个域的费用基本相同，即 $n_d^{Ney}=n(N_dS_d(y)/\sum_{d=1}^{D}N_dS_d(y))$。

内曼分配结合最优和比例分配的优点，当各域费用差不多，域规模和域内方差差异较大时，内曼分配较好。

由上可知，对于计划域，样本量分配取决于域规模、域内方差和费用。传统的分层抽样的方法也可以应用到域估计中。当域的费用和方差差异不大，比例分配较好。当域内方差差异较大但调查费用基本相同时，内曼分配方法较好。当域规模、域内方差和费用差异较大时，最优分配较好。但是，上述域样本量分配方法来源于分层抽样设计，主要是为了满足总体的估计精度，容易导致某些域的样本量很少甚至为零，无法满足域的估计精度。在总样本量固定的情况下，如何兼顾总体和域的估计精度呢？在实践中，若调查目的还关注各个域的比较研究，一般使用等量分配。

(四)等量分配

等量分配是将总样本量在各个域之间平均分配，即 $n_d=n/L$。样本量对域的估计精度影响较大，若各个域的费用和方差系数差异不大，等量分配便于域的比较，提高域的估计精度。但是，若域内方差和抽样比例差异较大，等量分配不利于域的比较，降低总体的估计精度。

因此，上述四种方法都不能很好地兼顾总体和域的估计精度，Bankier 提出一种 Power 分配，可以很好地兼顾总体和域的估计精度[5]。

(五)Power 分配

Power 分配通过构造损失函数兼顾各个域的估计精度，即 $F=\sum_d(X_d^qCV(Y_d))^2$。其中 $CV^2(Y_h)=V(Y_h)/Y_h^2$ 是域的方差系数，X_h 测量域的重要性(可以是目标变量 y、y 的替代值或与 y 相关的辅助变量)，q 是(0,1) 的常数，$0\leqslant q\leqslant 1$。

在 $\sum_d n_d=n$ 下，最小化损失函数 F，得到域的最优分配量是：$n_d=n\frac{S_dX_d^q/\overline{Y}_d}{\sum_d S_dX_d^q/\overline{Y}_d}$，其中 S_d^2 是层内方差，$S_d^2=\sum_i^{N_d}(y_{di}-\overline{Y}_d)^2/(N_d-1)$，$\overline{Y}_d=Y_d/N_d$。令 $CV_d=S_d/\overline{Y}_d$ 是域 d 的变异系数，则 Power 分配是 $n_d^{power}=n\frac{CV_dX_d^q}{\sum_d CV_dX_d^q}$。

由上可知，Power 分配通过控制各个域的估计精度的加权和，兼顾总体和域的估计精度。但是，该方法受 q 取值的影响。当 $q=1$ 时且 $X_d=N_d\overline{Y}_d$ 时，

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

是内曼分配。当 $q=0$，各域的变异系数差异不大时，是等量分配。当 $X_d=N_d$，$q=1$，$CV_d=CV$ 时，是比例分配。当 $X_d=N_d$，$q=1/2$，$CV_d=CV$ 时，$n_d=n\sqrt{N_d}/\sum_d\sqrt{N_d}$，称为“平方根分配”。Bankier 通过数值案例说明当 q 为 0.3 或 0.5 时效果较好。

上面介绍了 5 种样本量分配方法，其中比例分配、最优分配和内曼分配，与传统方法相同，有助于提高总体的估计精度，等量分配有助于提高域的估计精度，便于域之间的比较。而 Power 分配更有助于提高域和总体的估计精度，是兼顾域和总体估计精度的方法。

在实践中，若某些域，按照上述方法得到的域样本量大于域总量，即 $n_d>N_d$。此时，需要对这些域进行大于 100% 抽样的修正[3]。例如，第一个域的样本量 $n_1>N_1$，则将第一个域的样本量设为 N_1，然后将 $n-N_1$ 在剩下的域中进行分配。

下面，使用中国家庭追踪调查（CFPS）2010 年的调查数据，对域的各种样本量分配方法进行实证研究[6-7]。

三、实证研究

CFPS 是北京大学中国社会科学调查中心主持的追踪调查项目，抽样设计比较复杂。首先将 25 个省份分成两类：一类为省级层次的推断总体，用以满足省级推断的要求，包含辽宁省、上海市、河南省、广东省和甘肃省五个省市，也称为“大省”；其余 20 个省份为非省级推断总体。两类样本数据加权得到 25 个省份总体的有效估计，用来推断全国。分别在 5 个“大省”和 1 个“小省”中采用三阶段、不等概的整群抽样设计。因此，CFPS 是多阶段不等概率的复杂抽样设计。同时，CFPS 也是一个追踪调查，目前已经完成了 2010 年的基线调查，以及 2012 年、2014 年和 2016 年三轮追踪调查。

下面，使用 CFPS2010 成人数据库中的再抽样样本的体质量指数（BMI）数据对样本量分配方法进行研究。CFPS2010 数据库共有 21 812 个成人样本，其中 BMI 缺失 556 个，缺失比例是 2.5%，缺失比例较小，因此本文对缺失数据采用删除法，视为随机缺失。

（一）Power 分配比较研究

Power 分配兼顾域和总体的估计精度，但是 q 的取值影响最终的样本量分配结果和效率。此处，使用 CFPS2010 成人数据库中的体质量指数（BMI）数据对不同 q 值下的 Power 分配进行研究。首先，将数据分为老年层（年龄在 60 岁及以上）和中青年层。假设总样本量是 2 000，使用 Power 分配确定各域的样本量，为了便于操作，使用域规模作为辅助变量，即 $n_d^{\text{power}}=n\dfrac{CV_dN_d^q}{\sum_d CV_dN_d^q}$，$d=1,2$，当 q 取不同值（0，0.1，0.2，…，1）时，Power 样本量分配方法下的方差系数的比较表和图如下。

表 1　Power 样本量分配方法在不同 q 值下不同的分配结果的比较表

q	n_1	n_2	CV_1	CV_2	CV	DEFF
0	1 047	953	0.434	0.464	0.384	1.433
0.1	977	1 023	0.454	0.447	0.371	1.341
0.2	907	1 093	0.476	0.431	0.360	1.262
0.3	838	1 162	0.501	0.417	0.350	1.196
0.4	770	1 230	0.528	0.405	0.342	1.140
0.5	705	1 295	0.557	0.394	0.335	1.095
0.6	642	1 358	0.589	0.384	0.330	1.059
0.7	582	1 418	0.623	0.375	0.326	1.032
0.8	526	1 474	0.661	0.367	0.322	1.012
0.9	473	1 527	0.702	0.360	0.321	1.000
1	424	1 576	0.746	0.354	0.320	0.995

注：n_1、CV_1 是老年域的样本量和方差系数，n_2、CV_2 是中青年域的样本量和方差系数，CV 和 DEFF 是总体的方差系数和设计效应。

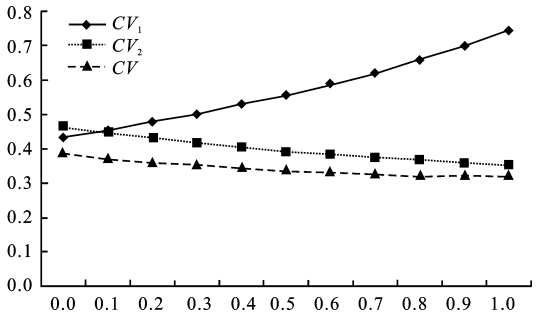

图 1　在不同 q 值下 Power 分配结果的比较图

由表 1 和图 1 可知，随着 q 的增加，老年域的方差系数不断增加，且增加速度不断加大，而中青年域的方差系数不断减少，且减少速度逐渐降低。总体的方差系数不断减少，但是减少速度在逐渐降低。

由于两个域的方差系数差异不大，所以当 q 趋向于 0 时，各个域的估计精度差异较小，Power 分配趋向于等量分配，各个域的精度较高，但是总体的估计精度较差。当 q 趋向于 1 时，Power 分配趋向比例分配，总体的估计精度最高，但各个域的估计精度差异较大。

由上可知，因此 Power 分配介于等量分配和比例分配之间，兼顾总体和域的估计精度。

（二）域样本量分配方法实证研究

下面，使用 3 种 Power 分配方法，即 Power 分

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

配 1($q=0$)、Power 分配 2($X_d=N_d, q=1/2$)、Power 分配 3($X_d=N_d, q=1/2, CV_d=CV$),以及比例分配、内曼分配和等量分配,共 6 种分配方法,对老年域和中青年域的样本分配方法进行实证研究,假设域内采用随机抽样方法,得到表 2 的结果。

表 2　不同样本量分配下老年和中青年域的精度比较

域	总样本量	分配方法	n_1	n_2	CV_1(%)	CV_2(%)	CV(%)	DEFF
老人和中青年域	20	比例分配	4	16	7.411	3.753	3.353	0.998
老人和中青年域	20	内曼分配	4	16	7.411	3.753	3.353	0.998
老人和中青年域	20	等量分配	10	10	4.683	4.748	3.976	1.404
老人和中青年域	20	Power1	10	10	4.683	4.748	3.976	1.404
老人和中青年域	20	Power2	6	14	6.049	4.013	3.462	1.064
老人和中青年域	20	Power3	6	14	6.049	4.013	3.462	1.064
老人和中青年域	200	比例分配	36	164	2.459	1.167	1.054	0.996
老人和中青年域	200	内曼分配	36	164	2.459	1.167	1.054	0.996
老人和中青年域	200	等量分配	100	100	1.462	1.497	1.253	1.407
老人和中青年域	200	Power1	99	101	1.469	1.490	1.248	1.395
老人和中青年域	200	Power2	63	137	1.852	1.278	1.098	1.081
老人和中青年域	200	Power3	64	136	1.837	1.283	1.101	1.086
老人和中青年域	2 000	比例分配	357	1 643	0.744	0.351	0.318	0.996
老人和中青年域	2 000	内曼分配	360	1 640	0.741	0.352	0.318	0.996
老人和中青年域	2 000	等量分配	1 000	1 000	0.397	0.460	0.383	1.449
老人和中青年域	2 000	Power1	994	1 006	0.399	0.459	0.382	1.440
老人和中青年域	2 000	Power2	630	1 370	0.536	0.388	0.332	1.089
老人和中青年域	2 000	Power3	636	1 364	0.533	0.389	0.333	1.093
老人和中青年域	4 000	比例分配	713	3 287	0.496	0.234	0.212	0.996
老人和中青年域	4 000	内曼分配	721	3 279	0.493	0.235	0.212	0.996
老人和中青年域	4 000	等量分配	2 000	2 000	0.219	0.315	0.260	1.506
老人和中青年域	4 000	Power1	1 987	2 013	0.221	0.314	0.260	1.495
老人和中青年域	4 000	Power2	1 260	2 740	0.336	0.262	0.223	1.101
老人和中青年域	4 000	Power3	1 271	2 729	0.333	0.262	0.223	1.105

注:n_1、CV_1 是老年层的样本量和方差系数,n_2、CV_2 是中青年层的样本量和方差系数,CV 和 DEFF 是总体的方差系数和设计效应。

由表 2,得到以下结论:

1. 无论哪种分配方法,随着总样本量的增加,域和总体的估计精度增加,说明样本量越大,总体和域的估计精度越高。

2. 在各种分配方法中,比例分配和内曼分配下总体的估计精度最高,但是域的估计精度较低。

3. 由于老年域和中青年域的 BMI 的方差系数差异不大,因此等量分配和 Power1 分配结果相近,尤其是样本量较小时,等量分配得到的域的估计精度近似相同,便于域之间的比较。但是,等量分配会降低总体的估计精度,因此总体的估计精度最低。

4. Power2 分配和 Power3 分配方法的估计精度相似,虽然使总体的估计精度有所下降,但增加了域的估计精度,是兼顾域和总体的估计精度的样本分配方法。在实践中,域和总体的方差系数往往是未知的,所以 Power3 分配,即平方根分配方法较好。

5. 从抽样设计的设计效应可知,比例分配和内曼分配更多地关注总体的估计精度,设计效应接近 1,其余分配方法兼顾总体和域的估计精度,提高了域的估计精度,设计效应大于 1,尤其是等量分配和 Power1 分配,设计效应最高。

四、结 论

在抽样调查中,域估计一直受到各界的重视。在实践中,最好的方法还是在抽样设计阶段尽量兼顾总体域的估计精度,尤其是计划域,本文在总样本量固定的情况下研究域的样本量分配方法。首先介绍了域的比例分配、最优分配和内曼分配,这三种方法有助于提高总体的估计精度,但是当样本量较小时,降低域的估计精度。然后,介绍了等量分配,该方法有助于提高域的估计精度,便于域的比较,但是降低总体的估计精度。最后,介绍了兼顾总体和域

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

的估计精度的 Power 分配。本文使用 CFPS2010 数据库对各种样本量分配方法进行实证研究。研究发现，无论是什么样的抽样设计，充足的样本量都有助于提高总体和域的估计精度；比例和内曼分配提高总体的估计精度，一般情况下无法兼顾域的估计精度；当各个域的方差系数差异不大时，等量分配提高域的估计精度，便于域之间的比较，但是降低总体的估计精度。Power 分配虽然略微降低了总体的估计精度，但提高了域的估计精度，是兼顾总体和域的估计精度的分配方法。其中，平方根 Power 分配表现最好，不仅兼顾总体和域的估计精度，而且只涉及域规模，操作性强。

Power 分配方法通过最优化各个域的方差系数的加权和控制总体的方差系数，并没有控制总体的方差系数。当某个域的样本量非常少，方差系数非常大时，Power 分配降低总体的估计精度。在实践中，如果要同时控制域和总体的估计精度，需要使用最优化方法。

参考文献：

[1] Rao J N K. Small Area Estimation[M]. New York：Wiley，2003.

[2] 冯士雍. 抽样调查应用与理论中的若干前沿问题[J]. 统计与信息论坛，2007(1).

[3] 金勇进，杜子芳，蒋妍. 抽样技术[M]. 北京：中国人民大学出版社，2012.

[4] Kish L. Survey Sampling[M]. New York：Wiley，1965.

[5] Bankier M D. Power Allocations：Determining Sample Sizes for Subnational Areas[J]. The American Statistician，1988，42(3).

[6] Yu X，Lv P. The Sampling Design of the China Family Panel Studies[J]. Chinese Journal of Sociology，2015，1(4).

[7] 吕萍. 抽样信息在复杂调查数据中的应用研究[J]. 统计研究，2017(1).

The Research on Sample Size Allocation of Domain

LV Ping[1]，GUO Dan-bo[2]

(1. Institute of Social Science Survey，Peking University，Beijing 100871，China；

2. College of Mathematics and Statistics，Changshu Institute of Technology，Changshu 215500，China)

Abstract：The domain estimation is one of the research hotspots in sampling survey，the key problem is the sample size. In practical survey，there are planned domains and unplanned domains. If the domains are planned domains，the sample design is more effective which taken into account the accuracy of the population and domain estimation，and the key problem is the allocation of domain sample size. There are five kinds of sample size allocation methods，including proportional allocation，optimal allocation，Neyman allocation，equal allocation and power allocation. Then，we use the CFPS2010 data to compare the five sample size allocation methods. The study further illustrate，the proportional allocation，optimal allocation，Neyman allocation are helpful to improve the estimation accuracy of population，but not estimation accuracy of domain. Equal allocation improves the estimation accuracy of the domain，is helpful in the comparison between domains when the variance coefficients of each domain have little difference，but it will reduce the estimation accuracy of population. Although the power sample allocation method reduces the estimation accuracy of population slightly，it improves the estimation accuracy of the domain significantly. It is a better sample allocation method considering both the estimation accuracy of population and domain.

Key words：domain estimation；proportional allocation；Neyman allocation；equal allocation；power allocation

(责任编辑：张治国)

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net