

中国综合社会调查（CGSS）
第二期（2010-2019）
抽样方案

目 录

一、调查背景.....	3
二、调查目标总体	3
三、抽样设计原则	3
四、抽样设计中的几个问题	4
(一) 关于分层	4
(二) 各阶段抽样单元	4
(三) 样本量的界定及分配	5
五、具体设计.....	6
(一) 必选层的样本抽取	6
(二) 抽选层的样本抽取	8
六、最终样本的构成	10
七、样本权数的确定	10
(一) 设计权数	10
(二) 实际执行情况调整	11
(三) 最终权数的确定	11
八、主要估计量	11
(一) 各小层内部目标量的估计	12
(二) 全国目标量的估计	13
九、估计量的方差估计	14
十、附录：抽选层 100 个样本初级单元	16

中国综合社会调查抽样设计方案

(2010)

一、调查背景

我国是世界上人口最多的国家，随着我国市场经济的发展，社会正在发生巨大的变革，经济发展使得对全国社会多方面信息的需求日益突出。及时、全面、客观地了解我国收入、医疗卫生、教育、失业保障等方面的信息，将为国家宏观调控政策的制定和企业的更好发展提供保障，对于促进城乡社会的协调发展、保持国民经济可持续发展都有着重大的现实意义。

中国综合社会调查是一项全国性的大型调查项目，调查总体是全国城镇居民和农村居民。本次调查的总体要求：（1）能够全面了解我国城乡社会发展情况；（2）能够对城市群体和农村群体进行对比分析；（3）能够在地理概念或者区域发展水平方面体现我国社会发展的地域差异性。

本方案着重介绍此次调查项目的抽样方案设计以及样本数据的权数调整方法。设计方案充分考虑了全国及不同地域估计的需要，对调查总体进行了科学、细致的分层，在一定程度上提高了估计精度；特别是在城乡样本配比、设立自我代表层、样本数据加权等问题的处理上体现了本方案的科学性、高效性、可操作性等特点。

二、调查目标总体

此次调查的目标总体为全国 31 个省、自治区、直辖市（不含港澳台）的所有城市、农村家庭户。

三、抽样设计原则

首先，作为全国性的抽样调查，整体方案必须是严格的概率抽样，要求样本对全国及某些指定的城市或地区具有代表性。其次，抽样方案必须保证具有较高

的效率，即在相同的样本量条件下，方案设计应使抽样误差尽可能小，调查精度尽可能高。最后，方案必须具有较强的可操作性，不仅便于具体抽样的实施，也要求便于后期的数据处理。

四、抽样设计中的几个问题

（一）关于分层

根据本次调查的研究需要，将调查总体分为两大类：一是必选层，该层总体为入选大城市的市辖区家庭户；二是抽选层，该层总体为除去必选层市辖区以外全国所有家庭户。

1、必选层

根据调查需要及以往调查经验，本次调查对那些发展处于国内领先水平的大城市将特殊对待，将该类城市市辖区家庭户作为单独一层进行设计，作为必选层。

对于大城市的界定，着眼于直辖市、省会城市和副省级城市共 36 座城市，从经济水平、教育水平及城市开放性程度等角度出发，选取 GDP、拥有教师总数、外国直接投资（FDI）实际使用外资金额这三个总量指标进行考察，采用因子分析方法确定排名前五的城市进入必选层。需要特别说明的是：该层的调查对象为这些城市的市辖区居民，该层最终调查单元均划为城市家庭户。

2、抽选层

抽选层的调查总体由必选层以外的城市、农村家庭户组成。为了便于在后期数据分析中采用平衡半样本进行方差估计，对初级抽样单元的分层划分打破省级地域限制，进一步增加分层的层数。将抽选层划分为区层和县层（包含县级市和县），采用人口密度、非农业人口比重和人均地区生产总值三个指标，在区层和县层中分别进行因子分析，得到区层和县层内各个区县的综合因子得分；在对综合因子得分进行排名的基础上将区层进一步分为 19 层，县层划进一步分为 31 层，抽选层共计细分划分为 50 个层。

（二）各阶段抽样单元

本次调查采用分层三阶段概率抽样，视所在层情况，各阶段抽样单元略有不同，详见表 1。

表 1 各阶段抽样单元

	第一阶段抽样单元	第二阶段抽样单元	第三阶段抽样单元
必选层	街道	居委会	家庭户
抽选层	区、县级市、县	居委会、村委会	家庭户

这样设计的原因在于：对于必选层，选择街道作为初级抽样单元可以细化抽样框，使得样本点相对分散，有利于总体信息的采集，避免由于抽样框过粗而导致样本有偏。对于抽选层，全国区、县级市、县的数量较多，以其作为初级抽样单元比较合适。

（三）样本量的界定及分配

1、目标样本量

目标样本量是指在一定精度要求下，实现调查目标所必需的样本量。根据以往调查经验，本次调查设定目标样本量为 12000 户，其中必选层 2000 户，抽选层 10000 户。后续部分涉及的样本量分配均立足于目标样本量。

2、各阶段样本量分配

由于必选层与抽选层设计的各阶段抽样单元略有不同，因此，分别介绍两者各阶段样本量的分配。

对于必选层，该层总样本量为 2000 户，计划抽取 40 个初级抽样单元(街道)，每个初级抽样单元（PSU）抽取 2 个二级抽样单元（居委会），每个二级抽样单元（SSU）中抽取 25 个家庭户。

对于抽选层，该层总样本量为 10000 户，计划抽取 100 个 PSU（区、县级市、县），每个 PSU 中抽取 4 个 SSU（居委会、村委会），每个居委会（村委会）中抽取 25 个家庭户。

最终，本方案共需抽取 140 个 PSU，480 个 SSU。

3、样本城乡分配

为了能切实反映全国范围城乡社会的真实情况，并在后续研究中能够对城乡进行对比研究，需要保证样本在城乡分配上与实际情况一致。首先明确本方案中的城乡样本概念，本方案设计基于这样的假设：居委会的家庭户为城市居民，村委会的家庭户为农村居民。

统计资料显示，目前我国城市常住人口数与农村常住人口数基本持平，由于城市居民主体的各方面差异相对明显，方差较大，因此将样本量的城乡分配比例确定为 6：4。

根据第一阶段样本量的分配结果，必选层中共抽取 80 个居委会，共计 2000 户，因此抽选层内城乡家庭户数需分别为 5200 和 4800 户才能满足 6：4 的要求。由于抽选层每个 PSU 下抽取 4 个 SSU（居委会或村委会），每个 SSU 内最终抽样单元的目标样本量均为 25，因此对城乡样本比例的控制，主要是使得抽选层居委会与村委会下的城乡样本比例达到 5200：4800，也即抽选层中的居委会与村委会样本个数比约为 208：192。

为了实现样本二级单元 208：192 的目的，需要根据样本初级单元的城市化水平（非农业人口比重）分配样本居委会和村委会的数量。这里采用分别在区层和县层样本初级单元内，根据各个区县的城市化水平（用非农人口比重表示）进行分段，形成若干个区间，然后根据不同的区间对样本区县中的居委会、村委会个数进行分配的方法来实现样本居委会和村委会的比例要求。

4、接触样本量

如果回答率达到 100%，则调查时需要的接触样本量即为有效样本量，但现实中无回答现象不可避免。要克服这个困难，可以采取替换样本的方法，但样本替换在实际操作中存在缺陷，因此本方案采取利用膨胀系数扩大样本量的方法，对第三阶段样本量进行放大。

根据往年调查经验，发达城市的市辖区居民由于种种原因，回答率在 50% 左右，即膨胀系数在 2 左右，因此在必选层每个二级单元抽取 50 户家庭，该层接触样本量扩大至 4000；抽选层的居民群体回答率高于必选层，但其内部还有差异，大体上城市居民的回答率在 65% 左右，农村居民的回答率高于城市居民，大致在 85%左右，因此，对于抽选层，在每个居委会抽取 38 户，在每个村委会内抽取 30 户。

五、具体设计

（一）必选层的样本抽取

1、必选层入样城市的确定

首先对全国 36 座城市(包括直辖市、省会城市、副省级城市)市辖区的 GDP、拥有教师总数、外国直接投资(FDI)实际使用外资金额这三项指标进行因子分析,最终确定 5 个城市进入必选层。这里给出进行因子分析之后,综合得分排名前五的城市名单(见表 2)。

表 2 36 座城市中前 5 座城市

排名	城市名称
1	上海
2	北京
3	广州
4	深圳
5	天津

2、初级单元的抽取

根据最新的全国行政规划,这 5 座城市市辖区总数为 67 个,如果以 67 个市辖区为初级阶段抽样框,初级单元个数有限,可能会导致样本在分布上过于集中,影响样本对该层总体的代表性,为了使样本点分布较为分散,将初级抽样单元细化为街道,由此,必选层的抽样框为入选城市市辖区的街道,且该层人口规模均只采用城市人口。

以必选层入选城市市辖区的街道作为抽样框,以各街道的城市人口规模为辅助信息,采取与人口规模成比例的 PPS 抽样抽取 40 个街道作为该层的 PSU,这里:最终抽样单元为家庭户,理论上应该以街道的户数作为辅助信息,但目前这方面信息缺乏,因而用街道人口数作近似处理。

具体步骤说明如下:设共有 N 个初级单元(街道),第 i 个初级单元的人口规模为 M_i ($i=1,2,\dots,N$),首先按地理区域将街道进行排序并产生累计人口数列,即排序第一位二级抽样单元为 M_1 ,对应的累积人口为 M_1 ;排序第二位的单元人口数为 M_2 ,对应的累积人口为 M_1+M_2 ,依次进行,则排序第 i 的单元对应的人口数为 M_i ,对应的累积人口为 $\sum_{j=1}^i M_j$ ($i=1,2,\dots,N$),需要注意的是:这里的 i 是指排序后单元对应序号。接下来,对第一个单元赋以 $1\sim M_1$ 共 M_1 个代码;对第二个单元赋以 $M_1+1\sim M_1+M_2$ 共 M_2 个代码,……,对第 i 个单元赋以

$$\sum_{j=1}^{i-1} M_j + 1 \sim \sum_{j=1}^i M_j \text{ 共 } M_i \text{ 个代码, 人口总数为 } M_0 = \sum_{i=1}^N M_i。$$

若已知初级单元样本量为 n ，则首先可以确定抽样间隔 k ，理论上 $k = \frac{M_0}{n}$ ，实际中一般取最接近于 $\frac{M_0}{n}$ 的整数。接下来从 $1 \sim k$ 范围内随机地产生一个整数 r 作为抽样起点，则代码为 $r, r+k, \dots, r+(n-1)k$ 所对应的单元即为被抽中的初级单元，且第 i 个抽样单元被抽中的概率为 $Z_i = \frac{M_i}{M_0}$ 。

3、二级单元的抽取

在确定初级抽样单元（街道）后，二级单元抽样框为样本街道内所有居委会名单，同时收集各个居委会人口数。二级单元的抽取方法：在入选街道内，采用与各居委会人口规模成比例的系统 PPS 抽样抽取 2 个居委会。

具体抽取过程同初级单元部分。

4、最终单元的抽取

最终单元的抽取在每个入选的居委会中进行，在每个居委会内，按家庭户的门牌号进行排序，采取等概率系统抽样抽出 50 户家庭作为最终调查单元。这里，实际接触样本量在调查时视回答率状况分批投放，以达到既满足目标样本量，同时各二级样本单元（居委会）下最终样本（家庭户）数量差异不至于悬殊。必选层样本抽取情况见表 3。

表 3 必选层样本抽取情况

区域	初级单元数	居委会数	目标样本量	实际接触样本量
必选层	40	$40 \times 2 = 80$	$40 \times 2 \times 25 = 2000$	$40 \times 2 \times 50 = 4000$

（二）抽选层的样本抽取

抽选层的初级抽样单元总体为：除去必选层城市市辖区以外的区、县级市、县，根据《中国统计年鉴 2009》的行政区划设置信息，除去必选层 67 个区后，抽选层共有 792 个区和 2003 个县级市和县。

1、抽样框的构建

为了实现样本分布与总体分布的一致，根据《全国分县市人口统计资料 2008

年》的户籍人口统计，区层和县层的人口规模比约为 7：18，差异较大，为了区县两层样本初级单元分配数量差异，采用与人口规模平方根成比例的分配方法，将抽选层中的 50 个层划分为 19 个区层和 31 个县层。

在区层和县层中，以人口密度、非农业人口比例、人均地区生产总值三个重要指标对区县进行因子分析，首先利用综合因子得分法对区层和县层的区县进行排序，然后根据区县的个数基本等分为 19 层和 31 层。抽选层中区层共有 792 个初级单元不能被 19 整除，将余数 13 归入区层的最后一层，结果为前 18 个区层每层含 41 个初级单元，第 19 层含 54 个初级单元。县层做法类似，2003 个初级单元不能被 31 整除，将余数 19 归入县层的最后一层中，结果为前 30 个县层每层含 64 个初级单元，第 31 层含 83 个初级单元。具体分配结果见附录 excel 表。

2、初级单元的抽取

在上述 50 个小层中，以各个初级单元的综合因子得分排序，各个初级单元的人口数为辅助信息，按照与各初级单元人口数成比例的系统 PPS 抽样方式分别在各小层中抽取 2 个样本区或县。

3、二级单元的抽取

在每个样本区或县中，抽取 4 个村委会或居委会，二级单元的具体抽取要根据初级单元的类别来确定，这是本方案实现城乡样本配比的关键环节，根据入样初级单元的类别确定应在该初级单元内抽取的居委会、村委会数目。一般说来，城市化水平较高的区（县），居委会个数较多、村委会个数较少，因此，城市化水平不同的区县，居委会与村委会抽取个数之间的配比亦不同，经测算，具体的居委会、村委会分配标准见表 4。

表 4 初级单元内二级单元的分配标准

分配标准	居委会	村委会
95%及以上	4	0
50%~95%	3	1
15%~50%	2	2
15%以下	1	3

在每个入样初级单元内部分别构建居委会、村委会抽样名单，同时收集该居、村委会人口数。根据上表的分配标准，以人口数作为规模辅助信息，按照与人口规模成比例的 PPS 抽取相应个数的居委会、村委会，确保每个初级单元下有 4

个二级单元。

4、最终户单元的抽取

最终户单元的抽取在每个入选的居委会或村委会中进行，要求每个二级单元内部达到 25 户的目标样本，最终抽选层目标样本量为 10000。

但在实际调查中要涉及接触样本量的膨胀问题，正如在样本量界定中提到的，最终接触样本量在每个居委会（村委会）内必须进行扩大，由于城乡居民的回答率不同（大约分别在 65%、85% 左右），因此将居委会内的接触样本量扩大至 38 户，每个村委会内的接触样本量扩大至 30 户。具体抽取方法仍采用等概率系统抽样。

5、户内调查对象的抽取

在入选的户内，列出所有 18 岁及以上人口，随机抽取一人，作为最终的调查对象。

六、最终样本的构成

本次调查目标样本量为 12000，必选层样本量为 2000，抽选层样本量为 10000。在必选层中，抽取 40 个初级单元，每个初级单元内抽取 2 个二级单元，在每个二级单元内目标样本量达到 25 户；在抽选层中，共抽取 100 个初级单元，400 个二级单元，其中居委会 208 个，村委会 192 个。全部目标样本城乡比为 6：4，基本与我国实际情况相吻合。

本次调查的最终接触样本量为 17664，其中必选层 4000，抽选层 13664。在必选层中，在初级单元、二级单元个数不变的情况下，将每个二级单元内部的接触样本量增加至 50 户；在抽选层中，初级单元数 100、每个初级单元内二级单元数为 4，在每个入样居委会内将接触样本量扩大至 38 户，在每个入样村委会内部将接触样本量扩大至 30 户，因此，抽选层接触样本量为 13664，其中城市居民样本为 7904，农村居民样本为 5760。

七、样本权数的确定

（一）设计权数

设计权数是指每个样本单元在调查总体中所代表的单元数，它等于该样本单

元入样概率的倒数。由前面确定的抽样方法，必选层与抽选层样本的设计权数形式相似，但表达的内容有所区别。统一的设计权数公式可表达为：

$$w_{1hijk} = \frac{M_h}{M_{hi}} \times \frac{M_{hi}}{M_{hij}} \times \frac{D_{hij}}{25} \quad (h=1,2,3,\dots,51)$$

其中，下标 1 代表设计权数， $h=1$ 代表必选层， $h=2,3,\dots,51$ 分别对应于抽选层中的 50 个小层。 M 代表人口规模总数， D 代表每个二级单元中包含的三级单元数（户数）。

其他各符号及下标的含义分别解释如下：在必选层中， i 代表第 i 个初级单元（街道）， j 代表第 j 个二级单元（居委会）， k 代表第 k 个三级单元（家庭户）；在抽选层中， i 代表第 i 个初级单元（区、县）， j 代表第 j 个二级单元（居委会、村委会）， k 代表第 k 个三级单元（家庭户）。

（二）实际执行情况调整

本调查为了保证目标样本量，对接触样本量进行扩大，调查的执行力度将影响调查样本的权数，因此需要再次进行权数调整。设某个居委会（村委会）最后样本调查的执行量为 n ，则该家庭户的实际执行调整权数为：

$$w_{2hijk} = \frac{25}{n}$$

（三）最终权数的确定

每个最终调查单元的最终权数等于每个家庭户的初始设计权数与执行情况调整权数的乘积，即：

$$w_{hijk} = w_{1hijk} \times w_{2hijk}$$

其中， h （ $h=1,2,3,\dots,51$ ）代表第 h 小层，1 代表设计权数，2 代表执行调整权数。

八、主要估计量

利用加权调整后得到的最终权数 w_{hijk} ，可以得到各种类型目标量的估计公式。在这里，我们记第 h 小层第 i 个样本区、县（第 1 小层为样本街道）第 j 个

样本居委会、村委会中第 k 个家庭户中被调查者的目标变量为 y_{hijk} 。

（一）各小层内部目标量的估计

1、总体总值 Y 的估计

$$\hat{Y}_h = \sum_{S_{rh}} w_{hijk} y_{hijk}$$

其中， S_{rh} 是第 h 小层内回答单元所组成的集合。

2、总体均值 $\bar{Y}$ 的估计

$$\hat{\bar{Y}}_h = \frac{\sum_{S_{rh}} w_{hijk} y_{hijk}}{\sum_{S_{rh}} w_{hijk}}$$

3、总体比例 P 的估计

$$\hat{P}_h = \frac{\sum_{S_{rh} \cap C_h} w_{hijk}}{\sum_{S_{rh}} w_{hijk}}$$

其中， C_h 是第 h 小层内具有某种特征的单元集合。

4、总体比率估计量的估计

$$\hat{R}_h = \frac{\sum_{S_{rh}} y_{hijk} w_{hijk}}{\sum_{S_{rh}} x_{hijk} w_{hijk}}$$

其中， x_{hijk} 、 y_{hijk} 为该比率估计量所涉及的两指标的样本回答值。

5、域估计——针对特定群体的估计

定义示性函数 $I_{hijk} = \begin{cases} 1 & \text{属于该群体} \\ 0 & \text{不属于该群体} \end{cases}$

该特定群体均值估计量为：

$$\hat{\bar{Y}}_{Th} = \frac{\sum_{S_{rh}} I_{hijk} y_{hijk} w_{hijk}}{\sum_{S_{rh}} I_{hijk} w_{hijk}}$$

（二）全国目标量的估计

1、总体总值 Y 的估计

$$\hat{Y}_z = \sum_h \sum_{S_{rh}} \frac{M_0}{M_h} w_{hijk} y_{hijk} \quad h = 1, 2, 3, \dots, 51$$

其中， M_0 表示全国总人数， M_h 表示第 h 层的人口总数。

2、总体均值 $\bar{Y}$ 的估计

$$\hat{Y}_z = \sum_h \sum_{S_{rh}} \frac{M_0}{M_h} \bar{Y}_h$$

3、总体比例 P 的估计

$$\hat{P} = \frac{\sum_h \sum_{S_{rh} \cap C_h} \frac{M_0}{M_h} w_{hijk}}{\sum_h \sum_{S_{rh}} w_{hijk}}$$

其中， C_h 是第 h 层内具有某种特征的单元集合。

4、总体比率估计量的估计

$$\hat{R} = \frac{\sum_h \sum_{S_{rh}} \frac{M_0}{M_h} y_{hijk} w_{hijk}}{\sum_h \sum_{S_{rh}} \frac{M_0}{M_h} x_{hijk} w_{hijk}}$$

其中， x_{hijk} 、 y_{hijk} 为该比率估计量涉及的两指标的样本回答值。

5、域估计——针对特定群体的估计

定义示性函数 $I_{hijk} = \begin{cases} 1 & \text{属于该群体} \\ 0 & \text{不属于该群体} \end{cases}$

该特定群体均值估计量为：

$$\hat{Y}_T = \frac{\sum_h \sum_{S_{rh}} \frac{M_0}{M_h} I_{hijk} y_{hijk} w_{hijk}}{\sum_h \sum_{S_{rh}} \frac{M_0}{M_h} I_{hijk} w_{hijk}}$$

九、估计量的方差估计

由于本方案采取的是分层三阶段复杂抽样，采用常规的方差计算方法，过程十分繁杂，且需要用到大量的辅助信息。结合本次调查将初级单元划分为 50 个层，每层抽取 2 个初级单元的设计特点，我们将采用复杂样本方差估计方法中的平衡半样本方法（BRR）对目标量的方差进行估计，从而达到简化计算复杂样本方差估计的目的。

假设对总体 $N = \sum_{h=1}^L N_h$ ，采用分层随机抽样，每层有放回地简单随机抽取 2 个单元，设 y_{h1} 和 y_{h2} 是第 h 层的样本观测值（ $h=1, \dots, L$ ）。现从每层中抽取一个单元，组成一个半样本，总共有 2^L 个这样的半样本。

$\bar{Y}$ 的基于第 α （ $\alpha=1, 2, \dots, 2^L$ ）个半样本的估计为：

$$\bar{y}_{st,\alpha} = \sum_{h=1}^L W_h (\delta_{h1\alpha} y_{h1} + \delta_{h2\alpha} y_{h2})$$

其中， $\delta_{h1\alpha} = \begin{cases} 1, & \text{如果第} h \text{层中的第一个单元被选入第} \alpha \text{个半样本} \\ 0, & \text{其它} \end{cases}$ ， $\delta_{h2\alpha} = 1 - \delta_{h1\alpha}$ 。

所有 2^L 个这样的估计量的平均恰好是 $\bar{y}_{st}$ ，即

$$\frac{1}{2^L} \sum_{\alpha=1}^{2^L} \bar{y}_{st,\alpha} = \bar{y}_{st}$$

假设这个较小的半样本子集包含了 k 个半样本，有：

$$v_k(\bar{y}_{st}) \triangleq \frac{1}{k} \sum_{\alpha=1}^k (\bar{y}_{st,\alpha} - \bar{y}_{st})^2 = v(\bar{y}_{st}) + \frac{1}{2k} \sum_{h=1}^L \sum_{h'>h}^L \left[\sum_{\alpha=1}^k \delta_h^{(\alpha)} \delta_{h'}^{(\alpha)} \right] W_h W_{h'} d_h d_{h'}$$

因此，如果所选择的 k 个半样本对于所有的 $h < h' = 1, \dots, L$ 都满足条件

$$\sum_{\alpha=1}^k \delta_h^{(\alpha)} \delta_{h'}^{(\alpha)} = 0$$

则 $v_k(\bar{y}_{st})$ 就正好等于 $v(\bar{y}_{st})$ 。从而，基于这 k 个半样本便可实现对 $V(\bar{y}_{st})$ 的估计，而不必基于全部的 2^L 个半样本。满足上式的 k 组半样本称为平衡半样本。

Fay 对平衡半样本进行了修正，即 Fay's BRR，引入 Fay 因子 $0 \leq \varepsilon < 1$ ，即估计量变为：

$$\bar{y}_{st,\alpha} = \sum_{h=1}^L W_h (\delta_{h1\alpha} y_{h1} + \delta_{h2\alpha} y_{h2})$$

$$\delta_{h1\alpha} = \begin{cases} 1+\varepsilon, & \text{如果第}h\text{层中的第一个单元被选入第}\alpha\text{个半样本} \\ \varepsilon, & \text{其它} \end{cases}$$

方差估计量变为:

$$v_k(\bar{y}_{st}) \triangleq \frac{1}{k(1-\varepsilon)^2} \sum_{\alpha=1}^k (\bar{y}_{st,\alpha} - \bar{y}_{st})^2$$

当层数 L 很大时, 即使利用正交矩阵得到完全正交的平衡半样本个数已经大大减少, 所需要的计算量仍然十分庞大, 此时可用部分平衡半样本的方法, 即 k 部分平衡半样本方法。

方差估计量是:

$$v_k(\bar{y}_{st}) \triangleq \frac{1}{k(1-\varepsilon)^2} \sum_{\alpha=1}^k (\bar{y}_{st,\alpha} - \bar{y}_{st})^2$$

在本次调查中, 抽选层在 50 个区县层中共抽取 100 个样本区县, 在每个样本区县中平均抽取 4 个居委会或村委会。根据抽样设计, 有两种平衡半样本方法:

1. 初级样本单元(区县)层次上的平衡半样本方法

按照抽样设计, 将抽选层的区县分成 50 层, 在每一层中抽取一定数目的样本区县。若是在此基础上应用平衡半方法, 则需要将每一层的区县随机分成 2 组。此处我们选取 60 或 64 阶的正交半样本随机组, 利用分组平衡半样本方法, 得到目标变量的估计量和方差估计量分别是:

目标变量的估计量是

$$\bar{y}_{st,\alpha} = \sum_{h=1}^L W_h (\delta_{h1\alpha} y_{h1} + \delta_{h2\alpha} y_{h2})$$

其中

$$\delta_{h1\alpha} = \begin{cases} 1+\varepsilon(\frac{n_h-m_h}{m_h})^{0.5}, & \text{如果第}h\text{层中的第一个单元被选入第}\alpha\text{个半样本} \\ 1-\varepsilon(\frac{n_h-m_h}{m_h})^{0.5}, & \text{其它} \end{cases}$$

方差估计量是:

$$v_k(\bar{y}_{st}) \triangleq \frac{1}{k(1-\varepsilon)^2} \sum_{\alpha=1}^k (\bar{y}_{st,\alpha} - \bar{y}_{st})^2$$

此处选取 Fay 因子 $\varepsilon = 0.5$ 或 0 。

相关系数和比例估计量和方差估计量计算方法同上。

2. 二级样本单元（居委会或村委会）层次的平衡半样本方法

若以 100 个区县为层，在每个层中将 4 个二级样本单元（居委会或村委会）随机分成 2 组，利用分组平衡半样本方法得到目标变量的估计量及其方差估计量。我们利用 104×104 阶正交 Hadamard 矩阵得到完全正交的平衡半样本。由此，得到目标变量的估计量及其方差估计量分别是

估计量是

$$\bar{y}_{st,\alpha} = \sum_{h=1}^L w_h (\delta_{h1\alpha} y_{h1} + \delta_{h2\alpha} y_{h2})$$

$$\delta_{h1\alpha} = \begin{cases} 1 + \varepsilon \left(\frac{n_h - m_h}{m_h} \right)^{0.5}, & \text{如果第 } h \text{ 层中的第一个单元被选入第 } \alpha \text{ 个半样本} \\ 1 - \varepsilon \left(\frac{n_h - m_h}{m_h} \right)^{0.5}, & \text{其它} \end{cases}$$

方差估计量是：

$$v_k(\bar{y}_{st}) \hat{=} \frac{1}{k(1-\varepsilon)^2} \sum_{\alpha=1}^k (\bar{y}_{st,\alpha} - \bar{y}_{st})^2$$

上述计算以目标变量的均值估计量为例，相关系数和比例估计量和方差估计量计算方法同上。

十、附录：抽选层 100 个样本初级单元

省份	市	区县	小层	大层	居委会	村委会
浙江省	杭州市	上城区	1	区层	4	0
辽宁省	沈阳市	和平区	1	区层	4	0
江苏省	常州市	钟楼区	2	区层	4	0
浙江省	温州市	鹿城区	2	区层	3	1
山东省	济宁市	市中区	3	区层	3	1
浙江省	杭州市	江干区	3	区层	3	1
江苏省	南京市	六合区	4	区层	3	1
浙江省	杭州市	余杭区	4	区层	2	2
吉林省	长春市	朝阳区	5	区层	3	1
山西省	太原市	迎泽区	5	区层	4	0
湖北省	武汉市	江夏区	6	区层	4	0
贵州省	贵阳市	云岩区	6	区层	3	1
河北省	唐山市	丰南区	7	区层	2	2
黑龙江省	哈尔滨市	道外区	7	区层	3	1

江苏省	连云港市	新浦区	8	区层	3	1
青海省	西宁市	城北区	8	区层	3	1
四川省	成都市	温江区	9	区层	3	1
浙江省	湖州市	吴兴区	9	区层	3	1
吉林省	吉林市	昌邑区	10	区层	3	1
山东省	泰安市	泰山区	10	区层	3	1
山东省	日照市	岚山区	11	区层	2	2
西藏自治区	拉萨市	城关区	11	区层	3	1
辽宁省	铁岭市	银州区	12	区层	3	1
辽宁省	葫芦岛市	连山区	12	区层	2	2
广西省	柳州市	柳北区	13	区层	3	1
辽宁省	朝阳市	双塔区	13	区层	3	1
广东省	汕头市	潮南区	14	区层	4	0
贵州省	遵义市	红花岗区	14	区层	3	1
江西省	上饶市	信州区	15	区层	3	1
海南省	海口市	秀英区	15	区层	2	2
四川省	乐山市	五通桥区	16	区层	2	2
重庆市	重庆市	涪陵区	16	区层	2	2
四川省	泸州市	龙马潭区	17	区层	2	2
广西省	玉林市	玉州区	17	区层	2	2
河南省	信阳市	浉河区	18	区层	2	2
河南省	商丘市	睢阳区	18	区层	2	2
湖北省	孝感市	孝南区	19	区层	2	2
贵州省	安顺市	西秀区	19	区层	2	2
山东省	济宁市	汶上县	1	县层	3	1
福建省	漳州市	东山县	1	县层	3	1
江苏省	扬州市	江都市	2	县层	2	2
湖北省	襄樊市	老河口市	2	县层	3	1
山东省	烟台市	莱州市	3	县层	2	2
福建省	三明市	永安市	3	县层	3	1
山西省	运城市	河津市	4	县层	2	2
江西省	宜春市	丰城市	4	县层	2	2
河南省	漯河市	临颍县	5	县层	2	2
湖北省	襄樊市	宜城市	5	县层	2	2
吉林省	吉林市	桦甸市	6	县层	2	2
黑龙江省	黑河市	逊克县	6	县层	2	2
河北省	沧州市	肃宁县	7	县层	2	2
河南省	周口市	郸城县	7	县层	1	3
江苏省	徐州市	新沂市	8	县层	2	2
黑龙江省	七台河市	勃利县	8	县层	2	2
吉林省	白城市	洮南市	9	县层	2	2
安徽省	阜阳市	太和县	9	县层	1	3

山西省	长治市	襄垣县	10	县层	2	2
湖南省	邵阳市	邵东县	10	县层	1	3
河北省	邯郸市	磁县	11	县层	1	3
浙江省	宁波市	奉化市	11	县层	2	2
四川省	南充市	营山县	12	县层	2	2
江西省	鹰潭市	余江县	12	县层	2	2
河南省	周口市	扶沟县	13	县层	1	3
陕西省	延安市	志丹县	13	县层	2	2
四川省	眉山市	丹棱县	14	县层	2	2
湖南省	怀化市	洪江市	14	县层	2	2
黑龙江省	哈尔滨市	双城市	15	县层	2	2
黑龙江省	哈尔滨市	延寿县	15	县层	2	2
安徽省	蚌埠市	固镇县	16	县层	1	3
河南省	洛阳市	宜阳县	16	县层	1	3
山东省	临沂市	沂水县	17	县层	1	3
黑龙江省	齐齐哈尔市	克山县	17	县层	2	2
宁夏自治区	吴忠市	同心县	18	县层	2	2
安徽省	宣城市	绩溪县	18	县层	2	2
湖南省	邵阳市	隆回县	19	县层	1	3
陕西省	汉中市	城固县	19	县层	2	2
江西省	赣州市	兴国县	20	县层	2	2
湖北省	襄樊市	南漳县	20	县层	2	2
四川省	宜宾市	宜宾县	21	县层	1	3
福建省	福州市	永泰县	21	县层	2	2
湖北省	恩施土家族苗族自治州	来凤县	22	县层	1	3
重庆市	重庆市	奉节县	22	县层	1	3
安徽省	黄山市	歙县	23	县层	1	3
江西省	赣州市	会昌县	23	县层	2	2
内蒙古自治区	呼和浩特市	土默特左旗	24	县层	1	3
吉林省	松原市	扶余县	24	县层	1	3
广西省	崇左市	大新县	25	县层	1	3
陕西省	延安市	宜川县	25	县层	2	2
新疆自治区	喀什地区	英吉沙县	26	县层	1	3
陕西省	铜川市	宜君县	26	县层	2	2
云南省	玉溪市	新平彝族傣族自治县	27	县层	1	3
湖南省	郴州市	汝城县	27	县层	1	3
甘肃省	兰州市	榆中县	28	县层	1	3
重庆市	重庆市	城口县	28	县层	1	3
云南省	大理白族自治州	剑川县	29	县层	1	3
广西省	柳州市	三江侗族自治县	29	县层	1	3
云南省	昆明市	寻甸回族彝族自治县	30	县层	1	3
湖南省	怀化市	中方县	30	县层	1	3

云南省	昭通市	巧家县	31	基层	1	3
甘肃省	庆阳市	镇原县	31	基层	1	3
合计					208	192