China Family Panel Studies

CFPS 中国家庭动态跟踪调查

技术报告系列：CFPS-1

系列编辑：谢宇　责任编辑：胡婧炜

中国家庭动态跟踪调查
抽样设计

谢宇　邱泽奇　吕萍

2012.12.20

1. 调查对象和目标样本量

CFPS 调查的对象为中国（不含香港、澳门、台湾以及新疆维吾尔自治区、西藏自治区、青海省、内蒙古自治区、宁夏回族自治区、海南省）25 个省/市/自治区的满足项目访问条件的家户和样本家户中满足项目访问条件的家庭成员。在 2010 年的基线调查中，居住在传统居民住宅内的、家中至少有一人拥有中国国籍的一个独立经济单元，便可视为一个满足项目访问条件的家户。[1] 满足项目访问条件的家庭成员则指样本家户中经济上联系在一起的：A.与该家庭有血缘/婚姻/领养关系的直系亲属；B.与该家庭有血缘/婚姻/领养关系且居住满 3 个月的非直系亲属；C.与该家庭没有血缘/婚姻/领养关系但居住满 6 个月的其他成员。

CFPS 样本所在的 25 个省/市/自治区的人口覆盖了中国除香港、澳门、台湾外总人口数的 94.5%，由于覆盖范围如此广泛，因此可以将 CFPS 样本视为一个全国代表性样本。

CFPS 重点关注中国社会经济、教育、家庭、人口和健康等方面的变迁，为了更好的了解这一点，不仅需要从宏观层面上把握中国的整体变迁，还需要在微观层面上了解中国的几个典型省/市的在这些方面的变化状况。因此，在抽样设计上，我们首先将全国 25 个省/市/自治区分成两类：一类省市为在省级层次的推断样本，用以满足省级推断的要求。我们选择了 5 个省/市，分别为辽宁、上海、河南、广东、甘肃，也称为大样本省（以下简称为“大省”）。二类省市为上述 5 个省/市外的 20 个省/市/自治区，这一类省市的样本量不能够在省级层次进行推断，也称小样本省（以下简称为“小省”）。从这两类样本数据的加权可以得到对 25 个省/市/自治区总体的有效估计，进而用来推断全国。25 个省/市/自治区的分类见表 1。

表 1. 全国 25 个省市的分类

类型	省市自治区
一类省市（自我代表省市，“大省”）	上海市 辽宁省 河南省 甘肃省 广东省
二类省市（非自我代表省市，“小省”）	江苏省、浙江省、福建省、江西省、安徽省、山东省、河北省、山西省、吉林省、黑龙江省、广西壮族自治区、湖北省、湖南省、四川省、贵州省、云南省、天津市、北京市、重庆市、陕西省

[1] 最初我们还要求受访家庭户中至少有一名成员在抽样社区居住时间满 6 个月，但在执行过程中，这一条件被取消，实际被这一条件过滤掉的仅有极少数家户。

由上，本次基线调查共有 6 个子总体，即 5 个“大省”和 其它“小省”。根据中国和世界上一些大型的抽样调查的经验及 2008、2009 年对北京市、上海市、广东省三个省市预调查的经验，在考虑调查经费和估计量精度的基础上，确定两类省市的样本量：5 个“大省”的目标样本量分别是 1600 户，“小省”的目标样本量总共是 8000 户，共 16000 户。

2. 抽样设计总原则

CFPS 样本是一个采用内隐分层（implicit stratification）方法抽取的多阶段概率样本（multi-stage probability sample）。采用多阶段抽样设计既是为了减少调查的运作成本，也是考虑到中国社会的社会背景差异。CFPS 中的每个子样本都通过三个阶段抽取得到。

抽样过程中的前两个阶段使用官方的行政区划资料。中国的行政区划结构有两个重要特征：首先它是严格分等级的；其次，它覆盖了中国所有人口。由于上海不同于其它“大省”，所以，上海样本的抽取被作为特例处理。

因为中国的经济发展一直存在地理上的差异，所以抽样设计中需要着重强调的就是地理代表性（geographic representation）。通过内隐分层，可以确保样本很好地代表了这 25 个省份。而且，在每个省份中，省会城市作为隐含分层被挑选出来。当城乡差异存在且有意义时，城－乡区别总是被用来进行多阶段的区域层次上的隐含分层。一般地，区、街道办事处、或居委会指的是城市地区；相应地，县、乡镇或者村则指的是农村地区。除城－乡区别外，一个用于测量社会经济地位（SES）的连续变量也被用于进行内隐分层。根据数据的可获得性，所选用的排序变量依次为地方人均 GDP、非农人口比例或人口密度。

第三个阶段在入选的样本村/居中，利用村级调查地图得到的住户列表清单制作末端抽样框，按照随机起点的循环等距抽样方式，以扩大样本量的方法抽取家户样本，以确保每个样本村居的能够完成目标的 25 户家庭。

3. 各阶段抽样

3.1 第一阶段抽样

3.1.1 “大省”的抽样

对辽宁、河南、甘肃和广东这 4 个“大省”，每个省内的所有区（若为城市）或县（若为农村）构成了一个抽样框（sampling frame）。为了形成内隐分层，区/县按照辅助信息进行

以下排序：

⑴ 全省所有市级行政区以省会城市为首、其它所有地级市按照社会经济地位（SES）降序排列。

⑵ 每个市级行政区被区分成三个部分：区、县级市和县。在上述各部分的内部，区/县级市/县各自按照社会经济地位（SES）降序排列。这些区/县级市/县构成了本调查的初级抽样单位（PSU）。

对于上述 4 个抽样框，采用如附录 A 所描述的系统 PPS 抽样方式（systematic probability proportional to size sampling）从每个抽样框中抽取 16 个 PSU。附录 B 给出了用于构建这些抽样框的资料来源。

上海市只有 19 个区县。第一阶段的抽样在比区/县低一级的行政区划上实施，即街道办事处（若为城市）和乡镇（若为农村）层次上。为形成内隐分层，所有的街道办事处、乡镇按照辅助信息进行以下排序：

⑴ 全市 19 个区县按照社会经济地位（SES）降序排列。

⑵ 每个区/县分成街道、镇和乡区三个部分。

⑶ 每个区/县的每一部分内，街道、镇和乡——上海的初级抽样单位（PSU）——按照社会经济地位（SES）降序排列

在这个抽样框中，采用与规模成比例（proportional-to-size）的系统抽样方式（见附录 A）抽取 32 个 PSU（参见附录 A）。附录 C 给出了用于构建该抽样框的资料来源。

3.1.2 “小省”的抽样

对于作为 CFPS 样本剩余组成部分的“小省”的样本，这些省份中的区（若为城市）或者县（若为农村）构成了一个抽样框。为了形成内隐分层，区/县按照辅助信息进行以下排序：

⑴ 20 个省级行政区按照社会经济地位（SES）降序排列。

⑵ 每个省内，将全省所有市级行政区以省会城市居首、其它地级市按照社会经济地位（SES）降序排列。

⑶ 将每个地级市或同级行政区划区分成三个部分：区、县级市和县。在每一部分内，

所有的区/县级市/县——20 个小省的初级抽样单位（PSU）——按照社会经济地位（SES）降序排列。

在这个大抽样框中，采用附录 A 给出的与规模成比例（PPS）的系统抽样方式抽出 80 个 PSU。附录 D 给出了用于构建该抽样框的资料来源。小省的样本区县的分布如表 2 所示。

表 2. “小省”的样本区县在各省市的分配

省市	样本区/县数	省市	样本区/县数
北京市	1	重庆市	2
福建省	2	江西省	3
黑龙江省	5	陕西省	3
山西省	7	广西省	3
安徽省	3	湖北省	3
浙江省	3	云南省	4
天津市	1	贵州省	5
江苏省	3	湖南省	6
吉林省	3	山东省	7
河北省	8	四川省	8

3.2 第二阶段的抽样

3.2.1 村/居抽样框

村居抽样框由第一阶段 144 个样本区/县和 32 个样本街道/乡镇的所有村/居的组成。为了建立有效、准确的村级抽样框，项目组收集了 144 个区/县和上海市 32 个样本街道/乡镇的村居资料，包含省/市名称、地级市名称、区/县名称、街道/乡镇名称，村/居名称，村/居行政区划代码，村/居最新常驻人口数，村/居最新常住人口数，村/居外出打工人数，村/居人口密度，备注。

在对各样本区/县、街道/乡镇的村居资料进行整理后（上海市的 32 个街道/乡镇分布在 18 个样本区/县中，即上海统计了 18 个区/县的村居资料），共得到 43805 个村居，平均每个区县的村居个数是 272 个，各样本区县的村居数分类统计表和统计图如下：

表 3. 样本区/县的村居数据统计表

村居数	区县数
100 以下	31
100~200	41
200~300	38
300~400	22
400~500	11
500~600	8
600 以上	11
合计	**162**

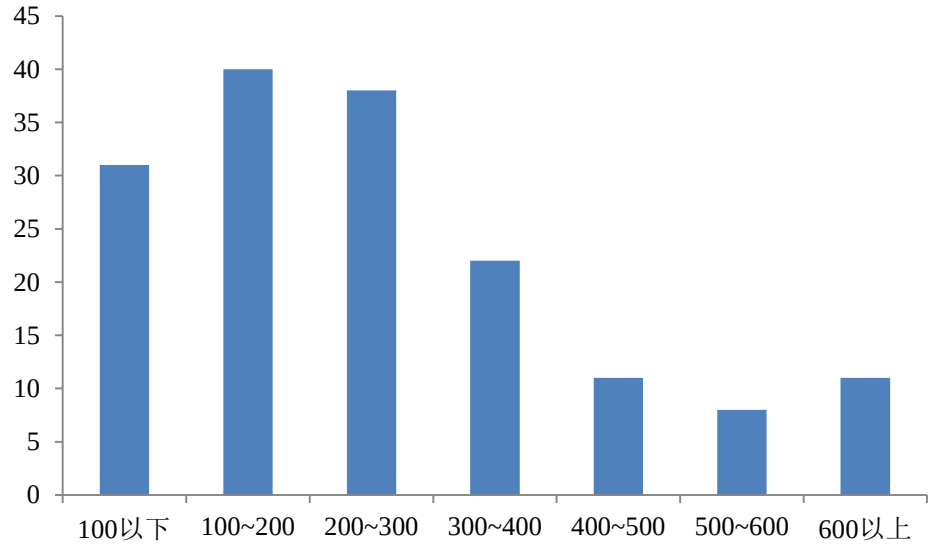

图 1. 样本区/县的村居规模统计图

由上可知，各个样本区/县的村/居数差距较大，其中村/居个数最少的只有 18 个，最多的有 1224 个。而且样本区/县内村/居人口数的差异同样较大，有些村/居的人口数不足 100 人，有些村/居的人口数 20000 人以上。CFPS 在此阶段仅对人口数小于 300 的小规模村/居进行了就近合并，[2] 即，若某一村/居的人口数小于 300，则与邻近的村/居中规模较小的村居合并，直到人口规模超过 300 为止，原则上村与居不能交叉合并，也不能跨街道/乡镇合并。合并的基本原则如下：

⑴ 村/居的合并在同一个乡级单位内进行，除非特殊情况不要跨乡级单位合并。

⑵ 考虑到邻近的村居经常有多个，在合并过程中，我们规定：①若此小村/居是村委会，

[2] 我们对大社区进行了抽样后的事后拆分，具体见下文。

则与邻近人口数比较少的村委会合并；若合并一个后还是不到 300，则按照相同的原则继续合并，直到人口数大于或等于 300 为止；若周围没有村委会，可以与邻近的居委会合并。② 若此小村/居是居委会，则与邻近人口数比较少的居委会合并；若合并一个后还是不到 300，则按照相同的原则继续合并；若周围没有居委会，可以与邻近的村委会进行合并。

⑶ 若小村/居人口数很少，邻近的村/居都十分远，则不进行合并。若某村/居与邻近村/居合并后人口数接近 300，则无需再次合并。若整个乡镇/街道合并后仍不到 300 也无需再次进行合并。

表 4 是最终合并村居情况的统计表：

表 4. 合并村居统计表

省	人口小于 300	合并成的村居个数
安徽	0	0
江西	0	0
福建	51	35
广东	19	13
广西	1	1
贵州	16	12
河南	63	54
湖北	7	5
湖南	39	30
江苏	9	2
四川	126	45
云南	30	13
浙江	21	19
重庆	4	4
北京	0	0
天津	3	3
黑龙江	28	19
吉林	9	8
辽宁	11	4
山东	820	508
陕西	279	161
河北	292	202
甘肃	158	74
山西	525	227
合并	**2511**	**1439**

将上述合并后的村居视为抽样框中的虚拟村居，其人口数为合并村居的人口数和，区域面积为合并村居的区域面积和。由此，得到样本区/县或样本街道/乡镇的村居抽样框。

3.2.2 村/居抽样

为了提高样本的代表性，我们在理想情况下也可以借助辅助信息对村居抽样框进行排序。但在村居层面上，我们并没有 SES 信息可以利用，所以基本是按照计生委提供的顺序，不做更改。有些地方提供的村居委会的排名是按照国统局的代码，这些代码在制作时有一定的经济发展意义。

在除上海外的所有省份中，从每个 PSU 中随机抽取 4 个村/居。在上海，从每个 PSU 中的中随机抽取 2 个村/居。为了形成内隐分层，村/居按照辅助信息进行以下排序：

⑴ 在合适的情况下，将一个 PSU 划分成三个部分：街道办事处、镇和乡。

⑵ 在每个部分中，将所有街道办事处/镇/乡分成两部分：居委会和村委会。在居委会和村委会内部，进行排序。

从每个被抽取的 PSU 组成的抽样框内，采用系统 PPS 抽样方式（见附录 A）抽出 2 个（上海）或 4 个（其他 24 个省份）社区。

最终共得到 640 个村/居样本，其中合并村/居有 8 个，占 1.25%。

需要注意的是，由于从获得各级抽样框到抽取样本会有一定的时间间隔，期间会有极少部分样本村/居拆迁、搬迁、取消等，此时需要在相同乡镇/街道中用相同规模的邻近村/居对其进行替换。在实际调查中，替换的村居共有 11 个（样本区/县更新时在样本区/县的重新抽样的情况不计算在内）。

3.3 第三阶段抽样

3.3.1 末端抽样框[3]

为了控制抽样框误差，提高末端抽样框的质量，此次调查没有用常规的户籍花名册的方式制作末端抽样框，而是通过地图地址法制作末端抽样框，即按照统一的规则, 通过绘制样本村/居的地图，确定每一个住址的地图信息和住户信息，并对其中的空户、非住户、商

[3] 关于末端抽样框制作的更详细信息可参见丁华等，2011，《地图地址抽样框制作手册》；丁华，2012，《中国家庭动态跟踪调查 2010 年基线调查末端抽样框制作（CFPS-2）》。

用、商住两用、一宅多户和一户多宅等特殊问题进行处理，最终得到末端抽样框。[4] 在整理抽样框的过程中，需要特别注意：

⑴ 一宅多户和一户多宅

一宅多户，即一个住宅地址对应着两个或两个以上的住户，例如兄弟分家后仍同住、某间房子出租给多人、老人与成家后的儿女同住但经济独立等。此时按照地址列表清单会遗漏目标单元，产生低覆盖的抽样误差。所以，需要对地址中的多个住户进行拆分编号，并在备注中标明一宅的第几户，以避免实际调查中利用 CAPI 系统的住户过滤系统再次进行住户过滤。注意，此处的一宅多户是实质意义上的一宅多户，一个住户拥有两个户口本的情况不是一宅多户。在实际绘图阶段，判断某个住宅是否为一宅多户可以通过观察门铃、邮箱、独立的进口、电表的个数得知，还可以进行询问相关知情人获得具体的信息。

一户多宅，即一个住户拥有两套或两套以上的房子，若直接按照地址列表清单进行抽样会产生过覆盖的抽样误差。此时，需要按照户主姓名进行筛选，若确实属于一户多宅的情况，需要对其进行合并。注意，此时的一户多宅是实质意义上的一户多宅，若是一家人有两套房子，但是一套房子闲置，不是实质意义上的一户多宅，空置的房子需要标注为空户。同样的，若是另一套房子出租，也不是实质意义上的一户多宅，出租的房子仍是正常住户，但需要在备注中对各种情况进行说明。

⑵ 末端抽样框的边界问题

末端抽样框的界限一般是村级单元的行政区域界限。但是对于一些无明确区域界限范围的村级单位，例如村村混合、村居混合的情况，确定末端抽样框的界限有一定的难度。以村居混合为例，基本的原则是属地原则：若村居混合不严重，即村委会地界中只有少量的居民，居委会地界中只有有少量的村民，此时视村委会地界中的少量居民为该村委会村民，居委会地界中的少量村民为该居委会居民，以此为基础制作住宅分布图；若村居混合现象一般，即村委会与居委会只有交界地带村居混合现象十分严重，且人数很多，则可以在交界地带借

[4] 在制作末端抽样框的过程中，为了提高末端抽样框的整理效率和末端抽样框的信息量，住宅列表清单除了基本的住宅信息外，需要包含住宅状态栏，并在其中标明正常住户（CFPS 需要正常调查的住户），商用（包含商店、医院、药店、商铺等）、商住（包含商住两用、门市、药店等）、非住宅（包含已拆、体育场、公园、车棚、仓库、旅馆、军营、福利院、养老院、寺庙、监狱等）、空房（例如空院、空户、搬迁等），一宅几户（例如一宅两户或一宅三户，并按右手原则对其编号）。一户多宅（例如一户两宅或一户三宅等，需要在每个地址下标明常住的地址），备注中表明具体的描述内容或其他信息。

助主要干道、河流、建筑等按照与村委会和居委会人口数成比例的方式划分村委会和居委会，然后制作相应的住宅分布图和末端抽样框；若混合现象十分严重，即村委会和居委会完全混合，无法区分界限，此时由于制作末端抽样框的原则是属地原则，可以视混合村居的范围为样本村委会或居委会的范围，并依此制作住宅分布图，若该地的流动人口比较少，也可以借助经过村委会或居委会相关人员核查后的户籍花名册制作末端抽样框，无需制作住宅分布图，但是需要得到一张总的住宅分布的概略图。

⑶ 住宅地址类型等无法确定的问题

在实际调查中，由于某些地方统计能力有限，以及人力和物力等原因，无法对当地的商用、商住两用、空户、非住宅、一宅多户、一户多宅的情况进行核实。此时，需要在尽量核查的基础上，按照最大化的原则处理，即无法核查清楚的住宅地址视为正常住户，无法核查的一户多宅视为多个正常住户，无法完全确认是商用还是商住两用的视为商住两用。

⑷ 大社区的拆分问题

在村居抽样框中，由于村居规模相差较大，需要对大社区（人口规模大于一万的社区）进行拆分，因为大社区人口规模较大，入样概率高，若大量入样会扭曲样本结构，另外大量的大社区样本会给村级调查地图的绘制工作带来很大的负担。但是，这将耗费很大的人力、物力和财力，所以在实际操作中，仅对抽中的大社区进行事后拆分，即对抽中的大社区样本增加一个抽样阶段，首先在社/区居委会工作人员的帮助下根据该社区主要的道路、桥梁、河流、建筑物等将社区居委会拆分为几个自然区域，然后根据各个自然区域的住宅类型和人口数，按照与人口数成比例的不等概抽样方式随机抽取一个自然区域，制作村级调查地图和末端抽样框。虽然该方法提高了工作效率、降低了成本，但是增加一个抽样阶段会增加抽样误差，降低估计精度。

对上述由住宅列表清单得到的末端抽样框，为了提高末端抽样框的效率，按照村级调查地图的行走路线或从西北方向起始的顺时针方向对住户列表清单（末端抽样框）进行排序，得到经过排序后的末端抽样框。

最终得到的 640 个村居的末端抽样框按照省市进行整理后，得到住户、有效住户、一宅多户和一户多宅的统计表如下：

表 5. 末端抽样框中住户数量统计表

省市	所有住户	正常住户和商住两用	一宅多户	一户多宅
安徽省	13421	12906	1626	2
广西省	6008	5982	2	776
湖北省	22008	19807	1513	774
江苏省	12195	12001	849	998
江西省	4592	4537	2	219
云南省	17284	17038	461	632
重庆市	4202	4177	55	186
贵州省	21332	21189	44	2427
福建省	7995	6257	647	632
湖南省	20261	19015	173	1761
浙江省	8426	7983	423	582
四川省	18086	16771	80	1356
北京市	5490	5374	0	0
河北省	22266	20178	31	160
黑龙江省	53654	41409	70	89
吉林省	18437	18360	462	61
山东省	15312	14583	9	146
山西省	18957	17691	237	14
陕西省	7553	7230	48	0
天津市	4363	3850	14	9
上海市	87870	83629	1292	346
河南省	62063	58201	256	4334
广东省	62450	53874	31	4022
甘肃省	40917	37868	141	194
辽宁省	90080	82451	168	403
合计	**645222**	**592361**	**8634**	**20123**

由上可知，末端抽样框中有效住户（正常住户和商住两用）的比例约为 92%，[5] 非住宅、空户、商用的比例仅占 8%，[6] 一宅多户的比例是 1.34%，一户多宅的比例是 3.12%。

3.3.2 末端抽样

末端样本量的确定采用扩大样本量的方法，相对于替代法，扩大样本量的方法可以更准确的计算调查中各种率，以便与国际的同类调查比较。由于此次调查的目标总体是满足条件的家庭户，而末端抽样框包含空户、商用、非住宅、无法接触户和拒访户等，在扩大样本容

[5] 由于各地区的统计力量差异以及绘图员工作态度和能力的差异，得到的村级地图和列表清单的质量差异较大，故该部分数据存在误差。

[6] 部分村居由于实际执行的困难，并没有进行住宅类型（非住宅、空户、商用、正常住户和商住）的核查，在抽样过程中，按照最大化原则将其视为正常住户，因此 8% 并不是一个准确的数字。一户多宅和一宅多户的比例同样存在该问题。

量时需要综合考虑上述因素。最终每个样本村居的初访的样本量如下公式所示：

$$n = \frac{\text{目标量}}{\text{1-（拒访率+空户率+无法接触率+无效户率）}}$$

由于各个村/居的上述“率”各不相同，结合已有的调查经验，首先将样本区/县按照已有的数据和调查经验分为三类，然后在每个类中再按照村/居委会的性质以及国家统计局给出的村/居委会是否为城乡结合部的指标分为两类，各类村/居的目标接触量是：

表 6. 2010 年 CFPS 基线调查末端样本量

地区	类型	预计应答率	接触样本数量
低应答率地区	居委会（主城区和城乡结合部的村委会[7]）	60%	42
	其他村委会	70%	36
一般应答率地区	居委会（主城区和城乡结合部的村委会）	70%	36
	其他村委会	80%	32
高应答率地区	居委会	80%	32
	村委会	90%	28

在经过排序后的末端抽样框中，按照循环等距抽样方式抽取 28-42 不等的家庭户样本，最终获得住宅地址样本共 19986 户。在实际操作中，末端抽样的样本量将根据实际调查情况略有调整。

4. CFPS 再抽样样本数据

4.1 CFPS 再抽样样本数据样本量

由 CFPS 的抽样设计可知，CFPS 包含 6 个抽样框，全国的估计量由这 6 个抽样框的样本组合得到。由于五个“大省”是自我代表省，在全国样本中占有很大的比例，因此，无法直接利用 6 个抽样框的样本进行全国的数据分析。为了得到可用于全国推断的样本，此处将全国的数据进行调整，即在 CFPS 的抽样框中按照人口比例调整样本量在各子总体中的分配，分配结果如表 7 所示。

[7] 主城区、城乡结合部的划分由国家统计局设计管理司的城乡代码确定。

表 7. CFPS 全国再抽样样本分配表

层	省/市/自治区	人口统计	人口比例	区/县、乡镇/街道总数	再抽样区/县、乡镇/街道数	再抽样村/居数	目标户数（按有效户 25 户计算）
1	上　海	1858	1.51	32	4	8	200
2	辽　宁	4298	3.49	16	4	16	400
3	河　南	9360	7.6	16	8	32	800
4	广　东	9449	7.67	16	8	32	800
5	甘　肃	2617	2.13	16	2	8	200
6	北　京	1633	77.6	80	80	320	8000
	天　津	1115					
	河　北	6943					
	山　西	3393					
	吉　林	2730					
	黑龙江	3824					
	江　苏	7625					
	浙　江	5060					
	安　徽	6118					
	福　建	3581					
	江　西	4368					
	山　东	9367					
	湖　北	5699					
	湖　南	6355					
	广　西	4768					
	重　庆	2816					
	四　川	8127					
	贵　州	3762					
	云　南	4514					
	陕　西	3748					
合计		**123128**	**100**		**106**	**416**	**10400**

由上，CFPS 再抽样数据是在 CFPS 全部样本数据基础上通过再抽样的方式得到的，即在 CFPS 的 5 个自我代表省中按照与区/县、乡镇/街道人口数成比例的系统 PPS 抽样设计，抽取相应样本区/县、乡镇/街道，样本区/县、乡镇/街道的村/居样本、家户样本、个人样本为 CFPS 的再抽样样本。

4.2 抽样过程

实际抽样过程中，只需对 CFPS“大省”的第一阶段样本进行再抽样，包含两部分：

⑴ 广东省、辽宁省、甘肃省、河南省的第一阶段的再抽样

在广东省、辽宁省、甘肃省、河南省的样本区/县抽样框中，运用省/市和区/县的辅助信息对样本区/县进行排序，排序方式如下：

① 按照地级市的人均 GDP（在没有 GDP 数据的情况下，使用非农业人口比例），对样本区/县所在的地级市进行降序排列。

② 在各地级市中，将各地级市分成区层、县级市层、县层。

③ 分别在区层、县级市层和县层，对区、县级市和县按照其人均 GDP（在没有 GDP 数据的情况下，使用非农业人口比例；在没有 GDP 和非农业人口比例数据的情况下，使用人口密度）进行降序排列。

在经过排序的样本区县抽样框中，按照与样本区县的常住人口数成比例的方式抽取 CFPS 再抽样样本数据的样本区县。

⑵ 上海市的第一阶段的再抽样

上海市由于其特殊性，第一阶段的抽样单元是样本街道/乡镇，首先对样本街道/乡镇框进行排序，排序方式如下：

① 按照区县的人均 GDP（在没有 GDP 数据的前提下，使用非农业人口比例；在没有 GDP 和非农人口比例数据的前提下，使用人口密度），对样本街道/乡镇所在的区县进行降序排列。

② 在各区县内将街道/乡镇分为街道层和乡镇层，按照先街道层后乡镇层的顺序排列。

③ 在街道层和乡镇层中，按照街道或乡镇的人均 GDP（在没有 GDP 数据的情况下，使用非农人口比例；在没有 GDP 和非农业人口比例数据的情况下，使用人口密度）对街道或乡镇进行降序排列。

在经过排序的样本街道/乡镇抽样框中，按照与样本街道/乡镇的常住人口数成比例的方式抽取 CFPS 再抽样样本数据的样本街道/乡镇。

样本区县或街道/乡镇的所有 CFPS 的样本村/居、样本家户和样本个人组成 CFPS 的再抽样样本的数据库。

5. CFPS 权数

在使用 CFPS 数据的过程中，可以通过加权来适当地处理由于抽样设计、无应答及其它原因造成的研究个体（社区、家庭、个人）在 CFPS 2010 数据中具有不平等的被抽中概率、实际中的无应答以及选择性偏差等问题。《中国家庭动态跟踪调查 2010 年基线调查权数计算（CFPS-17）》讨论了 CFPS 各种权数的构建和使用方法。附录 E 给出了 CFPS 2010 权数数据库的变量名及变量标签。

附录 A. 含内隐分层的系统概率规模成比例（PPS）抽样方式

在可行的情况下，抽样框内的单位（units）按照行政边界和社会经济水平加以排序。排序后，采用系统概率规模成比例（PPS）抽样方式抽取样本。这种抽样程序内在的包含了按照被用来对抽样框内的单元进行排序的变量形成的分层。分层用于获得效率（effieciency）和提高估计精度（estimation precision）。

在前两个阶段，抽样在汇总层次（aggregate level）（如县或村）上进行。针对汇总单位（aggregate unit）的系统抽样要求抽样框内代表它们的抽样间距（intervals）与其人口规模成比例。例如，假设 A 省有 10 个区县，M_i 表示每个区县内的常住人口数。假设我们想采用与每个区县内常住人口规模成比例的概率方式抽取 3 个区县。即 $n=3$。附表 1 给出了 3 个区县的常住人口的数量以及具体的抽样过程。

附表 1. 10 个区县的 PPS 抽样示例

区/县 i	区/县 i 的人口数，M_i	累计人口数，T_i	抽取的编码 $R+(j-1)K$
1	1160	1160	
2	18160	19320	
3	8360	27680	22020
4	8840	36520	
5	12300	48820	
6	39440	88260	69486
7	12260	100520	
8	14680	115200	
9	10280	125480	116952
10	16920	142400	

$T=\sum_{i=1}^{10} M_i = 142400$，$n=3$，且 $K=\dfrac{T}{n}=47466$ (对 K 四舍五入取整)

在 1 到 K 的范围内随机选取一个整数，比如 22,020，那么包含 $R=22020$、$R+K=47466+22020=69486$、以及 $R+2K=47466+44040=91506$ 的区县（见 T_i 列）将进入样本。换言之，区县 3、6、9 被抽中。

附录 B.　第一阶段建立区/县抽样框所用的资料来源

建立区/县层次的抽样框所用的主要资料资源包括：

⑴ 2006-2008 年各省统计年鉴。

⑵ 2007 年中国人民共和国全国分县市人口统计数据。

⑶ 2007-2008 中国城市统计年鉴。

⑷ 2008 年中国县（市）社会经济统计年鉴。

⑸ 2009 年中华人民共和国行政区划简册。

附录 C. 第一阶段建立上海街道/乡镇抽样框所用资料来源

本阶段建立抽样框所用的主要资料资源包括：

上海市街道（市）或镇/乡（乡）行政管理资料由国家计划生育委员会和复旦大学提供。

附录 D.　第二阶段建立社区抽样框所用的资料来源

本阶段建立抽样框所用的主要资料资源包括：

抽中区/县当年行政管理数据，由国家计划生育委员会提供。

附录 E.　CFPS 权数变量名及变量标签

数据库	变量名	变量标签
社区问卷数据库	commid_base_weight	标准化后的村居问卷的抽样设计权数
	commid_non_weight	标准化后的村居问卷的无回答调整权数[8]
家庭问卷数据库	fam_base_weight	标准化后的家庭问卷的抽样设计权数
	fam_non_weight	标准化后的家庭问卷的无回答调整权数
	fam_post_weight	标准化后的家庭问卷的事后分层调整权数[9]
成人问卷数据库	adult_base_weight	标准化后的成人问卷的抽样设计权数
	adult_non_weight	标准化后的成人问卷的无回答调整权数
	adult_post_weight	标准化后的成人问卷的事后分层调整权数[10]
少儿问卷数据库	child_base_weight	标准化后的少儿问卷的抽样设计权数
	child_non_weight	标准化后的少儿问卷的无回答调整权数
	child_post_weight	标准化后的少儿问卷的事后分层调整权数11

[8] 不存在缺失的村居没有此变量。
[9] 按照城乡、家庭人口规模的事后分层，辅助信息来源于 2010 普查资料。
[10] 按照城乡、年龄、性别的事后分层，辅助信息来源于 2010 普查资料。
[11] 按照城乡、年龄、性别的事后分层，辅助信息来源于 2010 普查资料。