

CHFS 抽样设计说明

CHFS 的抽样设计包括两个方面：整体抽样方案和绘图与末端抽样方案。为了保证样本的随机性和代表性，同时达到 CHFS 着眼于研究家庭资产配置、消费储蓄等行为的目，抽样设计力求满足如下四个方面的要求：一是经济富裕地区的样本比重相对较大；二是城镇地区的样本比重相对较大；三是样本的地理分布比较均匀；四是尽可能节约成本。

1. 整体抽样方案

总体而言，本项目的整体抽样方案采用了分层、三阶段与规模度量成比例（PPS）的抽样设计。初级抽样单元（PSU）为全国除西藏、新疆、内蒙和港澳地区外的 2585 个市/县。第二阶段抽样将直接从市/县中抽取居委会/村委会；最后在居委会/村委会中抽取住户。每个阶段抽样的实施都采用了 PPS 抽样方法，其权重为该抽样单位的人口数（或户数）。为控制成本，本项目首轮调查的户数设定为 8000—8500 户。

从可操作性角度出发，各阶段样本数设定如下：首先，根据城乡以及地区经济发展水平，末端抽样的户数（即从每个居委会/村委会抽取的户数）设定在 20—50 户之间，其平均户数约为 25 户；其次，在每个市/县中抽取的居委会/村委会数量为 4；最后可以计算得到抽取的市/县个数约为 $8000 \div (4 \times 25) = 80$ 。

（1）第一阶段抽样

第一阶段抽样的目标是从 2585 个市县中抽取 80 个市县。同时，要求 80 个市县的地理分布相对均匀，并且富裕地区的样本不能过少。为达到该目的，我们将 2585 个市县按照人均 GDP 分成十层，在每个层内以市县人口数为权重，采用 PPS 抽样抽取 8 个市县，共抽得 80 个市县，样本涵盖全国 25 个省。表 1 列出了抽取的 80 个市县样本与总体的人均 GDP 描述统计。可以看出，样本与总体在人均 GDP 的分布上是非常接近的。

表 1 总体和 80 个市县样本人均 GDP 分布

人均 GDP	均值	标准差	中位数	Q25	Q75	峰度	偏度
总体	17334.8	17736.9	11370	7173	20263	3.2	17.64
样本	17809.2	19336.3	11349	7232	21143	3.5	20.41

注：Q25 和 Q75 分别表示 25% 和 75% 的分位数

上述抽样过程并没有考虑到抽取出的市县样本的地理分布情况。为了评估在上述抽样方案下样本的地理分布情况，我们通过随机模拟的方式将上述分层 PPS 抽样过程重复 1000 次，得到样本的平均地理分布情况（以东部、中部和西部城市占样本总体的比例衡量）。模拟结果如表 2 所示。

表 2 总体和样本的地理分布

	总体			样本平均（模拟 1000 次）		
	东部	中部	西部	东部	中部	西部
均值	0.343	0.272	0.384	0.367	0.306	0.327
标准差	—	—	—	0.023	0.023	0.023

由表 2 可以看出，在该抽样方案下，样本中东部、中部和西部市县占样本总体的比例十分稳定（其模拟的标准差很小），约为 37:30:33。与总体相比，样本中东部城市的比例更大。但由于我们的抽样方案仅要求样本分布相对均匀，而无需与总体一致，因此该抽样方案能满

足第一阶段的抽样要求。在最终抽取出的涵盖 25 个省份 80 个市县的样本中，东、中、西部省份的比例为 32:27:21。

(2) 第二阶段抽样

第二阶段抽样的目标是从市县中抽取居委会/村委会样本。抽样的关键在合理分配城镇样本和农村样本的比例。由于我国非农人口众多，若按照非农人口比例等比例分配城乡样本，则会出现城镇样本过少的情况。考虑到 CHFS 项目的研究主题是居民资产配置等家庭金融行为，必须按一定的原则分配样本，以达到多抽取城镇样本的目的。其具体实施方法如下：

第一，按照各市县的非农人口比例的分位数，将各市县分成 5 个组。分组的依据是各市县非农人口比重 20%、40%、60% 和 80% 的分位数。

第二，在非农人口比例最大的市县组中，居委会和村委会分配的样本比例是 4:0。

第三，在非农人口比例次大的市县组中，居委会和村委会分配的样本比例是 3:1。

第四，以此类推，在非农人口比例最低的市县组中，居委会和村委会分配的样本比例是 0:4

基于上述样本分配方式，在既定市/县内形成了城镇和农村两个抽样框。在城镇和农村抽样框内，其抽取的居委会和村委会样本数量是已知的。因此可以分别按照各居委会（村委会）的居民户数进行 PPS 抽样。80 个市县的城乡样本分配如表 3 所示。

表 3 城乡样本分配

城镇样本数	频数	比重(%)
0	15	18.75
1	10	12.25
2	15	18.75
3	15	18.75
4	25	31.25

从表 3 可以看出，在我们的城乡样本分配方案下，城镇样本数（居委会数量）为 0 的市县个数为 15，占整 80 个市县的 18.75%；而农村样本数（村委会数量）为零的市县为 25 个，占 80 个市县样本的 31.25%。上述统计数据表明该样本分配方案达到了多抽取城市样本的目的。按照上述方案，在第二阶段抽取出的 320 个居委会/村委会中，城镇样本与农村样本比例为：181:139。

(3) 第三阶段抽样

第三阶段抽样是 CHFS 项目的末端抽样阶段。其目的是从给定居委会/村委会的住户清单列表中抽取访问的住户。在这一阶段的抽样中，农村地区抽取的户数统一设定为 20 户。而在城市地区，我们收集了各社区的平均住房价格信息，并以此作为社区富裕程度的衡量指标。在此基础上，我们根据住房价格由高到低将各社区分成四个组，在住房价格最高的组分配 50 户样本；而在住房价格最低的组分配 25 个样本，以达到进一步抽取富裕家庭的目的。各城市社区的住户样本分配情况如表 4 所示。

表 4 城市社区户数分配

分配户数	社区数量	社区百分比(%)	社区累计百分比(%)
25	53	32.72	32.72
30	52	32.10	64.81
35	28	17.28	82.10
50	29	17.90	100.00

2. 绘图与末端抽样方案

(1) 绘制住宅分布图

本项目的末端抽样建立在绘制住宅分布图以及制作住户清单列表的基础上，借助“住宅分布地理信息”作为抽样框来进行末端抽样。末端抽样框的精度很大程度上取决于实地绘图的精度，因此，如何有效的提高绘图精度成为关键。

CHFS 的绘图采用项目组自行研发的地理信息抽样系统，借助 3G（遥感、GPS、GIS）技术解决了目标区域空间地理信息的采集问题。借助地理信息研究所提供的高精度数字化影像图和矢量地图，绘图员在野外通过电子平板仪加上 GPS 定位获取高精度的测量电子数据，并直接输入到计算机系统中，从而获得高质量矢量底图。考虑到地图数据的时效性，通过后期实地核查人工修正的方式对空间地理数字模型进行修正，建立起与现实地理空间对应的虚拟地理信息空间。

该系统除了使绘图工作人员能直接在电子地图上绘制住宅分布图外，还能储存住户分布信息，并辅助完成末端抽样工作，从而在最大程度上提高工作效率，减少绘图和末端抽样误差。此外，使用电子地理信息抽样系统也有利于保存住户信息资料，为进一步深化和改进项目的工作奠定了基础。该部分核心流程如下图 1 所示：

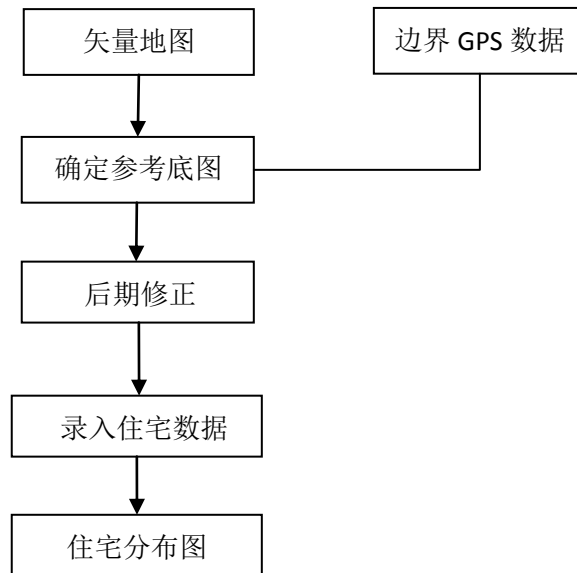


图 1 绘图核心处理流程

(2) 末端抽样

末端抽样基于绘图工作生成的住户清单列表采用等距抽样的方式进行。具体步骤如下：

第一，计算抽样间距，即每隔多少户抽选一个家庭。抽样间距的计算公式为：

$$\text{抽样间距} = \text{住户清单总户数} \div \text{设计抽取户数} (\text{向上取整})$$

若某社区共有住户 100 户，计划抽取 30 户， $100/30=3.33$ 。则抽样间距为 4。

第二，确定随机起点。随机起点的确定为抽样开始时，钟表上分针所处位置的个体数值。如此时时间为 15 时 34 分，则随机起点为 4；如此时时间为 12:03，则随机起点为 3。

第三，确定抽中住户。随机起点所指示的住户为第一个被抽中的住户。在上述例子中，假定随机起点为 3，则第一个被抽中的住户是编号为 3 的住户。则其他被抽中的住户依次为 7，11，15，19，...，依次类推，直至抽满 30 户为止。