CPFS 数据库文档

一、数据介绍

1. 数据库名称

CFPS(China Family Panel Studies)，中国家庭追踪调查

2. 采集数据的执行机构

北京大学中国社会科学调查中心

3. 调查的目的

中国家庭追踪调查（CFPS）是一项全国性、综合性的社会追踪调查项目，旨在通过追踪收集个体、家庭、社区三个层次的数据，反映中国社会、经济、人口、教育和健康的变迁，为学术研究和公共政策分析提供数据基础。

4. 数据的采集年份

2010 年,2012 年,2014 年,2016 年,2018 年

5. 目标总体

CFPS 的样本覆盖中国除香港、澳门、台湾、新疆、西藏、青海、内蒙古、宁夏和海南之外的 25 个省/市/自治区的人口。这 25 个省/市/自治区的人口约占全国总人口（不含港、澳、台）的 95%，因此，CFPS 可以视为一个具有全国代表性的调查。

6. 抽样框及抽样设计

考虑到中国社会有很大的地区差异，同时为了减少调查的运作成本，CFPS 抽样采用了内隐分层的（implicit stratification）、多阶段、多层次、与人口规模成比例的概率抽样方式（PPS）。行政区划和社会经济水平是主要的分层变量。在同级行政层以地方人均 GDP 作为社会经济水平的排序指标；在无法获得 GDP 指标的条件下，则采用非农人口比例或人口密度作为替代指标。

CFPS 每个子样本框的样本都通过三个阶段抽取得到。第一阶段样本（PSU）为行政性区/县，第二阶段样本（SSU）为行政性村/居委会，第三阶段（末端）样本（TSU）为家庭户。

CFPS 前两个阶段的抽样使用官方的行政区划资料，第三阶段则使用地图地址法构建末端抽样框，并采用随机起点的循环等距抽样方式抽取样本家户。考虑到每个地区的应答率，2010 年的实际操作参考了 2008 年和 2009 年预调查所得的预估应答率，采用按应答率比例扩大样本规模的方法，依据系统抽样原则共抽取了 19986 个居住地址，以保证获得预计的有效样本家户数量（见表 3）。

7. 样本量

省/市/自治区类型	省/市/自治区	目标户数	过度抽样比率
“大省”	上海市	1600	10.28
	辽宁省	1600	4.45
	河南省	1600	2.04
	甘肃省	1600	7.30
	广东省	1600	2.02
“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	8000	1.00

二、数据库基本情况介绍

1. 2010 年 CFPS

CFPS 2010 年基线数据库包含村居问卷数据库、家庭关系数据库、家庭问卷数据库、成人问卷数据库和少儿问卷数据库五个类型，分别对应了村/居问卷、家庭成员问卷、家庭问卷、成人问卷和少儿问卷的内容。

1.1 成人库

家庭成员关系库以 CFPS 界定的每个家庭成员为一行，家庭成员以 pid 标识，包括 2010 年基因成员及之后调查年新增的家庭成员的配偶(_s 系列变量)、父亲（_f 系列变量）、母亲（_m 系列变量）及子女（_c1-_c10 系列变量）的基本信息，而且同处一个家庭的成员拥有同样的家户号 fid。

1.2 少儿库

家庭经济库以家庭为单位，fid 为每个家庭的唯一标识符。

1.3 家庭成员关系库

家庭成员关系库以家庭成员为单位，包括 2010 年基因成员

1.4 家庭经济库

家庭经济库以家庭为单位，包括 2010 年基因成员所在原家庭

2. 2012 年 CFPS

2.1 成人库

成人库包括 2010 年界定出来的基因成员中 2012 年追踪调查时年龄处在 16 岁及以上的 个人和 2012 年新增家庭成员中年龄处在 16 岁及以上的个人。访问方式为面访(IWmode=1) 或电访(IWmode=0)， 问卷形式为长问卷(longform=1)和/或短问卷(shortform=1)。

2.2 少儿库

少儿库包括 2010 年界定出来的基因成员中 2012 年追踪调查时年龄处在 15 岁及以下的 个人，以及 2012 年新增家庭成员中年龄处在 15 岁及以下的个人。其中 10 岁及以上的少儿 既有家长代答问卷，也有少儿自答问卷；而 10 岁以下的少儿只有家长代答问卷。访问方式 依然为面访或电访，问卷形式为长问卷和/或短问卷。

2.3 家庭成员关系库

家庭成员关系库以家庭成员为单位，包括 2010 年基因成员及 2012 年新增家庭成员的配 偶、父母及子女的基本信息。2012 年家庭成员关系库中包括来自 13453 个家庭的 55014 条 个人样本。需要注意的是：这 55014 条观测并不代表着 55014 个独立个人，这是因为 2012 年家庭成员库将另组家庭成员分别放在原家庭列表和另组家庭列表中，并用是否在家 (co_a12_p=1 表示在家，0 表示离开原家庭)来表明该个体在 2012 年经济上属于哪个家户。

2.4 家庭经济库

家庭经济库以家庭为单位，包括 2010 年基因成员所在原家庭以及由 2010 年家庭因婚姻 变化、子女经济独立等原因所派生出来的另组家庭。

3. 2014 年 CFPS

3.1 成人库

成人库包括往期调查界定出来的家庭成员中 CFPS2014 调查时年龄处在 16 岁及以上的 基因及核心成员， 以及 2014 年新增家庭成员中年龄处在 16 岁及以上的基因和核心成员。访问方式为面访或电访，回答人可能是受访者自己(selfrpt=1)或家人代答(proxyrpt=1)。

3.2 少儿库

在 CFPS2014 中，有可能存在一个少儿出现多份代答问卷的情况。这是因为对于物理上离家的少儿来说，原家庭将会提供一份代答问卷，当异地追踪成功时，如果在异地有与该少儿同住的家长，将会提供另一份家长代答问卷。也即对于少儿来说，有可能存在两份家长代答问卷共存的情况。

3.3 家庭成员关系库

家庭成员关系库以家庭成员为单位，CFPS2012 家庭关系库为基础，通过 CFPS2014 家庭成员问卷和个人问卷的最新信息，重新构造 CFPS2012 到 CFPS2014 之间家庭成员构成及 其流动状态，补充、变更家庭关系和个人基本信息。为了体现人员跨家庭的流动性，将另组家庭成员分别放在原家庭列表和另组家庭 列表中，并用是否同灶吃饭(也即是否与相应家庭有经济联系，其中 co_a14_p=1 表示经济有 联系，0 表示经济独立)来表明该个体在 CFPS2014 调查时在经济上属于哪个家户。

3.4 家庭经济库

家庭经济库以家庭为单位，包括往期调查所界定出来的原生家庭以及在 2014 年调查时 发现由家庭因婚姻变化、子女经济独立等原因所派生出来的新组家庭。在 2014 年家庭经济 库的 13946 户中，有 1250 户为当年调查时所界定的另组家庭。访问方式为面访（IWmode=1） 或电访(IWmode=2)。

3.5 村居库

村居问卷在 CFPS2012 调查时暂停一轮，到 CFPS2014 时再次执行时，CFPS 样本所分布的村居已由基线时的 643 个扩充到 1976 个。

4. 2016 年 CFPS

4.1 跨年核心变量库

跨年核心变量库是个人层面的数据库，它包含了自 2010 年 CFPS 基线调查以来所有进入 CFPS 样本的个人基本信息。核心变量库中收录的变量可以分为三类：第一类是基线变量（time constant variables），如出生年、性 别、民族，这些变量对于每个个体样本来说只

有一个值。第二类是跨年变量（time varying variables），包括婚姻、收入（个人、家庭人均）、工作状态（在业、失业、退出劳动力 市场）、户口状况（城乡）、居住地城乡状态、是否经济上是一家、是否物理地址上居住在一起、教育(是否在学、最高学历、上学/离校阶段)、迁移。这些变量每轮调查都有一个相应的变量，各轮之间数值可能不等。第三类变量是与访问过程相关的其他变量，包括访问状 态、基因成员类型、死亡状态（死亡时间、死因编码）以及各轮权数。

4.2 成人库

与往年相比，CFPS2016 中成人和少儿库中共用模块变量的比例加大，两个数据库的相似程度比往年有了进一步提高。成人库包括往期调查界定出来的基因成员中 CFPS2016 调查 时年龄处在 16 岁及以上的个人，以及 2016 年新增家庭成员中年龄处在 16 岁及以上的个人。 少儿库包括往期调查界定出来的基因成员中 CFPS2016 调查时年龄处在 16 岁以下的个人， 以及 2016 年新增家庭成员中年龄处在 16 岁以下的个人，其中 10 岁及以下的家庭成员只有 家长的代答问卷，10 岁到 15 岁的家庭成员既有家长的代答问卷，也有个人的自答问卷。无论是成人库还是少儿库，问卷的实际回答人如果是受访者本人，则 selfrpt=1，表明使用了自答问卷；如果问卷的实际回答人是了解受访者情况的其他家人，则 proxyrpt=1，表明使用了代答问卷。在成人库中，代答问卷在以下两种情况下会启动，一是当家庭中有成员外出时，会先由原家庭成员完成一份代答问卷以捕捉外出个人的基础信息；二是当受访者本人由于身体的原因不适合自己回答问卷时（如老年失智症患者），也将由其家人替他完成一份代答问卷。

4.3 家庭成员关系库

家庭成员关系库以家庭成员为单位，包括 2010 年基因成员及之后调查年新增的家庭成员的配偶、父母及子女的基本信息。

4.4 家庭经济库

家庭经济库以家庭为单位，包括往期调查所界定出来的原生家庭以及在 2016 年调查时发现由家庭因婚姻变化、子女经济独立等原因所派生出来的新组家庭。

5. 2018 年 CFPS

CFPS2018 访问问卷包括家庭成员问卷、家庭经济问卷、个人自答问卷、个人代答问卷以及少儿家长代答问卷。

5.1 家庭成员关系库

家庭成员关系库以 CFPS 界定的每个家庭成员为一行，家庭成员以 pid 标识，包括 2010 年基因成员及之后调查年新增的家庭成员的配偶(_s 系列变量)、父亲（_f 系列变量）、母亲（_m 系列变量）及子女（_c1-_c10 系列变量）的基本信息，而且同处一个家庭的成员拥有同样的家户号 fid18。

5.2 家庭经济库

家庭经济库以家庭为单位，fid18 为每个家庭的唯一标识符。

5.3 个人库

个人库包括所有 10 岁及以上个人的问卷数据，个人样本由 pid 唯一标识。Pid 在跨年跨问卷的数据集中保持不变，个人层面的不同数据集都可以通过 pid 来进行链接。除了个人标识符之外，个人库还给出了个人所在的家户号 fid18，可以通过 fid18 作为链接变量进行跨库链接。与往期的成人库相比，CFPS2018 的个人库多出了 10-15 岁样本的个人自答问卷。此外，可以通过年龄的筛选条件保留 15 岁以上的个人样本构建与往期在结构上相似的成人库。

CFPS2018 个人库包含个人自答、个人代答（针对因为身体原因无法完成自答）和家庭代答（针对离家单元中的个人，一般由成员问卷的回答人统一代答）的样本（用变量 PROXYTYPE 进行标识），访问模式为面访和电访（用 self_iwmode 和 proxy_iwmode 进行标识）。

5.4 少儿家长代答库

少儿家长代答库包含的是所有 0-15 岁少儿的家长代答数据。少儿家长代答库的观测单位为少儿，每个被访问到的孩子为一行，以 pid 标识，其代答的家长以 respc1pid 来标识。

少儿家长代答库包括少儿家长代答样本，以及少量来自家庭代答的 0-15 岁少儿样本。数据库中同样用标识符变量 PROXYTYPE 对少儿家长代答（1）和家庭代答（2）进行区分。

注：关于各年份 CFPS 变量及数据库储存方式见附件