2019 年中国家庭金融调查（CHFS）数据使用说明

1 问卷数据变量说明

1.1 数据简介

2019 年中国家庭金融调查样本覆盖全国 29 个省（自治区、直辖市），343 个区县，1360 个村（居）委会，最终搜集了 34643 户家庭、107008 个家庭成员的信息，数据具有全国及省级代表性。中国家庭金融调查数据包含 3 个数据集：家庭数据集、个人数据集、master 数据集。

（1）数据集文件名中含有**“hh”**，代表家庭变量库，包含了问卷中**家庭**部分的数据，例如：家庭农业生产经营情况、住房资产拥有情况等。

（2）数据集文件名中含有**“ind”**，代表个人变量库，包含了问卷中**个人**部分的数据，例如：人口统计特征，个人工作及收入信息，保险与保障等。

（3）数据集文件名中含有**“master”**，代表非问卷变量库，包含了在问卷数据基础上衍生出来的样本地理信息、权重、综合变量等信息。

中国家庭金融调查数据存储格式包含 3 个版本：stata13 版本、stata14 版本、txt 版本。其中，stata13 版本数据建议使用 stata13 软件打开；stata14 版本数据建议用 stata14 及以上版本软件打开。

1.2 样本标识变量

样本标识变量分为**家庭样本标识变量（hhid）**和**个人样本标识变量（pline）**。

hhid 是标识**家庭**的变量，每一个 hhid 代表一个家庭，hhid 可唯一识别家庭，同一家庭在不同年度的 hhid 保持不变。**pline** 是标识每个家庭中**家庭成员**的变量，每一个 pline 代表一个家庭成员；hhid 和 pline 结合起来可唯一识别个体，同一家庭内各家庭成员在不同年度的 pline 保持不变。此外，数据库中还包含个人样本序号变量（pline_order），该变量主要用于与家庭问卷中涉及到加载家庭成员列表的题目进行匹配，例如 B2099a（谁参与了家庭工商业生产活动），其取值与 pline_order 对应，通过 pline_order 可以识别该题选择的家庭成员与个人数据库的对应关系。

1.3 变量缺失值说明

调查数据中的缺失总体可分为两类，**第一类为因为问卷逻辑跳转产生的信息缺失，**对于调查样本而言，有些问题不在逻辑范围内，本来就不需要回答，这类缺失不影响数据完整性。根据变量类型不同，这类缺失取值表现不同，其中**数值型变量取值表现为“ . ”，文本型变量取值表现为空白。**

第二类是在实际访问过程中，由于受访者认知差异、配合情况、访员理解或填答错误等原因，引起的缺失值，在执行过程中需要尽可能避免这类缺失。这些缺失值定义如下：

（1）数据值为“.d”，代表受访户不知道如何回答，由此造成的缺失。

（2）数据值为“.r”，代表受访户拒绝回答该问题，由此造成的缺失。

（3）数据值为“.e”，代表当核查员根据核查信息纠正了某个问题的取值，同时该问题产生一系列需要补充回答的问题，这些问题是受访户没有被询问到的，由此造成的缺失。

（4）数据值为“.n”，代表核查中发现受访户并未就该某个题给出答案，而为**访员臆答（自行猜测答案）**的结果，原始值不能直接被纳入数据库，由此造成的缺失。

1.4 变量中的其他选项

在调查问卷的单选题或多选题中，有些问题设置了**“其他（请注明）”选项**，对应取值设置为**“7777”**。若受访者选择了该选项，一般会继续追问需要注明的信息，该信息被存储在一个新的变量中，用原变量后加上后缀**_ex1 表示**。例如：[C1013c] 您寻找租住房源的主要方式？

1.中介机构（链家、贝壳等）　　2.黑中介（非正规中介）

3.熟人介绍（亲友、同学、同事等）　　4.街头广告

5.网络平台（58 同城、赶集网、BBS 论坛等）　　7777.其他（请注明）。

该题设置了“7777”选项，c1013c_ex1 表示受访户寻找租住房源的方式是其他途径（自己寻找房东，儿女帮忙租房等），且具体内容是访员注明的信息。

除此之外，个别多选题设置了表示**“以上都没有”的选项，对应取值设置为“7788”**，即对于问卷列出的所有选项，受访户均不符合。例如：[A3136ba] 去年，该工作为【CAPI 加载家庭成员姓名】提供了哪些现金福利？以现金形式发的或打到工资卡里的都算。［可多选］

1.交通费补贴　　2.餐费补贴

3.住房补贴　　4.通讯费补贴

5.出差补贴　　6.过节费

7777.其他（请注明）　　7788.以上都没有【跳至 A3136bc】

1.5 循环问题

对于所有循环询问的问题，命名规则为在原变量名后加上**后缀“_#”**；“#”代表第#次**循环**。例如，c2003_1 表示第一套房子的建筑面积；c2003_2 则表示第二套房子的建筑面积。

除此之外，问卷中还包含其他由于加载而产生的循环，如：[D5104a] 按照投资标的不同，您家拥有的基金主要是什么类型？［可多选］

1.股票型　　2.债券型

3.货币市场基金　　4.混合型

5.QDII 型　　6.商品型

7.指数型或 ETF　　7777.其他（请注明）

[D5107] 目前，您家拥有的【CAPI 逐一加载 D5104a 所选选项】基金的总市值是多少钱？（单位：元）

此变量循环了其加载变量 D5104a 的选项，d5104a 的选项 1 为“股票型”，选项 7777 为“其他（请注明）”，因此 d5107_1 表示股票型基金的总市值，d5107_7777 表示其他类型基金的总市值。

特殊的，在**“农业/工商业信贷”**部分的问题中，命名规则为：**农业相关的信贷**问题，原变量名后加上**“_1”**代表[1]；工商业相关的信贷问题，原变量名后加上**“_2”**代表。例如：

b3030d_1 表示农业生产经营的民间借款计划借款额，b3030d_2 表示工商业生产经营的民间借款计划借款额。b3004ba_#_*,此处的“#”表示依次循环 b3004ba 的加载变量 b3004c 的选项，“*”则表示农业/工商业的循环，即 b3004ba_#_*表示在第*类生产经营中，通过加载变量 b3004c 选项“#”的方式计划贷款金额。也就是说，b3004ba_1_1 表示农业生产经营中，计划通过网上银行（电脑端）申请的贷款金额，相应地，b3004ba_1_2 则表示工商业生产经营中，计划通过网上银行（电脑端）申请的贷款金额；b3004ba_2_1 表示农业生产经营中，计划通过手机银行（银行 APP 端）申请的贷款金额，b3004ba_2_2 则表示工商业生产经营中，计划通过手机银行（银行 APP 端）申请的贷款金额。

1.6 多项选择题

对于多项选择题，处理原则为将每一个选项转换为取值为 0 和 1 的哑变量。多项选择题的命名规则为在原变量名后加上**后缀“_*_mc”；“*”代表第*个选项**。

（1）非循环多项选择题

非循环多项选择题的命名规则为在原变量名后加上后缀“**_*_mc**”；“*”代表第*个选项。

例如：[A3118]上班采用的交通工具是？［可多选］

1.公共交通	2.私家车
3.打车	4.电动车或摩托车
5.自行车	6.步行
7.单位班车	7777.其他（请注明）

a3118_1_mc 表示是否勾选 a3118 的第一个选项“公共交通”；0 表示未选择，1 表示选择。a3118_2_mc 表示是否勾选 a3118 的第二个选项“私家车”；0 表示未选择，1 表示选择。

（2）循环多项选择题

循环多项选择题的命名规则为在原变量名后加上后缀“**_#_*_mc**”；**“*”代表第*个选项，“#”代表第#次循环**。

例如：b6001a_2_1_mc 表示在第二次循环（第二笔互联网借款）时是否勾选第一个选项“P2P 平台”，即此变量表示受访户家里的第二笔互联网借款是否来自 P2P 平台，0 表示未选择，1 表示选择。又如：e3005dh_7_2_mc 表示在第七次循环（日常消费）时是否勾选第二个选项“亲朋好友”，即此变量表示因日常消费产生的信贷需求，是否要通过亲朋好友渠道借入所需资金，0 表示未选择，1 表示选择。

[1] 数据库中不包括“农业信贷”部分的相关变量，也就是数据库中“农业/工商业信贷”模块不包括“_1”结尾的变量，这是由于这部分变量的权属归于中心合作单位，不包括在此次公开发布数据范围内。

1.7 插值变量

为了解决数据缺失问题，我们对部分重要的变量进行了插值处理。插值变量的相关说明如下：

以变量 d3109（您家持有的所有股票目前市值是多少）为例，在访问过程中若受访者没有回答所有股票目前市值的具体值，则进一步询问其区间值 d3109it（这些股票市值大概在下列哪个范围）。

（1）在 d3109 有取值且该取值合理的情况下，插值变量等于原变量。即：对于变量 d3109,若受访者回答该问题，此时 d3109_imp 等于 d3109。

（2）当 d3109 取值异常或者 d3109 需要回答但是表现为缺失时，根据受访者提供的 d3109it 信息以及其他辅助信息，建立模型进行插补。插值的方法主要有**逻辑插补**和**多重插补**。

对于插值后的变量，命名规则为原变量名加上后缀_imp。

1.8 地理信息相关变量说明

为了在保护受访者信息的前提下能尽可能满足学者研究的需要，CHFS 公开数据未披露样本所属省级以下具体地理信息。用户可在 **CHFS 公开数据**基础上进行研究，但**不得探索 CHFS 样本地市、区县、乡镇/街道、社区的具体名称、国标码等相关信息**。

公开数据集中，有部分变量与省市县地理信息相关，该类变量在数据集中对应四个变量：省份、省份国标码、城市伪码、区县伪码。例如：［A2019］【CAPI 加载姓名】现在的户口所在地是____省____市_____区/县？【CAPI 列出省市县列表】。该变量在数据集中有四个对应的变量：a2019_prov、a2019_prov_code、a2019_city_lab、a2019_county_lab。同时，**若家庭或个人数据集中该类变量的省市县地理信息与非问卷变量数据集（master 数据集）中地理信息一致，则对应的城市伪码和区县伪码保持一致。**

1.9 截尾处理

根据统计法规定，为保护受访者隐私，我们对**部分变量的极值**进行了截尾处理。将超过某一规定值的样本替换为该规定值，并同时给出一个截尾处理的哑变量，命名规则为变量名加上后缀_censor，即 varname_censor，1 表示进行了截尾处理，0 表示未进行处理。表 4 显示了已经进行截尾处理的相关变量的情况。

表 1　截尾处理变量列表

变量名	截尾标准	影响样本数
b2003a	5000000	14
b2003a_imp	5000000	14
b2003b	5000000	10
b2003b_imp	5000000	10
b2003d	5000000	7
b2003d_imp	5000000	7
b2050	5000000	63
b2050_imp	5000000	63

b2052	8000000	32
b2052_imp	8000000	32
b2055	5000000	7
b2055_imp	5000000	7
b2059	3000000	18
b2059_imp	3000000	18
b2063	3000000	8
b2063_imp	3000000	8
b2080	2000000	19
c2016_1	8000000	42
c2016_1_imp	8000000	42
c2016_2	8000000	5
c2016_2_imp	8000000	5
c2016_3	8000000	1
c2016_3_imp	8000000	1
c2016_4	8000000	76
c2016_4_imp	8000000	76
c2016_5	8000000	10
c2016_5_imp	8000000	10
c2016_6	8000000	1
c2016_6_imp	8000000	1
c3019a	3000000	19
c3019a_imp	3000000	19
c7052b	2000000	4
c7052b_imp	2000000	4
d1105	3000000	4
d1105_imp	3000000	4
d2104	2000000	9
d2104_imp	2000000	9
d3109	2000000	4
d3109_imp	2000000	4
d7110a	2000000	13
d7110a_imp	2000000	13
k1101	2000000	1
k1101_imp	2000000	1
k2102c	3000000	10
k2102c_imp	3000000	10
f1028	200000	1
f1028_imp	200000	1

2 master 数据集变量说明

2.1 资产、负债、收入、消费变量说明

CHFS 数据库中**综合变量**主要包含四个：家庭总收入（total_income）、家庭总消费（total_consump）、家庭总资产（total_asset）和家庭总负债（total_debt），都属于项目组自行汇总生成的家庭层面变量，对于同一户的每个家庭成员，这些变量取值是相同的。

家庭总收入包括工资性收入、农业收入、工商业收入、财产性收入和转移性收入。CHFS 数据库中，部分家庭收入为负数，主要由于生产经营性项目亏损或者金融市场投资亏损导致。

家庭总消费包括食品消费、衣着消费、居住消费、家庭设备服务消费、交通通信消费、教育文娱消费、医疗保健消费和其他消费。

家庭总资产包括金融资产和非金融资产。**金融资产**包括现金、存款、理财产品、股票、基金、债券、衍生品、非人民币资产、黄金、其他金融资产、借出款、社保账户余额。**非金融资产**包括农业资产、工商业资产、房屋资产、商铺资产、土地资产、车辆资产、车库资产和其他非金融资产。

家庭总负债按照负债成因主要分为农业负债、工商业负债、房屋负债、商铺负债、车辆负债、其他非金融资产负债、金融资产负债、教育负债、信用卡负债、医疗负债和其他负债。

2.2 权重变量

在我们的抽样设计下，由于每户家庭被抽中的概率不同。在分析总体情况时，需要使用抽样**权重进行调整**来推断总体情况。在 master 数据集中分别给出了**家庭样本权重（weight_hh）和个人样本权重（weight_ind）。进行家庭和个人层面的分析时，建议使用对应的样本权重进行调整。**

抽样权重计算过程如下：

第一，根据每阶段的抽样分别计算出调查市县被抽中的概率 p1、调查社区（村）在所属区县被抽中的概率 p2 、以及调查样本在所属社区（村）被抽中的概率 p3，分别计算出三阶段的抽样权重 w1=1/p1、w2=1/p2、w3=1/p3，最后得家庭样本的抽样设计权重 weight_hh=w1×w2×w3。

第二，由于我们的末端抽样单位是以家庭为单位，家庭内部每个个体的抽样设计权重相同。考虑到抽样设计的复杂性、实际调查过程中问题多样性，在使用抽样设计权重调整后，样本在城乡、性别、年龄结构等维度仍可能存在偏差。我们进一步使用官方公布数据，从城乡、性别、年龄结构等维度对抽样权重进行进一步调整，经过调整后得到个人调整权重 weight_ind。

2.3rural 变量

若 rural 等于 **1**，则代表**乡村**；若 rural 等于 **0**，则代表**城镇**。其中城乡的定义[2]，城镇包括城区和镇区。城区是指在市辖区和不设区的市，区、市政府驻地的实际建设连接到的居民委员会和其他区域。镇区是指在城区以外的县人民政府驻地和其他镇，政府驻地的实际建设连接到的居民委员会和其他区域。与政府驻地的实际建设不连接，且常住人口在 3000 人以上的独立的工矿区、开发区、科研单位、大专院校等特殊区域及农场、林场的场部驻地视为镇区。乡村是指城镇以外的区域。

2.4 地区分类变量

根据样本的地理信息，添加了不同口径下的地理变量，分别为 region（按东、中、西、东北部划分）、city_level（按一、二、三线城市划分）。

region，按东、中、西、东北划分[3]。**东部**包括：北京、天津、河北、上海、江苏、浙江、福建、山东、广东和海南。**中部**包括：山西、安徽、江西、河南、湖北和湖南。**西部**包括：内蒙古、广西、重庆、四川、贵州、云南、西藏、陕西、甘肃、青海、宁夏和新疆。**东北**包括：辽宁、吉林和黑龙江。

city_level，按一、二、三线城市划分[4]。一线城市包括：上海、北京、深圳、广州；新一线城市：成都、杭州、重庆、西安、苏州、武汉、南京、天津、郑州、长沙、东莞、佛山、宁波、青岛、沈阳；二线城市包括：合肥、昆明、无锡、厦门、济南、福州、温州、大连、哈尔滨、长春、泉州、石家庄、南宁、金华、贵阳、南昌、嘉兴、珠海、南通、惠州、太原、中山、徐州、绍兴、常州、台州、烟台、兰州、潍坊、临沂。其余城市归为三线及以下城市。

2.5 抽样地址与常住地址

master 数据集中的地理信息是指抽样地址，是指基于抽样设计，样本家庭第一次接受访问时的常住地址。**一般而言，对于同一家庭，在不同轮次调查的抽样地址保持不变。**对于每轮调查中首次被抽中的家庭，抽样地址等于常住地址；对于每轮调查中的追踪家庭，如果追踪家庭搬离了首次被抽中时的常住地，这时候抽样地址不再等于常住地址。据统计，总体而言，抽样地址相对于常住地址，1%左右的样本家庭抽样地址和常住地址在省市层面存在差异，5%在区县层面存在差异。

2.6 分卷变量

CHFS 问卷需要搜集的信息含量大，对受访者的配合要求较高。为此，问卷设计中，采用分卷的方式，将所有样本随机分成若干卷，主要是为了尽可能降低受访者回答负担。以 pab 变量为例，该变量表示将目标

[2] 《统计上划分城乡的规定》：http://www.stats.gov.cn/tjsj/tjbz/200610/t20061018_8666.html
[3] 《东西中部和东北地区划分方法》：http://www.stats.gov.cn/ztjc/zthd/sjtjr/dejtjkfr/tjkp/201106/t20110613_71947.htm
[4] 第一财经新一线城市研究所《2021 城市商业魅力排行榜》：https://www.yicai.com/news/101063860.html

样本随机分为 A、B 两卷，一些问题只需要 A 卷受访户需要回答，另外一些则只需要 B 卷受访户回答，而不需要目标受访户同时回答 A、B 卷适用的所有问题。**因此，适用于分卷的变量，样本量会适当减少。**分卷信息在问卷中有明确标识，请注意查看。

3 常见数据问题解答

本部分，我们梳理了 CHFS 数据用户在使用历年 CHFS 数据时咨询频率较高的问题，以帮助大家更恰当地使用 2019 年 CHFS 数据。

3.1 关于年龄变量

问：问卷中与出生年份相关的变量有很多，如果想计算受访家庭成员的年龄，该选用哪个变量呢？

答：2019 年家庭金融调查问卷中与家庭成员**出生年份**的变量有三个。第一个是变量 A1106c，该变量主要是对追踪受访户原家庭成员进行确认；第二个是变量 A1114，主要是对追踪受访户新增家庭成员进行询问；第三个是变量 A2005，该变量包含追踪家庭和新访家庭的出生年份信息。所以，在计算受访者年龄时，数据用户可以直接通过**家庭成员出生年份 a2005** 这一变量，计算受访者年龄（等于接受访问的年份减去出生年份）。

3.2 关于问卷分卷

问：数据库中 A、B 等分卷是什么含义，不同类别分卷具体的区别在哪？

答：CHFS 问卷需要搜集的信息含量大，对受访者的配合要求较高。为此，问卷设计中，采用分卷的方式，将所有样本随机分成若干卷，主要是为了尽可能降低受访者回答负担。以 pab 变量为例，该变量表示将目标样本随机分为 A、B 两卷，一些问题只需要 A 卷受访户需要回答，另外一些则只需要 B 卷受访户回答，而不需要目标受访户同时回答 A、B 卷适用的所有问题。分卷信息在问卷中有明确标识，请注意查看。分卷信息在问卷中有明确标识，如 A3100a—A3160a，问卷中明确标注了仅询问 A 卷城镇样本且是受访者本人，请数据用户特别注意查看**问卷中所有问题前后的适用条件和相关说明。**

3.3 关于家庭总收入负数

问：数据库中已经生成了单独的家庭负债 debt 变量，为什么总收入 total_income 还会有负值呢？

答：家庭收入为负数是**正常现象**，主要由于生产经营性项目亏损或者金融市场投资亏损导致。收入变量跟负债变量是两个单独核算的变量，收入为负数的原因跟负债没有直接联系。

3.4 关于家庭识别码 hhid

问： 同一个家庭编号 hhid 有多个重复值，为什么会这样？是同一家庭不同人员吗？

答：**hhid** 是标识**家庭**的变量。每一个 hhid 代表一个家庭；hhid 可唯一识别家庭。**pline** 是标识每个家庭中**家庭成员**的变量。每一个 pline 代表一个家庭成员；hhid 结合 pline 可唯一识别个体，同一家庭下的同一个家庭成员不同时期的 pline 保持不变。所以，在家庭数据集中 hhid 是唯一的、不可重复的；在个人数据集中包含一个家庭中的多个家庭成员，因此会出现 hhid 多个重复值，hhid 结合 pline 是唯一的、不可重复的。

3.5 关于多选题取值含义

问：一些变量，如 f6704，存在值为 1-2-3-6，是代表 1、2、3、6 四个选项都选了吗？

答：若数据集中多选题具体取值等于 1-2-3-6，则代表受访者在回答该题目时依次选择了 1、2、3、6 四个选项。

对于多项选择题，数据集中会有其对应的多项选择题拆分变量，处理原则为将每一个选项转换为取值为 0 和 1 的哑变量。多项选择题的命名规则为在原变量名后加上**后缀“_*_mc”；“*”代表第*个选项**。若“_*_mc”取值等于 1，则代表受访者回答该题时，选择了第*个选项。

3.6 关于 CHFS 数据统计结果与其他来源数据的差异

问：为什么中心的数据与统计局数据或者其他调查数据不一致？

答：CHFS 数据的统计结果与其他来源数据存在差异，可能的原因比较复杂，从数据收集阶段来看，调查**抽样方案、变量统计口径、概念相关定义**存在一定程度的差异；从数据清理阶段来看，数据基础清理、深度清理过程中**处理方法、插值模型**等内容存在差异；从数据分析阶段来看，各个**研究内容、研究方法、研究模型**存在差异。

若您在数据使用过程中有任何疑问，欢迎将问题发送至数据服务公共邮箱 contactus@chfs.cn，中心项目组成员会每周定期、集中对数据问题进行汇总回答。