CHFS2017 年家庭金融调查数据使用说明

一、 数据清理

1、清理过程

数据采集回来后，中心会对所有数据进行初步处理，主要处理包括：删除由于访员严重臆答、作弊的无效样本，删除无效变量，删除敏感数据；校正人为导致的重复样本编号，校正访员主动报备的人为误操作；合并追踪和新访数据，拆分家庭和个人数据，拆分多选；加注标签，加注问卷类型；清理备注和其他选项，数值题插值及相关规则确定等。

经过初步处理后数据生成可使用的版本。在数据使用过程中如再发现极值或异常值，中心会进行二次录音核查确认。如果没有录音，则通过创建模型处理极值和异常值，从而更新数据版本并告知用户。

2、数据存储

数据清理结束，生成的中国家庭金融调查数据存储在以下 3 个数据集里面。

1. 数据集文件名中含有“**hh**”，代表问卷中家庭部分的数据，例如：金融知识、基层治理与主观评价等；
2. 数据集文件名中含有“**ind**”，代表问卷中个人部分的数据，例如：人口统计特征（部分），个人工作及收入信息，保险与保障，家庭成员教育等；
3. 数据集文件名中含有“**master**”，代表非问卷变量数据，主要有权重变量、样本追踪情况变量、样本地理信息、综合变量（家庭总收入、家庭总消费、家庭总资产、家庭总负债）等。

除此之外原始数据使用格式对应为 dta 格式和 txt 格式，其中 dta 格式主要包含两个版本：stata13版本和stata14版本。其中，13版数据建议使用stata13软件打开；14版数据建议用stata14软件及以上版本软件打开，可根据自身需要自行选择数据版本。

二、 变量命名规则

家庭金融调查问卷数据的变量名由首位的字母及后面四位数字构成，不同的首字母对应问卷的不同部分（如，b-农业/工商业；d-金融资产），一些特别的标识性变量（id 变量、

城乡、省份等）则根据它们的含义被重新命名以便使用。部分变量在原变量名后加上后缀字母it，表示对前一个问题的数值范围追问。部分变量在原变量名后加上后缀 ex1，表示对该问题受访户选择“其他”选项的备注说明内容。部分变量在原变量名后加上后缀 ms，表示对前一个问题答案的二次确认。例如：“family01ms”该变量是指：再次确认老访户家庭成员人数；“family02ms”该变量是指：再次确认新访户家庭成员人数。“house01ms”该变量是指：再次确认受访户家的住房数量。

问卷中的每个问题前均给出了对应的变量名称，需要特别注意的变量及其命名规则说明如下：

1、id 变量

id 变量分为家庭变量（hhid）和个体变量（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。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 可唯一识别个体，同一个家庭同一个家庭成员不同时期的 pline 保持不变。

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，因此每户家庭代表的中国家庭数量也就不同。在推断总体的时候，需要通过权重的调整来真实准确地反映每户样本家庭代表的家庭数量，以获得对总体的正确推断。中国家庭金融调查的所有计算结果都经过抽样权重的调整。在 master 数据集中含有权重变量：“weight_hh”代表家庭权重，“weight_ind”代表个人权重。

其抽样权重的计算方法如下：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，最后得到该样本的抽样权重为 weight_hh =w1×w2×w3 。

考虑到样本在性别、年龄、地区等属性上与全国人口偏差较大，因此还会进行分组调整，基于国家统计局人口结构和总数，设定每个组相应的调整系数值。家庭权重（weight_hh）为没有经过调整的权重，个人权重（weight_ind）等于家庭权重乘以调整系数。

3、变量中的其他选项

在数据变量单选题或多选题中，一般情况下最后 1 个选项为“其他（请注明）”，选项设置为“7777”。若受访者选择了该选项，对应会多问 1 道填空题。在 2017 年家庭金融调查数据

中，该新生成的变量命名格式为：原变量后加上后缀_ex1，例如：f4103_ex1 表示受访户没有缴纳住房公积金的原因是其他相关信息，且具体内容是访员注明的信息。除此之外，个别多选题还会有 1 个选项为“以上都没有”，选项设置为“7788”。

4、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 c5003 等，在此不一一列出。

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

5、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。多项选择题分为两类：非循环多项选择题、循环多项选择题。

（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，b1004_1_mc 表示是否勾选 b1004 的第一个选项“粮食作物”；0 表示未选择，1 表示选择。非循环多选问题的变量列表如下：

表 1　非循环多选题处理说明

varname_*_mc	变量信息
0	第*个选项未选择
1	第*个选项已选择

（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，b3002a_2_1_mc 表示在第一次循环时是否勾选第二个选项“估计贷款申请不会被批准”；0 表示未选择，1 表示选择。循环多选问题的变量列表如下：

表 2　循环多选题处理说明

varname_*_#_mc	变量信息
0	在第#次循环中第*个选项未选择
1	在第#次循环中第*个选项已选择

6、插值变量

为了解决数据缺失问题，我们对所有题目里含“it”的题进行了插值处理。

插值变量的具体计算过程如下：

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有取值且该取值不为异常值的情况下，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取值为异常值或者d3109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回归插补。

对于插值后的变量，命名规则为原变量名加上后缀_imp，插值问题的变量列表如下：

表 3　插值问题的变量列表

变量名称				
c7052b_imp	d1105_imp	d6116_imp	f2006_#_imp	a3109_imp
c7060_imp	d2104_imp	d8104_imp	f2004_#_imp	a3136_imp
c7008_#_imp	d3103_imp	d8106_imp	f1031_imp	a3137_imp
c7008a_#_imp	d3109_imp	d9103_imp	f1029_imp	a3136b_imp
c7008b_#_imp	d3110_imp	d9105_imp	f1028_imp	a3136a_imp
c7009a_#_imp	d3116_imp	d9110a_imp	f1010_imp	a3125_imp
c7059_imp	d3116b_imp	d9110b_imp	f1008_imp	a3124_imp
c7058_imp	d3117_imp	d9108_imp	f1005_imp	a3123_imp
c7061_imp	d5107_imp	c2016_#_imp	f6503_imp	a3113_imp
c7062_imp	d5108_imp	c2023d_imp	f6502_imp	a3112_imp
c8002_imp	d5109_imp	c2023e_imp	f6204_imp	h2004_#_imp
c8002a_imp	d7106h_imp	c3019c_imp	f6203_imp	h2002_#_imp
c8005_imp	d7106j_imp	c3019e_imp	f6119_imp	h1004a_#_imp
c8005a_imp	d7110a_imp	c3024_imp	f6110a_imp	h1003c_imp
c8007_imp	d7112_imp	c3025_imp	f6106_imp	h1002b_imp
d4103_#_imp	k1101_imp	c2016_3_imp	f6103_imp	h1002a_imp
d4111_imp	k2102c_imp	c2016_4_imp	f4011_imp	g2004a_#_imp
d6100a_imp	k2208_imp	c3019a_imp	f4008_imp	g2003b_imp
c1011_imp	c2027d_#_imp	b2003e_imp	f4005_imp	g2003a_imp
c1014_imp	c2032_#_imp	b2051_imp	f2026_imp	g2003_imp
c1015a_imp	c2035a_#_imp	b2110_imp	f2025_imp	g1023d_imp
c2000f_#_imp	c2064_#_imp	b2055_imp	f2024_imp	g1022_imp
c2002d_imp	c3002_#_imp	b2003a_imp	e1006_imp	g1020_imp
c2011e_#_imp	c3002a_#_imp	b2003d_imp	e1022_imp	g1019a_imp
c2013_#_imp	c3017ca_imp	b2003b_imp	e4003_imp	g1019_imp

b2046_imp	b3004b_2_imp	b2101_imp	e2001b_imp	g1018_imp
b2050_imp	b3005b_2_imp	b2102_imp	e2002c_imp	g1017_imp
b2052_imp	b3005_2_imp	b2104_imp	e2002e_imp	g1016_imp
b2059_imp	b3006a_2_imp	b2105_imp	e3003c_imp	g1012_imp
b2063_imp	b3030d_2_imp	b2103_imp	e3005c_#_imp	g1011_imp
b2080_imp	b3030e_2_imp	g1001_imp	g1008_imp	g1004_imp
b2093_imp	b3031a_2_imp	g1006_imp	g1009_imp	g1005_imp
b2099_imp	b3045c_2_imp	g1006a_imp	g1010_imp	g1007_imp
b2100_imp	b3056a_2_imp			

7、截尾处理

根据统计法规定，为保护受访者隐私，我们对收入和资产的极值进行了截尾处理。将收入或资产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前缀 censor_，即 censor_varname，0 表示未进行处理；1 表示进行了截尾处理。

表 4　截尾处理说明

censor_varname	变量信息
0	未处理
1	已处理

表 5　截尾处理变量列表

变量名	截尾标准	影响样本数
b2003a	5000000	24
b2003a_imp	5000000	31
b2003b	5000000	14
b2003b_imp	5000000	16
b2003d	5000000	16
b2003d_imp	5000000	18
b2003e	3000000	9
b2003e_imp	3000000	11
b2052	8000000	44
b2052_imp	8000000	44
b2055	5000000	11
b2055_imp	5000000	14
b2059	3000000	18
b2059_imp	3000000	13
b2063	3000000	7
b2063_imp	3000000	9
b2080	2000000	9
b2080_imp	2000000	6
c2016_1	8000000	158
c2016_1_imp	8000000	166
c2016_2	8000000	28

c2016_2_imp	8000000	30
c2016_3	5000000	30
c2016_3_imp	5000000	32
c2016_4	5000000	9
c2016_4_imp	5000000	9
c2016_5	2000000	4
c2016_5_imp	2000000	4
c2016_6_imp	2000000	1
c3019a	5000000	13
c3019a_imp	5000000	18
c7052b	2000000	6
c7052b_imp	2000000	6
c7059	2000000	1
c7059_imp	2000000	1
c7062	3000000	1
c7062_imp	3000000	2
c8005	2000000	9
c8005_imp	2000000	9
d1105	3000000	11
d1105_imp	3000000	12
d2104	2000000	6
d2104_imp	2000000	9
d3103	2000000	5
d3103_imp	2000000	6
d3109	2000000	6
d3109_imp	2000000	7
d5107	3000000	2
d5107_imp	3000000	2
d7106h	2000000	1
d7106h_imp	2000000	1
d7110a	2000000	9
d7110a_imp	2000000	10
h2004_2	2000000	11
h2004_2_imp	2000000	12
k1101	2000000	2
k1101_imp	2000000	1
k2102c	3000000	11
k2102c_imp	3000000	12

8、其它特殊变量说明

表 6　特殊变量说明

变量归属	变量名称	变量取值	变量取值含义	补充说明
hh 数据集	**track**	0	新访户	

		1	追踪受访户	
ind 数据集	**hhead**	0	非户主	
		1	户主	
	ts003	1	昨天或现在在家	仅询问新访户 3 岁及以上家庭成员
		2	昨天或现在不在家	
master 数据集	**qc**	0	样本质量正常	
		1	样本质量不高	由于受访户客观回答的不知道/拒绝比例较高或者访员主动报备质量不高导致
	rural	0	城镇	
		1	农村	
	pab	a	A卷	受访户回答的是A卷问题
		b	B卷	受访户回答的是B卷问题

9、资产、负债、收入、消费变量说明

中心的综合变量主要包含四个：家庭总收入（total_income）、家庭总消费（total_consump）、总资产（total_asset）、家庭总负债（total_debt），四个综合变量，是单独计算相互独立的变量。

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、转移性收入。中心数据集中，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、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**包括存款、股票、基金、理财、债券、衍生品、非人民币资产、黄金、其他金融资产、现金、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、其他非金融资产。

家庭总负债按照负债成因主要分为金融资产负债、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、教育负债、信用卡负债、医疗负债和其他负债。

10、部分缺失值说明

实际访问受主客观因素影响，由于受访者主观知识以及态度、访员理解或填答错误等原因，不可避免地会产生少量的缺失值。这些缺失值定义如下：

1. 数据值为“.d”

受访户不知道如何回答。访员选择了“不知道”（该选项实际问卷题目中没有）。由于不知道造成的缺失。

2. 数据值为“.r”

受访户拒绝回答该问题，访员选择了“拒绝回答”（该选项实际问卷题目中没有）。由于拒绝回答造成的缺失。

3. 数据值为“.e”

核查专员由于校正了访员在题目（该题目含有逻辑跳转设置）上的错误填答，而该题目后续新生成的问题是实际受访户没有被询问到的，核查专员无法就此新生成的问题进行正确填答，则被设置为“.e”。

4. 数据值为“.n”

核查中发现受访户并未就该题给出答案，而为访员臆答的结果，因此由于无法校正而被设置为“.n”。

缺失值的类型主要分为两类，其具体处理方法如下：

第一类：部分问题只在特定的逻辑条件下才需要回答，不需要回答导致的缺失，属于合理逻辑范围下的合理缺失，不需要进行插值；

第二类：根据逻辑跳转，部分题目需要回答，但是没有收集到有效信息，对于这类缺失主要通过逻辑插补和回归插补的方式补充缺失值。

11、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公开数据不披露受访者省级以下具体地址信息。用户可在CHFS公开数据基础上进行研究，但不得探索CHFS地市、区县、乡镇、村居的具体名称、国标码等相关信息。

中心的hh/ind/master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】的户口是在哪个省/市/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hh/ind数据集中该类变量的省市县地理信息与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。