China Family Panel Studies

CFPS 中国家庭动态跟踪调查

技术报告系列：CFPS-15

系列编辑：谢宇　责任编辑：李汪洋

中国家庭动态跟踪调查
2010 年收入消费支出数据整理

沈艳　雷晓燕

2012.12.20

本报告旨在构造和清理 CFPS 的收入、消费支出数据，为之后的分析提供更为准确的收入支出信息。本报告分为两部分，其中，第一部分简单介绍 CFPS 数据，并给出城乡与权重等相关定义；第二部分则详细介绍收入与支出指标的清理与构造，以及与年鉴等其他统计资料的比较。

1. CFPS 数据简介

中国家庭动态跟踪调查 （Chinese Family Panel Studies，CFPS）是两年一期的跟踪调查数据，旨在通过对全国代表性样本村居、家庭、家庭成员的跟踪调查，反映中国的经济发展与社会变迁。数据的搜集和发布由北京大学中国社会科学调查中心和美国密歇根大学调查研究中心等机构合作完成，在 2008 和 2009 两年分别做了三个城市的预调查和追踪调查之后，2010 年开展了基线调查。

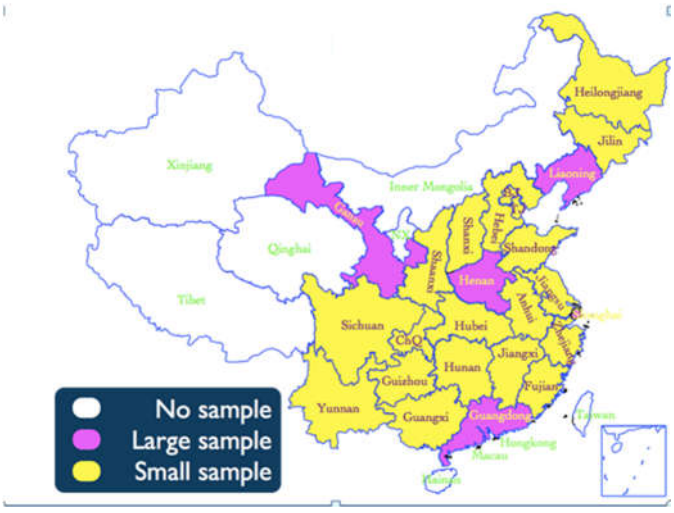

图 1. CFPS 样本分布

CFPS 借鉴美国 PSID 数据的结构，收集五个层面的数据：村居数据、家户数据、家户

成员数据、成人数据和儿童数据。数据调查覆盖除西藏、青海、新疆、宁夏、内蒙古、海南、香港、澳门、台湾的 25 省市，涵盖我国 95%的人口。具体的各省样本分布情况见上图 1。

整体数据含有六个子抽样框，其中五个用于省层面推断(广东、河南、甘肃、辽宁、上海)，另外二十省市采用一个样本框，用作全国推断，之后又合并形成二十五个省/市/自治区的全国样本框。具体的抽样设计方法采用分层、多阶段、与人口规模成比例的概率抽样方式（PPS），分三阶段抽取，其中第一阶段抽取 144 个行政性区（县）；第二阶段抽取 640 个行政性村（居）；第三阶段抽取 14,000 多个常住家庭户。在抽样设计上，常住家庭户采用国家统计局的界定，即家庭户户籍登记在、且家庭成员居住在样本村居；家庭户户籍登记在、且家庭成员离开样本村居未满 6 个月；家庭户户籍未登记在、但家庭成员居住在样本村居满 6 个月或以上。且对于所有村（居）样本，均采用地图地址法建构末端抽样框，所获得的抽样框为村（居）行政区划内排除了空址、商用地址后，有人居住的地址列表。

从调查单位层级来分，CFPS 2010 年基线调查的问卷包括三种，即村（居）问卷、家庭问卷和个人问卷。村居问卷的目的在于了解样本家庭所在的环境；家庭问卷的目的在于了解样本个体生活的家庭环境，包括家庭的社会关系网络，尤其是血缘和亲缘关系网络（家庭人口问卷）、生活设施、资产、社会经济活动、社会经济地位；个人问卷的目的在于了解样本个体的状况，包括个体从出生到有能力回答问题阶段的天赋和资质的以下方面：出生时的身体状况、成长的家庭环境、受教育的历史与现状、职业状况、经济状况、代际关系等等。目前 CFPS 2010 年基线调查的全国代表性样本数据来源于 25 个省市自治区、107 个行政性区（县）、424 个行政性村（居），共有 9500 个家户样本和 21760 个成人。

城乡差别是我们分析的重点，因此首先必须明确城乡划分的定义。比较常见的划分标准有三个，第一是按照国家统计局城镇、农村住户调查方案的定义，第二是按照居住地定义，第三是按照户口是农业户口或非农业户口进行定义。其中，国家统计局的城乡划分以我国的行政区划为基础，以民政部门确认的居民委员会和村民委员会辖区为划分对象，以实际建设[1]为划分依据。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[2]。乡村是指本规定划定的城镇以外的

[1] 实际建设是指已建成或在建的公共设施、居住设施和其它设施。
[2] 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。

区域。按照居住地定义的城乡划分则是通过家户所在地报告是村委会或居委会来加以划分，村委会被定义为农村户，居委会则被定义为城市户。第三种方法是根据家户户主户口状况定义，户主为农业户口的则记为农村户，反之则为城市户。

表 1. 不同城乡定义标准下家户数、人口数及其占比

	家户数	占比(%)	总人口数	总人口占比(%)
按统计局城镇、农村住户调查方案定义				
农村	4949	54.10	20250	56.50
城镇	4552	45.90	15877	43.50
总计	9501	100.00	36127	100.00
按居住地				
农村	6128	58.60	24889	62.10
城镇	3238	41.40	10725	37.90
总计	9366	100.00	35614	100.00
按户口				
农村	6075	62.10	24243	65.50
城镇	2592	37.90	8367	34.50
总计	8667	100.00	32610	100.00

表 1 报告的是按照三种方案分农村和城市的家户总数、人口总数及其分别的占比。可以看出，按居住地方案计算的农村户规模大于国家统计局方案的计算结果；由于户口类型的部分缺失，按户口计算的农村户数量小于按居住地计算的数量，但比重反而更高。在本文的后续讨论中，我们将选用国家统计局的城乡定义标准，以便于与国家统计局的数据结果进行相应的比对。

CFPS 一共包含了三类权重，分别是设计权重、无回答调整权重，以及在此基础上再基于国家统计局 2005 年数据年龄结构、性别结构、城乡结构调整的事后权重。上面的描述反映了事后权重调整的力度。本报告的描述和回归都采用事后权重，力图使得数据从年龄、性别和城乡等多个角度上都更具有全国代表性。

2. CFPS2010 收入、消费支出数据清理

在构造和清理收入和支出相关数据时，我们主要采用以下原则。第一，对于缺失数据的处理。我们的策略是尽可能使用所有的信息。例如，家户的工资性收入在家户层面有所询问，成人问卷也有涉及。在构造家户的工资性收入这一变量时，我们首先直接采用家户问卷中有关家户工资性收入的信息；如果一个家庭户缺失家户工资性收入，我们则进一步从成人问卷中搜寻相关信息来补齐家户工资性收入的缺失。尽管这一方法可能因为没有访问到所有成人而出现一些遗漏，但总比缺失全部工资性收入要好。只有在穷尽了不同问卷对于同一问题的所有问题后仍无法补齐所需信息时，我们才视该变量对于该家户为真正缺失。第二，对于非缺失数据的清理。这里我们的主要工作是判定变量的异常值，原则是尽可能少丢弃已有信息。比如，我们认为一个家庭户一年从政府所获得的所有转移性收入为个位数（在 10 元以内）的可能性很小，因此当家户政府转移性收入为个位数时，我们视该家户转移性收入值为缺省处理。为比较清理效果，在后面的讨论中，我们往往报告运用三组数据生成的统计资料，分别为没有进行任何处理的原始数据、去掉所有缺失值的数据，以及按照上述原则处理了异常值后的数据。

2.1 收入、消费支出指标的构造

家户的收入与消费支出信息主要在 CFPS2010 基线调查的家庭问卷中询问，同时成人问卷也涉及一些内容，如工资、医疗支出等。因此，在构造家户的收入与消费支出指标时，我们综合采用这两个层面的数据。为便于说明，我们用上标“#”表示该数据来自成人问卷，用上标“*”表示该数据来自家户问卷。

家户收入包含工资与经营性收入、财产性收入和转移性收入三大部分，具体构成如下：

工资与经营性收入	**= 工资性收入+非工资性及农业生产收入+经营企业收入**
工资性收入：	全家（含工资、奖金、补贴、分到个人名下的红利等）总收入*。若此项为“不知道”、“拒绝回答”、“不适用”或零值，而家户内所访到成人的工资性收入#之和大于 0，则用后者替代。
非工资性及农业生产收入：	将家户按是否从事农业生产分为两类，其中未从事农业生

产的家户取全家其他非工资性或农业生产收入*，从事农业生产的家户以农业生产的总收入*除去总支出*所得纯收入计入此单项[3]。

经营企业收入：　　参与经营的所有企业/公司去年的税后纯利润*与拥有的股份比例*之积

财产性收入　　= 出租房屋租金收入*+出租土地或其他生产资料租金收入*+出租其他东西租金收入*+存款利息收入*

转移性收入　　= 私人转移性收入+礼金/礼品+政府转移性收入+退休/养老金

私人转移性收入：　　家户内所访到成人从家人和亲友处得到经济帮助#的最大值与家户赡养支出*之差

礼金/礼品　　全家收到的礼金/礼品折合为现金收入*

政府转移性收入：　　全家从各级政府得到的补助（含实物和现金）*

退休/养老金：　　全家离/退休金/社会保障金/低保等收入*

需要补充说明的是，征地补偿、拆迁补偿、出卖财物收入、金融产品（股票、基金、债券）的市值变动均未计入家户收入。这里，征地补偿、拆迁补偿与出卖财物收入是与房屋相关的资产从实物形式变为现金形式而不视为收入；金融产品的市值变动本应计入资产性收入，但是由于问卷中关于本金部分的问题在本轮调查中并未收集，我们无法算出相应的市值变动。另外，问卷中对于“全家（含工资、奖金、补贴、分到个人名下的红利等）总收入”这一项除了问到具体数值外，同时问到了此项收入的大致区间，因此若按上述各项之和计算的家户收入为零，而此区间的下限大于零，则以区间的中值为此家户的收入。如家户加总收入为零，而“全家（含工资、奖金、补贴、分到个人名下的红利等）总收入”的自报区间为[2500, 5000]，则将家户收入调整为此区间的中值，即 3750 元[4]。

家户的消费支出包含日常支出（问卷中以月计的数值换算为以年计）和特殊支出（问卷中以年计）。我们将其中的相关项进行整合，一共分为食品支出、衣着支出、出行支出、通

[3] 因为原数据对农户没有区分非工资性收入和农业生产收入（问卷 F701 询问：“去年，您家其他非工资性收入或农业收入一共有多少元？”），所以将家户分为两类，对于农户我们通过农业生产收入与农业生产支出获得相应收入数据。

[4] 此步骤对 2 户家庭收入进行了调整。

信支出、文娱休闲支出、家庭日常/家电/服务支出、居住支出、医疗保健支出、教育支出和其他支出，共十大类，具体构成如下：

1. 食品支出*	家庭购买的各类食品
2. 衣着支出*	家庭在衣着服饰上的支出
3. 出行支出	用于日常交通的费用，如养车、加油/加气/加电、乘坐公共汽车交通的费用
4. 通信支出	用于如电话、手机、互联网接入、邮寄信件的费用
5. 文娱休闲支出*	家庭文化、娱乐、休闲支出
6. 家庭日常/家电/服务支出	= 购买日常用品支出*+购买家电支出*+家庭杂项商品、服务支出*
7. 居住支出	= 租房支出（不包括住房按揭）*+家庭居住支出（如物业、取暖等，不含住房按揭及房租) *
8. 医疗保健支出	若家户问卷中全家医疗保健支出*为“不知道”、“拒绝回答”、“不适用”或零值，而家户内所访到成人除去医保的个人医疗支出#之和大于 0，则用后者替代
9. 教育支出	若家户问卷中全家教育支出*为“不知道”、“拒绝回答”、“不适用”或零值，而家户内所访到成人的个人教育支出#之和大于 0，则用后者替代
10. 其他支出	= 自家婚丧嫁娶支出*+家庭其他支出*

因为家户问卷中的住房按揭、车辆按揭、其他按揭、购房建房支出、各项商业保险支出等各项不属于消费性支出，所以未计入家户消费支出。

2.2 收入的清理

对于家户收入数据，我们共进行了两个步骤的清理：步骤一，各单项收入中若有一项回答为“不知道”或“拒绝回答”，则将此家户剔除；步骤二，具体考查各单项收入，对于其

中的极端值，逐个仔细考察其它相关的信息，若确认所报数值不合常理，则将此家户剔除。

其中，各单项的回答情况，包括“不知道”或“拒绝回答”的家户数，见表 2、表 3。可以看出，私人转移性收入、礼品/礼金和工资性收入中“不知道”或“拒绝回答”的情形较多。

表 2. 单项收入回答情况（农村）

	不适用	拒绝回答	不知道	0 值	非缺失[5]	总户数
政府转移	4201	0	17	0	4933	4950
私人转移	3143	2	60	0	4888	4950
赡养费	0	0	17	4140	4933	4950
礼品/礼金	11	0	58	3076	4892	4950
退休/养老金	4144	0	9	0	4941	4950
企业纯收入	4826	0	29	13	4921	4950
所占股份	4862	0	6	7	4944	4950
非工资性与农业	1083	0	21	101	4929	4950
出租房屋	4877	0	0	0	4950	4950
出租土地/生产资料	4779	0	9	0	4941	4950
出租其他	4936	0	0	0	4950	4950
工资性收入	474	3	52	539	4895	4950

表 3. 单项收入回答情况（城镇）

	不适用	拒绝回答	不知道	0 值	非缺失[6]	总户数
政府转移	4098	0	19	0	4532	4551
私人转移	2996	3	48	0	4500	4551
赡养费	0	2	23	3592	4526	4551

[5] “非缺失”指该项为 0 或大于 0，以及“不适用”的情形。
[6] “非缺失”指该项为 0 或大于 0，以及“不适用”的情形。

表 3. 单项收入回答情况（城镇）（续）

	不适用	拒绝回答	不知道	0 值	非缺失[7]	总户数
礼品/礼金	0	4	68	2753	4479	4551
退休/养老金	3041	1	17	0	4533	4551
企业纯收入	4347	0	45	20	4506	4551
所占股份	4406	0	8	9	4543	4551
非工资性与农业	3155	1	10	37	4540	4551
出租房屋	4219	1	4	0	4546	4551
出租土地/生产资料	4414	0	2	0	4549	4551
出租其他	4546	0	0	0	4551	4551
工资性收入	108	4	27	447	4520	4551

表 4. 单项收入异常值判定标准

	农村	城镇
政府转移	单项总收入∈(0, 10)	单项总收入∈(0, 10)
私人转移	单项总收入最大值(80 万)[8]	无
赡养费	单项总支出≥12 万[9]	单项总支出≥12 万
礼品/礼金	无	无
退休/养老金	单项总收入∈(0, 10)[10]	单项总收入∈(0, 10)[11]
企业纯收入	大于千万的调整为万[12]	大于千万的调整为万
所占股份	1 调整为 100	1 调整为 100

[7] “非缺失”指该项为 0 或大于 0，以及“不适用”的情形。
[8] 此户除私人转移收入外，其余收入为千元。
[9] 在赡养费较高的几户中，我们发现他们的收入与赡养支出差距较大。
[10] 另外还有一个异常户，此户唯一家庭成员为一位老年农民，其成人问卷中被问及“未参加工作”的原因时并未选择“退休”，但退休金收入为农村最大值 80 万。
[11] 另外还有一个异常户，famindex 为 7404，其被采访到的成人无人退休，但有退休金收入 15 万。
[12] 企业/公司去年的税后纯利润采用了“元”与“万元”单位供访员选择，但根据我们对于单位为“万元”且纯利润大于一千元的家户的考查，认为这 19 户以及 1 个农村户（famindex 为 2826）为极小可能有大于千万元的企业利润收入，故将其单位调整为“元”。另外，“拥有的股份比例”采用的是百分制，我们将 2 户报告 1 的调整为 100。

表 4. 单项收入异常值的判定标准（续）

	农村	城镇
非工资性与农业	单项总收入∈(0, 10)	单项总收入∈(0, 10)或 单项总收入≤-60 万[13]
出租房屋	无	无
出租土地/生产资料	单项总收入∈(0, 10)	单项总收入∈(0, 10)
出租其他	无	无
工资性收入	单项总收入∈(0, 10)或 为最大值(1 千万)	单项总收入∈(0, 100)

根据以上标准，异常值的个数由表 5 所示，可以看到“政府转移收入”和“工资性收入”的异常值较多，其余单项异常值不足 10 户。

表 5. 单项收入异常值个数

	农村		城镇	
	较小异常值	较大异常值	较小异常值	较大异常值
政府转移	82	0	40	0
私人转移	0	1	0	0
赡养费	0	2	0	4
礼品/礼金	0	0	0	0
退休/养老金	4	1	2	1
企业纯收入	0	13	0	7
所占股份	2	0	0	0
非工资性与农业	3	0	7	1
出租房屋	0	0	0	0
出租土地/生产资料	13	0	1	0
出租其他	0	0	0	0
工资性收入	17	1	17	0

[13] 另外还有一个异常户，其该单项收入为 38 万，但并未从事农业生产。

总体样本量在经过了两步的清理后，样本量由 9501 户减少为 8849 户，农村与城镇样本的变化情况见表 6。

表 6. 收入清理的样本量变化

	原始数据	清理一	清理二
全国	9501	9026	8849
农村	4950	4711	4602
城镇	4551	4315	4247

2.3 消费支出的清理

对于家户消费支出的清理，我们分四步进行：首先，对于构成总消费的各单项进行考查，若有一项回答为“不知道”或“拒绝回答”，则将此家户标记为异常值；然后考查各单项支出的分布，将不合常理的最大值、最小值标记为异常值；接着，我们利用家户自报的总支出，从中减掉在构造家户消费支出时未用到的“购建房支出”、“购买商业保险支出”、“各类按揭支出”[14]，得到与我们定义一致的调整后自报支出，对于在前两步中标记为异常的家户，若调整后自报支出非缺失，则用其替代此家户消费支出，并取消对其异常值的标记；最后，我们将消费支出为零，且未从事农业生产的家户，以及加总的家户总消费与调整后自报支出差距过大的[15]标记为异常值。

其中，各单项的回答情况，以及“不知道”或“拒绝回答”的家户数情况，见表 7、表 8。可以看出，支出部分拒绝回答的情况很少，但食品支出、衣着支出、日常用品支出和出行支出中“不知道”的情形较多。

[14] 若自报支出，“购建房支出”、“购买商业保险支出”、“各类按揭支出”中有一项缺失，即“拒绝回答”或回答“不知道”，则将调整自报支出标记为缺失。
[15] 有 3 户，这两者比值小于 0.01。

表 7. 单项支出回答情况（农村）

	不适用	决绝回答	不知道	0 值	非缺失	总户数
食品支出	0	1	82	264	4867	4950
衣着支出	0	2	118	858	4830	4950
日常用品支出	0	1	97	523	4852	4950
家电支出	0	0	23	2569	4927	4950
家庭杂项商品、服务支出	0	0	39	3932	4911	4950
医疗保健支出	0	0	0	617	4950	4950
出行支出	0	0	74	1649	4876	4950
通信支出	0	0	46	529	4904	4950
租房支出	0	0	17	4749	4933	4950
家庭居住支出	0	0	27	2896	4923	4950
教育支出	0	0	0	2472	4950	4950
文娱休闲支出	0	0	29	4525	4921	4950
婚丧嫁娶支出	0	0	10	4294	4940	4950
其他支出	0	0	20	4155	4930	4950

表 8. 单项支出回答情况（城市）

	不适用	决绝回答	不知道	0 值	非缺失	总户数
食品支出	0	0	64	52	4487	4551
衣着支出	0	2	71	681	4478	4551
日常用品支出	0	0	96	308	4455	4551
家电支出	0	0	23	2532	4528	4551
家庭杂项商品、服务支出	0	0	29	3322	4522	4551
医疗保健支出	0	0	0	624	4551	4551
出行支出	0	1	58	1516	4492	4551
通信支出	0	0	38	216	4513	4551
租房支出	0	0	16	3989	4535	4551

表 8. 单项支出回答情况（城市）（续）

	不适用	决绝回答	不知道	0 值	非缺失	总户数
家庭居住支出	0	1	30	1508	4520	4551
教育支出	0	0	0	2076	4551	4551
文娱休闲支出	0	0	27	3437	4524	4551
婚丧嫁娶支出	0	0	16	3988	4535	4551
其他支出	0	0	18	3887	4533	4551

总体样本量在经过了四步的清理后，样本量由 9501 户减少为 9254 户，农村与城市样本的变化情况见表 9。

表 9. 支出清理的样本量变化

	原始数据	清理一	清理二	清理三、四
全国	9501	8898	8896	9362
农村	4950	4599	4599	4866
城市	4551	4299	4297	4496

2.4 CFPS2010 与统计年鉴的比对

在完成构造与清理的工作后，我们对 CFPS2010 基线调查收入和支出不同清理步骤之后两部分的交集数据和国家统计局的相关数据进行比对。这里我们共有两个样本，一个是原始样本，另一个是收入和支出都清理之后的样本。两个样本的样本量及城乡分布情况见下表。

表 10. 两个样本样本量及城乡分布表

	原始数据	清理后的数据
全国	9501	8768
农村	4950	4555
城市	4551	4213

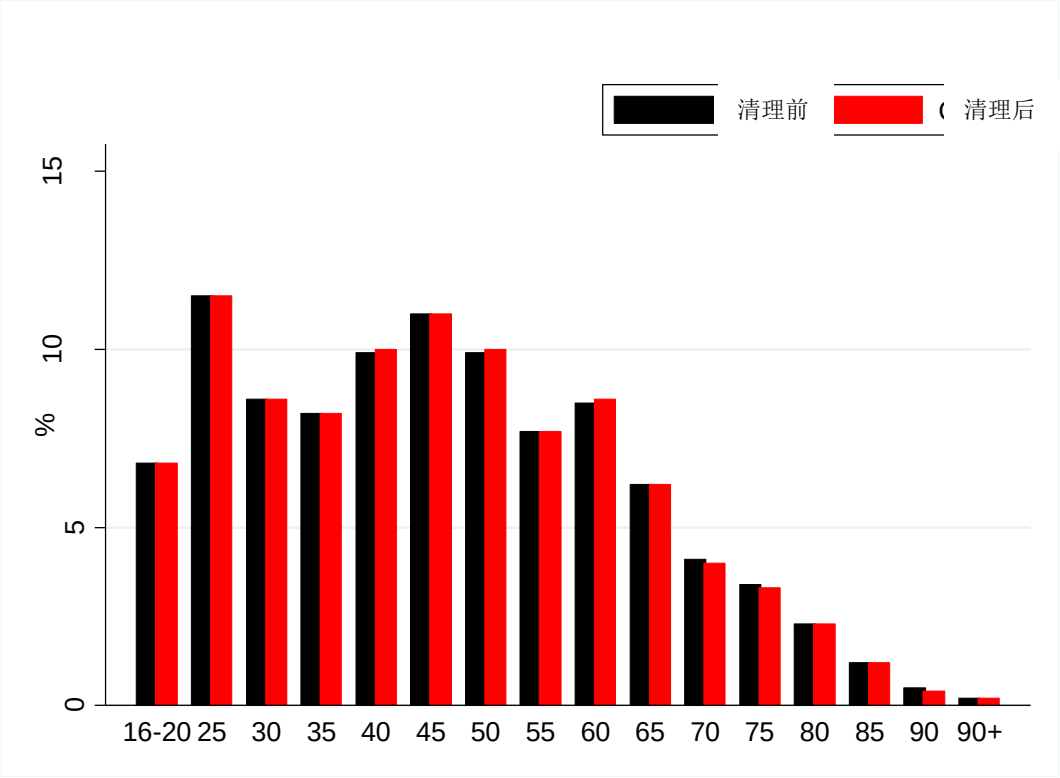

图 2. 清理前后成人年龄结构对比

基于成年人（16 岁及以上）的人口结构如图 2 所示，可以看出在清理前后成年人人口结构并未发生显著变化， [35,44]、[45,49]和[55,59]年龄段比例有稍许上升，70 岁及以上的人口比例稍有下降。同时，我们将清理前后全样本 0-14 岁人口占比、60 岁及以上人口占比与 2010 年人口普查公布的数据进行比对（见表 11）虽然少儿占比与老年人口占比略高，CFPS2010 年基线调查的人口结构与 2010 年人口普查的结果总体来说还是比较接近的。

表 11. CFPS2010 人口结构与 2010 年人口普查比对

	原始数据	清理后	2010 人口普查
% 0-14	17.0	17.0	16.6
% >= 60	14.6	14.3	13.3

接下来，我们分城乡报告这两个样本的收入和消费支出均值（见表 12）。可以发现，在清理后，从全国层面来看，收入和支出均呈上升趋势，但农村的人均收入略微下降。总体来看，清理本身带来的影响不大。

表 12. CFPS2010 收入和消费支出均值—分城乡

收入	原始数据	清理后的数据
全国	11628	12011
农村	6398	6421
城市	17802	18428
支出	原始数据	清理后的数据
全国	8979	9085
农村	5472	5582
城市	13120	13105

接下来，我们再将这两个样本的单项收入和支出均值与年鉴对比，并观察其收入和支出构成与年鉴的差异（表 13 和表 14）。

表 13. CFPS2010 收入及其分项与 2010 年年鉴的比对—均值水平

农村	原始数据	清理后的数据	2010 年鉴
人均收入	6398	6421	5153
转移性收入	1291	1315	398
工资、经营性收入	5049	5029	4588
财产性收入	53	72	167
城市	原始数据	清理后的数据	2010 年鉴
人均收入	17802	18428	18858
转移性收入	5158	5292	4515
工资、经营性收入	12444	12831	13911
财产性收入	199	299	432

表 14. CFPS2010 收入及其分项与 2010 年年鉴的比对—所占比例

农村	原始数据	清理后的数据	2010 年鉴
转移性收入	20.2	20.5	7.7
工资、经营性收入	78.9	78.3	89.0
财产性收入	0.8	1.1	3.2
城市	原始数据	清理后的数据	2010 年鉴
转移性收入	29.0	28.7	23.9
工资、经营性收入	69.9	69.6	73.8
财产性收入	1.1	1.6	2.3

从表 13 可以看出，对于农村户，清理后的收入均值与年鉴比较接近，但转移收入高于年鉴水平，财产性收入则又低于年鉴水平。在城市方面，清理后的数据人均收入有所上升，和年鉴的总收入数值比较接近[16]，但 CFPS2010 年基线调查的转移性收入高于年鉴，而工资和经营性收入、财产性收入则较低。再看收入构成，CFPS2010 年基线调查所得农村户的转移性收入占总收入达 21.5%，显著高于年鉴水平将近三倍，而城市户的收入构成则和年鉴比较类似，但 CFPS2010 年基线调查转移性收入和财产性收入比重较高，而工资及经营收入比重较低，这对于原始数据和清理后数据来说都是成立的。

再看这两个样本支出和年鉴的比对结果（见表 2.4.6）。可以发现，CFPS2010 年基线调查计算的农村户人均支出在原始数据和清理后的样本中都显著高于 2010 年年鉴水平，主要体现在医疗和家庭用品、设备、服务等方面较高的水平上，城市户的支出水平比 2010 年年鉴稍高，且原始样本和清理后数据相差不大，高出部分主要体现在医疗、家庭用品、设备、服务和其他方面较高的水平，但衣着支出则相对低于年鉴。

[16] 2010 统计年鉴中收入还包含了“股票等金融资产收入”，在 CFPS2010 基线调查问卷中虽然包含关于年底股票等金融资产的市值与本金的问题，但是调查数据中对本金一项的数据全部缺失，所以无法计算相应金融资产的收入。但是，根据家户对是否持有各类金融资产一问(F3)的回答，只有 6.4%的家户拥有金融资产。

表 15. CFPS2010 消费支出及其分项与 2010 年年鉴的比对—均值水平

农村	原始数据	清理后的数据	2010 年鉴
人均支出	5472	5582	3995
食品	1375	1418	1636
衣着	190	197	233
家庭用品/设备/服务	458	463	205
医疗	1096	1111	288
出行/通信	681	685	403
居住	277	294	805
教育/文娱休闲	601	617	341
其他	718	708	84
城市	原始数据	清理后的数据	2010 年鉴
人均支出	13120	13105	12264
食品	4141	4171	4479
衣着	597	609	1284
家庭用品/设备/服务	1489	1505	787
医疗	1475	1451	856
出行/通信	1575	1591	1683
居住	1263	1163	1228
教育/文娱休闲	1538	1550	1473
其他	1043	1088	474

表 16 展示了这两个样本分类细项比例与统计年鉴的比对结果。从表中我们可以看出，对于农村户来说，清理后的食品支出比例显著低于年鉴水平，而医疗支出则明显较高；对于城市户来说，CFPS2010 年基线调查支出结构与国家统计局类似，但衣着支出占比明显较低，而医疗、家庭用品、设备、服务及其他的占比则略高于年鉴水平。

表 16. CFPS2010 消费支出及其分项与 2010 年年鉴的比对—支出构成

农村	原始数据	清理后的数据	2010 年鉴
食品	25.1	25.4	41.0
衣着	3.5	3.5	5.8
家庭用品/设备/服务	8.4	8.3	5.1
医疗	20.0	19.9	7.2
出行/通信	12.4	12.3	10.1
居住	5.1	5.3	20.2
教育/文娱休闲	11.0	11.1	8.5
其他	13.1	12.7	2.1
城市	原始数据	清理后的数据	2010 年鉴
食品	31.6	31.8	36.5
衣着	4.5	4.6	10.5
家庭用品/设备/服务	11.3	11.5	6.4
医疗	11.2	11.1	7.0
出行/通信	12.0	12.1	13.7
居住	9.6	8.9	10.0
教育/文娱休闲	11.7	11.8	12.0
其他	8.0	8.3	3.9

2.5 对自产自消部分的回归调整

由于在家户问卷中，对于食品支出没有明显指出“包含自产自消”部分，即农户自己种植/养殖又自己消费的农产品/牲畜，因此回答者在报告收入和消费水平时极有可能忽略了这一部分，这将会造成对于从事农业生产活动家户的消费支出和收入水平的同时低估。为了处理这一问题，我们利用 CHIPs 2007 的农村样本对于从事农业生产活动家户的食品支出进行粗略调整。CHIPs 数据全称为“中国城镇居民家庭收入分配调查”，由中国社会科学院等机构合作收集。该数据包含城市、农村和移民三个样本。我们采用的是 CHIPs2007 年度的农村样本数据，村居变量来自村居问卷，住户变量来自住户问卷。

从 CFPS2010 数据中我们可以看到，在农村和城镇样本中，均有家户从事农业生产，但主要来自于农村户，占比为 77.2%，具体见表 17。

表 17. 从事农业生产家户的样本量

		农村	占比（%）	城镇	占比（%）	全国
是否从事	是	3823	77.2	1215	26.7	5038
农业生产[17]	否	1127	22.8	3336	73.3	4463
	总计	4950	100	4551	100	9501

CHIPs 2007 是我们目前可得的与 CFPS2010 年基线调查问卷最为接近的家户调查数据，但是两者在样本地理覆盖上有所不同：CHIPs 2007 是基于 11 省的数据，而 CFPS 涉及 25 省。在 CFPS 里从事农业生产家户的子样本中与 CHIPs 2007 相同的省份占此子样本的 53.9%，村居数占 54.8%， 所以这样调整的结果仅仅是近似，作为辅助参考两者具体省份的分布见表 18。

表 18. CHIPs 2007 与 CFPS2010 省份分布比较

	村居				家户			
	CHIPs 11 省		CFPS 25 省		CHIPs 11 省		CFPS 25 省	
	观测数	占比	观测数	占比	观测数	占比	观测数	占比
北京			1	0.3			1	0
天津			2	0.6			17	0.3
河北	50	6.3	31	9.4	497	6.3	515	10.2
山西			26	7.9			375	7.4
辽宁			1	0.3			20	0.4
吉林			9	2.7			129	2.6
黑龙江			11	3.3			119	2.4

[17] 有 2 户回答“不知道”，我们将其中的一个农村户替换为从事农业生产，城镇户替换为不从事。

表 18. CHIPs 2007 与 CFPS 省份分布比较（续）

	村居				家户			
	CHIPs 11 省		CFPS 25 省		CHIPs 11 省		CFPS 25 省	
	观测数	占比	观测数	占比	观测数	占比	观测数	占比
上海			2	0.6			5	0.1
江苏	100	12.6	9	2.7	999	12.6	133	2.6
浙江	99	12.5	9	2.7	979	12.4	76	1.5
安徽	88	11.1	12	3.6	871	11	200	4
福建			7	2.1			111	2.2
江西			11	3.3			168	3.3
山东			27	8.2			478	9.5
河南	100	12.6	31	9.4	995	12.6	459	9.1
湖北	89	11.2	5	1.5	887	11.2	90	1.8
湖南	10	1.3	15	4.6	99	1.3	182	3.6
广东	83	10.4	23	7	828	10.5	228	4.5
广西	17	2.1	12	3.6	165	2.1	212	4.2
重庆	50	6.3	5	1.5	500	6.3	87	1.7
四川	109	13.7	29	8.8	1078	13.6	538	10.7
贵州			17	5.2			321	6.4
云南			16	4.9			290	5.8
陕西			11	3.3			179	3.6
甘肃			7	2.1			105	2.1
总　计	795	100	329	100	7898	100	5038	100

我们将两组数据共有的变量进行了对比，可以看出，在村居层面上，CHIPs 的村居人均收入更高，村居人口较少，从事农业生产的人口比例更高；在家户人口结构上，CHIPs 的家庭规模稍大，少儿占比高，老年人口占比低；户主的教育水平显著高于 CFPS；对于人均收入、消费支出和食品支出，由于自产自消部分的影响，CHIPs 普遍高于 CFPS 水平。

表 19. CHIPs 2007 与 CFPS2010 村居变量比较

	CHIPs			CFPS		
	均值	村居数	家户数	均值	村居数	家户数
村居平均人均收入[18]	10.8	794	7888	7.6	322	4966
村居人口	2478	795	7898	3440	322	4966
村居农业人口(%)	50.7	795	7898	38.8	316	4862

表 20. CHIPs 2007 与 CFPS2010 从事农业生产家户变量比较

	CHIPs		CFPS			
			原始数据		清理后	
	均值	户数	均值	户数	均值	户数
家户						
总人口	3.98	7898	3.550	5038	3.550	4645
<=16 %	0.17	7898	0.149	5038	0.151	4645
>=60 %	0.12	7898	0.201	5038	0.195	4645
户主						
年龄	49.3	7898	50.30	5038	50.20	4645
男性	0.96	7898	0.809	5038	0.809	4645
教育年限	7.47	7644	4.920	4616	4.950	4281
文盲	3.1	7887	32	5035	29.50	4643
小学	29.8	7887	28.80	5035	26.30	4643
初中	49.7	7887	30.20	5035	28.30	4643
高中	16.1	7887	8	5035	7.300	4643
大专及以上	1.3	7887	1	5035	0.800	4643

[18] “村居人均收入”在 CHIPs 问卷中是一个范围变量，10 对应的收入范围为 3000-3500 元，11 对应的为 3500-4000 元；7 和 8 分别对应 1800-2000 元和 2000-2500 元。

表 20. CHIPs 2007 与 CFPS2010 从事农业生产家户变量比较（续）

	CHIPs		CFPS			
			原始数据		清理后	
	均值	户数	均值	户数	均值	户数
人均收入与消费						
收入	6822	7898	5861	5038	6004	4645
消费	5808	7898	5038	5038	5103	4645
食品支出	1751	7898	1223	5038	1251	4645
sEngel[19]	0.301		0.243	.	0.245	.
Engel[20]	0.372		0.300	.	0.304	.

我们利用 CHIPs 样本对于食品消费的预测模型进行了估计，最终采用的模型[21]是基于以上村居及家户层面变量对家户的 Engel 系数进行预测，由于食品支出与 Engel 系数是单调关系，因此可以由 Engel 系数还原家户的食品支出（此时包括自产自消部分）。在回归方程中，我们去掉了不显著的变量：户主性别、户主的教育程度、村居人口数，同时控制了省份虚拟变量。前三组回归与后三组回归采用了不同的回归方程：后三组中加入了人均消费对数值的平方，而前三组没有。同时，为了检验回归系数的稳健性，我们分别取了全样本、去掉上下 5%和去掉上下 10%的样本进行回归。可以看出，系数的符号与预期一致，在去掉可能的极端值后表现稳健。在加入人均收入对数平方后，回归方程的 R 方变大，其余变量保持显著性。

[19] 均值的比。
[20] 比的均值。
[21] 我们同时也对直接基于食品支出的预测模型进行了回归，但 Engel 系数的估计模型表现更好。

表 21. Engel 系数的预测模型估计

	(1)	(2)	(3)	(4)	(5)	(6)
	全样本	[.05, .95]	[.10, .90]	全样本	[.10, .90]	[.05, .95]
人均收入对数	-0.091***	-0.066***	-0.047***	0.132***	0.100***	0.081***
对数平方				-0.014***	-0.010***	-0.008***
户总人口	-0.016***	-0.011***	-0.008***	-0.019***	-0.012***	-0.010***
<=16 %	0.025**	0.019**	0.023***	0.018*	0.014	0.018**
>=60 %	0.012*	0.009	0.011*			
户主年龄	0.001***	0.001***	0.001***	0.001***	0.001***	0.001***
村农业人口%	-0.001***	-0.000***	-0.000***	-0.001***	-0.000***	-0.000***
村人均收入范围						
vill_inc==2	-0.058***	-0.040**	-0.011	-0.057***	-0.039**	-0.011
vill_inc==3	-0.078***	-0.052***	-0.029	-0.078***	-0.052***	-0.029*
vill_inc==4	-0.024	-0.012	0.002	-0.020	-0.008	0.005
vill_inc==5	-0.078***	-0.056***	-0.026	-0.075***	-0.053***	-0.024
vill_inc==6	-0.014	-0.009	0.005	-0.012	-0.007	0.007
vill_inc==7	-0.077***	-0.047***	-0.022	-0.075***	-0.047***	-0.021
vill_inc==8	-0.083***	-0.059***	-0.032**	-0.080***	-0.058***	-0.031**
vill_inc==9	-0.091***	-0.067***	-0.037**	-0.089***	-0.066***	-0.036**
vill_inc==10	-0.077***	-0.057***	-0.032**	-0.076***	-0.056***	-0.032**
vill_inc==11	-0.075***	-0.050***	-0.024*	-0.073***	-0.049***	-0.023
vill_inc==12	-0.078***	-0.061***	-0.035**	-0.074***	-0.059***	-0.033**
vill_inc==13	-0.061***	-0.047***	-0.024	-0.056***	-0.043***	-0.021
vill_inc==14	-0.050***	-0.037**	-0.020	-0.042**	-0.032*	-0.016
vill_inc==15	-0.064***	-0.036**	-0.017	-0.053***	-0.030*	-0.011
vill_inc==16	-0.043**	-0.039**	-0.018	-0.034*	-0.032*	-0.012
vill_inc==17	0.003	-0.007	0.007	0.023	0.007	0.019
vill_inc==18	0.064*	0.040	0.029	0.068**	0.046	0.035
Observations	7,888	7,102	6,312	7,888	7,102	6,312
R-squared	0.281	0.201	0.152	0.308	0.215	0.164

我们分别选取第一、四两组回归模型，对 CFPS2010 年基线调查从事农业生产户的 Engel 系数进行了预测，并将预测出的食品支出代替之前的食品支出，将自产自消部分同时计入家户收入。注意到，由于预测方程无法保证预测值不小于原食品支出，即使在 CHIPs 数据中，我们将预测食品支出与原值进行比较，也有近半数小于原值。当然，这与回归结果并不矛盾，因为线性回归在模型正确的情形下也只满足对于预测误差的均值为 0，所以总会出现误差为正或为负的情形。因此我们对于预测值小于原食品支出的家户没有进行调整，保持其原有的食品支出值。具体来说，我们分别对 CFPS 原始数据和清理后的数据进行了调整，基于预测模型一调整后的结果见表 23，基于预测模型四调整后的结果见表 224。在表 22 中，我们对预测的自产自消食品支出进行了描述，可以看到，模型四明显优于模型一。一方面，模型四的预测值大于原始值的比例大于模型一；另一方面，模型一预测值的最大值、最小值都更为极端。

表 22. CFPS2010 自产自消食品支出的预测值

	总户数	调整大于之前	均值	中值	最小值	最大值
模型一						
原始数据	5038	0.662	714	468	-470000	290000
清理后	4645	0.658	768	459	-470000	290000
模型四						
原始数据	5038	0.674	1102	473	-16381	53516
清理后	4645	0.669	1090	464	-16381	53516

经过模型一的调整，对于 CHIPs 数据本身，51.6%的家户预测支出不低于原始支出，对于 CFPS 数据这一比例在 68%左右（表 23）。农村的食品支出比原始值高出 1000 元左右，调整后的人均支出水平高于 CHIPs 2007 水平，也高于统计局 2010 年年鉴水平；城镇的食品支出上调了 200 余元，调整后的人均收入仍然在年鉴水平的可支配收入与收入（17175 与 18858 元）之间；农村和城镇的 Engel 系数大致为 0.377，高于 CHIPs 水平，但低于统计局水平。

表 23. 基于模型一对 CFPS2010 食品支出的调整

	2010 年鉴	CHIPs	CFPS 原始		CFPS 清理后	
调整后≥调整前		0.516		0.662	.	0.658
农村（人均）			.	.		.
消费支出	3993	5808	5576	6670	5582	6678
食品支出	1636	1751	1375	2469	1418	2515
Engel	0.409	0.301	0.247	0.370	0.254	0.377
城镇（人均）			.	.	.	.
消费支出	12265	.	13042	13296	13105	13362
食品支出	4479	.	4141	4395	4171	4429
Engel	0.365	.	0.317	0.331	0.318	0.331
人均收入		.	.	.	.	.
全国		.	11697	12406	12011	12717
农村	5153	6822	6421	7514	6421	7518
城镇	17175	.	17926	18181	18428	18685

模型四的调整，对于 CHIPs 数据本身，同样有近 51.5%家户预测值不低于原始值，而 CFPS2010 年基线调查的这一比例与模型一类似（见表 24）。同时，对于农村和城镇食品支出的调整幅度也与模型一相差不大；农村和城镇的 Engel 系数调整后水平略低于模型一，分别为 0.373 和 0.326。

表 24. 基于模型四对 CFPS2010 食品支出的调整

	2010 年鉴	CHIPs	CFPS 原始		CFPS 清理后	
			调整前	调整后	调整前	调整后
调整后≥调整前		0.515		0.674	.	0.669
农村（人均）			.	.		.
消费支出	3993	5808	5576	6561	5582	6646
食品支出	1636	1751	1375	2436	1418	2482
Engel	0.409	0.301	0.247	0.371	0.254	0.373

表 24. 基于模型四对 CFPS2010 食品支出的调整（续）

	2010 年鉴	CHIPs	CFPS 原始		CFPS 清理后	
			调整前	调整后	调整前	调整后
城镇（人均）			.		.	.
消费支出	12265	.	13042	13191	13105	13252
食品支出	4479	.	4141	4290	4171	4318
Engel	0.365	.	0.317	0.325	0.318	0.326
人均收入		.	.	.	.	.
全国		.	11697	12340	12011	12648
农村	5153	6822	6421	7481	6421	7485
城镇	17175	.	17926	18076	18428	18574

总的说来，根据预测模型的调整，农村户的消费支出和收入的调整较大，调整后的结果显著高于 2010 年年鉴水平，也高于 CHIPs 水平；城镇户的调整不明显，且调整后收入与消费支出的水平与年鉴较为一致。Engel 系数在农村和城镇均较原始值有一定上升，且高于 CHIPs 水平，但仍低于年鉴水平。

两模型的预测结果差别不大，模型一的 Engel 系数和食品支出的调整幅度略大，调整值不小于原始值的比例比模型四略低。综合来看，模型四的估计结果是最好的。由于我们没有很好的标准对于预测的异常值进行进一步的清理，所以我们仅列出对于自产自消的调整结果，有待在新的一期数据发布后进一步的完善。