

China Family Panel Studies



中国家庭追踪调查

技术报告系列: CFPS-44

系列编辑: 谢宇 责任编辑: 赵逸文

中国家庭追踪调查 2020 年权数调整报告

吕萍 谢宇 吴琼

北京大学中国社会科学调查中心

2023. 06

目录

一、CFPS2020 横截面权数.....	2
1.1 个人横截面权数	3
1.1.1 个人横截面设计权数	3
1.1.2 个人横截面无回答权数	3
1.1.3 个人横截面事后分层调整权数	4
1.1.4 个人横截面极值调整权数	4
1.1.5 个人横截面标准化权数	5
1.1.6 个人横截面权数的描述性分析	5
1.2 家庭横截面权数	8
1.2.1 家庭横截面基础设计权数	8
1.2.2 家庭横截面无回答权数	9
1.2.3 家庭横截面事后分层权数	9
1.2.4 家庭横截面极值权数	9
1.2.5 标准化权数	9
二、CFPS2020 追踪权数.....	10
2.1 个人追踪设计权数	10
2.2 个人追踪无回答权数	10
2.3 个人横截面事后分层调整权数	11
2.4 个人追踪极值调整权数	11
2.5 个人追踪标准化权数	11
2.6 CFPS2020 个人追踪样本的权数分布	12
三、CFPS2020 的权数使用.....	15
相关阅读	15

中国家庭追踪调查于 2010 年进行基线调查，之后每两年实施一轮追踪调查，截至 2020 年年底，共完成六轮全国调查。CFPS2020 的权数调整方法与 CFPS2018 年类似，即按照追踪规则，对 CFPS2020 年的基因成员进行权数调整，包含横截面权数和追踪权数。横截面权数针对横截面样本，即 2010 年的个人或后续轮次新采集到的新生血缘/领养子女以及所在的 家庭，包含家庭层面横截面权数和个人层面横截面权数。追踪权数针对追踪总体，即 2010 年的完访家庭中完成个人问卷（包括成人问卷和少儿问卷）且在 2020 年完成个人问卷和少儿问卷的个人，追踪权数只在个人层面存在。我们接下来分别介绍 CFPS2020 的横截面权数和追踪权数的调整过程。

一、CFPS2020 横截面权数

CFPS2020 横截面权数对应的**目标总体**是 2020 年全国家庭（家庭层面横截面权数）和 2020 年全国个人（个人层面横截面权数），适用于估计目标总体的相关研究，如估计 2020 年全国家庭的收入支出分布、2020 年全国个人的婚姻、教育等¹。**样本总体**是 2020 年的所有基因成员（2010 年个人以及各年新生的血缘/领养子女）及其所在家庭，分别对应个人横截面权数和家庭横截面权数。

2010 年基线调查时，我们采用多阶段不等概率的抽样方法，得到家庭层面和个人层面的具有近似全国代表性的样本。但随着时间的变迁，个人和家庭层面都产生了变化。在个人层面，新基因成员产生，已有基因成员会死亡；在家庭层面，新家庭因为婚姻、分家等原因增加，原有家庭因为基因成员死亡等消失。横截面样本总体在变动，CFPS 的调查设计使得我们的样本能基本反映中国人口与家庭的自然更替与变化，但需要对设计加权进行调整。同样，在调查过程中存在样本流失、拒绝访问、无回答、无联系等，因此需要进行无回答调整。其次，随着时间的变迁，调查样本中与研究目标变量相关的年龄、性别、城乡等变量的结构会发生变化，导致与全国的结构分布不同，因此需要进行结构调整。再次，CFPS2010 的复杂抽样设计，以及权数的多次调整，导致权数的变异增加，为了提高估计精度，需要对权数进行极值调整，进而得到极值调整权数。最后，需要对权数进行标准化转换，进而得到最终

¹ CFPS2010 年基线调查的对象为中国（不含香港、澳门、台湾以及新疆维吾尔自治区、西藏自治区、青海省、内蒙古自治区、宁夏回族自治区、海南省）25 个省/市/自治区的满足项目访问条件的家户和样本家户中满足项目访问条件的家庭成员。CFPS 样本所在的 25 个省/市/自治区的人口覆盖了中国除香港、澳门、台湾外总人口数的 94.5%，由于覆盖范围如此广泛，因此可以将 CFPS 样本视为一个全国代表性样本。相关信息可以参考 CFPS 技术报告系列：CFPS-01 中国家庭追踪调查抽样设计（CFPS 项目网站“文档中心”——>“技术报告”）。

权数。因此，对 CFPS 的权数需要进行设计权数调整、无回答权数调整、事后分层调整和权数的极值调整。

由上，对 CFPS2020 的横截面家庭样本和个人样本，进行加权调整，最终得到家庭横截面权数和个人横截面权数。

1.1 个人横截面权数

CFPS2020 的个人横截面权数的构建过程依次包含个人横截面设计权数、个人横截面无回答权数、个人横截面事后分层权数、个人横截面极值权数和个人横截面标准化权数。下面我们具体介绍各个步骤的调整过程。

1.1.1 个人横截面设计权数

CFPS 的个人层面的横截面样本总体来源于 CFPS2010 年的完访家庭中的基因成员（以下简称“CFPS2010 基因”）、CFPS2012-CFPS2020 的新进基因成员，不包含 CFPS2010-CFPS2020 年的死亡基因。其中 CFPS2010 基因成员的权数是 CFPS2010 基因成员的家庭无回答权数（以下简称“CFPS2010 基因权数”）；新进基因成员的权数来源于其同样为基因成员的父母的 CFPS2010 基因权数的均值，若仅有母亲（或父亲）是基因成员，则新进基因成员的权数为母亲（或父亲）的 CFPS2010 基因权数。死亡基因的数量在样本中占比较小，由于已经死亡，权数是零。由如上步骤所得到的权数是 CFPS2020 个人层面的横截面设计权数。

1.1.2 个人横截面无回答权数

由于在调查过程中存在样本流失，为了提高样本的代表性，我们需要进行权数调整，计算流失权数。本次调查使用家庭、个体层面的辅助信息，采用基于 logistic 模型的倾向权数计算方法计算流失权数。首先将个人样本分为完访样本和未完访样本，建立 logistic 回归模型，本次使用的变量是性别、家庭规模、是否有 65 岁及以上的老人、是否有 10 岁以下的儿童、城乡、2010 年是否完成个人问卷访问、婚姻状态、最高学历、是否居住在家、年龄，其中由于年龄变量的特殊性，在建模过程中年龄作为样条函数，计算倾向流失的权数调整系数是：

$$p_1 = 1 / (\exp(\beta_1 X) / (1 + \exp(\beta_1 X)))$$

由上得到，个人横截面无回答权数等于 CFPS2020 个人横截面设计权数与倾向流失的权

数调整系数的倒数的乘积。

1.1.3 个人横截面事后分层调整权数

随着时代的变迁，研究总体、样本总体不断变化，而样本个人的出生、死亡，以及调查过程中的无回答、拒绝访问、无法联系等，导致样本总体与研究总体的结构发生变化。其中，与研究目标变量相关的年龄、性别、城乡等变量的结构会发生变化，导致与全国的结构分布不同，因此需要对最终的样本总体进行结构调整。由于 2020 年是中国第七次人口普查（以下简称“七普”）年，全国的分城乡年龄性别的总量已知且可获得。因此，我们可以使用七普的分城乡年龄性别数据对 CFPS2020 样本进行完全事后分层调整，即在六个抽样框中用城乡（分为城镇和乡村）、性别（分为男和女）、年龄变量（0~9、10~19、20~29、30~39、40~49、50~59、60 岁及以上）进行完全事后分层调整，得到个人横截面事后分层调整权数系数，即各个总体的事后分层调整的系数是：

$$P_{\text{postmj}} = \frac{W_{mj}}{N_{mj}}$$

其中 W_{mj} 、 N_{mj} 分别表示第 m 个总体的第 j 个总体的权数和和七普总量。

由此得到个人横截面事后分层调整权数，即 CFPS2020 个人横截面设计权数、倾向流失的权数调整系数的倒数与个人横截面事后分层调整权数系数的乘积。

1.1.4 个人横截面极值调整权数

上述权数调整保证了样本单元加权与总体大致保持一致，然而权数调整过程中会导致权数波动变大，使得估计量的方差也随之变大，影响估计的效率，降低了目标变量的估计精度。为了使得权数波动在一个可控范围，可以使用权数截取，即对权数进行极值调整。

下面使用基于权数分布的分位数的权数截取方法，即选取权数分布的 a 分位数，将样本权数小于等于 a 均赋值为 a 分位点权数值，将样本单元最终权数大于等于 $1-a$ 均赋值为 $1-a$ 分位点权数值，由此得到极值权数，进而降低权数的变异性，提高估计精度，但是过大的权数截取，虽然会降低权数的变异性，但是也会增加偏差，进而增加均方误差。此处，我们首先使用权效应判断是否需要权数的极值调整。即，权效应越小，说明权数的变异性越小，最终的加权估计量的估计精度较高，反之则权数的变异性大，权数的质量较差，甚至不宜使用。由此我们计算五个“大省”和 1 个“小省”的权效应，进而发现仅上海市需要进行权数的

极值调整。由于极值权数会影响样本的分布，为了尽可能保持每个子总体的分布，此处根据最终权数的分布和权数变量，做微小的调整，即通过对最终的事后分层调整权数分布的分析，选取相应的分位数作为最小最大值，即满足

$$W'_{post} = \begin{cases} W_{\alpha}, W_{post} < W_{\alpha} \\ W_{1-\alpha}, W_{post} > W_{1-\alpha} \end{cases}$$

由此，得到五个大省和一个小省的个人横截面极值权数系数。

1.1.5 个人横截面标准化权数

由上，个人横截面设计权数、无回答调整系数、事后分层权数调整系数和极值权数调整系数的乘积，即为最终的个人横截面权数。

由于可以获得 2020 年七普人口数据，因此 CFPS2020 年可以获得上述个人横截面事后分层权数，而 CFPS2018 年的权数计算中由于没有普查数据，没有进行事后分层调整，因此为了与 CFPS2018 年的权数可比，此处获得个人横截面无回答调整权数，即个人横截面无回答权数是个人横截面设计权数、无回答调整系数、和极值权数调整系数的乘积。

综上，CFPS2020 的个人权数包含个人横截面无回答权数和个人横截面事后分层权数。由于 CFPS 抽样设计和追踪规则的复杂性，CFPS 借鉴国内调查（中国综合社会调查 CGSS 等）和国际调查（美国收入动态面板调查 PSID）的权数调整，在 CFPS2020 中给出标准化权数。下面，分别对上述每个子总体的个人横截面无回答权数、个人横截面事后分层权数和个人横截面极值权数进行标准化，即 CFPS2020 个人横截面标准化权数=子总体的 CFPS2020 个人权数/子总体 CFPS2020 个人权数的均值

由上，得到 CFPS2020 个人横截面标准化权数即最终的个人横截面无回答权数（发布数据集中的变量 `rswt_natcs20n`）和个人横截面事后分层权数（发布数据集中的变量 `rswtps_natcs20n`）。

1.1.6 个人横截面权数的描述性分析

下面利用获得的个人横截面无回答权数和个人横截面事后分层权数分析样本的年龄、性别、城乡加权分布与七普的年龄结构分布，见表 1 至表 4。其中 CFPS2020 个人样本来自于个人库（适用于 10 岁及以上个人）及少儿库（适用于 0-15 岁儿童）中不重复的样本。七普的比例来自于基线调查时涵盖的 25 省的七普数据。

表 1 CFPS2020 样本年龄加权分布与七普分布对比表

	样本比例	无回答权数比例	事后分层权数比例	七普比例 ¹
城市 0-9 男	4.37	4.41	3.97	3.97
城市 0-9 女	3.63	3.7	3.52	3.52
城市 10-19 男	2.98	3.02	3.9	3.9
城市 10-19 女	2.53	2.54	3.42	3.42
城市 20-29 男	3.28	3.93	4.33	4.33
城市 20-29 女	3.03	3.75	4.04	4.04
城市 30-39 男	3.7	3.78	5.75	5.75
城市 30-39 女	3.53	3.61	5.64	5.64
城市 40-49 男	3.43	4.21	4.95	4.95
城市 40-49 女	3.48	4.25	4.84	4.84
城市 50-59 男	4.14	4.59	4.71	4.71
城市 50-59 女	4.39	4.62	4.71	4.71
城市 60-男	5.53	5.82	4.87	4.87
城市 60-女	5.82	6.79	5.36	5.36
乡村 0-9 男	3.54	2.34	2.32	2.32
乡村 0-9 女	3.36	2.19	2.06	2.06
乡村 10-19 男	3.44	2.81	2.1	2.1
乡村 10-19 女	2.96	2.47	1.76	1.76
乡村 20-29 男	2.51	2.6	1.86	1.86
乡村 20-29 女	2.01	2.2	1.53	1.53
乡村 30-39 男	3.06	2.63	2.35	2.35
乡村 30-39 女	2.66	2.36	2.03	2.03
乡村 40-49 男	2.96	2.65	2.51	2.51
乡村 40-49 女	2.76	2.65	2.32	2.32
乡村 50-59 男	3.52	3.05	3.24	3.24
乡村 50-59 女	3.55	3.2	3.16	3.16
乡村 60-男	4.87	4.68	4.28	4.28
乡村 60-女	4.94	5.14	4.46	4.46

注：数据来源于 CFPS 基线调查时所涵盖的 25 省市。

表 2 CFPS2020 样本年龄加权分布与七普分布对比表

抽样子总体	分类	样本比例	无回答权数比例	事后分层权数比例	七普比例
上海市	0-9 岁	7.99	9.85	7.12	7.12
上海市	10-19 岁	7.23	5.59	5.53	5.53
上海市	20-29 岁	6.46	6.07	14.96	14.96
上海市	30-39 岁	11.45	9.77	20.25	20.25
上海市	40-49 岁	13.53	12.94	14.61	14.61
上海市	50-59 岁	13.37	13.55	14.15	14.15
上海市	60-岁	39.97	42.23	23.38	23.38
辽宁省	0-9 岁	9.24	8.63	7.27	7.27
辽宁省	10-19 岁	8.43	6.69	7.81	7.81

辽宁省	20-29 岁	9.42	11.5	9.59	9.59
辽宁省	30-39 岁	12.04	10.24	15.02	15.02
辽宁省	40-49 岁	13.48	15.27	15.74	15.74
辽宁省	50-59 岁	18.71	17.5	18.85	18.85
辽宁省	60-岁	28.67	30.16	25.72	25.72
河南省	0-9 岁	17.21	11.87	14.67	14.67
河南省	10-19 岁	13.98	17.8	14.83	14.83
河南省	20-29 岁	10.58	9.17	10.43	10.43
河南省	30-39 岁	13.23	14.51	14.41	14.41
河南省	40-49 岁	11.39	10.95	12.82	12.82
河南省	50-59 岁	14.39	16.73	14.77	14.77
河南省	60-岁	19.22	18.97	18.08	18.08
甘肃省	0-9 岁	15.9	10.77	13.16	13.16
甘肃省	10-19 岁	11.98	9.68	11.53	11.53
甘肃省	20-29 岁	12.31	13.92	12.05	12.05
甘肃省	30-39 岁	13.77	13.93	14.13	14.13
甘肃省	40-49 岁	12.89	15.55	14.71	14.71
甘肃省	50-59 岁	16.58	17.85	17.39	17.39
甘肃省	60-岁	16.58	18.3	17.03	17.03
广东省	0-9 岁	19.92	22.01	13.14	13.14
广东省	10-19 岁	13.12	9.53	10.74	10.74
广东省	20-29 岁	13.6	13.77	16.9	16.9
广东省	30-39 岁	13.5	11.69	19.25	19.25
广东省	40-49 岁	9.57	9.8	14.95	14.95
广东省	50-59 岁	13.74	13.23	12.66	12.66
广东省	60-岁	16.54	19.97	12.35	12.35
25 省 ¹	0-9 岁	14.9	12.64	11.87	11.87
25 省	10-19 岁	11.92	10.85	11.18	11.18
25 省	20-29 岁	10.83	12.48	11.77	11.77
25 省	30-39 岁	12.95	12.38	15.77	15.77
25 省	40-49 岁	12.63	13.76	14.62	14.62
25 省	50-59 岁	15.61	15.46	15.82	15.82
25 省	60-岁	21.17	22.43	18.97	18.97

注：此处 25 省指的是 CFPS 基线调查时所涵盖的 25 省市

表 3 CFPS2020 样本性别加权分布与七普分布对比表

总体	分类	样本比例	无回答权数比例	事后分层权数比例	七普比例
上海市	男性	50.12	47.95	51.77	51.77
上海市	女性	49.88	52.05	48.23	48.23
辽宁省	男性	48.15	50.59	49.92	49.92
辽宁省	女性	51.85	49.41	50.08	50.08
河南省	男性	51.5	52.27	50.15	50.15
河南省	女性	48.5	47.73	49.85	49.85

甘肃省	男性	52.97	51.04	50.76	50.76
甘肃省	女性	47.03	48.96	49.24	49.24
广东省	男性	52.22	51.72	53.07	53.07
广东省	女性	47.78	48.28	46.93	46.93
25 省 ¹	男性	51.34	50.53	51.15	51.15
25 省	女性	48.66	49.47	48.85	48.85

注：此处 25 省指的是 CFPS 基线调查时所涵盖的 25 省市

表 4 CFPS2020 样本城乡加权分布与七普分布对比表

总体	分类	样本比例	无回答权数比例	事后分层权数比例	七普比例
上海市	城市	90.08	89.14	89.3	89.3
上海市	乡村	9.92	10.86	10.7	10.7
辽宁省	城市	56.4	66.07	72.14	72.14
辽宁省	乡村	43.6	33.93	27.86	27.86
河南省	城市	50.64	51.7	55.43	55.43
河南省	乡村	49.36	48.3	44.57	44.57
甘肃省	城市	38.79	46.97	52.23	52.23
甘肃省	乡村	61.21	53.03	47.77	47.77
广东省	城市	60.12	66.78	74.15	74.15
广东省	乡村	39.88	33.22	25.85	25.85
25 省 ¹	城市	53.84	59.01	64.03	64.03
25 省	乡村	46.16	40.99	35.97	35.97

注：此处 25 省指的是 CFPS 基线调查时所涵盖的 25 省市

由表 1 至表 4 可知，经过加权调整，尤其是事后分层调整后，样本的年龄、性别、城乡分布与七普更为接近，估计精度更高。

1.2 家庭横截面权数

CFPS2020 的家庭横截面权数基于个人横截面的基础设计权数，在此基础上包含家庭横截面设计权数、家庭横截面无回答权数、家庭横截面事后分层权数、家庭横截面极值权数和家庭横截面标准化权数。下面介绍家庭横截面权数的调整过程。

1.2.1 家庭横截面基础设计权数

由于 CFPS 追踪的是基因成员，因此家庭横截面基础设计权数为 CFPS2020 的家庭中的所有基因成员的 CFPS2020 的个人基础设计权数的均值。

1.2.2 家庭横截面无回答权数

由于家庭层面的无回答权数，是指家庭经济问卷的无回答权数。由于家庭经济问卷在访问中存在样本流失、样本无回答、拒绝访问、无法联系、身体原因无法访问等问题，因此需要计算家庭层面的无回答权数。此处，使用家庭或社区层面的辅助信息，基于 logistic 模型的倾向权数计算方法计算家庭横截面无回答权数。首先，将家庭样本分为完访样本和未完访样本，然后建立 logistic 回归模型，使用的变量是家庭规模、是否有 65 岁及以上的老人、是否有 10 岁以下的儿童、城乡、家庭城乡变量、区县的属性（是否为区），家庭所在的区域（东部、中部和西部），计算家庭横截面无回答调整系数是：

$$p_2 = 1 / (\exp(\beta_1 X) / (1 + \exp(\beta_1 X)))$$

所以家庭横截面无回答权数是 CFPS2020 的家庭基础设计权数与家庭无回答调整系数的乘积。

1.2.3 家庭横截面事后分层权数

随着时代的变迁，研究总体、家庭样本总体不断变化，而家庭样本分裂、消亡以及调查过程中的无回答、拒绝访问、无法联系等，导致样本总体与研究总体的结构发生变化。而 2020 年是普查年，各个总体的家户数量与七普的家户数量差异较大。因此，我们使用七普的各省的家庭户数进行完全事后分层调整，得到个人横截面事后分层调整权数系数，进而得到家庭横截面事后分层权数。

1.2.4 家庭横截面极值权数

上述权数调整会导致权数波动变大，使得估计量的方差也随之变大，降低了目标变量的估计精度。为了使得权数波动在一个可控范围，我们使用基于权数分布的分位数的权数截取方法（取分位数是 1% 和 99%）的权数截取方法，进而得到家庭横截面的极值调整。

1.2.5 标准化权数

由上，同样得到家庭横截面权数包含家庭横截面无回答权数（不进行事后分层调整，但是进行极值调整）和家庭横截面事后分层权数。

分别对上述家庭横截面权数进行标准化，得到最终的 CFPS2020 的家庭横截面无回答权

数（变量名 fswt_natcs20n）和家庭截面事后分层权数（变量名 fswtps_natcs20n）。具体的标准化方法可以参考“个人横截面标准化权数”部分。

二、CFPS2020 追踪权数

因为 CFPS 的家庭是对基因成员的所在的家庭进行访问，家庭样本由于基因成员的死亡、婚姻、迁移会发生变化，无论从人数还是结构上已经不再具有可比性，因此不再计算家庭追踪权数。

CFPS2020 个人追踪权数是为了研究目标总体的变动情况，其中目标总体是 CFPS2010 年的全国 25 省市的所有个人，样本总体是 2010 年个人问卷完成访问且 2020 年也完成个人问卷访问的基因成员。因此 CFPS2020 追踪权数基于 CFPS2010 年的完访个人权数（CFPS2010 年的个人事后分层权数），并在此基础上得到 2010 到 2020 年由于样本流失导致的无回答权数，最终得到 CFPS2020 的个人追踪权数。下面介绍 CFPS2020 个人追踪权数的计算方法。

2.1 个人追踪设计权数

由上，CFPS2020 个人追踪权数针对的样本总体是 CFPS2010 年完成个人问卷的个人样本，因此 CFPS2020 年的个人追踪设计权数是个人在 CFPS2010 年的个人事后分层权数。

2.2 个人追踪无回答权数

以 CFPS2010 完成个人访问的样本为基础，CFPS2020 的个人访问中不可避免的存在无回答²，此处仍旧基于 logistic 模型的倾向权数计算方法计算流失权数。首先将 2020 年的个人样本分为完访样本和未完访样本，建立 logistic 回归模型，使用的辅助变量是性别、家庭规模、家庭中是否有 65 岁以上老人、家中住房类型、家庭中是否有 10 岁以下儿童、城乡、婚姻状态、学历、是否在家、年龄，其中因为年龄变量的特殊性，在建模过程中年龄作为样条函数，计算倾向流失的权数调整系数是：

$$p_3 = 1 / (\exp(\beta_1 X) / (1 + \exp(\beta_1 X)))$$

所以，CFPS2020 的个人追踪无回答权数就是 2010 年个人基础权数与 2020 年个人流失权数调整系数的乘积。

² 我们在计算时将去世导致的无回答也视为流失。

2.3 个人横截面事后分层调整权数

随着时代的变迁，目标总体、样本总体不断变化，而样本个人的出生、死亡，以及调查过程中的无回答、拒绝访问、无法联系等，导致样本总体与目标总体的结构发生变化。其中，与研究目标变量相关的年龄、性别、城乡等变量的结构会发生变化，导致与全国的结构分布不同，因此需要对最终的样本总体进行结构调整。由于我们研究的追踪总体仍旧是 2010 年的目标总体，因此仍旧是使用 2010 年的第六次人口普查（以下简称“六普”）数据进行事后分层调整。与横截面权数计算相同，使用六普的分城乡年龄性别数据进行完全事后分层调整，即在六个抽样框中用城乡（分为城镇和乡村）、性别（分为男和女）、年龄变量（0~9、10~19、20~29、30~39、40~49、50~59、60 岁及以上）进行完全事后分层调整，得到个人追踪事后分层调整权数系数，进而得到个人追踪事后分层权数，即 CFPS2020 个人追踪设计权数、倾向流失的权数调整系数的倒数与个人追踪事后分层调整权数系数的乘积。

2.4 个人追踪极值调整权数

与个人横截面权数相同，经过权数调整，虽然保证了样本单元加权与总体大致保持一致，然而权数调整会导致权数波动变大，影响估计的效率，降低了目标变量的估计精度。为了使得权数波动在一个可控范围，需要对权数进行极值调整。同样地，此处使用基于权数分布的分位数的权数截取方法。首先，通过计算六个总体的权效应，发现仅上海市需要进行权数的极值调整。由此，得到五个大省和一个小省的个人追踪极值权数系数。

2.5 个人追踪标准化权数

由上，个人追踪设计权数、无回答调整系数、事后分层权数调整系数和极值权数调整系数的乘积，即为最终的 CFPS2020 个人追踪权数。

同样地，个人追踪权数包含个人追踪无回答权数（不进行事后分层调整）和个人追踪事后分层权数。对 CFPS 的个人追踪权数进行标准化，即为最终的 CFPS2020 的个人追踪标准化权数，即 CFPS2020 个人追踪无回答权数（发布数据集中的 `rswt_natpn1020n` 变量）和个人追踪事后分层权数（发布数据集中的 `rswtps_natpn1020n` 变量）。

2.6 CFPS2020 个人追踪样本的权数分布

下面，使用 CFPS2020 个人追踪权数和个人追踪样本进行分析，由于个人追踪权数仅针对 2010 年完访且 2020 年完访的基因成员，研究的总体仍旧是 2010 年的总体。因此我们通过对比发现，经过个人追踪权数加权后，样本的分布在年龄、性别和城乡上与 2010 年的 25 省市近似一致，加权效果较好。

CFPS2020 年追踪样本的年龄、性别、城乡分布与六普数据的对比，见表 5 至表 8。

表 5 CFPS2020 样本年龄加权分布与六普分布对比表

	样本比例	无回答权数比例	事后分层权数比例	六普比例 ¹
城市 0-9 男	2.51	2.34	2.5	2.5
城市 0-9 女	2.19	1.99	2.11	2.11
城市 10-19 男	2.27	3.43	3.4	3.4
城市 10-19 女	2	2.75	3.08	3.08
城市 20-29 男	2.01	2.11	4.91	4.91
城市 20-29 女	2.39	2.49	4.82	4.82
城市 30-39 男	2.73	3.43	4.62	4.62
城市 30-39 女	3.29	3.85	4.44	4.44
城市 40-49 男	3.51	3.51	4.53	4.53
城市 40-49 女	4.1	3.99	4.26	4.26
城市 50-59 男	2.98	2.93	2.94	2.94
城市 50-59 女	3.28	3.41	2.86	2.86
城市 60-男	2.4	2.81	2.89	2.89
城市 60-女	2.63	3.04	3.04	3.04
乡村 0-9 男	5.62	4.64	3.45	3.45
乡村 0-9 女	4.63	3.63	2.87	2.87
乡村 10-19 男	4.09	4.85	3.49	3.49
乡村 10-19 女	3.99	4.88	3.08	3.08
乡村 20-29 男	2.97	2.87	3.68	3.68
乡村 20-29 女	3.51	2.98	3.7	3.7
乡村 30-39 男	4.05	3.57	3.56	3.56
乡村 30-39 女	4.5	3.84	3.42	3.42
乡村 40-49 男	5.71	4.56	4.25	4.25
乡村 40-49 女	6.37	5.2	4.22	4.22
乡村 50-59 男	4.61	4.56	3.21	3.21
乡村 50-59 女	4.67	4.7	3.09	3.09
乡村 60-男	3.5	3.57	3.71	3.71
乡村 60-女	3.49	4.08	3.85	3.85

注：数据来源于 CFPS 基线调查时所涵盖的 25 省市。

表 6 CFPS2020 样本年龄加权分布与六普分布对比表

总体	分类	样本比例	无回答权数比例	事后分层权数比例	六普比例
上海市	0-9 岁	7.61	8.61	6.2	6.2
上海市	10-19 岁	6.77	6.88	7.29	7.29
上海市	20-29 岁	11.22	11.3	22.55	22.55
上海市	30-39 岁	14.1	10.48	17.59	17.59
上海市	40-49 岁	15.31	12.91	15.98	15.98
上海市	50-59 岁	23.56	23.54	15.32	15.32
上海市	60-岁	21.43	26.28	15.07	15.07
辽宁省	0-9 岁	9.91	9.07	7.26	7.26
辽宁省	10-19 岁	9.97	13.56	9.91	9.91
辽宁省	20-29 岁	10.55	9.6	15.33	15.33
辽宁省	30-39 岁	15.22	14.26	15.99	15.99
辽宁省	40-49 岁	20.81	17.5	19.26	19.26
辽宁省	50-59 岁	19.48	20.01	16.82	16.82
辽宁省	60-岁	14.06	15.99	15.43	15.43
河南省	0-9 岁	18.14	15.24	14.45	14.45
河南省	10-19 岁	12.64	18.2	14.44	14.44
河南省	20-29 岁	11.19	10.53	16.74	16.74
河南省	30-39 岁	13.02	11.89	14.22	14.22
河南省	40-49 岁	18.74	18.01	16.16	16.16
河南省	50-59 岁	12.98	12.88	11.27	11.27
河南省	60-岁	13.28	13.26	12.73	12.73
甘肃省	0-9 岁	15.82	11.99	11.09	11.09
甘肃省	10-19 岁	14.8	16.62	16.6	16.6
甘肃省	20-29 岁	10.87	10.92	15.66	15.66
甘肃省	30-39 岁	14.9	16.02	15.58	15.58
甘肃省	40-49 岁	21.39	21.43	18.39	18.39
甘肃省	50-59 岁	12.82	13.04	10.25	10.25
甘肃省	60-岁	9.4	9.97	12.44	12.44
广东省	0-9 岁	18.72	13.39	10.34	10.34
广东省	10-19 岁	16.73	19.24	16.11	16.11
广东省	20-29 岁	10.09	10.74	21.95	21.95
广东省	30-39 岁	11.55	11.08	17.7	17.7
广东省	40-49 岁	19.3	17.96	15.25	15.25
广东省	50-59 岁	13	13.43	8.91	8.91
广东省	60-岁	10.61	14.16	9.73	9.73
25 省 ¹	0-9 岁	14.94	12.6	10.93	10.93
25 省	10-19 岁	12.35	15.91	13.05	13.05
25 省	20-29 岁	10.88	10.45	17.12	17.12
25 省	30-39 岁	14.57	14.69	16.04	16.04
25 省	40-49 岁	19.69	17.26	17.27	17.27
25 省	50-59 岁	15.54	15.6	12.1	12.1

25 省	60-岁	12.02	13.5	13.49	13.49
------	------	-------	------	-------	-------

注：此处 25 省指的是 CFPS 基线调查时所涵盖的 25 省市

表 7 CFPS2020 样本性别加权分布与六普分布对比表

总体	分类	样本比例	无回答权数比例	事后分层权数比例	六普比例
上海市	男性	49.72	46.81	51.5	51.5
上海市	女性	50.28	53.19	48.5	48.5
辽宁省	男性	44.78	48.81	50.63	50.63
辽宁省	女性	55.22	51.19	49.37	49.37
河南省	男性	49.05	48.87	50.51	50.51
河南省	女性	50.95	51.13	49.49	49.49
甘肃省	男性	50.55	51.74	51.08	51.08
甘肃省	女性	49.45	48.26	48.92	48.92
广东省	男性	50.61	49.21	52.15	52.15
广东省	女性	49.39	50.79	47.85	47.85
25 省 ¹	男性	48.96	49.18	51.16	51.16
25 省	女性	51.04	50.82	48.84	48.84

注：此处 25 省指的是 CFPS 基线调查时所涵盖的 25 省市

表 8 CFPS2020 样本城乡加权分布与六普分布对比表

总体	分类	样本比例	无回答权数比例	事后分层权数比例	六普比例
上海市	城市	77.64	78.44	89.3	89.3
上海市	乡村	22.36	21.56	10.7	10.7
辽宁省	城市	47.09	56.31	62.15	62.15
辽宁省	乡村	52.91	43.69	37.85	37.85
河南省	城市	35.69	35.25	38.52	38.52
河南省	乡村	64.31	64.75	61.48	61.48
甘肃省	城市	13.29	24.89	35.94	35.94
甘肃省	乡村	86.71	75.11	64.06	64.06
广东省	城市	49.74	56.92	66.17	66.17
广东省	乡村	50.26	43.08	33.83	33.83
25 省 ¹	城市	38.29	42.07	50.4	50.4
25 省	乡村	61.71	57.93	49.6	49.6

注：此处 25 省指的是 CFPS 基线调查时所涵盖的 25 省市

由表 5 至 表 8 可知，经过加权调整，尤其是事后分层调整后，样本的年龄、性别、城乡分布与六普更为接近，估计精度更高。

三、CFPS2020 的权数使用

CFPS2020 年包含横截面权数和追踪权数，其中横截面权数包含个人截面权数和家庭截面权数，追踪权数包含个人追踪权数。其中个人截面权数包含个人截面无回答权数（`rswt_natcs20n`）和个人截面事后分层权数（`rswtps_natcs20n`），家庭截面权数包含家庭截面无回答权数（`fswt_natcs20n`）和家庭截面事后分层权数（`fswtps_natcs20n`），个人追踪权数包含个人追踪无回答权数（`rswt_natpn1020n`）和个人追踪事后分层权数（`rswtps_natpn1020n`）。

在研究中，研究者根据自己的研究目的选取不同的权数。其中 CFPS2020 年家庭横截面权数是为了估计 2020 年的全国家庭总体情况，例如 2020 年的全国家庭住房拥有率等。CFPS2020 年个人横截面权数是为了估计 2020 年的全国个人总体情况，例如 2020 年全国离婚率等。2020 年个人追踪权数是为了分析 2010 年到 2020 年个人的变动情况，2020 年 CFPS 数据库可以研究 2010、2012、2014、2016 和 2018 年的个体变动情况，也可以研究 2010 年、2020 年两年的个体的变动情况，当然可以研究任意两年或多年的个体情况，但是需要注意 CFPS2020 的个人追踪权数仅是针对 2010 年完访的且在 2020 年也完成访问的个人，若部分调查年没有数据，可以视为缺失数据，按照缺失数据的分析方法进行插补、删除或加权调整。例如研究 2010 年到 2018 年的 5 年的婚姻变动情况，若其中 2012 年没有完成访问，可以根据 2010 和 2014 的婚姻数据或 2012 年的家庭成员数据或 2012 年家中其他人员的数据等进行插补后，再进行纵向分析。根据 CFPS 的研究目的，CFPS 无法分析其他轮次与 2020 年的面板数据（如 2012、2016 和 2020）。

CFPS2020 年的权数包含无回答权数和事后分层权数，建议在研究中使用事后分层权数。

相关阅读

CFPS 用户手册：<http://www.issp.pku.edu.cn/cfps/wdxx/yhsc/index.htm>

CFPS 技术报告：<http://www.issp.pku.edu.cn/cfps/wdxx/jzbg/index.htm>

CFPS-01 中国家庭追踪调查抽样设计

CFPS-17 中国家庭追踪调查 2010 年基线调查权数计算

Xie, Y. L., Ping (2015). "The sampling design of the China family panel studies (CFPS)." *Chinese Journal of Sociology* 1(4): 471-484.

CFPS 小课堂 | 权数使用 Q&A：<http://www.issp.pku.edu.cn/cfps/cjwt/qz/1356863.htm>