China Family Panel Studies

CFPS 中国家庭追踪调查

技术报告系列：CFPS-17

系列编辑：谢宇　责任编辑：胡婧炜

中国家庭追踪调查
2010 年基线调查权数计算

（第二版）

吕萍　谢宇

2013.7.17

CFPS 2010 对五个“大省”（上海市、河南省、甘肃省、辽宁省和广东省）和一个“小省”（25 省市中的其他省市）共 6 个总体分别计算了家庭问卷、成人问卷和少儿问卷三个数据库的权数，权数的调整包含抽样设计权数、无回答调整权数、事后分层调整权数以及在加权过程为了保证样本的精度的权数的极端值调整。本报告将对权数调整过程及结果进行具体介绍。

1 CFPS 2010 权数调整过程

CFPS 2010 的抽样、访问流程及各个环节中需要进行的权数调整如下图 1 所示：

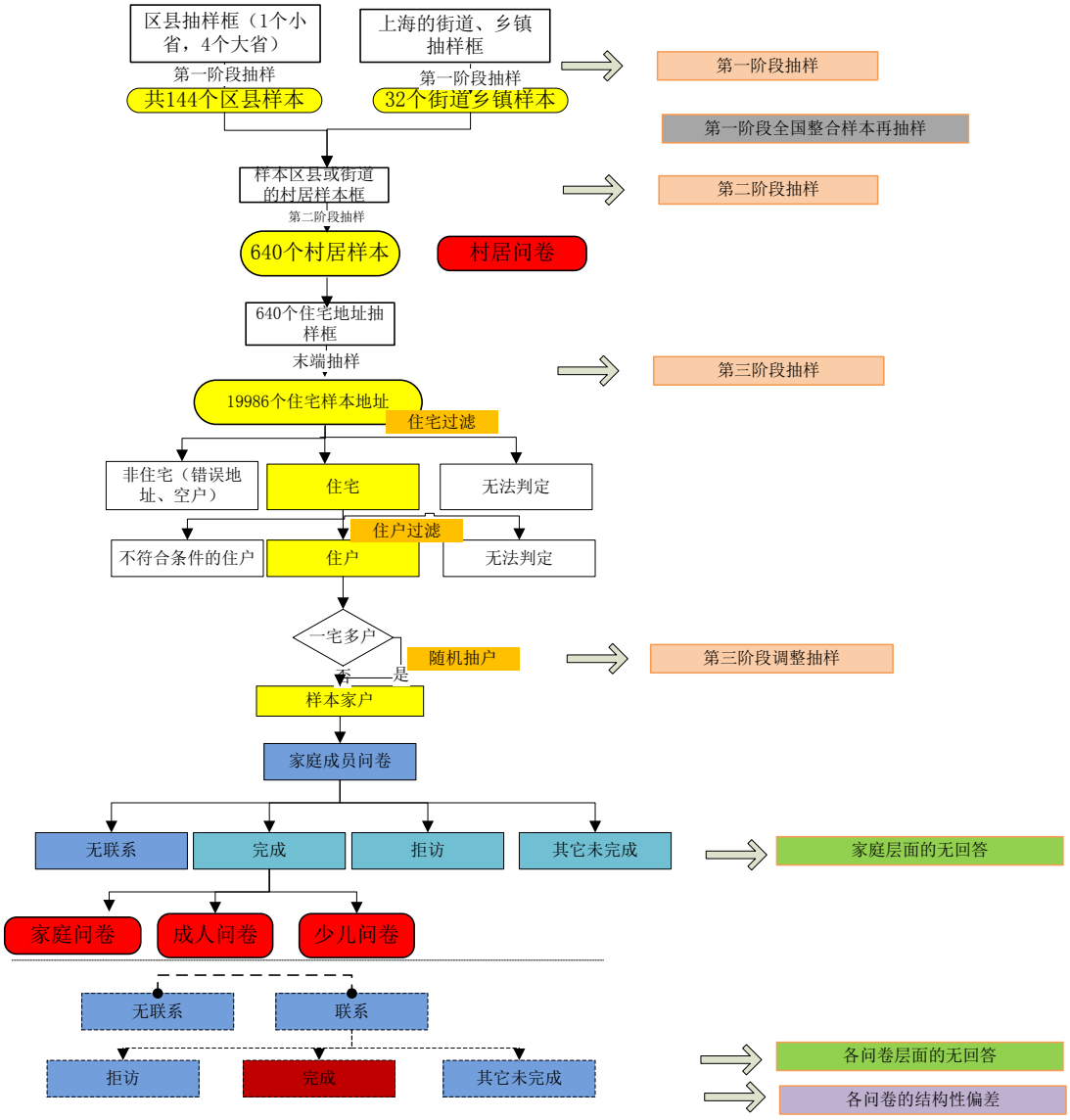

图 1. CFPS 2010 抽样和访问流程图

1.1 抽样设计权数

抽样设计权数是抽样过程中各抽样阶段的抽样概率的倒数，即为第一阶段、第二阶段、第三阶段抽样和第三阶段调整抽样概率的乘积的倒数。其中, 第三阶段调整抽样主要针对末端抽样框不完善、同一个抽样地址下有多个满足条件的家户导致的抽样框误差。在实际调查过程中，CFPS 采用随机抽取一户的方法弥补该误差。

由此，抽样设计权数是：

$$
\begin{aligned}
W_d &= \frac{1}{p_1 p_2 p_3 p'_3} \\
&= \frac{N_i}{m_i N_{ij}} \times \frac{N_{ij}}{m_{ij} N_{ijk}} \frac{N_{ijk}}{N'_{ijk}} \times \frac{H_{ijk}}{h_{ijk}} \times \frac{S_{ijkh}}{s_{ijkh}} \\
&= \frac{N_i H_{ijk} S_{ijkh}}{m_i m_{ij} N'_{ijk} h_{ijk} s_{ijkh}}
\end{aligned}
$$

其中，p_1、p_2、p_3、p'_3分别表示第一阶段、第二阶段、第三阶段和第三阶段调整抽样的概率。N_i是第i个抽样框的总人口数，N_{ij}是第i个抽样框的第j个样本区县（上海为街道/乡镇）的人口数，m_i是第i个抽样框的样本区县的个数。N_{ijk}是第i个抽样框的第j个样本区县（上海为街道/乡镇）中第k个样本村居的人口数，N'_{ijk}是第i个抽样框的第j个样本区县（上海为街道/乡镇）中第k个样本拆分后选中的区域的人口数。m_{ij}是第i个抽样框的第j个样本区县（上海为街道/乡镇）的样本村居个数。H_{ijk}是第i个抽样框的第j个样本区县（上海为街道/乡镇）中第k个样本村居中的所有家户数，h_{ijk}是第i个抽样框的第j个样本区县（上海为街道/乡镇）中第k个样本村居的有效样本家户数。S_{ijkh}, s_{ijkh}分别表示第i个抽样框的第j个样本区县（上海为街道/乡镇）中第k个样本村居中第h个样本地址下的有效家户数和抽取的样本家户数。

需要注意的是，在抽样设计中，为了保证抽样效率，CFPS2010 对大村居进行了拆分，对小村居进行了合并。若为合并的村居，N_{ijk}表示合并的村居的人口数。若为拆分村居，N'_{ijk}是拆分后的样本的人口数。为了简化运算，若该村居无需拆分，则定义N_{ijk}与N'_{ijk}相等。

此次调查包含的六个子总体（25 省、甘肃省、辽宁省、广东省、河南省、上海市）的

抽样设计权数均按照上述公式计算。

对于全国整合样本[1]，其抽样设计权数是

$$W'_d = \frac{1}{p_1 p'_1 p_2 p_3 p'_3}$$
$$= \frac{N_i}{m_i N_{ij}} \times \frac{N'_i}{m'_i N'_{ij}} \times \frac{N_{ij}}{m_{ij} N_{ijk}} \frac{N_{ijk}}{N'_{ijk}} \times \frac{H_{ijk}}{h_{ijk}} \times \frac{S_{ijkh}}{s_{ijkh}}$$
$$= \frac{N_i N'_i H_{ijk} S_{ijkh}}{m_i m'_i m_{ij} N'_{ij} N'_{ijk} h_{ijk} s_{ijkh}}$$

其中，p'_1表示在第一阶段区县（上海为街道/乡镇）样本基础上的再整合区县（上海为街道/乡镇）样本被抽取的概率，N'_i表示第i个样本区县（上海为街道/乡镇）样本抽样框的人口数，N'_{ij}表示第i个样本区县（上海为街道/乡镇）样本抽样框中第j个再整合区县（上海为街道/乡镇）样本的人口数，m'_i表示第i个样本区县（上海为街道/乡镇）样本抽样框的再整合区县（上海为街道/乡镇）样本的个数。

1.2 无回答调整权数

在 CFPS 2010 调查中，区县、村居层面没有无回答，因此 CFPS 2010 的无回答调整主要在家庭层面和问卷层面，此处的家庭层面指家庭成员问卷层面，问卷层面指家庭问卷、成人问卷、少儿问卷层面。在本次调查中，定义完成家庭成员问卷的家庭为完访家庭。

1.2.1 家庭层面的无回答调整权数

家庭层面的无回答采用加权组调整的方法，即在各个抽样框的村居层面进行加权组调整。各个加权组中的调整系数采用 AAPRO 的应答率 RR1，即

$$P_n^1 = \frac{I_{ijk}}{I_{ijk} + R_{ijk} + NC_{ijk} + O_{ijk} + UE_{ijk}}$$

其中I_{ijk}、R_{ijk}、NC_{ijk}、O_{ijk}、UE_{ijk}分别为各个抽样框中各个样本村居层面的家庭成员问卷的完成数量、拒访的样本数量、无联系的样本数量、其他未完成的样本数量、不确定是否符合访问条件的样本数量。

[1] 整合样本也称再抽样样本。

1.2.2 问卷层面的无回答调整权数

本次调查的需要加权的问卷包含家庭问卷、成人问卷和个人问卷。

⑴ 家庭问卷的无回答调整权数

在 CFPS 2010 年调查中，家庭层面完成家庭成员问卷后继续完成家庭问卷的比例达到99%。因此，采用样本村居层面的加权组调整的方法，家庭问卷的无回答调整系数是

$$P_n^{\,2} = \frac{n'_{fam\_ijk}}{n_{fam\_ijk}}$$

其中 $n_{fam\_ijk}, n'_{fam\_ijk}$ 分别表示第 i 个抽样框的第 j 个样本区县（上海为街道/乡镇）的第 k 个样本村居的完成家庭成员问卷的家户中家庭问卷的有效家户数和完成家庭问卷的家户数。

所以，家庭问卷的无回答调整权数是

$$W_{nf} = \frac{1}{p_1 p_2 p_3 p'_3 P_n^1 P_n^2}$$

对于全国整合样本，家庭问卷的无回答调整权数是

$$W'_{nf} = \frac{1}{p_1 p'_1 p_2 p_3 p'_3 P_n^1 P_n^2}$$

⑵ 个人问卷的无回答调整权数

如图 1 所示，在实际访问过程中，CFPS 对完成家庭成员问卷的家户进行个人问卷的访问，其中个人问卷包含成人问卷和少儿问卷。由于个人层面有家庭成员问卷信息可以利用，为了提高权数的精度，在个人层面采用两阶段的基于 logistic 模型的倾向权数计算方法，进而得到成人和少儿层面的无回答调整系数。

首先将个人样本分为联系样本和非联系样本，利用数据中的辅助信息，建立 logistic 回归模型，则个人问卷联系层次的倾向应答概率是

$$\hat{p}_1 = \exp(\beta_1 X) / (1 + \exp(\beta_1 X))$$

建模过程中用的变量如表 1 所示。

表 1． 变量名列表

变量名	标签	类型
city	城乡	分类
county	区县	分类
TB1B_A_P	年龄	分段
TB2_A_P	性别	分类
TB3_A_P	婚姻状况	分类
TB4_A_P	最高学历	分类
TB6_A_P	现在是否住在家中	分类
num	家庭人口数	连续
ifold	是否有老人	分类
ifchild	是否有儿童	分类
house	房屋所有情况	分类

注意，由于年龄变量在社会经济中有非常大的作用，此处将年龄变量视为样条函数进行处理，并基于各个子总体的数据特征决定年龄变量样条函数节点的选择。

同理，在联系上的样本中，将样本分为拒访样本和非拒访问样本，建立 logistic 回归模型，则个人问卷联系上的样本中拒绝访问的样本的倾向应答概率是

$$\hat{p}_2 = \exp(\beta_2 X) / (1+\exp(\beta_2 X))$$

所以，个人问卷的个人层面的无回答调整系数是 $\hat{p} = \hat{p}_1(1-\hat{p}_2)$

由此，个人问卷的无回答调整权数是

$$W_{ng} = \frac{1}{p_1 p_2 p_3 p'_3 P_n^1 P_n^2 \hat{p}}$$

对于全国整合样本，个人问卷的无回答调整权数是

$$W_{ng} = \frac{1}{p_1 p_1^{'} p_2 p_3 p'_3 P_n^1 P_n^2 \hat{p}}$$

1.3 事后分层调整权数

由于抽样设计的复杂性、实地调查过程中问题的多样性、样本无回答的存在，CFPS2010 在某些关键变量上存在样本结构性偏差，导致得到的估计量有偏。为了调整该结构性偏差，减小抽样误差，提高估计精度，需要对成人和少儿样本数据进行事后分层调整。

在个人问卷层面，性别、年龄、城乡是非常重要的指标。因此，在六个抽样框以及整合

样本的成人和少儿问卷数据中，用城乡（分为城镇和农村）、性别（分为男和女）、年龄变量（分为 16~19、20~29、30~39、40~49、50~59、60~69、70~79、 80 岁以上，共 8 类）进行完全事后分层调整。由于 CFPS 个人问卷中年龄、性别有极少量的缺失，采用均值和中位数插补方法对其进行插补。各个总体的各个抽样框的事后分层调整的系数是：

$$\mathrm{p}_{\mathrm{postmj}}=\frac{W_{mj}}{N_{mj}}$$

其中W_{mj} 、 N_{mj} 分别表示第 m 个总体的第 j 个抽样框的权数和以及总量。

由此，个人问卷的事后分层调整权数是

$$W_{postg}=\frac{1}{p_1p_2p_3p'_3P_n^1P_n^2\hat{p}p_{postmj}}$$

对于全国整合样本，个人问卷的事后分层调整权数是：

$$W_{postg}=\frac{1}{p_1p_1'p_2p_3p'_3P_n^1P_n^2\hat{p}p_{postmj}}$$

1.4 权数的极值调整

由上得到成人和少儿问卷的事后分层调整权数和家庭问卷的无回答调整权数。但是，在实际利用权数进行目标变量的估计过程中，权数差异太大会导致过大的方差，影响估计的效率，因此需要对最终的权数进行极值调整。但是，权数的极值调整会带来一定的偏差，因此在调整中一定要注意降低均方误差。

我们从两部分对权数的极值进行调整：

1.4.1 过程中极值权数调整

在进行权数的无回答和事后分层调整时，若无回答、事后分层调整导致权数差异过大，则说明此无回答和事后分层调整将带来大的方差，降低抽样效率。因此，在权数调整过程中需要使每一次的调整系数都在一定的范围内，即在设计权数基础上的无回答调整、事后分层调整不要太大或是太小，我们使其满足：

$$(W_{non}/\bar{W}_{non})/(W_{base}/\bar{W}_{base})\in[1/L,L]$$

或

$$(W_{post}/\bar{W}_{post})/(W_{base}/\bar{W}_{base})\in[1/L,L]$$

其中，令 L 为 3， W_{base}、W_{non}、W_{post} 分别表示抽样设计权数、无回答和事后分层调整权数。由此，得到过程中极值无回答、极值事后分层的调整权数是 $w^{extr}{}_1$。

1.4.2 最终极值权数调整

上述权数调整只是为了使每次无回答、事后分层调整时权数差异不要太大，但是最终的权数是抽样设计权数和无应答、事后分层调整系数的乘积。因此，同样需要对最终的权数进行极值调整，以保证估计效率。通过对最终的事后分层调整权数分布的分析以及经验研究发现，用事后分层权数分布的 0.05 和 0.95 分位数作为最小最大值的极值点是比较好的选择。由此，对上述事后分层调整权数进行极值调整，即满足

$$W'_{post}=\begin{cases}W_{0.05}, W_{post}<W_{0.05}\\ W_{0.95}, W_{post}>W_{0.95}\end{cases}$$

由此，得到该部分的调整极值系数是 $w^{extr}{}_2$。

所以，极值调整需要在各个总体的家庭问卷的无回答调整权数和成人问卷、少儿问卷的事后分层调整权数的基础上同乘以 $w^{extr}{}_1 w^{extr}{}_2$。

1.5 最终调整权数

在调查数据的加权调整中，要求最终权数的和等于总体，而通过极值调整后的权数的和不再与总体相等，因此需要对上述权数进行再次调整。此处，采用简单的将各个总体视为均匀总体的方法对五个“大省”以及一个“小省”分别进行极值调整，即同乘各个子抽样框的极值调整因子 $w^{adjustm}$。，使其满足：

$$\sum_s w^{extr}{}_{post} w^{adjustm} = N_m$$

其中 N_m 是各个子总体的总量。

由此，得到家庭问卷的最终调整权数是

$$W_f = \frac{1}{p_1 p_2 p_3 p'_3 P_n^1 P_n^2} w^{extr}{}_{1f} w^{extr}{}_{2f} w_f{}^{adjustm}$$

全国整合样本的家庭问卷的最终权数是

$$W'_f = \frac{1}{p_1 p_1' p_2 p_3 p'_3 P_n^1 P_n^2} w^{extr}_{1f} w^{extr}_{2f} w_f^{adjustm}$$

个人问卷的最终调整权数是

$$W_g = \frac{1}{p_1 p_2 p_3 p'_3 P_n^1 P_n^2 \hat{p}p_{postmj}} w^{extr}_{1g} w^{extr}_{2g} w_g^{adjustm}$$

全国整合样本的个人问卷的最终调整权数是：

$$W_g = \frac{1}{p_1 p_1' p_2 p_3 p'_3 P_n^1 P_n^2 \hat{p}p_{postmj}} w^{extr}_{1g} w^{extr}_{2g} w_g^{adjustm}$$

由上，分别得到五个“大省”、一个“小省”、一个整合样本共 7 个总体的家庭、成人、少儿的权数。最终的权数数据库包含 25 省市完全样本的权数数据库、25 省市整合样本的权数数据库，每个数据库中包含家庭、成人和少儿三个权数数据。其中 25 省市完全样本的权数数据库的权数为“大省”和“小省”的权数数据库的合并。

2 权数调整后的图表比较

2.1 权数分布图

由上，本次共有 7 个总体的家庭、成人、少儿三套问卷的权数。其中 25 省市完全样本以及 25 省市整合样本的家庭、成人、少儿数据库的权数分布图如下所示。

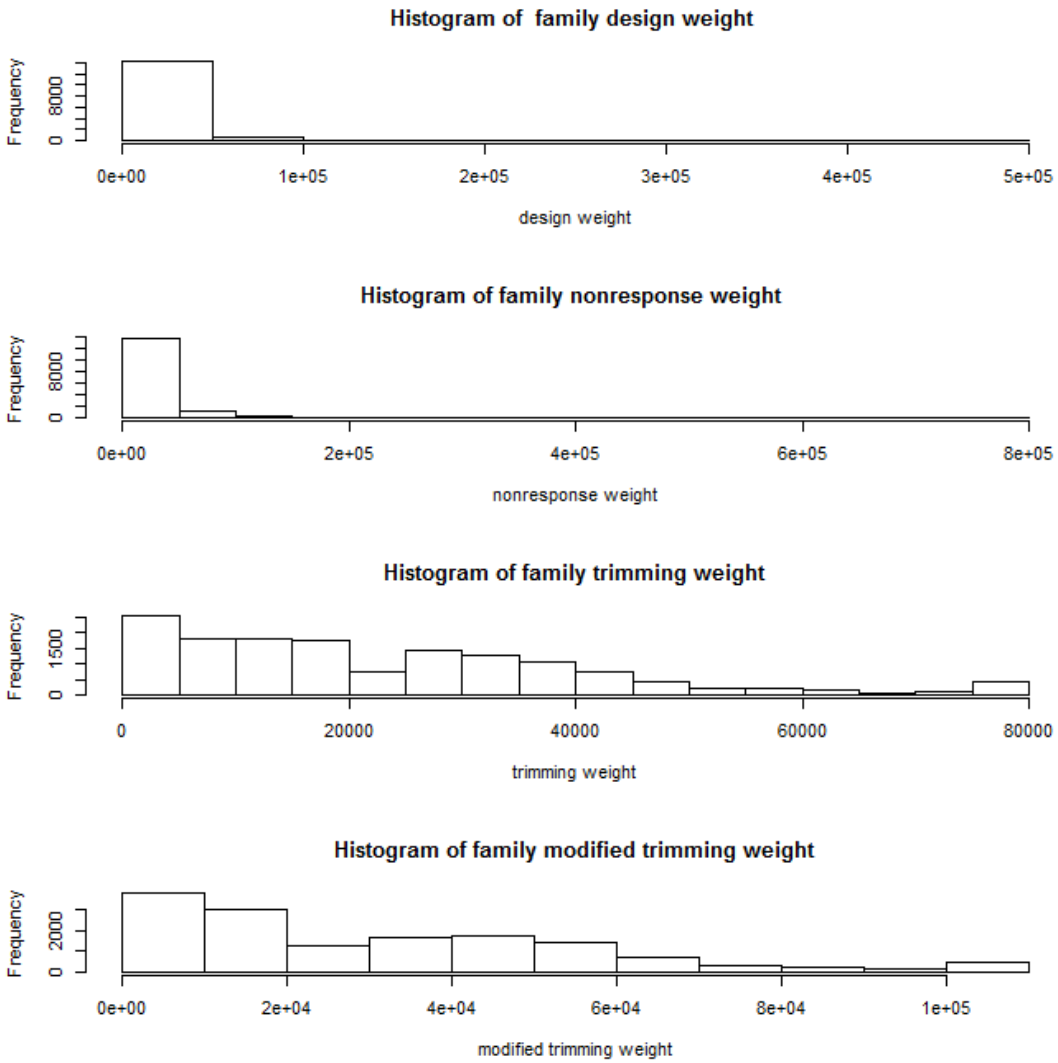

图 2.　完全样本家庭问卷的抽样设计、无回答、极值调整和最终权数的分布图

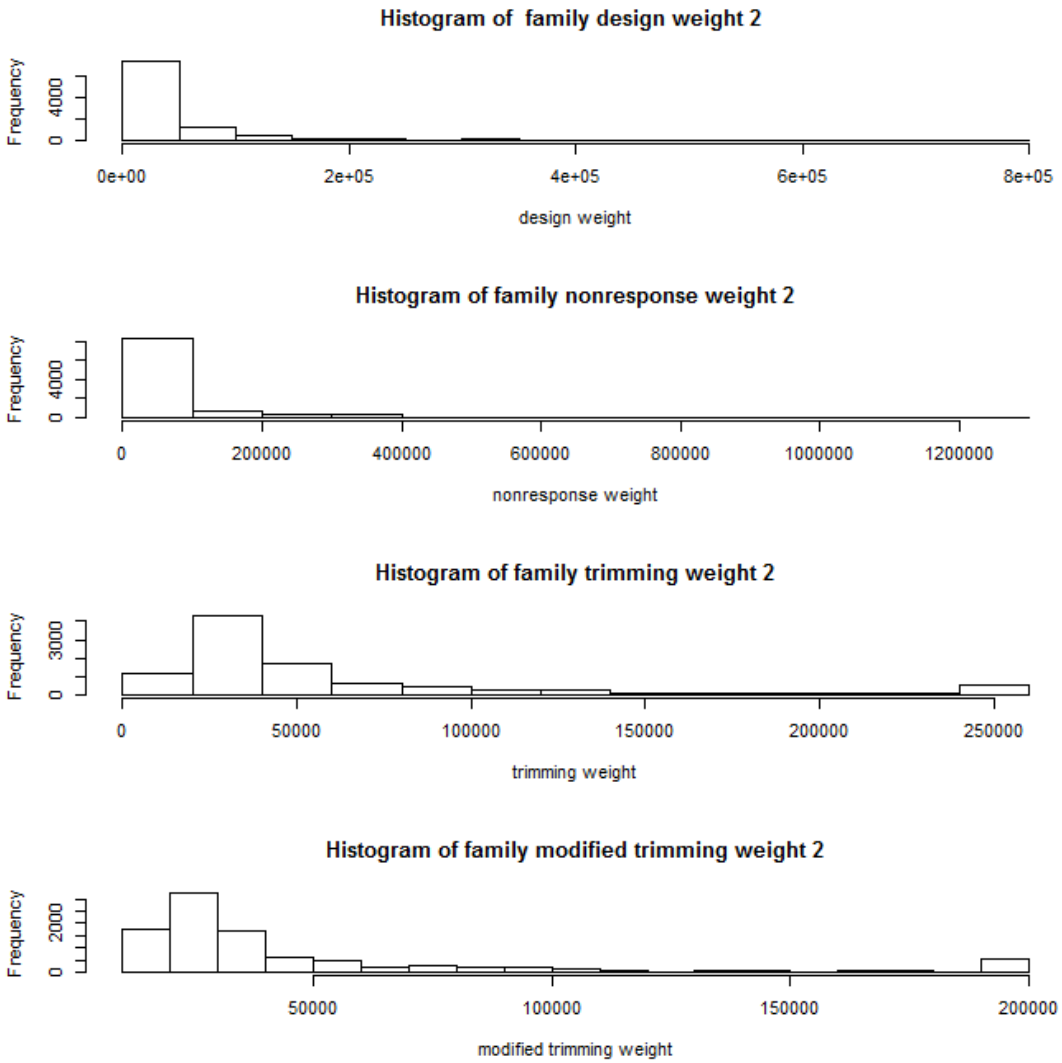

图 3.　整合样本家庭问卷的抽样设计、无回答、极值调整和最终权数的分布图

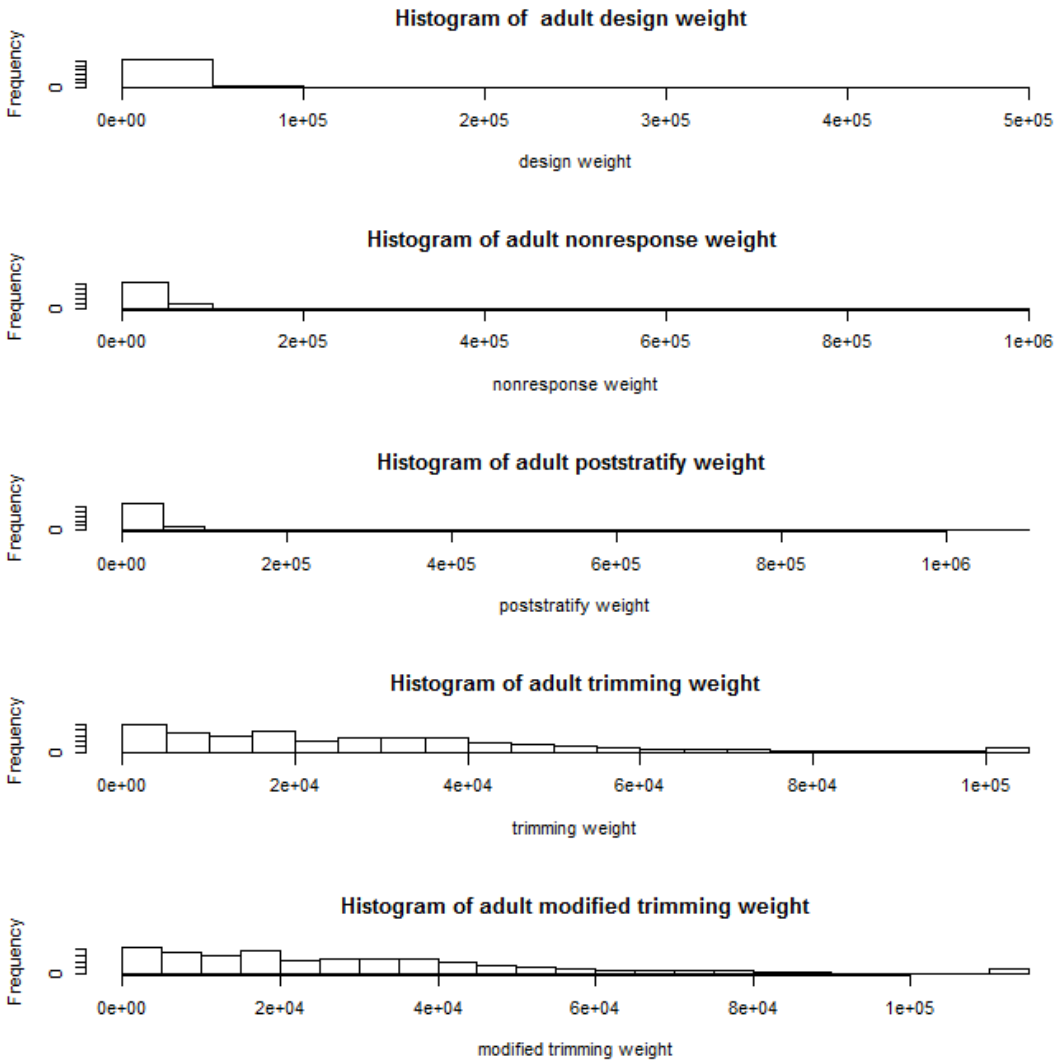

图 4.　完全样本成人问卷的抽样设计、无回答、事后分层、极值调整和最终权数的分布图

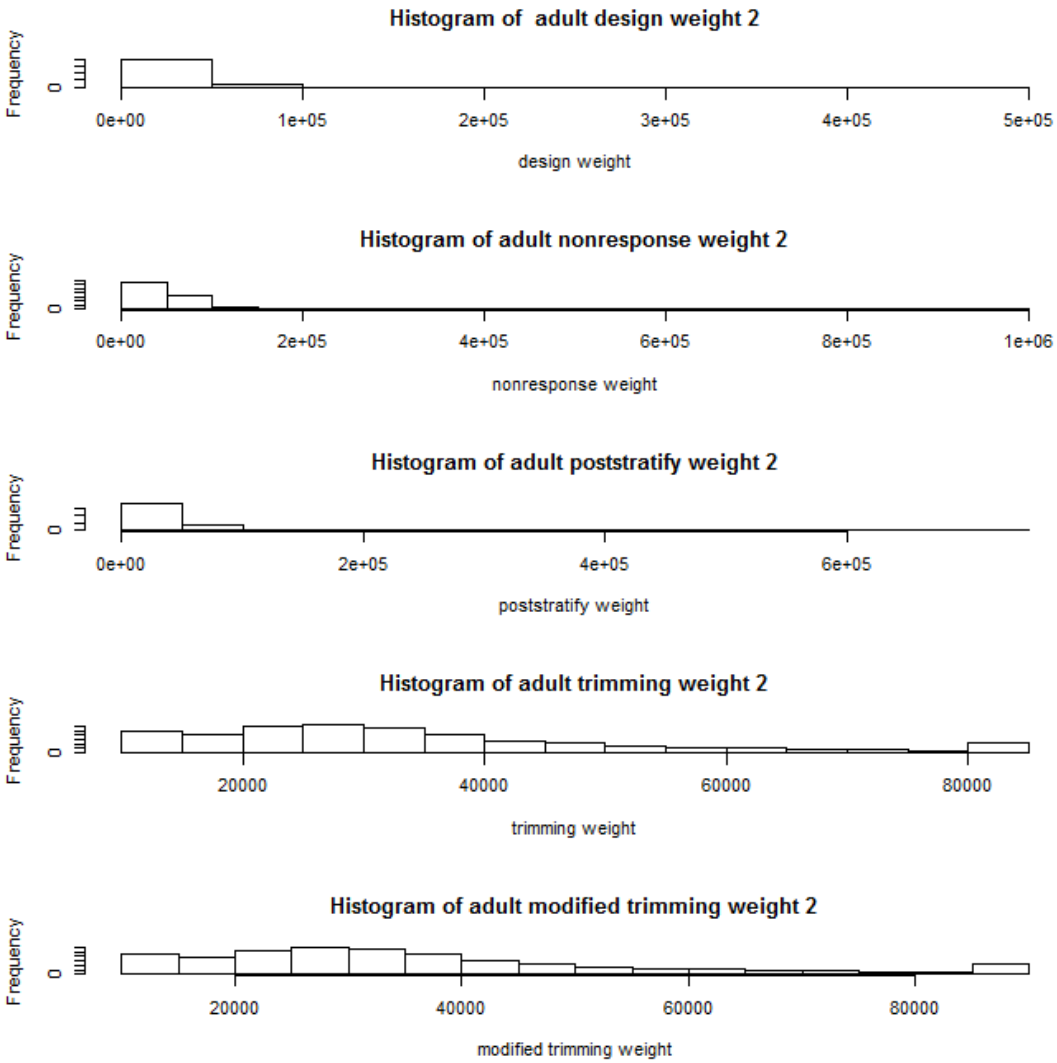

图 5.　整合样本成人问卷的抽样设计、无回答、事后分层、极值调整和最终权数的分布图

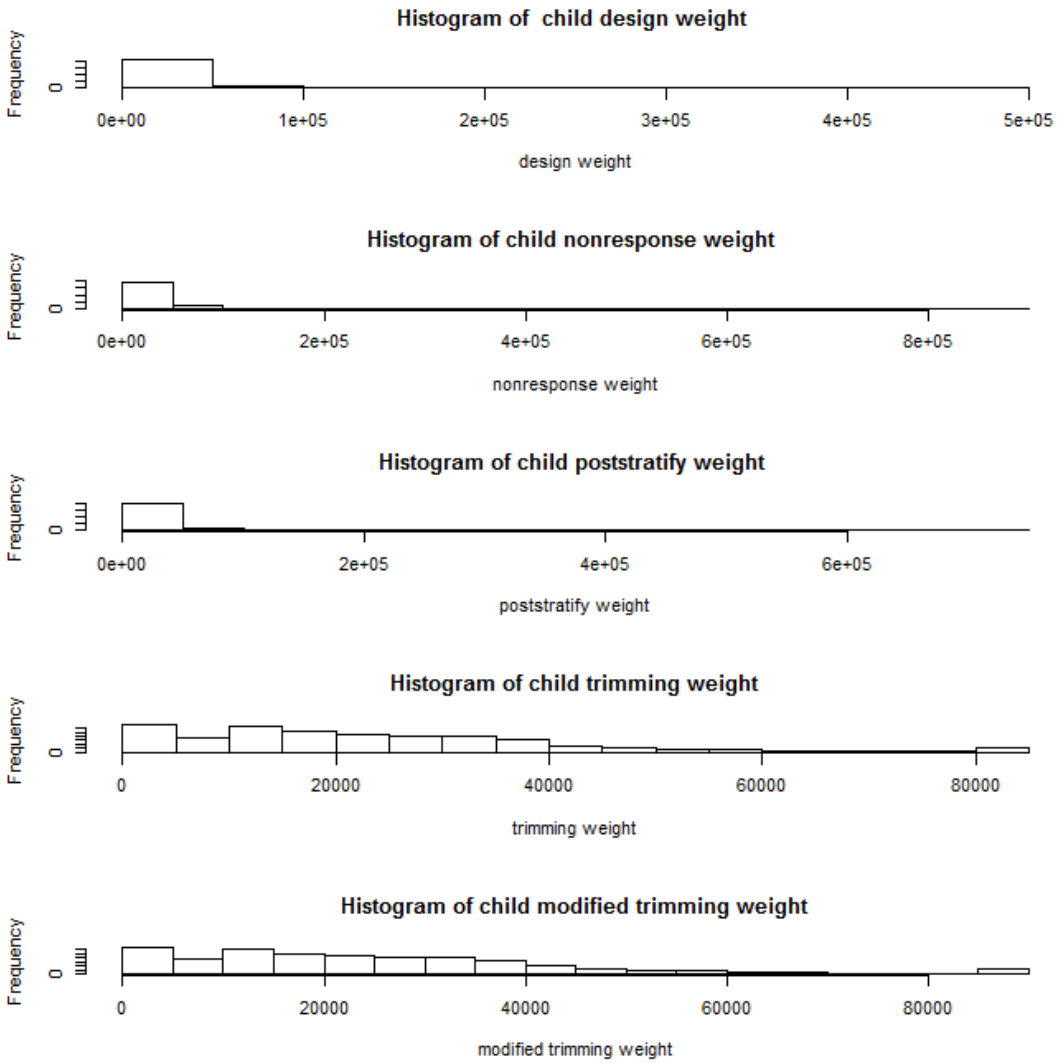

图 6.　少儿问卷的抽样设计、无回答、事后分层、极值调整和最终权数的分布图

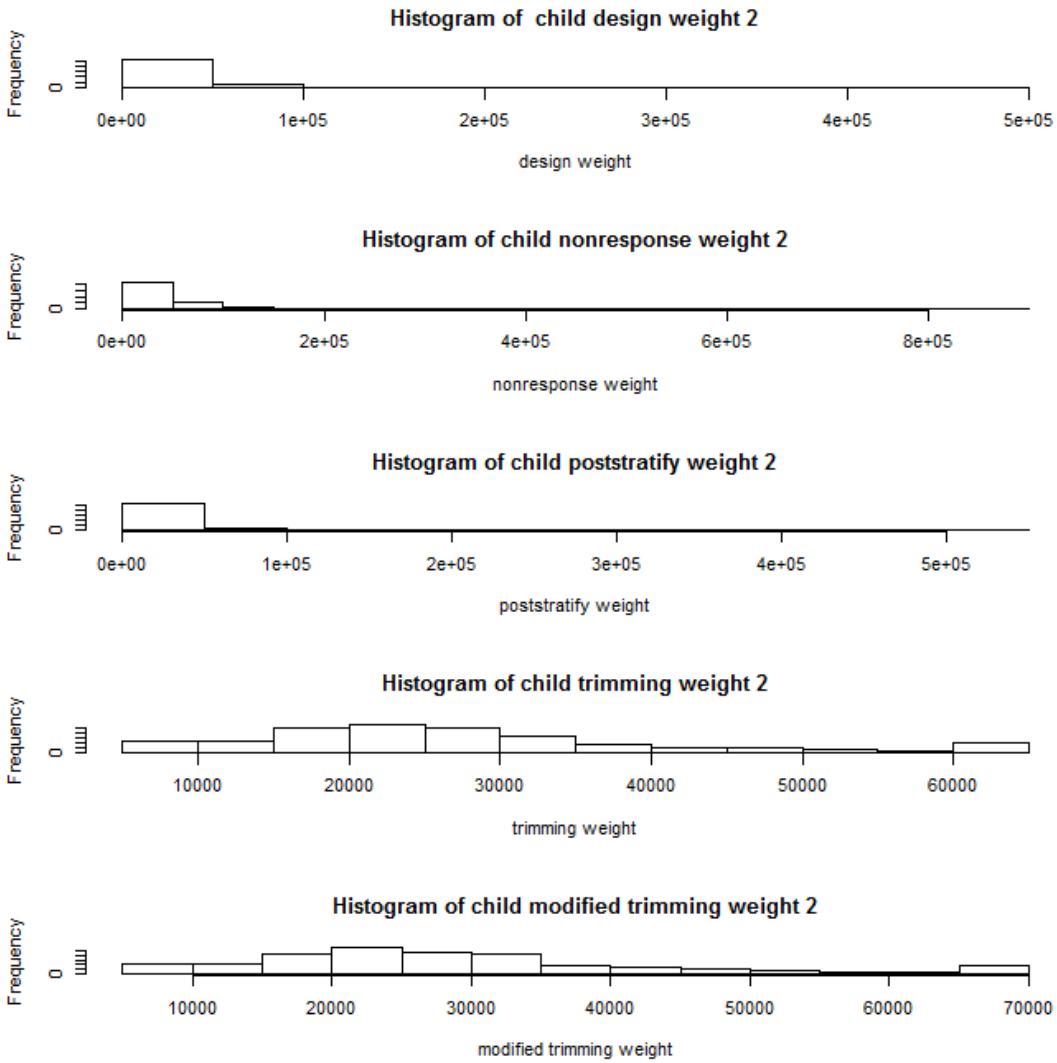

图 7. 整合样本少儿问卷的抽样设计、无回答、事后分层、极值调整和最终权数的分布图

由图 2—图 7 可以看出，经过抽样设计、无回答、事后分层权数调整后，权数分布变动较大，对权数进行极值调整则使权数的分布变得较为均匀，降低了方差，提高了估计精度。

各个问卷的六个抽样框的最终权数的分布如图 8—图 10 所示：

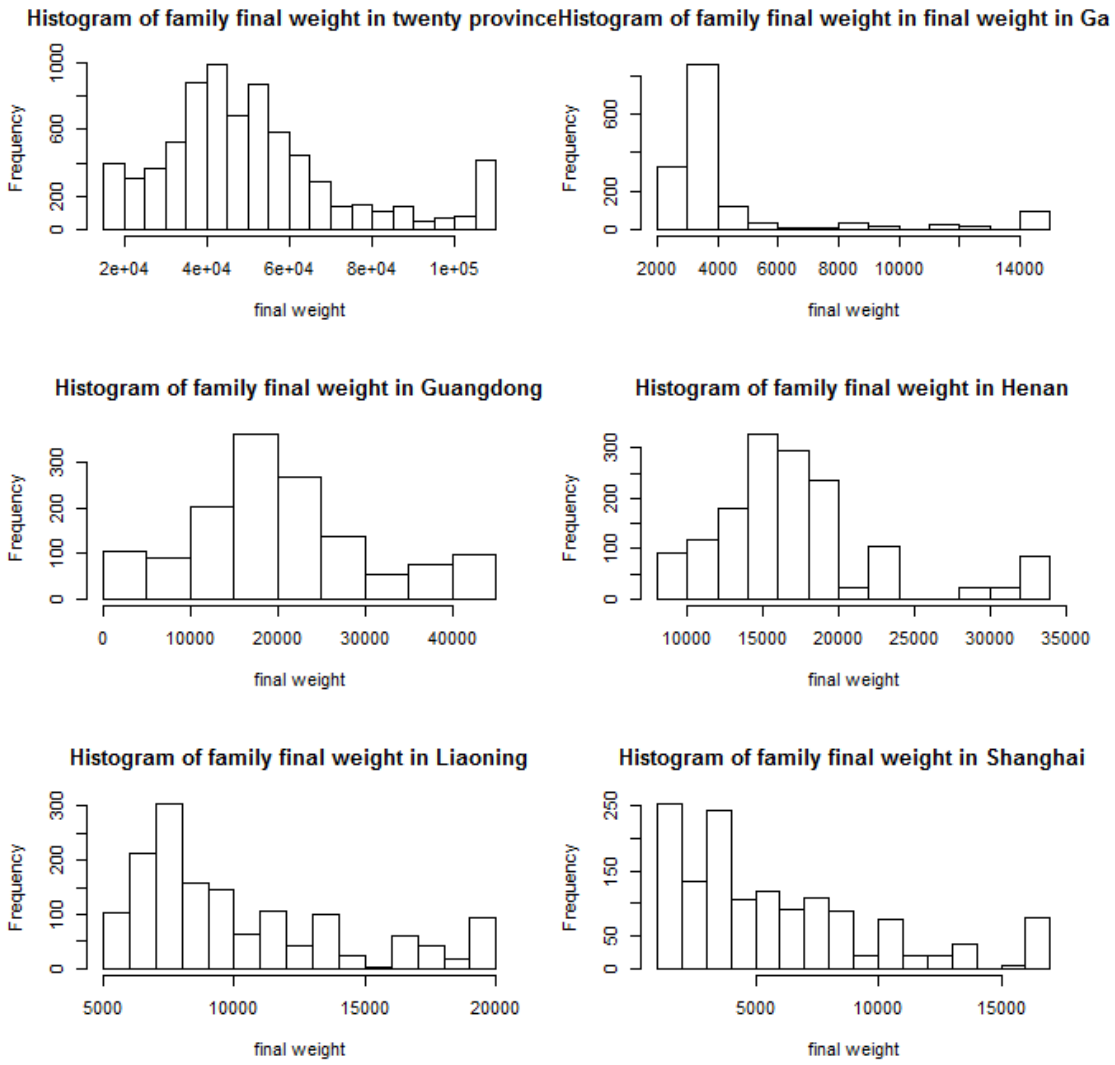

图 8.　家庭问卷的 6 个抽样框最终的权数分布

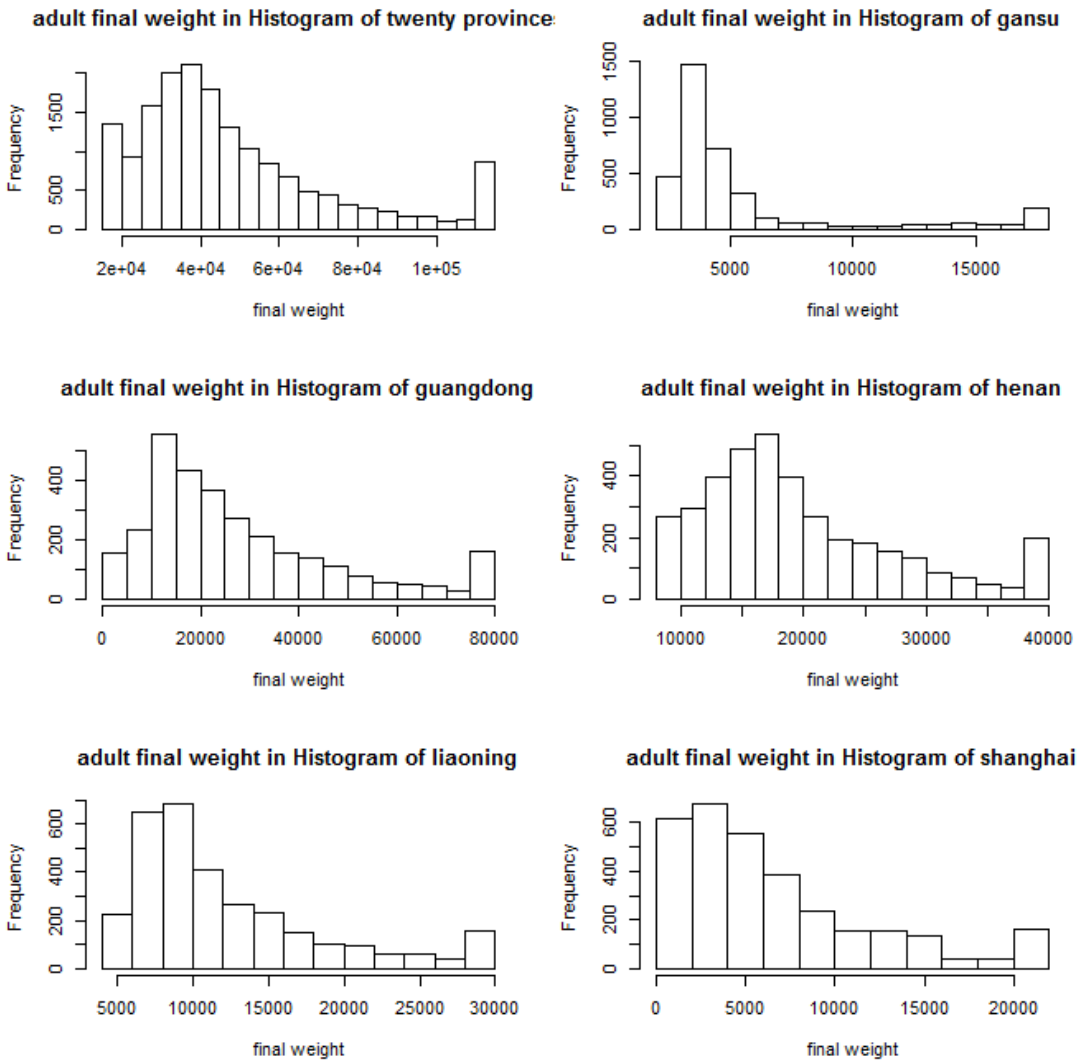

图 9.　成人问卷的 6 个抽样框最终的权数分布

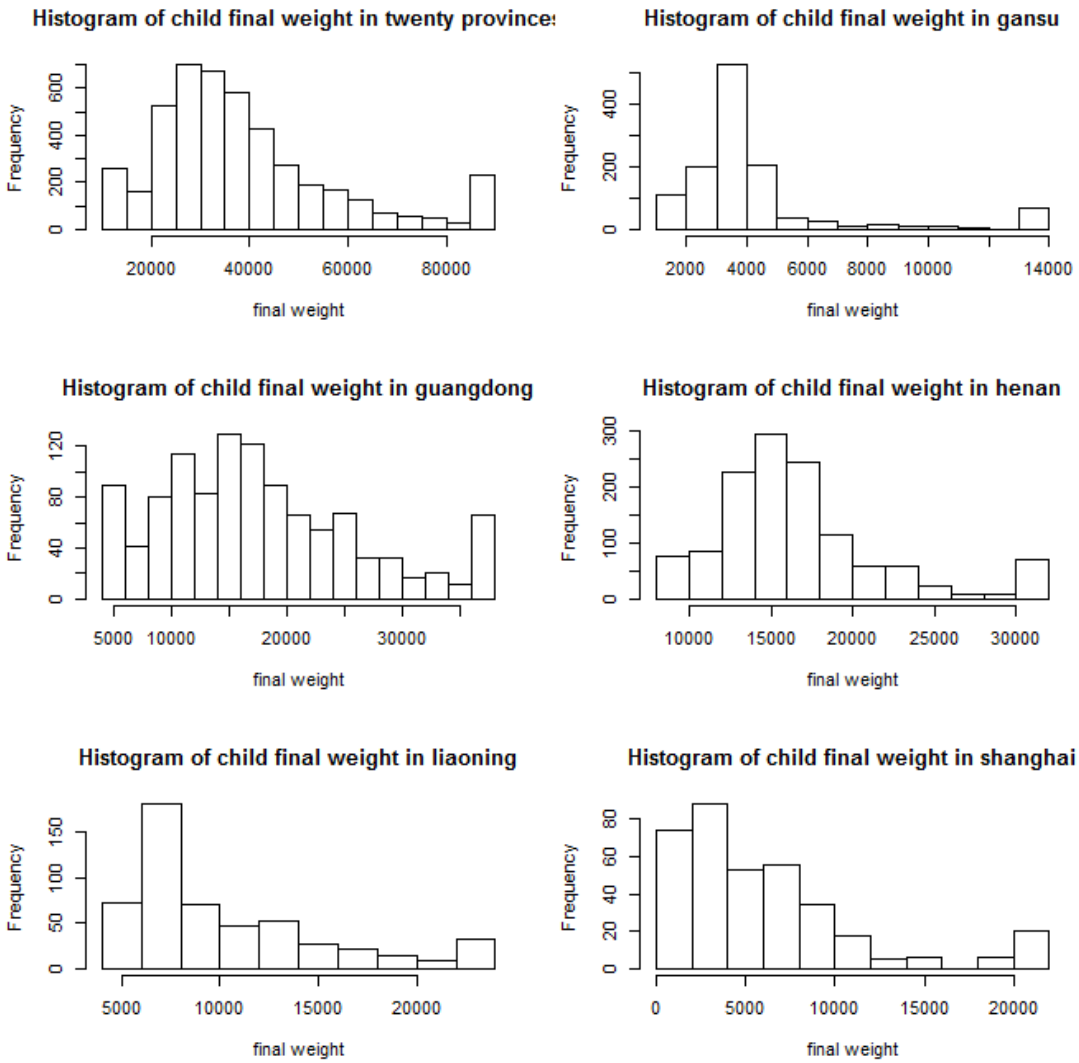

图 10.　少儿问卷的 6 个抽样框最终的权数分布

由六个抽样框的权数分布可以看出，经过抽样设计、无回答、事后分层和极值权数调整后，权数呈现偏态分布。本次调查的极值权数不够理想，最终的极值调整导致权数分布末端的频数较大，此为本报告的极值权数调整需要改进的地方，这将在今后的权数调整中进行改进。

2.2　个人问卷分城乡、年龄、性别的权数分布图

在家庭问卷层面，由于 CFPS 调查的家庭定义与国家统计局的家庭定义略有不同，我们未对家庭问卷进行事后分层调整。但为了保证精度，我们根据家户总量对家庭问卷的最终权数进行了略微调整，使权数和等于总体。

在个人问卷层面，即在各个子总体的成人和少儿问卷数据库中，为了避免结构性偏差，

提高估计精度，按照分城乡、年龄、性别进行事后分层调整。下图为 25 省市完全样本和 25 省市整合样本的分城乡、年龄、性别的权数分布图。

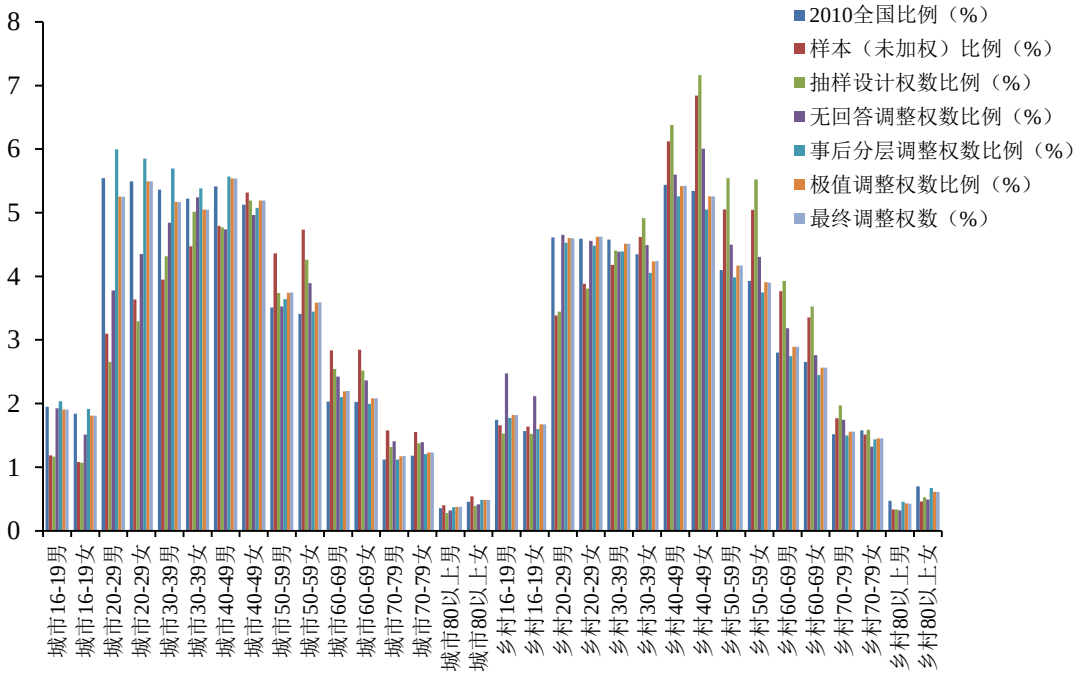

图 11.　成人问卷分城乡、年龄、性别的权数分布图（完全样本）

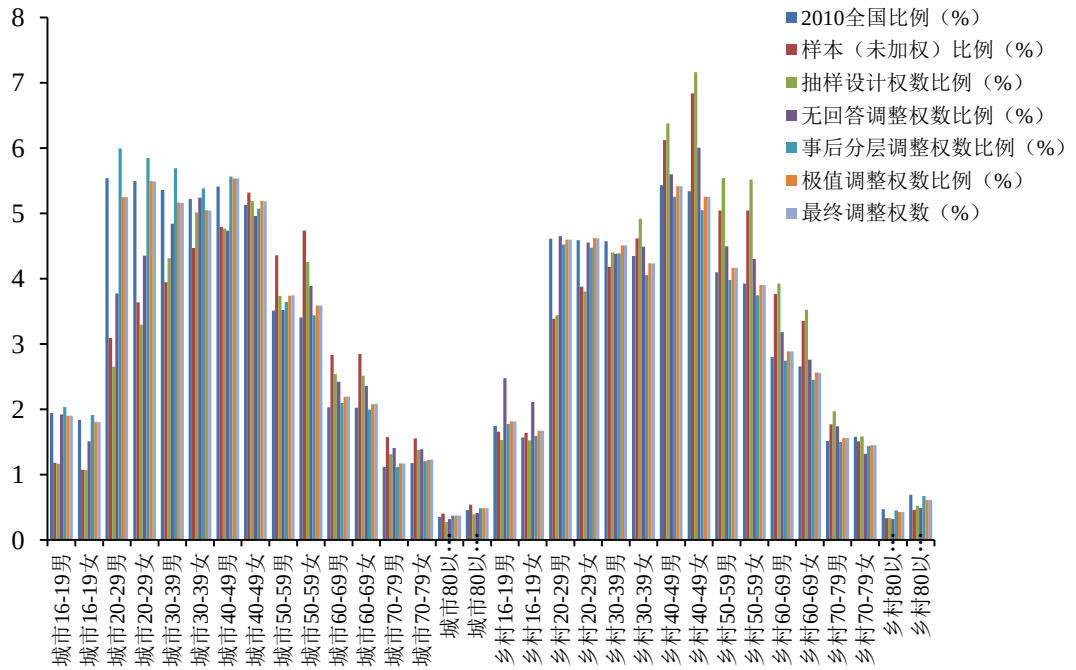

图 12.　成人问卷分城乡、年龄、性别的权数分布图（整合样本）

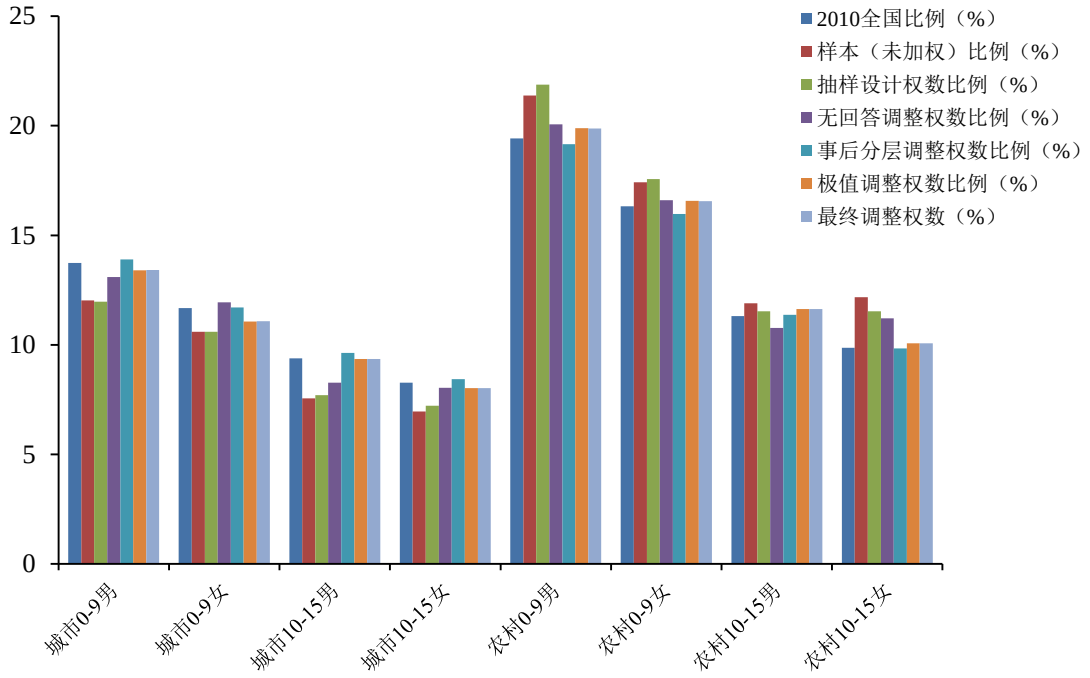

图 13.　少儿问卷分城乡、年龄、性别的权数分布图（完全样本）

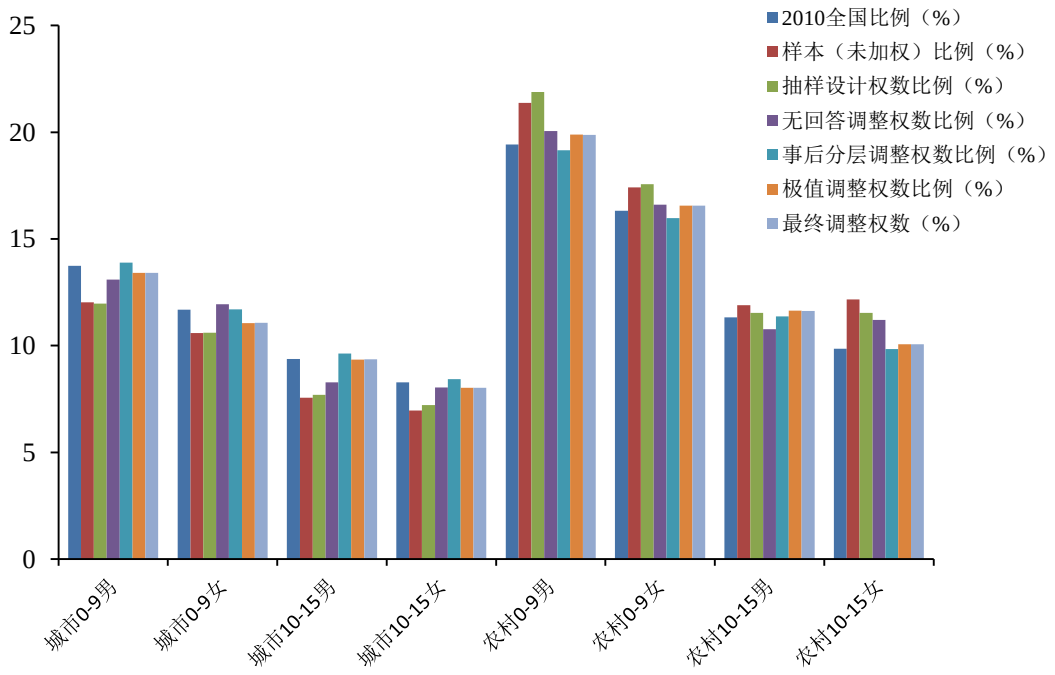

图 14.　少儿问卷分城乡、年龄、性别的权数分布图（整合样本）

由上述分城乡、年龄、性别的权数分布图可以看出，经过抽样设计、无回答和事后分层调整权数后，分城乡、年龄、性别的权数分布与总体一致。虽然经过极值处理后，权数的分布稍有变化，但是相对于未加权的样本分布依然更接近总体分布，提高了估计效率，减小了权数的差异，降低了方差，提高了估计精度。

3. CFPS 2010 权数数据库中的权数变量说明及其分布

3.1 家庭问卷权数数据库

- ✧ fid：家户样本代码
- ✧ subpopulation：抽样子总体
  - subpopulation =1：上海市子总体
  - subpopulation =2：辽宁省子总体
  - subpopulation =3：河南省子总体
  - subpopulation =4：甘肃省子总体
  - subpopulation =5：广东省子总体
  - subpopulation =6：其它省市子总体
- ✧ fswt_nat：家庭权重-全国完全样本
- ✧ fswt_res：家庭权重-全国整合样本

全国完全样本和整合样本的家庭问卷权数数据库中权数的均值、标准差和极差如下：

表 2. 家庭问卷权数的均值、标准差和极差

变量名	均值	标准差	极差
fswt_nat	31444.58	25909.2	106056.5
fswt_res	48164.46	47207.44	186088.5

3.2 成人问卷权数数据库

- ✧ pid：个人 id
- ✧ subpopulation：抽样子总体
  - subpopulation =1：上海市子总体
  - subpopulation =2：辽宁省子总体
  - subpopulation =3：河南省子总体
  - subpopulation =4：甘肃省子总体
  - subpopulation =5：广东省子总体
  - subpopulation =6：其它省市子总体
- ✧ rswt_nat：个人权重-全国完全样本
- ✧ rswt_res：个人权重-全国整合样本

全国完全样本与整合样本的成人问卷权数数据库的权数的均值、标准差和极差如下：

表 3. 成人问卷权数的均值、标准差和极差

变量名	均值	标准差	极差
rswt_nat	30815.46	25608.24	110219.2
rswt_res	36602.74	19781.57	76127.41

3.3 少儿问卷权数数据库

- ✧ pid：个人 id
- ✧ subpopulation：抽样子总体
  - ■ subpopulation =1：上海市子总体
  - ■ subpopulation =2：辽宁省子总体
  - ■ subpopulation =3：河南省子总体
  - ■ subpopulation =4：甘肃省子总体
  - ■ subpopulation =5：广东省子总体
  - ■ subpopulation =6：其它省市子总体
- ✧ rswt_nat：个人权重-全国完全样本
- ✧ rswt_res：个人权重-全国整合样本

全国完全样本与整合样本的少儿问卷权数数据库的权数的均值、标准差和极差如下：

表 4. 少儿问卷权数的均值、标准差和极差

变量名	均值	标准差	极差
rswt_nat	25282.28	19219.39	84961.92
rswt_res	29208.75	14879.85	58374.49