中国家庭追踪调查数据
权数使用常见问题

吕 萍

北京大学中国社会科学调查中心

2016年7月17日

CFPS权数使用常见问题

一、CFPS权数的含义、使用和注意事项

二、CFPS全国总样本和全国再抽样样本数据库的使用问题

三、CFPS系统性缺失变量的使用

四、CFPS的strata和PSU

五、软件操作

一、CFPS权数的含义、使用和注意事项

（一）目标总体

1、全国25个省市

2、省级推断单位（自我代表层）

- 上海市：**subpopulation** =1
- 辽宁省：**subpopulation** =2
- 河南省：**subpopulation** =3
- 甘肃省：**subpopulation** =4
- 广东省：**subpopulation** =5

3、再抽样总体（全国25省）（subsample=1）

（二）研究对象

- 中国（除香港、澳门、台湾、新疆维吾尔自治区、西藏自治区、青海省、内蒙古自治区、宁夏回族自治区和海南省等省区之外）的25个省市自治区的家庭户和家庭户中的所有满足调查条件的家庭成员。

- 上海市的家庭户和家庭户中的所有满足调查条件的家庭成员
- 辽宁省的家庭户和家庭户中的所有满足调查条件的家庭成员
- 河南省的家庭户和家庭户中的所有满足调查条件的家庭成员
- 甘肃省的家庭户和家庭户中的所有满足调查条件的家庭成员
- 广东省的家庭户和家庭户中的所有满足调查条件的家庭成员

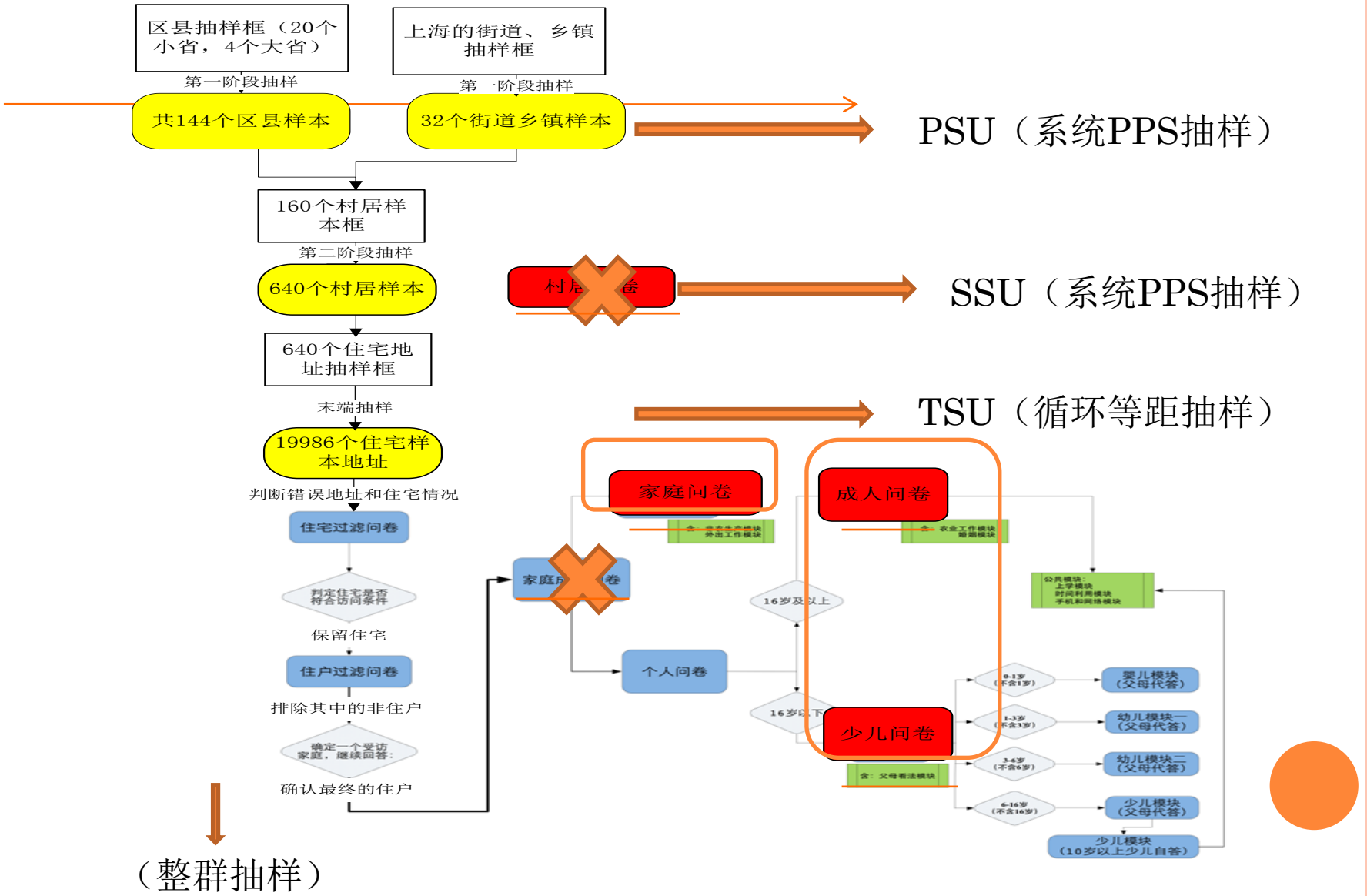
（二）研究问题

1、家庭总体情况（家庭收入、支出、资产等）

2、个人总体情况（教育、婚姻、工作、健康、认知等）

（四）抽样设计和抽样过程

采用多阶段不等概率的系统PPS整群抽样设计：



（五）问卷数据库和权数

- 家庭经济问卷数据库　研究家庭经济问题（家庭总体）
- 个人问卷数据库　研究个人问题（个人总体）

家庭关系数据库、村居问卷数据库作为辅助信息。

1、2010年初访调查数据库和权数（截面权数）

（1）全国25个省市（2010年数据）

- Fswt_Nat：家庭权重-全国样本
- Rswt_Nat：个人权重-全国样本

（2）省级推断单位（自我代表层）

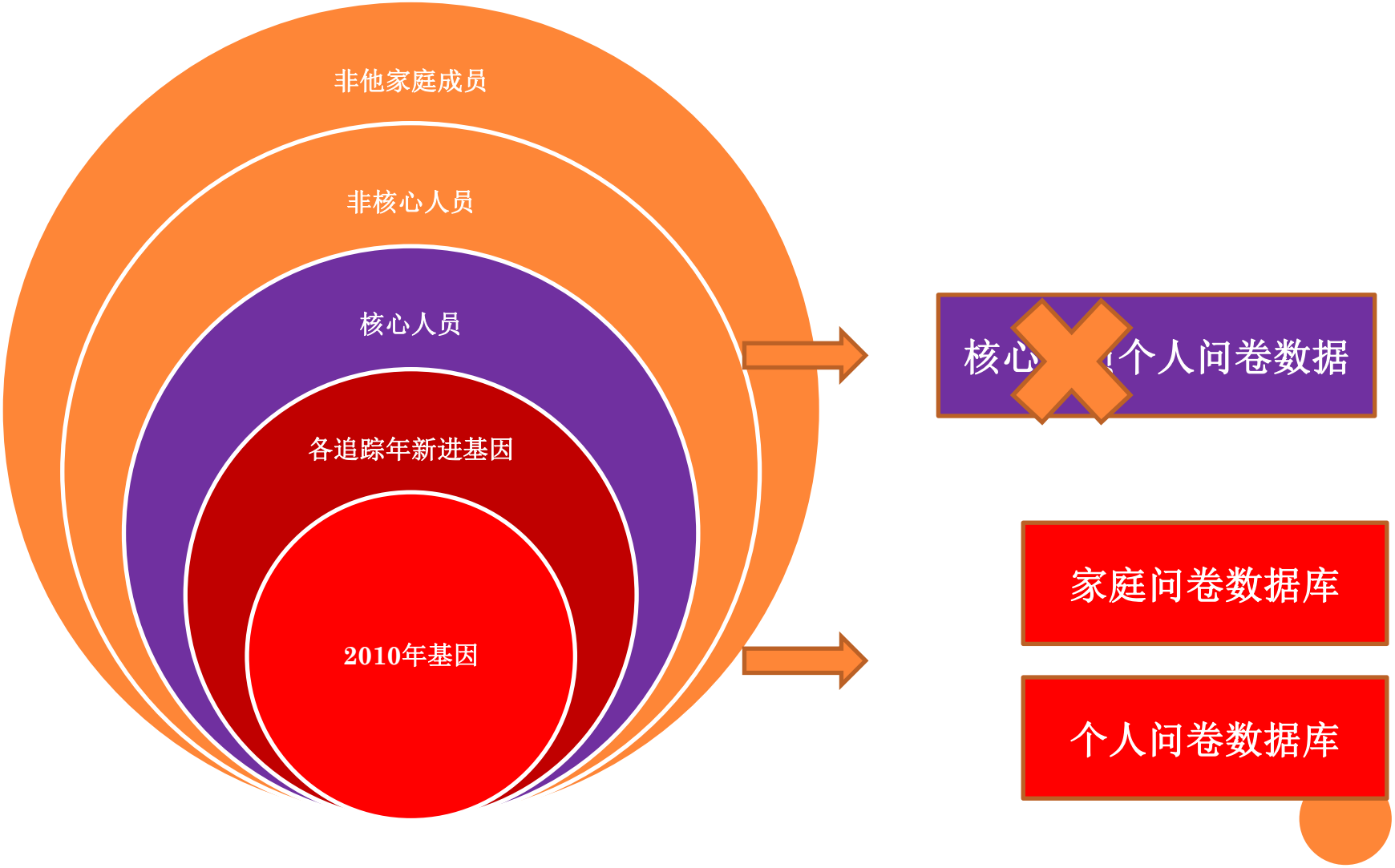
- 上海市：**subpopulation** =1
- 辽宁省：**subpopulation** =2
- 河南省：**subpopulation** =3
- 甘肃省：**subpopulation** =4
- 广东省：**subpopulation** =5

（3）再抽样总体（全国25省）（subsample=1）

- Fswt_Res　　家庭权重-全国再抽样样本
- Rswt_Res　　个人权重-全国再抽样样本

2、追踪数据库

（1）CFPS的人员类型和问卷数据库



核心成员问卷信息都可以作为辅助信息使用

追踪数据库分析

（1）横截面分析：分析调查年家庭经济和个人总体情况。

- 调查年数据（2010基因+各年新进基因）
- 调查年横截面权数。

（以往各年的数据、核心成员个人数据、家庭关系数据、村居数据作为辅助信息）

方法：可以将存在横截面权数的数据库提取出来使用。

（2）纵向分析：仅分析2010年的基因成员变动情况。

- 调查年及其以往各年数据（或调查年与基线调查数据）
- 调查年纵向权数。

方法：可以将存在纵向权数的数据库提取出来使用。

注：家庭层面纵向权数或将取消。

2012年追踪数据库（截面分析）

（1）全国25个省市（2012年数据）

- fswt_natcs12(家庭横截面权数:全国总样本)
- rswt_natcs12(个人横截面权数:全国总样本)

（2）省级推断单位（自我代表层）

- 上海市：**subpopulation** =1
- 辽宁省：**subpopulation** =2
- 河南省：**subpopulation** =3
- 甘肃省：**subpopulation** =4
- 广东省：**subpopulation** =5

（3）再抽样总体（全国25省）（subsample=1）

- fswt_rescs12(家庭横截面权数:全国再抽样样本)
- rswt_rescs12(个人横截面权数:全国再抽样样本)

2012年追踪数据库（纵向分析）

（1）全国25个省市（2012+2010数据）

- fswt_natpn1012 (家庭面板权数:全国总样本)
- rswt_natpn1012 (个人面板权数:全国总样本)

（2）省级推断单位（自我代表层）

- 上海市：**subpopulation** =1
- 辽宁省：**subpopulation** =2
- 河南省：**subpopulation** =3
- 甘肃省：**subpopulation** =4
- 广东省：**subpopulation** =5

（3）再抽样总体（全国25省）（subsample=1）

- fswt_natpn1012 (家庭面板权数:全国再抽样样本)
- rswt_natpn1012 (个人面板权数:全国再抽样样本)

2014年追踪数据库（截面分析）

（1）全国25个省市（2014数据）

- fswt_natcs14(CFPS2014家庭横截面权数：全国总样本)
- rswt_natcs14(CFPS2014个人横截面权数:全国总样本)

（2）省级推断单位（自我代表层）

- 上海市：**subpopulation** =1
- 辽宁省：**subpopulation** =2
- 河南省：**subpopulation** =3
- 甘肃省：**subpopulation** =4
- 广东省：**subpopulation** =5

（3）再抽样总体（全国25省）（subsample=1）

- fswt_rescs14(CFPS2014家庭横截面权数：2010年全国再抽样样本)
- rswt_rescs14(CFPS2014个人横截面权数：2010年全国再抽样样本)

2014年追踪数据库（纵向分析）

（1）全国25个省市（2014+2010 或 2014+2012+2010）

- fswt_natpn1014 (家庭面板权数:全国总样本)
- rswt_natpn1014 (个人面板权数:全国总样本)

（2）省级推断单位（自我代表层）

- 上海市：**subpopulation** =1
- 辽宁省：**subpopulation** =2
- 河南省：**subpopulation** =3
- 甘肃省：**subpopulation** =4
- 广东省：**subpopulation** =5

（3）再抽样总体（全国25省）（subsample=1）

- fswt_natpn1014 (家庭面板权数:全国再抽样样本)
- rswt_natpn1014 (个人面板权数:全国再抽样样本)

总体	指示	2010家庭权数	2010个人权数	2012家庭截面权数	2012家庭追踪权数	2012个人截面权数	2012个人追踪权数
甘肃省	subpopulation =4	Fswt_Nat	Rswt_Nat	fswt_natcs12	fswt_natpn1012	rswt_natcs12	rswt_natpn1012
广东省	subpopulation =4	Fswt_Nat	Rswt_Nat	fswt_natcs12	fswt_natpn1012	rswt_natcs12	rswt_natpn1012
河南省	subpopulation =3	Fswt_Nat	Rswt_Nat	fswt_natcs12	fswt_natpn1012	rswt_natcs12	rswt_natpn1012
辽宁省	subpopulation =2	Fswt_Nat	Rswt_Nat	fswt_natcs12	fswt_natpn1012	rswt_natcs12	rswt_natpn1012
上海市	subpopulation =1	Fswt_Nat	Rswt_Nat	fswt_natcs12	fswt_natpn1012	rswt_natcs12	rswt_natpn1012
全国	所有数据	Fswt_Nat	Rswt_Nat	fswt_natcs12	fswt_natpn1012	rswt_natcs12	rswt_natpn1012
全国再整合	subsample=1	Fswt_Res	Rswt_Res	fswt_rescs12	fswt_respn1012	rswt_rescs12	rswt_respn1012

（1）同一年各个总体的家庭情况：

CFPS 2010年家庭人均纯收入各个总体的均值和中位数的估计量

研究总体	样本量	均值	标准差	中位数	标准差
全国	14798	9842	517	5998	268
全国再抽样	9661	9780	564	5995	294
上海	1405	24861	1788	17352	1575
辽宁	1478	12321	1414	8757	1018
河南	1506	6370	504	4444	271
甘肃	1537	7606	1285	4101	410
广东	1394	10379	1641	6312	556

注：使用2010年各个总体的家庭人均纯收入和各个总体的截面权数

（2）不同年总体的比较：

CFPS 2010年和2012年家庭人均纯收入的截面数据估计量比较

	2010估计量(标准误）	2012估计量(标准误）	比率（2012/2010）
均值	9842(517)	11726（416）	1.19
25%的分位数	3059(148)	3441（165）	1.12
50%的分位数	5998(268)	8111（252）	1.35
75%的分位数	11222(603)	14937（466）	1.33

注：2010年的估计量使用2010年家庭人均纯收入和2010年权数； 2012年的估计量使用2012年家庭人均纯收入和2012年截面权数

（3）追踪年纵向分析（变动分析）

CFPS2010-2012年收入组纵向分析

2010年	2012年收入组			
收入组	0～25%	25%～50%	50%～75%	75%～100%
0～25%	45(1.23)	26(0.81)	18(0.86)	11(0.72)
25%～50%	31(1.29）	32(1.11)	25(1.17)	12(0.92)
50%～75%	21(1.10)	25(1.09)	31(1.06)	23(1.12)
75%～100%	12(1.14)	14(0.89)	22(1.20)	52(2.05)

注：使用2012的纵向权数、2010-2012年的收入组数据
缺失数据需要插补，少量缺失情况下可以不考虑

二、全国总样本和全国再抽样样本数据库的使用问题

全国总样本：

- 抽样设计总体
- 五个大省过抽样，必须加权使用

全国再抽样样本：

- 满足分析者及时分析的需求，CFPS2010调查后得到。
- 不加权使用
- 由于CFPS抽样的复杂性，同样给出了权数

判断方式：

- 抽样误差越小越好
- deff

例：

CFPS 2010年家庭人均纯收入的均值估计量

研究总体	样本量	均值	标准差
全国	14798	9842	517
全国再抽样	9661	9780	564

（1）标准误差：517<564

（2）deff全国=11.5，deff全国再抽样=8.65.

有效样本量分别是1292和1123

三、CFPS系统性缺失变量的使用

多种访问模式和问卷类型导致的变量缺失问题

- CFPS调查过程中面访、网访、电访多种调查方式
- 访问过程包含代答和自答两种类型

注：首先保证最终数据库只包含唯一的数据

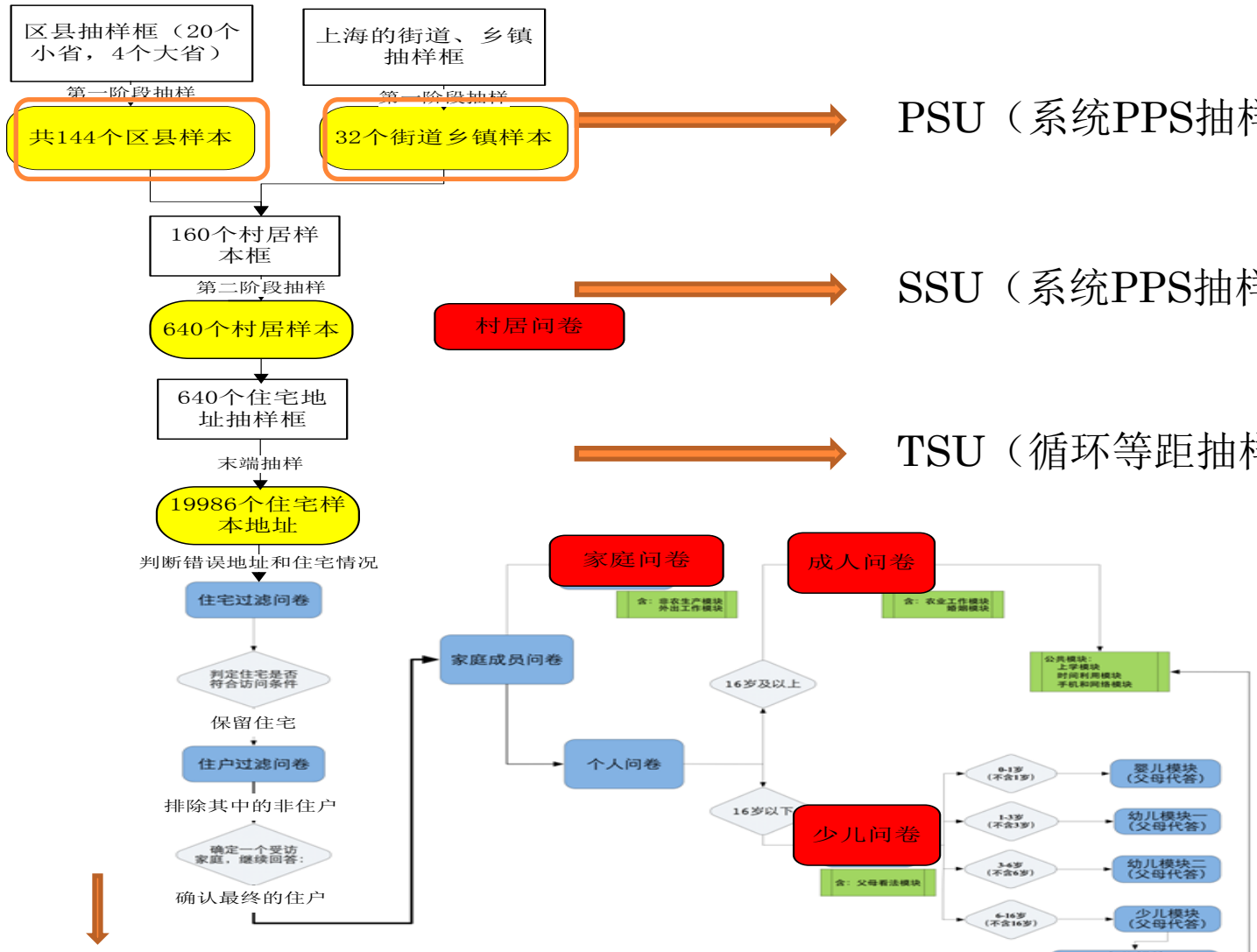
处理方式：

- 若缺失较少，进行缺失值处理，不影响权数使用
- 若缺失较多，在样本量足够的情况下，针对某些变量做加权调整
- 若缺失较多，样本量不足以分析，（1）考虑其他替代问题或删除变量（2）仅用于辅助信息，不进行估计

四、CFPS的STRATA和PSU

CFPS抽样设计：

在充分利用辅助信息对抽样总体进行有效排序的基础上的多阶段不等概率的系统PPS整群抽样设计。


区县抽样框（20个小省，4个大省）
上海的街道、乡镇抽样框
第一阶段抽样
共144个区县样本
32个街道乡镇样本
PSU（系统PPS抽样）
160个村居样本框
第二阶段抽样
640个村居样本
村居问卷
SSU（系统PPS抽样）
640个住宅地址抽样框
末端抽样
TSU（循环等距抽样）
19986个住宅样本地址
判断错误地址和住宅情况
住宅过滤问卷
判定住宅是否符合访问条件
保留住宅
住户过滤问卷
排除其中的非住户
确定一个受访家庭，继续回答：
确认最终的住户
家庭问卷
成人问卷
家庭成员问卷
个人问卷
16岁及以上
16岁以下
少儿问卷
公共模块：上学模块 时间利用模块 手机和网络模块
0-1岁（不含1岁）
婴儿模块（父母代答）
1-3岁（不含3岁）
幼儿模块一（父母代答）
3-6岁（不含6岁）
幼儿模块二（父母代答）
6-16岁（不含16岁）
少儿模块（父母代答）
少儿模块（10岁以上少儿自答）

（整群抽样）

CFPS各阶段抽样信息

省市	甘肃省	广东省	河南省	辽宁省	上海市	全国	全国再抽样
估计总体	自代表省	自代表省	自代表省	自代表省	自代表省	全国总体	全国总体
PSU	区县	区县	区县	区县	街道乡镇	区县或街道乡镇	区县或街道乡镇
抽样方法	系统PPS	系统PPS	系统PPS	系统PPS	系统PPS	系统PPS	系统PPS
初级单元数	16	16	16	16	32	176	106
SSU	村居	村居	村居	村居	村居	村居	村居
抽样方法	系统PPS	系统PPS	系统PPS	系统PPS	系统PPS	系统PPS	系统PPS
二级单元数	64	64	64	64	64	640	416
TSU	家户	家户	家户	家户	家户	家户	家户
抽样方法	系统抽样	系统抽样	系统抽样	系统抽样	系统抽样	系统抽样	系统抽样
三级有效数	1600	1600	1600	1600	1600	16000	10400

- CFPS是一个复杂抽样设计：

（1）参数估计量：使用最终权数可以得到无偏或近似无偏的估计量

（2）方差估计量：采用复杂方差计算方法（泰勒级数法和重抽样方法）

- 多阶段抽样设计在一定条件下可以使用第一阶段的层和群信息。

- CFPS目前的数据库：不考虑隐分层（高估方差）

（1）strata：6个抽样框（5个大省+1个小省）

（2）PSU：上海是街道乡镇；其他是区县（cluster）

The SURVEYMEANS Procedure

Data Summary

Number of Strata	**6**
Number of Clusters	**176**
Number of Observations	**14798**
Sum of Weights	**465316871**

Statistics

Variable	N	N Miss	Mean	Std Error
indinc_net	**13851**	**947**	**9841.924901**	**610.755101**

Data Summary

Number of Strata	88
Number of Clusters	176
Number of Observations	14798
Sum of Weights	465316871

Statistics

Variable	N	N Miss	Mean	Std Error
indinc_net	13851	947	9841.924901	521.136511

Replicate weight（重权数）：

- 为了保密性，数据库中不包含层和群信息。
- 数据库中使用weight_1-weight_n个权数变量，隐含层和群的信息。
- 只能使用重抽样方差估计方法

五、软件操作

数据分析软件

CFPS是一个复杂抽样设计，考虑复杂抽样设计方法。

SAS
Stata
R
SPSS
。。。

SAS软件

```
proc surveymeans   data=shuju;
strata stratum;
 cluster psu;
weight weight;
varname   var;
run;

proc surveymeans   data=shuju;
strata  subpopulation;
cluster  psu;
weight  Fswt_Nat;
varname indinc_net ;
run;
```

```
proc surveyfreq data=shuju;
strata stratum;
cluster psu;
weight weight;
Tables  var;
run;

proc surveyreq  data=shuju;
strata stratum;
cluster psu;
weight weight;
model .....;
run;

proc surveylogistic;
.....
```

STATA软件操作

---

svyset [pweight=*varname*], strata (*varname*), psu (*varname*)

svyset [pweight= Fswt_Nat], strata (subpopulation), psu (psu)

svy: mean indinc_net

svy: tab indinc_net

svy: reg

svy: logit

……

svyset psu [pweight] [, strata(varname) fpc(varname)] [|| ssu , [, strata(varname) fpc(varname)] ...

R软件操作

Design=svydesign(ids=~psu,strata=~stratum,fpc=~fpc,data=shuju,weight=~weight)

Svymean(~age,design)

Svytotal

Svyglm

.....

Design=svydesign(ids=~psu,strata=~subpopulation,data=shuju,weight=~ Fswt_Nat)

svymean(~indinc_net,design)

SPSS软件操作

- spss complex samples

谢谢！