China Family Panel Studies

CFPS

中国家庭追踪调查

技术报告系列：CFPS-24

系列编辑：谢宇　责任编辑：张聪

中国家庭追踪调查

2012 年综合变量：

生育子女数量和子女具体信息

穆峥　谢宇

2014.8.26

1、问题描述：子女数量和子女具体信息无法直接获取

生育行为是个人生命历程中的重要事件。生育水平、子女性别的组合模式、子女年龄差别的分布都极大地影响着个人的工作状态和整体生活模式。因此，生育子女数量和子女具体信息是相关分析中非常重要的一组综合变量。

不过，这些信息并没有直接作为 2012 年 CFPS 成人问卷中的访问题目出现在问卷中，我们只能从 2012 年 CFPS 家庭关系问卷提供的 T2 信息表中获取关于受访者各子女的具体信息。基于问卷设计形式的这种特点，且由于部分受访者在访问过程中并未按子女的年龄顺序回答各子女的信息，我们无法简单直观地从 T2 信息表中获取关于受访者子女数量和各子女具体信息的数据。因此，我们需要对家庭关系问卷中的信息进行相应处理，才能得到包括被访者生育子女数量、最大和最小子女在内信息的一组变量。

2、解决方案

2.1 子女信息库的生成

我们仅保留个人编码（pid）、个人年龄（tb1b_a_p）、个人出生年（tb1y_a_p）和子女的所有信息，由此生成一个子女信息数据库，以备重新排序和编号。

通过观察新生成的子女信息数据库，可以发现有 1181 位被访者的个人编码（pid）出现了两行观测值，而并非唯一存在的。导致这一现象的主要原因是这部分受访者的子女情况在 2010 年访问结束后发生了改变，具体变化情况可以总结为以下几个方面：其一，子女由于某些原因离开或者回归了家庭，导致这些子女在 2010 年和 2012 年两次调查所收集的个人家庭内部编号发生了变化； 其二，尽管 2010 年和 2012 年两次调查的子女个人家庭内部编号未产生变化，但是子女的某些其他个人信息发生了变化；其三，所有子女信息在 2010 年和 2012 年均为缺失值，家庭关系数据库中的其他信息发生变化；其四，2010 年子女信息为缺失值，2012 年生成子女信息，这可能是在 2010 年访问结束后出生的新生儿或 2010 年访问结束后新住进家庭中的子女。

除上述 1181 位受访者之外的其他受访者的个人编码（pid）均是唯一存在的，只具有一行观测值。这其中包括了完全没有子女（interviewyear_latest_c*=-8）、子女只在 2010 年被访（interviewyear_latest_c*=2010）以及子女只在 2012 年被访（interviewyear_latest_c*=2012）等情形。

综合以上分析，前述 1181 位个人编码（pid）不是唯一存在的受访者可按其子女的最近被访年份信息被划分为以下几类可能情形：

被访年份 1 / 被访年份 2	-8	2012
-8	（1）2010 年和 2012 年都没有孩子	（2）2010 年被访后有新生儿出生或者有子女新住进家庭中
2010	NA	（3）2010 年和 2012 年均对子女收集了信息

对于情形（1），无论保留哪一行观测值都是没有差别的；而对于情况（2）和情况（3），为得到 2012 年最新的子女情况信息，应该去掉 2010 年这一行的观测值（interviewyear_latest_c*=2010）。基于这样的处理之后，样本量被减去 1181，处理后的数据库中包括 53,895 个具有独特个人编码（pid）的受访者。

2.2 新生成子女年龄变量（chdage）和父母年龄变量（page）

在将样本中个人编号重复的观测值删除后，我们进一步考虑子女的具体信息。由于调查问卷中预留的 10 名子女并不一定按照年龄大小进行排序，所以各位子女在原问卷中的编号并不能代表子女的年龄长幼，无法直接基于调查问卷来确定被访者最小和最大子女的信息。为了对受访者的子女按照其年龄长幼重新进行编号，我们需要先基于家庭关系数据库中既有的子女编号（这一编号在进行数据格式转化时被命名为 rchdno）将子女信息数据的格式转化为长表，从而生成一个新的数据库。这个新数据库中各个变量包含了被访者 10 名潜在子女的全部信息：即对于每一项子女信息，每一个被访者都有 10 个观测值。然后，基于子女年龄在各被访者内部进行重新排序，并将数据重新转化为短表形式，从而实现对子女基于年龄长幼重新进行编号。

在基于子女年龄进行重新编号之前，我们需要进一步确认子女的正确年龄。通过观察我们发现子女的年龄存在以下多种情况：

出生年（tb1y_a_c） / 年龄（tb1b_a_c）	正常取值	缺失值（-8）	其他异常值（-9, -2, -1）	其他异常值（“.”）
正常取值	（1）	（2）	NA	NA
缺失值（-8）	（3）	（4）	NA	NA
其他异常值（-9, -2, -1）	NA	NA	（4）	NA
其他异常值（“.”）	NA	NA	NA	（4）

为了最大程度地确认或者还原子女的正确年龄，需要针对以上不同情形分别处理：对于情况（1）、（2）和（4），可以直接使用原有孩子年龄变量（tb1b_a_c）；但对于情况（3），则需

要通过出生年（tb1y_a_c）对年龄进行推断（tb1b_a_c）。基于这一过程，可以生成新的子女年龄变量（chdage）。

在生成子女年龄变量（chdage）之后，还需要根据父母与子女的年龄差别进行筛选，进一步调整子女年龄变量。在进行调整之前，我们先对父母年龄进行类似上述对子女年龄的处理：即对于父母年龄（tb1b_a_p）和出生年（tb1y_a_p）均为正常取值、缺失值或异常值的情况，直接保留父母年龄变量；对于父母年龄为缺失值或异常值，但父母出生年有正常取值的，则使用出生年计算父母年龄。经以上处理可以生成新的父母年龄变量 page。

我们将合理的父母与子女年龄差界定为 15-50 岁。对于父母与子女年龄差别过小或者过大的个人，我们均将相应的子女年龄处理为误填，并相应地将此类子女的各项信息（包括其年龄）处理为缺失值（-8），从而在此基础上进一步调整子女年龄变量（chdage）。但如果新生成的父母年龄变量（page）仍为缺失值的，我们则不对子女信息进行此步处理，而是遵从数据库中既有的子女信息。

2.3 子女数量变量（nchd）的生成及子女具体信息的确定

经过以上步骤的处理，生成了一个相对可靠的子女年龄变量 chdage。基于这一变量，可以在每个被访者内部对其子女按照年龄长幼进行重新排序。我们先依次按照父母个体编号（pid）、子女年龄（chdage）和问卷中所填的子女家庭内顺序（rchdno）重新对各子女信息进行排序，并在此排序基础上为每个子女生成一个新的家庭内部编号（cageord）；然后再基于 cageord 将数据重新转化为宽数据格式。此时，每一位被访者又拥有 10 组子女信息变量，并且这些变量后缀的编号代表了子女年龄的长幼之分：数字越大，则子女年龄越小。

实际上，绝大多数受访者的子女数量都远小于 10，致使这 10 组变量中（尤其是后缀编号较大的变量）有很多缺失值。因此，需要进一步确定子女数量和最大、最小年龄子女信息，以更有效率地建立子女数据库。首先，为确认子女数量，我们需要确定将哪些子女信息作为基本信息，即如果此项信息缺失，则认为此子女是不存在的。但具体将哪些信息作为子女信息取决于我们对信息可靠性的信任程度。因此，我们根据依据信息的多少生成了三个版本的子女数量变量：

1. 依靠所有可用子女信息，这包括：孩子*家庭内部编码（code_a_c*），2012 年孩子*样本编号（pid12_c*），2010 年孩子*样本代码（pid10_c*），孩子*属相（tb1a_a_c*），2012 年孩子*年龄（tb1b_a_c*），孩子*出生（年）（tb1y_a_c*），孩子*出生（月）（tb1m_a_c*），孩子*

性别（tb2_a_c*），2012 年孩子*是否健在（alive_a_c*），2012 年孩子*去世的原因（deathreason_c*），2012 年孩子*婚姻状况（tb3_a12_c*），2012 年孩子*最高学历（tb4_a12_c*），2012 年孩子*户口类型（qa301_a12_c*），2012 年孩子*户口所在地（qa302_a12_c*），2012 年孩子*是否在家住（tb6_a12_c*），2012 年孩子*离家的原因（tb601_a12_c*），2012 年离家人（孩子*）的居住区域（outpers_where12_c*），2012 年离家人（孩子*）的省国际码（tb602acode_a12_c*），2012 年孩子*离家的年份（leavingtime_y_c*），2012 年孩子*离家的月份（leavingtime_m_c*），2012 年孩子*是否与该家庭同灶吃饭（co_a12_c*），以及孩子*最近一次访问时间（inerviewyear_latest_c*）。只要其中有一项信息为有效数值，则认为这一子女是存在的。由此生成的子女数量变量为 nchd1（基于 n=53,895 的样本，nchd1 的均值为 1.30，标准差为 1.39，最小值为 0，最大值为 10）。

2. 将子女年龄（chdage*）和性别（gender*）作为子女基本信息，即只要这两项信息中有一项为有效数值，则认为这一子女是存在的。由此生成的子女数量变量为 nchd2（基于 n=53,895 的样本，nchd2 的均值为 1.30，标准差为 1.38，最小值为 0，最大值为 10）。生成这一变量的逻辑在于，如果被访者连子女年龄和性别这样的基本信息都无法提供，我们对其提供的其他子女信息也要保持审慎态度，所以我们在计算子女数量时可以采取适当保守的处理方式。

3. 仅将子女年龄（chdage*）当作子女基本信息，即只要子女年龄没有合理取值，我们则认为此子女是不存在的。按照此规则生成的子女数量变量为 nchd3（基于 n=53,895 的样本，nchd3 的均值为 1.27，标准差为 1.34，最小值为 0，最大值为 10）。这一变量主要是为了进一步生成各次序子女的具体信息。具体而言，我们基于子女信息变量名后缀编号为 1 的变量得到最大年龄子女信息，并通过子女信息变量名后缀编号为已生成家庭子女数量（nchd）的变量为各个家庭生成一套最小年龄子女信息变量（均以-yog 为后缀。值得注意的是，既有数据库中所显示的最大年龄子女信息和最小年龄子女信息都是基于具有有效子女年龄（chdage*）的子女进行排序的，因此，它们与以上生成的第三种子女数量（nchd3）相呼应，子女年龄缺失或者无效的将无法进入排序。最后，为了更为方便地对成人的生育行为数据进行使用或者分析，可以将这个子女信息数据库通过个人编码（pid）合并至成人数据库。

附录：

附录表 1：子女数量版本 1（nchd1）分布
（n=53,895）

数量（个）	频次	比例
0	19,948	37.01
1	13,052	24.22
2	12,060	22.38
3	4,988	9.26
4	2,177	4.04
5	957	1.78
6	467	0.87
7	187	0.35
8	26	0.05
9	18	0.03
10	15	0.03

附录表 2：子女数量版本 2（nchd2）分布
（n=53,895）

数量（个）	频次	比例
0	20,177	37.44
1	12,900	23.94
2	12,008	22.28
3	4,979	9.24
4	2,171	4.03
5	959	1.78
6	463	0.86
7	206	0.38
8	24	0.04
9	4	0.01
10	4	0.01

附录表 3：子女数量版本 3（nchd3）分布
（n=53,895）

数量（个）	频次	比例
0	20,276	37.62
1	13,090	24.29
2	12,071	22.40
3	4,939	9.16
4	2,093	3.88
5	888	1.65
6	384	0.71
7	144	0.27
8	6	0.01
9	2	0.00
10	2	0.00

附录表 4：父母亲年龄（page）与子女数量版本 1（nchd1）描述

父母年龄	均值	标准差	频次
15	0.00	0.06	594
16	0.00	0.09	635
17	0.01	0.10	748
18	0.02	0.17	682
19	0.03	0.18	686
20	0.06	0.26	757
21	0.13	0.37	825
22	0.20	0.46	1021
23	0.30	0.56	1056
24	0.43	0.62	1021
25	0.51	0.68	1084
26	0.61	0.71	903
27	0.74	0.77	825
28	0.90	0.87	737
29	1.01	0.80	756
30	1.05	0.84	792
31	1.13	0.79	676
32	1.33	0.86	643
33	1.33	0.80	696
34	1.40	0.79	710
35	1.48	0.80	587
36	1.47	0.84	707
37	1.48	0.80	685
38	1.65	0.87	716
39	1.65	0.86	847
40	1.62	0.82	823
41	1.63	0.81	895
42	1.69	0.85	948
43	1.69	0.81	813
44	1.73	0.86	1023
45	1.80	0.86	755
46	1.76	0.83	880
47	1.88	0.99	888
48	1.81	0.95	860
49	1.86	0.96	996
50	1.93	1.00	869
51	1.94	0.93	397
52	1.86	0.95	565
53	1.88	0.99	497
54	1.87	1.02	663
55	1.96	1.00	783
56	1.95	0.94	689
57	1.92	0.95	741

58	2.03	1.07	848
59	2.20	1.08	716
60	2.18	1.12	755
61	2.24	1.09	559
62	2.31	1.02	568
63	2.41	1.17	616
64	2.48	1.17	493
65	2.64	1.24	508
66	2.65	1.29	480
67	2.79	1.37	431
68	3.00	1.48	424
69	3.06	1.54	337
70	3.30	1.58	359
71	3.23	1.58	327
72	3.20	1.58	327
73	3.49	1.67	259
74	3.48	1.72	309
75	3.43	1.69	264
76	3.61	1.69	269
77	3.62	1.86	259
78	3.50	1.92	229
79	3.77	1.79	228
80	3.43	1.96	183
81	3.70	2.05	148
82	3.53	2.02	137
83	3.54	2.08	109
84	3.28	1.94	123
85	3.75	2.29	91
86	3.43	2.14	69
87	3.68	2.14	65
88	3.22	2.23	45
89	3.63	2.14	30
90	3.56	1.94	34
91	2.90	1.89	31
92	4.33	2.19	30
93	3.13	2.01	24
94	4.56	1.88	9
95	3.00	2.00	5
96	2.60	1.67	5
97	3.33	2.08	3
98	1.33	1.53	3
99	5.00	3.61	3
101	2.00	1.41	2
103	2.00	0.00	1
104	3.00	0.00	1

112	5.00	0.00	1
总和	1.58	1.38	44,191

附录表 5：父母亲年龄（page）与子女数量版本 2（nchd2）描述

父母年龄	均值	标准差	频次
15	0.00	0.06	594
16	0.00	0.09	635
17	0.01	0.09	748
18	0.02	0.15	682
19	0.03	0.17	686
20	0.05	0.25	757
21	0.12	0.34	825
22	0.19	0.45	1021
23	0.29	0.55	1056
24	0.41	0.61	1021
25	0.50	0.67	1084
26	0.59	0.71	903
27	0.71	0.77	825
28	0.86	0.80	737
29	0.99	0.81	756
30	1.03	0.77	792
31	1.12	0.79	676
32	1.31	0.87	643
33	1.32	0.80	696
34	1.40	0.79	710
35	1.47	0.80	587
36	1.47	0.83	707
37	1.48	0.80	685
38	1.64	0.87	716
39	1.64	0.86	847
40	1.62	0.82	823
41	1.63	0.81	895
42	1.69	0.85	948
43	1.69	0.81	813
44	1.73	0.86	1023
45	1.79	0.86	755
46	1.76	0.84	880
47	1.87	0.99	888
48	1.81	0.93	860
49	1.86	0.96	996
50	1.93	1.00	869
51	1.94	0.94	397
52	1.86	0.95	565
53	1.88	1.00	497
54	1.85	0.97	663
55	1.95	0.96	783
56	1.95	0.94	689
57	1.92	0.95	741

58	2.03	1.07	848
59	2.19	1.08	716
60	2.17	1.12	755
61	2.24	1.09	559
62	2.31	1.02	568
63	2.41	1.17	616
64	2.48	1.17	493
65	2.64	1.24	508
66	2.64	1.28	480
67	2.79	1.37	431
68	2.99	1.48	424
69	3.03	1.50	337
70	3.28	1.57	359
71	3.23	1.58	327
72	3.19	1.56	327
73	3.49	1.67	259
74	3.46	1.69	309
75	3.43	1.69	264
76	3.61	1.69	269
77	3.61	1.85	259
78	3.48	1.90	229
79	3.75	1.79	228
80	3.40	1.91	183
81	3.70	2.03	148
82	3.50	1.96	137
83	3.52	2.04	109
84	3.28	1.92	123
85	3.73	2.24	91
86	3.38	1.99	69
87	3.63	2.05	65
88	3.22	2.23	45
89	3.57	1.96	30
90	3.56	1.94	34
91	2.90	1.89	31
92	4.07	1.95	30
93	3.00	1.82	24
94	4.44	1.67	9
95	3.00	2.00	5
96	2.60	1.67	5
97	3.33	2.08	3
98	1.33	1.53	3
99	5.00	3.61	3
101	2.00	1.41	2
103	2.00	0.00	1
104	3.00	0.00	1

112	5.00	0.00	1
总和	1.57	1.37	44,191

附录表 6：父母亲年龄（page）与子女数量版本 3（nchd3）描述

父母年龄	均值	标准差	频次
15	0.00	0.00	594
16	0.00	0.04	635
17	0.01	0.09	748
18	0.02	0.13	682
19	0.03	0.17	686
20	0.05	0.25	757
21	0.11	0.33	825
22	0.19	0.45	1021
23	0.29	0.54	1056
24	0.41	0.60	1021
25	0.50	0.67	1084
26	0.59	0.71	903
27	0.71	0.77	825
28	0.86	0.80	737
29	0.99	0.81	756
30	1.03	0.77	792
31	1.12	0.79	676
32	1.30	0.87	643
33	1.32	0.80	696
34	1.39	0.78	710
35	1.47	0.80	587
36	1.46	0.83	707
37	1.47	0.80	685
38	1.64	0.87	716
39	1.64	0.86	847
40	1.61	0.81	823
41	1.62	0.81	895
42	1.69	0.85	948
43	1.68	0.81	813
44	1.73	0.86	1023
45	1.79	0.86	755
46	1.75	0.83	880
47	1.87	0.99	888
48	1.79	0.92	860
49	1.85	0.97	996
50	1.91	1.00	869
51	1.92	0.93	397
52	1.85	0.95	565
53	1.86	0.99	497
54	1.82	0.96	663
55	1.93	0.96	783
56	1.93	0.94	689
57	1.91	0.95	741
58	2.00	1.07	848
59	2.17	1.08	716

60	2.15	1.11	755
61	2.18	1.08	559
62	2.29	1.02	568
63	2.38	1.17	616
64	2.44	1.16	493
65	2.58	1.24	508
66	2.58	1.24	480
67	2.74	1.38	431
68	2.88	1.46	424
69	2.90	1.55	337
70	3.14	1.53	359
71	3.12	1.56	327
72	3.01	1.51	327
73	3.25	1.59	259
74	3.28	1.68	309
75	3.18	1.65	264
76	3.39	1.63	269
77	3.40	1.86	259
78	3.19	1.88	229
79	3.36	1.80	228
80	3.13	1.78	183
81	3.38	2.02	148
82	3.18	1.83	137
83	3.20	1.96	109
84	2.99	1.82	123
85	3.14	2.12	91
86	2.93	1.87	69
87	3.14	1.96	65
88	2.73	1.85	45
89	2.87	1.76	30
90	3.12	1.89	34
91	2.55	1.59	31
92	3.43	1.74	30
93	2.58	1.69	24
94	3.78	1.48	9
95	2.80	2.28	5
96	2.60	1.67	5
97	3.00	2.00	3
98	1.00	1.00	3
99	4.00	3.00	3
101	1.00	0.00	2
103	2.00	0.00	1
104	3.00	0.00	1
112	0.00	0.00	1
总和	1.54	1.33	44,191

附录表 7：最小子女年龄（cageyog）分布（n=33,619）

数量（个）	频次	比例
0	699	2.08
1	1,200	3.57
2	1,095	3.26
3	1,175	3.5
4	917	2.73
5	853	2.54
6	783	2.33
7	759	2.26
8	782	2.33
9	596	1.77
10	628	1.87
11	587	1.75
12	618	1.84
13	716	2.13
14	625	1.86
15	594	1.77
16	622	1.85
17	782	2.33
18	718	2.14
19	671	2
20	789	2.35
21	822	2.45
22	984	2.93
23	1,073	3.19
24	911	2.71
25	1,069	3.18
26	769	2.29
27	684	2.03
28	610	1.81
29	661	1.97
30	878	2.61
31	706	2.1
32	624	1.86
33	746	2.22
34	682	2.03
35	579	1.72
36	629	1.87
37	523	1.56
38	495	1.47
39	527	1.57
40	544	1.62
41	458	1.36
42	407	1.21

43	312	0.93
44	238	0.71
45	187	0.56
46	173	0.51
47	164	0.49
48	141	0.42
49	160	0.48
50	148	0.44
51	56	0.17
52	45	0.13
53	42	0.12
54	56	0.17
55	45	0.13
56	44	0.13
57	51	0.15
58	32	0.1
59	22	0.07
60	24	0.07
61	16	0.05
62	14	0.04
63	10	0.03
64	14	0.04
65	9	0.03
66	8	0.02
67	6	0.02
68	5	0.01
69	1	0
70	1	0
71	2	0.01
72	2	0.01
85	1	0