China Family Panel Studies

CFPS 中国家庭动态跟踪调查

技术报告系列：CFPS-13

系列编辑：谢宇　责任编辑：胡婧炜

中国家庭动态跟踪调查
2010 年综合变量（3）：年龄、婚姻最佳变量

张春泥　许琪　孙妍

2012.12.20

CFPS 2010 个人问卷设计的一大优势是采用回顾（retrospective）的方式收集了受访者生命历程中重要事件（如教育、婚姻）的起止时间，这一设计满足了社会科学领域对事件史数据日益增长的需求，尤其是 CFPS 的婚姻史模块，详尽询问了受访者经历各种婚姻变化的时间，弥补了同类数据在这一领域的空白，为专门研究婚姻及婚姻与其他事件关系的研究者提供了数据。但是，这一设计对记录事件发生时间的准确性提出了要求。由于 2010 年是以回顾的方式收集婚姻事件的时间，所以受访者的记忆偏差导致数据中出现了时间的逻辑顺序不合理、同一信息不同来源填答不一致等问题。在数据采集结束后，CFPS 数据团队对数据填答的质量进行了评估，并对个别常用的时间变量做了后期更正。为了保持原始填答，我们通过额外生成最佳变量（X_best）的方式来保存更改后的取值。由于时间和人力有限，我们仅就成人库的出生年、初婚结婚时间、初婚配偶出生年份、婚姻状态生成了最佳变量。在 2012 年的调查中，我们还会对这些重要变量的已有填答进行确认，并会在日后的数据发布中更新最佳变量。

1. 问题描述

1.1 受访者的出生年在不同问卷中填答不一致

成人受访者的出生年信息有三个来源：一是家庭成员问卷以代答的方式询问了家中每一位成员的出生年（tb1y_a_p），若受访者不记得具体出生年，则用调查年与填报的年龄（tb1b_a_p）相减，或根据填报的属相（tb1a_a_p）计算对应的出生年；二是受访者在填答成人问卷时填报的出生年（qa1y）；三是该受访者的配偶在填答配偶出生年时填报的受访者的出生年（qe606y 或 qe210y）。理论上，对同一个受访者的出生年，这三个来源的填答应该一致，但由于家庭成员问卷、受访者的个人问卷、受访者配偶的个人问卷是由不同的人填答的，CAPI 系统未设计对不同问卷的填答的一致性及逻辑性检查，因此，出生年变量仍存在缺失、各来源填答不一致（见表 1）、填答的出生年与其他生命事件（如婚姻）在时间上的逻辑关系不合理等问题。

表 1. 受访者出生年在不同问卷中填答不一致的样本数（N= 33,600）

	频数
个人问卷自答出生年不等于家庭成员问卷代答的出生年	983
其中，填答相差在 3 岁以上	249
个人问卷自答出生年不等于初婚/现任配偶填答的配偶出生年	1,645
其中，填答相差在 3 岁以上	511

注：表中的统计不包括相关变量含缺失值的样本，下同。

1.2 受访者本人填答的初婚年份与初婚配偶填答的初婚年份不一致

如果受访者及其配偶均为初婚，且两人均有个人问卷，在将受访者与其配偶的数据匹配后，二人各自填答的初婚的年份（qe605y）应一致。但我们发现，仍有相当数量的初婚夫妻填答的初婚年份不一致。不一致的原因有两种：一、由于 T2 表代码错误所造成的夫妻匹配错误，即 T2 表中填答的受访者配偶并非其配偶，而是其他家庭成员；二、由于记忆错误等原因导致的信息填答错误（见表 2）。对于第一种错误，我们已通过对家庭关系数据库的手工清理，纠正了错误的代码；[1] 对第二种错误，我们需要在两个来源的初婚时间中做出判断。

表 2. 受访者本人填答的初婚年份与初婚配偶填答的初婚年份不一致的样本数

	频数
初婚夫妻双方填答的初婚年份不一致	3,434
其中，填答相差在 3 岁以上	792
初婚夫妻双方均填答了初婚年份的样本数	21,720

注：此处的统计基于 T1-T3 表清理后的匹配样本

1.3 根据出生年份和初婚年份计算的初婚年龄不合理

当受访者的出生年份、初婚年份均不为缺失值时，研究者可以通过将两个年份相减得到受访者的初婚年龄。由于《婚姻法》对适婚的初始年龄有严格的规定，初婚年龄取值过小（如小于 16 岁，见表 3）可能是由于初婚年份或出生年份的填答错误造成，而非事实上的早婚。我们需要检查初婚年龄过小的案例，如发现明显错误，则要做出适当的修正。

[1] 具体过程详见许琪、张春泥、孙玉环、胡婧炜、吕萍，2012，《中国家庭动态跟踪调查 2010 年家庭关系数据库清理（CFPS-7）》。

表 3. 根据出生年份和初婚年份计算的初婚年龄不合理的样本数

初婚年龄	用个人问卷自答的出生年计算	用家庭成员问卷代答的出生年计算	用配偶个人问卷的配偶出生年计算
17 岁	609	614	369
16 岁	335	336	206
14-15 岁	243	248	150
13 岁以下（含负数）	211	214	191

1.4 婚姻状态不一致

在核查婚姻时间和配偶出生年份时，发现少部分受访者在个人库里的婚姻状态与其在家庭关系库中的婚姻状态不一致，或者这些受访者在 T2 表中有健在的配偶，但婚姻状态却填报为未婚、离异或丧偶。造成这一问题的原因包括：一、个人问卷和家庭成员问卷是由不同的人填答的，其中有填答者的信息填答有误；二、T2 表匹配错误，把其他家庭成员的名字填在了原本没有配偶的受访者的配偶位置上；三、个别访员或受访者对 T2 表填答规则的理解有偏差。按照设计，离异和丧偶的受访者的 T2 表配偶位置上应该留空，但有的填答人仍把已离异或已去世的配偶名字填在了 T2 表配偶位置上。

2. 更正方案：最佳变量的生成规则

由于 CFPS 2010 对上述信息的填答是多来源的，且可以参照其他与时间相关的变量来推断事件先后的逻辑关系，因此，我们能够对一部分出生年份、初婚年份、婚姻状态的填答错误做出判断并更正。我们后期更正的基本原则是：在保留原始填答的基础上根据多方面信息额外生成取值合理的最佳变量。这里的“最佳”并非是唯一正确的取值，而是建立在已有填答的基础上，结合多个信息来源，根据逻辑关系判断出来的目前最合理取值。研究者可以自行取舍是使用原始填答还是使用最佳变量。

针对以上提及的几个问题，我们生成了 4 个最佳变量。最佳变量的命名均为原变量名加“_best”后缀。这 4 个最佳变量是：qa1y_best，qe605y_best，qe606y_best，qe1_best。接下来，本报告将介绍各个最佳变量的生成规则。

2.1 qa1y_best：您的出生日期修正（年）

最佳的出生年的更正分为手工更正和程序更正两个阶段。在手工更正阶段，数据团队对存在本人与父母年龄差不合理、结婚年龄不合理等问题的样本进行了核查，如果确认这些问题是由于受访者的出生年（qa1y）填报错误造成的，则将修改后的出生年记录在 qa1y_best 中。如果受访者有多个不同来源的出生年，则选取的规则是：首选 2011 年追访个人问卷自答的出生年，如果该出生年为缺失或取值不合理，则依次对照 2010 年家庭成员问卷初访的代答出生年、2011 年家庭成员问卷代答的出生年、受访者配偶所填答的配偶出生年三个变量，选择其中最为合理的一个作为 qa1y_best 的取值。手工更正阶段仅清理了通过核查匹配或逻辑错误连带发现的出生年错误的样本，并未对所有存在出生年填答不一致的情况作专项核查及更正，后者通过程序更正完成。程序生成 qa1y_best 的规则是：先比较经手工更正过的个人问卷自答出生年 qa1y_best、家庭成员问卷代答的出生年 tb1y_a_p（或调查年 cyear-年龄 tb1b_a_p）和配偶填答的该受访者出生年 qe606y（初婚）或 qe210y（非初婚）三个变量，如三个变量中两个及以上的取值一致，则用该取值生成 qa1y_best。如果三个变量的取值均不一致（含缺失），我们先利用初婚年（如有）与出生年相减，计算初婚年龄。假定正常的初婚年龄介于 18-40 岁，如果用某一来源的出生年计算得到的初婚年龄小于或大于这个区间，则认为该出生年有可能不合理，选用更为合理的出生年来作为 qa1y_best。如果三个取值不一致的出生年有两个以上均合理，则优先使用个人问卷填答的出生年 qa1y 作为 qa1y_best 的取值，其次使用家庭代答出生年，最后使用配偶填答的配偶出生年。如上述来源的出生年均缺失或均不合理，则 qa1y_best 赋值为缺失。

2.2 qe605y_best：您与初婚配偶的结婚日期（年）修订
qe606y_best：您初婚配偶的出生年月是（年）修订

由于相当数量的被访家户中，受访者与其配偶均生成了个人问卷，我们通过匹配夫妻样本，利用双方均为初婚的夫妻均填答同一婚姻信息的特点，来对婚姻模块中初婚结婚年（qe605y）和初婚配偶的出生年（qe606y）两个变量进行校验，并做了后期修正，修正后的变量名为 qe605y_best 和 qe606y_best。

对初婚结婚年变量校验时所使用到的逻辑条件包括：⑴ 用初婚结婚年份和最佳出生年份计算得出的初婚结婚年龄须在 18-40 岁之间；⑵ 初婚生育的第一个孩子的出生年份不应

早于初婚结婚年份。在这两个条件下，我们评估了初婚夫妻双方填答的初婚结婚年份的合理性。在双方填答一致的情况下，qe605y_best 取值等于原始取值。在双方填答不一致的情况下，我们选取满足逻辑条件的一方填答作为最佳初婚结婚年份（qe605y_best）的值，当两条逻辑条件有冲突时，优先以初婚年龄区间为判断标准。若双方填答不一致且均不满足逻辑条件或双方填答不一致但均满足逻辑条件，则在最佳变量上保留本人填答的初婚结婚年份。对本人为初婚、配偶为再婚，则是根据类似的逻辑条件检查本人初婚年份和配偶目前婚姻的结婚年份（qe210y）之间的填答出入，并按照各自填答的合理性来修改。对本人目前不在婚（丧偶、离婚、同居）、本人和配偶均不为初婚、初婚配偶该信息缺失的情况，在 qe605y_best 中保留原本人填答的值。

对初婚配偶出生年份的校验，所使用到的逻辑条件包括：⑴ 初婚夫妻双方，受访者本人填答的配偶出生年份应该与配偶自己填答出生年份一致；类似地，一方为初婚，另一方为再婚时，初婚一方填答的初婚配偶出生年份应该与再婚一方填答的自己的出生年份一致。⑵用初婚配偶出生年份和最佳初婚结婚年份计算出来的初婚配偶的初婚年龄应在 18-40 岁之间。我们以初婚配偶自己填答的最佳出生年（qa1y_best）为标准对本人填答的初婚配偶出生年进行更正，并将更正结果保存在最佳变量 qe606y_best 中。当初婚配偶自己填答的最佳出生年不合理或缺失时，则在 qe606y_best 中保留受访者本人填答的合理的初婚配偶出生年，或使用家庭关系库中配偶出生年的合理代答值（t1y_a_s 或 cyear-t1b_a_s）。若经更正后个别案例的 qe606y_best 取值的合理性仍存疑，则仍然保留原取值。

2.3 qe1_best：您现在的婚姻状态的修正

我们将成人问卷中受访者本人填答的目前婚姻状况（qe1）和家庭成员问卷中由某一家庭成员代答的婚姻状况（tb3_a_p）进行比对，并参照家庭关系库中夫妻互认、父母-子女互认等匹配情况，对 qe1 填答的真实性进行了人工判断，并将更正后的结果保存在变量 qe1_best 中，而不改动 qe1 原来的取值和取值所对应的婚姻模块。

3. 小结

表 4 比较了经过程序更正后生成的最佳变量与原变量在取值上的变动。取值的变动既包

括对不合理取值的更正，也包括对缺失值的填补。在四个变量中，初婚年的取值存疑比例最高，达 14.4%（用取值不一致或不合理的样本数除以应填答的样本数），次之为初婚配偶出生年，再次为出生年，婚姻状况的取值存疑比例最低。我们根据已有的信息对取值存疑的样本进行了取值更正，并计算了更正的样本占取值存疑样本的比例，即表 4 中的纠正比例。从该比例中我们可以看到，出生年和婚姻状况能够更正的样本较少，初婚年和初婚配偶出生年能够更正的样本较多。表 4 中亦列出最佳变量无法更正的情况，即在 qe605y 和 qe606y 的填答中，夫妻双方的填答不一致但都合理（无法确定孰对孰错）以及夫妻双方填答不一致且都不合理。对“取值不一致但都合理”的情况，我们建议研究者任选夫妻任意一方填报的答案。对“取值不一致且都不合理”的情况，我们建议研究者可以考虑在分析时剔除这部分样本（该部分样本的编号见附录）。

表 4. 经过更正的样本数

最佳变量	原变量	应做填答的样本数	取值不一致或不合理的样本数	更正取值的样本数	取值不一致但都合理	取值不一致且都不合理	纠正比例（%）
qa1y_best	qa1y	33,600	2,431	150			6.2
qe605y_best	qe605y	29,165	4,191	1,091	2,936	164	26.3
qe606y_best	qe606y	29,165	3,345	3,022	309	14	90.3
qe1_best	qe1	33,600	299	30			10.0

注：纠正比例是用更正取值的样本数除以该变量存在不一致或不合理取值的样本数。

需要重申的是，最佳变量的取值不一定是最真实的取值，只是后期根据已有信息推断出来最合理的取值。对于已有信息缺失或信息之间存在逻辑矛盾后期难以判断的样本，我们只能保留原始取值。所以，最佳变量只是做出了有限的更正，研究者应根据其研究性质自行决定是否使用最佳变量。

由于时间和人力的有限，我们目前尚未对出生月份、初婚月份、以及其他婚姻时间（如离婚时间、丧偶时间）做出清查和更正。但为了更正其他婚姻时间可能存在的填答错误及确认最佳变量取值的合理性，2012 年调查的婚姻模块的设计新增了确认模块，对初婚及 2010 年各婚姻状态的时间均进行了填答确认，未来将根据新收集的数据更正最佳变量的取值及生成其他婚姻时间的最佳变量。

附录：qe605y 或 qe606y 夫妻双方填答不一致且均不合理的样本清单

序号	sampleid	qa1y_best	qe605y_best	qe606y_best
1	450356102	1945	1963	1947
2	450358101	1955	1987	1964
3	450358103	1964	1984	1955
4	450131101	1977	2005	1980
5	450131103	1980	2004	1977
6	450304102	1984	1998	1976
7	420022101	1936	1950	1936
8	420022102	1936	1951	1936
9	320155101	1951	1978	1949
10	320236101	1954	1970	1955
11	320236102	1955	1963	1954
12	360280104	1956	1974	1967
13	360294101	1957	1971	1957
14	360294102	1957	1970	1957
15	360129102	1947	1964	1942
16	530447101	1945	1972	1952
17	530447102	1952	1973	1945
18	530214101	1949	1998	1960
19	530394103	1976	1985	1972
20	530309102	1984	2000	1982
21	530310106	1992	2008	1988
22	500183102	1930	1947	1926
23	520032101	1952	1971	1944
24	520032102	1944	1970	1952
25	520203101	1980	1996	1976
26	520384102	1951	1968	1946
27	520402101	1964	1981	1953
28	520436102	1961	1969	1949
29	520535101	1943	1970	1943
30	520535105	1943	1968	1943
31	350045102	1951	1962	1944
32	350154102	1956	1972	1952
33	350124101	1950	1974	1954
34	350124102	1954	1973	1950
35	350173102	1947	1964	1939
36	350217101	1975	1982	1971
37	350220101	1940	1960	1948
38	430367101	1939	1954	1941

序号	sampleid	qa1y_best	qe605y_best	qe606y_best
39	430367102	1941	1953	1939
40	430455101	1947	1974	1954
41	430455102	1954	1973	1947
42	430481102	1943	1958	1939
43	430520101	1951	1995	1952
44	430527101	1948	2006	1938
45	430712101	1938	1947	1933
46	430712102	1933	1939	1938
47	330255101	1968	1988	1962
48	330255102	1962	1987	1968
49	330087102	1949	1966	1938
50	510229101	1961	1987	1953
51	510229102	1953	1985	1961
52	510859103	1937	1953	1933
53	510650102	1951	1967	1947
54	510905101	1969	1980	1963
55	510353103	1931	1947	1930
56	510375102	1972	1983	1967
57	510396101	1985	1999	1984
58	510396102	1984	2000	1985
59	510398101	1955	1971	1954
60	510398102	1954	1969	1955
61	510403102	1952	1965	1949
62	510546102	1968	1980	1967
63	130750101	1966	1988	1962
64	130750102	1962	1987	1966
65	130713101	1942	1964	1941
66	130713102	1941	1963	1942
67	130935104	1973	1989	1966
68	130907103	1946	1958	1933
69	130504101	1949	1959	1946
70	130393101	1954	1995	1954
71	130595101	1937	1953	1936
72	130595102	1936	1951	1937
73	230585101	1953	1969	1946
74	230544101	1949	1973	1951
75	230544102	1951	1975	1949
76	230625101	1969	1984	1967
77	230121101	1974	1988	1974
78	230121102	1974	1981	1974
79	230244103	1948	1969	1948
80	230244104	1948	1970	1948
81	370292102	1949	1966	1947

序号	sampleid	qa1y_best	qe605y_best	qe606y_best
82	370277103	1947	1964	1942
83	370210103	1923	1941	1927
84	370219106	1934	1947	1928
85	140614102	1948	1962	1940
86	140617101	1953	1983	1959
87	140617102	1959	1982	1953
88	140531101	1952	1968	1945
89	140784102	1974	1990	1970
90	140845102	1984	1999	1973
91	140554102	1950	1963	1944
92	140320101	1971	1987	1965
93	140629103	1943	1961	1944
94	140251102	1931	1940	1927
95	140254101	1949	1964	1942
96	140436101	1958	1989	1961
97	140436102	1961	1986	1958
98	610347101	1969	1985	1962
99	610237101	1963	1980	1962
100	511014101	1942	1974	1947
101	511014102	1947	1973	1942
102	310138102	1939	1981	1955
103	311058102	1946	1964	1932
104	311058103	1932	1963	1946
105	311355102	1932	1949	1932
106	311771102	1933	1953	1935
107	311942101	1959	1981	1959
108	311942102	1959	1982	1959
109	311948102	1942	1968	1944
110	312116102	1945	1959	1941
111	312497101	1946	1969	1947
112	312497102	1947	1972	1946
113	312558101	1938	1956	1935
114	312558102	1935	1966	1938
115	410342102	1940	1957	1939
116	410569101	1979	1999	1977
117	410569103	1977	1998	1979
118	410586101	1952	1980	1951
119	410586103	1951	1981	1952
120	411659102	1946	1960	1944
121	411660102	1969	1984	1964
122	410810101	1947	1968	1938
123	410872101	1925	1937	1901
124	411210103	1985	1997	1984

序号	sampleid	qa1y_best	qe605y_best	qe606y_best
125	411214101	1940	1957	1936
126	411217101	1970	1990	1970
127	411217102	1970	1989	1970
128	440472102	1978	1994	1972
129	440316101	1958	1990	1979
130	440493101	1946	1994	1963
131	440493104	1963	1993	1946
132	441396102	1965	1982	1961
133	441487101	1946	1975	1950
134	441487102	1950	1976	1946
135	441767102	1982	1999	1970
136	441610104	1963	2004	1962
137	441621101	1937	1954	1931
138	440592102	1942	1987	1962
139	621289101	1959	1984	1955
140	621289102	1955	1983	1959
141	621290102	1943	1950	1939
142	620787101	1945	1961	1937
143	620274101	1948	1967	1954
144	621141101	1968	1984	1968
145	621141102	1968	1985	1968
146	620004101	1965	1986	1964
147	620004102	1964	1987	1965
148	620191101	1945	1963	1948
149	621617104	1974	1991	1971
150	621764101	1946	1962	1946
151	621764102	1946	1961	1946
152	621765102	1979	1996	1976
153	621766102	1963	1970	1961
154	621769101	1965	1975	1966
155	621769102	1966	1976	1965
156	621805101	1940	1953	1957
157	620771101	1967	1983	1962
158	620772101	1943	1963	1949
159	620772102	1949	1965	1943
160	620447101	1952	1972	1953
161	620447102	1953	1973	1952
162	620403101	1948	1971	1947
163	620403103	1947	1974	1948
164	621489101	1985	2001	1980
165	621549103	1978	1989	1975
166	620129103	1954	1969	1950
167	621570102	1976	1993	1972

序号	sampleid	qa1y_best	qe605y_best	qe606y_best
168	621577103	1948	1964	1942
169	620336102	1941	1958	1936
170	620900102	1944	1958	1940
171	621233102	1965	1981	1960
172	621030102	1971	1986	1958
173	211106101	1940	1960	1937
174	211106102	1937	1966	1940
175	211505102	1973	1988	1966
176	211508101	1945	1966	1940
177	211508102	1940	1965	1945
178	210719101	1930	1964	1939