China Family Panel Studies

CFPS 中国家庭追踪调查

技术报告系列：CFPS-41

系列编辑：谢宇　责任编辑：赵逸文

中国家庭追踪调查
文本编码技术报告

王祎睿　谷丽萍　戴利红　吴琼

2021. 06

目录

社会调查中的大部分信息是通过数值型方式来记录的，譬如表示收入的数字或表示某个选项的数字。数值型数据便于记录，也便于后期整理和发布，因此被广泛采用。但是，数值型数据并不适用于所有场景，因为它对输入信息的格式有着严格要求，适用于应答较为固定的封闭题型的场景。在其他一些场景中，假如我们想要获得更为丰富的信息，应该选择文本型数据的采集方式。

文本型数据既可以采用结构化的采集方式，也可以采用非结构化的采集方式。它们有各自的优势和缺陷。文本型数据的结构化采集方法就是采用封闭题型设计，让受访者或访员直接从列表或选项中选择信息。结构化封闭题型的优势在于其获取的数据格式规范，后期处理成本低；但其缺陷也较为明显，一是需要受访者或者访员在现场进行归类，如果归类原则较为复杂，存在分类错误的风险；二是现场分类的类别一般不能过多，这样的限制也会影响采集信息的丰富程度。文本型数据的非结构化采集方法不提供给受访者或访员固定选项，而是通过一系列的开放型问题进行文本信息获取，譬如询问受访者的工作单位名称、职位名称、工作的具体内容等。与高度结构化的固定选项题型相比，开放文本题能够采集更为丰富的信息，但是，开放文本的采集方式更容易造成数据的不完整性。以职业信息为例，如果受访对象的职业文本缺失关键信息，编码员在后期可能无法通过访员记录的文本来进行准确有效的编码。除了数据的完整度有待核查之外，因为原始的文本信息格式不够规范，同时调查不会将受访者的隐私信息直接共享给数据用户，所以文本题的数据处理过程比数值型变量更复杂。针对开放型文本信息，普遍做法是依据一定的原则是将文本信息转化为结构化编码。

CFPS 作为一项综合型社会调查项目，采集了多种类型的文本信息，其中包括职业、行业、疾病、行政职务、方言等。下面我们对每一种类型的文本信息的采集方式、数据资源以及编码规则进行介绍。

一. 职业

1.1 采集方式

职业信息的采集方式在历年有所变化。CFPS2010 在职业文本的采集上最为特别，既直接询问了受访者工作的具体内容，又让受访者对自己的职业进行分类，也即 2010 年 CFPS 综合采用了结构化和非结构化两种方式来采集职业信息。涉及到的职业信息有受访者的当前主要工作（G307），非农工作（H405），第二职业（G701），兄弟姐妹职业（B309）以及第

一份工作。2012 年，CFPS 针对受访者在两次调查之间从事的所有工作按照其雇佣性质（受雇、非农自雇、家庭帮工）进行逐一提问。职业信息的采集均通过自由文本的方式输入，不再需要受访者提供直接的分类。与 2010 年不同，CFPS 在 2012 年没有直接提问受访者当前的主要工作，而是在询问完所有工作之后，在 G7 最主要工作部分，在调查中按照一定的规则生成了“主要工作单位名称”，详细说明见《中国家庭追踪调查 2012 年当前主要职业综合变量的建构 》。2012 年，CFPS 还在 V 部分（父母信息）模块采集了受访者 14 岁时的父母职业信息；并对在婚人员在 E2 模块其配偶的职业情况。自 2014 年开始，CFPS 职业信息采集的方式都保持一致，采用自由文本的方式，采集内容包括主要工作（GE 模块），实习工作（GA 模块），以及配偶工作。除此之外，CFPS 在 2018 年还额外采集了第一份工作的职业信息。表 1 展示了不同年份采集的职业信息。

为了从数据源上保证职业文本的有效性，CFPS 项目组针对职业文本的采集和编码做了以下工作：（1）在访员培训中讲解采集文本型数据的注意事项。（2）在访问过程中对职业文本数据进行实时核查，并将不符合编码规范的文本记录发给访员，要求提供反馈。（3）将访员反馈的文本信息会被纳入编码范围。这一质量控制的过程适用于职业、行业和疾病文本的采集。编码的判断过程可能会结合多个变量，如主要工作和第一份工作的职业编码要结合这份工作的“工作单位名称”和“工作单位做什么”，实习与兼职的职业编码要结合这份工作的“工作单位做什么”。

1.2 编码规则

CFPS2010 至 CFPS2018 中的职业编码采用的标准以《中华人民共和国国家标准职业分类与代码》（GB/T6565-2009）为基础。该编码表采用从粗到细的三级分类制。例如，针对最终分类（细类）为“哲学研究人员”的观测，其第 2 级分类为“科学研究人员”，第 1 级（粗）分类为“专业技术人员”。CFPS 基于该分类原则，进行了重组，将该编码表的原始 1 级分类代码 （包含 0 到 9 以及代表军人的 X 类和不便分类人员的 Y 类）用如下代码表示：1 替换 0；2 替换 1 和 2；3、4、5 类不变；6 替换了 6、7、8 和 9；7 替换了 X 类，9 替换了 Y 类，另增加 8 代表无职业者。并从格式上将国标码的横杠链接模式改为五位数字码，譬如按国标码模式是 2-11，在 CFPS 数据采集系统中是 2.1.1，在发布数据集中转换成发布码 20101。此外，在 CFPS 的编码中，“-7”表示无法分类的职业描述，该代码只能由编码判定员给出，主要针对记录信息与职业无关、十分不详等职业描述不清的情况；“99999”表示文本无效导

致的无法编码。从 2014 年开始，职业编码体系针对“打工”和“工人”等职业描述，加入了以下三个编码：“99700”表示零工、临工、散工、打工、工作、上班、务工；“99800”表示工人；“99900”表示技术工人、技工、技术员、操作工。同时，CFPS 还保留了问卷设计中的 3 个通用代码：“-1”不知道、“-2”拒绝回答、“-8”不适用。职业编码的代码和标签详见 CFPS 项目网站中“数据文档”页面的《职业-职业威望转换说明.xlsx》 。

1.3 编码工作的组织模式

CFPS2010 至 CFPS2016 的编码全过程均采用的是人工编码。2010 年成人问卷中的 G303 和 G305 设计为封闭性选择题，G307、G308、H405、H406 是访员在计算机辅助调查系统（CAPI）下根据 CFPS 职业和行业代码字典，采用查询法对受访者的职业和行业进行现场分类编码。其余问题则均设计为开放性问题，访员根据受访者的回答详细记录职业的重要信息。CFPS2010 的职业编码过程可以参考技术报告《中国家庭追踪调查 2010 年职业行业编码》。职业编码的工作模式是“双向独立验证并判定”。它的具体方法是在第一阶段对每条文本信息由两位编码员进行独立编码，如果两位独立编码员的编码结果一致则直接通过。2010 年至 2016 年第一阶段的编码全是人工编码。从 2018 年开始，我们在编码的第一阶段启用了一位人工编码员，同时由我们的工作人员进行计算机辅助编码[1]。结果不一致时需要引入第三位经验较为丰富的编码员，如果该编码员的结果与之前任意一位一致则采用此编码，当三个结果均不相同时由编码管理员（一般为资深编码员）审核并决定用前三人中谁的值，或者赋予除了这三个人给的值之外的值。

1.4 CFPS2010-CFPS2018 职业编码变量名分布

在 CFPS2010 至 CFPS2018 中，职业变量的发布数据中的变量名汇总如下表。编码变量一般为原始文本信息之后添加 code，如果问卷中的文本变量名在跨年间发生变化，编码变量也会有所差异。

表 1 CFPS2010 至 CFPS2018 职业变量的变量名

内容	CFPS2010	CFPS2012	CFPS2014	CFPS2016	CFPS2018
实习/兼职			QGA401CODE	QGA401CODE	QGA401code
第一份工作	QG601_OCCU				KGD4code
主要工作			QG303CODE	QG303CODE	QG303code

[1]吴琼, 戴利红, 张婧申. 机器学习在社会调查职业编码中的应用[J]. 调研世界，2019(09)：56-60

配偶/同伴职业-上期婚姻	TB5_CODE_A_S	qe209bcode_best	QEA203CODE	QEA203CODE	QEA203code
配偶职业-婚姻史			EEB4022_A_1CODE	EEB4022_A_1CODE	EEB4022_A_1code
受雇工作		qg411code_a_1-10			
非农自雇		qg510code_a_1-10			
家庭帮工		qg609code_a_1-10			
父亲职业	TB5_CODE_A_F FOCCUPCODE（综合变量）	14 岁时父亲职业编码:qv103code_best			
母亲职业	TB5_CODE_A_M MOCCUPCODE（综合变量）	14 岁时母亲职业编码:qv203code_best			
主要职业（综合变量）		job2012mn_occu			
职业类别	QG307CODE				
第二职业	QG701_OCCU				
非农工作	QH405CODE				
兄弟姐妹职业	QB309_OCCU_1-15				
孩子职业	TB5_CODE_A_C1-10				

1.5 衍生变量：职业威望

为了方便用户更好地使用职业变量，CFPS 将职业国标码体系（Chinese Standard Classification of Occupations，CSCO）转换为国际标准职业分类代码（International Standard Classification of Occupation，ISCO-88），并依据 ISCO-88 职业分类代码建构了国际标准职业社会经济指数（International Socio-Economic Index of Occupational Status，ISEI）、标准国际职业声望量表（Treiman’s Standard International Occupational Prestige Scale，Treiman’s SIOPS）两套职业社会经济地位测量指标，以及成人问卷受访者职业现在工作、主要工作的 EGP 职业分类代码（Erikson and Goldthorpe’s Class Categories，EGP）。2010 年职业威望系列变量的解释说明见《CFPS-10 中国家庭追踪调查 职业社会经济地位测量指标构建》。由职业编码转换成的职业威望的对应表见 CFPS 项目网站中“数据文档”页面的《职业与职业威望转化说明.xlsx》。我们生成的职业威望系列变量的变量名如下表。

表 2 CFPS2010 至 CFPS2018 职业威望变量的变量名

	职业编码变量	ISCO	ISEI	SIOPS	EGP
2010	现在工作 QG307CODE	QG307ISCO	QG307ISEI	QG307SIOPS	QG307EGP
	第一份工作 QG601_OCCU	QG601_ISCO	QG601_ISEI	QG601_SIOPS	

	第二份工作	QG701_ISCO	QG701_ISEI	QG701_SIOPS	
	QG701_OCCU				
	非农工作	QH405ISCO	QH405ISEI	QH405SIOPS	
	QH405CODE				
	父亲主要职业	FOCCUPISCO			
	FOCCUPCODE				
	母亲主要职业	MOCCUPISCO			
	MOCCUPCODE				
2012	母亲职业	QV203_ISCO			
	qv203code_best				
	父亲职业	QV103_ISCO			
	qv103code_best				
	配偶职业	QE209B_ISCO			
	qe209bcode_best				
	非农工作	SG411_ISCO			
	SG411CODE_BEST				
2014	实习	*QGA401COD	*QGA401CODE	*QGA401CODE	
	QGA401CODE	E_ISCO	_ISEI	_SIOPS	
	主要工作	*QG303CODE_	*QG303CODE_I	*QG303CODE_	*QG303CODE
	QG303CODE	ISCO	SEI	SIOPS	_EGP
2016	实习	QGA401CODE	QGA401CODE_	QGA401CODE_	
	QGA401CODE	_ISCO	ISEI	SIOPS	
	主要工作	QG303CODE_I	QG303CODE_IS	QG303CODE_S	QG303CODE_
	QG303CODE	SCO	EI	IOPS	EGP
2018	实习	QGA401CODE	QGA401CODE_	QGA401CODE_	
	QGA401CODE	_ISCO	ISEI	SIOPS	
	第一份工作	KGD4CODE_IS	KGD4CODE_IS	KGD4CODE_SI	
	KGD4CODE	CO	EI	OPS	
	主要工作	QG303CODE_I	QG303CODE_IS	QG303CODE_S	QG303CODE_
	QG303CODE	SCO	EI	IOPS	EGP

注：*为暂时未发布变量，将在相应数据集的下次数据更新时发布。用户如需提前使用，请发信至项目组服务邮箱。

二． 行业

2.1 采集方式

2010 年，我们行业编码的依据是受访者的工作属于什么行业；2012 年及以后，我们的行业编码的依据是受访者的工作单位属于什么行业。CFPS2010 涉及行业编码的题目有两种提问方式，一是直接询问受访者的工作属于哪个行业；二是询问受访者的工作内容，据此判

断相应的行业。2012 年，我们在共用模块的【G4 受雇】中询问受雇单位是做什么的，在【G5 非农自雇】模块中询问受访者主要做什么生意，在【G6 不拿工资为家庭经营活动帮工】模块中询问受访者参与的家庭经营活动主要生产什么产品或者从事什么经营活动。

2014 年后提问形式都保持一致。我们在【GA 实习与兼职】模块提问了实习单位从事的活动,在【GE 主要工作】模块提问了工作单位主要是做什么的。2018 年我们还采集第一份工作的单位信息。

2.2 编码规则

行业编码的主要依据是受访者所在的单位以及受访者对单位主要经营活动描述。当受访者没有单位信息时，我们则根据其工作内容来进行行业信息的判断。CFPS 的行业编码采用的标准行业代码使用的是《国民经济行业分类》（GB/T 4754-2002），该编码表将国民经济行业划分为 20 类。CFPS 事后编码完全采用了该编码表既有的分类和代码，并在此基础上添加了一个类别“21”，表示不便分类的其他行业。此外，CFPS 事后编码同样保留了问卷设计的 3 个通用代码：“-1”不知道、“-2”拒绝回答、“-7”职业描述不清，无法分类、“-8”不适用。行业编码的代码和标签见附录 1。

2.3 编码工作的组织模式

CFPS2010 的行业编码过程可以参考《中国家庭追踪调查 2010 年职业行业编码》。行业编码的工作模式是“双向独立验证并判定”。它的具体方法是在第一阶段对每条文本信息由两位编码员进行独立编码，如果两位独立编码员的编码结果一致则直接通过。2010 年至 2016 年第一阶段的编码全是人工编码。2018 年，第一阶段启用了一位人工编码员，同时由我们的工作人员进行计算机辅助编码。结果不一致时需要引入第三位经验较为丰富的编码员，如果该编码员的结果与之前任意一位一致则确定该编码为最终编码，当三人结果均不相同时由编码管理员（一般为资深编码员）审核并决定用前三人中谁的值，或者赋予除了这三个人给的值之外的值。CFPS 项目组在调查季访问进行中会通过实时核查系统，把采集不符合规范的文本发送给访员，访员会审核并反馈更新值。完成以上的四遍编码后，会有两位编码员对访员反馈中的文本进行人工编码，项目组的工作人员会做判断和选值，用其选定的数据替换原始数据。

2.4 CFPS2010-CFPS2018 行业编码变量名分布

在 CFPS2010 至 CFPS2018 中，职业变量的发布数据中的变量名汇总如下表。

表 3 CFPS2010 至 CFPS2018 行业变量的变量名

工作单位做什么（行业）	CFPS2010	CFPS2012	CFPS2014	CFPS2016	CFPS2018
实习/兼职			QGA4CODE	QGA4CODE	QGA4code
第一份工作	QG601_IND				KGD3code
主要工作			QG302CODE	QG302CODE	QG302code
受雇		qg410code_a_1-10			
非农自雇		qg509code_a_1-10			
家庭帮工		qg608code_a_1-10			
工作属于哪个行业	QG308CODE				
第二职业	QG701_IND				
非农工作	QH406CODE				
兄弟姐妹工作	QB309_IND_1-15				

三. 行政/管理职务

3.1 采集方式

行政/管理职务体现了职业中的权威地位。除了 2012 年，CFPS 均采集了受访者的行政/管理职务相关信息，我们描述行政/管理职务的文本与职务的部门以及下属数量结合，进行行政/管理职务的编码。CFPS 先询问受访者是否有行政/管理职务，对于给出肯定应答的受访者再询问其行政/管理职务是什么。为了从数据源上保证编码文本的有效性，CFPS 项目组会在历年追踪调查开展前的访员培训中，讲解采集行政/管理职务类数据的具体注意事项。

3.2 编码规则

行政/管理职务变量的编码来源是附录 2 的《行政/管理职务编码体系表》。历年行政/管理职务的编码逻辑可参考《中国家庭追踪调查 2010 行政/管理职务综合变量的建构》。

3.3 编码工作组织形式

行政/管理职务编码过程中需要用到 GE 模块的其他变量，所以首先资深编码员会把编码可能需要的变量全部提给编码员，并讲解《中国家庭追踪调查 2010 行政/管理职务综合变量的建构》的要点。行政/管理职务的具体编码过程为：首先由两位编码员对每条文本信

息进行独立编码，如果两位独立编码员的编码结果一致则直接通过，如果结果不一致，则引入第三位经验较为丰富的编码员。该编码员的结果与之前任意一位一致则确定该编码为最终编码，如果编码结果与之前两位编码员的结果都不一样，则由其确定用这三遍编码的哪个值。

3.4 CFPS2010-CFPS2018 行政/管理职务历年变量名分布

从 2010 年到 2018 年中，行政/管理职务变量的发布数据中的变量名汇总如下表。

表 4 CFPS2010 至 CFPS2018 行政/管理职务变量的变量名

	CFPS2010	CFPS2012	CFPS2014	CFPS2016	CFPS2018
是否有行政管理职务	QH407	QG309	无 QG14	QG14	QG14
职务编码	*QH407CODE	QG310CODE	QG1401CODE	QG1401CODE	QG1401code

注：*为暂时未发布变量，将在相应数据集的下次数据更新时发布。用户如需提前使用，请发信至项目组服务邮箱。

四. 职业期望

4.1 采集方式

职业期望是对于未来希望从事的职业的想象和描述，并很有可能影响今后实际的职业选择。CFPS 以开放性问题的形式采集了职业期望信息，即由访员根据受访者的回答详细记录未来职业的重要信息。职业期望有两种类型：一是个人对自己的职业期望，二是父母对孩子的职业期望。对于 10 岁以上的样本，我们在大部分年份针对正在上学的人群询问其将来最希望从事的具体职业是什么。对于 0-15 岁的孩子，我们在不同年份针对不同阶段的样本询问其父母对孩子的职业期望。

4.2 编码规则

建构职业期望分类表的基本原则包括：第一，与国家标准职业分类和代码保持一致。CFPS 的职业编码使用的是《中华人民共和国国家标准职业分类与代码》（GB/T6565-2009）。一方面，职业期望的类别必须是可以具体到三级代码的职业分类，要么是某一个具体的职业小类，要么是多个职业小类的集合。另一方面，所有的职业分类都可以进入职业期望分类表。第二，每一职业期望类别拥有足够的个案数。这为下一步研究提供基本的数据支持。第三，每一类别代表社会经济地位和性别差异。换句话说，如果某几项职业期望类别之间不存在地

位或性别构成的差异，我们便可以进行合并。这是因为我们更希望把握和体现受访者不同的职业期望。由此，基于 CFPS 既有的职业期望原始数据和国家标准职业分类体系，我们建构了一套职业期望变量的编码来源，即《职业期望编码体系表》。该分类表包括 27 类职业期望类别。其中，前 23 类为详细职业期望类型，如“国家机关、党群组织、事业单位负责人”、“企业负责人”，并给出了它们分别对应的国家标准职业分类；第 24-26 类为粗略职业期望类型，分别是“读书”、“为人民服务”、“打工”；最后一类是不便分类的其他从业人员。之所以单独给出粗略职业期望类别，主要的考虑是这三类反映了不同的职业取向，且拥有一定的个案数。职业期望编码体系表的详细类别见附录 3。

4.3 编码工作组织形式

职业期望编码采取手动集中编码和自动集中编码的方式进行。其中，手动集中编码指的是在调查结束后，由专业编码员采用双向独立验证并判定的方式根据对职业的理解和编码列表的掌握情况选择相应的职业编码；自动编码指在上述职业编码完成后，由专业编码员借助编码软件进行全自动的职业期望编码。后者是因为职业期望分类表是以职业分类为基础编制而成的。

4.4 CFPS2010-CFPS2018 职业期望变量在发布库中的变量名

从 2010 年到 2018 年中，职业期望变量的发布数据中的变量名汇总如下表。

表 5 CFPS2010 至 CFPS2018 疾病变量的变量名

少儿	CFPS2010	CFPS2012	CFPS2014	CFPS2016	CFPS2018
孩子自报	wm601code	*KS801code	*KS801code	KS801CODE	QS801_B_2code
父母对孩子	wd101code	*wd101code	wd101code	WD101CODE	WD101code
成人	2010	2012	2014	2016	2018
成人自报		KS801	KS801CODE	KS801CODE	QS801_B_2code

注：*为暂时未发布变量，将在相应数据集的下次数据更新时发布。用户如需提前使用，请发信至项目组服务邮箱。

五. 疾病

5.1 采集方式

CFPS 在各轮次分别采集了成人和少儿的疾病信息。总的来说，我们对二者的问法有所不同：在少儿问卷的疾病题目中，CFPS 对于初次参加访问的儿童询问家长孩子患过的最严重的疾病，在后续调查中询问过去 12 个月孩子的患病情况。在成人问卷的疾病题目中，CFPS 采集受访者被医生诊断的慢性疾病名称。

5.2 编码规则

疾病编码的编码来源是《中国家庭追踪调查疾病编码》，详见官网数据文档中的《CFPS 疾病编码.xlsx》。

5.3 编码工作组织形式

疾病编码的工作模式是双向独立验证并判定。针对疾病文本，CFPS 项目组在调查季访问进行中会通过实时核查系统把被判断为不符合规范的文本反馈给访员，访员会审核并反馈更新值。完成以上的四遍编码后，会有两位编码员对访员反馈中的文本进行人工编码，项目组的工作人员会做判断和选值，用其选定的数据替换原始数据。

5.4 CFPS2010-CFPS2018 疾病变量在发布库中的变量名

从 2010 年到 2018 年，疾病变量的发布数据中的变量名汇总如下表。

表 6 CFPS2010 至 CFPS2018 疾病变量的变量名

	过去 12 个月最严重的疾病	出生至今最严重的疾病
CFPS2010 少儿		WC501A_LBL
CFPS2012 少儿	WC501、WC501CODE	WC501_2010、WC501_2010CODE
CFPS2014 少儿	WC5CODE	WC5_2010CODE
CFPS2016 少儿	PC5_CODE	PC5_2010CODE
CFPS2018 父母代答	WC5_B_1CODE	WC5_2010CODE

	第一种慢性疾病名称	第二种慢性疾病名称
CFPS2010 成人	QP404ACODE	QP404BCODE
CFPS2012 成人	QP403A	QP403B
CFPS2014 成人	QP402ACODE	QP402BCODE

CFPS2016 成人	QP402ACODE	QP402BCODE
CFPS2018 个人自答	QP402ACODE	QP402BCODE

5.5 相关变量：身体不适

在医生诊断的疾病之外，CFPS 还询问了受访者身体不适的情况，它没有标准的编码规则。2010-2014 年均是选择题，询问受访者过去两周内的主要身体不适情况，但 2010 年的选项与后面两轮有所不同（见下表）。2016 和 2018 年是文本型，我们通过提取关键词信息采用程序进行编码，并对剩余样本实施人工编码。

表 7 身体不适变量编码体系表

2010 身体不适的固定选项		2012 和 2014 身体不适的固定选项		2016 年及以后文本型身体不适的编码表	
数值	数值对应的身体不适	数值	数值对应的身体不适	编码	编码对应的身体不适
1	发烧	1	发烧	1	发烧
2	疼痛	2	疼痛	2	血糖高/血糖低/糖尿病
3	腹泻	3	腹泻	3	肩颈不适
4	咳嗽	4	咳嗽	4	呼吸系统不适
5	心慌/心悸	5	上不来气	5	口腔不适
6	其他【请注明】___	6	无法集中注意力	6	眼部不适
7	无自觉症状	7	步行困难	7	妇科问题
		8	心慌/心悸/心口痛	8	心脏不适
		77	其他【请注明】____	9	感冒
		78	以上都没有	10	血压低/血压高
				11	肠胃不适
				12	头部不适/睡眠不佳
				14	腰部不适
				15	中暑
				16	全身不适、疲劳
				17	腿部不适
				77	其他【请注明】

六. 死亡原因

6.1 采集方式

CFPS2010 至 CFPS2018 的死亡变量出现在个人问卷和成员问卷。CFPS2010 在成员问卷中采集过世父母的死亡原因，受访者兄弟姐妹的死亡原因变量没发布。在个人问卷的婚姻模块，CFPS 也会根据受访者婚姻状况询问受访者在某一段婚姻中配偶的死亡原因。从

CFPS2012 开始，我们会问受访者所在家庭成员去世的原因。

6.2 编码规则

CFPS 死亡原因编码的编码来源是《中国家庭追踪调查死亡原因编码》，详细见附录 4。

6.3 编码工作组织形式

CFPS2010 至 CFPS2018 的死亡原因编码均是调查中的现场人工编码，即访员在访问过程中，直接询问受访者的亲属的死亡原因，由访员根据受访者的原话判断这位亲属死亡原因，并从访问系统的死亡原因列表选择对应的编码。

6.4 CFPS2010-CFPS2018 死亡原因变量在发布库中的变量名

从 2010 年到 2018 年，死亡原因变量的发布数据中的变量名汇总如下表。

表 8 CFPS2010 至 CFPS2018 死亡原因变量的变量名

		2010	2012	2014	2016	2018
个人问卷	父亲	QB401				
	母亲	QB501				
	初婚配偶	QE604				
	前任配偶	QE404		EEB302	EEB302	EEB302
	刚过世配偶	QE505	QE513	EEB408	EEB408	EEB408
	上期配偶		QE205、QE308、QE406、QE506	QEA210	QEA210	QEA210
成员问卷	成员		deathreason_p	ta401_A14_p	ta401_a16_p	ta401_a18_p
	父亲		deathreason_f	ta401_A15_f	ta401_a16_f	ta401_a18_f
	母亲		deathreason_m	ta401_A16_m	ta401_a16_m	ta401_a18_m
	配偶		deathreason_s	ta401_A17_s	ta401_a16_s	ta401_a18_s
	孩子 1		deathreason_c1	ta401_A14_c1	ta401_a16_c1	ta401_a18_c1
	孩子 2		deathreason_c2	ta401_A14_c2	ta401_a16_c2	ta401_a18_c2
	孩子 3		deathreason_c3	ta401_A14_c3	ta401_a16_c3	ta401_a18_c3
	孩子 4		deathreason_c4	ta401_A14_c4	ta401_a16_c4	ta401_a18_c4
	孩子 5		deathreason_c5	ta401_A14_c5	ta401_a16_c5	ta401_a18_c5
	孩子 6		deathreason_c6	ta401_A14_c6	ta401_a16_c6	ta401_a18_c6
	孩子 7		deathreason_c7	ta401_A14_c7	ta401_a16_c7	ta401_a18_c7
	孩子 8		deathreason_c8	ta401_A14_c8	ta401_a16_c8	ta401_a18_c8
	孩子 9		deathreason_c9	ta401_A14_c9	ta401_a16_c9	ta401_a18_c9
	孩子 10		deathreason_c10	ta401_A14_c10	ta401_a16_c10	ta401_a18_c10

七. 专业

7.1 采集方式

CFPS2010 到 CFPS2014 采集到的专业数据都是封闭型的固定选项选择题，即访员直接报出专业分类，受访者做选择； 2016 年起此部分的采集方式发生了改变，CFPS2016 和 CFPS2018 的问卷直接采集受访者提供文本信息。

专业问题针对如下教育阶段和学校类型展开提问：职业初中、普通中专、成人中专、职业高中或技工学校。

7.2 编码规则

2016 年后的专业编码的编码依据是教育部发布的《中等职业学校专业目录》（2010 年修订版）。我们根据该目录，将原始文本概括为 18 个类别及“其他”，如下表所示：

表 9 专业编码体系表

2010 专业的固定选项		2012 和 2014 专业的固定选项		2016 年及以后文本型专业的编码表	
数值	数值对应的专业	数值	数值对应的专业	编码	编码对应的专业
1	农林类	1	制造大类	1	农林牧渔类
2	资源与环境类	2	资源开发与测绘大类	2	资源环境类
3	能源类	3	水利大类	3	能源与新能源类
4	土木水利工程类	4	交通运输大类	4	土木水利类
5	加工制造类	5	医药卫生大类	5	加工制造类
6	交通运输类	6	材料与能源大类	6	石油化工类
7	信息技术类	7	财经大类	7	轻纺食品类
8	医药卫生类	8	土建大类	8	交通运输类
9	商贸与旅游类	9	生化与药品大类	9	信息技术类
10	财经类	10	艺术设计传媒大类	10	医药卫生类
11	文化艺术与体育类	11	文化教育大类	11	休闲保健类
12	社会公共事务类	12	旅游大类	12	财经商贸类
13	师范类	13	电子信息大类	13	旅游服务类
77	其他【请注明】	14	轻纺食品大类	14	文化艺术类
		15	公安大类	15	体育与健身类
		16	法律大类	16	教育类
		17	农林牧渔大类	17	司法服务类
		18	环保、气象与安全大类	18	公共管理与服务类
		19	公共事业大类	19	其他
		77	其他		

7.3 编码工作组织形式

专业编码的编码方式主要是利用统计软件，对原始信息进行文本分析，提取关键词信息，建立关于专业方面的编码字典，对文本进行机器编码，把专业的原始文本归到 18 类；对于无法用程序归类的文本，我们的处理方式是：编码员对剩余样本实施人工编码。

7.4 CFPS2010-CFPS2018 专业变量在发布库中的变量名

在 CFPS2010 至 CFPS2018 中，专业变量在发布数据中的变量名汇总如下表。

表 10 CFPS2010 至 CFPS2018 专业变量的变量名

	上学模块 1/学校基本情况	教育史
CFPS2010 成人	KR440	
CFPS2010 少儿	KR440	WH403
CFPS2012 成人	KR301（初中）、KRA401（高中）	KW402（初中）、KW502（高中）
CFPS2012 少儿	KR301（初中）、KRA401（高中）	KW402（初中）、KW502（高中）
CFPS2014 成人	KRA401（高中）	KW402（初中）、KW502（高中）
CFPS2014 少儿	KRA401（高中）	KW402（初中）、KW502（高中）
CFPS2016 成人	*PS501code（高中）	*KW502_B_1code（高中）、 *KW502_B_2code（高中）
CFPS2018 个人自答	QS401code（初中）、 QS501_b_1code（高中）	KW1002_B_1code（初中）、 KW1002_B_2code（高中）
CFPS2018 父母代答	WS401code（初中）、 WS501_b_1code（高中）	

注：*为暂时未发布变量，将在相应数据集的下次数据更新时发布。用户如需提前使用，请发信至项目组服务邮箱。

八. 学科

8.1 采集方式

前一节介绍的专业信息只针对高等教育之前的学业阶段，而学科只针对高等教育阶段（大专、本科、硕士和博士）。CFPS2010 到 CFPS2014 采用封闭型的固定选项方式采集学科信息，即访员直接报出学科分类，受访者做出选择；从 2016 年开始，CFPS 让受访者直接汇报自己的学科文本信息，访员记录受访者汇报的文本。

8.2 编码规则

学科编码依据国务院学位委员会、教育部于 2011 年公布的《学位授予和人才培养学科目录(2011 年)》的学科门类。该门类把我国高校的学科分成了 13 类，即哲学、经济学、法学、教育学、文学、历史学、理学、工学、农学、医学、军事学、管理学和艺术学。我们在整合学科门类时，将“艺术学”归到“其他”，其他门类保持不变，如下表所示。

表 11 学科编码体系表

编码	编码对应的学科
1	哲学
2	经济学
3	法学
4	教育学
5	文学
6	历史学
7	理学
8	工学
9	农学
10	医学
11	军事学
12	管理学
99	其他

8.3 编码工作组织形式

学科编码的编码方式主要是利用统计软件，对原始信息进行文本分析，提取关键词，建立关于学科方面的编码字典，把学科的原始文本归到 12 类；对于无法用程序归类的文本，我们的处理方式是：编码员对剩余样本实施人工编码。

8.4 CFPS2010-CFPS2018 学科变量在发布库中的变量名

在 CFPS2010 至 CFPS2018 中，学科变量的发布数据中的变量名汇总如下：

表 12 CFPS2010 至 CFPS2018 学科变量的变量名

	上学模块 1/学校基本情况	教育史
CFPS2010 成人	KR601(大专、本科)、 KR801(硕士、博士)	QC402(大专)、QC302(本科)、 QC202(硕士)、QC102(博士)
CFPS2012 成人	KR501(大专)、KRA601(本科)、 KR701(硕士)、KR801M(博士)	KW602(大专)、KW702(本科)、 KW802(硕士)、KW902(博士)

CFPS2014 成人	KR501(大专)、KRA601(本科)、KR701(硕士)、KR801M(博士)	KW602(大专)、KW702(本科)、KW802(硕士)、KW902(博士)
CFPS2016 成人	*PS701code(大专)、*PS9code(本科、硕士、博士)	*KW602_B_1code(大专)、*KW702_B_1code(本科)、*KW702_B_2code(本科)、*KW802_B_1code(硕士)、*KW902_B_1code(博士)
CFPS2018 个人自答	QS701_B_1code(大专)、QS9code(本科、硕士、博士)	KW1003_A_1code(大专)、KW1003_A_2code(本科)、KW1003_A_3code(硕士)、KW1003_A_4code(博士)

注：*为暂时未发布变量，将在相应数据集的下次数据更新时发布。用户如需提前使用，请发信至项目组服务邮箱。

九. 学校类型

9.1 采集方式

CFPS2010 至 CFPS2018 在教育史模块、上学模块或学校基本情况模块采集了受访者的学校信息。我们根据院校名称，对高等教育院校的类型进行编码。

9.2 编码规则

学校编码的编码规则是我们根据中国专科及以上院校的实际情况做的分类。如下表所示，CFPS2010 和 CFPS2014 皆把院校编到了 14 类和其他。虽然题干提问的是就读的本科院校，但回答中包含大中专院校、高职院校、成教院校等多类教育形式不同的院校。因此，此次编码对各类学校采取明细分类，无明细分类需求的用户可自行合并。2016 年开始，我们把 2010 至 2014 年学校编码中的“全国重点”进行细分，包括“1、全国重点院校（985 高校，第一批次录取）”和“2、全国重点院校（非 985 的 211 院校，第一批次录取）”，其他编码依次往后挪一个类别。

表 13 学校编码体系表

CFPS2010-CFPS2014		CFPS2016、CFPS2018	
1	全国重点	1	全国重点院校（985 高校，第一批次录取）
2	普通重点	2	全国重点院校（非 985 的 211 院校，第一批次录取）
3	二本	3	普通重点院校（第一批次录取）
4	三本	4	普通本科院校（第二批次录取）
5	部队院校	5	三本院校（第三批次录取）
6	艺体院校	6	部队院校（提前批录取）
7	海外院校	7	艺术、体育类院校

8	大专高职	8	海外院校
9	中专	9	高职与大专院校
10	夜大与函授	10	中专院校 11
11	自考	11	成人教育院校（夜大和函授）
12	广播电视大学	12	自考院校
13	网络教育院校	13	广播电视大学
14	党校	14	网络教育院校
99	其他	15	党校，代码
		99	其他

9.3 编码工作组织形式

学校变量的具体编码为双向独立验证并判定。

9.4 CFPS2010-CFPS2018 学校变量在发布库中的变量名

在 CFPS2010 至 CFPS2018 中，学校变量的发布数据中的变量名汇总如下表。

表 14 CFPS2010 至 CFPS2018 学校变量的变量名

	学校基本情况	上学模块 1	教育史
CFPS2010 成人			COLLEGETYPE (来源：C301 您读的是哪类本科？)
CFPS2012 成人		*KRA603CODE	
CFPS2014 成人		KRA603CODE	
CFPS2016 成人	PS1CODE_COLLEGE		
CFPS2018 个人自答	QS1_B_1CODE		

注：*为暂时未发布变量，将在相应数据集的下次数据更新时发布。用户如需提前使用，请发信至项目组服务邮箱。

十. 方言

CFPS 方言编码的主要依据是《中国语言地图集》（The Language Atlas of China，以下简称《地图集》）。《地图集》由中国社会科学院语言研究所和澳洲人文科学院合作，由中国社会科学院语言研究所李荣、熊正辉、张振兴担任主编，于 1983 年开始编制，1987 年完成。《地图集》在全面的语言学调查的基础上，按古入声字、古浊声母字的演变规律对汉语方言进行分类，相比其他分类方法更为科学，已成为方言学界实际上的学科标准。《地图集》有中文和英文版本，中文版由香港朗文（远东）出版公司于 1987 年和 1991 年分两次出版。

方言编码被设定为限制性数据，并不在公开发布的数据集中，用户需要使用的话请填写我们的限制性数据申请表。方言编码的详细说明见《技术报告系列：CFPS-28 中国家庭追踪调查方言编码》。

我们采用了双向独立验证并判定（Two-way Independent Verification with Adjudication）的方式进行编码。第一轮编码由三个编码员分别单独对每一个受访者所填写的方言信息进行编码，若结果一致，则保留；若不一致，则由另一位经验较为丰富的编码员结合 CFPS 数据中的其他信息，重新确定所属编码类别编码。经统计，2012 年成人库中的 QZ104 变量，不同编码员之间的匹配率 83.04%，不匹配的情况在二次编码时结合多变量信息已得到很好解决，可编码的样本达到 99.88%。编码时，编码员通过被访者填写的文字信息，并结合其所在区县，按照《中国语言地图集》进行编码。整个过程遵循以下基本原则：a）受访者的回答为“本地话”：按照其所在区县的方言类型编码；b）受访者回答出的方言类型与其所在区域的方言不符：以受访者回答为准；c）非单一方言及少数民族语言：统一编码为 99（代表无法编码）；d）受访者的回答为“家乡话”：参照其出生地及 3 岁时户口所在地信息编码。

附录

附录 1 行业编码体系表

Code	Label
1	农、林、牧、渔业
2	采矿业
3	制造业
4	电力、燃气及水的生产和供应业
5	建筑业
6	交通运输、仓储和邮政业
7	信息传输、计算机服务和软件业
8	批发和零售业
9	住宿和餐饮业
10	金融业
11	房地产业
12	租赁和商务服务业
13	科学研究、技术服务和地质勘查业
14	水利、环境和公共设施管理业
15	居民服务和其他服务业
16	教育
17	卫生、社会保障和社会福利业
18	文化、体育和娱乐业
19	公共管理和社会组织
20	国际组织
21	军队
99	无法编码

附录 2 行政/管理职务编码表

Code	Label
0	无职务
1	公共部门基层行政/管理职务
2	市场部门基层行政/管理职务
3	公共部门中层行政/管理职务
4	市场部门中层行政/管理职务
5	公共部门高层行政/管理职务
6	市场部门高层行政/管理职务
7	公共部门顶层行政/管理职务
8	市场部门顶层行政/管理职务
-7	无法分类

-8	不适用
-2	拒绝回答
-1	不知道

附录 3 职业期望编码体系表

Code	Label
1	国家机关、党群组织、事业单位负责人
2	企业负责人
3	科学研究人员
4	工程技术人员
5	飞机和船舶技术人员
6	卫生专业技术人员
7	经济和金融业务人员
8	法律专业人员
9	教学人员
10	文学艺术工作人员
11	体育工作人员
12	新闻出版和文化工作人员
13	其他专业技术人员
14	行政办公和其他办事人员
15	安全保卫和消防人员
16	餐饮、旅游和健身娱乐场所服务人员
17	运输服务人员
18	社会服务和居民生活服务人员
19	农业生产人员
20	机械、电子、电力设备制造加工和维修人员
21	运输设备操作人员及有关人员
22	其他设备操作人员及有关人员
23	军人
24	读书
25	为人民服务
26	打工
27	不便分类的其他从业人员
-1	不知道
-2	拒绝回答
-8	不适用
-7	职业描述不清，无法分类
-9	缺失
28	工人
29	看他/她自己

*注：2010 年的职业期望体系表没有“28 工人”、“29 看他/她自己”。我们在往后轮次里加入了这两个类别。

附录 4 死亡原因编码体系表

编码	含义	编码	含义
1	传染病和寄生虫病	17.0.81	其他意外事故和有害效应
1.0	传染病	17.0.82	自杀
1.0.1	伤寒和副伤寒	17.0.83	被杀
1.0.10	败血症	2	肿瘤
1.0.11	流行性乙型脑炎	2.0	恶性肿瘤
1.0.12	流行性出血热	2.0.18	鼻咽癌
1.0.13	麻疹	2.0.19	食道癌
1.0.14	病毒性肝炎	2.0.20	胃癌
1.0.15	艾滋病	2.0.21	结肠、直肠和肛门癌
1.0.2	痢疾	2.0.22	肝癌
1.0.3	肠道其他细菌性传染病	2.0.23	肺癌
1.0.4	呼吸道结核	2.0.24	乳腺癌
1.0.5	其他结核	2.0.25	宫颈癌
1.0.6	钩端螺旋体病	2.0.26	膀胱癌
1.0.7	破伤风	2.0.27	白血病
1.0.8	百日咳	2.1	良性肿瘤
1.0.9	脑膜炎球菌感染	2.2	其他肿瘤
1.1	寄生虫病	3	血液、造血器官及免疫疾病
1.1.16	疟疾	3.0.28	贫血
1.1.17	血吸虫病	3.1	血液、造血器官及免疫的其他疾病
10	肌肉骨骼和结缔组织疾病	4	内分泌、营养和代谢疾病
11	泌尿生殖系统疾病	4.0	糖尿病
11.0.52	肾小球和肾小管间质疾病	4.1	内分泌、营养和代谢的其他疾病
11.0.53	前列腺增生	5	精神障碍
11.0.54	泌尿生殖系统的其他疾病	6	神经系统疾病
12	妊娠、分娩和产褥期并发症	6.32	脑膜炎
12.0	直接产科原因计	6.33	神经系统的其他疾病
12.0.55	流产	7	循环系统疾病
12.0.56	妊娠高血压综合征	7.0	急性风湿热
12.0.57	梗阻性分娩	7.1	心脏病
12.0.58	产后出血	7.1.34	慢性风湿性心脏病
12.0.59	母体产伤	7.1.35	高血压性心脏病
12.0.60	产褥期感染	7.1.36	急性心肌梗死
12.0.61	间接产科原因计	7.1.37	其他冠心病
12.0.62	妊娠、分娩和产褥期的其他情况	7.1.38	肺原性心脏病
13	起源于围生期的某些情况	7.1.39	其他心脏病
13.0.63	早产儿和未成熟儿	7.1.40	其他高血压病
13.0.64	新生儿产伤和窒息	7.1.41	脑血管病
13.0.65	新生儿溶血性疾病	7.1.42	循环系统的其他疾病
13.0.66	新生儿硬化病	8	呼吸系统疾病
13.0.67	起源于围生期的其他情况	8.0.43	肺炎

14	先天畸形、变形和染色体异常	8.0.44	慢性下呼吸道疾病
14.0.68	先天性心脏病	8.0.45	尘肺
14.0.69	其他先天畸形、变形和染色体异常	8.0.46	呼吸系统的其他疾病
15	诊断不明	9	消化系统疾病
16	其他疾病	9.0.47	胃和十二指肠溃疡
17	损伤和中毒外部原因	9.0.48	阑尾炎
17.0.70	机动车辆交通事故	9.0.49	肠梗阻
17.0.71	机动车以外的运输事故	9.0.50	肝疾病
17.0.72	意外中毒	9.0.51	消化系统的其他疾病
17.0.73	意外跌落	-1	不知道
17.0.74	火灾	-2	拒绝回答
17.0.75	由自然环境因素所致的意外事故	-8	不适用
17.0.76	淹死	-9	缺失
17.0.77	意外的机械性窒息	-10	无法判断
17.0.78	砸死		
17.0.79	由机器切割和穿刺工具所致的意外事故		
17.0.80	触电		