应用 Kish 表入户抽样被访者年龄结构扭曲问题研究*

张丽萍

提要：本文以实际调查数据和概率理论为基础，研究抽样调查入户抽样阶段各种统计口径人口的年龄结构。在对比登记人口、Kish 选样表人口和被调查人口的年龄结构特点的基础上，分析 Kish 表的理论概率分布，发现了低龄人口入选比例低和 Kish 选样表中入选人口与被访者年龄结构扭曲的原因。为解决 Kish 表应用的实际问题，对 Kish 表进行了仿真研究并提出对现有 Kish 表抽样过程的改进建议。

关键词：入户抽样　Kish 表　年龄结构扭曲　抽样调查

一、问题的提出

抽样是科学研究的重要方法。样本能否具有很好的代表性且能推断总体取决于抽样是否科学。概率样本是保证抽样科学和样本具有较好代表性的必要条件。获得概率样本的重要前提是有完整、准确的抽样框。在一般的社会科学调查中，往往很难获得全部被访者完整、准确的抽样名单，这一方面是由于建立被访者抽样框需要高额成本，另一方面由于及时维护、更新抽样框中的个人信息非常困难。在具体抽样实施过程中，往往采取多阶段、分层、整群等抽样方法来确保在抽样科学的基础上降低抽样成本。在多阶段入户调查抽样设计中，最常用的抽取最终被访者的方法是：首先根据家庭户抽样框抽取家庭户，然后再对户内适合的调查对象进行抽样。由于家庭户规模大小不同，不同规模家庭中适合的调查对象被抽中的概率也不同，这样就造成了入户抽样后被访者特定指标分布与总体分布不同的问题。对于这种问题的解决方法只能是通过对样本进行概率加权，但由于多阶段、分层、整群等抽样设计的复杂性以及无应答等问题，对每个样本进行加权变得非常复

* 本文为中国社会科学院重大项目“2008 年中国社会状况综合调查”课题阶段性成果。

杂。因此，在入户抽样阶段应尽量减少抽样偏差。为了确保被访者抽样偏差的最小或样本加权的简单易行，需要研究一套科学可行的方法来解决上述问题。

Kish 表是基什(L. Kish)针对入户抽样的上述问题在 20 世纪 40 年代末根据美国的人口和家庭情况设计的。目前 Kish 表已经广泛地应用在世界各国入户抽样调查中。然而 20 世纪中期设计的 Kish 表是否适合目前中国国情或其他人口和家庭特征，需要仔细检验和深入研究。本文就是从 Kish 表在中国的具体应用问题出发，研究中国目前应用 Kish 表入户抽样被访者年龄结构扭曲问题并提出解决的办法，目的是尽量减少入户调查的抽样偏差。

二、研究数据来源与方法

(一)研究数据来源

为了研究应用 Kish 表在入户抽样数据的代表性和可能的系统偏差以及在当前中国的调查实地操作时所面临的问题，本文以“2008 年中国社会状况综合调查”入户登记表数据为例，分析被访者抽样分布偏差的来源。

“2008 年中国社会状况综合调查”是中国社科院社会学研究所于 2008 年 5 月至 9 月实施的，本调查采用多阶段、分层、系统抽样方式，成功入户访问了 7139 位年龄在 18－69 岁的城乡居民(其中 7046 位被访者是应用 Kish 表在家庭户中抽样获得的)，样本覆盖全国 28 个省市自治区的 134 个县(市、区)、251 个乡(镇、街道)和 523 个村(居委会)。这次调查是以 2000 年人口普查的县(市、区)统计资料为基础进行抽样框设计，具体抽样过程是：第一步，采用城镇人口比例、居民年龄、教育程度、产业比例 4 大类指标 7 个变量，对东中西部的 2797 个县(市、区)进行了聚类分层，在划分好的 37 个层中，采用 PPS 方法抽取 134 个县(市、区)；第二步，在抽中的每一个县(市、区)中，采用 PPS 方法抽取 2 个乡(镇、街道)；第三步，在抽中的每一个乡(镇、街道)中采用 PPS 方法抽中 2 个村(居委会)；第四步，收集抽中村(居委会)中所有居民个人或家庭的名单资料；第五步，在此抽样框中，采取 PPS 方法抽中被访住户。对于一户中有多个家庭居住的，按随机数表抽取其中一个家庭访问；如

果抽中的住户是集体户，则按集体户抽样，使用随机数表抽取被访者；第六步，对于抽中家庭，将该家庭中所有人的情况填在《家庭人口登记表》中，包括与答话人的关系、性别、年龄；第七步，把《家庭人口登记表》中 18—69 岁并且可接受访问的人口按“先排男性，后排女性，在同一性别中，按年龄由大到小排列”的规则进行排序，并按此顺序将成员的性别和年龄填在《Kish 选样表》中；第八步，用 Kish 表进行入户抽样。根据上述抽样步骤和方法，抽取被访者，入户登记的基本情况见表 1。

表 1　　调查登记（家庭户抽样部分）基本结果

类别	人数	比例（%）
登记人口（0—98 岁）	27338	
登记人口（18—69 岁）	21115	100.0
其中：可以接受访问	14948	70.8
家庭户中抽中被访者	7046	
不能接受访问	6170	29.2
长期出差	241	1.1
外出打工	3557	16.8
外出上学	701	3.3
外出参军	67	.3
临时生病	215	1.0
残疾不能接受访问	125	.6
其他	1221	5.8
不清楚	43	.2

本项研究之所以采用上述数据，一方面是由于本次调查的抽样设计完全按照概率样本的抽样调查进行科学设计，另一方面是本次调查的数据除了包括被访者的调查信息外，还包括家庭登记人口的信息、可接受访问者与不可接受访问者信息、Kish 选样表登记人口信息等。这些数据为 Kish 选样过程的研究提供了非常丰富和翔实的原始个案数据资料，使该项研究成为可能。

（二）Kish 选样表的基本原理与发展

由于直接获得个人名单在绝大多数调查中不仅存在数据获取困难，而且存在数据质量问题，所以目前调查户内个人通常采用 Kish 选样表（以下简称 Kish 表）来进行户内抽样。正如基什所说，入户抽样不但可以避免住户中的被访者有机会对问题展开讨论，而且同一户内的

被访者对某些问题的回答会相似，同时也避免同一户内多次访问(Kish，1949；Kish 表选样过程和基本原理见 Kish，1965)。

Kish 表的优点是在理论上坚持随机抽样，而且经过巧妙设计，使每一位适合的调查候选对象有不为零的入选概率。对于入户抽样，Deming 表(Deming，1960)是 12 种表格轮流使用，效果与 Kish 表(1949)差不多。电话调查的发展也对入户抽样提出了不同的要求，Kish 表在家庭人口登记和选样过程中的复杂性上对电话调查的形式提出挑战，在这一领域发展了一些新的调查方法来进行户中选样，在保证代表性的同时向相对简化的方向发展。T—C 方法(Troldahl &Carter，1964)是将 Kish 表(1949)的 8 个表简化为 4 个，并对性别加以控制，操作简单了，被应用在电话访问上，但这种方法过于简化，样本的代表性也有扭曲(Bryant，1975)。随着电话访问的应用日益广泛，有研究者(Bryant，1975；Hagan &Collier，1983)先后对 T—C 方法进行了改进，被称为 T—C—B—H 方法，被广为采用。另外，最近生日法(Salmon &Nichols，1983)因为操作简便在欧美很多著名的民意调查中被使用(以上均转引自洪永泰，1996)。

对于 Kish 表的使用也有研究者指出，其表格复杂，需要先调查被访户中的人口结构，才能确定要选谁为被访者，所以对访员的训练和素质要求较高；此外，过于复杂的表格以及对被访家庭姓名、性别和年龄的询问一方面增加访问者和被访者的负担，另一方面入户后的合格者的登记既费时也冒犯被访家庭的隐私，而且也容易造成拒访(洪永泰，1996)。还有研究指出 Kish 表在不同文化国家的适用性问题。基于西方社会的 Kish 表在津巴布韦的实际调查中，由于家庭人口多、扩大家庭比较常见等问题，使选样表中登记的人口规模非常大；同时入户登记还存在年长的答话者无法记清年龄的问题；还会由于家庭中的权威者是对外的“发言人”，而随机选出的被访者对回答问卷无所适从，等等，所以使用 Kish 表需要考虑调查国家的文化背景(McBurney，1988)。匈牙利研究者尼密斯(Nemeth，2002)对针对美国人口的年龄结构所设计的 Kish 表在其他国家的适应性问题提出质疑，并根据自己国家的情况对 Kish 表中数字的排列顺序进行了调整。

(三)评价方法

为了评价应用 Kish 表入户选样可能存在的问题，本项研究主要采

用概率论的基本分析方法，除此之外，在分析样本的偏差或代表性时，把年龄和性别作为指标分析某一年龄性别的人口在样本中的分布是否与总体一致并进行比较，即

$$a_{[i,j]} - A_{[i,j]} \to 0$$

其中：$a_{[i,j]} = \frac{age_{i,j}}{tpop}$，$A_{i,j} = \frac{AGE[i,j]}{Tpop}$，$i = 18, 19 \cdots\cdots 69$，$j = 1, 2$

$a_{[i,j]}$ 为样本的年龄结构，i 为年龄，j 为性别，如 $a_{[24\,1]}$ 为样本中所有 24 岁男性在调查人口中的比例。$A_{[i,j]}$ 为总体的年龄结构。样本与总体的年龄结构如果一致，则二者的差为 0。

为了更有效地衡量不同抽样方案的代表性，引入一个指标，即寻找抽样后的年龄结构 $a_{[i,j]}$ 与 Kish 选样表的年龄结构 $A_{[i,j]}$ 误差最小的方案，用 e 来表示，把 e 称为离差系数，e 的值越低，表示样本与总体的差异越小。

$$e = \sqrt{\frac{\sum (a_{[i,j]} - A_{[i,j]})^2}{n}} \times 10000$$

三、Kish 表抽样的概率分析

(一)调查登记人口与调查对象年龄结构比较

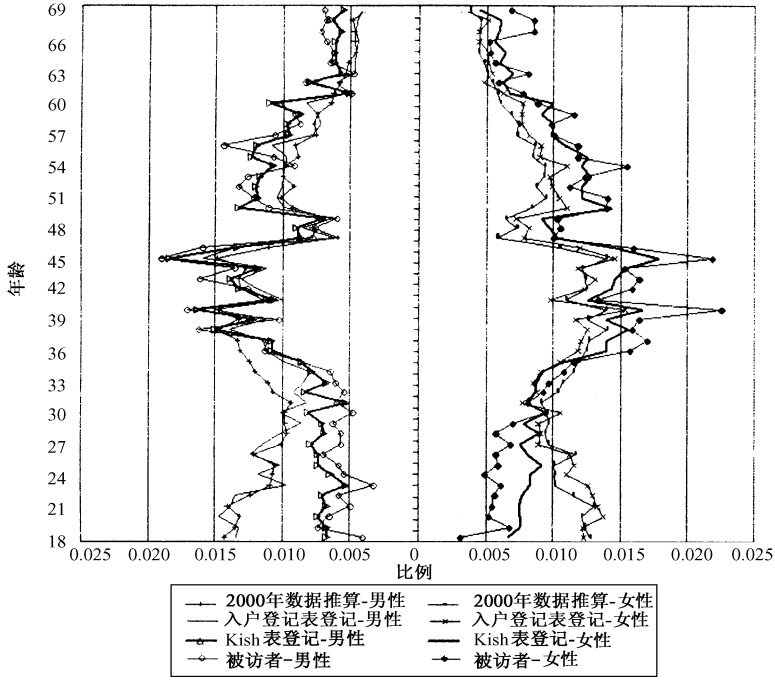
为了分析样本的代表性，把年龄和性别作为指标分析某一年龄性别的人口在样本与总体中的分布是否一致，把入户登记表、入户选样表和被访者三组人口的年龄结构与 2000 年人口普查相比，发现 2008 年调查的入户登记表登记的家庭成员的年龄结构除了女性 20－25 岁比例偏高、男性 31－37 岁比例偏低以外，其他年龄的分布与 2000 年的数据非常一致，也就是说，入户登记表中登记的人口基本能反映总体的年龄结构。

虽然入户登记表中登记的人口年龄结构与人口普查的年龄结构非常接近，但入户登记表人口与入户选样表人口年龄结构之间的差距很大，尤其是青壮年人口比例明显偏低。从表 1 可见，在实际调查时，18－69 岁人口为 21115 人，在登记人口中可以接受访问的人口为 14948 人，占 18－69 岁登记人口的 70.8%，有 29.2%的人由于各种原因不能

接受访问，其中属于无法接触的包括出差、打工、上学参军等，在入户登记表成员中的比例分别为 1. 1%、16. 8%和 3. 6%，外出打工人口比例最高。还有部分成员属于无能力回答，包括临时生病、残疾等原因，比例很低，分别为 1%和 0. 6%。

从年龄结构来看，在不能接受访问的 6170 人中，18—39 岁人口中为4444 人，占不能接受访问的人口的 72%（其中的 46. 5%为外出打工），而且年龄越低，比例越高。另外，外出上学的人口也集中在 24 岁以下，这样在调查中，登记人口中无法接受访问者主要集中在青壮年人口中，直接造成这部分人在可以接受调查的人口中比例偏低，虽然最初的入户登记与总体分布基本一致的，但很多低龄人口无法接受访问。

通过入户选样，抽出被访者，从图 1 发现，与入户登记表的人口年龄结构相比，入选 Kish 选样表的青壮年人口比例已经减少很多，更进一步，



注：图中 2000 年人口普查数据的年龄为 2000 年某年龄人口在 2008 年的年龄，如图中 18 岁为 2000 年 13 岁人口推算而来。

图 **1** **2008** 年全国社会状况调查入户登记表、**Kish** 选样表及被访者年龄结构

在入户选样后这一人口比例继续减少，也就是说，Kish 选样表进行入户抽样后，被访者的年龄结构与选样表中人口的年龄结构并不吻合。

为了分析 Kish 选样表的登记人口与选样后人口结构存在差异的问题，我们把 Kish 选样表登记人口数据假设为总体，研究抽样过程中样本与总体之间的差异。

在访问之前，每一份问卷的 Kish 表选择 8 个表中的哪一个表都是事先指定好的，但是在调查后发现，有接近 2%的问卷不是使用事先指定的表号，这与调查时的一些实际情况有关，例如实际调查中给定的问卷编号访问时不一定都是家庭户，有时遇到集体户，采用的是集体户的抽样方式。另外，调查期间因为特殊情况对样本进行了调整，造成家庭户中不同表号分配的比例有所变化。为了分析 Kish 表的选样过程，对现有的 7046 个家庭户重新分配 Kish 选样表的表号，不同类型的表可以按照设计时相同的概率被抽中，这样获得的数据作为假设总体的模拟数据分析入户抽样过程。

（二）户内选样与直接抽取个人的概率对比

对入户选样过程中不同特征的人的入选概率如何计算？基什对 Kish 表更多的是从操作流程介绍，并没有明确指出不同年龄概率的计算方法以及影响概率的具体参数，所以需要对选样过程与年龄别入选概率之间的关系进行探讨。

关于入户抽样的代表性，洪永泰对按户抽样和按人抽样的代表性进行了分析，指出先抽户再抽人造成被抽中者作为个体的代表性受到扭曲，相关的研究也证实这一点（转引自洪永泰，1996），从本文的模拟数据来看，如果把 Kish 表中的人口结构当作总体的结构，在调查中 7046 户的 14948 人中，以 24 岁男性为例，24 岁男性 98 人，如果不考虑户，直接抽人，抽中概率为 0.0066，而如果在户中抽人，24 岁男性的抽中概率为 0.0050（具体计算可以采用公式（1），结果见表 2），可见先抽户再抽人改变了人的入选概率。

$$a_{[i,j]} = \sum_{k=1,2\cdots,6} H_k \frac{age_{k[i,j]}}{tpop_k} (i = 18\cdots69,\ j = 1,2,\ k = 1,2,\cdots,6) \quad (1)$$

公式（1）中，i 为年龄，j 为性别；$a_{[i,j]}$ 为样本的某一年龄性别的入选概率；k 为入户选样表中人数；H_k 为总体中入户选样表人数从 1 到 k

人的户数比例；$age_{k[i,j]}$ 是不同人数的入户选样表中某一年龄性别的人数；$tpop_k$ 是不同人数的入户选样表中人数。

表 2 按户人口数分类的入选概率

选样表入选人数 k	户数	户比例 H_k	人数 $tpop_k$	人比例	24 岁男性 $age_{k[24,1]}$	24 岁男性在户内的比例	24 岁男性被抽中的实际比例
1	1846	. 262	1846	. 123	6	. 0033	. 0009
2	3342	. 474	6684	. 447	14	. 0021	. 0010
3	1155	. 164	3465	. 232	37	. 0107	. 0018
4	585	. 083	2340	. 157	30	. 0128	. 0011
5	95	. 014	475	. 032	9	. 0189	. 0003
6	23	. 003	138	. 009	2	. 0145	. 0000
	7046	1	14948	1	98	. 0066	. 0050

(三)家庭结构与 Kish 表之间关系的概率分析

Kish 表设计了用 A、B1、B2、C、D、E1、E2、F 代表 8 种抽选表，从设计角度来看，按照给定的概率分配表号在选样后，住户中可接受访问的人都有相同的概率被抽中(见表 3p_{kl} 理论分布值)。

为了分析选样过程的入选概率，运用模拟数据按照 Kish 表对抽样过程进行仿真，模拟实际调查的操作流程，抽出模拟的样本。在模拟样本数据中 24 岁男性有 25 人，计算 $a_{[24,1]}$ 在样本中的抽中概率，$a_{[24\ 1]}$ = 25/7046=0. 0036，这说明在模拟数据中经过 Kish 表选样后，24 岁男性被抽中的概率降低了。

Kish 表的抽样与选样表中的人数、不同年龄、性别的人在选样表中的位置密切相关。使用 Kish 表抽样后，计算入选概率的因素其实是更加复杂了，匈牙利学者尼密斯使用公式(2)对某一年龄的入选概率的进行了分析。

$$a_{[i,j]} = \sum_{k=1,2,\cdots,6} H_k \left(\sum_{l=1,2,\cdots,k} p_{kl} a_{k[i,j]} \right)$$

$$(i = 18 \cdots 69,\ j = 1, 2,\ k = 1, 2, \cdots, 6,\ l = 1, 2, \cdots, k) \qquad (2)$$

公式(2)中：i 为年龄；j 为性别；$a_{[i,j]}$ 为样本的某一年龄性别的入

选概率，k 为入户选样表中人数；H_k 为规模为 k 的住户的入选比例；l 为入户选样表中的位次；p_{kl} 是选样表中 k 人中成人 l 的入选概率；$a_{k[i,j]}$ 为 k 人中入选的第 l 人年龄为 i 性别为 j 的概率，其中 $a_{k[i,j]}=\frac{age_{k[i,j]}}{tpop_{kl}}$；$p_{kl}$ 是 k 人住户中第 l 人的入选概率（$k=1, 2, \cdots, 6$，$l=1, 2, \cdots, k$）。

由公式(2)可见，通过 Kish 表选中的被访者与以下变量有关：

1）住户中可接受访问的人数 k；

2）k 个人的住户中第 l 人在入户选样表中的位置（先排男性，后排女性；在同一性别中，先排年龄大者，后排年龄小者）；

3）该住户被分配的抽样表号，不同的抽样表中可以抽中的被访人的编号是不同的。如果是分配表 A，那么不论是住户中有几人接受访问，都是排在第 1 位的被访者接受访问。

按照 Kish 表的设计，理论上住户中可接受访问的人都有相同的概率被抽中（见表 3 p_{kl} 理论分布），如 3 人户家庭中 p_{31}、p_{32}、p_{33} 都是 0.333。但实际抽出的结果（见表 3 p_{kl} 实际分布）与理论设计还是有些差异，例如在我们的调查中，住户中可接受访问的人数为 3 时，这 3 个人的入选概率并不都是 0.333，而分别是 0.333、0.323 和 0.344；可接受访问为 4 人时，4 人入选概率也不都是 0.25，而是第 1 人和第 3 人入选概率高一些，p_{41}、p_{43} 分别为 0.279 和 0.255。

按照尼密斯（Renata Nemeth）的公式(2)，用表 3 的数据计算入选概率，计算过程分别使用 p_{kl} 的理论分布和实际分布。

1）按照 p_{kl} 的理论分布计算，$a_{[24,1]}=0.00493$；

2）按照 p_{kl} 的实际分布计算，$a_{[24,1]}=0.00473$。

实际上，采用 Kish 表进行模拟抽样，$a_{[24,1]}$ 的结果是 0.0036。就是说，使用公式(2)计算的入选概率与实际结果存在着差异，需要按照选样的步骤进一步分析公式(2)。在公式(2)使用的各个参数中，H_k 是固定不变的，每个人在选样表中的顺序也是固定的，那么 p_{kl} 是否可以进一步细分呢？

事先指定 Kish 表号目的是保证所有家庭户中选取的 Kish 表号的比例与设计时一致，但是并不知道分配到指定某一表号的家庭有多少人入选，在使用 Kish 表实际抽样时，利用模拟数据分析发现 p_{kl} 的实际

表 3 24 岁男性 Kish 选样入选概率分析

k	H_k	l	pop_{kl}	$age_{kl[24.1]}$	p_{kl} 理论分布	p_{kl} 实际分布
1	. 262	1	1846	6	1	1
2	. 474	1	3342	11	. 5	. 497
		2	3342	3	. 5	. 503
3	. 164	1	1155	5	. 333	. 333
		2	1155	30	. 333	. 323
		3	1155	2	. 333	. 344
4	. 083	1	585	1	. 25	. 279
		2	585	24	. 25	. 219
		3	585	5	. 25	. 255
		4	585		. 25	. 248
5	. 014	1	95		. 167	. 137
		2	95	7	. 167	. 126
		3	95	2	. 25	. 274
		4	95		. 167	. 147
		5	95		. 167	. 316
6	. 003	1	23		. 167	. 130
		2	23	1	. 167	. 130
		3	23	1	. 167	. 130
		4	23		. 167	. 217
		5	23		. 167	. 261
		6	23		. 167	. 130

分布与选中的 Kish 选样表的表号密切相关。例如入选的人数 k 为 3 的家庭，在 Kish 表号不同时，抽中的第 l 人也不相同，所以 k 人中第 l 人的入选概率 p_{kl} 应为：

$$p_{kl} = \sum p_{kl(kish)} \tag{3}$$

公式(3)中：$k=1, 2, \cdots, 6$，$l=1, 2, \cdots, k$；kish 为 Kish 选样表中的表号 A，B1，B2…F。

表 4 是以入选人数 k 等于 3 为例时，选样表中入选人数的分布与入选概率情况。利用公式(3)以 $k=3$，$l=3$ 为例计算入选概率。当

Kish 表号为 E1, E2 和 F 时, 选样表中 k 为 3 人中的第 3 人入选概率为: $p_{33} = p_{33(E1)} + p_{33(E2)} + p_{33(F)} = 0.081 + 0.088 + 0.174 = 0.344$, 而 p_{33} 的理论值应为 0.167 + 0.083 + 0.083 = 0.333。

p_{kl} 与 Kish 选样表中分配的选样表号有关, 那么公式(2)可以改进为:

$$a_{[i,j]} = \sum_{k=1..6} H_k \left(\sum_{l=1..k} p_{kl(kish)} a_{kl(kish)[i,j]} \right) \quad (4)$$

以 $a_{[24,1]}$ 的计算为例, 使用对公式(2)改进后的公式(4)计算模拟数据中某一年龄性别人口的抽中概率, 从表 4 第(3)部分可知 Kish 表中登记的 24 岁男性人口的分布情况, 分别从表 4 的第(1)和第(2)部分得到 $tpop_{kl(kish)}$ 和 p_{kl}, 按照公式(4)计算 $a_{[24,1]}$。如 $age_{32(c)[24,1]}$ 为 3, 且 $tpop_{32(c)}$ 为 192, $a_{32(c)[24,1]}$ 为 3/192, 与 $p_{32(c)}$ 为 0.166 相乘, 依此类推, 计算结果 $a_{[24,1]}$ 为 0.0036。这个结果与实际的仿真抽样的结果是相同的, 就是说, 使用公式(4)可以计算出与抽样模拟一致的入选概率。

表 **4**　　选样表中入选人数分布与 **24** 岁男性被访者分布

住户中可接受访问的人数(k=3 为例)	被选中人编号(1)	Kish 选样表号								
		A	B1	B2	C	D	E1	E2	F	总计
(1)选样表中被抽中人分布	1	186	93	106					385	
	2				192	181				373
	3						94	102	201	397
(2)选样表中被抽中人入选概率	1	.161	.081	.092						.333
	2				.166	.157				.323
	3						.081	.088	.174	.344
(3)24 岁男性入选情况	1	1	1	1					3	5
	2	5	5	5	3	3	1	1	7	30
	3					1			1	2

从不同家庭 Kish 选样表中成员的入选概率 p_{kl} 的理论分布到实际分布, 都无法真实地模拟总体中年龄别入选概率, 而在 p_{kl} 把每个成员在 Kish 表中的位置以及能否入选考虑进来时, 则把选样过程真实地模拟了出来。家庭选样表成员结构、年龄等都是计算年龄别入选概率的

重要参数，可以尝试改变这些参数值来分析入户抽样后样本与总体的差异。

四、Kish 表应用改进的仿真分析

基什在设计选样表时参照的美国 20 世纪 40—50 年代增长型的人口年龄结构，与我国目前人口分布以及我们调查时入户选样表登记人口的年龄结构已经明显不同（见表 5、表 6）。拿我们调查使用的 18—69 岁分组与普查时的年龄结构相比，由于外出务工、上学等原因，选样表登记的可以接受访问的年轻人比例低，而中老年人口比例高。在经过 Kish 表入户抽样后，年轻人比例进一步降低，而老年人比例继续提高。

表 **5** 不同来源数据年龄分布状况

年龄	美国(1946)	2000 年人口普查
21—29	22. 8	18. 8
30—44	33. 9	34. 8
45—59	25. 6	19. 4
60 岁及以上	17. 7	20. 9

表 **6** Kish 选样表及抽样结果与人口普查年龄结构比较

年龄	2000 人口普查	Kish 选样表	Kish 表模拟抽样结果
18—29	25. 4	17. 8	14. 2
30—44	38. 1	38. 1	37. 3
45—59	27. 4	34. 8	34. 3
60—69	10. 7	13. 0	14. 2

Kish 表中的人口是按照性别和年龄分层后排列的，原则是“先排男，后排女，同一性别，按年龄由大到小”，这样的排列顺序是否会增加年长者和男性的机会呢？我们改变 Kish 选样表的排列顺序，采用几种抽样方式进行仿真，对比样本与 Kish 选样表中登记人口假设总体之间年龄结构的差异，目的是希望入户抽样的结果能够真实地反映登记表的年龄结构。

以下几种抽样方案主要是改变不同家庭中入选 Kish 表中的人的排列顺序，即：

1）原有抽样方式是在选样表中按照男在前、女在后，同一性别年龄大在前，年龄小在后的顺序排列后的模拟抽样结果，这个方案简称为原方案。

2）方案一是按照男在前、女在后，同一性别年龄小在前，年龄大在后。与原有方案相比改变的是年龄的排序，即同一性别中把年龄小排在前面。

3）方案二是按照不考虑性别，年龄小在前，年龄大在后。与原有方案相比没有考虑性别，直接按照年龄排序把年龄小排在前面。

（一）对不同仿真方案的总体评价

根据不同方案对模拟数据进行仿真，然后对不同方案仿真结果中的年龄结构、离差系数及性别比等指标进行评价。

第一，从年龄结构来看，汇总样本的 $a_{[i,j]}$ 后，按照原方案 Kish 表设定的抽样方式，与 Kish 选样表的登记人口的年龄结构相比，33 岁以下样本的比例偏低，例如，登记表中 18－33 岁男性在所有人口中的比例为 11.2%，而原方案中的比例则为 8.9%，按照方案一和方案二中分别为 9.2%和 9.1%，青壮年人口的入选比例都要高一些。

第二，从离差系数 e 来看，Kish 表规定的抽样方案样本与总体年龄结构的离差系数分别为 12.4 和 14.8；方案一把最年轻的男性排列在最前面，男性的 e 最小为 11.2，但女性的 e 最高，为 14.9；而方案二的男性和女性的 e 分别为 14.7 和 14.4。从不同年龄组的 e 来看，方案二中 18－33 岁组的分性别的 e 在几种抽样方案中最低，也就是说，这一方案抽样后青壮年人口样本与登记表中的年龄结构相对接近，34－49 岁和 50－59 岁的女性的 e 也是最低的，但是方案二中 34－49 岁和 50－59 岁的男性的 e 比另外两个方案高。

第三，除了考察不同年龄性别人口在总体中的比例外，男女性别比与总体是否一致也是对各种方案考察的一个重要指标，对比几种方案，方案二（不考虑性别，直接按照年龄从低到高排序）的性别比与 Kish 表中登记人口的性别比相差最小，而原方案的性别比为 86.5，低于登记人口性别比，即样本中女性比例高。分年龄组来看，方案二的性别比虽然与登记表有所差异，但是差异也最小，男性入选比例在几种抽样方案

中都是最高的。

从改变 Kish 选样表中人的排列顺序的仿真方案结果来看，在模拟数据中人的不同排列方式会对样本的结构产生影响，也就是说在模拟数据中，被选中人不是等概率抽中的，被选中人的编号位置不同，选中的概率也不一样。仿真结果显示，分层的方式不同，对年龄和性别结构的影响也不同（见图 2）。

表 **7** 不同方案样本年龄结构

	原方案		方案一		方案二		Kish 选样表登记	
	男	女	男	女	男	女	男	女
18—33 岁	. 089	. 114	. 092	. 110	. 091	. 109	. 112	. 129
34—49 岁	. 197	. 236	. 196	. 238	. 203	. 228	. 188	. 213
50—69 岁	. 178	. 186	. 175	. 189	. 182	. 187	. 178	. 181
18—69 岁	. 464	. 536	. 464	. 536	. 476	. 524	. 477	. 523

表 **8** 不同选样方案离差系数

	原方案		方案一		方案二	
	男	女	男	女	男	女
18—33 岁	16. 6	14. 9	15. 3	15. 3	14. 7	14. 4
34—49 岁	11. 6	17. 8	10. 6	18. 7	12. 7	13. 4
50—69 岁	8. 0	11. 1	6. 4	10. 1	8. 4	9. 9
18—69 岁	12. 4	14. 8	11. 2	14. 9	12. 1	12. 6

表 **9** 选样表登记人口与不同方案样本性别比 （女= 100）

	原方案	方案一	方案二	kish 选样表登记人口
18—33 岁	77. 5	83. 6	83. 4	86. 7
34—49 岁	83. 6	82. 6	89. 2	88. 4
50—69 岁	95. 5	92. 9	97. 2	98. 1
18—69 岁	86. 5	86. 5	90. 8	91. 3

（二）不同仿真方案中排列顺序对单一年龄组的抽中概率影响分析

前面的分析是对不同方案抽样结果的评价，为了更清晰地分析这

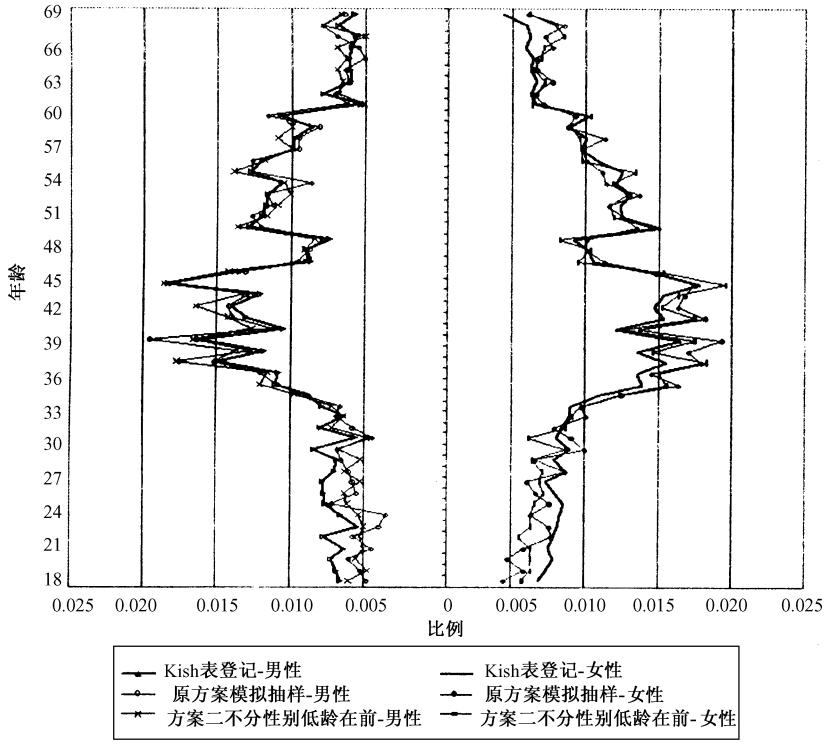


图 **2**　不同抽样方案年龄结构

些方案对于青壮年人口入选概率的影响，以 **24** 岁人口为例进行剖析。由表10可见，不同的抽样方案中这一年龄组的入选人口是不同的，无论哪种方式都无法保证这一年龄在样本中的分布与总体一致，只能是差异最小。

从具体的抽样过程来看，在不同抽样方案中 24 岁人口在选样表中的顺序有所不同，直接影响抽样结果(详见表 11)。

首先，从选样表中 k 为 1 人的家庭来看，不管是什么方案，他们都是被选中者。而入户选样表中只有 1 人能接受访问的比例过高，在所有的被访者中超过了 1/4，而他们中女性接近 60%，从年龄上看，超过 83%是 35 岁以上，也就是说，无论采用什么样的抽样方式，有 1/4 的人是肯定要被抽中的，如果希望提高年轻男性的入选比例的话，这 1/4 的人会对样本的年龄和性别结构产生很大影响，而改变抽样方案只能是调整另外 3/4 的家庭中接受访问者的年龄和性别结构。

表 **10** **24** 岁人口在选样表及不同抽样方案中的分布情况

	总人数	24 岁人数		概率	
		男	女	男	女
Kish 表登记人口	14948	98	124	. 0066	. 0083
Kish 表原抽样方案	7046	25	44	. 0036	. 0062
方案一	7046	37	45	. 0053	. 0064
方案二	7046	38	44	. 0054	. 0063

表 **11** **24** 岁人口在不同方案下的分布

k	*l*	原方案	方案一	方案二	*k*	*l*	原方案	方案一	方案二
1	1	6	6	6	5	2	7	3	4
2	1	11	14	10		3	2		3
	2	3		4		4			
3	1	5	37	33		5			
	2	30		4	6	1		1	
	3	2				2	1	1	2
4	1	1	27	13		3	1		
	2	24	3	17		4			
	3	5				5			
	4					6			
5	1		6	2	总计		98	98	98

其次, 选样表中有超过 2 人的, 选样方案对他们的位置产生了影响, 以 *k* 为 3 为例, 选样表中登记了 24 岁的男性 37 名。在原方案中排在第 1、2、3 位的分别有 5 人、30 人和 2 人, 排在第 2 位的最多。方案一是按性别分层, 年龄最小排在最前, 37 人全部排在第 1 位。方案二不考虑性别, 直接按照年龄分层, 分别有 33 人和 4 人排在第 1 和第 2 位。从抽中情况看, 原方案中被抽中的是排在第 1 位的选样表号为 A 和 B1 的 2 人、排在第 2 位的选样表号为 C 和 D 的 6 人以及排在第 3 位的选样表号为 F 的 1 人, 排在第 2 位的 30 人只有 6 人才能被抽中, 所以入选的人数相对较少。方案二和方案三中超过 30 人排在第 1 位, 其中表号为 A、B1、B2 的 17 人被抽中。

第三, 运用公式(2)分别计算 24 岁人口中处于选样表中不同人数的家庭中的抽样情况, 具体分布见图 3。选样表中有 2 人, 与原方案相

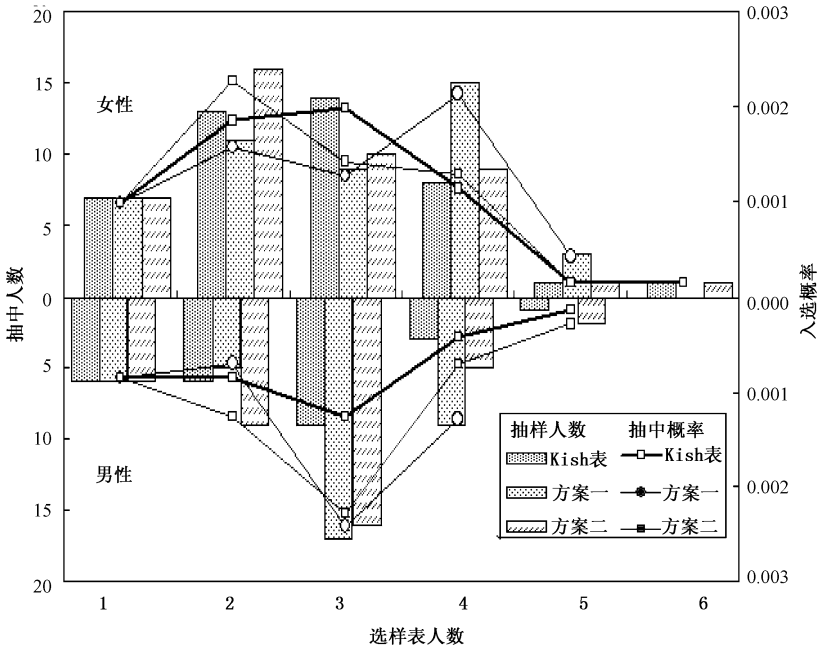


图 **3**　不同方案下 **24** 岁人口样本概率与人数

比，方案一男性变化不是很大，女性略有下降；方案二性别不分层，年龄从小到大排序后，样本中这一年龄的男性和女性都增加了，其中，2 人选样表的家庭男女的入选概率都提高了；3 人选样表中男性增加，女性下降；4 人选样表的家庭男女都略有增加，但幅度不大。

总之，从几种方案样本入选情况和对 24 岁这一年龄抽样过程进行分解和结果分析，把男性排在前面实际为增加了男性的入选机会，把年龄大的排在前面也是增加了他们的入选概率。方案一的分层排序方式是年轻男性入选比例大幅上升，方案二不做性别分层而是直接以人的年龄排序，从性别角度来看，样本的分布相对均衡。使用 Kish 表选样，改变选样表内的人的排列顺序，在模拟数据中，排序其实是改变了入选的机会，所以把总体中或可接受访问的比例较低的排在前面，这也可以理解为增加了他们入选的权重。但这种增加不是单纯的增加，与选样表登记人数的比例等关系非常密切。Kish 表规定的排序方式是在 20 世纪 40、50 年代按美国的年龄结构设计的，通过先排男性、先排年长者增加了这一部分人的入选概率。

表 12 不同方案入选情况(以 24 岁男性入户选样表中 3 人为例)

K= 3	位次 l	Kish 表号								
		A	B1	B2	C	D	E1	E2	F	
原方案	1	1		1					3	5
	2	5	5	5	3	3	1	1	7	30
	3					1			1	2
方案一	1	6	5	6	3	4	1	1	11	37
	2									
	3									
方案二	1	6	5	5	3	4	1	1	8	33
	2			1					3	4
	3									

注: 框内为抽中人口。

五、讨论与建议

首先, 从理论上看, Kish 表设计了 8 种抽选表, 样本按照给定的概率分配表号, 在选样后, 住户中可接受访问的人都有不为零的概率被抽中。然而, 从 Kish 表的基本原理和理论概率分布来看, Kish 表确实暗含对家庭成员被抽中概率加权的作用, 也就是, Kish 表对不同家庭结构(可以接受访问的人数的结构)中具有相同特征的人群抽中的概率不同。这种加权的作用实际上是调整被访者的缺失和无法访问所带来的偏差。因此, 户内抽样的家庭结构不同, 相同特征人群被抽中的概率不同, 这与选样表表号的分配和选样表登记人口的排列顺序有关。

其次, 从中国数据实证研究来看, 关于 Kish 表应用或误用造成年龄结构扭曲的问题得到了证实。实证数据研究结果表明, 在目前中国的入户抽样确实存在比较严重的年龄结构扭曲, 形成原因与人口流动造成青壮年人口比例过低有关。同时, 入户抽样进一步扭曲了年龄结构。

第三, 从仿真结果来看, 针对中国目前的实际情况和具体问题是可以进行改进和降低年龄结构扭曲问题的。最有效的改进方式之一是改变家庭成员在 Kish 表中的排序规则。对 Kish 选样表的仿真分析表明, 改变选样表中人的排列顺序, 把比例低的排在前面, 这样不但增加了这部分人的入选概率, 同时还可以对选样表中的其他部分做出尝试, 一类是改变 Kish 表数字的分布, 比如, 选样表内数字以随机数的形式出现。

还有一类是改变 Kish 表数字分布比例，如匈牙利研究者尼密斯的研究，而本项仿真研究认为，在不修改 Kish 表的情况下，改变选样表被访对象的排序规则也同样可以增加不易访问对象的入选概率。

鉴于 Kish 表在中国目前入户抽样可能存在的年龄结构扭曲问题，建议从操作流程上，增加可接受访问的人数、选样表号保持与设计时的一致（监控调查中没有使用的选样表号，以便在追加时轮换，保证随机性），对于家庭结构这一参数的改善主要是提高可以接受访问的人数，尤其是降低 1 人户的比例。

参考文献：

洪永泰，1996，《户中抽样之研究》，台北：五南图书出版公司。

Binson, D., J. A. Canchola & J. A. Catania 2000, "Random Selection in a Telephone Survey: A Comparison of the Kish, Next-birthday, and Last-birthday Methods." *Journal of Official Statistics* 16.

Bryant, B. E. 1975, "Respondent Selection in a Time of Changing Household Composition." *Journal of Marketing Research* 12.

Deming, W. E. 1960, *Sample Design in Business Research*. New York: John Wiley and Sons, Inc.

Hagan, D. E. & C. M. Collier 1982, "Must Respondent Selection Procedures for Telephone Surveys Be Invasive?" *Public Opinion Quarterly* 47.

Kish, Leslie 1949, "A Procedure for Objective Respondent Selection within the Household." *Journal of the American Statistical Association* 44.

—— 1965, *Survey Sampling*. New York: John Wiley and Sons, Inc.

Lavrakas, P. J. 1993, "Telephone Survey Methods: Sampling, Selection and Supervision." *Applied Social Research Methods Series* 7.

Levy, P. S. & S. Lemeshow 1999, *Sampling of Populations*. New York: John Wiley and Sons, Inc.

McBurney, Peter 1988, "On Transferring Statistical Techniques Across Cultures: The Kish Grid." *Current Anthropology* 29.

Nemeth, Renata 2002, "Respondent Selection within the Household - A Modification of the Kish Grid." (http://www.math.uni-klu.ac.at/stat/Tagungen/Ossiach/Nemeth.pdf)

Oldendick, R.W., G.G. Bishop, S.B. Sorenson & A.J. Tuchfarber 1988, "A Comparison of the Kish and Last Birthday Methods of Respondent Selection in Telephone Surveys." *Journal of Official Statistics* 4.

Salmon, C.T. & J.S. Nichols 1983, "The Next-birthday Method of Respondent Selection." *Public Opinion Quarterly* 47.

Troldahl, V.C & R.E Carter, Jr. 1964, "Random Selection of Respondents Within Households in Phone Survey." *Journal of Marketing Research* 1.

作者单位：中国社会科学院社会学研究所

责任编辑：杨　可

Abstract: After 30 years of reform and opening to the outside, Chinese institutions have changed greatly. During this process, we often find that a new institution, which was created by Tongbian mechanism from bottom to top, or created by central or local governments directly, should be pushed by some series of large-scale reform movements from up to down, and then put into effect from policy text. In this article, we will analyze those institution changing mechanisms and processes by three cases during past ten years in China. This mechanism is called Tongbian, and the process is separated into two steps: carrying out formal performance and achieving actual performance.

Abstract: Based on an empirical study in Jingshan county, Hubei province in September 2008, the paper launched a study of the elderly suicide. There were 206 suicide cases occurred in the past 30 years in six villages. The suicide rate and suicide proportion of rural elderly are alarmingly high, and they are also rising. In the light of change of inter-generational relations in current rural China, the paper summarizes the types of the elderly suicide, analyses the causes of high suicide rate and high suicide proportion of rural elderly, and shows their changing trends.

Abstract: Based on the survey data and probability theory, the paper studied the age structure of household during sampling survey. By comparing the age structure characters of the registered population, Kish table population and the respondent population, the author calculated the theoretical probability distribution of Kish table and found the reason of low proportion of the younger age group in the household sample table and the distorted age structure. For dealing with the problem of Kish table application, using the computer simulation methods, the author suggested the refine methods of the Kish table by reorder the distribution of Kish table population.

REVIEW