低年龄人口数据质量的分析与评价

翟振武　陶　涛

【摘　要】文章从人口统计数据系统存在的矛盾出发，对低年龄人口数据质量进行了分析与评价，并利用人口系统以外的教育数据来对人口数据进行了校验。结果表明，在考虑死亡的情况下，2000 年人口普查数据中对应的0～3 岁人口平均每个年龄组漏报了近 340 万人。这种漏报对中国生育水平、老龄化程度、未来劳动力供给状况、未来育龄人群人数、男女婴儿死亡率和出生性别比等重要人口指标的准确性产生了重要影响。

【关键词】低年龄人口　数据质量　教育数据　人口指标

【作　者】翟振武　中国人民大学人口与发展研究中心，教授；陶　涛　中国人民大学人口与发展研究中心，博士研究生。

一、背景和目的

数据是进行人口学指标测量的基础，其质量如何至关重要。目前中国公认的最权威的人口数据来源是人口普查、1%人口抽样调查和历年 1‰人口抽样调查，一方面因为它是全国性的数据，另一方面因为它是国家统计局组织实施、全民动员的结果。然而，即便如此，中国人口学界仍然在很多问题上颇具争议，有很多基础性问题仍然没有搞清楚。例如，中国的生育水平一直是个谜，有人认为目前的生育水平太高，应该更加严格地控制生育行为，也有人呼吁生育水平太低，需要赶快放开二胎。那么，生育水平到底是多少？是高了还是低了？应该采取怎样的措施才是适当的？再例如，中国出生性别比，尤其是省级出生性别比水平也是一个一直争论不休的问题。中国女婴的生存状况真的这么差吗？中国的性别失衡到底有多严重？会带来怎样的后果？此外，中国未来劳动力到底是多少？是短缺还是过剩？老龄化程度与速度如何？以上各种基本而重要的指标的确定和争论都与中国普查数据中的低年龄数据质量密切相关。因此，对普查数据中低年龄人口数据质量的评估是十分重要的，只有在对数据质量和误差有确切把握的情况下，才能正确和恰当地使用这些数据，并得出符合实际的结论。

关于人口漏报的问题目前已有很多学者进行过探索。周皓（2003）曾经在对 2000 年第五次人口普查数据的调查时点进行调整以后，重新估计了 1990 年第四次人口普查数据的漏报率，结论认为，1990 年“四普”的漏报率远高于“四普”公布的漏报率，其中低年龄段的漏报

非常严重。崔红艳、张为民(2002)从几个方面对第五次人口普查直接登记的 124 337 万人进行评估,指出普查登记人口确实存在一定程度的漏报,漏报人口超过 2 000 万。于学军(2002)对第五次人口普查有关数据进行粗略分析后发现,官方公布的人口总数与年龄结构存在一定程度的不吻合,并指出第五次人口普查公布的未成年人口数量过低,由此估计的总和生育率甚至低于政策生育水平,令人难以置信。乔晓春(2002)通过对全国和各省人口普查第一、二号公报的数据和文字进行解剖和分析,提出人口普查存在的一些问题,但注意力主要集中在暂住人口和总人口的漏登上,对低年龄人口没有给予特别关注。王广州(2003)以 1982 和 1990 人口普查资料为基础,运用人口存活分析方法对中国第五次人口普查数据存在的重报问题进行了分析和研究。王金营(2003)在估计 1990～2000 年分性别、年龄死亡概率的情况下,利用年龄移算方法对 1990 和 2000 年两次人口普查的漏报率给予评估,并对漏报人口进行年龄、性别拆分。然而,这些研究都是以传统"人口"数据为基础进行间接估计工作。传统的人口数据,无论是普查数据,还是抽样调查数据,无论是统计局组织的调查,还是计生委组织的调查,都不能解决出生人数、婴幼儿人数的大量漏报问题。这些数据,如果缺乏漏报程度的信息,就无法为研究提供可靠的支持。即使是以不同来源的人口调查数据进行相互校验,由于这也是"人口系统"内的数据,它们都无法避免共同的缺陷,即出生数据敏感,漏报严重。在这些数据基础上所作的各种估计的可靠性当然也就受到很大影响。张为民、崔红艳(2003)利用全国小学生入学人数反推历年出生人数的方法是一个新的很有价值的思路。但是,或许因为数据的可得性问题,他们只使用了每年《中国教育年鉴》公布的入学人数,而没有详细考察历年在校各年龄的小学生人数,所以,他们的估计和论证显得粗略。本文试图利用人口系统以外的教育数据来对人口数据进行详细校验,以期对普查数据中低年龄人口数据进行较为合理的评估。

二、人口统计数据的观察

(一)出生性别比忽高忽低,可能存在漏报

通过对 2000 年以来的中国人口出生性别比的观察可以看出,数据呈现不规律的忽高忽低(见图 1),而在历次普查数据中都可以发现,1～4 岁低年龄组人口的性别比异常偏高,而 5 岁及以上人口的性别比尽管依然偏高,但较 1～4 岁组偏低,并趋于稳定(见图 2、图 3),这预示着可能存在低年龄组女婴漏报的情况(翟振武、杨凡,2009)。

图 1　2000～2007 年中国人口出生性别比

(二)同批人 7 年间上升过千万,确实存在大量漏报

为验证这一情况,我们利用 2000～

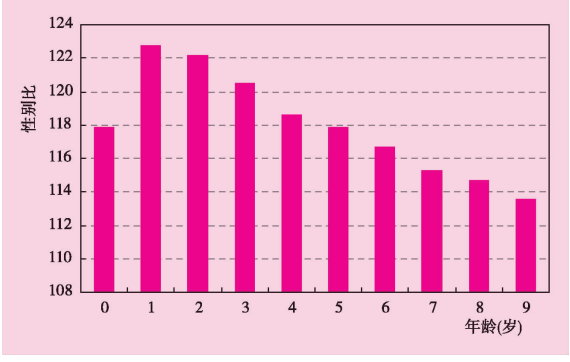

图 2　2000 年中国 0～9 岁人口分年龄性别比

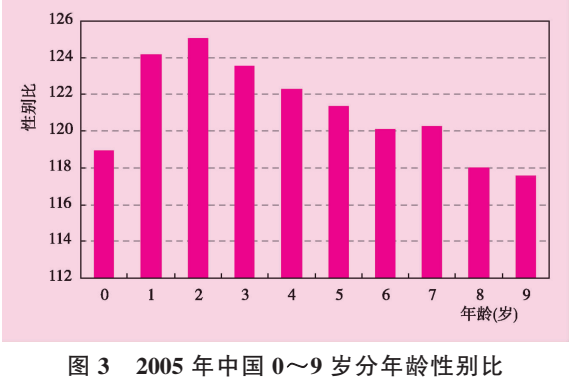

图 3　2005 年中国 0～9 岁分年龄性别比

2008 年《中国人口统计年鉴》公布的数据，根据各年的抽样比推断总体进行同批人的比较（由于单个年龄组的同批人数据可能会受到相邻年龄人口误报的影响，故采用 5 岁组数据进行比较），结果表明，2000 年 0～4 岁组人口为 6 898 万，而这批人的数量在各年的相应年龄组中逐渐上升，2007 年 7～11 岁组人口数已经上升至 8 169 万（见表 1）。亦即，在不考虑死亡影响的情况下，人口统计系统内的同一批人自出生以来的 7 年间上升了近 1 300 万，低年龄人口数据确实存在大量漏报的情况。

从图 4、图 5 可以看出，2000 年 0～4 岁 5 个年龄组的同批人，2000 年普查时约为 6 900 万，在 2007 年的统计中就变成了近 8 200 万，且每年的统计数据都比上一年要高出许多。考虑到死亡因素，每年同批人的统计数据应当逐年减少，即使由于抽样误差等原因有所波动，但连续 7 年每年都向同一个方向抽偏的概率是极小的。因此可以断定，低年龄组人口确实存在漏报，且漏报数量可观，8 年间 5 个年龄组共漏报了 1 271.3 万人，平均每个年龄组漏报 254.3 万人。由于 2007 年的 7～11 岁人口也还存在漏报的情况，因此，仅从人口系统本身的数据来看，低年龄组平均每个组就至少漏报了 250 万人。

关于利用抽样比反推人口数的方法，理论上是假定每个年龄抽样比与总样本的抽样比

表 1　历年人口统计数据回推

年份	2000 年 0～4 岁组被调查人口（人）			抽样比	反推总体（人）		
	男	女	合计		男	女	合计
2000（0～4）	37648694	31329680	68978374	1.000000	37648694	31329680	68978374
2001（1～5）	37294	31497	68791	0.000963	38726895	32707165	71434060
2002（2～6）	38893	32600	71493	0.000988	39365385	32995951	72361336
2003（3～7）	39818	33093	72911	0.000988	40301619	33494939	73796559
2004（4～8）	39662	33083	72745	0.000966	41057971	34247412	75305383
2005（5～9）	577003	483661	1060664	0.013250	43547396	36502717	80050113
2006（6～10）	40350	33321	73671	0.000907	44487321	36737596	81224917
2007（7～11）	40156	33366	73522	0.000900	44617778	37073333	81691111

一致，但在很多情况下，存在因实际调查的偏差而导致的该年龄段实际抽样比与当年总人口抽样比不太吻合的情况。对此，我们做以下几点说明。

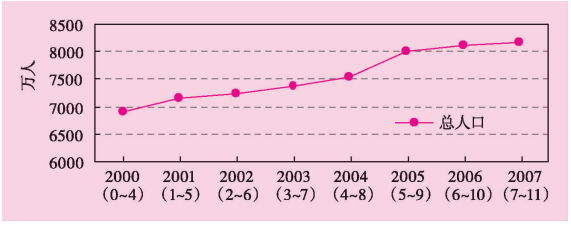

图 4　历年总人口统计数据回推

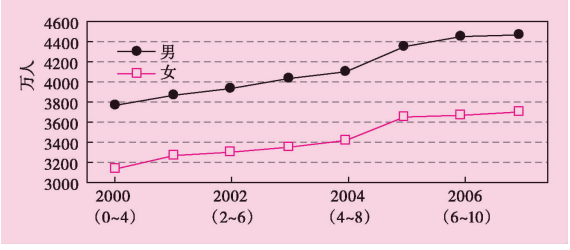

图 5　历年分性别人口统计数据回推

1.关于抽样比。由于历年 1‰人口抽样调查是采用多阶段、分层、整群、概率比例的抽样方法，调查是以调查小区为最终抽样单位，小区内全部的户都进行入户调查，户内的人口全部进行登记。从理论上讲，这种方法所抽出样本中的人口年龄结构应该是总体的一个无偏的缩影，样本的年龄结构和总体的年龄结构应当是一致的。也就是说，各个年龄段的抽样比应当同总人口的抽样比一致。

2.关于误差。由于实际抽样调查存在误差是不可避免的，因此，用样本中每个单岁年龄组的人数按抽样比推算这个年龄组全部人口时，也肯定会有误差，数值不会精确，但不妨碍我们的推断。首先，我们采用的是 2000 年 0～4 岁这个 5 岁组的同批人，亦即，通过扩大年龄范围，在一定程度上缓解了由抽样误差和非抽样误差所造成的误差。其次，由于调查时对调查小区内所有户的人口都进行调查，所以，低年龄人口人数的偏少，主要是漏报造成的。而随着该同批人的年龄增长，成为高年龄组后，漏报减少，在以后的调查中人数逐渐增多，这一过程恰好反映了被漏报的人口逐渐浮出水面的过程。 再次，也是最重要的，如果每年的每个年龄人口抽样存在误差和波动，那么这种误差和波动也应当是随机的、不规律的、有高有低的。如果某一年龄段人数在以后历次调查中都呈现趋势性的单调上升，那么，这种同批人总数逐年上升的现象就并不是由于抽样比误差的波动产生的偶然现象，而是当年调查时低年龄人口确实存在漏报。

3.关于研究目的。本文的目的并不是简单推算低年龄人口数在每一年的绝对水平，因为用一系列已经被怀疑漏报的调查样本去推断每一年低年龄人口总体的绝对水平肯定是不准确的。然而，尽管绝对水平不准确，但其相对变化所反映出来的规律性却是我们真正想关注的。我们的真正目的是利用这种人口系统内部调查数据所显现出来的同批人逐年上升的趋势来对低年龄组人口是否漏报，以及漏报的大致程度（不是精确水平）进行判断。

不仅 0～4 岁的低年龄组整体存在漏报的问题，即使将各个年龄组分开，无论是普查、1%人口抽样调查还是 1‰人口抽样调查，也无论任何年份，我们都能发现同样的现象，那就是任意低年龄组的同批人在任意统计年份的数量都会比上年的统计数量多出很多。也就是说，这并不是由于偶尔抽样偏差产生的巧合，而是确确实实存在低年龄组人口漏报。下面仅

表 2　2000 年与 2005 年单岁组同批人比较

年龄	2000 年(万人)			2005 年(万人)		
	总	男	女	总	男	女
0(5)	1379	746	633	1512	829	683
1(6)	1150	633	516	1473	804	670
2(7)	1401	770	631	1640	895	745
3(8)	1445	790	656	1661	899	762
4(9)	1522	826	697	1719	928	790

表 3　2001 年与 2007 年单岁组同批人比较

年龄	2001 年(万人)			2007 年(万人)		
	总	男	女	总	男	女
0(6)	1412.56	757.53	655.04	1481.67	800.11	681.56
1(7)	1250.05	682.76	567.19	1542.44	852.67	689.67
2(8)	1280.06	701.87	578.30	1498.56	815.11	683.33
3(9)	1479.65	797.51	682.14	1643.67	905.67	738.00
4(10)	1515.47	827.62	687.85	1699.33	926.33	773.00
5(11)	1618.17	862.93	755.35	1785.11	962.00	823.11

以 2000、2005 年单岁组比较和 2001、2007 年单岁组比较为例(见表 2、表 3)。

从表 2 可以看出，全国 1%人口抽样调查（2005 年）与人口普查数据(2000 年)相比，低年龄任意单岁组都在 5 年间有所上升。而从表 3 可以看出，两个 1‰抽样调查数据的比较也呈现出了同样的规律，低年龄任意单岁组都在 6 年间有所上升。再取任何年份的任何两个数据（无论是普查、1%还是1‰人口抽样调查)，也都遵从这一规律，无一例外(计算结果略)。这充分说明了漏报是真实存在的，而不是由于各年抽样偏差产生的偶然现象。

三、教育数据的推算

在中国各种分年龄人口数据中，除了“人口系统”内部来源数据外，还存在着独立于“人口系统”之外的数据，即教育统计数据。我们利用人口系统以外的相对独立的教育统计数据来进一步说明低年龄人口漏报的情况并进行推算。为排除早上学和早毕业的影响，我们选择 2007 年教育统计数据中的 7～10 岁人口进行分析，并与 2000 年 0～3 岁人口相比较。假设不考虑入学率的影响，最大可能地假设入学率为 100%，同时先不考虑死亡因素的影响。

如图 6～8 所示，虽然是同一批人，但 7 年后的教育数据平均每个年龄比普查时的人数多出 300 万左右，其中男性人数每组平均多出 150 多万，女性人数每组平均多出 165 万。具体到每个年龄组，同一批人在 2007 年教育统计中的人数分别比 2000 年普查中 0 岁、1 岁、2 岁、3 岁的人数多出 240 万、449 万、280 万、291 万人。其中男性分别多出 123 万、222 万、127 万、136 万，女性分别多出 118 万、227 万、153 万、155 万。

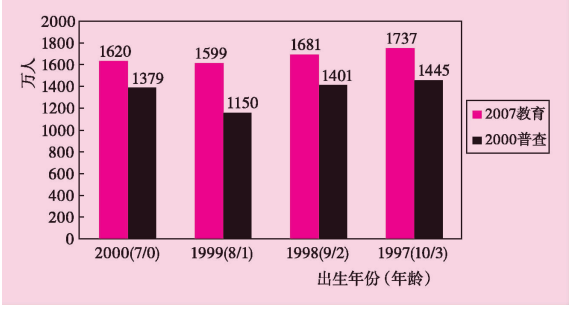

图 6　教育数据与普查数据之差(两性合计)

为使数据估算更为准确，我们将这几个年龄组的人按照 2000 年生命表推算回 0

岁，应当有 6 886 万人，而将 2000 年人口普查数据中对应的 0～3 岁人口也按照 2000 年生命表推算回 0 岁，却只有 5 527 万人，少了 1 350 多万人，平均每个年龄组漏报了近 340 万人（见表 4～6）。

四、对人口各指标的影响

假定教育数据反映的人数更真实，用教育数据将低年龄人口数据补齐，可以分析低年龄数据质量对各个人口指标的影响。结果表明：

1.由于漏报低年龄人口，在“人口统计系统”的数据基础上计算的生育率水平偏低了。根据 2000 年人口普查资料直接计算的 2000 年中国的总和生育率为 1.218。很显然，普查数据中 0 岁孩子的漏报使 2000 年的总和生育率大大偏低了。利用教育数据和其他来源数据校正和调整 2000 年的总和生育率，一些学者做过很多研究，尽管方法各异，但所获得的结果远高于直接计算的 2000 年总和生育率（1.218）。

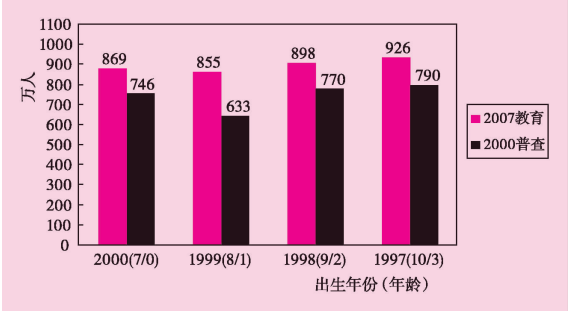

图 7　男性教育数据与普查数据之差

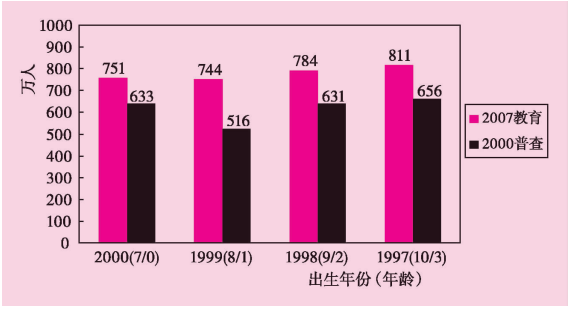

图 8　女性教育数据与普查数据之差

表 4　2007 年 7～10 岁人口数（教育数据）

	7 岁		8 岁		9 岁		10 岁	
	男	女	男	女	男	女	男	女
2007 年	8688311	7509565	8549795	7435315	8976131	7835995	9259918	8107177
还原至 0 岁	8987923	7804821	8851178	7731308	9298896	8151456	9598953	8437066

表 5　2007 年 7～10 岁人口数（普查数据）

	0 岁		1 岁		2 岁		3 岁	
	男	女	男	女	男	女	男	女
2000 年	7460206	6333593	6332425	5162822	7701684	6309027	7897234	6557101
还原至 0 岁	7595501	6472675	6502673	5330263	7922470	6524783	8137647	6792996

表 6　2007 年 7～10 岁人口教育数据与普查数据之差

	7 岁		8 岁		9 岁		10 岁	
	男	女	男	女	男	女	男	女
2007 年	1228105	1175972	2217370	2272493	1274447	1526968	1362684	1550076
还原至 0 岁	1392423	1332147	2348504	2401045	1376426	1626674	1461306	1644070

2.由于低年龄组漏报的3 000多万,人口老龄化程度被过高估计。以2000年数据为例,直接调查到的60岁以上人口为1.2998亿,65岁以上人口为0.883亿,总人口为12.426亿,老年人口比例分别为10.46%和7.10%。将漏报的低龄人口(0～9岁)加入总人口后,老年人口比例分别降到10.21%和6.94%。同样,低年龄组漏报也会影响少年儿童比例和总抚养比的水平。

3.目前对未来劳动力数量的预测偏低了。当这批低年龄组人口随着年龄的增长而逐步进入劳动力市场时,如果按照现行人口系统的统计数据,就会低估未来10年新进入劳动力年龄的人口数量(3 000万左右)。对未来劳动力数量的判断正确与否会直接影响到人们对就业形势的判断。

4.对未来主要育龄人群(20～30岁)的预测偏低。当这批低年龄组女性人口随着年龄的增长而逐步进入生育期时,如果按照现行人口系统的统计数据,就会低估未来育龄人群数量,进而在预测中低估预测年的出生人数。

5.对男女婴死亡率尤其是女婴死亡率的估计偏高。由于漏报的低年龄人口中,女性多于男性,因此,把2000年人口普查数据中的低年龄人口补齐后,女婴死亡率下降幅度大于男婴死亡率下降幅度,甚至在2岁组出现了逆转,即男婴死亡率由调整前的低于女婴死亡率,变成了比女婴死亡率高(见表7)。

表7 2000年婴儿分性别死亡情况(用教育数据调整)

年龄	年平均人口(人)			死亡人口(人)			调整年平均人口(人)			调整死亡率(‰)			原死亡率(‰)		
	合计	男	女	合计	男	女	合计	男	女	合计	男	女	合计	男	女
0	12644523	6896316	5748208	340085	155564	184521	15369093	8288739	7080355	22.13	18.77	26.06	26.9	22.56	32.1
1	12752979	7017055	5735925	31815	16646	15169	17502529	9365559	8136970	1.82	1.78	1.86	2.49	2.37	2.64
2	14232523	7799459	6433064	22749	12421	10328	17235622	9175885	8059738	1.32	1.35	1.28	1.6	1.59	1.61
3	14839309	8077190	6762119	17338	9584	7754	17944685	9538496	8406189	0.97	1.00	0.92	1.17	1.19	1.15

表8 2000年婴儿死亡率男女比例情况(用教育数据调整)

年龄	男女死亡率之比	
	调整前	调整后
0	0.702804	0.718771
1	0.749433	0.786237
2	0.77698	0.823925
3	0.764131	0.807273

尽管调整后0～1岁女婴死亡率仍然比男婴死亡率要高,但男女死亡率之比升高了,亦即,二者之间的差距缩小了(见表8)。由于男女漏报的不同,使普查数据显现出来的男女婴死亡率差距比真实情况要大。女婴的生存状况依然堪忧,但并没有普查数据显示的那么大。

6.由于同样的原因,在"人口统计系统"的数据基础上计算的出生性别比偏高了。从图9可以看出,2000年0～2岁人口性别比如果按照"人口统计系统"数据计算,已经超过了120,而如果用教育数据进行调整,则都在117以下。尽管都是严重偏高,但严重程度还有区别,这

种区别将直接导致对出生性别比偏高问题的不同判断。

五、小　结

人口系统调查(普查、1%、1‰抽样调查)中的人口低年龄组数据自 2000 年来,任意年份、任意年龄组的同批人口数都随着时间的推移而逐年增长,人口系统数据本身存在矛盾。这种矛盾并非抽样偏差造成的偶然现象,同批人的任意比对都无一例外地呈现随调查年份的增长,证明漏报确实存在。

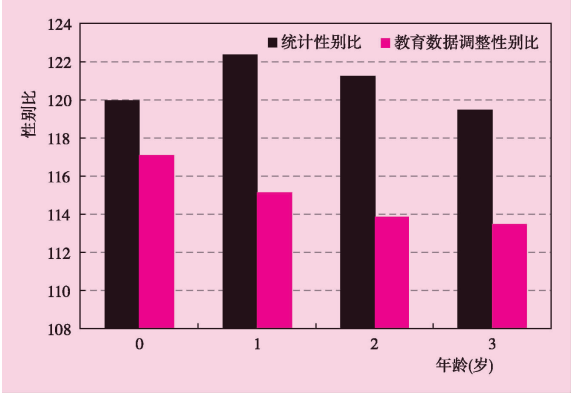

图 9　2000 年婴儿分年龄性别比(用教育数据调整)

利用人口系统以外的相对独立的教育统计数据来对低年龄人口进行推算,教育系统的数据平均每年比人口普查数据多出 300 万人左右。如果考虑死亡等情况,将教育数据和普查数据都用 2000 年生命表还原至 0 岁再进行比对,则差距扩大到平均每个年龄组 340 万人。

低年龄组人口数据与生育水平、老龄化程度、未来劳动力供给状况、未来育龄人群人数、男女婴死亡率及出生性别比等重要人口指标有着密切的联系。由于低年龄组的人口漏报,使得上述指标均偏离了真实情况,从而会妨碍我们对中国的人口形势做出正确的判断,进而妨碍中国人口领域的一系列制度安排的效果。

参考文献:

1. 周皓(2003):《我国第四次人口普查漏报情况的重新估计——基于第五次人口普查的分析》,《人口研究》,第 2 期。
2. 崔红艳、张为民(2002):《对 2000 年人口普查人口总数的初步评价》,《人口研究》,第 4 期。
3. 于学军(2002):《对第五次全国人口普查数据中总量和结构的估计》,《人口研究》,第 3 期。
4. 乔晓春(2002):《从"主要数据公报"看"第五次人口普查"存在的问题》,《中国人口科学》,第 4 期。
5. 王广州(2003):《对第五次人口普查数据重报问题的分析》,《中国人口科学》,第 1 期。
6. 王金营(2003):《2000 年中国第五次人口普查漏报评估及年中人口估计》,《人口研究》,第 5 期。
7. 张为民、崔红艳(2003):《对中国 2000 年人口普查准确性的估计》,《人口研究》,第 4 期。
8. 翟振武、杨凡(2009):《中国出生性别比水平与数据质量研究》,《人口学刊》,第 4 期。

(责任编辑:朱　犁)