对第五次人口普查数据重报问题的分析

王广州

【摘　要】 文章以1982年和1990年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予以分析和研究，以期为“五普”数据调整提供参考。

【关键词】 人口普查　重报　数据质量

【作　者】 王广州　中国人口信息研究中心，副研究员。

众所周知全国第五次人口普查（简称“五普”）存在比较严重的漏报问题，但由于缺乏详尽的漏报人口年龄结构等数据信息，故对漏报人口的分布和判断较困难。目前相对一致的看法是：漏报人口主要在20岁以下，特别是0～9岁儿童漏报可能是主要因素（于学军，2002）。由于漏报严重，对相关人口状态和过程的推断将会存在很大的偏差。人口数据分析、调整和间接估计等任务将异常繁重，尤其是对人口数量、质量、结构、水平、过程和发展趋势等一系列问题的研究和评判将直接影响人口与计划生育政策的制定和实施，影响未来中国人口的发展过程和趋势。同时，由于人口过程具有周期长、不可逆和滞后性，因此，任何决策失误都会对未来中国社会发展造成不堪设想的后果。除了漏报问题以外，“五普”是否也存在着重报问题？“五普”并没有给出可供参考的数据（乔晓春，2002），这就给正确评价“五普”数据带来了一定的困难。为了弥补人口普查数据的缺陷就必须对人口普查数据的问题所在进行全面、详细的分析和评价，然后根据已有的可靠信息对“五普”数据进行重新修正，以期达到对中国人口发展状况、发展过程的全面认识和准确把握，以防错误信息对人口问题判断的误导。本文试图对“五普”存在的重报问题予以论述，试图为今后“五普”数据年龄结构调整、生育水平的判定及死亡等问题的研究提供参考。

一、数据与方法

（一）数据

由于中国由计划经济向市场经济体制的转换，社会、经济、文化构成向多元化方向发展，受户籍制度改革和社会结构转型的影响，人口信息的构成和收集遇到前所未有的困难，这不仅是数据量增加带来的困难，更是由人口结构的复杂性对人口信息产生巨大影响，因此高质量数据获得难度加大。纵观大规模人口调查的历史，中国已进行了5次人口普查和多次大规模的人口抽样调查，人口数据信息日益丰富，尤其是1982年人口普查（简称“三普”）数据质量之高是前所未有的，这对分析和判断以前和以后的人口普查与人口调查起到重要的作用。1990年人口普查（简称“四普”）也是一次较高质量的普查，因此，上述数据对分析、研究“五普”数据的质量，如数据漏报和重报问题具有重要意义。本文所用数据源于《中国1982年人口普查资料（电子计算机汇总）》、《中国1990年人口普查资料》和《中国2000年人口普查资料》，在进行数据质量分析过程中所用数据不包括香港、澳门特别行政区和台湾省。

（二）方法

为了分析“五普”存在的重报问题，本文从年龄结构分析出发，采用的方法是根据已有的高质量人

口数据推算2000年人口状况，将推算结果与2000年实际普查数据进行比较分析，以期达到对“五普”数据的数据质量进行比较全面衡量和认识的目的。在分析方法上采用存活分析法、对普查时点进行调整和数据汇总比较三个步骤。

第一，存活分析方法的应用可以达到分析2000年人口普查中相应人口数据的准确程度的目的。存活分析的基本表达式为：${}_nP_{t_2}(x+n)={}_nP_{t_1}(x)*[{}_nL(x+n)/{}_nL(x)]$。式中$x$的取值范围是0岁到90岁；${}_nP_{t_1}(x)$是在$t_1$时刻年龄在$x$岁至$x+n$岁的人口数；${}_nP_{t_2}(x+n)$是在$t_2$时刻年龄在$x+n$岁至$x+2n$岁的人口数；${}_nL(x)$为确切年龄在$x$至$x+n$队列存活人年数；${}_nL(x+n)$为确切年龄在$x+n$至$x+2n$队列存活人年数。

第二，由于“五普”与“三普”、“四普”的普查时点不同，为了使存活分析更接近“五普”时的人口状态，相应地将存活分析的时点调整到与“五普”相同，即11月1日。各年龄组数据时点人口计算方法是：${}_nP'_{t_{11}}(x)={}_nP_{t_1}(x)+[{}_nP_{t_2}(x+n)-{}_nP_{t_1}(x)]*0.25$。式中${}_nP'_{t_{11}}(x)$是在2000年11月1日年龄在$x$岁至$x+n$岁的人口数；${}_nP_{t_1}(x)$是在2000年7月1日年龄在$x$岁至$x+n$岁的人口数；${}_nP_{t_2}(x+n)$是在2001年7月1日时年龄在$x+n$岁至$x+2n$岁的人口数。

第三，通过总量和分组的方法将推测数据与“五普”数据进行分析和比较。

另外，需要说明的是由于“三普”现役军人年龄结构是按5岁分组，缺少按1岁分组的数据资料，但为了尽可能准确地反映普查数据间的数量关系，对“三普”现役军人采取按“四普”公布现役军人年龄结构进行分解的方式得到1982年1岁组现役军人的年龄结构。由于各年龄组现役军人的相对数量较少，且现役军人年龄结构变化不会很大，因此，现役军人年龄结构按1岁组分解后的年龄别人口数不会对总人口的年龄结构带来很大的分解误差。

（三）方法的有效性分析

对人口普查数据质量的检验事关重大，为了表明推算方法的有效性和研究问题的需要，使用高质量数据进行检验就显得尤为重要，这不仅关系到对推算准确性的评价，而且涉及到对事后质量检验的可靠性评价。由于可获得的高质量数据只有“三普”和“四普”的数据，因此，本文采用“三普”和“四普”数据进行分析方法的有效性检验。在用“三普”数据对“四普”8～90岁人口进行存活分析时，考虑到“四普”死亡漏报问题，将男、女平均预期寿命分别设定为68.3岁和71.3岁，对1982与1990年间各年度的男性和女性预期寿命采取线性差值的方法得到。根据上述假定的推测结果见表1。从表1可以看到，根据“三普”数据推测1990年8～90岁男性人口为49 355.6万，女性人口为46 641.3万，男女合计为95 996.9万；相应的，1990年实际普查数男性为49 294.9万，女性为46 481.0万，男女合计为95 775.9万。推测值与实际调查的差值为221万人，即推测数据比实际普查数多了221万人。根据1990年质量抽查漏报率为0.7%，重报率为0.1%，0～100岁及以上的人数净误差为0.6%，即推算可能漏报791万人，净误差为678万。因此，从推算结果可以断定1990人口普查漏报的发生，并与事后抽样结果的结论是一致的。此外，根据推测可以看到18～60岁数据质量相对较高，净误差为196.0万人，而60岁以上的数据质量相对较差。由于“三普”和“四普”调查时点相同，不涉及时点调整问题，因此，从推算符合精度上来看相对较高。总之，数据检验表明上述分析方法可以比较确切地反映问题所在。

表1 1990年人口普查数据与1982年推测数据比较 万人

	1990年普查数		据1982年普查推算数	
	男	女	男	女
8～90岁	49 294.9	46 481.0	49 355.6	46 641.3
合计	95 775.9		95 996.9	
其中				
18～90岁	38 620.7	36 453.0	38 873.4	36 799.6
合计	75 073.7		75 673.0	
18～60岁	34 415.5	31 762.0	34 470.0	31 903.5
合计	66 177.5		66 373.5	
61～90岁	4 205.2	4 691.0	4 403.4	4 896.1
合计	8 896.2		9 299.5	

二、重报问题分析

（一）2000年人口普查存在比较严重的重报问题

虽然从时间的有效性角度看，“四普”数据好于“三普”。但由于“四普”也存在一定的漏报问题，漏报率为0.7%，相比之下“三普”的漏报率仅0.56‰，因此从数据的有效性和完整性上看，1982年人口普查是一个难得的准确数据标准。此外，由于死亡水平变化相对比较缓慢且方向确定，而且对总人口的变化影响相对比较确切，因此，通过恰当的假定对现存人口的预测误差不会很大。正是出于上述考虑，为了比较全面验证和推断“五普”数据存在的主要问题，本文将采取以“三普”为主、“四普”为辅，结合部分“五普”数据来分析“五普”的重报问题。

根据“五普”数据计算男性和女性人口的预期寿命分别为70.61和74.45岁，考虑到死亡漏报和男性和女性预期寿命的差距问题，将2000年男性和女性预期寿命估计为70和73岁，对1982或1990年与2000年间各年度的男性和女性预期寿命同样采取线性差值的方法得到。基于上述数据、算法和基本思路，得到2000年11月1日中国18～90岁人口推测值（见表2）。

从表2可以看出，与1982年的推算结果相比，2000年人口普查得到18～90岁的人口数量多了1 238.4万，其中，18～60岁多了1 483.3万，而61～90岁则少了244.9万。如果不考虑时点转换问题，直接用1982年数据推算到2000年7月1日，18～90岁人口推算结果是男性45 525.9万人，女性为43 353.7万人，合计88 879.6万人，与2000年人口普查的误差为1 027万人。然而，事后人口普查数据质量抽样结果认为，2000年人口普查漏报率为1.8%，即漏报2 246万人，比较“四普”漏报情况，这里推算的结果与实际调查结果存在数据矛盾现象。如果根据1990年人口普查数据推算，与推算结果比较2000年人口普查10～90岁人口多报了2 673万人，其中18～90岁人口多报1 461万人。如上所述，由于根据1990年人口普查质量抽查的漏报率0.7%推算，1990年可能漏报791万人，那么，如果假定1990年所有漏报人口都发生在0～80岁之间（80岁以上人口比重相对较小，忽略不计），可以推断“五普”10～90岁人口至少多报1 800万。鉴于上述推测可以确信“五普”存在比较严重的重报问题。如果不考虑10～90岁人口的重报与漏报的相抵问题，年龄结构数据的差错率可能会相当大。

表2　2000年人口普查数据与人口推测值比较　　万人

	2000年普查数		据1982年普查推算		据1990年普查推算	
	男	女	男	女	男	女
10～90岁	55 649.0	52 880.5	—	—	54 384.0	51 472.5
合计	108 529.5		—		105 856.5	
其中						
18～90岁	45 978.0	43 928.7	45 412.2	43 256.1	45 281.6	43 164.2
合计	89 906.7		88 668.3		88 445.8	
18～60岁	40 136.5	37 755.1	39 466.0	36 942.3	39 546:1	37 019.8
合计	77 891.6		76 408.3		76 565.9	
61～90岁	5 841.5	6 173.6	5 946.2	6 313.9	5 735.5	6 144.4
合计	12 015.1		12 260.1		11 879.9	

（二）2000年人口普查重报人口年龄分布特征

经过使用1982年人口数据和1990年人口数据的推算值与2000年人口普查数据可以看出人口重报问题的存在。由于1982年人口普查数据资料质量相对较高，可以此作为数据推测与判断的依据，通过比较推测值和实际普查值看出“五普”误差的分布情况（见图）。从图可以看出，重报问题主要发生在45岁以下，尤其是30～40岁和20岁以下（见表3）。

从表3可以看出误差最大的年龄组依次为38、30、31、32岁，尤其是37、38、39和40岁连续出现负值，即推算低估现象，而1982年对1990年进行推测误差却没有发生连续为负值或正值的情况，因此可以断定2000年普查重报现象的发生，绝非运算方法带来的误差。

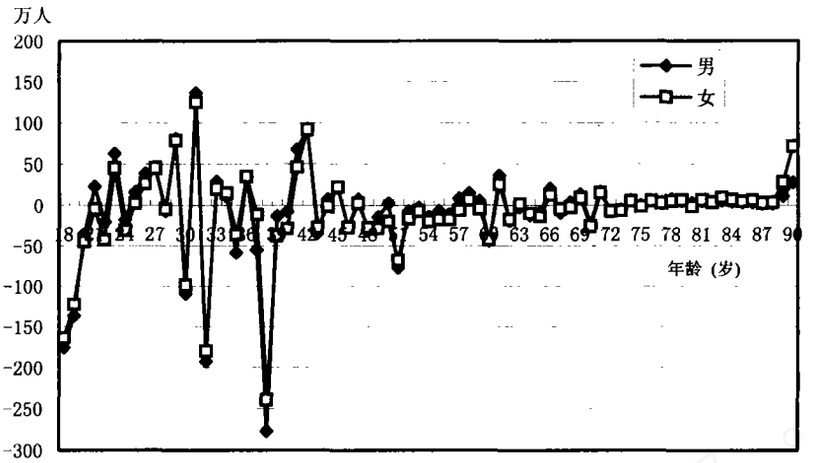

图 2000年人口普查数据与1982年人口普查数据推算值误差分布

注：根据1982年人口普查数据推算。误差＝1982年预测值－2000年普查值。

表3 2000年人口推测值与人口普查数据误差比较 万人

年龄	2000年人口普查数		2000年人口推测值		估计误差	
	男	女	男	女	男	女
30	1 444.0	1 362.0	1 334.7	1 263.7	－109.3	－98.3
31	1 286.0	1 220.5	1 422.6	1 345.5	136.6	125
32	1 429.3	1 345.9	1 237.0	1 166.2	－192.3	－179.7
33	1 116.3	1 060.6	1 145.2	1 081.2	28.9	20.6
34	1 278.3	1 208.1	1 288.8	1 222.8	10.5	14.7
35	1 282.9	1 200.7	1 224.4	1 164.5	－58.5	－36.2
36	1 241.1	1 177.4	1 275.6	1 212.4	34.5	35
37	1 439.8	1 350.3	1 385.1	1 338.7	－54.7	－11.6
38	1 081.5	1 012.6	804.6	773.4	276.9	－239.2
39	581.8	561.0	568.5	523.6	－13.3	－37.4
40	762.0	707.7	754.6	678.7	－7.4	－29
小计	12 943.0	12 206.8	12 441.1	11 770.7	－501.9	－436.1
合计	25 149.8		24 211.8		－938	

注：同图。

（三）对2000年人口普查总人口数量的再认识

如果“五普”10～90岁人口重报高达1 800万以上的话，那么，可以确信总人口的重报会更大一些，因此，我们不得不对总人口数量和人口漏报问题重新审视。假定人口普查事后抽样漏报率准确，且漏报全部发生在10～90岁年龄组，那么，10～90岁年龄组人口重报将高达4 000万，总人口将在12.26亿左右。如果假定漏报全部发生在0～9岁年龄组，人口重报也将高达1 800多万，那么，总人口可能将在12.48亿左右。由此可见，如果普查事后抽样漏报率准确，那么，2000年人口普查中人口重报数量可能在1 800万～4 000万之间，总人口数量则应在12.26亿～12.48亿之间。如果普查事后抽样漏报率存在偏差，那么，“五普”的总人口数量将无从判断。

最后需要说明的是，虽然上述分析可以断定人口重报问题的发生和问题的严重性，但由于对死亡水平假定和年龄结构时点调整可能影响推断的精度。无论精度如何，如果上述重报问题存在且比较严重，那么，不仅影响到对2000年总人口的推断，而且几乎需要对所有数据进行重新调整和修正，尤其是对0～9岁年龄组数据的修正直接关系到对中国人口生育水平的判断和未来人口政策的制定。

参考文献：

1. 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。
2. 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。
3. 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。
4. 国务院人口普查办公室、国家统计局人口统计司编(1985)：《中国1982年人口普查资料(电子计算机汇总)》，中国统计出版社。
5. 国务院人口普查办公室、国家统计局人口统计司编(1993)：《中国1990年人口普查资料》，中国统计出版社。
6. 国务院人口普查办公室、国家统计局人口和社会科技统计司编(2002)：《中国2000年人口普查资料》，中国统计出版社。

（责任编辑： 朱 犁）