

人口年龄结构数据异常的检验

黄荣清, 肖周燕

(首都经济贸易大学 人口经济研究所, 北京 100026)

摘 要: 本文提出可以利用人口年龄结构的差分符号和各年龄人口与相邻两侧人口均值的差的变化来检验人口年龄数据的申报质量, 并以此来判断特定年龄上人口的“堆积”与“流失”。

关键词: 人口年龄数据; 异常; 年龄尾数

中图分类号: C92 **文献标识码:** A **文章编号:** 1000-4149 (2009) 02-0001-08

Test on Data Exception of Age-specific Population Structure

HUANG Rong-qing, XIAO Zhou-yan

(Institute of population and Economics, Capital University of Economics and Business, Beijing 10026, Chian)

Abstract: The paper is proposed that the quality of age-specific population data declaration can be tested using difference sign of age-specific population structure and the variation of the difference between the data of a certain age group and the average of the two age groups flanking it, based on which the population of “age-heaping” and “age-eavading” in a certain age group can be estimated.

Keywords: age-specific population data; exception; age mantissa

人口年龄结构数据作为人口研究最基础性的数据, 不仅反映了当前的人口状态和过去的人口变动, 而且是分析未来人口发展趋势的前提和基础。只有建立在可靠的人口数据基础上的研究才能有助于正确把握和理解现实的人口问题, 因此, 人口学者一直非常重视人口年龄结构数据的质量。就目前来说, 检验人口年龄结构数据的准确性方法主要有三种, 分别是玛叶指数、韦伯指数和联合国综合指数, 另外, 还有如年龄偏好指数, 年龄准确性指数等。中国学者在应用这些方法时, 从分析中国人口的实际变动情况中, 发现了这些方法的局限性^[1]。但至今还没有提出一个更好的、简单而有效的, 可以替代它们的检验方法。本文提出, 可以利用人口年龄结构的差分符号和各年龄人口与相邻两侧人口均值差的变化来把握过去的人口变动倾向和用它们的尾数符号分布来检验人口年龄数据的申报质量, 并以此来判断特定年龄上人口的“堆积”与“流失”。

一、年龄别人口数据变动特征

在一般情况下, 一个区域的人口数随着年龄而变化, 应该是有一定规律性的。我们假设区域

收稿日期: 2008-11-23

作者简介: 黄荣清 (1946-), 江苏无锡人, 首都经济贸易大学人口经济研究所教授, 博士生导师, 研究方向为人口与劳动统计、中国少数民族人口、城市化和城市人口等。

人口是封闭的 (即无迁移发生), 先来观察理论上的稳定人口的年龄结构。由人口学而知, 对于一个稳定人口来说, 它的年龄结构永远保持不变, 为

$$P(x) = b \cdot \exp(-rx) l(x) \quad (1)$$

这里, b 表示稳定人口的出生率, r 表示稳定人口的增长率, $l(x)$ 表示在固定的死亡水平下, 对应的生命表中从出生到 x 岁年龄的存活率。则 $P(x)$ 的导数有

$$P'(x) = b \cdot \exp(-rx) \cdot [-r \cdot l(x) + l'(x)]$$

如果人口是稳定增长型或静止型, $r \geq 0$, 由于 $l'(x) < 0$, 所以

$$P'(x) < 0$$

如果人口是稳定减少型, $r < 0$, 设在 x_0 年龄上, $l'(x_0) / l(x_0) = r$, 则当

$$x < x_0 \text{ 时, } P'(x) > 0$$

当

$$x > x_0 \text{ 时, } P'(x) < 0$$

根据上面的推导, 如果人口是稳定增长型或静止型, 在每一个年龄上, 差分 $\Delta_F P_X - P_{X-2} < 0$, 如果人口是稳定减少型, 则差分符号分为两部分, 在年龄开始阶段, $\Delta_F > 0$, 过了某一年龄后, $\Delta_X < 0$ 。

在实际中, 由于出生、死亡率水平的变动, 纯粹的稳定人口是不存在的。但当出生、死亡率水平的变动幅度不大时, 稳定人口所呈现出的年龄结构特征是存在的。以下, 我们来讨论当出生、死亡发生变动时人口的年龄结构特征。设 t 年 x 岁的人口数为 $P_X(t)$, $x+1$ 岁的人口数为 $P_{X+1}(t)$, 有

$$P_X(t) = B(t-x) \times p_0(t-x) \times p_1(t-x+1) \times \dots \times p_{X-1}(t-1) = \prod_{j=0}^{X-1} p_j(t+j-x)$$

这里, $B(t-x)$ 表示 $t-x$ 年的新生人口, $p_X(t)$ 表示 t 年 x 岁的存活率。则 $P_{X+1}(t)$ 可写成

$$P_{X+1}(t) = \prod_{j=0}^X p_j(t+j-x-1)$$

所以, $P_{X+1}(t)$ 和 $P_X(t)$ 之比为

$$P_{X+1}(t)/P_X(t) = [B(t-x-1)/B(t-x)] \cdot p_X(t-1) \cdot \prod_{j=0}^{X-1} p_j(t+j-x-1)/p_j(t+j-x) \quad (2)$$

现在, 我们来观察 (2) 式右边各项。首先, $p_X(t-1) < 1$, 对于连乘号内的因子, $p_j(t+j-x-1)/p_j(t+j-x)$ 表示上一年度 $(t+j-x-1)$ j 岁的存活率和这一年度 $(t+j-x-1)$ j 岁的存活率之比, 如果死亡率随年代变动是下降的, 则在同一年龄的存活率应该是增大的, 所以也有 $p_j(t+j-x-1)/p_j(t+j-x) < 1$, 自然也有 $\prod_{j=0}^{X-1} p_j(t+j-x-1)/p_j(t+j-x) < 1$, 这样, 当 $B(t-x) > B(t-x-1)$, 即和上年比较, 每年的新生人口在不减少的情况下, 总有 $P_{X+1}(t)/P_X(t) < 1$, $P_{X+1}(t) < P_X(t)$, 或者说 $\Delta_F P_X - P_{X-1} < 0$, 差分符号为负。

进一步可以推得, 若 t 年年龄区间 $(x, x+a)$ 年龄段内, 正的差分值符号共有 $n(t)$ 个, 则若干年后, 在 $t+\Delta$ ($\Delta > 0$) 年, 在年龄段内 $(x+\Delta, x+a+\Delta)$ 年龄段内, 差分为正号的数量只会减少, 不会增加, 即 $n(t+\Delta) \leq n(t)$ 。

实际得到的人口年龄别数据, 在一些年龄可能出现 $\Delta_X > 0$, 在另一些年龄则可能出现 $\Delta_X < 0$ 的情况。我们把出现 $\Delta_X > 0$ 称为“异常”值, 出现异常情况的年龄称为异常年龄。由于人口实际变动的复杂性, 这里的所谓“异常”, 只是指和稳定增长型或静止型的年龄人口数据呈现出的不同变动特征而言, 并不一定是数据有问题, 不正确。例如, 对于缩减型人口来说, 尤其是在低年龄, 出现 $\Delta_X > 0$ 是自然而然的。对总体上呈增长趋势的人口来说, 在个别年份, 也会出现新生人口减少导致在若干年内一直保持 $\Delta_X > 0$ 。

下面我们分别考察全国和各省 (区) 五次普查中年龄人口数出现的异常情况。从 1953~2000

年五次人口普查数据来看，出现异常次数有增长的态势。1953 年出现异常次数共 22 次，1964 年出现异常次数有所减少，为 18 次，到了 1982 年、1990 年出现异常次数增加，均为 30 次，2000 年出现异常次数达到 37 次。从表 1 还可看到，各省区在 2000 年出现异常总数均大于 30 次，其中，新疆、吉林、四川、上海年龄人口数出现异常次数超过 40 次。在 1990 年人口普查数据中，出现异常次数最多的是新疆、天津和河南，出现次数在 35 次以上，江西、湖南、广东等省区相对来说较少，出现异常数不超过 30 次。1982 年第三次人口普查数据中，除新疆超过 40 次外，其他省区出现异常次数都在 35 次以下。在 1964 年人口普查中，各省区出现异常次数明显低于其他各次人口普查，除新疆、河南和陕西等于或超过 30 次外，其他省区异常出现次数都在 30 次以下。在 1953 年人口普查中，出现异常次数较少的是吉林、山东、广东和福建，它们出现异常的次数在 20 次以下，新疆仍然是出现异常次数最多的区域（见表 1）。

表 1 五次普查时各地区人口在各年龄尾数上出现的“异常”次数

地区	2000 年			1990 年			1982 年			1964 年			1953 年		
	合计	最大	最小	合计	最大	最小	合计	最大	最小	合计	最大	最小	合计	最大	最小
北京	36	5	2	30	4	0	32	5	2	26	4	1	22	4	1
天津	39	5	2	36	5	2	33	4	2	—	—	—	29	5	1
河北	38	6	1	30	5	0	32	5	1	29	5	1	29	5	1
山西	35	6	1	32	5	1	31	4	2	27	5	1	26	4	1
内蒙	37	5	0	29	5	2	29	5	2	27	6	1	30	5	1
辽宁	36	5	2	32	5	1	29	4	0	27	5	0	22	4	1
吉林	40	7	2	34	5	0	31	5	0	21	3	0	20	4	0
黑龙江	37	6	2	27	5	1	26	4	0	23	5	0	23	4	0
上海	42	6	3	33	6	1	28	6	0	22	3	1	23	4	0
江苏	39	6	2	29	6	1	35	7	1	26	5	1	25	4	1
浙江	35	5	3	33	6	1	28	5	1	24	6	1	25	5	1
安徽	39	8	0	30	7	1	34	5	2	29	5	1	26	4	2
福建	33	5	1	29	6	0	28	5	1	20	4	0	17	3	1
江西	32	6	2	23	5	1	24	4	1	17	3	0	23	3	0
山东	32	5	2	30	5	1	32	5	1	27	5	1	20	3	1
河南	39	9	1	35	7	2	35	5	2	32	6	2	25	5	0
湖北	39	5	2	30	6	1	35	6	2	28	5	1	25	6	0
湖南	34	6	1	28	5	1	33	5	2	25	5	0	33	7	0
广东	33	7	0	29	5	1	26	4	0	20	4	0	20	6	1
广西	37	9	0	30	5	1	26	5	1	19	4	0	28	8	0
海南	37	10	0	35	9	1	—	—	—	—	—	—	—	—	—
重庆	36	6	1	—	—	—	—	—	—	—	—	—	29*	5*	1**
四川	40	7	1	29	4	1	29	4	1	26	5	1	23	4	0
贵州	36	8	0	28	6	1	29	5	1	26	4	1	27	6	0
云南	34	7	2	29	5	1	28	4	1	29	5	0	31	9	0
西藏	37	10	0	25	6	0	28	6	0	—	—	—	—	—	—
陕西	36	6	1	30	5	2	31	4	2	30	6	2	32	4	2
甘肃	32	7	0	29	6	2	30	5	1	24	4	1	23	4	1
青海	36	8	0	29	6	1	29	4	1	27	6	1	29	4	2
宁夏	37	10	0	31	7	1	29	5	1	26	5	0	27*	5*	0*
新疆	41	10	0	41	10	0	43	9	0	48	9	0	40	10	0

数据来源：根据 1953~ 2000 年全国五次人口普查数据得到的计算结果。

注：①这里的“最大”和“最小”是指在各年龄尾数上出现的异常次数的最大值和最小值。

②1953 年数据中，* 为热河省数据；** 为昌都地区数据。

显然，年龄别人口数据异常出现的次数多少和过去人口复杂的变动有着密切关系。异常情况出现少，说明人口随着年龄变化稳定，符合稳定增长型或静止型年龄人口数据呈现出的变动特

征; 异常次数出现多, 说明年龄人口数值高低起伏, 变化不稳定。总的说来, 第一、二次人口普查全国的数据异常出现的次数少于第三、四、五次人口普查异常出现的次数, 各省区也大致是这种情况。在历次人口普查中, 新疆出现的异常次数总是最高。

下面, 我们来讨论根据人口年龄数据的异常, 判别人口数据年龄报告的质量。

二、年龄报告准确性和年龄尾数指向的检验

1. 年龄报告的准确性

实际人口的年龄结构, 由于各年的出生情况不一样, 也由于受迁移的影响, 在一些年龄上出现了差分符号为正, 或者说在一些年龄上出现“异常”完全是可能的。按照前面的分析, 异常的出现, 发生在以下几种情况^①:

(1) 在一定时期内, 出生人口逐年减少; (2) 在一定时期内, 虽然总趋势是出生人口不断增加, 但在个别年份也出现了出生人口减少; (3) 由于迁移的作用, 一些年龄段 (例如青壮年) 人口增加。

上面的几种情况, (1) 和 (3) 的情况是影响某个年龄段, 而不是单个年龄; (2) 影响 1~ 2 个特定年龄。但不管是哪种情况, 只要它不是周期性的, 即不会发生在特定的年龄尾数上。所以, 我们总是可以假定, 年龄人口数据的异常值的出现, 在各年龄尾数上, 应该是随机地平均分配的。从全国五次各地区人口普查数据来看, 这些异常值的分布较为平均, 除了 1990 年和 2000 年人口普查在“0”年龄尾数上较为突出外, 在其他年龄尾数为平均分布。但是, 通过比较全国各地区五次人口普查异常值, 我们发现, 某些区域的年龄人口异常值出现并不是平均分配的, 而是存在一定指向。例如, 2000 年新疆普查数据, 异常出现次数最多的是以“0”为年龄尾数, 共出现了 10 处, 其次是以“1”、“4”和“7”为年龄尾数, 分别出现了 9, 8, 7 处, 其他年龄尾数较少出现或不出现。除新疆外, 其他省区在不同年份也普遍存在这种年龄尾数指向问题。下面, 我们来讨论异常出现特定年龄尾数指向的检验问题。

设某个年龄别人口数据在各年龄尾数“ i ”上出现异常次数分别为 $n_i (i = 1, 2, 3 \dots 0)$, 则全部异常出现的次数 $N = \sum N_i$ 。假设异常的发生对各年龄尾数没有指向, 是平均分配的, 这也就意味着异常发生的概率在每个年龄尾数上应该相同。由于出现总的异常次数为 N , 出现的频率为 $N/100$, 则在各个年龄尾数出现理论频数为 $N/10$ 。

对年龄尾数指向的检验, 类似概率论中的古典问题: 设共有 10 个容器, 其编号分别为 0, 1, 2...9, 现在, 我们向这 10 个容器投球, 共投了 S 次, 球投入容器内的次数为 N , 第 i 个容器内的次数为 $N_i (i = 0, 1, 2 \dots 9)$, $N = \sum N_i$, 现在, 我们要问, 各个容器被投入的可能性是否相同?

卡方检验是对样本的频数分布所来自的总体分布是否服从某种理论分布或某种假设分布所作的假设检验, 因此, 此处采用卡方检验方法来检验其假设, 检验年龄尾数的实际分布与假设是否一致。可以计算卡方值:

$$\chi^2(9) = \sum \frac{(N_i - N/10)^2}{N/10}$$

通过卡方值检验, 我们可以肯定或者否定上述的假定是否成立。如果肯定, 则认为假定成立, 即异常值出现是随机的, 人口调查中无特定年龄尾数指向; 如果假定不成立, 则认为这是由于报告的特定年龄尾数指向造成的。进一步, 由于总体出现异常的概率为 $N/100$, 各个年龄尾数出现异常的频率为 $N_i/10$, 这样, 可以通过检验各个年龄尾数出现的频率 $N_i/10$ 与 $N/100$ 有否显著

^① 由于死亡率变动通常呈稳定下降, 所以我们这里未对人口死亡因素的影响加以讨论, 但即使加上死亡变动的影响, 只要这种变动不是以 10 岁为周期, 也不会改变本文的结论。

性差别，从而判定尾数的分布是否均匀，在特定年龄是否存在年龄尾数指向^[2]。

下面，首先来判断全国各省区五次人口普查年龄别人口数据异常发生是否存在年龄尾数指向问题（见表2）。假设检验是基于“小概率原理”构建的一种检验方式，通过事先给出的概率，即显著性水平（ α ）来检验，为此，我们通过给定显著性水平的卡方值

表2 五次人口普查年龄尾数异常分布的卡方检验值					
年	2000	1990	1982	1964	1953
全国（N）	8.67	5.33	2.66	5.33	4.36
地区平均值	10.15	7.82	5.75	7.20	8.95
标准差	6.23	3.75	3.28	1.97	5.70

资料来源：根据全国各地五次人口普查数据计算得出。

来判断地区年龄人口数据是否存在特定年龄指向问题。不仅如此，通过年龄尾数的检验，我们还能更进一步了解全国各地区五次人口普查数据质量。在前面分析中，已经指出年龄别人口数据的质量和异常出现的次数多少有密切关系，异常出现次数多，说明年龄别人口数据有待进一步研究。但需要明确的是，异常出现的次数只能在某种程度上说明数据质量，还需进一步检验，因为异常出现的次数还可能与人口自身的变动有关，通过年龄尾数的检验，如果发现某一年份年龄尾数上分布不均匀，确实存在某些年龄尾数的指向问题，才说明异常出现的次数并不是人口变动造成的，才认定数据质量存在问题。

由于假设检验是基于“小概率原理”构建的一种检验方式， α 越小，统计量值落于拒绝域内的概率就越小，表示原假设与样本值之间的矛盾越显著，在多数统计书中显著性水平（ α ）通常取0.01 或 0.05，但这并不意味着 α 只能取 0.01 或 0.05。 α 取值的大小，说明的是拒绝域内的概率， α 取 0.1 或 0.2 也并不是不可以，只是说明犯第二类错误可能大一些。为此，可以用小概率来否定，通过选择 α 不同取值，将 α 临界值与卡方值进行比较，不但判断人口年龄数据是否存在年龄尾数指向问题，进而判断年龄人口数据质量。

通过对卡方值和 X^2 不同显著性水平的临界值进行比较，利用小概率事件来否定，我们可以做出如下判断^①：

（1）当卡方值大于 16.919，说明原均匀分布的假设发生的可能性不到 5%，应该拒绝假设。或者反过来说，至少有 95% 的把握认为某一区域在某一时点年龄人口数据存在严重的年龄尾数指向问题，年龄人口数据质量差。

（2）当 $12.242 < X^2 < 16.919$ ，说明有 80% 的概率可以认为某一区域在某一时点年龄人口数据存在年龄尾数指向问题，年龄人口数据质量较差。

（3）当 $9.414 < X^2 < 12.242$ ，表明仅有 60% 不到 80% 的概率可以认为某一区域在某一时点年龄人口数据可能存在年龄尾数指向问题，但还不算严重，年龄人口数据质量属可以接受。

（4）当 $X^2 < 9.414$ ，说明不到 60% 的概率认为某一区域在某一时点年龄人口数据存在年龄尾数指向问题，异常次数的产生更大的可能是随机误差造成的，表明年龄人口数据质量好。

表2 的数据表明，从全国五次人口数据来看，尽管异常出现的总次数在五次普查中不断增加，但经过卡方检验，就全国而言，五次人口普查数据质量较好。不同时期各地区数据质量存在差异。从表2 五次普查 X^2 的值来看，第二、第三、第四次人口普查地区的平均值和标准差都较小，说明这三次普查的年龄申报质量要高于第一次和第五次人口普查。大部分省区五次普查人口年龄数据质量都是好的或可以接受的。1953 年有 5 个省区数据质量较差，2000 年甘肃和贵州的年龄人口数据存在一定问题，还有 6 个省区数据质量较差。结合判断标准，五次人口普查数据质量差或有问题的省区见表3。

① 说明：不同显著性水平下的临界值： $X^2_{0.001}(9) = 27.877$ ； $X^2_{0.05}(9) = 16.919$ ； $X^2_{0.10}(9) = 14.684$ ； $X^2_{0.20}(9) = 12.242$ ； $X^2_{0.40}(9) = 9.414$

在各次普查中，新疆五次人口普查数据年龄尾数异常卡方值都远远大于 16.919，因此，我们有 99% 以上的把握认为新疆的人口普查数据存在严重的年龄尾数指向问题。我们注意到，在各次人口普查中，数据质量差的有新疆和西藏。在第五次人口普查时，又增加了一些地区（主要是西部的少数民族地区）。在以往的研究中，一般都认为全国各地区中，只有新疆的人口数据质量较差，其他地区的人口数据质量都属于好的或比较好的。但黄荣清分析 1990 年少数民族人口数据质量时^[3]，指出许多少数民族存在着在年龄尾数“0”的堆积，印证了民族地区人口数据存在着质量问题。

2. 特定年龄尾数指向

上面，我们利用了人口年龄数据的一阶差分符号的分布来判断人们在年龄申报上是否存在着特定年龄尾数的指向，估计了人口年龄报告的准确性。一般来说，对特定的年龄指向，总对应着特定的年龄尾数“回避”。但利用一阶差分，有时我们还难于判别在特定年龄尾数上出现的异常，到底是“堆积”还是“回避”？因为对 $\Delta F = P_x - P_{x-1} > 0$ ，如果它的报告不准确，既可能是由于在 x 年龄上的堆积，也可能是由于在 $x-1$ 年龄回避上造成的，且在 x 年龄上的堆积数，不一定就是从 $x-1$ 年龄上得到，而可能是从其他年龄上得到，同样，在 $x-1$ 年龄上的回避数，不一定就堆积到 x 年龄上，也可能堆积到其他年龄，所以，对年龄申报的准确性问题，我们还可以进一步深入探讨。设

$$D_x = P_x - (P_{x-1} + P_{x+1})/2$$

D_x 表示某一年龄的人数和这个年龄左右两边算术平均数的差^①。我们通过观察 D_x 的年龄尾数符号的分布来判断在该年龄尾数是否有“异常”现象。

对于理论上的稳定人口， D_x 的符号在各个年龄上的尾数分布，应该是很有规律的。例如，对静止人口（静止人口也可以看成是稳定人口）来说，在儿童期， D_x 的符号为负，在青壮年期， D_x 的符号大部分应该为正，而当稳定人口增长率达到一定程度，则 D_x 的符号又大部分会转变为负。对于实际人口，在有些场合，可能会出现正号数量多于负号数量，在另外的场合，则可能相反。但不管是什么情况，我们总是假定人口申报尾数的堆积，只发生在尾数的符号为正的较多，人口申报尾数的回避，只发生在尾数的符号为负的较多。

设在全部 S 个年龄上，（ S 应该是 10 的倍数，若研究对象能确定人口数的年龄共 $S+1$ 个，例如，0~ 90 岁，则 $S= 90$ ，0~ 100 岁，则 $S= 100$ ）， D_x 符号为正共出现了 N 次（则符号为负的次数出现 $S- N$ 次），在各个年龄尾数上符号为正出现的次数为 N_i （ $i= 0, 1 \dots 9$ ）次，如果人口的年龄变动不是周期变动，则我们总可以假定在各个年龄尾数上出现正号的概率是相等的。在检验时，若我们否定了这个假定，则我们可以认为出现正号的概率在各年龄上是不同的，并判定在那些出现正号概率大的年龄尾数上人们申报有“偏好”，人口年龄有“堆积”，而对那些出现正号概率小的年龄，人们申报年龄有“回避”倾向。

对特定年龄尾数指向检验，可以这样检验：由于全年龄出现正号（或负号）的次数为 N 次，

表 3 全国各地区五次人口普查数据质量情况

地区	2000 年	1990 年	1982 年	1964 年	1953 年
湖南	—	—	—	—	差
广东	—	—	—	—	差
广西	差	—	—	—	差
海南	差	差	*	*	*
贵州	较差	—	—	—	—
云南	—	—	—	—	差
西藏	差	差	差	*	*
青海	差	—	—	—	—
甘肃	较差	—	—	—	—
宁夏	差	—	—	—	*
新疆	差	差	差	差	差

注：* 表示无数据。

① D_x 也可写成差分形式： $D_x = (\Delta x - \Delta x) / 2$

则总体出现的频率为 N/S (本文为 $N/100$), i 年龄尾数出现正号 (或负号) 的次数为 N_i 次, 则它出现的频率为 $N_i/(S/10) = 10 N_i/S$, 设频率有正态分布, 在水平 α 下, 检验假设

$$H_0: 10 N_i/S = N/S; H_1: 10 N_i/S > N/S$$

$$Pr(10 N_i/S - N/S > K | 10 N_i/S = N/S) = \alpha$$

若 $10 N_i/S - N/S > K$, 则拒绝假设, 说明 i 年龄尾数出现正号 (或负号) 的频率与全部年龄出现正号 (或负号) 的频率有显著性差别, 在这年龄上人口有堆积 (或流失)。

经过检验, 质量差的地区 D_x 都通过了显著性水平为 0.05 的检验, 说明 D_x 出现正号的概率在各年龄上是不同的, 意味着这些区域存在年龄指向问题。经过比较发现, 质量差的地区年龄喜好主要集中在“0”和“8”尾数上, 此外, 新疆、西藏等地区在“2”和“5”上也存在喜好, 说明这些区域在“0”、“8”及“2”或“5”年龄尾数上存在申报“偏好”, 表明在这些年龄尾数上的人口有“堆积”。除了年龄偏好外, 质量差的地区在年龄尾数为“1”和“9”存在明显的“回避”现象, 部分地区如新疆、云南等在年龄尾数为“3”或“7”也存在明显“回避”, 说明这些年龄上人口有流失。

3. 特定的年龄尾数指向发生的原因

前面指出, 年龄尾数指向主要集中在“0”、“8”、“1”和“9”上, 其中, “0”和“8”存在年龄偏好, “1”和“9”存在年龄回避。正是由于对年龄尾数“0”和“8”存在偏好, 所以才出现“1”和“9”年龄指向回避问题, 它们之间是相互依存的。尾数“0”偏好有两种情况, 一种是夸大年龄, 即在年龄尾数为“8”或“9”的情况下, 或是记忆模糊或是其他原因, 直接记入尾数为“0”的年龄段, 如 78、79 岁或 88、89、99 岁直接登记为 80、90 或 100 岁, 把不应为“0”为尾数的年龄直接记入尾数为“0”的年龄段, 这种情况容易造成“0”尾数上的年龄堆积。由于将接近“0”的尾数直接记入“0”尾数年龄段, 造成以“0”为尾数的人口数过多, 就会出现 $\Delta > 0$ 的异常情况; 另一种情况是由于对“1”年龄指向的回避, 而造成“0”尾数上的堆积, 这是经常出现的现象, 由于在某些特殊事件如退休, 或是其他原因, 往往造成尾数为“1”年龄指向的回避问题, 如男女职工分别在 60 岁和 50 岁退休, 为了继续工作存在已经 51、61, 但仍然申报为 50 或 60 岁的情况。由于存在“1”年龄尾数上的回避, $P(1)$ 过小的情况就会出现, 而在 $P(2)$ 正常的情况下, 就会出现 $\Delta > 0$ 的异常情况。

前面数据分析指出, 除了以“1”和“9”尾数存在年龄回避问题外, 在某些地区还存在以“7”尾数年龄回避以及“5”尾数年龄堆积的问题。以尾数“7”为年龄指向, 通常可以解释为是对尾数“8”的年龄偏好造成。尾数“5”年龄堆积则主要是由于人们记忆上的原因, 通常对“5”存在偏好, 方便记忆形成的。依照上面的分析, 通常以尾数“1”和“9”为年龄指向是由于在“0”和“8”尾数的申报堆积有关, 对尾数“7”年龄指向主要也是由于对尾数“8”的偏好造成的。以尾数“5”为年龄指向主要是由于记忆上的偏好。正是由于对某些年龄的偏好或回避, 所以形成部分地区年龄尾数上的指向问题。对于前面列举的地区, 我们可以有 95% 以上的把握认为存在着年龄指向问题。如果再放宽一些标准, 例如在 80% 的置信水平上, 则可以有更多的地区, 特别在“5”年龄尾数上有申报堆积以及“7”年龄尾数上存在回避。可见, 正是由于对“0”“8”“5”上的偏好以及在“1”“9”“7”上的回避, 造成了某些地区年龄尾数上的指向, 从而导致地区数据质量存在问题。

三、异常发生的年龄区间讨论

上面我们探讨了年龄别人口数异常发生的年龄尾数指向问题, 然而, 事实上, 年龄别人口异常发生次数总会在一定的年龄区间, 下面我们再讨论年龄别人口数异常的年龄指向区间问题。

表 4 全国各地区五次年龄别人口数在各年龄段出现的异常次数

年龄段	2000 年		1990 年		1982 年		1964 年		1953 年	
	平均	最大	平均	最大	平均	最大	平均	最大	平均	最大
0~ 10	7.61	11	5.03	8	7.21	9	5.04	7	3.59	6
11~ 20	4.03	6	6.97	9	4.24	7	2.41	6	4.14	7
21~ 30	6.26	8	4.47	6	4.17	7	4.70	8	4.21	6
31~ 40	4.87	6	3.63	6	3.62	5	3.67	6	3.69	6
41~ 50	4.81	8	3.13	6	3.62	7	3.44	6	3.31	5
51~ 60	2.84	5	3.13	7	3.41	6	2.96	6	3.48	6
61~ 70	3.35	6	2.27	4	2.03	5	1.52	6	1.34	4
71~ 80	1.58	4	0.70	3	0.59	4	0.26	5	0.48	4
81~ 90	0.26	3	0.07	2	0.17	4	0.22	4	0.17	4
91~ 100	0.87	4	1.07	4	1.24	3	1.81	6	1.48	4

资料来源：同表 1。
说明：这里的“最大”指各地区的比较。

对年龄别人口数据异常的年龄区间指向类似于前面讨论年龄尾数的讨论。下面，我们先来看异常出现常常发生的年龄区间。我们将全国各地区五次年龄别人口数在各年龄段出现的异常次数绘制成表 4。由表 4 可知，2000 年我国大陆 31 个地区中，年龄别人口数异常发生最多的是在 0~ 10 岁；1990 年年龄别人口数发生异常最多的在 11~ 20 岁；1982 年全国各地人口数异常发生最多的同样在 0~ 10 岁；1964 年异常情况发生最多的在 21~ 30 岁；1953 年各地人口数异常发生最多的同样集中在 11~ 20 岁。通过全国各地五次人口异常数在年龄段的比较发现，0~ 10 岁是年龄别人口数异常发生最多的年龄区间，其次为 11~ 30 岁。原本以为年龄人口数异常发生应该会在高龄区间 91~ 100 岁，因为人到高龄，容易产生记忆错误，会错报年龄，从而使人口数据发生异常。然而，通过比较分析发现，尽管五次人口普查中 91~ 100 岁高龄组人口数存在异常现象，但与其他年龄组比较，并不是异常发生的主要年龄区间，而且 2000 年这种情况还得到了缓解。另外，年龄人口发生异常集中在 11~ 30 岁，应该从社会经济发展实践，如上学、工作、参军等事件年龄来考虑，从而可能引发年龄申报的误差。从 1982 年以来 0~ 10 岁低年龄组人口数异常发生的年龄区间是可以解释的，这应该从 1982 年以后出生人数逐年减少来寻找原因，除此以外，与年龄申报误差是否也有一定关系。据翟振武和陈卫在考察 1990 年代生育水平研究时，所得到的在历次普查中，低年龄组人口存在漏报，且漏报情况越来越严重^[4]。这可能也是导致低年龄异常次数增多的原因。

四、小结

最后简单地阐述一下本文所用的检验人口数据质量和目前常用的检验方法的区别。

常用的检验方法通常是这样构造的：根据一定的假设（例如稳定人口，或人口随年龄呈线性变化等），构造一个理论上的人口年龄结构，把它和实际人口作比较。预先给出一个临界值，观察实际人口与理论人口的偏离程度，如果偏离在临界值以内，则认为数据质量好或较好，否则，则认为数据质量差。问题在于人口的实际运动非常复杂，如中国人口，在不同时期有不同的特征，从人口再生产类型的增加到减少，中间还有 20 世纪 60 年代的大起大落，如果研究区域人口，还必须考虑迁移因素，很难用一定的模式来套中国人口的变动。这样，作为评判标准的临界值的确定就很难，定得太大，判定就变得模糊，定得太小，就会误判，把本来质量高的数据判定为有问题。

本文的检验方法，完全与上面的思路不同。这里把出现在某一年龄上的数据特征（例如一阶差分大于 0）看作一个随机发生的事件，人口变动对年龄数据的影响，只要不是周期性的，这种事件在各个年龄尾数上，应该是机会相等的，所以是均等分配的。与上面方法不同的是，这里并没有预先设定各个年龄尾数上应该发生的概率，但按照统计判定的习惯作法，设定了显著性水平。实际出现的事件，如果大于规定的范围，表示出现这样的现象可能性很小而（下转第 15 页）

只有对农村残疾人提供公益性康复服务，才能让他们享受到这些服务，这就需要政府加大资金投入力度。对于不同的残疾类型和残疾人群，需要采取有针对性的政策。

原中国残联主席邓朴方指出，中国残疾人大部分生活在农村，发展残疾人事业的着眼点和着力点应放到农村^[10]。在构建和谐社会和建设新农村的大背景下，农村残疾人已成为社会各界关注的一个焦点。《中共中央、国务院关于促进残疾人事业发展的意见》、我国签署联合国《残疾人权利公约》、实施《残疾人就业条例》以及举办 2008 年残奥会等一系列重要举措，都将为我国农村残疾人事业的不断发展带来新的契机。

参考文献:

[1] 国家统计局，第二次全国残疾人抽样调查领导小组. 第二次全国残疾人抽样调查主要数据公报（第一号） [EB/OL]，2006 <http://www.cdpf.org.cn/sytj/content/2008-04-07/content-84239.htm>

[2] 丛晓峰，唐斌尧. 转型期残疾人社会支持的实践模式研究 [J]. 北京科技大学学报（社会科学版），2003，（3）：2.

[3] 桂世勋. 更多地关爱独生子女夭折或残疾的家庭 [J]. 人口研究，2004，（1）：29- 30

[4] 许琳，王蓓，张晖. 关于农村残疾人的社会保障与社会支持现状研究 [J]. 南京社会科学，2006，（5）：97- 105.

[5] 陆杰华. 贫困地区人力资源开发与消除贫困研究 [J]. 人口研究，1998，（1）：60- 63.

[6] 朱宗福，郑青立. 我国农村残疾人职业康复浅析 [J]. 中国康复，1993，（3）：141- 142.

[7] 第二次全国残疾人抽样调查办公室. 第二次全国残疾人抽样调查主要数据手册 [M]. 北京：华夏出版社，2007.

[8] 同 [7].

[9] 周国富. 全方位推进城乡统筹发展 [N]. 人民日报，2004- 11- 16

[10] 邓朴方. 残疾人扶贫工作的经验与前瞻 [J]. 中国残疾人，2004，（5）：34.

[责任编辑 王树新]

(上接第 8 页)

加以否定，说明这种事件的发生，由特定的年龄指向，因而数据质量有问题。

与玛叶指数、韦伯指数和联合国综合指数等检验方法相比，本方法计算更为简单。对于那些公认为数据质量较高、较低的对象人口，用指数方法得到的结论，使用本方法也能得到相同的结论，而对于那些数据质量值得怀疑的对象（例如西藏，人口中绝大部分是文盲，藏族又没有如汉族一样，有记生肖作为年龄的习惯），指数方法未能判定出它的人口数据质量问题，本方法可以判定，说明本方法更加灵敏，有效。

本文从全国各地五次人口普查数据，分析了年龄人口数据质量问题，主要是误报的问题，并讨论了年龄尾数指向问题。研究发现，从整体来说，尽管全国人口普查数据并不存在年龄误报问题，数据质量可以接受，但是如果具体到各个地区而言，不少地区还是存在年龄指向问题，数据存在不同程度的质量问题。在那些存在数据质量问题区域，存在以“1”和“9”以及“0”和“8”为尾数的年龄指向。其中，“1”和“9”为尾数年龄回避，“0”和“8”为尾数年龄偏好。正是由于对“0”“8”“5”上的偏好以及在“1”“9”“7”上的回避，造成了某些地区年龄尾数上的指向，从而导致地区数据质量存在问题。

参考文献:

[1] 翟振武. 人口数据分析方法及其应用. 北京：外文出版社，1992. 13- 27.

[2] 黄荣清. 中国百万人口以上民族年龄报告准确性. 中国人口科学，1993，（5）.

[3] 黄荣清. 年龄别死亡率数据异常的检验与讨论. 人口与经济，2003，（6）.

[4] 翟振武，陈卫. 1990 年代中国生育水平研究. 人口研究，2007，（1）.

[责任编辑 王树新]