年龄别死亡率数据异常的检验与讨论

黄荣清
(首都经济贸易大学 人口经济研究所, 北京 100026)

摘要: 由人口调查结果计算得到的年龄别死亡率, 一般不能直接应用于人口研究中, 还需要对原始数据进行检验和修匀。本文提出了用死亡率的差分来检验、判定调查死亡率异常的发生, 并用它分析第五次人口普查中全国和各省区人口死亡率的数据质量, 在此基础上, 讨论了发生异常的原因。

关键词: 人口死亡率; 异常; 年龄尾数
中图分类号: C921 文献标识码: A 文章编号: 1000- 4149 (2003) 06- 0010- 06

Examining of and Discussion about Exceptional Death Rate by Age

HUANG Rong-qing
(Institute of Population and Economics, Capital University of Economics and Business, Beijing 100026)

Abstract: Without essential examination and smooth, the death rate by age from population survey can't be used directly in population research. This paper proposed to examine the occurrence of exceptional death rate through death rate difference method, and then judge data quality about population death rate by province. Based on this, the paper discussed why exceptional death rate occurred.

Keywords: population death rate; out of way; last digit of age

人口死亡率, 作为人口学的基础数据之一, 广泛地应用于各种人口研究中。因为是“基础”, 所以它的数据质量就非常重要。由人口普查的结果, 可以计算得到人口年龄别死亡率, 并由此编制出生命表来。有些学者认为, 关于人口死亡率的工作就可到此为止。其实并不然, 上述工作只是普查死亡率研究工作的一步, 真正要把普查的死亡率数据用到人口研究中去, 还有很多工作要做。打个比喻来说, 上述工作在工业产品生产中只是完成了从原料到毛坯的工序。产品的毛坯虽然重要, 但它尚不能直接使用, 还需要经过一系列加工, 如切削, 磨平, 抛光等工序, 最后通过质量检验才是制成品。死亡率的计算工作也同样, 设想一个由普查结果直接计算得到的死亡率, 如果它在各年龄上高低起伏 (很可能会有这种情况发生, 本文图 6 就是具体例子) 不经加工, 把它用于人口预测、保险精算中, 肯定是不合适的。由原始数据计算得到的人口死亡率, 同样也要经过一系列加工过程, 包括检验、修匀等才能用于实际, 而其中第一步工作就是数据质量检验。本文以下将用死亡率的差分来检验死亡率的数据质量, 并由此找出由“五普”得到的全国和各省区的人口死亡率存在的问题, 并讨论产生这些问题的原因。

收稿日期: 2003- 09- 03
作者简介: 黄荣清 (1946-), 江苏无锡人, 研究员, 首都经济贸易大学人口经济研究所所长。

一、年龄别死亡率变动的异常和异常年龄

在一般情况下，人口死亡率的大小随年龄而变化，是非常有规则的。从出生到青少年阶段，即从0岁到10~14岁，死亡率随年龄增加而减少，以后，死亡率就随年龄增加而上升。在上升阶段，开始只是缓慢增加，到了老年阶段则速度加快。从图形上看，死亡率曲线呈“U”型（高死亡力场合）或“J”型（低死亡力场合）。但有时也会出现这样的情况：在青年（一般在20~40岁）某个阶段，从某个年龄起，死亡率增加速度加快，过了若干年龄后又迅速回落，然后又回到原来的变动轨道。在图形上，在“U”型或“J”型曲线上出现了一个“隆起”现象。但从我国的死亡率数据看，这种现象并不明显，我们不妨把它看成只是特殊现象。对一般死亡率的变化，用数学符号来表示，若 m_x 表示年龄 x 岁的死亡率，m_x' 表示死亡率的导数，则

当 $x < x_0$ 时 $m_x' < 0$

$x > x_0$ 时 $m_x' > 0$

如果用差分形式来表示，设 $\triangle m_x = m_x - m_{x-1}$（$x = 1, 2, 3, \cdots$）则

当 $x < x_0$ 时 $\triangle m_x < 0$

$x > x_0$ 时 $\triangle m_x > 0$

从实际数据看，x_0 在10~14岁之间。但不同人口在不同时期，其位置可能会发生一些小的变化，有时在10岁、11岁，有时出现在12岁、13岁。为了简单起见，在以下计算中，一律取 $x_0 = 12.5$。

由调查资料直接计算得到的死亡率数据，由于有随机误差（因为死亡发生可以看成是个随机事件）和非随机误差存在，并不总是符合上述规律的，即在 $x < x_0$ 的某些年龄，可能出现 $\triangle m_x > 0$；而在 $x > x_0$ 的某些年龄，可能出现 $\triangle m_x < 0$。我们把出现上述的情况认为是“异常”，出现异常情况的年龄称为异常年龄。下面我们来看一下全国和各省（区）死亡率出现的异常情况。

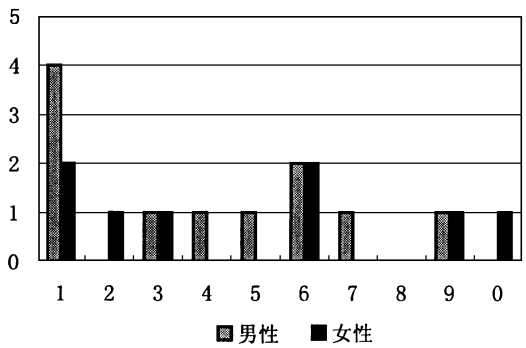

图1 各年龄尾数出现的“异常”频数（全国，2000）

从全国的数据看，男性出现“异常”的次数共有11次，女性出现“异常”的次数共有8次（见图1）。如果把全年龄分为1~10岁，11~20岁，…91~100岁等10个年龄段，则男性在91~100岁年龄段出现了5次，31~40岁年龄段出现了2次，在其他年龄段未出现异常或只发生了1次；而女性除在31~40岁年龄段出现了3次外，其他年龄段最多出现1次。而在31~40岁年龄段，出现异常的年龄并不是连续的，而是间断发生的。这也说明了我国人口死亡率曲线中在青年阶段并不存在“隆起”现象。

表1是2000年普查中，31个省（市、自治区）死亡率数据出现的“异常”次数。男女合计的死亡率出现异常次数较少的是江苏、浙江、福建、山东、四川，这些省区出现异常次数都在10次以下，而异常次数出现较多的是新疆、宁夏、青海、西藏，异常次数出现都在25次以上。男性死亡率异常次数出现较少的地区有浙江、四川、江苏，出现次数在12次及12次以下，出现较多的地区是新疆、河南、天津、青海、西藏、宁夏，出现次数都在25次以上。女性死亡率出现异常较少的地区是安徽、浙江、山东、贵州等，出现异常较多的地区是青海、新疆、海南、宁夏、西藏，其中以西藏为最多，共出现了37次异常值。

明显地，死亡率数据的质量和异常出现的次数多少有密切关系。异常情况出现少，说明死亡率随年龄变化的数据平滑、稳定，数据质量较高；异常次数出现多，说明死亡率值高低起伏，变

化不稳定，数据质量较差。总的来说，全国的数据异常出现的次数较少，说明数据质量是不错的。从各省区看，东部沿海，特别是华东地区的几个省，数据质量较好，而西部地区的数据质量较差。

表 1　我国大陆各省（市、自治区）人口死亡率在各年龄出现的“异常”次数（2000 年）

地区	男女合计			男性			女性		
	合计	最大	最小	合计	最大	最小	合计	最大	最小
北京市	18	3	0	21	5	0	17	4	0
天津市	19	6	0	30	7	0	22	4	0
河北省	15	4	0	18	4	0	15	5	0
山西省	14	4	0	24	5	1	19	5	0
内蒙古	16	3	0	17	2	1	16	3	1
辽宁省	13	3	0	14	4	0	18	3	1
吉林省	14	4	0	14	4	0	16	4	0
黑龙江	20	4	0	21	4	1	19	4	0
上海市	13	3	0	15	3	1	24	4	0
江苏省	7	1	0	12	2	0	13	3	0
浙江省	7	2	0	11	3	0	11	2	0
安徽省	14	3	0	21	4	0	10	2	0
福建省	10	3	0	17	4	0	19	5	0
江西省	15	3	0	15	3	0	16	4	0
山东省	10	2	0	13	3	0	12	3	0
河南省	21	8	0	21	8	0	17	5	0
湖北省	15	4	0	23	5	0	18	4	0
湖南省	19	8	0	22	8	0	18	6	0
广东省	11	3	0	14	5	0	14	3	0
广　西	11	2	0	13	3	0	17	3	0
海南省	23	6	0	28	5	1	31	6	1
重庆市	12	2	0	18	4	1	16	3	0
四川省	10	2	0	11	3	0	12	3	0
贵州省	11	2	0	19	5	1	12	3	0
云南省	15	4	0	13	4	0	20	4	0
西　藏	33	7	1	34	7	0	37	6	1
陕西省	15	3	0	22	4	1	18	3	1
甘肃省	16	3	0	19	3	0	17	4	0
青海省	30	6	1	32	6	2	29	5	1
宁　夏	29	6	1	35	6	0	31	5	2
新　疆	25	6	1	27	7	0	30	8	1

数据来源：根据 2000 年人口普查得到的死亡率计算结果。

注：这里的“最大”和“最小”是指比较在各年龄尾数上出现的异常数。

二、异常发生的年龄尾数指向

1. 年龄尾数的指向

人口死亡率的异常值的出现，在各年龄尾数上，并不是平均分配，而是有一定指向的。从全国男性死亡率来看（见图 1），异常出现次数最多的是以“1”为年龄尾数，共出现了 4 次，其次是以“6”为年龄尾数，出现了 2 次，在其他年龄尾数上不出现或出现 1 次。从女性死亡率看，异常出现较多的是发生在以“1”和“6”为尾数的年龄，其他年龄尾数亦较少出现或不出现。上述情况，在各省区也普遍存在。下面，我们来讨论异常出现特定指向的检验问题。

设某个人口死亡率在各年龄尾数“i”上出现异常次数分别为 n_i（i= 1，2，3，……，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特定指向，或者说发生的概率相同，则在各年龄尾数上出现的概率为 $p= n/100$ 其方差为 $D= p(1- p)$。在每个年龄段相当于作 1 次实验，10 个年龄段相当于作 10 次试验，则样本标准差 $s= \sqrt{D/10}= \sqrt{p(1- p)/10}$。设置信概率为 0.95，则 95% 的置信区间为

$$\frac{|P_i - P|}{s} \leqslant 1.96$$

$$|P_i - P| \leqslant 1.96s$$

$$P - 1.96s \leqslant P_i \leqslant P + 1.96s$$

或者说　　$n/10 - 19.6s \leqslant n_i \leqslant n/10 + 19.6s$

即是说，如果 n_i 总是在 $n/10 - 19.6s$ 和 $n/10 + 19.6s$ 之间，则我们可以认为上述假设是成立的，即异常的出现在各年龄尾数的可能性是相同的。实际数据中出现不同的 n_i，只是由于随机变动引起的。反之，若出现有 n_i 在置信区间（$n/10 - 19.6s$，$n/10 + 19.6s$）之外，则认为死亡率数据异常的发生在各年龄尾数的机会均等的假设不成立，特别是如果出现 $n_i > n/10 + 19.6s$，则我们称“i”为死亡率异常特定指向的年龄尾数。

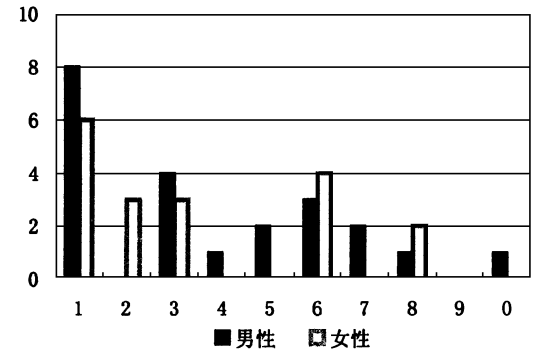

图 2　各年龄尾数出现的“异常”频数
（湖南，2000 年）

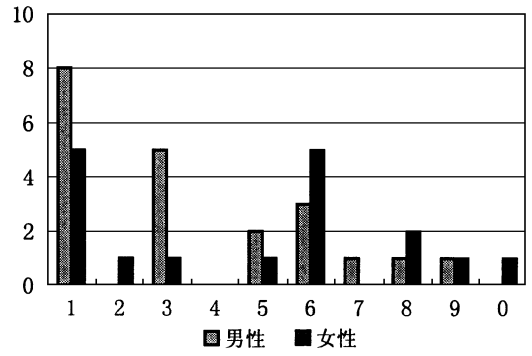

图 3　各年龄尾数出现的“异常”频数
（河南，2000）

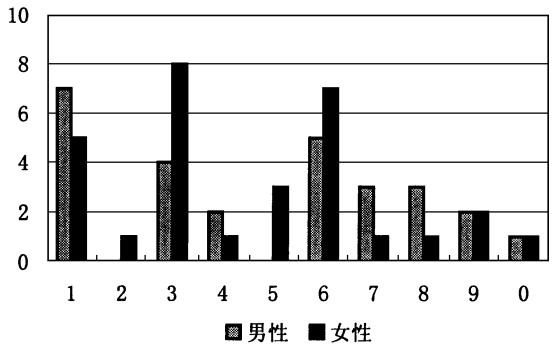

图 4　各年龄尾数出现的“异常”频数
（新疆，2000）

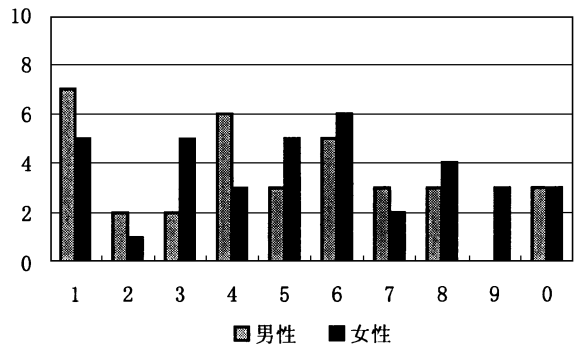

图 5　各年龄尾数出现的“异常”频数
（西藏，2000 年）

下面，我们来看一下全国各省（市、自治区）死亡率异常发生年龄尾数指向检验结果。

男性：以尾数“1”为指向的有河南、湖南、云南、新疆、西藏、贵州；以尾数“6”为指向的有广东，另外天津出现了以“2”为特定的年龄指向。

女性：以年龄尾数“1”为指向的有湖南、河北；以尾数“6”指向的有河南、福建；比较特殊的，新疆以尾数“3”为指向。

男女合计：湖南、河南以尾数“1”为指向，天津、河南以年龄尾数“6”为指向，新疆在尾数“1”和“6”为特定指向。

由上面的特定年龄尾数指向可以知道，特定的年龄尾数指向，与异常发生数 n 的大小并无直接关系，而与 n_i 与 $n/10$ 差的大小有关系。例如，全国男性死亡率在尾数“1”上异常出现了 4

次，可以认为在尾数“1”上有年龄指向，宁夏在尾数“1”上出现了6次，却并没有认为它在“1”上有特定的年龄指向，只认为它是随机出现的。这是因为宁夏死亡率异常次数出现了35次。在出现异常的每个年龄机会相等的假设下，n在它的95%置信区间内。全国男性死亡率在尾数“1”上虽然只出现了4次，但却在均等分配假定的置信区间之外。

另外，我们还要注意到上述说的各地区的年龄指向是有条件的，即在95%的置信概率下。如果置信概率增大，如在99%以内，则有特定年龄尾数指向的地区就不会有那么多，反之，如果置信概率缩小，如假定在80%条件下，则会有更多的地区有特定的年龄尾数指向出现。

有一个现象是过去研究中未被察觉的。通常我们认为西部少数民族地区人口年龄申报常常有堆积现象，所以自然就猜想这些地区会出现死亡率的异常，并出现特定的年龄指向。中原地区由于年龄申报堆积不太严重，所以估计死亡率也不会出现特定的年龄指向。本文的研究却发现了这样的现象：在一些中原地区，如湖南、河南，死亡率异常向特定年龄尾数的指向，要比西部地区还要强。例如把置信概率定在99%，则可认为西藏、新疆等并没有特定的年龄尾数指向，而湖南、河南却依然很明显。这是因为虽然西藏、新疆的死亡率异常次数出现较多，但它们分散在各个年龄尾数上，河南、湖南的死亡率出现的异常次数虽然不算多，但它们却非常集中（见图2~图5）。

2. 特定的年龄尾数指向发生的原因

所谓死亡率异常，则是按照死亡率变化的一般规律，在应该出现$\triangle m_x > 0$的年龄上，却出现了$\triangle m_x < 0$；而在应该出现$\triangle m_x < 0$的年龄上却出现了$\triangle m_x > 0$。按一般规律，在绝大部分年龄都是$\triangle m_x > 0$，所以以下我们只讨论异常$\triangle m_x < 0$的情况。由于

$$\triangle m_x = m_x - m_{x-1}$$

当$\triangle m_x < 0$，则说明m_x数据过小，或m_{x-1}过大。

死亡率m_x是死亡人口D_x与平均人口P_x之比。在不考虑死亡人口漏报和重报（或人口的漏报和重报）的情况下，人口和死亡人口的年龄误报，且人口和死亡人口在不同年龄误报的比例不同，会导致死亡率m_x的误差产生。我们需要补充的是，利用人口普查数据计算死亡率，即使在同一年龄上死亡人口和人口误报的比例相同，但在不同年龄上有不同的误报，同样也会导致m_x误差的产生。以年龄尾数为“1”的差分$\triangle m_{(1)} = m_{(1)} - m_{(0)}$为例，对于$m_{(0)} = 2D_{(0)}/(P_{(0)} + P_{(0)}^{(-1)})$，（这里$P_{(0)}^{(-1)}$表示普查前一年的“0”为尾数的人口。）由于$P_{(0)}^{(-1)}$是用$P_1$来估算的，如果死亡人口和在以“0”为结尾的年龄上有堆积，而在以“1”为结尾的年龄上无堆积，那么，在$D_{(0)}$人口堆积程度大于或等于$P_{(0)}$人口的堆积，则计算出的死亡率$m_{(0)}$就会过大。类似地，对于$m_{(1)} = 2D_{(1)}/P_{(1)} + P_{(1)}^{(-1)}$只要在以“1”为年龄尾数的$D_{(1)}$有年龄回避或回避程度小于或等于$P_{(1)}$人口的回避程度，而在$P_{(2)}$，如又有少量的人口堆积，则就会导致$m_{(1)}$过小的情况出现。

由上面的检验可以知道，死亡率异常的出现大部分发生在以“1”和“6”为年龄尾数的年龄上。依照上面的分析，这种异常的产生正是和人口在“0”和“5”尾数的申报堆积有关。对于前面列举的地区，我们可以有95%的把握认为它们存在着人口和死亡人口申报的年龄堆积。如果再放宽一些标准，例如在80%的置信度上，则可以肯定有更多的地区，特别在“0”，“5”年龄尾数上有申报堆积。有一个现象还有待于解释，死亡率异常的年龄尾数在一些地区不相同，男性在“1”的尾数（即在尾数“0”上堆积）较女性多，而女性在“6”的尾数（即在尾数“5”上堆积）较男性多。

三、异常发生的年龄区间

上面我们讨论的是死亡率异常发生的年龄尾数指向问题，下面我们再讨论死亡率异常的年龄区间指向问题。

表 2　各省（市、自治区）人口死亡率在各年龄段出现的异常次数

年龄段	男女合计		男性		女性	
	平均	最大	平均	最大	平均	最大
1~ 10	1. 7	5	2. 2	4	1. 8	5
11~ 20	1. 6	6	2. 1	5	2. 7	4
21~ 30	2. 9	6	3. 4	6	3. 5	6
31~ 40	2. 6	5	2. 5	5	3. 1	6
41~ 50	1. 2	4	1. 4	5	1. 5	3
51~ 60	0. 7	3	1. 0	4	1. 3	5
61~ 70	0. 6	3	0 6	3	0. 8	4
71~ 80	0. 5	4	0 5	3	0. 5	4
81~ 90	0. 9	5	1. 4	5	0. 8	5
91~ 100	3. 5	6	4. 7	8	2. 8	5

资料来源：同表 1。

注：这里的“最大”指 31 个地区的比较

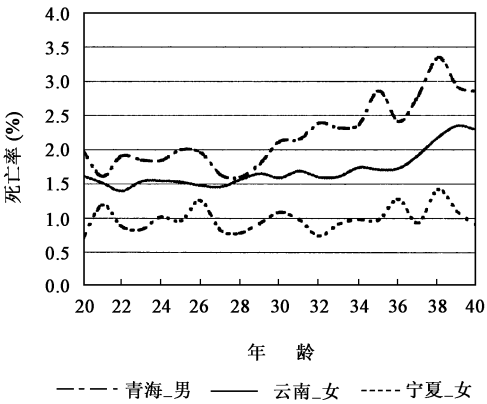

图 6　20- 40 岁年龄别死亡率

对死亡率异常的年龄区间指向类似于上面对年龄尾数的讨论。下面，我们先来一般地看一下异常的出现常常发生在那些年龄区间。由表 2 可知，我国大陆 31 个地区中，男女合计，男性的死亡率异常发生最多的是在 91~ 100 岁。而女性则在 21~ 30 岁，31~ 40 岁年龄区间为多。对于高龄区间 91~ 100 岁，异常的出现几乎可以肯定是死亡申报的误差所致。而对于 21~ 40 岁，我们还需要观察一下这种异常是否由于特殊的死亡类型，即在青年阶段死亡率曲线有“隆起”的现象存在造成的。下面我们来选择在 21~ 40 岁异常次数出现最多的几个地区，青海男性、云南女性、宁夏女性来观察他们的年龄别死亡率数据（图 6）变化。

观察图 6，云南、宁夏女性的年龄别死亡率完全没有“隆起”的迹象，异常的出现完全是年龄别死亡率波动所致。青海男性在 22~ 25 岁似乎有“隆起”的迹象，但如果考虑到误报的存在，则至少“隆起”部分并不突出。所以，一般地我们也可以认为，女性死亡率变动的异常现象，也是由于死亡报告误差造成的。

男性死亡率发生异常情况是在高龄区间居多。高龄区间误差大，本来就是可以想像的，这是因为人到高龄，容易产生记忆错误，会错报年龄；另外，高龄时人口较少，死亡率的随机误差也大。对于高龄，只要死亡率基本符合死亡率变动规则，即随年龄上升而上升，则还是可认为数据是基本可用的（当然也要修匀）。如果高龄死亡率数据与死亡率变动规律不同，则对原始数据不是修匀，而是要修正了。下面，我们来考察一下高龄死亡率的情况。

为了减少死亡率变动的影响，我们来观察 5 岁组的死亡概率，以${}_5Q_x$ 表示（x，x+ 5）的死亡概率，我们来观察$\triangle {}_5Q_x = {}_5Q_x - {}_5Q_{x-5}$的值，可以认为，如果$\triangle {}_5Q_x < 0$，则数据有误。对全国及各省区数据观察可以发现，以下地区的高龄人口死亡率有误。

男性：青海在 85 岁以上，甘肃和西藏在 90 岁以上，全国及大部分地区在 95 岁以上的数据都有误。

女性：山西、西藏、甘肃、宁夏在 95 岁以上。

从上面的情况可以知道，在高龄死亡率，女性的数据质量要好于男性。

从上面的讨论中，我们已经知道直接由人口普查计算得到的死亡率数据质量，一般多少还是有些问题的，在有些地区，问题还较多。所以，在实际工作中使用死亡率时，还需对原数据做一番修匀加工工作。

[责任编辑　崔凤垣]