【人口分析方法与应用】

人口年龄结构模型和它的应用

黄荣清

【摘　　要】提出了人口年龄结构模型——以年龄为自变量、累计的年龄百分比为因变量的函数形式,通过使用中国历次人口普查数据和其他一些数据对模型的验证,表明模型是成立的。累计的年龄百分比模型经过两次对数变换后,可以表示为线性函数的形式。以这一模型为基础,进一步构建了人口百分比、前后两次人口普查对应的年龄人口百分比之比等数学函数的表达式。在使用人口普查资料检验时发现,模型虽然能很好地拟合累计的年龄人口百分比曲线,但当人口的年龄波动较大时,年龄百分比模型的残差就会变大。由此可以得出这样的结论:用一个简单的数学函数要准确表示一般的年龄人口百分比是做不到的,若没有其他数据支持,根据一次普查的数据要全面准确判断人口普查数据的准确性也是不可能的。为了判别普查数据的报告误差,把普查数据拆分为估计值、偏离值和误报三部分,这里的估计值就是年龄百分比的模型值。本研究证明,在封闭人口条件下,年龄偏离系数(偏离值与估计值之比)是个常数。利用这个性质,可以用两次普查的百分比模型值计算实际人口的年龄存活率,并通过估计年龄偏离系数,估计出普查的误报。利用上述模型,本文估算了1982年全国人口普查的年龄误报情况。根据估算1982年7-91岁的年龄误报有683万人,年龄误报率为6.74‰。由于年龄误报,一些年龄的报告人口比估计的实际人口多,它们主要出现在中青年期,即青年期(24岁和25岁)和中年期,共计340万人,而一些年龄的报告人口少于估计的实际人口,它们主要分布在青年期(17、18岁和21岁),共计342万人。

【关 键 词】年龄结构模型;人口普查;普查数据修正

【作者简介】黄荣清,首都经济贸易大学人口经济研究所教授、博士生导师(北京　100070)。

【原文出处】《人口与经济》(京),2023.6.56~70

人口以规模和结构衡量,在各种结构中,又以人口年龄结构最为重要。年龄是人口研究中最常用的基础变量,人口研究的许多领域都是以年龄为变量而展开的。例如人们熟知的年龄别生育率、年龄别死亡率、年龄别劳动参与率,等等,这些指标分别是人口生育研究、人口死亡研究、就业研究中最重要的基础指标,而这些指标都必须以分年龄人口为基础计算得到。如果人口数据不准,那么以此为基础计算出的结果和得出的结论的可信度就令人存疑了。所以,一个国家和地区的各年龄的人口数以及它所占总人口的比,即人口年龄结构数据的准确性特别受人重视。

在使用人口数据时,首先要检查所用的数据是否准确可靠,这是人口学专业必须的基础训练。人口学学者对人口数据质量的重视程度,可以从人口统计分析教科书的内容看出:这些教科书一般都是从人口数据的来源和对数据质量的评估开始的。在人口数据质量的评估方面,人口学家提出了一些方法,如检验在某个年龄尾数上报告是否有偏好或排斥的方法有惠普尔指数、迈耶尔指数、联合国的年龄—性别准确性指数。黄荣清提出了用差分或者用每个年龄的数量与它左右两边年龄的人口数之差的符号分布来检验年龄尾数上是否有偏好或排斥[1-2]。此外,可以通过两次普查数据的一致性,即对应年龄人口比的大小来判定普查中是否存在漏报和误报,等等。可以说,在判断人口数据的准确性方面,人口学界已经积累了许多方法,虽然这些方法尚有改进的余地。但我们同时也注意到,即使检验出数据有质量问题,如何修正却没有被很好解决,甚至可以说有点束手无策。例如,我们可以判定某个人口的数

据在尾数为“0”的年龄上有堆积,但我们并不能确定是在每个尾数为“0”的年龄都有重报,还是在部分年龄上有重报;即使我们能确定在某个年龄上有重报,例如在40岁,但我们并不能确定是40岁以前还是40岁以后的人的误报,若我们已经确定是年龄高报,即40岁年龄以前的人报告到40岁,那到底是39岁,还是38岁,或者是35岁、36岁的人误报,还是35岁到39岁的人都有误报?若进一步问,他们误报的数量和比例是多少?诸如此类的问题,从现有的研究来看,我们并不能清晰地回答。

究其原因,实际上我们对人口数量在各年龄之间的相互联系,或者说人口年龄结构数量变化的规律还没有清楚的认识。在一些特殊的条件下,我们有已知的阐述年龄结构的模型。如在生育水平和死亡水平保持不变,且两者相等的条件下,人口的年龄结构等于生命表中的静止人口年龄结构;在放宽相等的条件,保持死亡水平、生育水平不变的条件下有稳定人口年龄结构。但上述模型都是建立在理论假设下,现实人口中,死亡水平、生育水平保持长期不变几乎不存在,尤其如近现代的中国,经历了百年翻天覆地的变化,生育水平、死亡水平都发生了急剧的变化,显然是无法用稳定人口模型来解释中国人口的年龄结构及其变化的,所以,我们需要设计一个更加普遍适用的模型来刻画人口年龄结构的特征,并解释在现实的人口统计中出现的种种问题。

一、年龄结构模型

人口年龄结构模型,就是以年龄为自变量、年龄结构为因变量的数学函数。

1. 累计的年龄百分比模型

年龄结构常常以某一年龄的人口占总人口的比重来表示。设 x 年龄的人口为 p_x,x 岁及以上的人口为 P_x[①],其表达式如下:

$$P_x=\sum_x^{\infty}p_x \tag{1}$$

$$P=\sum_0^{\infty}p_x \tag{2}$$

另设 $a_x=\frac{p_x}{P}$,$A_x=\frac{P_x}{P}(x=0,1,\cdots,n)$。由上面的定义可知,$a_x$ 表示 x 岁年龄的人口占总人口的百分比,A_x 表示 x 岁及以上年龄的人口占总人口的百分比。很显然,A_x 有以下性质:①$A_0=1$,$1>A_x>0$(当 $x>0$)。②$A_x\geqslant A_{x+j}$(当 $j>0$),即 A_x 随年龄增加而变小。③当 $x\to\infty$ 时,$A_x\to 0$。

根据以上性质可以知道,随年龄 x 的变化 A_x 的值由1逐渐变化到很小的一个数,当接近于0时,A_x 在年龄之间的差别就会变得很小而不易识别。这时,我们可以通过对 A_x 进行对数变换,即取 $\ln(1/A_x)$,经过变换后,年龄之间的差别就会非常明显。

下面,我们来构造累计百分比的连续函数 $\ln(1/A(x))$,在每一个年龄点上满足以下公式:

$$\ln\left(\frac{1}{A(x)}\right)=\ln\left(\frac{1}{A_x}\right)\quad x=0,1,2,\cdots,n \tag{3}$$

假设 $\ln(1/A(x))$ 有函数形式 $x^b\exp(C(x))$,这里我们虽然不知道 $C(x)$ 的函数形式,但由数学分析可知,在满足一定条件下,一般的函数都可以展开成泰勒级数,即多项式的形式:

$$C(x)=C_0+C_1x+C_2x^2+C_3x^3+\cdots+C_kx^k \tag{4}$$

设 $\exp(C_0)=A$,则:

$$\ln\left(\frac{1}{A(x)}\right)=A\cdot x^b\exp(C(x))\quad x=0,1,2,\cdots,n \tag{5}$$

其中,$C(x)=C_1x+C_2x^2+C_3x^3+\cdots+C_kx^k$,是关于 x 的 k 次多项式。A、b 是待定参数,$A\cdot x^b\exp(C(x))$ 被称为累计百分比模型。

设模型 $\ln(1/A(x))$ 和观测值 $\ln(1/A_x)$ 有以下关系:

$$\ln\left(\frac{1}{A_x}\right)=\ln\left(\frac{1}{A(x)}\right)(1+D_x) \tag{6}$$

其中,D_x 表示模型值和观测值在 x 岁的相对偏差。

将公式(5)代入上式,两边取对数得:

$$\ln\left(\ln\left(\frac{1}{A_x}\right)\right)=a+b\ln(x)+C(x)+D_x\quad a=\ln(A) \tag{7}$$

即:

$$D_x=\ln\left(\ln\left(\frac{1}{A_x}\right)\right)-(a+b\ln(x)+C(x)) \tag{8}$$

为了估计出模型参数,设参数 a、b、$c_j(j=1,2,\cdots,k)$ 使残差平方和达到最小,即:

$$\Delta(a,b,c_1,c_2,\cdots,c_k)=\min\sum D_x^2=\min\sum\left(\ln\left(\ln\left(\frac{1}{A_x}\right)\right)-(a+b\ln(x)+C(x))\right)^2\quad x=1,2,3,\cdots,n \tag{9}$$

$a+b\ln(x)+C(x)$ 为线性函数的形式,用最小二乘法可以估计出对应的参数,从而计算出相应的模

型值 $\ln(\ln(1/A(x)))$。对于公式(9)的方程,根据1982年全国人口数据估计出的参数见表1[②]。

表1　　1982年人口回归方程的参数

名称	a	b	C_1	$100C_2$	100^2C_3	100^3C_4	100^4C_5
数值	-3.860	0.718	0.099	-0.255	0.324	-0.156	0.020

由公式(9)估计出模型参数,就可以分别算出模型值 $\ln(1/A(x))$ 和 $A(x)$,进一步可以计算观测值 $\ln(1/A_x)$、A_x 与相对应模型值的残差。

用各次全国人口普查的数据来检验,上述累计百分比人口模型都是成立的。这里仅以第三次全国人口普查(1982年)数据估计出模型参数,对观测值 $\ln(1/A_x)$、A_x 和模型值 $\ln(1/A(x))$、$A(x)$ 进行比较(见图1和图2),可见两者是非常接近的。

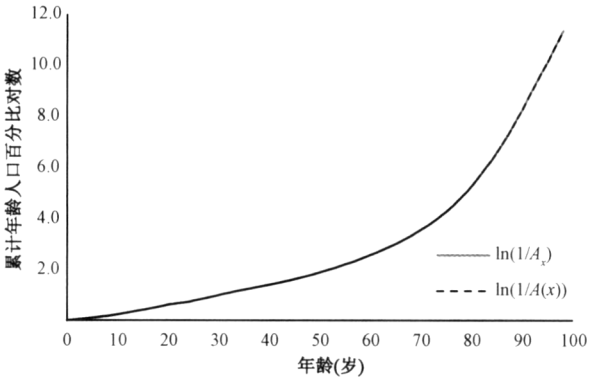

图1　累计年龄别人口百分比对数的观测值和模型值(1982年)

数据来源:由1982年第三次全国人口普查资料计算得出。

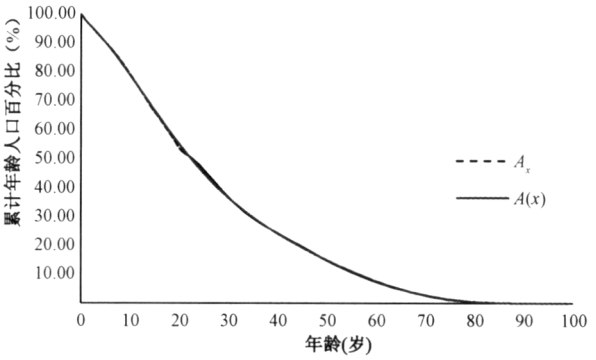

图2　累计年龄别人口百分比的观测值和模型值(1982年)

数据来源:由1982年第三次全国人口普查资料计算得出。

在统计学中我们通常用确定系数 R^2 来评价模型对观测值的拟合效果,设观测值为 Y_x,模型值为 $Y(x)$,Y_x 的均值为 $\bar{Y}$,则有:

$$R^2=\frac{\sum(Y(x)-\bar{Y})^2}{\sum(Y_x-\bar{Y})^2} \tag{10}$$

R^2 越接近于1(如用百分数表示则为100%),表明模型曲线对观测值的拟合越好。对1982年人口普查数据的拟合效果如表2所示。

表2　　拟合曲线的 R^2 值

观测数据	R^2(%)
$\ln(\ln(1/A_x))$	99.99
$\ln(1/A_x)$	100.03
A_x	100.01
a_x	94.55

前三个模型值的 R^2 几乎都接近于100%,说明曲线拟合得很好,模型的精度很高。用其他各次人口普查的数据拟合可得出相同的结果,可见模型是成立的。

2. 年龄百分比模型

以下我们来讨论年龄百分比 a_x,并定义对应的模型值 $a(x)$。

由于 $a_x=A_x-A_{x+1}\approx A(x)-A(x+1)\ x=0,1,2,\cdots,n$,由拉格朗日中值定理可知:

$$A(x)-A(x+1)=-A'(x+\beta),\beta\in(0,1) \tag{11}$$

由于 $\ln\left(\frac{1}{A(x)}\right)'=-\frac{A'(x)}{A(x)}$,所以:

$$A'(x)=-A(x)\cdot\ln\left(\frac{1}{A(x)}\right)\left(\frac{b}{x}+C'(x)\right) \tag{12}$$

其中,$C'(x)$ 为 K 次多项式 $C(x)$ 的导数,为 $k-1$ 次多项式。由公式(9)估计出的参数就可以分别算出 $A(x)$、$\ln(1/A(x))$ 和 $b/x+C'(x)\ x=1,2,3,\cdots,n$ 各项值,从而确定 $A'(x)$ 的值。

在公式(11)中,β 是区间(0,1)中的一个数,对不同的$(x,x+1)$区间内,β 值可能会有所不同,精确地确定 β 值有一定的难度,从应用的角度看,我们可以简单地取 $\beta=0.5$。

定义 a_x 的模型值 $a(x)$ 如下:当 $x=0$ 时,$a(0)=1-$

$A(1)$;当 x 为其他年龄时,$a(x)=-A'(x+0.5)$,即:

$$a(x)=A(x+0.5)\cdot\ln\left(\frac{1}{A(x+0.5)}\right)\left(\frac{b}{x+0.5}+C'(x+0.5)\right)\quad x>0 \qquad (13)$$

以下是1982年人口的观测值 a_x 和模型值 $a(x)$ 的比较,见图3。

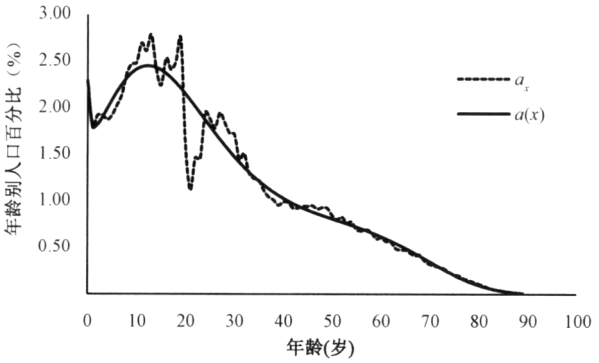

图3 年龄别人口百分比 a_x 的观测值和模型值(1982年)

数据来源:由1982年第三次全国人口普查资料计算得出。

与累计百分比模型相比,观测值 a_x 和模型值 $a(x)$ 的残差要大得多。以1982年全国人口普查数据的确定系数 R^2 作比较,累计百分比模型中 R^2 接近100%,而年龄百分比模型不到95%(见表2最后一行)。累计模型的拟合效果明显要比年龄百分比模型好得多,这是因为模型通常是通过过滤并消除掉观测数据高低的起伏后显示其均匀变化的结果。但是,在"过滤和消除"过程中,它常常并不区分(或很难区分)观测数据出现的高低起伏到底是由实际数据本身的高低起伏引起的还是由于测量错误(对人口普查的数据来说,是报告错误)引起的。对累计百分比的数据来说,由于各年龄的报告误差和波动经过累计后相互抵消,即累计本身就在过滤,这样累计百分比的观测数据的起伏就变小,观测值和模型值就会很接近。而年龄百分比模型是由累计百分比模型"派生"出的,累计百分比模型已经消除了这些起伏,数据图形显示是光滑的,在年龄百分比的观测数据中,当数据变化不稳定时,图形显示为高低起伏,a_x 和 $a(x)$ 的残差就会明显。

当 a_x 相邻年龄的值差距不大时,模型值 $a(x)$ 和观测值 a_x 就会比较接近。图4为中国人口在1953年 $a(x)$ 和 a_x 的比较,可以看出,1953年年龄百分比的观测值和模型值除了在18岁附近残差较大外[③],在其他年龄是相当接近的($R^2=98\%$),而1982年观测值 a_x 与 $a(x)$ 的差别比较大。究其原因,是由于30岁以前的多数的年龄别人口比例观测值 a_x 波动比较大会导致其与模型值的差很大,故1982年年龄别百分比模型的残差却比1953年要大。

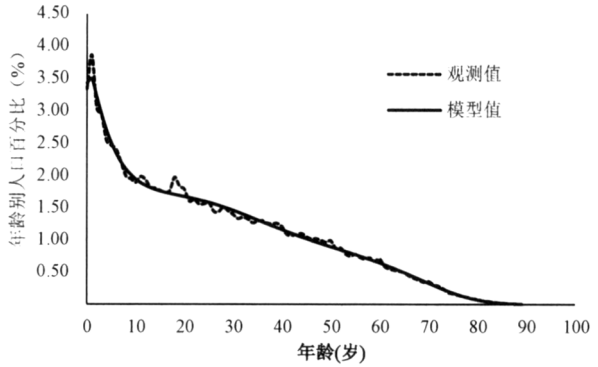

图4 年龄别人口百分比 a_x 的观测值和模型值(1953年)

数据来源:国务院人口普查办公室,国家统计局人口统计司.中国1982年人口普查资料[M].北京:中国统计出版社,1985:539-541。

从实际来说,无论是从动员、准备、组织实施和数据处理方面来比较,1982年普查要比1953年普查充分得多,从经验上说,1982年普查数据的质量应该比1953年的普查数据质量要高,因此模型的残差高产生的原因不是由于报告误差引起的,而是由于这两年的数据起伏不同引起的。

在大部分观测值接近理论值,而有个别观测值距离理论值较大时,统计学有这样的处理方法:即把这些点作为异常值排除在外再进行拟合,如1953年的数据,把年龄在18岁这点排除后,模型曲线的拟合度就会提高。但如1982年的情况,在许多年龄点都和模型值有较大距离的情况下,上述方法就不适用了。因为年龄百分比观测值的高低起伏,可能是报告错误所致,也可能是人口年龄结构自身的特征所致。如果加以剔除,就无法反映数据的真实性了。下面我们将通过进一步研究,找出区分并揭示观测值的报告错误和年龄起伏的大小的方法。

考虑观测值 a_x 的构成,在封闭状态下,假定报告错误只存在年龄误报,无漏报和重报发生,可以认为观测值和模型的残差 Δ_x 由两部分构成:一部分为实际数据(真值)与模型值的差,表示实际数据自身的大小变化,记为 $\Delta_x^{(1)}$;另一部分为测量错误(或报告错误),记为 $\Delta_x^{(2)}$。a_x 和 $a(x)$ 有如下关系:

$$a_x=a(x)+\Delta_x=a(x)+\Delta_x^{(1)}+\Delta_x^{(2)} \quad x=0,1,2,\cdots,n \tag{14}$$

由于 $A_x=\sum_{j=x}^{\infty}a_j=A_x+\sum_{j=0}^{\infty}\Delta_j=A_x+\sum_{j=0}^{\infty}a_j\left(\frac{\Delta_j}{a(j)}\right)$，$\Delta_x/a(x)$ 或正或负，其绝对值 $|\Delta_x/a(x)|<1$，$\sum a(j)\left(\frac{\Delta_j}{a(j)}\right)<\sum a(j)=A(x)$。实际上，由于 $\sum a(j)(\Delta_j/a(j))$ 中正负相抵，$\sum a(j)\left(\frac{\Delta_j}{a(j)}\right)$ 要比 $A(x)$ 小很多，这就解释了为什么在一些年龄百分比中残差较大，而在累计的年龄百分比中残差很小的原因。

残差 $\Delta_x^{(1)}$、$\Delta_x^{(2)}$ 的性质是完全不同的。$\Delta_x^{(1)}$ 是真值的一部分（尽管真值并不知道），为 $a(x)$ 对年龄百分比真值的修正，$\Delta_x^{(2)}$ 为报告错误，即观测值 a_x 与真值的差。由于人口总量不变，则有 $\sum\Delta_x^{(2)}=0$。只有区分和分别确定 $\Delta_x^{(1)}$、$\Delta_x^{(2)}$ 的大小，才能正确判定人口普查数据报告的质量和确定各年龄百分比的实际值。但如果只有一次普查的数据，我们只是知道 $\Delta_x^{(1)}$ 和 $\Delta_x^{(2)}$ 之和，还无法区分和确定 $\Delta_x^{(1)}$、$\Delta_x^{(2)}$。

$a_x-\Delta_x^{(2)}$ 为真值，它的变化不总是单调的，可能有高低起伏，而模型值 $a(x)$ 由于它是从模型 $A(x)$ 推导出的一个用简单函数来表示年龄别人口比例随 x 均匀变化的数值，它或大于 a_x 或小于 a_x。可以认为是 $a_x-\Delta_x^{(2)}$ 的估计值，由公式（14）整理可得：

$$a_x-\Delta_x^{(2)}=a(x)+\Delta_x^{(1)}=a(x)\cdot(1+h_x) \quad x=0,1,2,\cdots,n \tag{15}$$

定义 $h_x=\frac{\Delta_x^{(1)}}{a(x)}$ 为偏离度或偏离系数，它反映估计值与真值的偏离程度。

二、两次普查的人口年龄结构关系模型

由上面的讨论我们知道，一个人口的累计百分比的两次对数，可以转化为以线性函数的形式来表示。下面我们并不讨论一般意义上的两个人口年龄结构的关系，而是讨论其中之一：同一地域的人口在不同时点的年龄结构的关系，或者说同一地域在两次普查时的人口年龄结构的关系。

在讨论前，先对模型作一些补充说明：在上面讨论的模型中，作为自变量的年龄是从 0 岁开始的，对象人口是全部人口，但模型也适用于部分人口，函数的起始年龄并不限于 0，也可以用其他年龄。例如，以 x_0 作为起始年龄，则对象人口即为 x_0 岁及以上人口。这时公式（3）的函数形式为：

$$\ln\left(\frac{1}{A(x)}\right)=A(x-x_0)^b\exp(C(x-x_0)) \quad x=x_0,x_0+1,x_0+2,\cdots,x_0+n \tag{16}$$

通过变量代换，令 $x-x_0\to x$，这样上式就变为：

$$\ln\left(\frac{1}{A(x+x_0)}\right)=Ax^b\exp(C(x)) \quad x=0,1,2,\cdots,n \tag{17}$$

即，不管以哪个年龄作为起始点，通过变量代换，函数表达式可以和以 0 作为起始年龄的表达式有相同的形式。

公式（17）中的 $C(x)$ 为 k 次多项式，k 取值大小可根据自变量 x 的区间长度和观测数据变动的复杂程度来规定。

设两次普查的时点相同，例如都在年中，两次普查的时间间隔长度为 T 年，则上一次普查 x 岁人口，在下一次普查时对应的人口为 $x+T$ 岁。前次普查的年龄区间为 $(0,\infty)$，后次普查对应的年龄区间为 (T,∞)，记前次普查 x 岁累计人口百分比的观测值为 $A_x(0)$，后次普查对应的累计人口的百分比的观测值为 $A_{x+T}(T)$，对应的模型分别为 $A(x,0)$ 和 $A(x+T,T)$，由公式（3）和（17）可知，$\ln(1/A(x,0))$ 和 $\ln(1/A(x+T,T))$ 分别可有如下的形式：

$$\ln\left(\frac{1}{A(x,0)}\right)=A_0x^{b_0}\exp(C_0(x)) \tag{18}$$

$$\ln\left(\frac{1}{A(x+T,T)}\right)=A_Tx^{b_T}\exp(C_T(x)) \quad x=0,1,2,\cdots,n \tag{19}$$

由此，两次普查累计人口的百分比对数之比为：

$$R(x)\approx\frac{\ln\left(\frac{1}{A(x+T,T)}\right)}{\ln\left(\frac{1}{A(x,0)}\right)}=A\cdot x^b\cdot\exp(C(x)) \quad x=1,2,\cdots,n \tag{20}$$

其中，$A=A_T/A_0$，$b=b_T-b_0$，$C(x)=C_T(x)-C_0(x)$ 为 k 次多项式。

这时，观测值 $\ln(A_{x+T}(T))/\ln(A_x(0))$ 和模型有如下关系：

$$\frac{\ln(A_{x+T}(T))}{\ln(A_x(0))}=\frac{\ln(A(x+T,T))(1+D_x(T))}{\ln(A(x,0))(1+D_x(0))} \tag{21}$$

所以，$\ln(A_{x+T}(T))/\ln(A_x(0))=\ln(1/A(x+T,T))/\ln(1/A(x,0))\cdot(1+D_x)=A\cdot x^b\cdot\exp(C(x))\cdot$

$(1+D_x)\ x=1,2,\cdots,n$ (22)

其中,$1+D_x\approx(1+D_x(T))/(1+D_x(0))$。

对公式(22)两边取对数,可得到类似公式(4)的形式:

$$\ln\left(\frac{\ln(A_{x+T}(T))}{\ln(A_x(0))}\right)=\ln A+b\ln x+C(x)+D_x \quad (23)$$

D_x 为 $\ln(\ln(A_{x+T}(T))/\ln(A_{x+T}(T)))$ 与 $\ln A+b\ln x+C(x)$ 的差,在使 $\sum D_x^2$ 达到最小的条件下估计出参数 A、b 及 k 次多项式 $C(x)$ 的系数 $c_j(j=1,2,\cdots,k)$,由估计出的参数,就可算出模型值 $R(x)$。

分别算出 $A_x(0)$ 和 $A_{x+T}(T)$ 的模型值 $A(x,0)$ 和 $A(x+T,T)$,把模型值之比作为观测值之比,具体为:

$$R_1(x)=A_{x+T}(T)/A_x(0)\approx A(x+T,T)/A(x,0) \quad (24)$$

$A(x,0)$ 和 $A(x+T,T)$ 的导数之比为:

$$\frac{A'(x+T,T)}{A'(x,0)}=A(x+T,T)\cdot\ln\left(\frac{1}{A(x+T,T)}\right)\cdot\frac{\frac{b_T}{x}+C'_T(x)}{A(x,0)}/\ln\left(\frac{1}{A(x,0)}\right)\cdot\left(\frac{x}{b_0}+C'_0(x)\right) \quad (25)$$

由上式可知,两次普查对应年龄累计百分比的导数之比可分解为以下三个部分组成:①累计百分比对数之比,即(20)式中的 $R(x)$;②累计百分比之比,可由公式(24)来决定;③当分子和分母的累计百分比模型参数分别算出后,第三部分为:

$$R_2(x)=\frac{\frac{b_T}{x}+C'_T(x)}{\frac{b_0}{x}+C'_0(x)} \quad (26)$$

其中,$C'_T(x)$ 和 $C'_0(x)$ 的 k 次多项式的导数,由导数之比可算出 $R_2(x)$。但我们也可以用其他方法估计,由于分子多项式和分母的多项式都已确定,则 $R_2(x)$ 也可算出。也可以用下面的方法,即存在这样的多项式:

$$\frac{b}{x}+C(x)=\frac{\frac{b_1}{x}+C'_T(x)}{\frac{b_0}{x}+C'_0(x)} \quad (27)$$

其系数 b、C_0、C_1、$\cdots$、C_k 可由待定系数法确定。

最后,再来讨论两次普查人口年龄百分比之比的观测值 $a_{x+T}(T)/a_x(0)$ 和模型值 $a(x+T,T)/a(x,0)$ 的关系。

$$\frac{a_{x+T}(T)}{a_x(0)}=\frac{A_{x+T}(T)-A_{x+T+1}(T)}{A_x(0)-A_{x+1}(0)}\approx\frac{A(x+T,T)-A(x+T+1,T)}{A(x,0)-A(x+1,0)} \quad (28)$$

对等式右边来说,由柯西中值定理(数学上或称第二中值定理)可知,则存在正数 β,$0\leqslant\beta\leqslant1$,则有:

$$\frac{A(x+T,T)-A(x+T+1,T)}{A(x,0)-A(x+1,0)}=\frac{A'(x+T+\beta,T)}{A'(x+\beta,T)} \quad (29)$$

在公式(29)中,β 是区间(0,1)中的一个数,对不同的$(x,x+1)$区间内,β 值会有所不同,从应用的角度来看,简单地取 $\beta=0.5$,并把它作为百分比的模型值。这样,结合公式(25)和公式(26),$a_{x+T}(T)/a_x(0)$ 的模型值为:

$$\frac{a(x+T,T)}{a(x,0)}=\frac{1-A(x+T+1,T)}{1-A(1,0)}\quad x=0 \quad (30)$$

$$\frac{a(x+T,T)}{a(x,0)}=R(x+0.5)\cdot R_1(x+0.5)\cdot R_2(x+0.5)\ x>0 \quad (31)$$

也可以用这样的方法来估计 $R_2(x+0.5)$ 的值:由观测值 $a_{x+T}(T)$ 和 $a_x(0)$ 和已计算出的 $R(x+0.5)$、$R_1(x+0.5)$ 的值,先估计出(28)式右边的参数 b、C_0、C_1、$\cdots$、C_k。

设 $y=x+0.5$,$F_x=(a_{x+T}(T)/a_x(0))/(R(x)/R_1(x))x=1,2,\cdots,n$,在 F_x 与 $R(x+0.5)$ 的差的最小平方的条件下,

$$\min\Delta(b,C_0,C_1,\cdots,C_k)=\min\ \sum(F_x-B/y+C_0+C_1y+C_2y^2+\cdots+C_ky^k)^2 \quad (32)$$

由此进行参数估计,进而可以计算出 $R_2(x+0.5)$ 的值。

下面,我们来讨论年龄百分比的观测值 $a_{x+t}(T)/a_x(0)$ 和模型值 $a(x+T,T)/a(x,0)$ 的关系。

设前次普查 x 岁人口为 $p_x(0)$,后次普查对应的人口为 $p_x(T)$,假定两次普查的人口报告数准确,$P(0)$为前一次普查时的总人口,$P(T)$为后一次普查时 T 岁及以上的人口。$R_0=P(T)/P(0)$,$SR_x=p_{x+T}(T)/p_x(0)$,R_0 为对应的总人口之比,SR_x 表示上次普查 x 岁的人口到下一次普查 $x+T$ 岁的存活率。则以下关系成立:

$$p_x(0)=P(0)\cdot a(0) \quad (33)$$

$$p_{x+T}(T)=P(T)\cdot a_{x+T}(T) \quad (34)$$

$$SR_x=\frac{P_{x+T}(T)}{P_x(0)}=\frac{P(T)\cdot a_{x+T}(T)}{P(0)\cdot a_x(0)}=\frac{P(T)}{P(0)}\cdot\frac{a_{x+T}(T)}{a_x(0)}$$

$$= R_0 \cdot \frac{a_{x+T}(T)}{a_x(0)} \quad (35)$$

可以推出：

$$\frac{SR_x}{R_0} = \frac{a_{x+T}(T)}{a_x(0)} \quad (36)$$

另一方面，若普查报告无误，由公式（15）可知，$a_x(0)$、$a_{x+T}(T)$和对应的模型值 $a(x,0)$、$a(x+T,T)$有如下的关系：

$$a_x(0) = a(x,0) \cdot (1+h_x(0)) \quad x=0,1,2,\cdots,n \quad (37)$$

$$a_{x+T}(T) = a(x+T,T) \cdot (1+h_{x+T}(T)) \quad x=0,1,2,\cdots,n \quad (38)$$

所以，

$$\frac{a_{x+T}(T)}{a_x(0)} = \frac{a(x+T,T)}{a(x,0)} \cdot \frac{1+h_{x+T}(T)}{1+h_x(0)} \quad (39)$$

由公式（31）可知，$a(x+T,T)/a(x,0)$是 $a_{x+T}(T)/a_x(0)$的模型估计：

$$a_{x+T}(T)/a_x(0) = a(x+T,T)/a(x,0) \quad (40)$$

比较公式（36）和公式（39）可以得到这样的结论：当观测存活率用模型存活率估计时，前一次普查 x 岁百分比的偏离度 $h_x(0)$等于后一次普查 $x+T$ 岁的偏离度 $h_{x+T}(T)$。反过来也可以认为，当两次普查对应年龄的偏离度相等，或放宽条件，两者的值接近时，可以用两次普查 x 岁的模型存活率来估计观测存活率。

这里我们得到了一个很有意思的结果：当人口百分比在相邻年龄有较大的变化时，用百分比的模型值来估计实际值会有较大的误差。但两次普查对应年龄百分比的比的模型值就能准确地估计实际值。另外，还有一个重要的结果：在封闭的条件下，某个年龄百分比的模型值，随时间推移会发生改变，例如从 0 到 T 年，模型值会从 $a(x,0)$改变到 $a(x+T,T)$，但它的偏离度并不会随时间的推移发生改变。

三、1982 年全国人口普查年龄报告误差

1982 年第三次全国人口普查，我国开始按照现代人口普查的内容设计，采用了先进的技术手段，在当时的社会环境下，基层组织对人口管理有效，人口流动的规模很小。在全国动员、上下重视和努力下，人口普查非常成功，获得的调查数据质量很高，这一结论得到国内外学术界普遍的认可[④]。但普查数据可靠与否，需要通过和其他数据是否一致，包括后来的普查数据是否一致来检验。在 1982 年普查以后，我国在 1987 年举行了全国 1% 人口抽样调查，在 1990 年又举行了第四次全国人口普查。通过对比，人们发现在死亡数据方面，1982 年数据存在着漏报，在年龄人口方面，存在着漏报和误报的问题。这里不讨论死亡数据漏报问题，主要讨论年龄误报问题。

下面，我们来观察 1982 年和 1990 年前后两次普查对应年龄的人口比：设 1982 年 x 岁的人口为 $p_x(0)$，它和 1990 年 x+8 岁的人口相对应，设为 $p_{x+T}(T)$（T=8），对应的人口比为 $p_{x+T}(T)/p_x(0)$，如果普查报告的人数无误，这个比表示上次普查 x 岁的人经过 T 年后存活的比率，简称存活率。1982 年各年龄的人与 1990 年对应人口之比和存活率[⑤]见图 5。

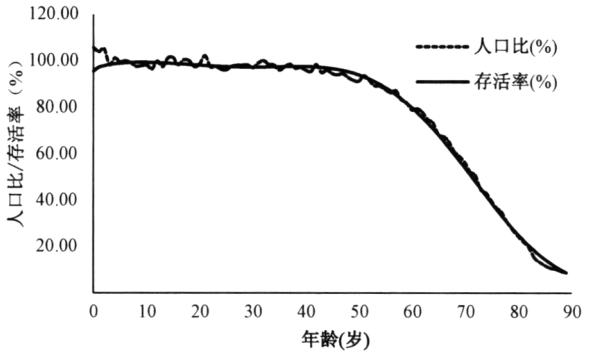

图 5　1982–1990 年对应年龄的人口比和存活率

资料来源：由中国 1982 年人口普查资料、中国 1990 年人口普查资料计算得出。

观察图 5 可以知道，两次普查对应年龄的人口比并不是一条光滑曲线，它是在存活率曲线上下波动的，这在 40 岁以前特别明显（见图 6）。

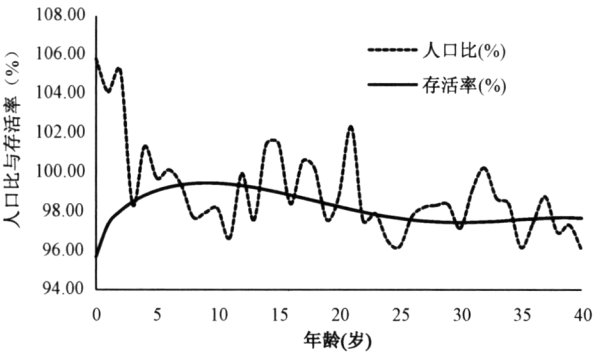

图 6　0–40 岁年龄人口比和存活率（1982–1990 年）

资料来源：由中国 1982 年人口普查资料、中国 1990 年人口普查资料计算得出。

理论上说，除出生人口外，其他年龄的人口随时间的变化完全由死亡率决定。死亡率的变化一般是

很稳定的。由于中国人口基数很大,除了高龄人口外,各个年龄人口数都很大,例如在1982年,50岁以下每个年龄人口都在800万以上,人口死亡率(或者存活率)的随机误差非常小⑥,出现上述波动,可以认为是由于人口误差引起的。从这里似乎可以得出结论:普查的人口数据其实也不是很准确的,是有一定差错的。从1982年0-40岁的年龄看(见图6),这些差错是非常明显的。

人口报告的错误最明显地是出现在两次普查的人口比大于1的年龄上。在人口封闭的情况下,人口随时间(或者说年龄)推移发生变化,由于死亡的发生,只会变少,存活率肯定是小于1。如果在某个年龄x岁人口比大于1的情况出现,可能是以下情况导致的:上次普查x岁的人口有漏报,或是后一次普查$x+T$岁的人口有重报;当然也可能是人口年龄误报导致,即上一次普查x岁的人报告到其他年龄上去,导致报告人数少于实际人口,或后一次普查非$x+T$岁的人在普查时报告到了$x+T$岁,导致在$x+T$岁报告人数大于实际人口。

在1982-1990年两次普查对应年龄的人口比中,有11个年龄大于1。其中,在0-6岁有5个年龄,在14-32岁,有6个年龄的人口比大于1。由于1982年0-6岁人口少于1990年的8-14岁人口,很可能是在1982年普查时,一部分家庭未按计划生育的规定"超生",因为担心受罚而瞒报了人口。到了1990年,这些被瞒报的人口都已进入上学年龄,在这以前有些家庭已经作出应对,如已经缴过了罚款,或采取了其他办法,已不必再隐瞒或不再隐瞒,所以出现了1990年8-14岁的人口多于1982年0-6岁人口。这里,我们以1990年8-14岁人口为基础并考虑了死亡的影响对1982年的0-6岁人口进行了调整。在其他年龄,则认为只是由于年龄报告的错误造成的。以下,我们来估计年龄报告的误差。

设1982年x岁的年龄百分比的观测值为$a_x(0)$,其模型值为$a(x,0)$,对应的1990年$x+T$岁的年龄百分比观测值为$a_{x+T}(T)$,模型值为$a(x+T,T)$。按照公式(14)可知:

$$a_x(0)=a(x,0)+\Delta_x(0)=a(x,0)+\Delta_x^{(1)}(0)+\Delta_x^{(2)}(0)\quad x=0,1,2,\cdots,n \tag{41}$$

$$a_{x+T}(T)=a(x+T,T)+\Delta_{x+T}(T)=a(x+T,T)+\Delta_{x+T}^{(1)}(T)+\Delta_{x+T}^{(2)}(T)\quad x=0,1,2,\cdots,n \tag{42}$$

其中,$\Delta_x(0)$、$\Delta_{x+T}(T)$为观测值与模型值的残差。$\Delta_x^{(1)}(0)$、$\Delta_{x+T}^{(1)}(T)$表示x岁观测值与模型值的偏差,$\Delta_x^{(2)}(0)$、$\Delta_{x+T}^{(2)}(T)$为报告数据的误差。$\Delta_x(0)$、$\Delta_{x+T}(T)$可以通过计算来确定,而$\Delta_x^{(1)}(0)$、$\Delta_{x+T}^{(1)}$、$\Delta_x^{(2)}(0)$和$\Delta_{x+T}^{(2)}(T)$为未知。

由公式(41)和公式(42)可知:

$$a_x(0)-\Delta_x^{(2)}(0)=a(x,0)\left(1+\frac{\Delta_x^{(1)}(0)}{a(x,0)}\right) \tag{43}$$

$$a_{x+T}(T)-\Delta_{x+T}^{(2)}(T)=a(x+T,T)+\Delta_{x+T}^{(1)}(T)=\frac{a(x+T,T)(1+\Delta_{x+T}^{(1)}(T))}{a(x+T,T)} \tag{44}$$

根据前面的论述,我们将偏离度或偏离系数h_x定义为$\Delta_x^{(1)}/a(x)$,当$\Delta_x^{(1)}(0)/a(x,0)=\Delta_{x+T}^{(1)}(T)/a(x+T,T)$时,$a(x+T,T)/a(x,0)=(a_{x+T}(T)-\Delta_{x+T}^{(2)}(T))/(a_x(0)-\Delta_x^{(2)}(0))$

$$h_x=\Delta_x^{(1)}(0)/a(x,0)=\Delta_{x+T}^{(1)}(T)/a(x+T,T) \tag{45}$$

下面,我们来估计h_x。在公式(43)中,先假定$\Delta_x^{(2)}(0)=0$,即人口报告无误,这样模型偏差等于模型残差。$\Delta_x^{(1)}(0)=\Delta_x(0)$,可算出偏离系数:

$$h_x(0)=a_x(0)/a(x,0)-1 \tag{46}$$

同样,由公式(44)得出:

$$h_x(T)=\frac{a_{x+T}(T)}{a(x+T,T)}-1 \tag{47}$$

算出的$h_x(0)$和$h_x(T)$见图7。

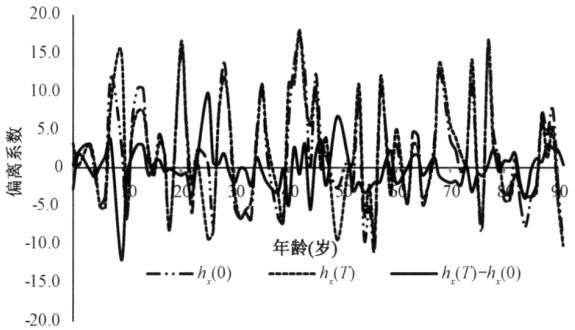

图7 偏离系数$h_x(0)$和$h_x(T)$的对比

由图7可以看出,尽管$h_x(0)$和$h_x(T)$的起伏很大,但在80岁以前,两者却很接近,用$h_x(0)$和$h_x(T)$的平均来估计偏离度h_x。

$$h_x=0.5(h_x(0)+h_x(T)) \tag{48}$$

把h_x代回到公式(43)和公式(44)中,可得1982年和1990年各年龄百分比的报告误差和实际值。

根据1982年和1990年全国人口普查数据,1982年0-6岁的人口中有5个年龄少于对应的1990年

的 *8-14* 岁的人口。据此,本文以 1990 年 8-14 岁的人口为基础,按照 1982-1990 年的平均死亡率,反推估计 1982 年 0-6 岁人口,并把它作为 1982 年的实际人口,将其和 1982 年 0-6 岁的调查人口之差作为人口漏报数,则 1982 年该年龄组漏报人口为 527 万人。

在理论上,年龄上限可以无限大,这样两次普查在年龄上可以一一对应。但在实际的统计中,总是会设定一个年龄上限。公布的 1982 年普查和 1990 年普查的年龄上限是 100 岁,该年龄是半开区间,即包括了 100 及以上的人口,所以能和 1990 年一一对应的是 1982 年的 91 岁以下的年龄。以下的数据都是指 1982 年 0-91 岁年龄的数据。

由公式(43)可知,$\Delta_x^{(2)}(0)=a_x(0)-a(x,0)(1+h_x)$,可估计算出各个年龄百分比的报告误差。

在 7-91 岁的年龄中,年龄误报人口有 683 万人,年龄误报率为 6.74‰。由于年龄误报,一些年龄的报告人口比估计的实际人口多,它们主要出现在中青年期,即青年期(24 岁和 25 岁)和中年期,共计 340 万人,而一些年龄的报告人口少于估计的实际人口,它们主要分布在青年期(21 岁和 17、18 岁),共计 342 万人。而多报人数比率较高的主要是高龄(80 岁以上),而少报人数比率较高的除了 21 岁,大多在 60-79 岁年龄段⑦。按 1982 年人口普查事后质量抽样调查结果,年龄误报率为 6.15‰[3]。相比于本文的计算结果,两者还是很接近的。

四、小结和讨论

人口年龄结构是人口研究中最重要的基础数据。人口年龄结构数据的准确性,直接影响到各项人口研究的质量。人口结构形形色色,不同地域在不同时点的人口年龄结构都不同。本文提出了人口年龄结构模型——以年龄为自变量、累计的年龄百分比为自变量的函数形式,并以中国历次人口普查的数据和其他一些数据进行了验证,表明模型是成立的。这个结果是很有意义的。因为根据模型,一个人口的累计百分比的两次对数,通过变量替换后,可以表示为线性函数的形式。由于线性函数的可传递性,任何两个人口的累计百分比的两次对数都可以用线性函数来联系。利用这个结果,可以方便地对不同的人口年龄结构进行区分和归类,也可以利用已知的人口年龄结构来推测未知的或信息缺失的另一个人口的年龄结构。

由人口的年龄累计百分比模型可以推导出年龄百分比模型。年龄百分比模型值也可称为人口年龄变化的估计值,而百分比模型值与观测值差的分布可以作为人口报告是否有特定年龄尾数堆积的判定方法。

本文分析表明,许多人口的年龄分布不是单调均匀变化的,所以,要用一个简单函数来准确地表示出各种人口的年龄百分比几乎是不可能的。特别是如我国的人口在不同年龄的比重起伏很大时,模型值和观测值出现较大的残差是必然的。但可以把年龄百分比模型值作为实际人口的年龄百分比的估计值。本文提出了这样的概念:每个年龄的人口通常可分成估计部分和偏离部分。这两部分的特点是,在人口封闭的条件下,随着时间的推移,各年龄的人口占总人口的比重发生改变,该年龄的估计部分的比重也随之变动,但它的偏离部分与估计部分的比(本文名之为偏离度)是几乎不变的。由此,就可以用两次普查对应年龄人口比的预估值来估计两次普查对应年龄人口的存活率,并估计出人口普查的年龄报告误差。即把实际的年龄人口分成估计部分和偏离部分概念的引入,是估计人口普查报告误差的关键。

但年龄人口的偏离度并不是调查数据直接给出的,它是由现今的人口年龄结构数据或者历史的出生人口数据得到的。它的准确性取决于原始数据的准确性。从表面上看,偏离系数和数据精度的估计相互依赖,成了解不开的结。这里需要注意的是,虽然年龄偏离系数是由调查数据估计出来的,但可利用的调查数据往往不是只有一个。我们可以利用数据质量较高的调查,或者比较分析不同的调查结果,从而估计出较为可靠的年龄偏离系数。

本文 1982 年普查在高龄部分(80 岁以上)的报告误差较大,估计出的年龄偏离系数误差也会比较大。但由于高龄人口数较少,对总的误报数影响较小。这里就不加讨论了。

注释:

①在本文中,x 岁年龄别人口比例用小写字母表示,如 p_x,x 岁及以上的人口用大写字母字母表示,如 P_x。字母加下标表

示观测值,如 a_x,表示 x 岁人口占总人口的百分比观测值;$a(x)$表示 x 岁人口占总人口的百分比模型值。当要表示某个时刻时,观测值用下标加括号,如 $a_x(T)$来表示,模型值则用二元函数的形式 $a(x,T)$来表示。

②观测数据和模型值的对比限于篇幅省略,备索。

③在 18 岁附近出现较大残差,很可能是报告错误引起的。

④1984 年 3 月在北京召开的"中国 1982 年人口普查北京国际讨论会"上,出席会议的人口学家对这次人口普查的质量一致地给予了很高的评价。

⑤这里暂时把根据模型推算的普查人口之比看作存活率,后面将证明这一推算是成立的。

⑥死亡率标准误差的计算可参见:蒋庆琅. 寿命表及其应用[M]. 上海:上海翻译出版公司,1984:49-50。

⑦各年龄段误报人数和误报人数比率详细数据可向作者索取。

参考文献:

[1]黄荣清. 中国人口普查中人口年龄报告准确性的检验[J]. 人口研究,2009(6):30-41.

[2]黄荣清,肖周燕. 人口年龄结构数据异常的检验[J]. 人口与经济,2009(5):1-8,15.

[3]马安. 对中国 1982 年人口普查资料质量的评估[C]//李成瑞. 中国 1982 年人口普查北京国际讨论会论文集,1984.

Population Age Structure Model and Its Application

Huang Rongqing

Abstract: In this paper, a model of population age structure is proposed, which is a function taking age as the independent variable and the cumulative age percentage as the dependent variable. The model is verified by the data of Chinese population censuses and other data. This model can be expressed as a linear function after two logarithmic transformations. On the basis of this model, this paper further constructs the expressions of mathematical functions such as the percentage of population and the ratio of the percentage of age population corresponding to the last two censuses. After the test of census data, the model can fit the cumulative age-population percentage curve well, but when the age of the population fluctuates greatly, the residual error of the percentage model will become larger. It can be concluded that it is impossible to accurately represent the general age population percentage with a simple mathematical function, and it is impossible to fully and accurately judge the accuracy of the census data from a single census without other data support. In order to solve the reporting error of the census data, the census data is divided into three parts: the estimated value, the deviation value and false value. The estimated value here is the model value of the age percentage. This study shows that the age deviation coefficient (the ratio of the deviation value to the estimated value) is a constant under closed population conditions. Using this property, we can use the percentage model values of the two censuses to calculate the age survival rate of the actual population, and estimate the census false values by estimating the age deviation coefficient. Finally, the model is used to estimate the age misstatement of the 1982 population census in China. In 1982, it was estimated that 6.83 million people aged 7-91 years were misstated, and the age misstatement rate was 6.74‰. Due to age misstatement, the reported population of some ages is larger than the estimated actual population, and they occur mainly in adolescence (young adulthood (24 and 25 years old) and middle age), totaling 3.4 million, while the reported population of some ages is smaller than the estimated actual population, and they are mainly distributed in young adulthood (17, 18 and 21 years old), totaling 3.42 million.

Key words: age structure model; population census; census data revision