

中国人口动态参数的识别

蒋正华

中国人口数据的来源及可靠性

中国人口数据登记历史悠久。早在四千多年前，就已经有了全国人口数字的记载。历代皇朝曾建立了各种户口登记制度，留下了许多宝贵史料。特别是明朝的户帖，登记内容相当详细，实已开了近代人口普查的先河。总的说来，古代人口登记只重在劳力及总人口数字，太平盛世，这些数据比较接近实际，战祸绵延则漏报情况严重。即使在数据质量较高的时期，由于当时统治者进行户口清查的目的在于征收赋税、分派徭役，百姓避之唯恐不及，脱漏在所难免；有些地方官也故意少报人口，以减轻地方负担或增加中饱私囊的份额。根据历代户均人口、消费粮食等资料的对比，并参照史书或当时人的评论，可以估计古代人口总数登记在比较准确的时期漏报率也在10%以上；特殊情况，如汉末三国时期，大量户口为豪强霸占成为家兵、荫户，全国人口漏登率更可能达到50%以上。

清乾隆以后，当时的政府针对户口报告不实的情况采取了许多严厉的措施，使人口总数统计大大接近实际。国民党统治时期曾对13个省进行过直接调查，并根据其他资料估计全国总人口数；据报告，1948年为465 237 773人，这项数字的可靠性没有充分的证据予以评价，若与此后可靠的报告资料比较，漏报率约为12%强，数据质量较以前略有改善，但仍缺乏全面的出生、死亡数字，更不用说详细的人口结构资料了。

解放以后，健全了人口统计制度，并于1953年进行了中国第一次全国性的现代化人口普查，1964年、1982年又相继进行了第二次和第三次全国人口普查。将每年的登记人口数字与普查结果相比较，人口总数报告的一致性很好。以1952年及1953年年末人口登记数的平均值与1953年中人口普查数比较，仅相差2.82%；以1963年末和1964年末人口登记数的平均值与1964年中人口普查数比较，相差仅0.54%；1981年与1982年末人口登记数的平均值与1982年中普查数的差额更减少到0.1%。普查数略高于经常性登记数字主要是由于在普查中调查人员逐户逐人访问，对人口总数的确定更加准确。根据1982年普查后的质量抽查结果，第三次人口普查的人口重报率为0.71%，漏报率为0.56%，大大低于许多统计工作比较健全国家1~2%的误差率。第一次人口普查项目较少，集中在年龄、性别、民族等最基本特征上，在当时缺乏经验的情况下，调查项目的集中保证了这些数据的准确性，除了西藏等极少数地区外，占总人口96%以上的人口均经过直接访问，并有当地群众的协助，获得被访问者的充分合作。普查数据经检验，联合国综合指数为20.19，玛叶指数为1.16，韦伯指数为102.5，说明1953年普查的年龄、性别报告数据是可信的。1964年举行的第二次全国人口普查在前一次

的基础上增加了三项调查项目,成为九项,但仍然没有出生和死亡的数据。由于有了前一次的经验,在年龄、性别等项目的报告完全性、准确性方面均有更高的质量。利用这一次人口普查的数据与1953年人口年龄、性别结构比较可以看出,这两次普查绝大部分年龄区间都十分吻合,只有在1953年普查时14岁到17岁的年龄显然少报,而18岁的人口显然多报,其原因是不难解释的,因为1953年的普选规定,凡年满18岁者具有公民权,可以参加选举,这就造成14到17岁的人口倾向于高报为18岁。在农村地区,由于过去就有报虚岁的习惯,这一现象的发生更为普遍。1964年年龄性别数据经检验联合国综合指数为19.72,玛叶指数为0.41,韦伯指数为101.9。1982年人口普查数据的高质量现已为国际人口学、统计学界所普遍公认,无需再作讨论。但由于60年代到70年代出生率的波动,人口年龄结构受到干扰,死亡率的波动相对来说对年龄结构影响较小。这是由于死亡率的变化对各年龄人口均发生作用,而出生率却只作用于一个年龄的人口。这样,对1982年人口年龄、性别数据作同样的计算时,联合国综合指数为28.46,玛叶指数为3.00,韦伯指数为102.0,大于上一次人口普查结果。为了消除人口年龄结构受出生率波动干扰的影响,取出生率的相对值为权,对人口数据作加权处理,可以计算得到修正后的联合国综合指数,如表1所示:

表1 修正后的联合国综合指数

普查年份	1953年	1964年	1982年
修正联合国综合指数	—	17.96	19.04

经过修正后,1982年联合国综合指数与1964年结果相近。可见,原来的联合国综合指数计算并未真正反映年龄、性别报告的质量,而反映了强烈扰动的后果。总的说来,这三次全国人口普查的质量是高的,特别是人口年龄、性别的报告是完全可以信赖的。从全国范围来说,三次普查的完全性也十分接近。

出生率和死亡率只是从1954年起才有正式的记录,在此之前的数据均是估计值。历次人口普查中,则只有1982年才列入了出生和死亡的调查项目。与人口年龄报告准确性相比,出生与死亡的报告误差较大,其中尤以出生报告误差较为显著。出生、死亡的登记时间序列数据波动很大,其中既有真实的变化,也包含了各个不同时期漏报率不同的干扰。总的说来,死亡率的漏报除在1959到1962年的4年期间较为严重外,准确性还是相当高的。出生登记的误差则从70年代起比较明显,漏报率逐渐上升,1976年后由于普遍提倡只生一个孩子,使出生漏报率有较大的跃升,特别是三胎以上的出生多有出生时瞒报,经过一年或二年后报为迁入人口的现象。这种出生的漏报和延迟进入户口登记对总人口数的影响虽然不大,但严重地影响了出生率和迁移率的登记数字,使国内省际迁移登记数据失实,出生率低估。这些现象均可在本文的计算过程中得到证实。将1981年人口普查中所获得的出生、死亡数据与登记数据相比较,可以看出出生漏报同样也对婴儿死亡率有严重影响,出生不久即死亡的婴儿可能在出生登记和婴儿死亡登记两方面同时被遗漏掉,这使当年婴儿死亡率的登记数比普查数约低五分之一到四分之一,但从全体人口的粗死亡率来看,影响不大,因此可以认为,除了婴儿死亡的登记外,其余年龄死亡漏报很少。上海等地曾采用了双重登记制度来检验日常登记制度中婴儿死亡率的漏报现象,获得了与1982年普查类似的结论,可见1982年普查所取得的

按年龄别死亡数据可靠性较高,这一点在普查后的复查中得到证实,也可以从普查中采取的种种严密措施得到间接的说明。

中国的国际迁移很少,即使在迁移量最大的一些时期,估计净迁移率也不超过万分之二或三,许多集体迁移的人群中年龄分布也比较分散,对全国性人口、年龄构成的影响不大,可以忽略。

各种抽样调查也从不同角度提供了有关人口动态参数的资料,但它们与登记数据相比具有不同的覆盖范围,因此不具备可比性,只有利用来作为各种局部地区数据的对比,籍以对全国性的资料提供一类间接性的评价信息,在评价时也必需考虑到抽样误差的影响。

根据中国目前人口数据的情况,从分析全国出生率、死亡率的角度,可以将数据分成三类:

第一类:普查人口年龄、性别构成,公安部门历年登记总人口数,准确性最高,误差可以忽略。

第二类:普查出生、死亡报告,误差很小,从全国范围而言是可靠的,只需作很小的调整即可使用。

第三类:户籍管理部门登记的出生和死亡数,有一定的误差,在各个时期误差程度不等,但其趋向是正确的,在一般情况下完全可以使用。

为了校正中国人口的出生率和死亡率,一个合乎逻辑的结论是尽可能利用第一类数据,适当地加入从第二类数据中获得的可靠信息来校正第三类数据。本文作者正是基于这样的认识,提出了适合中国人口数据特征的方法。

校正中国人口出生率、死亡率的方法

针对上述情况,利用1953年、1964年和1982年人口普查全国人口性别、年龄报告数据,1982年人口普查取得的上一年死亡资料可以辨识出这三次普查年之间各年人口年龄、性别结构,分性别的完全生命表,由此可以得到每年的死亡率与出生率。其方法如下,分为四步:

(一)求得1981年生命表:

从1982年人口普查资料可以得到1982年中按年龄和性别的人口数据以及1981年死亡人口的性别、年龄结构,分别以 $P_a^{1982, M}$, $P_a^{1982, F}$, $D_a^{1981, M}$, $D_a^{1981, F}$ 表示1982年中男、女性a岁人数和1981年男、女性a岁死亡人数,于是1981年生命表可由自修正迭代程序求得。迭代过程为,首先任意假设一个各年龄的留存率 SR_a^0 作为迭代的初始值,于是1981年中a岁人口数的零次迭代值可由

$$P_a^{1981, 0} = P_{a+1}^{1982} / SR_a^0 \dots\dots\dots (1)$$

求得,由此得到1981年a岁死亡率一次迭代值

$$m_a^1 = D_a^{1981} / P_a^{1981, 0} \dots\dots\dots (2)$$

根据这一组按年龄别死亡率即可建立起一个生命表,并从生命表参数求出留存率的一次迭代

值 SR_a^1 :

$$SR_a^1 = L_{a+1}^1 / L_a^1 \dots\dots\dots (3)$$

将 SR_a^1 代入(1)式以取代 SR_a^0 即可求得 $P_a^{1981,1}$, 依次类推可以建立起第二个生命表求得二次迭代值 SR_a^2 , 等等。这一迭代过程进行到第 n 次迭代得到的 $P_a^{1981,n}$ 与第 $n+1$ 次迭代所得到的 $P_a^{1981,n+1}$ 之差小于任意已知值时, 即

$$\left| P_a^{1981,n+1} - P_a^{1981,n} \right| \leq \varepsilon \dots\dots\dots (4)$$

其中 ε 为给定的一个很小的数字, 则迭代停止, 求得了1981年分年龄别的年中人口, 同时也得到了第 n 次迭代产生的生命表, 这一生命表即1981年生命表。以上过程可分别对男、女性别的人口求解得到分性别生命表和1981年人口。迭代解法已由作者证明是收敛的, 其解存在且唯一。

(二) 建立参数估计模型:

若已有两个普查年的人口年龄数据和其中一年的生命表(例如, 从上一节可得到1981年中人口年龄构成和生命表, 再加上1964年人口年龄构成即可作为参数估计的基本数据), 则可写出人口分年龄数 $P(a)$ 和 $P_{n_1}(a)$ 以及 n_1 年的留存率 $SR_{n_1}(a)$ 数列, 当迁移可以忽略时, 显然有

$$P_{n_1}(a+n_1) = P(a) \prod_{j=0}^{n_1-1} SR_j(a+j) \dots\dots\dots (5)$$

对两边取对数, 得

$$\ln \frac{P_{n_1}(a+n_1)}{P(a)} = \sum_{j=0}^{n_1-1} \ln SR_j(a+j) \dots\dots\dots (6)$$

令

$$\ln \frac{P_{n_1}(a+n_1)}{P(a)} = T(a)$$

$$\ln(SR_j(a+j)) = R_j(a+j)$$

则式(6)可写为

$$T(a) = \sum_{j=0}^{n_1-1} R_j(a+j) \dots\dots\dots (7)$$

附录中将证明, 在两次普查年间任何一年生命表函数 $SR(a)$ 的变换 $R(a)$ 可以表示为

$$\hat{R}_j(a) = \left(\sum_{i=0}^{n_2} C_{ji} \left(\frac{a}{100} \right)^i \right) R^*(a) \dots\dots\dots (8)$$

式中

$$R^*(a) = \ln(SR_{n_1}(a))$$

即是基准生命表参数 $SR(a)$ 的变换。式(8)中的 n_2 由使用者根据估计精度的要求来选择。最

高人口年龄在本文中取为100岁。现在,我们希望确定参数C,使各年生命表确定后,使后一次人口普查年龄人口倒推到前一次人口普查时刻的人口与普查统计数误差最小,即

$$\min \left| P_{n_1}(a+n_1) - P(a) \prod_{j=0}^{n_1-1} \widehat{SR}_j(a+j) \right| \dots\dots\dots (9)$$

式中 $\widehat{SR}_j(a+j)$ 为j年a+j岁人口的生命表留存率估计值,由于对绝对值取最小不易在数学上实现,(9)式改为另一种二次型的目标函数,即

$$\begin{aligned} F(C_{j1}) &= \sum_{a=0}^{100-n_1} \left(T(a) - \sum_{j=0}^{n_1-1} \widehat{R}_j(a+j) \right)^2 \\ &= \sum_{a=0}^{100-n_1} \left(T(a) - \sum_{j=0}^{n_1-1} \left(\sum_{L=0}^{n_1} C_{j1} \left(\frac{a+j}{100} \right)^L R^*(a+j) \right) \right)^2 \end{aligned} \dots\dots (10)$$

根据试算的结果,考虑前二阶矩已有足够的精度,故取目标函数为:

$$F(C_{j0}, C_{j1}) = \sum_{a=0}^{100-n_1} \left(T(a) - \sum_{j=0}^{n_1-1} \left(C_{j0} + C_{j1} \left(\frac{a+j}{100} \right) \right) R^*(a+j) \right)^2 \dots\dots\dots (11)$$

令(11)式为最小,即可解得两次普查年间各年的参数 C_{j0}, C_{j1} 并由此获得各年生命表和人口年龄构成。双参数估计模型于是可写为:

$$\begin{cases} \min_C F(C) = \min_C \left(\frac{1}{2} C^T (A^T A) C - I^T A^T C \right) \\ S.t. \cdot \left(C_{j0} + C_{j1} \left(\frac{a+j}{100} \right) \right) > 0 \end{cases} \dots\dots\dots (12)$$

式(12)中符号的表示与常用数学符号相同。

(三) 参数估计模型的求解:

式(12)所表示的是一个凸规划问题,约束条件保证了留存率将小于1。该规划问题可用迭代方法求解,迭代的初值由中国统计年鉴粗出生率、粗死亡率的记录给定,初值的变化并不影响最终的解,但由于实际计算机的精度,解的收敛值可能会有变化,因此我们选取登记值作为出发点。但不管怎样,不同解的收敛值相差都不大。迭代时每次沿F在 C_k 点的最速下降方向改变一个最优步长。由于式(12)的约束是开集,一般说来有可能无解,因此将该问题略作修改,变为:

$$\begin{cases} \min \left(\frac{1}{2} C^T (A^T A) C - I^T A^T C \right) \\ S.t. \cdot C_{j0} \geq 0, C_{j1} \geq 0 \end{cases} \dots\dots\dots (13)$$

这是一个典型的凸规划问题,最优解存在,且是唯一的。

(四) 解的修正:

由于人口普查选取的标准时间为7月1日, 所得到的人口年龄结构均相应于年中时刻, 直接按照(13)式解的结果得到 n 年下半年到 $n+1$ 年上半年的一年出生率和死亡率, 而不是按阳历年计的参数值, 因而与统计年鉴发表的数据缺乏可比性, 必须进行适当的修正。按照常用的人口学方法, 我们认为一年内各年龄死亡人数是均匀分布的, 按生育率抽样调查的结果, 受到中国婚俗及其他因素的影响, 上半年出生人数约为全年出生人数的46%, 下半年出生人数约占全年出生数的54%。这样, 由于1981年全年出生人数已有人口普查所得到的数据, 可从1980年起向1953年逐年倒推出修正后的按阳历年计出生人口数, 并由此得到各年出生率, 死亡率则可由算术平均加以修正, 这样, 可以获得最后的解。

计算结果及比较分析

计算结果与其他学者的估计比较如表2、3所示。

从表2可见, 中国的出生登记完全性比较好。即使在出生漏报率最高的1961年也只是28.01%, 一般均在10%以下。很大一部分的出生漏报是属于出生后不久死亡, 既不报出生, 又不报死亡而造成的。由于这一原因造成的漏报率, 对死亡报告影响极大, 造成表2中50年代许多年死亡漏报率达到40%左右, 而对出生漏报率影响较小。这显然是由于两者的基数不同所致。60年代以后, 出生和死亡的漏报率大大下降, 表明了我国统计工作的不断完善。1963年以前的死亡漏报, 经过第二次全国人口普查前的户口整顿得到了补正, 此后的报告经过这次整顿后质量大有提高。第二次人口普查前进行的户口整顿中发现应销户口未销者共约800万人。按本文估计1963年前全部死亡人口漏报约为600万人(不计出生后不久死亡、未登记户口者)。由于户口整顿中还发现200万人的漏登, 故总的漏报情况与本文估计十分接近。这既说明那次户口整顿是很成功的, 也说明了本文方法的正确性。从1976年以后, 出生漏报现象比较严重, 这与当时的实际情况是吻合的。1980年以后, 统计工作进一步完善, 利用了抽样调查等多种方式修正了公布的数据。因此, 年鉴发表的出生率、死亡率低估的现象甚少。

与科尔、班尼斯特、卡洛等人所得结果比较, 本文作者所作的估计更加切合历次人口普查各年龄人口、历年出生率、死亡率变化趋向的特征, 与历年所发生的各种事件相印证, 也更符合当时的实际情况。例如, 三年经济困难时期人口死亡率的变化从1959年就有较大的上升, 而并不如其他三名作者估计的那样集中在1960年。在困难时期非正常死亡总人数约为1700万人, 这与从其他资料所作的估计比较一致, 要比国外一些学者的估计低得多。本文方法的可靠性还可由各年总人口的估计、出生人数、死亡人数与其他来源的资料比较说明, 此处从略。

表2

中国历年人口出生率(1953—1980)

年份 (公元年)	年鉴数 (%)	本文作者估计数 (%)	出生漏报率 (%)	班尼斯特估计数 (%)	卡洛估计数 (%)	科尔估计数 (%)
1953	37.00	39.56	6.47	42.24	40.87	43.1
1954	37.97	39.39	3.60	43.44	41.91	44.4
1955	32.62	37.32	12.59	43.04	41.37	41.3
1956	31.90	35.92	11.19	39.89	38.28	40.2
1957	34.03	36.84	7.63	43.25	41.45	41.1
1958	29.22	31.77	8.03	37.76	36.22	37.7
1959	24.78	27.86	11.06	28.53	27.24	28.3
1960	20.86	24.24	13.94	26.75	25.65	25.2
1961	18.02	25.03	28.01	22.43	21.70	22.3
1962	37.01	39.65	6.66	41.02	39.79	40.9
1963	43.37	46.23	6.19	49.79	48.69	47.3
1964	39.14	43.63	10.29	40.29	39.82	40.7
1965	37.88	39.51	4.13	38.98	38.77	39.7
1966	35.05	36.54	4.08	39.83	39.52	38.3
1967	33.96	34.85	2.55	33.91	33.34	34.1
1968	35.59	37.78	5.80	40.96	40.35	39.1
1969	34.11	37.50	9.04	36.22	35.75	36.5
1970	33.43	35.84	6.72	36.98	36.38	37.2
1971	30.65	33.75	9.19	34.87	34.32	33.5
1972	29.77	31.51	5.52	32.45	31.69	32.4
1973	27.93	29.95	6.74	29.85	29.46	30.1
1974	24.82	27.25	8.92	28.08	27.91	27.1
1975	23.01	24.64	6.62	24.79	24.65	25.3
1976	19.91	22.84	12.83	23.05	23.14	22.5
1977	18.93	21.40	11.54	21.04	21.08	21.5
1978	18.25	21.20	13.92	20.73	20.81	21.2
1979	17.82	20.49	13.02	21.37	21.57	20.9
1980	18.21*	18.91	3.70	17.65	—	18.5

*1980年年鉴公布的出生率已经过对登记数修正

表3

中国历年人口死亡率(1953—1980)

年 份 (公元年)	年 鉴 数 (‰)	本文作者估计数 (‰)	漏 报 率 (%)	班尼斯特估计数 (‰)	卡洛估计数 (‰)	科尔估计数 (‰)
1953	14.00	20.70	32.37	25.77	18.99	25.5
1954	13.18	23.78	44.58	24.20	17.96	29.1
1955	12.28	22.54	45.52	22.33	22.31	22.4
1956	11.40	21.52	47.03	20.11	16.85	20.8
1957	10.80	20.53	47.39	18.12	13.24	19.0
1958	11.98	20.06	40.28	20.65	15.98	20.4
1959	14.59	26.91	45.78	22.06	9.20	23.3
1960	25.43	31.58	19.47	44.60	40.76	38.8
1961	14.24	24.38	41.59	23.01	27.03	20.5
1962	10.02	17.83	43.80	14.02	18.28	13.7
1963	10.04	16.35	38.59	13.81	21.22	13.0
1964	11.50	14.93	22.97	12.45	20.82	13.5
1965	9.50	13.04	27.15	11.61	10.26	11.1
1966	8.83	11.62	24.01	11.12	12.27	10.4
1967	8.43	10.40	19.94	10.47	9.14	9.9
1968	8.21	9.91	17.15	10.08	12.38	9.6
1969	8.03	9.54	15.83	9.91	8.91	9.4
1970	7.60	8.80	13.64	9.54	8.02	8.9
1971	7.32	8.23	11.06	9.24	7.73	8.6
1972	7.61	7.68	0.91	8.85	9.09	8.9
1973	7.04	7.54	6.63	8.58	6.39	8.3
1974	7.34	7.50	2.13	8.32	9.61	8.6
1975	7.32	7.43	1.48	8.07	7.62	8.6
1976	7.25	7.38	1.75	7.84	9.21	8.5
1977	6.87	7.22	4.85	7.65	7.76	8.1
1978	6.25	6.93	9.81	7.51	7.37	7.3
1979	6.21	6.74	7.86	7.61	8.33	7.3
1980	6.34	6.46	1.86	7.65	—	7.3

注: 根据本文作者的估计, 1954年中国人口平均期望寿命男性为46.38岁, 女性为49.37岁, 到1957年分别上升为49.46岁和52.68岁。从1958年起略有下降, 1960年降到最低值, 男性38.02岁, 女性39.33岁。此后平均期望寿命稳定上升, 到1971年男性达62.37岁, 女性64.70岁, 这与其他抽样调查的结果十分一致。

参考资料:

蒋正华《人口预测模型及动态参数识别》第三次全国人口普查科学讨论会论文, 1983年12月。

(作者单位: 西安交通大学人口研究所)