中国人口动态估计的方法与结果

蒋正华　李　南

一、中国人口数据来源及可靠性

1949年以前，中国没有全国性的人口统计数据，当时政府的内政部、海关和邮局从各自不同的目的出发曾对全国人口数作过估计，但都没有经过全面调查的出生率和死亡率数据，1910年与1948年的两次人口普查也由于各种原因只调查了部分省市就中途停止，因此，可以说在1949年前没有比较可靠的人口动态资料。1953年是中国第一次全国性的现代化人口普查，由于组织工作严密，此次普查总的来说可靠性较高。第一次全国人口普查的主要目的是为普选作准备，因而项目较少，集中在年龄、性别、民族等最基本的特征上，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可靠性，除了西藏等地较少数地区外，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。普查数据经过检验，联合国综合指数为20.9，玛叶指数为1.16，韦伯指数为102.5，说明1953年的普查中年龄、性别报告数据是可靠的。1964年举行了第二次全国人口普查，在前一次的基础上增加了3个调查项目，成为9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、性别等方面的报告完全性、准确性均有更高质量。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年人口普查数据质量之高现已为国际人口学界、统计学界所公认，无需再作讨论。总的来说，这三次口普查质量是高的，特别是人口年龄、性别的报告是完全可以信赖的，从全国范围来看，三次普查的完全性也十分接近。故本文以此作为一项基本数据用以估计其他人口参数。

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，只有在1982年才列出了出生与死亡的调查项目。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。死亡登记的时间序列数据波动很大，其中虽包含了各个不同时期漏报的干扰，但其总的趋势是反映真实变化的，可以在本

文中作为参考数据使用，另外由于1982年普查的高质量，本文使用1982年死亡报告求得1981年生命表做为一项基本数据用于估计。1982年普查所得的按龄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，对全国人口及其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，但它与登记数据相比具有不同的复盖范围，因此不具备可比性，只能利用来作为各种局部地区数据的对比，对全国性资料提供一类间接性的评价信息，在评价时也还必须考虑到抽样误差的影响。

根据以上情况，本文采用三次普查人口年龄、性别数据进行适当而必要的、最小量的调整后作为一项基本数据，1982年按龄死亡登记作为一项基本数据，1953—81年历年粗死亡率作为参考数据进行估计。

二、估计方法

利用1953、1964、1982年人口普查全国年龄、性别报告数据，1982年人口普查取得的上一年死亡资料可以辩别出这三次普查年之间各年的总和与分性别的按龄留存率，进而得到各年的人口年龄、性别结构，总和的与分性别的完全生命表，进而再得到各年的出生率与死亡率。其方法如下：

（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_{am}^{1982}、P_{aF}^{1982}、D_{am}^{1981}、D_{aF}^{1982} 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，首先注意假设一个按龄留存率 SR_a^0 作为迭代初值，于是1981年中a岁人口数的零次迭代值可由

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

求得，由此得到1981年a岁死亡率一次迭代值

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

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 $SR'a$：

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

将 SR^1a 代入(1)式以取代 SR_a^0 即可求得 $P_a^{1981\cdot 1}$，依次类推可以建立起第二个生命表求得第三次迭代值 SR_a^2 等等。这一迭代过程进行到第n次迭代得到的 $P_a^{1981\cdot n}$ 与第n+1次迭代所得到的 $P_a^{1981\cdot n+1}$ 之差小于给定值 ε（$\varepsilon \ll 1$）时：

$$| P_a^{1981\cdot n+1} - P_a^{1981\cdot n} | \leqslant \varepsilon \qquad (4)$$

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，这一生命表即是1981年生命表。以上过程可分别对求解得到分性别的生命表和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。有关这一方法的详细论述见参考资料，

（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、从上一年推出的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$Pa_1(a+n_1)$以及n_1年的留存率$SPn_1(a)$数列，按以上指示，迁移可以忽略时：

$$Pn_1(a+n_1)=P(a)\prod_{j=0}^{n_1-1}SR_j(a+j) \tag{5}$$

$$\ln\frac{Pn_1(a+n_1)}{P(a)}=\sum_{j=0}^{n_1-1}\ln SR_j(a+j) \tag{6}$$

令 $T(a)=\ln\frac{Pn_1(a+n_1)}{P(a)}$ $Rj(a+j)=\ln[SR_j(a+j)]$

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

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

$$\hat{R}_j(a)=\left(\sum_{l=0}^{n_2}C_{jl}\left(\frac{a}{10_0}\right)\right)R^*(a) \tag{8}$$

式中$R^*(a)$是基准生命表$SR^*(a)$的变换。n_2的取法也在附录中给出，最高人口年龄本文中取为100岁。现在，我们希望确定参数C_{jl}，使各年生命表确定后，使后一次人口普查各年龄人口倒推到前一次人口普查时刻相应年龄人口与普查统计数误差最小，即

$$\min\left|Pn_1(a+n_1)-P(a)\prod_{j=0}^{n_1-1}\widehat{SR}_j(a+j)\right| \tag{9}$$

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。由于一致逼近在数字上不易实现，（9）式改为另一种2次型目标函数而将一致性要求留待下步进行：

$$F(C_{j*})=\sum_{a=0}^{100-n_1}\left(T(a)-\sum_{j=0}^{n_1-1}\hat{R}_j(a+j)\right)^2$$

$$=\sum_{a=0}^{100-n_1}\left[T(a)-\sum_{j=0}^{n_1-1}\left(\sum_{l=0}^{n_2}C_{jl}\left(\frac{a+j}{100}\right)^l R^*(a+j)\right)\right]^2 \tag{10}$$

本文原用前二阶矩，目标函数为

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

令（11）最小，即可解得两次普查年间各年的参数C_{j0}、C_{j1}。双参数模型于是可写为

$$\begin{cases} \min\limits_{c} F(c) = \min\limits_{c} \left[\frac{1}{2} C^T (A^T A) C - I^T \bar{A}^T C\right] \\ S.t. \quad \left[C_{j0} + C_{j1}\left(\frac{a+j}{100}\right)\right] > 0 \end{cases} \tag{12}$$

式中符号详见附录。

（三）参数模型求解。

附录中证明了式（12）所表示的是一个凸规划问题，约束条件保证了留存率为正且小于1，由于式（12）的约束为开集，一般说来可能无解，因此将问题改为

$$\begin{cases} \min\left[\frac{1}{2} C^T (A^T A) C - I^T \bar{A}^T C\right] \\ S.t \quad C_j \geqslant 0 \quad C_{j1} \geqslant 0 \end{cases} \tag{13}$$

若取后一次普查为标准表，在留存率不大于标准表对应留存率假定下，还可将约束写为

$$S.tCjo \geqslant 1 \quad Cj_1 \geqslant 0$$

这是一个典型凸规划问题，最优解存在且唯一。但在约束起作用时，最优解必在约束集边界达到，因此本文不求其最优解，而是按附录中方法先找一个C的经验初值CB，然后求以上问题的对CB的最小偏离最优解。CB先由历年的死亡率按近似方法求出，再作经验校正，校正的死亡率不少于已知的各年死亡率，如不满足，重新校正。

（四）一致优化。

由于（13）的解并不满足（9）式要求，即按（13）中求得的各年生命表倒推时各年龄上误差分布不一致，需进行一致优化，这里的目标函数为

$$\min_{\overline{SR}} \left[\alpha (T(a) - \Pi \overline{SR})^2 + (1-\alpha) \Sigma (\overline{SR} - \hat{SR})^2\right]$$

其中α为权系数，但并不是人为地给定，而是由对一致误差的要求确定，这将在附录中说明。一致优化公式为：

$$\overline{SR}_k(i+k) = \sqrt[n]{\frac{1 - E/TS_i}{1 - R_{2i}/TS_i}} \cdot \hat{SR}_k(i+k) \tag{14}$$

符号说明与推导过程见附录。

这里的一致优化过程可使各年龄上的误差一致地小于预先给定的任意正数E，E越小对$\hat{SR}$的修正越大，数E不能过小使$\overline{SR}$对$\hat{SR}$偏离太大从而使最终生成的生命表与标准表相比过于奇异。由于1982年普查年龄报告误差为6.15‰，要求E<6.15‰并无意义，本文中对1964—81年E=7‰，对1953—64年E=10‰，一致优化的结果$\overline{SR}$即是各年生命表中的留存率函数，可由$\overline{SR}$按一般人口学方法生成各年生命表，再求出各年按龄别人口数。

（五）解的修正。

由于人口普查选取的标准时刻为7月1日，按以上所得到的人口年龄结构均相应于年中时刻，而各年生命表也是本年7月1日至下一年7月1日间生命表，由此算出的各年出生率与死亡率也均为本年7月1日到下一年7月1日间的出生率与死亡率，这样的生命表与出生率、死亡率不是按公历年计的值，因而与《中国统计年鉴》和其他资料不具备可比性，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人数的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年逐年推出修正后的按公历年计出生人口数，并由此得到各年出生率。按龄死亡率则可由算术平均加以修正，得出各年死亡率及总和与分性别的生命表。本文口径与《中国统计年鉴》一致，即包括大陆29个省、市、自治区。

三、计算结果

利用本文的方法可以取得从1953年到1981年每年的总和及分性别生命表，分性别和年龄别的人口年龄构成，每年出生婴儿数，出生率及死亡率；1953年以前的出生率及死亡率亦可作出估计，但其误差无法估计。表1与2中分别列出了本文作者对历年出生率、死亡率的估计，并与班尼斯特、卡洛和科尔的估计数作了对比。1980年以后的统计年鉴数字已通过各种资料的分析由统计部门作了校正，从本文提出的方法来说，看1982年以后的数据需从下一次人口普查结果中取得原始资料，因此不进行分析。

表2　　中国历年人口出生率（1953—1980）

年份（公元年）	《中国统计年鉴》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	26.76	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	29.35	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	17.63	—	18.5

* 1980年公布的出生率已经过对登计数修正。

从表1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年也只有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，是造成表2中50年代许多年死亡漏报率达到40%左右的主要原因，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在不断的完善。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

表2　　中国历年人口死亡率（1953—1980）

年份（公元年）	《中国统计年鉴》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	18.94	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	9.92	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	1.76	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

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《中国统计年鉴》公布的出生率、死亡率估低的现象甚少。

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，而并不如其他三名学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。

四、后记

本文的初步结果曾由国务院人口普查办公室和中国人口学会与西安交通大学人口研究所联合召开专家会议进行了讨论，国家统计局、公安部、国家计划生育委员会、中国科学院、

中国社会科学院、中国人民大学、北京经济学院、航天部、中国人口丛书总编会等单位和有关部门的专家学者就本文的方法及结果提出了许多宝贵的意见，对本文的最后修改定稿很有益处，作者谨向上述单位及参加专家会议的同志们致谢。

（作者工作单位：西安交通大学人口研究所）

附录　计算方法

（一）最小二乘估计。

1.表示方法

由于该估计参数多于已知约束数，本文建立了适合最小二乘法而又具有原参数与表示出参数前n阶矩相等意义的表示方法：

若f(x)与g(x)为（a、b)上定号可积函数f(x)、$g_1(x)>0$

$$|f(x)-g(x)|=|\Delta g(x)|\ll 1 \qquad (1)$$

$$\hat{f}(x)=\left(\sum_{i=0}^{n}C_i x^i\right)g(x) \qquad (2)$$

则可选Ci使f(x)与$\hat{f}(x)$进n阶矩相等.

命　$F(C^i)=\int_a^b[(f(x)-\hat{f}(x))^2/g(x)]dx \qquad i=0\sim n$

$$\frac{\alpha F}{\alpha C_i}=-2\left[\int_a^b(f(x)-\hat{f}(x))x^i dx\right]=0$$

则　$$\int_a^b x^i f(x)dx=\int_a^b x^i\hat{f}(x)dx \qquad (3)$$

由(3)与(2)有：

$$AC=B \qquad (4)$$

其中　$A=(a_{ij})_{n'\times n'}$　$C=(C_i)_{n'\times 1}$　$B=(b_j')_{n'\times 1}$　$n'=n+1$

$$a_{ij}=\int_a^b x^i x^j g(x)dx \qquad b_j=\int_a^b x^j f(x)dx$$

$$C_0=\frac{1}{A}\begin{vmatrix}\int_a^b f(x)dx & a_{01}\cdots a_{0n}\\ \vdots & \\ \int_a^b x^n f(x)dx & a_{n}\cdots a_{nn}\end{vmatrix}$$

又　$f(x)=g(x)+\Delta g(x)$

$$C_0=1+\frac{1}{A}\begin{vmatrix}\int_a^b\Delta f(x)dx\cdots\cdots a_{0n}\\ \vdots \qquad\qquad \vdots\\ \int_a^b x^n\Delta g(x)dx\cdots a_{nn}\end{vmatrix}=1+\Delta_0 \qquad |\Delta_0|\ll 1$$

同理有

$$C_i=\frac{1}{A}\begin{vmatrix}a_{00}\cdots a_{0,i-1}\cdot\int_a^b\Delta g(x)dx_1\cdots a_{0n}\\ \cdots\\ a_{n0}\cdots a_{n,i-1}\cdot\int_a^b x^n\Delta g(x)dx_1\cdots a_{nn}\end{vmatrix}=\Delta_i \qquad |\Delta_i|\ll 1,\ i\geqslant 1$$