

中国的初婚初育模型研究

黄荣清 开 昕

人口学是研究人口再生产及其运动规律的。人口学历来把婚姻作为主要研究内容,是因为婚姻和生育有密切关系。在我国的人口学研究中,虽然对婚姻已有过不少著述,但总的来说,大部分还是描述性的,理论上定量研究婚姻模式尚少。本文从介绍寇尔的初婚模式开始,进而提出另一改进模型,并用实际数据对两种模型作了比较。在第二部分,本文把初婚时间与正常性生活开始时间区别开来,并分析了我国妇女正常性生活开始的时间分布。最后,给出了一套从初婚到初育的估测方法。

一 初婚模型

(一) 寇尔的初婚模型。在人口学研究中,寇尔(Coale)的同期人已婚比例模型是比较著名的。他在研究欧洲各国妇女的年龄别已婚比例时发现,若以一定的初婚年龄为起点,通过调整横轴(年龄轴)、纵轴(已婚比例),则各国的已婚比例曲线几乎重合。根据这一现象,他以1865~1869年瑞典的人口数据为基础,作了一个标准年龄别(X_s)的初婚率(g_s)、已婚率(G_s)和累积已婚年数(Z_s)的时间序列模型。即

$$x_s = \frac{a - a_0}{k},$$

$$G_s(x_s') = \int_0^{x_s'} g_s(x_s) dx_s,$$

$$Z_s(x_s') = \int_0^{x_s'} G_s(x_s) dx_s.$$

各个口群体的妇女年龄别初婚比例,通过调整一定的初婚年龄起点 a_0 、横轴比例 R 和纵轴比例 c ,就可与标准的初婚率分布相重合。由于在实际中,初婚年龄起点很难确定,所以 a_0 只表示发生一定数量的初婚年龄; k 表示在标准时间序列表中发生的一定的初婚率在特定的人口群体中需要的年数。若 $k < 1$,则表示初婚速度比标准快; $k = 1$ 时,则为标准速度; $k > 1$,则初婚速度比标准慢。 c 为终身已婚率。为了计算特定人口群体的3个系数,首先计算5岁组已婚率之比 R_1, R_2, R_3 ,这里

$$R_1 = \frac{M_{15-19}}{M_{20-24}}, \quad R_2 = \frac{M_{20-24}}{M_{25-29}}, \quad R_3 = \frac{M_{25-29}}{M_{30-34}}$$

其中 M_i 表示已婚率。

根据 R_1, R_2, R_3 的值进行组合,当 $R_1 > 1 - R_3$ 时把 R_1 和 R_2 组合,否则就把 R_2, R_3 一起组合,根据一定的表来推算 a_0 和 k 。 c 为由算出的 a_0 和 k ,根据模式时间表推算的已婚率除以实际已婚率之比。寇尔认为,由25~29岁的已婚率就可求出 C

$$C = \frac{M_{25-29}}{\frac{k}{5} [Z_s(\frac{30-a_0}{k}) - Z_s(\frac{25-a_0}{k})]}$$

1971年,寇尔和麦克尼尔把由上述经验数据作成的初婚率模型标准时间表改写成下列概率密度函数

$$g_1(x_1) = (0.19465) \exp[-0.174(x_1 - 6.06) - \exp(-0.288(x_1 - 6.06777))]$$

他们的推导过程是这样的:

$$\text{令: } r(a) = \frac{g(a)}{1-G(a)}$$

表示a岁妇女初婚的可能性,由标准的经验曲线可得:

$$r(x) = 0.174 \exp(-4.411 \exp(-0.309x)) \quad (1.1)$$

由于在(1.1)式中,当 $x \rightarrow \infty$ 时, $r(x) \rightarrow 0.174$,所以可令 $k=0.174$ 。对于一般的初婚频率,寇尔认为它可以是两个随机变量之和。一个为可能结婚变量 ξ_1 (它的分布为 $g_1(x)$),另一个是从可能结婚到实际结婚变量 η (它的分布为指数分布,即 $\eta = e^{-k(a-x)}$),则两个随机变量之和的分布为

$$g(a) = k \int_0^a g_1(x) e^{-k(a-x)} dx$$

其中, $g(a)$ 为结婚率,它可以由实际数据来确定,从而可以确定 ξ_1 的分布 $g_1(x)$,由此得到 $g_1(a) = g(a) + \frac{1}{k} g'(a)$;而 ξ_1 又可看成是两个随机变量 $\xi_1 = \xi_2 + \eta$ 之和,则 ξ_1 的分布为:

$$g_1(x) = k \int_0^a g_2(x) e^{-k(a-x)} dx$$

其中 $g_2(x)$ 为 ξ_2 的分布。类似地有

$$g_n(x) = k \int_0^a g_{n-1}(x) e^{-k(a-x)} dx$$

其中 $g_{n-1}(x)$ 为 ξ_{n-1} 随机变量的分布。

经过这样无限地推导,他们发现 $g_n(x)$ 越来越近于正态分布。然后,他们把这样一个由类似正态曲线的函数形式作为概率密度函数,即:

$$g(x_1) = 0.195 \exp[-0.174(x_1 - 6.06) - \exp(-0.288(x_1 - 6.06))] \quad (1.2)$$

$$G(x_1) = \int_0^{x_1} g(x_1') dx_1' \quad (1.3)$$

$$\text{其中 } x_1 = \frac{x - a_0}{k} \quad (1.4)$$

若以 c 表示50岁以上妇女的终身结婚率,把(1.2)、(1.4)代入(1.3)式,并考虑到纵轴的坐标交换 c ,则以年龄为变量的初婚率分布为

$$G(x) = C \int_{x_0}^x 0.1946 \exp \left\{ (-0.174) \left(\frac{x - x_0}{k} - 6.06 \right) - \exp \left[(-0.288) \left(\frac{x - x_0}{k} - 6.06 \right) \right] \right\} d \left(\frac{x - x_0}{k} \right)$$

其密度函数分布为

$$g(x, x_0, k, c) = \frac{0.1946}{k} \exp \left\{ \left(-\frac{0.174}{k} \right) (x - x_0 - 6.06k) \right. \\ \left. - \exp \left[\left(-\frac{0.288}{k} \right) (x - x_0 - 6.06k) \right] \right\} \quad (1.5)$$

由(1.2)式可得随机变量 ξ 的均值为11.37, 所以该人口的平均初婚年龄为

$$SMAM = x_0 + 11.37k \text{ 或 } k = \frac{SMAM - x_0}{11.37}$$

x_0 表示起始初婚年龄。

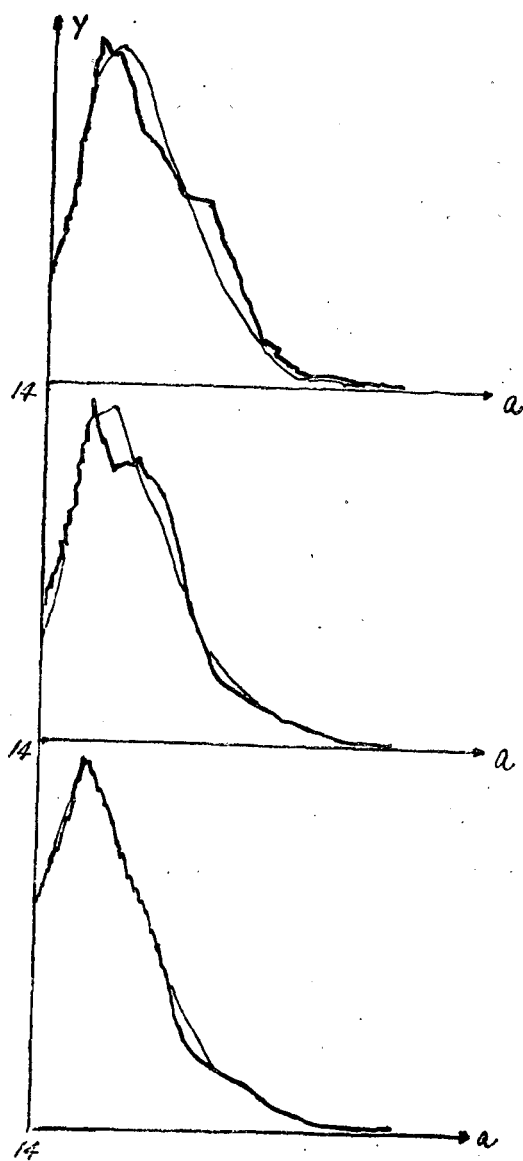


图1 1981年35岁、40岁、45岁(自上而下)同期群妇女按龄初婚率曲线
图中粗线表示实际数据值, 细线表示模型值。

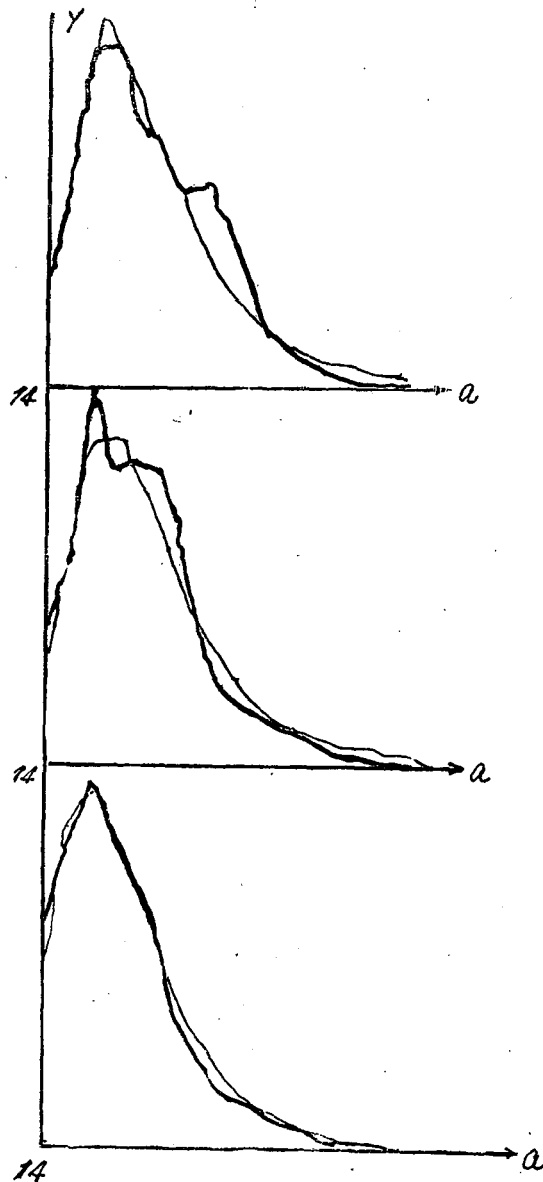


图2 1981年35、40、45岁(自上而下)同期群妇女按龄初婚率曲线

注: 图2为对数正态分布初婚模型的拟合曲线。粗线表示原始数据值, 细线表示模型值。

① 拟合时, 式(1.5)的幅度系数 $\frac{0.1946}{k}$ 以一参数 r 来代替, 式(1.5)就成为三参数模型, 参数为 r, k, x_0 。

因此只要知道了终身结婚率 c 、平均初婚年龄SMAM,并规定了初婚起始年龄,就决定了此人口的已婚分布。

据上述模型,用中国人口的初婚数据进行检验时,由于我国妇女基本上都结婚,所以 $c=1$ 。初婚模型主要取决于参数 a_0 和 k ,利用1982年中国1‰人口生育率抽样调查资料,直接对(1.5)式进行曲线拟合,其结果是好的(见表1、图1)。

(二)对数正态分布的初婚模型。寇尔和麦克尼尔的初婚模型被认为对许多人口是适用的。但也有人作过实证研究,认为此模型对某些人口并不太适合(即模型曲线和实际结果拟合得不太理想)。笔者则认为,寇尔在模型的推导上不够简捷,并且其中的一些常数都是经验取得的,对不同人口可能会有不同。这里我们换了一个思路来推导这个初婚模型。

人口的已婚变化是与该人口的未婚变化相对应的。设某个人口开始时都未婚,随着年龄增大,不断有人加入可能结婚的队伍。同时,由于有人结了婚,退出了未婚队伍,所以未婚队伍越来越小。在开始时,已婚人口尚少,尽管未婚队伍在缩小,但结婚队伍却在增大。到一定年龄后,结婚人数增多,未婚人数减少,可以补充到可能结婚队伍中的人口也不断减少,最后趋于零,随之,未婚比例也趋于零。由于每个人从未婚到可能结婚,从未婚到结婚是随机事件,所以我们可以把人口的可能结婚、未婚看作是两个随机变量,其中 ξ 为可能结婚变量, η 表示未婚变量。由于 ξ 随年龄变化时呈“两头大,中间小”的形状,故可看成是具有正态分布的随机变量;而 η 表示人口中未婚比例变化,它越来越小,可看成是具有指数函数分布的随机变量,且 η 是 ξ 的函数,即,

$$\eta = e^{-\xi}$$

ξ 的密度分布为正态分布,它的函数形式为

$$f_{\xi}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{[(x-a_0)-\mu]^2}{2\sigma^2} \right\}$$

由于 $\xi = -\ln \eta$,可知 η 的密度分布为

$$f_{\eta}(y) = \frac{1}{\sqrt{2\pi}\sigma(y-a_0)} \exp \left\{ -\frac{[\ln(y-a_0)-\mu]^2}{2\sigma^2} \right\} \quad (1.6)$$

即一个人口的已婚密度分布为对数正态分布。

用实际数据可以验证上述模型的合理性。用同样的1‰抽样资料,对1981年全国20~46岁27个同期群初婚数据进行拟合,模型拟合结果良好(见图2及表2)。

表3给出了两个模型误差情况。总的看来,对以正态分布模型的精度比寇尔模型的精度高。

表1 模型拟合参数及误差平方和

周期群 年龄	参数 i	参数 k	起始年龄 x_0	$\Sigma(Y_i - \hat{Y}_i)^2$
35	0.397	0.489	14.00	0.0025
40	0.416	0.475	14.00	0.0027
45	0.424	0.399	14.00	0.0029

表2 对数正态分布模型拟合参数及误差

周期群 年龄	参数 σ	μ	起始年龄 a_0	$\Sigma(Y_i - \hat{Y}_i)^2$
35	0.512	1.78	14	0.0018
40	0.508	1.76	14	0.0027
45	0.555	1.67	14	0.0012

表3

两个模型拟合误差比较

$\sum_{i=1}^n (y_i - \hat{y}_i)^2$	同期群年龄	35	36	37	38	39	40
寇尔法		0.0025	0.0078	0.0039	0.0043	0.0024	0.0027
对数正态分布法		0.0018	0.0031	0.0038	0.0044	0.0028	0.0027
$\sum_{i=1}^n (y_i - \hat{y}_i)^2$	同期群年龄	41	42	43	44	45	46
寇尔法		0.0019	0.0019	0.0013	0.0018	0.0029	0.0079
对数正态分布法		0.0014	0.0014	0.0011	0.0018	0.0012	0.0032

关于人口的初婚模型推导还可以换一个角度来看。假设一个同期群人口可分成若干不同的小群体，各个小群体进入可能结婚队伍时间虽然不同，但她们未婚比例的变化都是按指数型分布的，即：

$$\eta_i = e^{-\xi_i}$$

$$\xi_i = -\log \eta_i \quad (i=1, 2, \dots, n)$$

这样同期群的可能结婚队伍是各个小群体队伍之和，即

$$\xi = \sum \xi_i = -\log(\eta_1 \eta_2 \dots \eta_n) = -\log(\eta)$$

在满足一定条件下，由中心极限定理

$$\frac{\xi - E(\xi)}{\sqrt{D\xi}}$$

近似地服从 $N(0, 1)$ 分布，故 ξ 近似地服从 $N(a, \sigma)$ (令 $a = E(\xi)$, $\sigma = \sqrt{D\xi}$)，所以初婚的分布近似地服从对数正态分布。

二 初婚与正常性生活

结婚意味着男女双方正常性生活的开始。这里所说的正常性生活开始，是指妇女从这个时间起，她的生育按自然生育概率生育第一胎。在实际中，婚姻和正常性生活开始的时间并不完全一致。例如，有的人虽然办了结婚登记手续，但由于种种原因，夫妻双方可能并未马上同居；或者虽然同居了，但由于某些原因，他（她）们会比按自然生育概率估算的生育第一胎时间推迟一些。在这种情况下，我们便认为正常性生活开始时间晚于结婚时间。另一方面，也有一些人可能在结婚以前就有性关系，怀孕以后才登记。这时就认为正常性生活时间先于结婚时间。应该说明，本文是从对生育影响的角度来衡量正常性生活开始时间的。这个概念本身是否合理，在这里不作讨论。

设年龄别自然生育概率为 P_i ，某个妇女在 i 岁开始有正常性生活，则在当年，她的生育概率为

$$P(i, 0) = (1 - \frac{280}{365}) \cdot p_i = (\frac{85}{365}) p_i \quad (2.1)$$

式 (2.1) 中的 $\frac{85}{365}$ 是因为若假定妇女在过第一次性生活时就怀孕，则她要过 280 天后才能

生育。若结婚是均匀分布的, 而且在每个年龄内生育概率一定, 则有 (2.1) 式。

第一年不生育, 第二年生育的概率为

$$p_{i+1} = (1 - \frac{85}{365} p_i) p_{i+1} = (1 - 0.233 p_i) p_{i+1}$$

则头两年生育的概率为

$$P(i, 1) = p_{i+0} + p_{i+1} = (1 - (1 - 0.233 p_i) (1 - p_{i+1}))$$

类似地, $n+1$ 年内生育概率为

$$P(i, n) = 1 - (1 - 0.233 p_i) (1 - p_{i+1}) \cdots (1 - p_{i+n}) \quad (2.2)$$

若写成连续形式, 则有

$$P(x_0, x) = \exp(-\int_{x_0+0.767}^x \mu(a) da), (x \geq x_0 + 0.767)$$

其中 x_0 为正常性生活初始年龄点, $\mu(x)$ 表示在年龄 x 点的自然生育力。若妇女的正常性生活的年龄分布为 $m'_1(x)$, 则到 x 岁, 妇女生头胎的生育率为

$$R(x) = \int_{x_0}^x m'(y) \cdot P(y, x) dy \quad (2.3)$$

在实际工作中, (2.3) 式的离散形式比较常用, 记

$$Q = \begin{bmatrix} Q(x_0, 0) & 0 & \cdots & 0 \\ Q(x_0, 1) & Q(x_0+1, 0) & \ddots & 0 \\ \vdots & \vdots & & \vdots \\ Q(x_0, n) & Q(x_0+1, n-1) & & Q(x_0+n, 0) \end{bmatrix}$$

为生育概率矩阵

$$\vec{m}' = (m(x_0), m(x_0+1), \cdots, m(x_0+n))^T$$

$$\vec{R} = (R(x_0), R(x_0+1), \cdots, R(x_0+n))^T$$

分别表示正常性生活的年龄分布和头胎累积生育率分布, \vec{m}' 、 \vec{R} 称为正常性生活向量和头胎生育向量, 它们都为列向量。这样 (2.3) 式可写为

$$\vec{R} = Q \cdot \vec{m}' \quad (2.4)$$

$$\text{或 } \vec{m}' = Q^{-1} \vec{R} \quad (2.5)$$

若 Q 确定以后 (Q 的确定见参考文献①), 则由 \vec{m}' 可求出 \vec{R} 。反过来, 由头胎生育分布 \vec{R} 也可求出 \vec{m}' 。

设正常性生活的分布也服从对数正态分布, 即

$$m'(x) = \frac{1}{\sqrt{2\pi} \sigma (x - x_0)} \exp \left(-\frac{(\ln(x - x_0) - \mu)^2}{2\sigma^2} \right) \quad (2.6)$$

利用1982年1‰人口生育率抽样调查资料中各同期群第一胎的按龄生育率数据可求 $m'(x)$, 其 μ 和 σ 及误差值见表4。

表4

正常性生活分布的参数值及误差值

周期群年龄 参数及误差	35	36	38	40	42	44	45	46
μ	1.746	1.723	1.673	1.768	1.756	1.647	1.590	1.561
σ	0.485	0.517	0.478	0.409	0.508	0.609	0.644	0.658
$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	0.0027	0.0033	0.0043	0.014	0.019	0.011	0.0063	0.007

由(1.6)式及(2.6)式,我们可以算出原按龄初婚率分布及正常性生活分布的平均年龄和峰值年龄。 $\bar{m}(i)$ 和 $\bar{m}'(i)$ 表示原按龄初婚率分布的平均值和正常性生活分布的平均值; $Ma(i)$ 和 $Ma'(i)$ 分别表示上述两个分布的峰值年龄,其值如表5所示。

表5

按龄初婚曲线的平均值、峰值与正常性生活曲线对应值的比较

周期群 比较	30	32	35	37	39	40	41	42	43
$\bar{m}(i)$	21.99	21.16	20.73	20.29	20.73	20.65	20.51	20.35	20.07
$\bar{m}'(i)$	21.62	21.02	20.45	20.15	20.15	20.37	20.49	20.58	20.42
$\bar{m}(i) - \bar{m}'(i)$	0.37	0.15	0.29	0.14	0.62	0.28	0.03	-0.23	-0.35
$Ma(i)$	19.63	19.65	18.55	18.74	18.74	18.50	18.42	18.28	18.20
$Ma'(i)$	19.31	18.79	18.53	18.19	18.59	18.95	18.90	18.47	17.87
$Ma - M'a$	0.32	0.87	0.01	0.55	0.15	-0.45	-0.48	-0.20	0.33

由于在模式计算中,总是假设婚姻、生育在每个年龄内为均匀分布,而实际上婚姻和生育可能比较集中在某几个月份,所以由模式计算的婚姻和正常性生活在时间上会存在误差(但不会太大)。由表5可知,在37岁以上同期群,其平均初婚年龄与正常性生活平均年龄基本一致,峰值初婚年龄和正常性生活峰值年龄也基本一致。但在37岁以下,平均初婚年龄值与其峰值有大于正常性生活相应两值的趋势。这说明,其中部分人实际性生活开始于结婚以前。而这种倾向在30岁以下表现得尤为明显,若再考虑到可以用避孕方法不使怀孕,或者怀孕可用流产终止生育,则婚前有性生活的比例还要大。

当然,在任何时期,婚前性生活情况或多或少是存在的,但婚前性生活有明显增大趋势。从全国看,可能发生于70年代中期,而城镇可能于70年代初即已开始。这一点,从初婚后当年生育率在不同年代的比较即可知道。在60年代以前,结婚后当年的生育率小于50%;60年代前期,由于受自然灾害影响,结婚当年生育率比较低;到60年代中期,当年生育率约为60~70%;70年代初在100%以内;而到70年代中期都高于100%;1981年达147.4%。城镇妇女结婚当年生育率,从70年代开始就大于100%。1970年为126.9%,1978年为153.8%,70年代10年平均为124.4%。根据笔者计算,妇女理论上的自然生育率在15~30岁为0.4~0.6之间,若假设妇女结婚就等于正常性生活开始,以0.5来粗略估计妇女婚后一年内的自然生育概率,则大约为 $0.5 \times \frac{85}{365} = 116.4\%$,这是理论上的最大值。由上面数字可以知道,70年代后期实际值却远远大于理论值,说明实际的性生活开始时间应早于登记的婚姻时间。

三 关于初育模型

由上面的讨论可以知道,初育可以由正常性生活的分布来决定。由于正常性生活的分布

和初婚分布都可以用对数正态分布函数来表示, 所以我们只要建立两个对数正态分布函数的关系, 再利用式(2.4)就可以建立初育模型。

设初婚和正常性生活开始的初始年龄相同, 初婚的密度分布为 $f_1(\mu_1, \sigma_1)$, 正常性生活开始的密度分布为 $f_2(\mu_2, \sigma_2)$ 。每个对数正态分布都有一个相对应的正态分布。记随机变量 ξ_1 有 $f_1(\mu_1, \sigma_1)$ 相对应的密度分布为 $g_1(\mu_1, \sigma_1)$, 随机变量 ξ_2 有 $f_2(\mu_2, \sigma_2)$ 相对应的密度分布为 $g_2(\mu_2, \sigma_2)$, 且 ξ_2 为 ξ_1 的线性函数

$$\xi_2 = a\xi_1 + b$$

则 ξ_2 的密度分布为

$$f_{\xi_2}(y) = \frac{1}{a} f_{\xi_1}\left(\frac{y-b}{a}\right) = \frac{1}{\sqrt{2\pi} a \sigma_1} \exp\left(-\frac{1}{2} \frac{[y-(a\mu_1+b)]^2}{(a\sigma_1)^2}\right)$$

$$\text{令 } \sigma_2 = a\sigma_1 \quad \mu_2 = a\mu_1 + b$$

$$\text{或 } a = \sigma_2/\sigma_1 \quad b = \mu_2 - a\mu_1$$

则我们便建立了两个正态随机变量的关系, 同时也确定了两个对应的对数正态分布的关系。由表2和表4知各个同期群妇女初婚分布参数 μ_1, σ_1 和正常性生活开始时间分布参数 μ_2, σ_2 。相对应的 a, b 值见表6。

表6 σ_2/σ_1 和 $\mu_2 - \sigma_2/\sigma_1 \mu_1$

周期群年龄 比 值	33	35	38	40	42	44	45	46
$\sigma_2/\sigma_1 = a$	1.02	0.947	1.1123	0.804	0.989	1.25	1.16	1.00
$\mu_2 - \frac{\sigma_2}{\sigma_1} \mu_1 = b$	-0.157	0.0641	-0.304	0.350	0.0586	-0.399	-0.341	-0.747

由表6可以知道, 当 $a > 1$ 时, 有 $b < 0$; $a < 1$ 时, 有 $b > 0$ 。在35岁以上, a 总是在1, b 总是在0左右。但在35岁以下, 其 a 和 b 变化较大。实际上由初婚预测初育时, 我们可以同时调整 a, b 值使模型值和实际值相符。

另外, 由对数分布函数可知, 其均值年龄 $x_{\text{均}}$ 和峰值年龄 $x_{\text{峰}}$ 分别为

$$y_{\text{峰}} = x_{\text{峰}} - a_0 = \exp(\mu - \sigma^2)$$

$$y_{\text{均}} = x_{\text{均}} - a_0 = \exp(\mu + \sigma^2/2)$$

则

$$\mu = \frac{1}{3} \ln(y_{\text{峰}} \cdot y_{\text{均}})$$

$$\sigma^2 = 2/3 \ln(y_{\text{均}}/y_{\text{峰}})$$

由实际的平均初婚年龄、峰值和初婚年龄算出对应的 μ, σ , 再由经验数据找到 a, b 值, 便可估计其初育分布。或者直接由经验估计正常性生活开始的平均年龄和峰值年龄, 利用式(2.6)来确定正常性生活的分布, 从而决定对应的初育分布。

参考文献:

- ①寇尔:《婚姻和生育各种新模型的发展》, 载(英)《人口》特辑, 1977。
- ②黄荣清、开昕:《同期群胎次——年龄别生育模型》, 载《人口与经济》, 1990年。第1、2期。
- ③王梓坤:《概率论基础及其应用》, 科学出版社, 1976年。
- ④《中国计划生育年鉴》, 人民卫生出版社, 1986年。
- ⑤1982年中国《千分之一人口生育率抽样调查资料》, 《人口与经济》特刊, 1983年。

(本文责任编辑: 郭汉英)

(作者工作单位: 北京经济学院人口所)