中国妇女生育模型研究*

谢韦克 黄荣清

生育率是人口学中重要的统计指标。许多年来，不少人口学家一直在尝试建立一个能准确描述生育率变化规律的数学模型。在各种模型中，科尔（Coale）的生育模型是最著名的。他的生育模型由2个子模型构成，一个是初婚模型，一个是已婚生育模型。黄荣清1990年提出了初婚对数正态分布模型。他用二个模型，即科尔的初婚模型与对数正态分布模型对1981年1‰生育资料进行拟合。总的来看，对数正态分布模型的精度比科尔模型的精度高。

黄荣清推导模型的思路比较简捷，而且符合初婚的实际过程。他的思路中最有启发性的一点是利用了随机变量函数的分布，不过，那篇文章并没有对模型本身进行更多的阐述。比如，两个参数μ，σ的意义。在把这个模型用于生育时，仅限于初育。另外，该文中的某些数学假设也欠严密。本文准备就这些问题进行更深入的研究。

一、生育模型

我们先讨论一胎生育。科尔在研究初婚模型时，把初婚频率表示为2个随机变量的和。一个为可能结婚变量，另一个是从可能结婚到实际结婚变量。借用科尔的提法，我们也用2个随机变量建立生育模型。一个是可能生育一胎变量ξ，它的分布表示在生育时间内某时点生育可能性的大小。另一个是从可能生育一胎到实际生育一胎变量η，简称实际生育变量，具体表现为一胎生育年龄。η取某个年龄值的概率可以用一胎年龄别生育率近似表示。由于生育可能性的大小总要通过生育年龄表现出来，所以可将实际生育变量η表示为可能生育变量ξ的函数，又由于从可能生育到实际生育总要经过一段时间，所以函数的形式可以表示为：

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

一旦ξ的分布确定，η的分布也就确定了。估计ξ的理论分布有一定的困难，因为这个概念比较抽象，而与可能生育有关的具体因素又很多。不过至少有一点可以肯定，ξ随生育时间的分布呈“两头小，中间大”的形状，故可以近似看成具有正态分布的随机变量。它的密度函数为：

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

a_0表示生育一胎的起始年龄。

* 本文系高校人口研究和培训项目（P04）第四次全国人口普查资料分析课题研究成果之一。

由概率论知识可以知道，如果$l_n = \xi$服从正态分布，则η服从对数正态分布。或者说，如果ξ服从正态分布，则$\eta = e^{\xi}$服从对数正态分布。推导过程也很简单，这里从略。η的密度函数为：

$$f\eta(x) = \frac{1}{\sqrt{2\pi}\,\sigma\,(x-a_0)} \exp\left\{-\frac{[l_n(x-a_0)-\mu]^2}{2\sigma^2}\right\}$$

一胎生育率随年龄的分布为对数正态分布。

二胎生育与一胎生育稍有不同。因为妇女只有在生了第一胎后才可能生第二胎。所以，妇女生育第二胎的年龄应该是2个随机变量的和，即第一胎的生育年龄与胎次间隔之和。但是，如果用这样的思路处理二胎生育，就大大增加了问题的复杂性。为简单计，我们用处理一胎生育的方法处理二胎生育，同样可以推导出二胎生育率随年龄的分布为对数正态分布。

利用1988年2‰生育资料，将数据分为城市、农村2个类型，对1943—1953年出生的妇女，即1988年35—45岁共20个同期群妇女的一胎及二胎年龄别生育率用最小二乘法进行对数正态分布拟合，即设实际的年龄别生育率为$\hat{y}_i$，模型值为y_i，选择参数μ，σ使$\sum(y_i-\hat{y}_i)^2$达到最小拟合结果参见表1—表4。

表 1　20个同期群妇女模型拟合的平均误差

胎　次	农　村	城　市
一　胎	0.0056	0.0079
二　胎	0.0098	0.0229

注：误差定义为$\sum(y_i-\hat{y}_i)^2$

表 2　45岁同期群妇女拟合结果

胎　次	类　型	μ	σ	误　差
一　胎	农　村	2.0063	0.3088	0.0056
	城　市	2.1206	0.3956	0.0056
二　胎	农　村	2.0815	0.3474	0.0027
	城　市	2.0653	0.4810	0.0045

表 3　40岁同期群妇女拟合结果

胎　次	类　型	μ	σ	误　差
一　胎	农　村	2.0242	0.3654	0.0046
	城　市	2.1882	0.3821	0.0064
二　胎	农　村	2.2110	0.3489	0.0092
	城　市	2.1573	0.5295	0.0289

表 4　40岁同期群妇女一胎生育率理论值与实际值的比较

年　龄	农　村		城　市	
	实际值	理论值	实际值	理论值
14—20	0.4099	0.4151	0.2431	0.2630
21—24	0.4438	0.4316	0.4710	0.4454
25—29	0.1126	0.1330	0.2152	0.2285
30—34	0.0141	0.0177	0.0371	0.0506
35—40	0.0024	0.0022	0.0062	0.0106

从拟合结果来看，一、二胎的生育模型基本是成立的。如果比较拟合精度，则一胎优于二胎，农村优于城市。从表4可以看出，理论值在尾部与实际值相比明显呈偏高趋势。

原因是很明显的。对数正态分布是概率分布，各胎次别终生（累计）生育率即便不等于1，至少也应该接近于1，这样才可能有较好的拟合结果。这个问题可以用规范化生育率的方法解决，我们在后面再谈。

二、参数η，σ意的义

评价一个模型的好坏，首要的一点当然要看模型与实际符合的程度。另一方面，模型中参数的意义，能否解释实际中不同现象，也是很重要的。我们来分析一、二胎生育模型中的参数μ，σ，看它们能否解释不同的生育模式。

我们定义ξ为可能生育变量，随生育时间的分布为正态分布，则μ就反映了可能生育变量的中间位置，σ就反映了可能生育变量的集中（离散）程度。μ大（小），ξ的均值就靠后（前）；σ大（小），ξ的分布就分散（集中）。因此，μ，σ最明显的实际意义就是，较大（小）的μ意味着较高（低）的平均生育年龄，较大（小）的σ意味着在平均生育年龄周围较低（高）的生育率或生育密度。不过，这种推断毕竟还比较粗糙。可能生育变量转变为实际生育变量，实际生育变量的分布为对数正态分布。其均值μ_a，标准差σ_a与μ，σ的关系为：

$$\mu_a = e^{\mu+\sigma^2/2}$$

$$\sigma_a^2 = e^{2\mu+\sigma^2}\ (e^{\sigma^2}-1)$$

即μ_a，σ_a既与μ有关，也与σ有关。

μ_a为实际生育变量的均值。它是从起始生育年龄a_0到平均生育年龄的时间。$a_0+\mu_a$即为平均生育年龄的理论值。σ对μ_a的影响可以这样理解：在μ相同的条件下，σ小，可能生育变量的分布集中在μ值周围，μ值附近生育的可能性就大，到达平均生育年龄的时间就短，反之亦然。这是符合中国生育的实际情况的。为了定量的分析μ，σ对μ_a的作用，利用级数展开，μ_a可以表示为

$$\mu_a = e^{\mu}\cdot e^{\sigma^2/2} = \left(1+\mu+\frac{1}{2!}\mu^2+\cdots\cdots\right)\left(1+\frac{\sigma^2}{2}+\frac{1}{2!}\left(\frac{\sigma^2}{2}\right)^2+\cdots\right)$$

由于$\frac{\sigma^2}{2}$取值一般在0.045～0.125之间，所以级数$e^{\sigma^2/2}$从第三项开始即迅速趋于零。亦即μ_a取值的大小基本由μ决定。或者说，μ_a值的92%左右由μ贡献。

σ_a为实际生育变量的标准差。较小（大）的σ_a表示在μ_a周围有较高（低）的生育率或生育密度。μ对σ_a的影响可以这样理解：μ小，可能生育变量的均值就小，若是二胎生育，μ小还意味着胎次间隔较小，而较早的生育时间与较短的生育间隔就会在平均生育年龄周围造成较高的生育密度。这也是符合中国生育的实际情况的。利用级数，σ_a^2可以表示为：

$$\sigma_a^2 = e^{2\mu}\cdot e^{\sigma^2}\ (e^{\sigma^2}-1) = e^{2\mu}\ \left(1+\sigma^2+\frac{1}{2!}\sigma^4+\cdots\cdots\right)\ \left(\sigma^2+\frac{1}{2!}\sigma^4+\cdots\cdots\right)$$

$$= e^{2\mu}\ \left(\sigma^2+\sigma^4\left(1+\frac{1}{2!}\right)+\cdots\cdots\right)$$

从上式可以看出，μ，σ对σ_a的作用都是不容忽视的。也就是说，μ对σ_a有不容忽视的作用。较大（小）的μ能造成σ有较明显的增加（下降）。这个公式的实际意义就在于：它揭示了可能生育变量平均值后移对“稀育”的重要作用。只要我们能有效地控制住“早育”，也就在相当程度上控制住了“密育”。

表 5　40岁同期群妇女一胎生育的三个指标理论值与实际值的比较

指标名称	农村		城市	
	实际值	理论值	实际值	理论值
平均生育年龄(岁)	21.43	22.09	22.80	23.60
生育年龄标准差	2.94	3.06	3.40	3.78
峰值生育年龄(岁)	20.00	20.62	21.00	21.71

生育中还有一个重要指标，即生育的峰值年龄。对数正态分布的极大值在

$$x = e^{\mu-\sigma^2}$$

处达到。则$a_0+e^{\mu-\sigma^2}$可以看作生育峰值年龄的理论值。可能生育变量的均值μ所对应的实际生育年龄a_0+e^{μ}恰好在$a_0+e^{\mu-\sigma^2}$与$a_0+e^{\mu+\sigma^2/2}$之间，亦即在生育峰值年龄与生育平均年龄之间，不过离生育平均年龄更近一些。

三个指标实际值与理论值的比较参见表5。相对误差基本在0.03左右。

可见，参数μ，σ确实可以解释不同的生育模式。这也就从从另一个角度说明了模型的合理性。

三、利用规范化生育率建生育模型

为了把对数正态分布用于高胎次生育率及总累计生育率，很自然的想法是利用规范化生育率，这是科尔、宋健等学者使用过的方法。令$f(x,i)$表示第i胎x岁生育率，令$TFR=\sum_x f(x,i)$，则：

$$h(x,i)=f(x,i)/TRF(i)$$

$h(x,i)$为第i胎x岁规范化生育率。令$f(x)=\sum_i f(x,i)$，$TFR=\sum_i\sum_x f(x,i)$，则

$$h(x)=f(x)/TFR$$

$h(x)$为x岁规范化总累计生育率。

显然，$\sum_x h(x,i)=1$，$\sum_x h(x)=1$

用对数正态分布对规范化生育率进行拟合，则胎次别生育率的理论值为对数正态分布的概率值乘以TFR（i），总累计生育率的理论值为对数正态分布的概率值乘以TFR。

把生育率变成规范化生育率并不改变生育年龄的均值与标准差，或者说，规范化生育率不改变生育的模式。

用对数正态分布对20个同期妇女的规范化生育率进行拟合，拟合结果参见表6—表9，及图1和图2。

表6 规范化生育率模型拟合的平均误差

胎次	农村	城市
总累计	0.0078	0.0070
一胎	0.0027	0.0075
二胎	0.0016	0.0051
三胎	0.0029	0.0038
四胎	0.0051	0.0047
五胎及以上	0.0121	0.0105

表 8 农村40岁同期群妇女生育率理论值与实际值的比较(利用规范化生育率)

年龄	一胎		总累计	
	实际值	理论值	实际值	理论值
14—20	0.4099	0.4160	0.5346	0.4511
21—24	0.4438	0.4349	1.2434	1.3836
25—29	0.1126	0.1175	1.2500	1.1771
30—34	0.0141	0.0126	0.4392	0.4026
35—40	0.0024	0.0011	0.0988	0.1207

表 7 40岁同期群妇女拟合结果
（利用规范化生育率）

胎次	类型	μ	σ	误差
总累计	农村	2.3840	0.3833	0.0039
	城市	2.4135	0.3439	0.0034
一胎	农村	2.0130	0.3475	0.0012
	城市	2.1717	0.3502	0.0009
二胎	农村	2.1944	0.3179	0.0009
	城市	2.0814	0.3909	0.0024
三胎	农村	2.2371	0.3218	0.0011
	城市	2.1816	0.3272	0.0014
四胎	农村	2.1074	0.3748	0.0022
	城市	2.0138	0.4121	0.0042
五胎及以上	农村	2.2842	0.3631	0.0064
	城市	2.0482	0.4606	0.0090

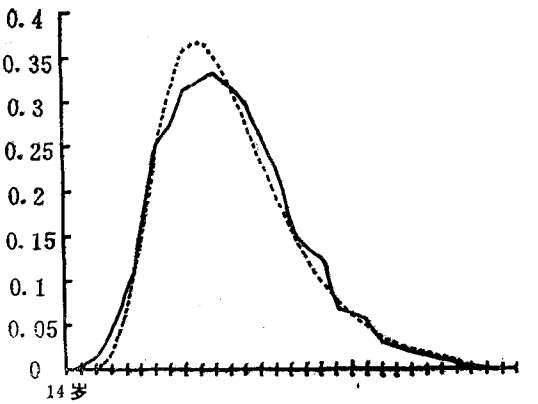

图1 1988年农村40岁同期群妇女总累计生育率曲线(实线为实际值，虚线为理论值)

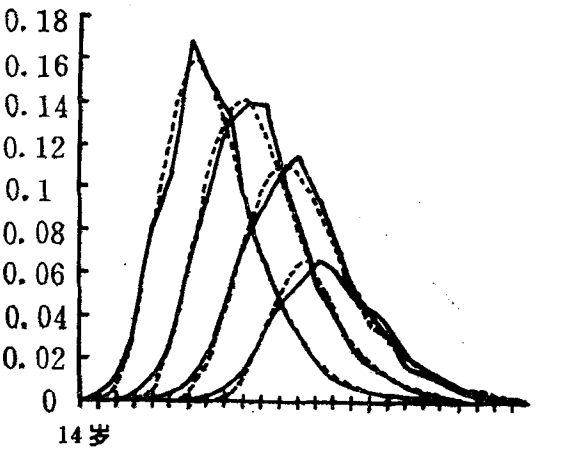

图2 1988年农村40岁同期群妇女胎次别生育率曲线(从左至右分别为1、2、3、4胎)

表 9 40岁同期群妇女一胎生育的三个指标理论值与实际值的比较（规范化生育率）

指标名称	农村		城市	
	实际值	理论值	实际值	理论值
平均生育年龄	21.43	21.59	22.80	23.33
生育年龄标准差	2.94	2.85	3.40	3.37
峰值生育年龄	20.00	20.63	21.00	21.76

从拟合结果来看，规范化生育率的拟合优度大大提高。绝大部分指标的拟合误差都在千分之几。还有12个指标(占总指标的10%)的拟合误差达到万分之几。从表8可以看出，生育率理论值在尾部偏高的趋势得到纠正。从表9与表5的比较可以看出，平均生育年龄、生育年龄标准差的实际值与理论值也更接近。

把规范化总累计生育率看作某个假设的胎次别生育率，则μ，σ值也可以解释不同的生育模式在总累计生育率上的反映。比如从表7可以看出，各胎次别生育率的σ值，农村均小于城市，而总累计生育率的σ值，农村却大于城市。这正是二种生育模式的合乎逻辑的结果。对农村生育模式来说，尽快的完成低胎次的生育，就可能有高胎次的高生育率。而高胎次的高生育率就会造成总累计生育率曲线的尾部偏高，下降缓慢。

规范化生育率的拟合结果说明，生育率与对数正态分布有较好的相似性。这种相似性从图形上可以这样解释：把生育率曲线按同一比例向上拉长或向下压缩，就能较好地接近某个对数正态分布曲线。

以上我们讨论的都是真实一代人的生育率，那么，对数正态分布能否拟合时期指标的规范化总和生育率呢？既然总和生育率是说明假设一代人的生育水平，这假设的一代人至少在

表 10 5个年份总和生育率拟合结果

年份	μ	σ	误差	TFR
1964	2.6031	0.5073	0.0345	6.176
1974	2.5677	0.4642	0.009	4.168
1980	2.4421	0.3860	0.0035	2.308
1987	2.3300	0.4102	0.0005	2.53
1989	2.3165	0.4257	0.0005	2.252

表 11 1987年总和生育率分类型拟合结果

类型	μ	σ	误差
农村	2.3317	0.4212	0.0002
城市	2.1215	0.3698	0.0025

表12　　1987年农村总和生育率

年　龄	实际值	理论值	年　龄	实际值	理论值
15—19	0.1301	0.1273	31	0.0932	0.0901
20	0.1565	0.1664	32	0.0735	0.0719
21	0.2308	0.2345	33	0.0586	0.0572
22	0.2757	0.2782	34	0.0428	0.0452
23	0.3098	0.2941	35	0.0349	0.0358
24	0.2852	0.2868	36	0.0293	0.0282
25	0.2832	0.2642	37	0.0220	0.0220
26	0.2230	0 2333	38	0.0189	0.0173
27	0.1806	0.2002	39	0.0141	0.0138
28	0.1732	0.1676	40	0.0116	0.0109
29	0.1335	0.1380	41—48	0.0377	0.0339
30	0.1164	0.1121			

表13　1989年胎次别生育率拟合结果

胎　次	μ	σ	误　差	TFR(i)
一　胎	2.0793	0.3305	0.0014	1.012
二　胎	2.2932	0.3739	0.0005	0.7148
三　胎	2.4323	0.3530	0.0007	0.3211
四　胎	2.5570	0.3559	0.0008	0.1196
五　胎及以上	2.7677	0.3427	0.0079	0.0826

理论上应该存在。用对数正态分布对1964、1974、1980、1987和1989年五个年份的规范化总和生育率进行拟合，拟合结果见表10，11，12，13。由于1987年农村生育率拟合误差达到万分之二，特将理论值与实际值的详细比较列于表12，以供有兴趣的读者参考。

从拟合结果来看，1964年总和生育率拟合结果较差。似乎应该有这样的结论，如果TFR＞4.5，则对数正态分布生育模型不适用。或者说，对于自然型生育的总和或总累计生育率，找不到与之相似的胎次别生育率。

最后，我们从另一个角度简单说明为什么初育年龄服从对数正态分布。对数正态分布可以描写大量非负独立随机变量的乘积。如果把初育年龄看成是随机变量，则某个妇女恰在某个年龄初育，确实是许多条件同时成立的结果。因为所有与怀孕、分娩有关的因素都要考虑进去。

继宋健等学者提出生育率的伽玛分布模型之后，本文又提出了对数正态分布模型。本文在方法上和对资料的利用上都更加充分，是利用概率分布研究生育模型的一个新成果。关于本模型与其它模型的比较以及概率分布模型的局限性，我们准备在另一篇文章里讨论。

参考文献

(1)黄荣清，亓昕："中国的初婚初育模型研究"，《中国人口科学》1990年第4期。

(2)《人口研究》，1991年第3期。

(3)《中国1990年人口普查10%抽样资料》，中国统计出版社，1992年。

(4)《概率论基础及其应用》，王梓坤著，科学出版社，1976年。

(5)《人口控制论》，宋健著，中国科技出版社，1985年。

（作者工作单位：谢韦克，北京医科大学卫生统计教研室；黄荣清，北京经济学院人口经济研究所）