《中国人口科学》1993年第6期（总第39期）

对概率分布生育模型的探讨

谢韦克

在中国对生育模型的研究中，概率分布有比较广泛的应用。这是由于概率分布生育模型的最大优点，是拟合效果较好，而且构造简单。但是，也还有一些重要的理论问题没有得到充分的讨论。本文准备就这些问题做进一步的探讨。

一、规范化生育率的意义与性质

由于任何概率密度函数与横轴所围面积都为1，而生育率曲线与横轴（年龄轴）所围面积或者大于1（总和或终身生育率），或者小于等于1（胎次别生育率），因此，在对年龄别生育率进行概率分布的曲线拟合时，必须对年龄别生育率进行规范化处理，即用总和或终身生育率去除每个年龄别生育率，以形成规范化年龄别生育率。从图形上看，这相当于对生育率曲线按同一比例压缩或拉伸，以使规范化生育率曲线接近某条概率分布曲线。有些人口学家利用概率分布建立生育模型时虽未明确指出这点，但他们模型的构造也隐含着类似于规范化处理的要求。比如，寇尔（A. J. coal）的已婚生育模型中的参数M，虽然他定义为实际的生育水平与自然生育水平的一个比值，但从数学的角度来看，M也起着对已婚生育率压缩的作用。可见，规范化生育率对于构造概率分布的生育模型起着十分重要的作用。但是对于这个重要概念，无论是其数学意义还是人口统计学意义，至今还没得到很好的解释，而仅仅成了一种数学处理方法。之所以发生这种情况，研究方法是一个重要原因。以往的生育模型研究，大都是针对总和生育率的。总和生育率并不反映真实一代人的生育过程，它是35个同批妇女生育率的横向组合。因此，如果想从中分析规范化生育率的意义也就十分困难。但是，如果我们的分析是针对同期群妇女的生育过程，则规范化生育率的数学意义与人口统计学意义就可以很清楚的揭示出来。

我们以同批妇女的一胎生育为例。为简单起见，假定有100名同期出生的妇女，这100名妇女从15岁到49岁完成她们一生的一胎生育过程。令AFR（x，1）为一胎年龄别生育率，B（x，1）为x岁生育一胎的妇女人数，则有：

$$AFR(x,\ 1)=B(x,\ 1)/100,\ x=15,\ 16\cdots,\ 49 \qquad (1)$$

一胎终身生育率LFR（1）为：

$$LFR(1)=\sum_{x=15}^{49}AFR(x,\ 1)$$

则规范化生育率SAFR（x，1）为：

$$SAFR(x,\ 1)=AFR(x,\ 1)/LFR(1)$$

显然有：

$$\sum_{x=15}^{49}SAFR(x,\ 1)=1$$

如果这100名妇女中有98名妇女生育第一胎，也即LFR（1）=0.98，则对生育率进行规范化处理的结果，就相当于在（1）式中把分母由100转成98，而分子不变。

$$SAFR(x,1)=\frac{AFR(x,1)}{LFR(1)}=\frac{B(x,1)/100}{98/100}=\frac{B(x,1)}{98}。$$

也就是说，规范化生育率只考虑那些生育了一胎的妇女，而不考虑未生育一胎的妇女。此即规范化生育率的人口统计学意义。因为所考虑的妇女都生育一胎，所以规范化生育率就没有生育水平的差别，而只有生育随年龄分布的差别。此即规范化生育率只反映生育模式而不反映生育水平的原因。

另一方面，假定这98名生育一胎的妇女构成一个集合W，$W=(w_1, w_2, \cdots, w_{98})$。令$\xi(w_i)$表示某个妇女初育的年龄，则由概率论知识可以知道，$\xi(w_i)$为一个随机变量。设$\xi(w_i)=x$，则$\xi(w_i)$取某个年龄的概率为：

$$P(x)=B(x,1)/98=SAFR(x,1),$$

而且有：

$$\sum_{x=15}^{49}P(x)=1。$$

因此，当我们对年龄别生育率进行规范化处理的时候，实际上是在构造一个概率空间，以使规范化生育率成为名符其实的概率。此即规范化生育率的数学意义，规范化生育率与年龄别生育率相差一个比例常数，而这比例常数恰为一胎终身生育率。如果终身不育的比例很小，也即一胎终身生育率十分接近于1，则对一胎年龄别生育率直接进行概率分布的曲线拟合也能得到较好的结果。虽然如此，两者在概念上仍有本质的区别。另一方面，如果累计生育率与终身生育率已相差很小，则利用累计生育率进行规范化处理也不会造成太大的误差。以上的分析可推广到其它胎次别的生育情况。

把以上分析用于规范化终身生育率并没原则上的困难。仍假定有100名同期出生的妇女。令B(x，i)为x岁生育第i胎的妇女人数，令LFR(x)为年龄别终身生育率，则有：

$$LFR(x)=\sum_i B(x,i)/100,\quad x=15,16,\cdots 49 \qquad (2)$$

终身生育率LFR为：

$$LFR=\sum_x\sum_i B(x,i)/100,$$

规范化年龄别终身生育率SLFR(x)为：

$$SLFR(x)=LFR(x)/LFR=\sum_i B(x,i)/\sum_x\sum_i B(x,i)。$$

因此，计算SLFR(x)就相当于在(2)式中把分母由100换成$\sum_x\sum_i B(x,i)$，而分子保持不变。也就是说，如果一名妇女一生生育了i胎，在SLFR(x)的计算中就要被看成i个生育了一胎的妇女，而如果一名妇女一生中一次也未生育，就要在计算中被排除出去。因此，规范化终身生育率具有与规范化胎次别生育率完全类似的数学意义与人口统计学意义。

对于时期指标总和生育率来说，虽然我们难以得出与上面类似的结果，但它至少是形式上的概率，而且具有与规范化终身生育率完全相同的两条性质。

1．规范化生育率不改变生育模式

我们知道，生育模式可以由：(1)峰值生育年龄；(2)平均生育年龄；(3)生育年龄的标准差决定。把年龄别生育率变成规范化生育率，上述三个指标并不改变。(1)是显然的，我们只证明(2)。

$$\text{平均生育年龄}=\frac{\sum_{x=15}^{49}(0.5+x)\cdot AFR(x)}{\sum_{x=15}^{49}AFR(x)}$$

$$=\frac{\sum_{x=15}^{49}(0.5+x)\cdot AFR(x)/TFR}{\sum_{x=15}^{49}AFR(x)/TFR}=\frac{\sum_{x=15}^{49}(0.5+x)\cdot SAFR(x)}{\sum_{x=15}^{49}SAFR(x)}$$

式中AFR（x），SAFR（x）分别为年龄别及规范化年龄别生育率。因此，用年龄别生育率和用规范化年龄别生育率计算出的平均生育年龄是一样的。

2. 由于SAFR（x）=AFR(x)/TFR，所以TFR=AFR(x)/SAFR(x)。假定规范化生育率严格服从某一密度函数为f（x）的概率分布，也即假定SAFR（x）$=\int_x^{x+1}f(t)dt$对任何x成立，则有：

$$TFR=AFR(x)\Big/\int_x^{x+1}f(t)dt \qquad (3)$$

当然，规范化生育率严格服从某一概率分布的假定是不符合实际情况的。但是，如果规范化生育率能比较好的接近某条概率分布曲线，则利用（3）式估计总和生育率也能得到较好的结果。

上述两条性质虽然简单，但在生育模型的实际应用中有十分重要的作用。

二、概率分布生育模型中参数的估计问题

如果我们有详尽的生育率数据，则一般用最小二乘法估计模型中的参数。这种估计尽管可以达到很高的精度，有一定的学术价值，但笔者本人并不认为有很大的实际意义。因为即便没有生育模型，我们仍然可以对详尽的生育率数据进行全面、正确的分析。也可以只用少数几个指标如峰值生育年龄、平均生育年龄等反映生育的主要特点。但是，如果缺少某些年龄组的生育率数据，则我们希望能用部分年龄组的生育率数据估计出模型中的参数；再利用生育模型把缺少的生育率数据估计出来。类似的问题在生育率的预测中也存在。比如，如果我们已经知道1992年25岁同批妇女的一胎年龄别生育率，我们想预测这批妇女25岁以后的年龄别生育率，问题同样归结为利用部分年龄组的生育率数据估计模型中的参数。所以，能否利用不完全的数据正确估计出模型中的参数，是判断一个生育模型实际意义大小的重要标准。

在缺少某些年龄组的生育率数据时如何估计模型中的参数，是中国人口学家曾毅首先提出来的。他在研究Brass-Gomperts相关生育模型时，提出了估计模型中参数的“四分位数与中位数解析法”。不过笔者认为，此方法还没完全解决问题。第一，由于没有全部年龄组的生育率数据，就不能计算出总和生育率，也就无法对年龄别生育率进行规范化处理。因此，该方法必须以知道总和生育率为前提，或者假定总和生育率可以用其它方法估计出来。这就增加了问题的难度。况且，如果我们已经知道总和生育率，只要缺少生育率数据的年龄组不是太多，则最小二乘法仍然成立。我们应该研究这样的方法，在利用部分年龄组的生育率数据估计模型中的参数时，可以不以已知总和生育率为前提，而把总和生育率也做为被估计对象。第二，此方法虽然给出了模型中参数与生育率的四分位数与中位数的解析表达式，但是没有给出在缺乏详尽的生育率数据时如何估计四分位数与中位数的方法。比如，如果数据的范围不包括第一个四分位数，那么此方法是否仍然成立？第三，数据的缺少具有程度上的区

别。如果缺少生育率数据的年龄组很少，则用简单的线性内插或线性外推也能得到很好的估计。另一方面，生育年龄组也有重要与不重要的区别。缺少45～49岁的生育率与缺少22～26岁的生育率显然是不一样的。第四，此方法实际上是参数估计的非最小二乘法，因此我们必须分析这种方法所造成的误差，特别是在缺乏详尽的生育率数据时所造成的误差。

由于伽玛分布与对数正态分布中的两个参数与极值点和数学期望之间有明确的解析表达式，比如，以对数正态分布为例，令x_e为数学期望，x_m为极值点，则有：

$$\begin{cases} x_e = e^{\mu+\frac{\sigma^2}{2}} \\ x_m = e^{\mu-\sigma^2} \end{cases}$$

因此有：

$$\begin{cases} \mu = \frac{1}{3} l_n (x_m \cdot x_e^2) \\ \sigma^2 = \frac{2}{3} l_n (x_e / x_m) \end{cases}$$

所以，如果规范化年龄别生育率严格服从伽玛或对数正态分布，则我们可以利用规范化生育率的峰值生育年龄，平均生育年龄以及初始生育年龄求出模型中的参数。由于规模化生育率不改变生育模式（性质1），所以我们可以直接用年龄别生育率求出平均生育年龄与峰值生育年龄，而不必进行规范化处理。同样，对于总和生育率TFR，在规范化生育率严格服从伽玛或对数正态分布的假定下，只要我们能求出密度函数f（x）中的参数，就可以不利用$TFR=\sum AFR(x)$，而利用$TFR=AFR(x)\Big/\int_x^{x+1} f(t)\,dt$来计算。

这种方法的重要意义在于：第一，只要有生育率数据的年龄组包含峰值生育年龄与平均生育年龄，则利用部分年龄组的生育率数据估计模型中的参数就有可能；第二，在利用部分年龄组的生育率数据估计模型中的参数时，可以不以已知总和生育率为前提，而利用性质2把总和生育率也做为被估计对象。

这种方法所要求的条件是：1．年龄别生育率经规范化处理后能比较好的接近某条概率分布曲线。如1989年生育率（见图1、图2）。2．部分年龄组的生育率数据必须准确，而且必须包含峰值生育年龄与平均生育年龄。

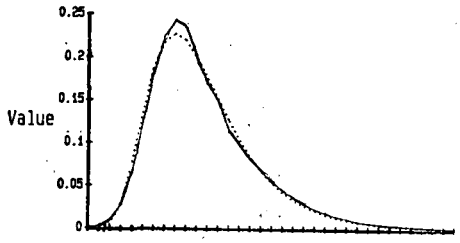

图1 1989年总和生育率、实线为实际值，虚线为理论值（下同）假定规范化生育率服从对数正态分布

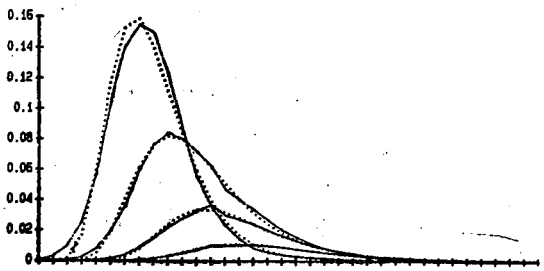

图2 1989年胎次别生育率，从左至右分别为1，2，3，4胎

由于Gomperts分布中的参数与极值点和数学期望之间没有解析表达式。所以上述方法不适用，而只能用“四分位数与中位数”方法。这种方法需要计算第一及第三个四分位数，

所以就需要有更多的年龄组数据。从这一点来说，伽玛与对数正态分布模型要比Brass-Gomperts相关生育模型有更多的优越性。

由于这种方法需要利用密度函数计算规范化生育率的理论值，而伽玛分布中伽玛函数的计算十分复杂。所以，如果从计算的简单性来说，对数正态分布又比伽玛分布具有更多的优越性。

假定1989年规范化生育率服从对数正态分布，我们用三种不同的方法估计模型中的参数。并利用（3）式估计生育水平（见表1～表3）。表中$\bar{T}$表示比值$AFR(x)\Big/\int_x^{x+1}f(t)\,dt$的均数。数据来源为第四次人口普查10%抽样资料。

表1　　最小二乘法拟合结果

胎次＼指标	μ	σ	a_0	x_m	x_a	$\bar{T}$	TFR	$\lvert TFR-\bar{T}\rvert$
总和	2.3186	0.4249	15	23.483	26.121	2.246	2.252	0.006
一胎	2.0793	0.3305	15	22.171	23.448	1.028	1.012	0.016
二胎	2.3924	0.3399	15	24.785	26.567	0.701	0.715	0.014
三胎	2.5177	0.3272	16	27.141	29.081	0.327	0.321	0.006
四胎	2.6333	0.3327	17	29.461	31.711	0.119	0.120	0.001

注：a_0表示起始生育年龄；x_m表示峰值生育年龄；x_a表示平均生育年龄。

表2　　非最小二乘法拟合结果（利用全部生育率数据）

胎次＼指标	μ	σ	a_0	x_m	x_a	$\bar{T}$	TFR	$\lvert TFR-\bar{T}\rvert$
总和	2.3182	0.4221	15	23.5	26.104	2.239	2.252	0.013
一胎	2.0920	0.2776	15	22.5	23.417	0.917	1.012	0.095
二胎	2.3825	0.3622	15	24.5	26.566	0.749	0.715	0.034
三胎	2.5157	0.2708	16	27.5	28.837	0.171	0.321	0.050
四胎	2.6311	0.3247	17	29.5	31.641	0.117	0.120	0.003

表3　　非最小二乘法拟合结果（利用部分生育率数据）

胎次＼指标	μ	σ	a_0	x_m	x_a	$\bar{T}$	TFR	$\lvert TFR-\bar{T}\rvert$	范围（岁）
总和	2.3183	0.4222	15	23.5	26.106	2.240	2.252	0.012	22～32
一胎	2.1011	0.2936	15	22.5	23.536	0.933	1.012	0.079	21～26
二胎	2.3827	0.3624	15	24.5	26.569	0.750	0.715	0.035	23～29
三胎	2.5255	0.2883	16	27.5	29.127	0.296	0.321	0.025	26～32
四胎	2.6277	0.3193	17	29.5	31.557	0.116	0.120	0.004	28～35

从表中可以看出，对于总和生育率，二胎及四胎生育率，非最小二乘法的结果还是比较好的。而对于一胎及三胎生育率，误差则比较大。究其原因，主要是峰值生育年龄的估计误

差所致。峰值生育年龄的估计虽然简单，但对它的估计实际上只利用了一个年龄组的数据，因此就不如平均生育年龄的估计值准确和稳定。我们可以考虑改用生育年龄的标准差估计模型中的参数，并对两种方法的误差进行比较。但这样也仍没解决问题，因为这种方法缺乏一种能调整参数的机制。表2及表3虽然说明了“解”的存在，具有一定的意义，但并不表示已经找到求解的方法。因为这些解是利用最小二乘法的结果“凑”出来的。比如，以1989年总和生育率为例。如果我们有18～32岁年龄组生育率，我们怎么知道用22～32岁年龄组的生育率数据计算平均生育年龄比较准确，又如何判断a_0为15岁呢？因此，我们必须提出一个标准，利用这个标准可以修正模型中的参数。笔者考虑，可以提出比值的方差最小原则，也即确定概率密度函数f（x；θ_1，θ_2）中的参数θ_1，θ_2，使比值

$T(x;\ \theta_1,\ \theta_2)=AFR(x)\Big/\int_x^{x+1}f(t;\ \theta_1,\ \theta_2)dt$的方差最小，实际上，若

$S(\theta_1,\ \theta_2)=\sum_x(SAFR(x)-\int_x^{x+1}f(t;\ \theta_1,\theta_2)dt)^2$越小，则每个比值

$ST(x;\ \theta_1,\ \theta_2)=SAFR(x)\Big/\int_x^{x+1}f(t;\ \theta_1,\ \theta_2)dt$ 越接近于1，ST（x；θ_1，θ_2）的方差也越小。由于AFR（x）与SAFR(x)只相差一个比例常数TFR。所以当ST(x；θ_1，θ_2）的方差小时，T（x；θ_1，θ_2）的方差也小；如果$S(\theta_1,\ \theta_2)=0$，则$T(x;\ \theta_1,\ \theta_2)$=TFR，也即T（x；$\theta_1$，$\theta_2$）的方差为0。反之亦然。因此，当比值的方差最小时，T（x；θ_1，θ_2）总的来说应该与TFR最接近。我们可以采用比值的均数（当然也可以考虑采用其它指标）做为TFR的估计值。所以，如果这个原则成立，则利用部分年龄组的生育率数据估计模型中的参数能否达到较好的效果，就取决于S（θ_1，θ_2）能否充分小。因此，最小二乘法的拟合优度就是判断生育模型好坏的重要标准。虽然我们不能在这里比较三个模型的拟合优度，但Brass-Gomperts相关生育模型的建立需要借助一个标准模式，而这个标准又随不同的生育模式而变化，这就给提高拟合优度造成了困难。因此，这就再一次说明伽玛与对数正态分布模型要比Brass-Gomperts相关生育模型有更多的优越性。

顺便说一句，不少文章把最小二乘法中的误差平方和写成：

$$\sum_x(SAFR(x)-f(x))^2$$

实际上是有问题的。因为SAFR（x）表示的是年龄区间（x，x+1）内的规范化生育率，而不是某一年龄点x上的生育率。f（x）应改为$\int_x^{x+1}f(t)dt$比较合适。

对于上面提出的问题，我们可以这样解决。首先假定初始生育年龄$a_0=15$，这可以由以往的生育资料并结合数据本身的特点推断，再假定峰值生育年龄$a_m=23.5$，这也可以由数据本身推断。因此，困难的只是如何确定平均生育年龄。一个简单而直观的解决办法是，由不同的年龄组范围计算出“平均生育年龄”，从这些“平均生育年龄”中找出使比值

$T(x;\ \theta_1,\ \theta_2)=AFR(x)\Big/\int_x^{x+1}f(t;\ \theta_1,\ \theta_2)dt$的方差最小的一个，以它做为平均生育年龄的估计值（参见表4）。

从表4可以看出，平均生育年龄取26.106时V（T）最小，这与最小二乘法的结果基本一致。但是比值的均数$\bar{T}$却是在平均生育年龄取26.682时与TFR的实际值2.252最接近，这说明利用$\bar{T}$估计TFR有时还不是最好的。另一方面，我们也可以考虑用传统的求极值的方法确

表4　　不同的年龄组范围计算出的平均生育年龄的比较

年龄组范围（岁）	平均生育年龄	μ	σ	V(T)	$\bar{T}$	$\lvert\bar{T}-TFR\rvert$
21～32	25.643	2.2898	0.3869	0.0415	2.213	0.039
22～32	26.106	2.3182	0.4222	0.0166	2.214	0.038
23～32	26.682	2.3521	0.4604	0.0653	2.270	0.018

注：V(T)表示$T(x;\theta_1,\theta_2)$的方差。

定参数的精确解。但是这要涉及到许多理论与计算问题，还必须就实际的生育率数据进行充分的比较与验证，由于篇幅所限，对于比值的方差最小原则问题我们在另外的文章里专门讨论。

本文讨论了规范化生育率的数学意义与人口统计学意义。并在曾毅研究的启发下进一步讨论了利用部分年龄组的生育率数据估计模型中的参数问题。利用规范化生育率的性质提出了参数估计的新的计算方法。

主要参考文献

1. 宋健，于景元著：《人口控制论》科学出版社，1985年。
2. 曾毅等："中国女性婚后离家模型"，《中国人口科学》1991. No. 1.
3. 谢韦克、黄荣清："中国妇女生育模型研究"，《人口与经济》1993. No. 1.

（本文责任编辑：朱　萍）

（作者工作单位：北京医科大学卫生与人口统计教研室）

（上接第44页）

（三）中国妇女的文化教育水平较低，整体文化程度只相当于小学四年级水平，它不仅直接影响到妇女的个人素质的高低，而且对下一代的素质也会产生影响。在中国家庭中，多数是妇女承担着教育和培养子女的重任，妇女本身素质过低，必须影响到下一代素质的提高，也不利于妇女地位提高的世代传递。

（四）尽管妇女的劳动地位指数很高，但是受个人素质的影响，妇女就业多集中于劳动密集型产业，而在技术密集型的高层次职业中占比例很低。近两年来妇女就业受到市场经济的强烈冲击，企业减员首先考虑的是女性，优化组合优化掉的也多是女性，女性就业难的问题日趋严重。时代对女性提出了更高的要求。

（五）中国妇女微观家庭地位相当高，而社会参与很低，这说明中国妇女还处于传统型向现代型的转变过程中。要实现这种转变，一方面要提高妇女自身的素质；另一方面要实现家务劳动社会化，把妇女从家务劳动中解放出来，使她们走向社会。

参考文献：

1. 冯立天主编《中国人口生活质量研究》，北京经济学院出版社，1992年9月。
2. 《中国妇女社会地位调查初步分析报告》，《妇女研究论丛》，1992年创刊号。

（本文责任编辑：朱　犁）

（作者工作单位：北京经济学院人口经济研究所）