极大似然估计

“极大似然估计法”是寻找未知参数估计量的最常用方法。在介绍这一方法的一般形式之前，我们先推导一个简单的二项分布的极大似然估计量。

例 二项分布

假设我们抛掷一枚有偏硬币 10 次。有偏的意思是对于这枚硬币，出现“正面”的概率 p 不一定是 $1/2$。令 X 为 10 次抛掷中出现“正面”的次数。那么随机变量 X 服从参数为 10 和未知的 $p \in [0,1]$ 的二项分布。假设我们得到了 3 次正面朝上的结果，则这一结果的概率为：

$$P_p(X=3) = \binom{10}{3} p^3 (1-p)^7$$

概率 p 是未知的，需要进行估计。

那么我们这时候问:**哪个 p 值最有可能呢?

下图展示了 $P_p(X=3)$ 作为 p 的函数的变化情况。我们可以看到，存在一个唯一的 p 值使得该概率最大，即 0.3。这个 p 值赋予了观察结果“3 次正面朝上”最大的可能性。在这种情况下，估计量 $\hat{p} = 0.3$ 就是极大似然估计量。

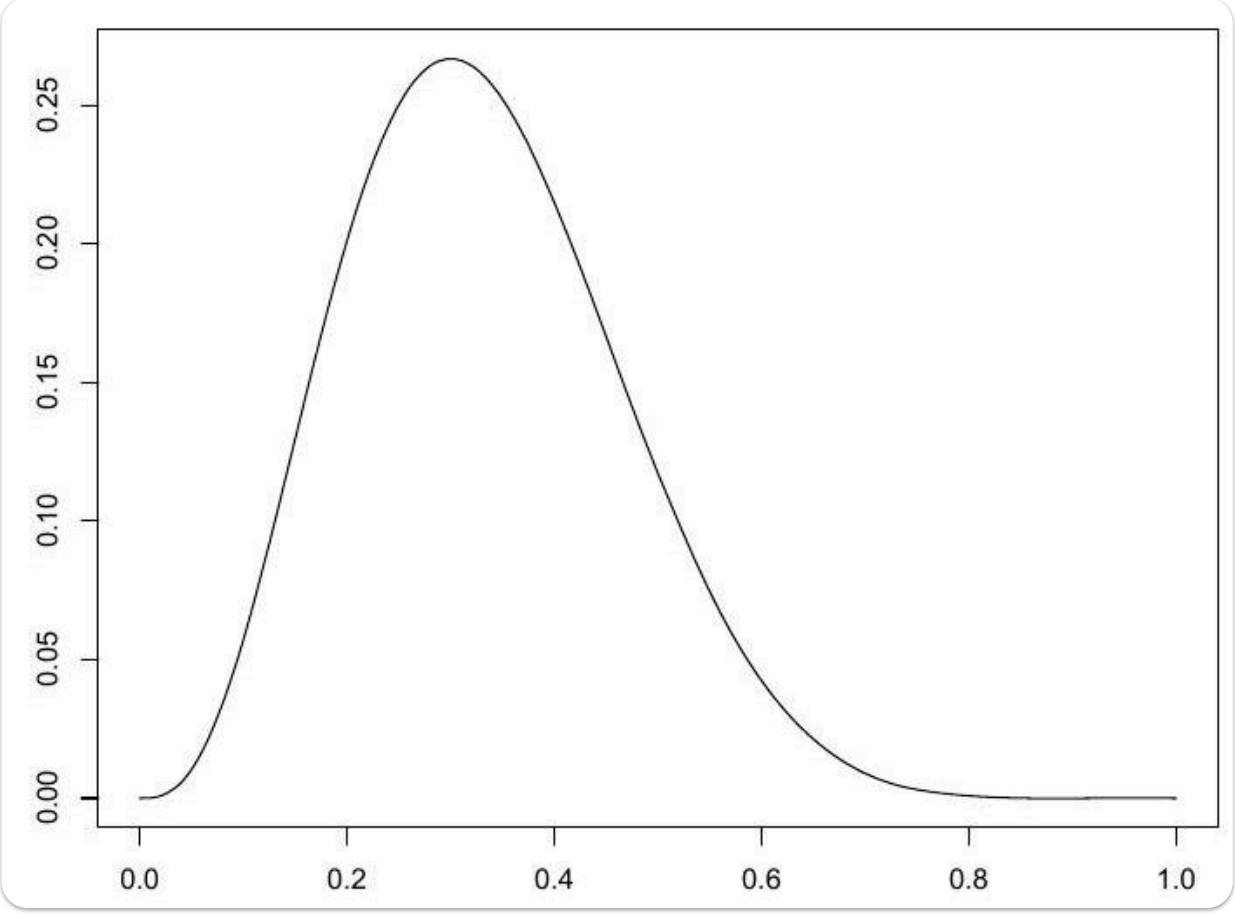


图:. $P_p(X=3)$ 随 p 变化的函数，其中随机变量 X 服从参数为 10 和 p 的二项分布。

极大似然方法需要一个由观察值的概率密度推导出的似然函数。对于随机变量 X 的概率密度 p_θ，我们指的是函数 $x \mapsto P_\theta(X=x)$（若 X 是离散的），或函数 p_θ 满足 $\mathrm{P}_\theta(X \in B) = \int_B p_\theta(x)dx$

（若 X 是连续的）。

似然函数

令 X 为具有依赖于参数 $\theta \in \Theta$ 的概率密度 p_θ 的随机向量。对于固定的 x，将函数

$$\theta \mapsto L(\theta; x) := p_\theta(x)$$

视为 $\theta \in \Theta$ 的函数（其中 Θ 是参数空间），称为似然函数。

通常，$X = (X_1, \ldots, X_n)$ 是一个具有独立同分布坐标 X_i 的向量。X 在 $(x_1, \ldots, x_n)$ 上的密度等于边际概率密度的乘积 $\prod_{i=1}^n p_\theta(x_i)$，似然函数则为：

$$\theta \mapsto L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^n p_\theta(x_i)$$

其中，p_θ 是一个 X_i 的（边际）密度。然而，似然函数的通用定义也适用于元素不独立或不同分布的观察向量。这里要注意哪怕是 $i.i.d.$ 样本也是要写成乘积, 因为考察的是整个随机向量的密度函数. 因此，我们更愿意将观察记为 x，而不是 $(x_1, \ldots, x_n)$，并将似然函数写为 $L(\theta; x) \equiv p_\theta(x)$。

极大似然估计和估计量

θ 的极大似然估计是使似然函数 $\theta \mapsto L(\theta; x)$ 达到最大值的 $T(x) \in \Theta$。极大似然估计量是相应的估计量 $T(X)$。

对于离散概率分布，极大似然估计可以描述为赋予观测值 x 最大概率的参数值。实际上，我们通过先固定的 x 然后最大化概率密度 $p_\theta(x) = P_\theta(X = x)$ (视为 θ 的函数) 来进行估计。

直观上，这一原则是合理的。虽然名字听起来不错, 叫做极大似然, 也就是说"最有可能"，但**极大似然估计量**并不一定是**最好的估计量**。所谓“最好的”估计量，是指均方误差最小的估计量, 这一点的定义需要把握住.

对于给定模型，计算极大似然估计量主要依赖于微积分工具。通常，我们对似然函数求导并将导数设为 0 然后找最大值。

为了减少不必要的计算量（尤其是在处理独立观测时），一个技巧是首先对似然函数取对数。

由于**对数是单调函数**，$\hat{\theta}$ 使得 $\theta \mapsto L(\theta; x)$ 达到最大值的条件与 $\theta \mapsto \log L(\theta, x)$ 达到最大值的条件是等价的（这里我们讨论的是使达到最大值处的参数值，而不是最大值本身！）。对于固定的 x，对数似然函数为：

$$\theta \mapsto \log L(\theta; x) = \log p_\theta(x)$$

如果 L 在 $\theta \in \Theta \subset \mathbb{R}^k$ 上可微分，并且在 Θ 的**内点**达到最大值，那么：

$$\frac{\partial}{\partial \theta_j} \log L(\theta; x)_{|\theta=\hat{\theta}} = 0, \quad j = 1, \ldots, k$$

这个方程组被称为似然方程组(注意是对数似然函数对各个参数的分量求偏导,得到的是个方程组)，但它不总是可以显式求解。如果必要的话，可以使用优化方法来近似求解。

但是从微积分工具的角度来说,方程组的解不仅仅是极大值，极小值和拐点也是似然方程的解。为了验证某个解是否确实是最大值，我们需要仔细考虑（对数）似然函数的形式。

一种方法是求解对数似然函数在该解处的二阶导数（或如果参数是多维的，则为 Hessian 矩阵）。如果函数在该点有极大(极小)值，二阶导数在该点将是负(正)的。对于高维参数，Hessian 矩阵的所有特征值都必须为负(正)。

如果观测 $X = (X_1, \ldots, X_n)$ 由独立同分布的子观测 X_i 组成，那么观测 x 的似然函数 $L(\theta; x)$ 是一个乘积：

$$L(\theta; x) = \prod_i p_\theta(x_i)$$

其中 p_θ 是某个 X_i 的（边际）密度。对数似然函数则为：

$$\theta \mapsto \log L(\theta; x_1, \ldots, x_n) = \log \prod_{i=1}^{n} p_\theta(x_i) = \sum_{i=1}^{n} \log p_\theta(x_i)$$

对数似然函数的导数，即所谓的**得分函数**，是各个观测得分函数的和。因此，似然方程为：

$$\sum_{i=1}^{n} \frac{\partial}{\partial \theta_j} \log p_\theta(x_i) = 0, \quad j = 1, \ldots, k$$

例 指数分布

令 $X = (X_1, \ldots, X_n)$ 为来自参数 $\lambda > 0$ 的指数分布的样本。则其观测值 $x_1, \ldots, x_n$ 的对数似然函数为：

$$\lambda \mapsto \log L(\lambda; x_1, \ldots, x_n) = \log \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = n \log \lambda - \lambda \sum_{i=1}^{n} x_i$$

参数空间为 $\lambda \in (0, \infty)$。对对数似然函数关于 λ 求导并令其等于 0，得到：

$$\frac{d}{d\lambda} \log L(\lambda; x_1, \ldots, x_n)_{|\lambda=\hat{\lambda}} = \frac{n}{\hat{\lambda}} - \sum_{i=1}^{n} x_i = 0$$

其解为 $\hat{\lambda} = 1/\bar{x}$。对数似然函数的二阶导数为：

$$\frac{d^2}{d\lambda^2} \log L(\lambda; x_1, \ldots, x_n) = -\frac{n}{\lambda^2}$$

它在所有 $\lambda > 0$ 时为负，因此似然函数在 $\hat{\lambda}$ 处确实有一个最大值。参数 λ 的极大似然估计量为 $\hat{\lambda} = 1/\bar{X}$。

例 二项分布

变量 X 定义为掷硬币 10 次中出现正面朝上的次数。X 服从参数为 10 且未知概率 p 的二项分布。观察到的值是 $x = 3$。对数似然函数等于以下函数：

$$\begin{aligned} p \mapsto \log L(p; x=3) &= \log\left(\binom{10}{3} p^3 (1-p)^7\right) \\ &= \log\binom{10}{3} + 3\log p + 7\log(1-p) \end{aligned}$$

最大似然估计量 $\hat{p}$ 是使该函数在 $[0,1]$ 上最大化的 p 值。
得到的解为 $\hat{p} = 0.3$。

对于一般的二项分布 X，其参数为 n 和 p，对数似然函数为：

$$p \mapsto \log L(p; x) = \log\binom{n}{x} + x\log p + (n-x)\log(1-p)$$

如果 $0 < x < n$，那么当 $p \downarrow 0$ 或 $p \uparrow 1$ 时，$\log L(p; x) \to -\infty$，因此对数似然函数在区间 $(0,1)$ 内取其最大值。由此可知，似然函数 $L(p; x)$ 也在 $(0,1)$ 内取其最大值。将对 p 的导数设为 0，得到一个解 $\hat{p} = x/n$。因此该解为最大似然估计量，$\hat{p} = x/n$。

如果 $x = 0$ 或 $x = n$，则 $L(p; x)$ 在 0 或 1 处有局部最大值。在这些情况下，最大似然估计量仍然可以写为 $\hat{p} = x/n$。因此，最大似然估计量为 $\hat{p} = X/n$。

例 正态分布

对于来自 $N(\mu, \sigma^2)$ 分布的样本 $X = (X_1, \ldots, X_n)$，对数似然函数为：

$$\begin{aligned} (\mu, \sigma^2) &\mapsto \log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(x_i-\mu)^2/\sigma^2} \\ &= -\frac{1}{2} n \log 2\pi - \frac{1}{2} n \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \end{aligned}$$

对于参数 $\theta = (\mu, \sigma^2)$，我们选择的自然参数空间(确定参数空间是非常重要的)是 $\Theta = \mathbb{R} \times (0, \infty)$。对数似然函数对 μ 和 σ^2 的偏导数为：

$$\begin{aligned} \frac{\partial}{\partial \mu} \log L\left(\mu, \sigma^2; x_1, \ldots, x_n\right) &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) \\ \frac{\partial}{\partial \sigma^2} \log L\left(\mu, \sigma^2; x_1, \ldots, x_n\right) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 \end{aligned}$$

将第一个方程设为 0，得到解 $\hat{\mu} = \bar{x}$。对于每个 $\sigma^2 > 0$，对数似然函数在 $\mu = \hat{\mu}$ 处确实有全局最大值，因为当 $\mu \to \pm\infty$ 时，对数似然值趋于 $-\infty$。

接着，我们将 $\mu = \hat{\mu}$ 代入第二个偏导数，将其设为 0，解出 σ^2 的似然方程，得到解 $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$。由于相同的原因，对数似然函数在该值处有最大值。（请注意，如果用关于 σ 的导数最大化对数似然函数而不是关于 σ^2 的导数，会得到 $\hat{\sigma}^2$ 的平方根作为 σ 的最大似然估计量。）

为了验证该（可微）对数似然函数在我们找到的似然方程解处是否有最大值，我们还可以求解对数似然函数在点 $(\hat{\mu}, \hat{\sigma}^2)$ 处的 Hessian 矩阵，该矩阵在此情况下为：

$$\frac{1}{\hat{\sigma}^4} \begin{pmatrix} -n\hat{\sigma}^2 & 0 \\ 0 & -n/2 \end{pmatrix}$$

该矩阵的两个特征值都是负的，因此对数似然函数在点 $\left(\hat{\mu}, \hat{\sigma}^2\right)$ 处有一个最大值。

对于参数 $\left(\mu, \sigma^2\right)$，最终的最大似然估计量等于

$$\left(\bar{X}, \frac{1}{n} \sum_{i=1}^n \left(X_i-\bar{X}\right)^2\right)=\left(\bar{X}, \frac{n-1}{n} S_X^2\right)$$

其中

$$S_X^2=\frac{1}{n-1} \sum_{i=1}^n \left(X_i-\bar{X}\right)^2$$

样本均值是 μ 的无偏估计量，但求出的最大似然估计量 $\hat{\sigma}^2$ 存在轻微的偏差(bias)。

由于这个小偏差，样本方差 $S_X^2=(n /(n-1)) \hat{\sigma}^2$ 通常是优选的。

然而，S_X^2 的均方误差大于 $\hat{\sigma}^2$ 的均方误差，不过谈及到底谁的均方误差最小方面，使用 $((n-1) /(n+1)) S_X^2$ 反而还要优于这两者。因为对于较大的样本数，差异很小，所以使用哪个估计量并不重要。

如果我们假设 μ 是已知的，可以得到另一个模型。此时参数为 $\theta=\sigma^2$，参数空间为 $(0, \infty)$。然后我们发现 σ^2 的最大似然估计量等于 $n^{-1} \sum_{i=1}^n \left(X_i-\mu\right)^2$。但是请注意，只有在假设 μ 已知的情况下，这才是一个估计量, 否则这个统计量含有参数!

如果（对数）似然函数的最大值不在参数空间的内部取得，那么最大似然估计 $\hat{\theta}$ 通常不是似然函数导数的驻点，而是一个局部最大值，且似然方程不成立。在其他一些例子中，似然函数并不是处处可微（甚至不连续），最大似然估计也不满足似然方程。下一个例子 阐述了这种情况。此外，似然函数可能有多个（局部）最大值和最小值。此时，似然方程可能有多个解。最大似然估计根据定义是似然函数的全局最大值。

例 均匀分布

设 $x=\left(x_1, \ldots, x_n\right)$ 为来自区间 $[0, \theta]$ 的均匀分布的观测样本，其中 $\theta>0$ 为未知数。我们希望使用最大似然估计量来估计参数 θ。由于观测值 $x_1, \ldots, x_n$ 落在区间 $[0, \theta]$ 内，因此必须有 $\theta \geq x_i$ 对于 $i=1, \ldots, n$。立刻可以推导出 $\theta \geq x_{(n)}$，其中 $x_{(n)}$ 是观测到的最大顺序统计量。

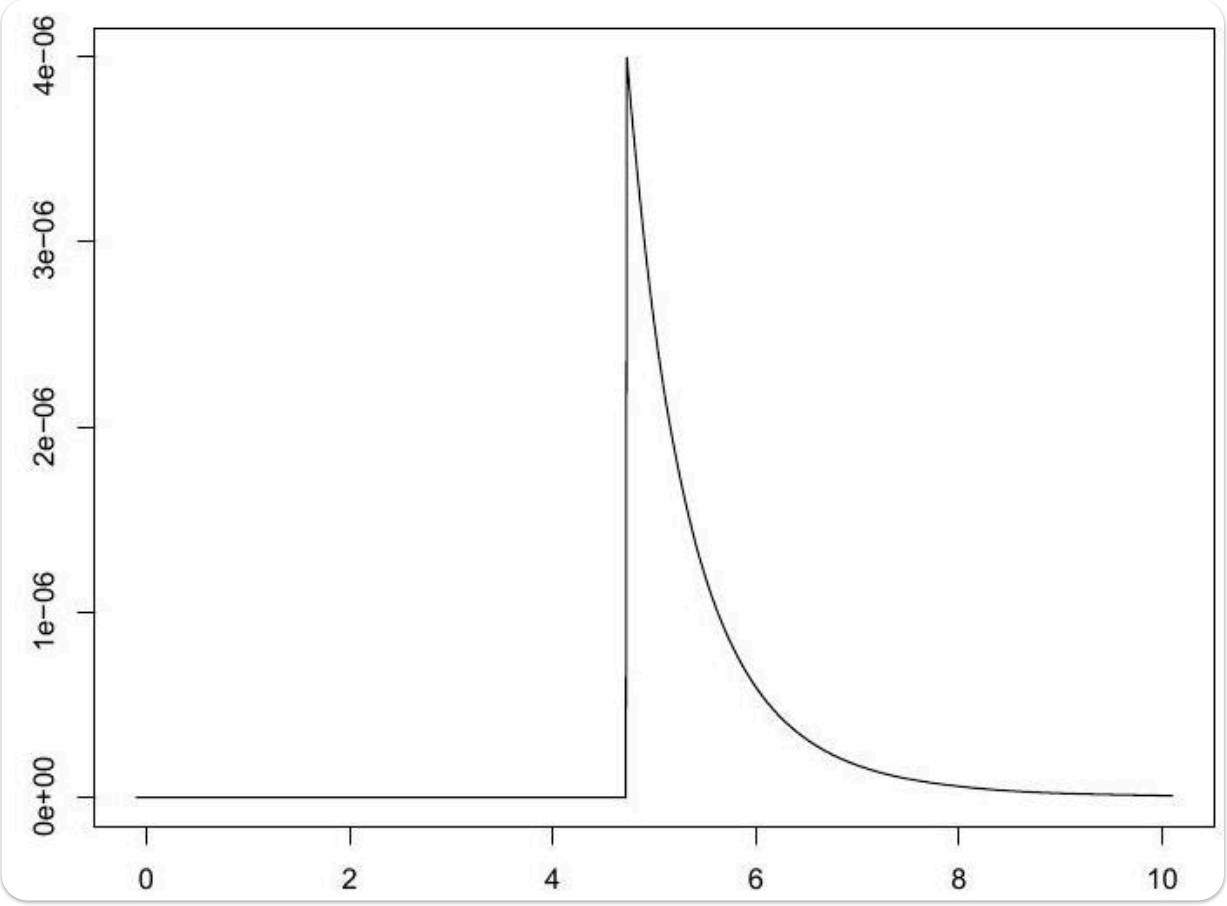
观测值 $x_1, \ldots, x_n$ 的似然函数等于 $X_1, \ldots, X_n$ 在 $x_1, \ldots, x_n$ 上的联合密度，视为 θ 的函数。由于 $X_1, \ldots, X_n$ 是独立同分布的，联合密度等于边际密度的乘积，该密度在区间 $[0, \theta]$ 上为 $1 / \theta$，在其他地方为 0。因此，似然函数为

$$\theta \mapsto L\left(\theta ; x_1, \ldots, x_n\right)=\prod_{i=1}^n \frac{1}{\theta} 1_{0 \leq x_i \leq \theta}=\left(\frac{1}{\theta}\right)^n 1_{x_{(1)} \geq 0} 1_{x_{(n)} \leq \theta}$$

第二个等号的处理很关键, 记住均匀分布要"收住"次序统计量的最大和最小值.

我们先把 $X_{(1)}$ 放到 0 之后, 之前都是 0, 没有意义, 现在我们来考虑 $X_{(n)}$, 这才是比较关键的统计量. 该 θ 的函数在 $\theta<x_{(n)}$ 时等于 0，因为此时指示函数 $1_{x_{(n)} \leq \theta}$ 等于 0。在 $\theta=x_{(n)}$ 时，该函数跳变至 $1 / \theta^n$。在 $\theta=x_{(n)}$ 处，似然函数以及对数似然函数相对于 θ 是不可微的。可以通过绘制似然函数关于 θ 的图像找到最大值。当 $\theta \geq x_{(n)}$ 时，似然函数等于递减函数 $\theta \rightarrow 1 / \theta^n$。下图展示了似然函

数的变化情况（关于 θ）。在 $x_{(n)}$ 处，似然函数是上半连续的，并且也是最大值；因此，θ 的最大似然估计量等于 $x_{(n)}$，相应的最大似然估计量是 $X_{(n)}$。



图来自区间 $[0, 5]$ 的均匀分布样本，样本量为 8 的似然函数的图像。最大似然估计量 $x_{(n)}$（尖峰位置）为 4.73。

例 限制条件下的正态分布

假设观测值 $X_1, \ldots, X_n$ 独立且服从均值为 μ，方差为 1 的正态分布，且我们知道 $\mu \geq 0$。对于 $X_1, \ldots, X_n$ 的一次观测 $x_1, \ldots, x_n$，在实数域上，似然函数在 $\bar{x}$ 处取得绝对最大值。然而，$\bar{x}$ 可能为负数，并且 $\mu \geq 0$，因此 $\bar{x}$ 不是最大似然估计量, 因为显然偏出了参数空间。如果 $\bar{x} \leq 0$，则在参数空间 $[0, \infty)$ 上，似然函数在 0 处取得局部最大值。最大似然估计量是当 $\bar{x} \geq 0$ 时为 $\bar{x}$，否则为 0。相应的最大似然估计量是 $\max(0, \bar{X})$。

从以上我们可以清楚地看到统计模型和最大似然估计量不仅由观测值的密度形式决定，也由参数空间的定义决定！

如果 $g : \Theta \to H$ 是一个双射的函数，我们可以用参数 $\eta = g(\theta) \in H$ 替代 $\theta \in \Theta$ 来对模型进行参数化。根据定义可以立即得出，如果 $\hat{\theta}$ 是 θ 的最大似然估计量，那么 $g(\hat{\theta})$ 是 η 的最大似然估计量。因此，对于任意函数 g，我们定义 $g(\theta)$ 的最大似然估计量为 $g(\hat{\theta})$。（该估计量最大化了轮廓似然函数 $L_g(\tau; x) = \sup_{\theta \in \Theta : g(\theta) = \tau} p_\theta(x)$, 之后会更加具体的提到）

在之前定义中，最大似然估计量基于最大似然估计值。在实际操作中，常常直接以随机变量 X 的形式书写（对数）似然函数，而不是以观测值 x 表示，并且直接通过对 θ 求最大值来推导估计量。这种简化的表示法在接下来的最大似然法应用示例中使用。之后也会进一步的有关于将此方法应用于回归模型的示例。

例 指数分布，续例

设 $X=(X_1,\dots,X_n)$ 是来自参数 $\lambda>0$ 的指数分布的样本。在之前中，我们已经证明 λ 的最大似然估计为 $\hat{\lambda}=1/\bar{X}$。由此我们可以很容易地推导出 $\mathrm{E}_\theta X_i=1/\lambda$ 的最大似然估计，即 $\widehat{\mathrm{E}_\theta X_i}=1/\hat{\lambda}=\bar{X}$。

例 伽马分布

设 $X=(X_1,\dots,X_n)$ 是来自伽马分布的样本，其概率密度函数为

$$p_{\alpha,\lambda}(x)=\frac{x^{\alpha-1}\lambda^\alpha e^{-\lambda x}}{\Gamma(\alpha)}$$

其中，$\alpha>0$ 和 $\lambda>0$ 是未知的形状参数和逆尺度参数，Γ 是伽马函数：

$$\Gamma(\alpha)=\int_0^\infty s^{\alpha-1}e^{-s}ds$$

$X_1,\dots,X_n$ 的对数似然函数为

$$\begin{aligned}(\alpha,\lambda)&\mapsto\log\prod_{i=1}^n\frac{X_i^{\alpha-1}\lambda^\alpha e^{-\lambda X_i}}{\Gamma(\alpha)}\\&=(\alpha-1)\sum_{i=1}^n\log X_i+n\alpha\log\lambda-\lambda\sum_{i=1}^n X_i-n\log\Gamma(\alpha)\end{aligned}$$

作为参数空间 $\theta=(\alpha,\lambda)$，我们取 $\Theta=[0,\infty)\times[0,\infty)$。为了确定 α 和 λ 的最大似然估计，我们求出对数似然函数相对于 λ 和 α 的偏导数：

$$\begin{aligned}\frac{\partial}{\partial\lambda}\log L\left(\alpha,\lambda;X_1,\dots,X_n\right)&=\frac{n\alpha}{\lambda}-\sum_{i=1}^n X_i\\\frac{\partial}{\partial\alpha}\log L\left(\alpha,\lambda;X_1,\dots,X_n\right)&=\sum_{i=1}^n\log X_i+n\log\lambda-n\frac{\int_0^\infty s^{\alpha-1}\log se^{-s}ds}{\int_0^\infty s^{\alpha-1}e^{-s}ds}\end{aligned}$$

（在相对于 α 的偏导数中，我们在积分号内对伽马函数 $\alpha\mapsto\Gamma(\alpha)$ 进行了微分，并使用了 $(\partial/\partial\alpha)s^\alpha=s^\alpha\log s$。）偏导数在最大似然估计 $(\hat{\alpha},\hat{\lambda})$ 处等于 0；这给出了两个似然方程。从第一个方程立即可得 $\hat{\lambda}=\hat{\alpha}/\bar{X}$。要注意这里使用的方法是, 把他作为结论(如正态分布时候一样)将其代入第二个似然方程(解方程组的常见方法)，得到

$$\sum_{i=1}^n\log X_i+n\log\hat{\alpha}-n\log\bar{X}-n\frac{\int_0^\infty s^{\hat{\alpha}-1}\log se^{-s}ds}{\int_0^\infty s^{\hat{\alpha}-1}e^{-s}ds}=0$$

此方程没有 $\hat{\alpha}$ 的显式解，但可以通过迭代方法在观测到的 $X_1,\dots,X_n$ 样本上进行数值求解。对于大多数数值算法，我们需要初始值作为求解方程的起始点。可以使用矩估计方法的估计值作为初始值。

我们将所得值 $\hat{\alpha}$ 代入方程 $\hat{\lambda}=\hat{\alpha}/\bar{X}$ 来确定 $\hat{\lambda}$。为了验证对数似然函数在解处是否取得极大值，我们必须计算在 $(\hat{\alpha},\hat{\lambda})$ 处的 Hessian 矩阵的特征值。如果两个特征值在 $(\hat{\alpha},\hat{\lambda})$ 处都为负，则 $(\hat{\alpha},\hat{\lambda})$ 为 (α,λ) 的最大似然估计。

例 应用: 计数细菌

在被污染的水中，细菌的数量无法用肉眼或显微镜直接计数。为了了解污染的程度，我们估计每分升水中细菌的集落形成单位（colony-forming units, CFU）数量。我们采用以下方法。假设被污染水中每分升的细菌集落形成单位数量服从参数为 μ 的泊松分布。为了估计水中的细菌集落形成单位数量，我们需要估计 μ。我们将被污染的水倒入 100 升的纯水桶中，充分混合，然后将水分成 100 个容量为 1 升的培养皿。接着我们检查每个培养皿是否形成集落。如果形成集落，则说明该分升水中至少有一个细菌集落形成单位；如果没有形成集落，则说明该分升水中没有细菌。令 X 表示被污染水中每分升的细菌集落形成单位数量，则可以表示为 $X = \sum_{i=1}^{100} X_i$，其中 X_i 表示第 i 个培养皿中的细菌集落形成单位数量。变量 $X_1, \ldots, X_{100}$ 是独立的，并服从参数为 $\mu/100$ 的泊松分布(这里要注意泊松分布的可加性, 同时要领会其中参数 μ 代表的含义)。

然而，我们无法直接观察 $X_1, \ldots, X_{100}$。我们实际观测到的是 $Y_1, \ldots, Y_{100}$，其中 Y_i 定义为

$$Y_i = \begin{cases} 0 & \text{如果第 } i \text{ 个培养皿中没有形成集落} \\ 1 & \text{否则} \end{cases}$$

观测量 Y_i 是独立的，并服从以下伯努利分布：

$$P(Y_i = 0) = P(X_i = 0) = e^{-\mu/100} \quad \text{且} \quad P(Y_i = 1) = 1 - e^{-\mu/100}$$

定义 $p := P(Y_i = 1) = 1 - e^{-\mu/100}$。伯努利分布参数 p 的最大似然估计可以通过列出似然方程并解得 p 来简单推导出来。基于样本 $Y_1, \ldots, Y_{100}$，该估计量为 $\hat{p} = \sum_{i=1}^{100} Y_i/100$。由于 $p = 1 - e^{-\mu/100}$，参数 μ 可以表示为 $-100\log(1-p)$，因此 μ 的最大似然估计量为

$$\hat{\mu} = -100 \log\left(1 - \sum_{i=1}^{100} Y_i/100\right)$$

例 自回归

最大似然法不仅限于独立观测值。我们通过自回归模型来展示这一点，该模型常用于分析随时间变化的变量：

$$X_i = \beta X_{i-1} + e_i$$

其中，β 是未知参数，$e_1, \ldots, e_n$ 是不可观测的随机波动，也被称为 "噪声" 。该模型与无截距的线性回归模型非常相似，只是这里的观测值 X_i 是通过 X_{i-1} 进行回归解释的。如果我们将 $i \in \{1, \ldots, n\}$ 看作表示时间的连续时刻，那么回归发生在 X_i 与序列过去的 X_{i-1} 之间，这就是 "自回归" 一词的来源。这里我们讨论的是一阶自回归(也就是说这里讨论的是相隔序数差为 1 的自回归, 普遍形式可以写为 $i, i+h$)模型；从一阶的方法出发, 去研究回归到过去多个变量的扩展形式也是显而易见的。

在这个模型中，数据点的顺序至关重要，将数据描绘成时间的函数是有益的。下图展示了向量 $(X_0, X_1, \ldots, X_n)$ 的三个可能的实现，以 i 为横轴，x_i 为纵轴绘制。所有三个实现都以 $x_0 = 1$ 开始，但此后根据模型 $X_i = \beta X_{i-1} + e_i$，通过独立的噪声生成，且 β 的值相同。统计问题是根据观察到的实现 $(x_0, x_1, \ldots, x_n)$ 来估计 β 的值。我们将使用最大似然法来解决这个问题。

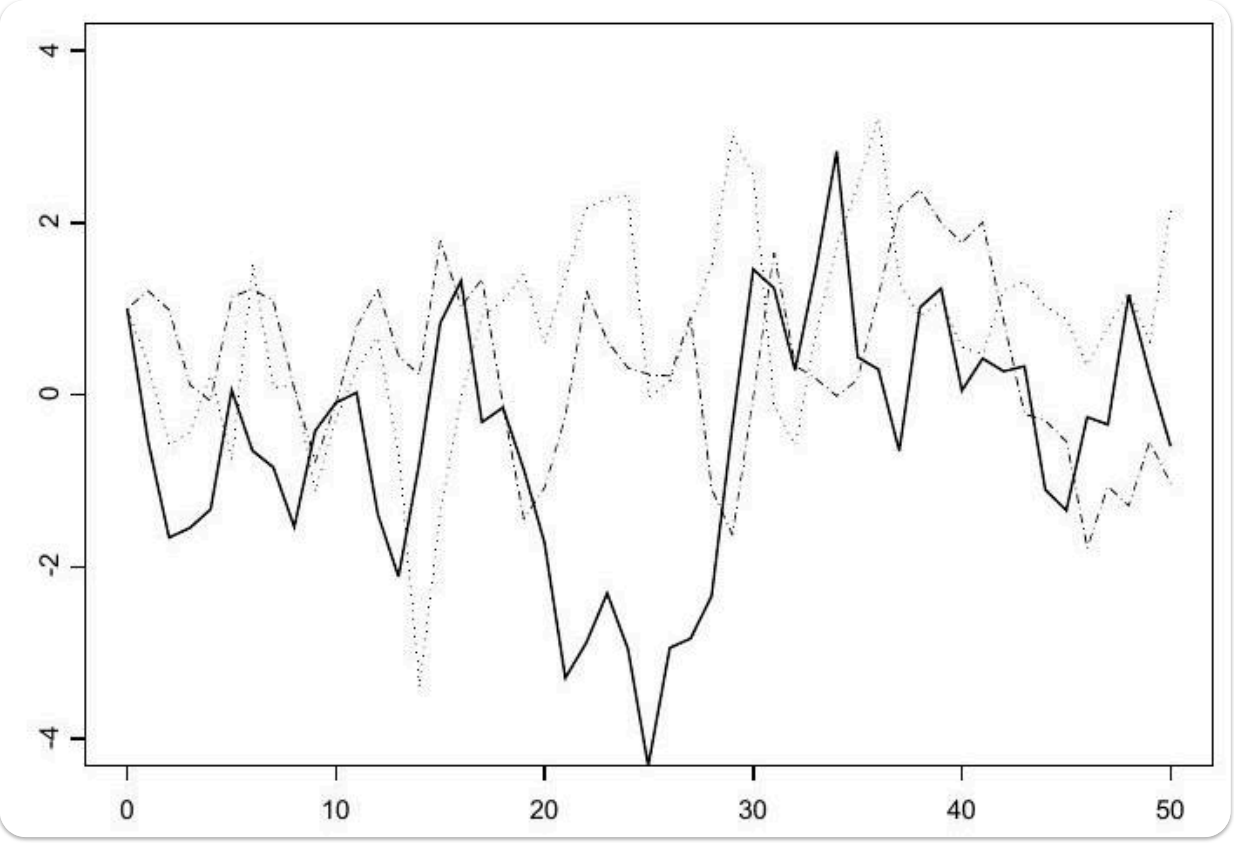


图: 根据自回归模型生成的向量 $(X_0, X_1, \ldots, X_{50})$ 的三个实现，具有标准正态噪声，$x_0 = 1$ 且 $\beta = 0.7$。每个图都是点 $\{(i, x_i) : i = 0, \ldots, 50\}$ 的线性插值。

我们通过假设 X_0 服从概率密度 p^{X_0} 来完成模型的描述，并假设噪声 $e_1, \ldots, e_n$ 是独立且服从正态分布 $N\left(0, \sigma^2\right)$ 的随机变量，并且它们与 X_0 独立。似然函数是观测向量 $X = (X_0, \ldots, X_n)$ 的联合概率密度。由于观测值 $X_0, X_1, \ldots, X_n$ 是相关的(虽然是一种带有随机的相关)，因此联合密度不是边缘密度的乘积。

然而，我们可以使用联合密度的一般分解公式(这里用到了条件概率的重要性质, 并且要注意第一项是 x_0 因为要注意这是起始项, 而后面的例如 x_n 是依赖于之前的, 所以我们才能写出有意义的"条件概率")：

$$
\begin{aligned}
&p^{X_0,\ldots,X_n}\left(x_0, \ldots, x_n\right) = p^{X_0}\left(x_0\right) p^{X_1|X_0}\left(x_1 \mid x_0\right) p^{X_2|X_0,X_1}\left(x_2 \mid x_0, x_1\right) \times \\
&\cdots \times p^{X_n|X_0,\ldots,X_{n-1}}\left(x_n \mid x_0, \ldots, x_{n-1}\right)
\end{aligned}
$$

这个公式将联合密度分解为条件密度的乘积，推广了独立观测值情况下的乘积公式。通过反复应用公式 $f^{X,Y}(x, y) = f^X(x) f^{Y|X}(y \mid x)$，可以证明这一公式。

在自回归模型中，已知 $X_0 = x_0, \ldots, X_{i-1} = x_{i-1}$ 的情况下，X_i 的条件密度等于 $\beta x_{i-1} + e_i$ 的密度，即具有期望 βx_{i-1} 和方差 $\operatorname{var} e_i = \sigma^2$ 的正态分布的密度。因此，似然函数的形式为：

$$
(\beta, \sigma) \mapsto L\left(\beta, \sigma; X_0, \ldots, X_n\right) = p^{X_0}\left(X_0\right) \prod_{i=1}^{n} \frac{1}{\sigma} \phi\left(\frac{X_i - \beta X_{i-1}}{\sigma}\right)
$$

我们尚未指定 X_0 的密度。由于该密度仅影响 $n+1$ 个因子中的一个，而 n 通常很大，因此相关的因子 $p^{X_0}\left(X_0\right)$ 被省略，分析是在 "给定 X_0 的值的条件下" 进行的。

使用这种定义的似然函数，可以通过与线性回归模型相同的计算方法（参见之后的章节）确定参数 (β, σ) 的（条件）最大似然估计。

最大似然估计量 $\hat{\beta}$ 最小化平方和 $\beta \mapsto \sum_{i=1}^{n} (X_i - \beta X_{i-1})^2$，并且等于

$$\hat{\beta} = \frac{\sum_{i=1}^{n} X_i X_{i-1}}{\sum_{i=1}^{n} X_{i-1}^2}$$

σ^2 的最大似然估计量为

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \hat{\beta} X_{i-1} \right)^2$$

根据初始观测值 X_0 的建模，基于无条件似然函数的最大似然估计量可能会采取略微不同的形式。

关于上一个例子的补充证明以及一些说明

在给定的自回归模型中：

$$X_i = \beta X_{i-1} + e_i, \quad i = 1, 2, \ldots, n,$$

其中误差项 e_i 独立且服从正态分布 $N(0, \sigma^2)$，并且与初始值 X_0 独立。由于 e_i 是正态分布的随机变量，且线性变换保持正态性，因此 X_i 条件于 X_{i-1} 也是正态分布的。

1. 条件概率密度为何为正态分布：

由于：

$$X_i = \beta X_{i-1} + e_i,$$

且已知 X_{i-1}，因此 X_i 的条件分布为，：

$$X_i \mid X_{i-1} \sim N\left(\beta X_{i-1}, \sigma^2\right).$$

这是因为 e_i 的分布为 $N(0, \sigma^2)$，将 X_{i-1} 视为已知量所以说是以此为条件，βX_{i-1} 为确定的均值，故 X_i 条件于 X_{i-1} 服从均值为 βX_{i-1}、方差为 σ^2 的正态分布。

2. 构建似然函数：

观测向量为 $X = (X_0, X_1, \ldots, X_n)$。由于 X_i 条件于 X_{i-1} 的条件密度已知，联合概率密度可以分解为：

$$p^{X_0, X_1, \ldots, X_n}(x_0, x_1, \ldots, x_n) = p^{X_0}(x_0) \prod_{i=1}^{n} p^{X_i \mid X_{i-1}}(x_i \mid x_{i-1}).$$

其中，$p^{X_i \mid X_{i-1}}(x_i \mid x_{i-1})$ 是条件密度函数，具体为：

$$p^{X_i \mid X_{i-1}}(x_i \mid x_{i-1}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \beta x_{i-1})^2}{2\sigma^2} \right).$$

因此，似然函数（忽略初始值 X_0 的密度）为：

$$L(\beta, \sigma; X_0, X_1, \ldots, X_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_i - \beta X_{i-1})^2}{2\sigma^2} \right).$$

为了简化计算，通常取对数似然函数：

$$\ell(\beta,\sigma)=-\frac{n}{2}\ln(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\beta X_{i-1})^2.$$

3. 最大似然估计的推导：

（a）估计 β：

对于正态分布来说最大化对数似然函数相当于最小化平方和(这一点在之后的回归分析时候会提到)：

$$S(\beta)=\sum_{i=1}^{n}(X_i-\beta X_{i-1})^2.$$

对 β 求导并令导数为零：

$$\frac{\partial S(\beta)}{\partial\beta}=-2\sum_{i=1}^{n}X_{i-1}(X_i-\beta X_{i-1})=0.$$

解得：

$$\hat{\beta}=\frac{\sum_{i=1}^{n}X_iX_{i-1}}{\sum_{i=1}^{n}X_{i-1}^2}.$$

（b）估计 σ^2：

将 $\hat{\beta}$ 代入平方和：

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\hat{\beta}X_{i-1})^2.$$

这是因为对于正态分布，最大似然估计量是残差平方和除以样本量 n。

4. 关于初始值 X_0 的处理：

初始值 X_0 的密度 $p^{X_0}(x_0)$ 未被指定，但由于 n 较大，且 X_0 仅影响似然函数中的一个因子，故在估计 β 和 σ^2 时，可忽略 $p^{X_0}(x_0)$，相当于在给定 X_0 的条件下进行估计。

例 应用：复合泊松过程

一家健康保险公司为其客户和医疗服务提供者报销已产生的医疗费用。每月初，公司希望估算该月需要预留多少资金以支付所有批准的索赔。为此，整理了一个包含过去 120 个月所有支付记录的数据集。

批准的索赔数量因月份而异，且取决于健康保险公司在该月的客户数量。我们定义 N_i 为第 i 个月批准的索赔数量，并假设 $N_1,\ldots,N_{120}$ 是独立的随机量，且满足

$$N_i\sim\text{Poisson}\,(\mu M_i),\quad i=1,\ldots,120$$

其中 $\mu>0$ 是一个未知参数，M_i 是公司在第 i 个月初的客户数量。假设 M_i 是已知的且不是随机的。

我们用 $C_{i,j}$ 表示第 i 个月第 j 个索赔的金额。第 i 个月的支付金额等于 $\sum_{j=1}^{N_i} C_{i,j}$。我们假设已支付索赔的金额是独立的随机变量，满足

$$C_{i,j} \sim \exp(\theta), \quad i = 1, \ldots, 120, j = 1, \ldots, N_i$$

其中 $\theta > 0$ 是一个未知参数。我们还假设索赔金额 $C_{i,j}$ 与索赔数量 N_i 相互独立。

根据上述模型的假设，可以确定下个月的预期支出。如果已知下个月的索赔数量为 n，那么预期支出为

$$\mathrm{E}_\theta \sum_{j=1}^{n} C_j = \frac{n}{\theta}$$

其中 $C_1, \ldots, C_n$ 是下个月获批的索赔金额。然而，总的索赔数量是未知的，且服从 $\mathrm{Poisson}(\mu M)$ 分布，M 是下个月的客户数量。此时，预期支出为

$$\mathrm{E}_{\mu,\theta} \left(\sum_{j=1}^{N} C_j \right) = \mathrm{E}_\mu \left(\mathrm{E}_\theta \left(\sum_{j=1}^{N} C_j \mid N \right) \right) = \mathrm{E}_\mu \left(\frac{N}{\theta} \right) = \frac{\mu M}{\theta}$$

在这个表达式中，我们首先计算给定 N 时 $\sum_{j=1}^{N} C_j$ 的期望，得到 N/θ，然后再取 N/θ 的期望。因此，当 θ 和 μ 已知时，下个月的预期支出为 $\mu M/\theta$。

参数 $\mu > 0$ 和 $\theta > 0$ 是未知的，需要通过数据来估计。我们使用最大似然法。为了推导似然函数，我们首先确定一个月内观测到的 $(C_1, \ldots, C_N, N)$ 的联合密度。我们将这个密度记为 $f_{\theta,\mu}$，其表达式为

$$\begin{aligned} f_{\theta,\mu}\left(c_1, \ldots, c_N, N = n\right) &= f_{\theta,\mu}\left(c_1, \ldots, c_n \mid N = n\right) P_\mu(N = n) \\ &= \left(\prod_{j=1}^{n} \theta e^{-\theta c_j} \right) e^{-\mu M} \frac{(\mu M)^n}{n!} \end{aligned}$$

我们假设不同月份和年份的观测值是独立的。过去十年数据集中的所有观测值的对数似然函数等于各个月份联合概率密度的乘积的对数：

$$\begin{aligned} (\mu, \theta) &\mapsto \log \left(\prod_{i=1}^{120} \left(\prod_{j=1}^{N_i} \theta e^{-\theta C_{i,j}} \right) e^{-\mu M_i} \frac{(\mu M_i)^{N_i}}{N_i!} \right) \\ &= \sum_{i=1}^{120} \log \left(\prod_{j=1}^{N_i} \theta e^{-\theta C_{i,j}} \right) + \sum_{i=1}^{120} \log \left(e^{-\mu M_i} \frac{(\mu M_i)^{N_i}}{N_i!} \right) \end{aligned}$$

第一项不依赖于参数 μ，第二项不包含参数 θ。因此，为了确定 θ 和 μ 的最大似然估计量，只需分别对第一项关于 θ 和第二项关于 μ 进行最大化即可。结果为

$$\hat{\theta} = \frac{\sum_{i=1}^{120} N_i}{\sum_{i=1}^{120} \sum_{j=1}^{N_i} C_{i,j}} \quad \text{和} \quad \hat{\mu} = \frac{\sum_{i=1}^{120} N_i}{\sum_{i=1}^{120} M_i}$$

支付金额的最大似然估计为

$$M \frac{\hat{\mu}}{\hat{\theta}} = M \frac{\sum_{i=1}^{120} \sum_{j=1}^{N_i} C_{i,j}}{\sum_{i=1}^{120} M_i}$$

在此示例中，我们假设参数 μ 和 θ 在每个月和每一年都是相同的。这样的假设其实是有问题的。事实上，由于通货膨胀，平均支付金额会增加，并且冬季的索赔数量通常比夏季多。因此，值得考虑让参数依赖于年份和月份。我们可以使用 12 个参数 $\mu_1, \ldots, \mu_{12}$ 来表示不同月份的参数。然而，很显然的是增加模型中的未知参数数量会降低估计的精度。