参数点估计与最大似然估计

Dezeming Family

2021 年 7 月 16 日

DezemingFamily 系列书和小册子因为是电子书，所以可以很方便地进行修改和重新发布。如果您获得了 DezemingFamily 的系列书，可以从我们的网站 [https://dezeming.top/] 找到最新版。对书的内容建议和出现的错误欢迎在网站留言。

20210718：完成第一版。暂时预留了两个疑问，有效性证明中我用了两种方法，但一种方式是错的，我没有列出错误方法和错误原因；相合性中的证明我认为资料中的证明方法有些冗杂，故暂时不贴出。

目录

一 参数估计

很多时候，我们已经知道总体的样本分布形式，但我们并不知道其中的一些参数。比如我们知道人的身高分布呈现某种正态分布，但我们不知道其方差和均值。

参数估计就是利用一些样本来去估计总体（见 DezemingFamily 的《样本估计》）的参数，比如我们想估计某个参数 θ，我们设 Θ 是参数空间，也就是 θ 能够取的所有值的集合。**参数点估计**就是根据样本 $X_1, X_2, ..., X_n$ 来估计 θ 的值。

点估计一般常用矩估计法和最大似然估计法。

二 矩估计法

所谓矩估计，就是使用样本的 k 阶原点矩来作为总体 k 阶原点矩，从而估计未知参数。

例如我们使用矩估计来估计方差，我们先得到样本期望和样本方差（见 DezemingFamily 的《样本估计》）：

$$\mu = E(X) = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{二.1}$$

$$\mu^2 + \sigma^2 = E(X^2) = \frac{1}{n}\sum_{i=1}^{\infty} X_i^2 = \frac{1}{n}\sum_{i=1}^{\infty}(X_i - \overline{X})^2 + \overline{X}^2 = \frac{n-1}{n}S^2 + \overline{X}^2 \tag{二.2}$$

得到方差的估计为：

$$\hat{\sigma}^2 = \frac{n-1}{n}S^2 \tag{二.3}$$

注意：

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{二.4}$$

可以看到，矩估计方差其实就是“使用样本作为总体”来估计总体方差。

三　最大似然估计法

同矩估计一样，我们已经有了一些样本数据，我们也知道总体的分布规律，但是有些参数并不知道，我们希望用这些样本来去估计总体分布的这些参数（在后面的例题有更直观的认识）。

最大似然估计的思路可以描述如下：

- 确定需要估计的参数 $\theta = (\theta_1, \theta_2, ..., \theta_m)$，设参数空间为 Θ。
- 确定样本集的 n 个样本，设样本为 $x_1, x_2, ..., x_n$。
- 在参数空间 Θ 中确定一个 $\hat{\theta}$，使得出现样本观测结果 $X_1 = x_1, X_2 = x_2, ..., X_n = x_n$ 的概率 $L(\theta)$ 最大。

我们称 $L(\theta)$ 为样本的似然函数。

因为样本之间都是独立的，对于离散随机变量而言：

$$L(\theta) = P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) \tag{三.1}$$

$$= \prod_{i=1}^{n} P(X_i = x_i) = \prod_{i=1}^{n} p(x_i, \theta) \tag{三.2}$$

对于连续随机变量而言，设在某参数 θ 下的概率密度函数为 $f(x, \theta)$，因为恰好落在某一值的概率为 0，因此我们考察的是样本 X_i 落在 $x_i < X_i < x_i + dx_i$，的概率，这个概率近似等于 $f(x_i, \theta)$，因此 $L(\theta)$ 就可以表示为：

$$L(\theta) = P(x_1 < X_1 < x_1 + dx_1, x_2 < X_2 < x_2 + dx_2, ..., x_n < X_n < x_n + dx_n) \tag{三.3}$$

$$= \prod_{i=1}^{n} f(x_i, \theta) \tag{三.4}$$

我们对似然函数求导时，为了求导方便（后面示例会看到）会直接使用对数似然函数，即 $\ln L(\theta)$，它们在同一个 θ 上达到最大值。当参数 θ 有多个时，就会构建最大似然方程组：

$$\frac{\partial \ln L(\theta)}{\partial \theta_i} = 0 \qquad i = 1, 2, ..., m \tag{三.5}$$

确认极值点以后，还要确定这个极值点是否是函数最大值点。

四　最大似然估计示例

4 1　示例一：分别使用矩估计和最大似然估计

设总体的概率密度为:

$$f(x) = \begin{cases} (\theta+1)x^{\theta}, & 0 < x < 1 \\ 0, & otherwise \end{cases} \tag{四.1}$$

其中，$\theta > -1$，$X_1, X_2, ..., X_n$ 是一组样本。

矩估计求 θ

如果计算一下 $F(x)$ 就可以知道无论 θ 取值如何，$F(1)$ 的值都是 1，因此说明这是个合理的概率密度函数。首先计算期望：

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_{0}^{1} x(\theta+1)x^{\theta}dx = \frac{\theta+1}{\theta+2} \tag{四.2}$$

期望的矩估计为 $\frac{1}{n}\sum_{i=1}^{n} x_i = \overline{X}$，因此可以得到矩估计量：

$$\frac{\theta+1}{\theta+2} = \overline{X} \tag{四.3}$$

求解得到 θ 值即可。

最大似然估计求 θ

$L(\theta)$ 的计算：

$$L(\theta) = \prod_{i=1}^{n}[(\theta+1)X_i^{\theta}] = (\theta+1)^n\Big(\prod_{i=1}^{n} X_i\Big)^{\theta} \tag{四.4}$$

两边同取对数，并求导：

$$\ln L(\theta) = n\ln(\theta+1) + \theta\sum_{i=1}^{n}\ln X_i \tag{四.5}$$

$$\frac{d\ln L(\theta)}{d\theta} = \frac{n}{\theta+1} + \sum_{i=1}^{n}\ln X_i = 0 \tag{四.6}$$

解得：

$$\hat{\theta} = -\frac{n}{\sum_{i=1}^{n}\ln X_i} - 1 \tag{四.7}$$

$$\left.\frac{d^2\ln L(\theta)}{d\theta^2}\right|_{\theta=\hat{\theta}} = -\frac{\Big(\sum_{i=1}^{n}\ln X_i\Big)^2}{n} < 0 \tag{四.8}$$

由上可知，$\ln L(\theta)$ 在 $\hat{\theta}$ 处取得最大值，因此 $\hat{\theta}$ 是 θ 的最大似然估计。我们可以看到，在这个例子中，最大似然估计和矩估计量并不一样。

4 2　示例二：正态分布最大似然估计

设总体服从正态分布：

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{四.9}$$

其中 μ 和 σ^2 都是未知参数，我们有一组观测值 $X_1, X_2, ..., X_n$，求最大似然估计。

首先先求出似然函数以及对数函数：

$$L(\mu,\sigma^2) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(X_i-\mu)^2}{2\sigma^2}} \tag{四.10}$$

$$\ln L(\mu,\sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 \tag{四.11}$$

然后分别对两个参数求导数：

$$\frac{\partial \ln L(\mu,\sigma^2)}{\partial \mu} = \frac{\sum_{i=1}^{n}(X_i-\mu)}{\sigma^2} = 0 \tag{四.12}$$

$$\frac{\partial \ln L(\mu,\sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{2\sigma^4} = 0 \tag{四.13}$$

解得：

$$\hat{\mu} = \overline{X} \tag{四.14}$$

$$\hat{\sigma}^2 = \frac{n-1}{n}S^2 \tag{四.15}$$

五　估计量的优劣评价

我们使用不同的估计方法就可以得到不同的参数估计值，但是如何评估我们估计的好坏呢？

5 1　无偏性

我们计算一下期望的期望值：

$$E(\hat{\mu}) = E(\overline{X}) = \mu \tag{五.1}$$

说明期望的估计是总体期望的无偏估计。

我们再计算一下前面正态分布估计方差的估计式的期望（在《样本估计》中推导过，这里给出更简单的推导过程）：

$$E(\hat{\sigma}^2) = E(\frac{n-1}{n}S^2) = E[\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2] \tag{五.2}$$

$$= \frac{1}{n}\sum_{i=1}^{n}E(X_i^2) - E(\overline{X}^2) \tag{五.3}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\mu^2+\sigma^2) - (\mu^2+\frac{\sigma^2}{n}) = \frac{n-1}{n}\sigma^2 \tag{五.4}$$

关于上面 $E(\overline{X}^2)$ 的计算方法，可以用 $Var(\overline{X}) + E(\overline{X})^2$ 来得到，$Var(\overline{X})$ 的计算方式可以参考《样本估计》。

可以看到，最大似然估计来得到的方差并不是实际的总体方差，但是当样本量 n 趋近于无穷的时候，最大似然估计得到的估计值就会与总体方差相同，因此这个估计叫做有偏估计中的渐进无偏估计。

注意，只要总体均值存在，样本均值就是总体均值的无偏估计；总体方差存在，则样本方差就是总体方差的无偏估计。关于样本均值和样本方差的计算方法可以参考《样本估计》。

5 2　有效性

无偏性只是基本要求，但其实只要对任意常数 $a_1, a_2, ..., a_n$ 满足 $\sum_{i=1}^{n} a_i = 1$，则就能得到 $\sum_{i=1}^{n} a_i X_i$ 是总体均值的无偏估计，但是这样很明显不准确。

当我们需要估计一个量 θ 时，我们得到了两个无偏估计 θ_1 和 θ_2，我们计算这两个估计方法的方差 $D(\theta_1)$ 和 $D(\theta_2)$，方差小的，则认为其有效性更好（详细解释一下，我们选用不同的估计方法得到无偏估计，但某种方法估计的方差比较大，也就是说我每次取不同的样本估计 100 次，估计结果更发散，则说明估计效果并不会很好）。

我们以上面的均值为例，$E(\sum_{i=1}^{n} a_i X_i) = \mu$：

$$D(\theta) = E(\sum_{i=1}^{n} a_i X_i - \mu)^2 \tag{五.5}$$
$$= \frac{1}{n^2} E(na_1 X_1 - na_1\mu + na_2 X_2 - na_2\mu + ... + na_n X_n - na_n\mu)^2 \tag{五.6}$$
$$= \frac{1}{n^2} E(na_1(X_1 - \mu) + na_2(X_2 - \mu) + ... + na_n(X_n - \mu))^2 \tag{五.7}$$
$$= \frac{1}{n^2} E(\sum_{i=1}^{n} n^2 a_i^2 (X_i - \mu)^2 + \sum_{i \neq j} n^2 a_i a_j (X_i - \mu)(X_j - \mu)) \tag{五.8}$$

因为样本之间是独立的，所以协方差为 0，即 $E((X_i - \mu)(X_j - \mu)) = 0$，也就是说上式可以化简为：

$$D(\theta) = \frac{1}{n^2} E(\sum_{i=1}^{n} n^2 a_i^2 (X_i - \mu)^2) \tag{五.9}$$
$$= \sum_{i=1}^{n} a_i^2 E((X_i - \mu)^2) \tag{五.10}$$
$$= \sum_{i=1}^{n} a_i^2 D(X) \tag{五.11}$$

这里的 a_i 越平均，得到的平方和就越小，我这里给出一个不完美的简单理解性证明：我们假设只有两个样本，权重分别设为 $(0.1, 0.9)$ 以及 $(0.5, 0.5)$，那么得到结果：

$$0.1 * 0.1 + 0.9 * 0.9 = 0.82 \tag{五.12}$$
$$0.5 * 0.5 + 0.5 * 0.5 = 0.5 \tag{五.13}$$

也就是说，$a_i = \frac{1}{n}$ 时会得到最有效的总体均值无偏估计。

5 3 相合性（一致性）

若某组参数估计 $\hat{\theta} = (\theta_1, ..., \theta_m)$ 是参数 θ 的一致性估计，则对于任意 $\epsilon > 0$，都有：

$$\lim_{n \to \infty} P(|\hat{\theta} - \theta| \geq \epsilon) = 0 \tag{五.14}$$

其实就是随着样本量逐渐增加，估计值和真实值越来越接近，而且可以证明得到，若 $\lim_{n\to\infty} E(\hat{\theta}) = \theta$，$\lim_{n\to\infty} D(\hat{\theta}) = 0$，则 $\hat{\theta}$ 是 θ 的一致性估计。

参考文献

[1] 吴臻, 刘建亚. 概率论与数理统计 [M]. 山东大学出版社, 2004

[2] https://wenku.baidu.com/view/213455baf68a6529647d27284b73f242326c31d1.html