第 3 章　点 估 计

统计分析的一个基本内容是利用样本对总体分布或总体的数字特征进行推断，这一过程通常称为**统计推断**．统计推断主要分为两大类，即**参数估计**与**假设检验**．根据估计的形式不同，参数估计又分为**点估计**和**区间估计**两种类型．我们将在第 4 章介绍假设检验的有关问题，并将主要探讨点估计，而将区间估计放到了第 5 章．文献中关于点估计的求解方法有很多，其中矩估计法、极大似然估计方法及最小二乘方法是三种常见的点估计方法，本章我们主要讨论矩估计和极大似然估计方法，最小二乘方法我们将在第 6 章讨论．

3.1　点估计与优良性

在许多实际问题中，根据历史经验可以认为总体的分布类型已知，而其中的参数是未知的，即总体 X 来自参数分布族 $\{P_\theta(x):\theta\in\Theta\}$．这时，可以通过利用样本 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 估计参数 θ 来推断总体的分布情况．在另外一些问题中，有时可能仅仅关心总体的某些数字特征（也称为参数），如总体的均值 $E(X)$、方差 $\mathrm{Var}(X)$ 等，这时也可以利用样本给出其估计．这些问题都可以归结为参数的点估计问题．

3.1.1　点估计的概念

定义 3.1.1　设 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 是来自总体 X 的一个样本，θ 为总体的未知参数，若用统计量 $\hat{\theta}=\hat{\theta}(\boldsymbol{X})=\hat{\theta}(X_1,X_2,\cdots,X_n)$（$\hat{\theta}$ 与 θ 的维数相同）来估计参数 θ，则称 $\hat{\theta}=\hat{\theta}(\boldsymbol{X})$ 为参数 θ 的**点估计量**．若将其中的样本换成样本观测值 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$，即有 $\hat{\theta}=\hat{\theta}(\boldsymbol{x})=\hat{\theta}(x_1,x_2,\cdots,x_n)$，称 $\hat{\theta}=\hat{\theta}(\boldsymbol{x})$ 为参数 θ 的**点估计值**．

通常，参数 θ 的点估计量和点估计都通称为 θ 的**点估计**或**估计**．从定义中可以看出，未知参数 θ 的点估计量可以有很多，因此，点估计的一个重要内容就是在一定的优良准则下给出一个“好”的估计量．下面讨论估计的优良性标准．

3.1.2　无偏性

定义 3.1.2　设 $\hat{\theta}=\hat{\theta}(\boldsymbol{X})$ 是参数 θ 的点估计量，若 $E(\hat{\theta})=\theta$，则称 $\hat{\theta}=\hat{\theta}(\boldsymbol{X})$ 为参数 θ 的**无偏估计量**；若 $E(\hat{\theta})\neq\theta$，则称 $E(\hat{\theta})-\theta$ 为估计量 $\hat{\theta}$ 的**偏差**，记为 $\mathrm{Bias}(\hat{\theta})$；若 $\lim\limits_{n\to\infty}E(\hat{\theta})=\theta$，则称 $\hat{\theta}=\hat{\theta}(\boldsymbol{X})$ 为参数 θ 的**渐近无偏估计量**．

无偏性的定义可以改为 $E(\hat{\theta}-\theta)=0$．由于 $\hat{\theta}$ 为随机变量，因此用 $\hat{\theta}$ 估计 θ 时，二者之间在很多时候都存在偏差，这种偏差可能为正，可能为负，可能大，可能小．无偏性的含义就是这些偏差的平均值为 0，即不存在系统偏差．

例 3.1.1　设 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 为来自总体 X 的一个样本，记 $E(X)=\mu$，$\mathrm{Var}(X)=\sigma^2$，

试证明样本均值 $\bar{X}$ 和修正的样本方差 $S^2=\dfrac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$ 分别为 μ 和 σ^2 的无偏估计，样本方差 $S_n^2=\dfrac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$ 为 σ^2 的渐近无偏估计.

证　因为 $E(\bar{X})=\dfrac{1}{n}\sum_{i=1}^{n}E(X_i)=\mu$，所以 $\bar{X}$ 是 μ 的无偏估计．又因为修正的样本方差

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

所以

$$E(S^2)=\frac{1}{n-1}\left[\sum_{i=1}^{n}E(X_i^2)-nE(\bar{X}^2)\right]=\frac{1}{n-1}\left[\sum_{i=1}^{n}(\sigma^2+\mu^2)-n\left(\frac{\sigma^2}{n}+\mu^2\right)\right]=\sigma^2,$$

故 S^2 为 σ^2 的无偏估计．而

$$E(S_n^2)=\frac{n-1}{n}\cdot E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right]=\frac{n-1}{n}\sigma^2\to\sigma^2\quad,\quad n\to\infty,$$

因此 S_n^2 为 σ^2 的渐近无偏估计.

例 3.1.2　设 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 为来自总体 $X\sim N(\mu,\sigma^2)$ 的样本，试问 $\bar{X}^2$ 是否为 μ^2 的无偏估计量？

解　由例 3.1.1 可知，$\overline{X}$ 是 μ 的无偏估计量，即 $E(\bar{X})=\mu$，因此

$$E(\bar{X}^2)=\mathrm{Var}(\bar{X})+[E(\bar{X})]^2=\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i)+\mu^2=\frac{\sigma^2}{n}+\mu^2\neq\mu^2,$$

故 $\bar{X}^2$ 不是 μ^2 的无偏估计量.

3.1.3　有效性

若参数 θ 同时存在多个无偏估计，如何选择一个好的无偏估计量？一个自然的想法就是方差越小越好，因为方差越小，该无偏估计在参数 θ 真值的附近波动程度就越小.

定义 3.1.3　设 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 为来自总体 X 的一个样本，$\hat{\theta}_1(\boldsymbol{X})$ 与 $\hat{\theta}_2(\boldsymbol{X})$ 都是 θ 的无偏估计量，若有

$$\mathrm{Var}_\theta(\hat{\theta}_1)\leqslant\mathrm{Var}_\theta(\hat{\theta}_2),\quad\forall\theta\in\Theta,\tag{3.1.1}$$

且至少对某一个 $\theta_0\in\Theta$，有 $\mathrm{Var}_{\theta_0}(\hat{\theta}_1)<\mathrm{Var}_{\theta_0}(\hat{\theta}_2)$，则称 $\hat{\theta}_1(\boldsymbol{X})$ 比 $\hat{\theta}_2(\boldsymbol{X})$ **有效**.

例 3.1.3　设 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 为来自总体 X 的样本，记 $E(X)=\mu$，$\mathrm{Var}(X)=\sigma^2$，常数 $c_i>0$，$i=1,2,\cdots,n$，$\sum_{i=1}^{n}c_i=1$，试证明在 μ 的形如 $\sum_{i=1}^{n}c_iX_i$ 的无偏估计中，$\bar{X}=\dfrac{1}{n}\sum_{i=1}^{n}X_i$ 是最有效的.

证　由于

$$E\left(\sum_{i=1}^{n}c_iX_i\right)=\sum_{i=1}^{n}c_iE(X_i)=\sum_{i=1}^{n}c_i\mu=\mu\sum_{i=1}^{n}c_i=\mu,$$

故 $\sum_{i=1}^{n}c_iX_i$ 为 μ 的无偏估计量. 又因为对于 $\forall\mu\in R$，有

$$\mathrm{Var}\left(\sum_{i=1}^{n}c_iX_i\right)=\sum_{i=1}^{n}c_i^2\cdot\mathrm{Var}(X_i)=\sigma^2\sum_{i=1}^{n}c_i^2\geqslant\frac{\left(\sum_{i=1}^{n}c_i\right)^2\sigma^2}{n}=\frac{\sigma^2}{n},$$

当且仅当 $c_1=c_2=\cdots=c_n=\dfrac{1}{n}$ 时，上述不等式中的等号成立，因此在 μ 的形如 $\sum_{i=1}^{n}c_iX_i$ 的无偏估计中，$\bar{X}=\dfrac{1}{n}\sum_{i=1}^{n}X_i$ 是最有效的.

3.1.4　均方误差准则

对于无偏估计而言，可以通过比较方差来判断其优劣，但有些时候，也可以通过牺牲一定的无偏性来换取方差的大幅下降. 这时可以通过均方误差准则评价估计量的优劣.

定义 3.1.4　设 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 为来自总体 X 的一个样本，$\hat{\theta}=\hat{\theta}(\boldsymbol{X})$ 是参数 θ 的点估计量，$\hat{\theta}$ 的均方误差定义为

$$\mathrm{MSE}_\theta(\hat{\theta})=E(\hat{\theta}-\theta)^2. \tag{3.1.2}$$

均方误差是评价点估计的最一般的标准，反映了估计量 $\hat{\theta}$ 与参数真值 θ 之间的平均差异程度，自然希望估计量 $\hat{\theta}$ 的均方误差越小越好. 那么，是否存在估计量 $\hat{\theta}^*$，使得对所有的估计量 $\hat{\theta}$，有

$$\mathrm{MSE}_\theta(\hat{\theta}^*)\leqslant\mathrm{MSE}_\theta(\hat{\theta}),\qquad\forall\theta\in\Theta. \tag{3.1.3}$$

答案是否定的，即这样的 $\hat{\theta}^*$ 是不存在的. 因为若这样的 $\hat{\theta}^*$ 存在，则对某个 $\theta_0\in\Theta$，取 $\hat{\theta}\equiv\theta_0$，有 $\mathrm{MSE}_{\theta_0}(\hat{\theta})=0$，因此 $\hat{\theta}^*=\theta_0$，a.s.，由于 θ_0 具有任意性，因此这样的 $\hat{\theta}^*$ 是不存在的.

尽管使得均方误差一致达到最小的最优估计是不存在的，但我们可以对估计的合理性提出一些正则性的要求，如可以在某个估计类中去寻求这样的最优估计.

经简单计算可知，

$$\mathrm{MSE}_\theta(\hat{\theta})=\mathrm{Var}_\theta(\hat{\theta})+[E_\theta(\hat{\theta})-\theta]^2=\mathrm{Var}_\theta(\hat{\theta})+[\mathrm{Bias}(\hat{\theta})]^2, \tag{3.1.4}$$

即估计量 $\hat{\theta}$ 的均方误差等于 $\hat{\theta}$ 的方差与 $\hat{\theta}$ 偏差的平方之和. 显然，若 $\hat{\theta}$ 是 θ 的无偏估计，则有 $\mathrm{MSE}_\theta(\hat{\theta})=\mathrm{Var}_\theta(\hat{\theta})$. 因此相对于无偏估计方差最小原则而言，利用均方误差评价估计量的优劣更具有一般化，如有些时候，可以通过牺牲一定的无偏性来换取方差的大幅下降.

3.1.5　相合性

定义 3.1.5　设 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 为来自总体 X 的一个样本，$\hat{\theta}_n=\hat{\theta}_n(\boldsymbol{X})$ 是参数 θ 的点估计量，若对 $\forall\varepsilon>0$，有

$$\lim_{n\to\infty} P_\theta\{|\hat{\theta}_n-\theta|<\varepsilon\}=1，或\ \lim_{n\to\infty} P_\theta\{|\hat{\theta}_n-\theta|\geqslant\varepsilon\}=0，\tag{3.1.5}$$

则称 $\hat{\theta}_n$ 是参数 θ 的（弱）**相合估计量**，记作 $\hat{\theta}_n \xrightarrow{P} \theta$.

定义 3.1.6　设 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 为来自总体 X 的一个样本，$\hat{\theta}_n=\hat{\theta}_n(\boldsymbol{X})$ 是参数 θ 的点估计量，若

$$P_\theta\{\lim_{n\to\infty}\hat{\theta}_n=\theta\}=1，\tag{3.1.6}$$

则称 $\hat{\theta}_n$ 是参数 θ 的**强相合估计量**，记作 $\hat{\theta}_n\to\theta$，a.s..

由 1.6 节的内容可知，强相合性可以推出弱相合性．在绝大多数情况下，弱相合性能够满足统计分析的需要，因此本书主要讨论弱相合性．

相合性的直观含义是随着样本容量 n 的不断增大，估计量 $\hat{\theta}$ 不断逼近参数 θ 的真值. 相合性是估计量的一个最基本的要求，如果随着样本容量 n 的不断增大，用 $\hat{\theta}$ 估计 θ 的精度没有提高，说明该估计就不是一个好的估计.

例 3.1.4　设 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 为来自总体 $X\sim N(\mu,\sigma^2)$，试证明 $S^2=\dfrac{1}{n-1}\sum\limits_{i=1}^{n}(X_i-\bar{X})^2$ 为 σ^2 的相合估计.

证　由于 $\dfrac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$，因此

$$E\left[\frac{(n-1)S^2}{\sigma^2}\right]=n-1，\ \mathrm{Var}\left[\frac{(n-1)S^2}{\sigma^2}\right]=2(n-1)，$$

从而

$$E(S^2)=\sigma^2，\ \mathrm{Var}(S^2)=2(n-1)\cdot\frac{\sigma^4}{(n-1)^2}=\frac{2\sigma^4}{n-1}.$$

对 $\forall\varepsilon>0$，当 $n\to\infty$ 时，有

$$0\leqslant P\{|S^2-\sigma^2|\geqslant\varepsilon\}\leqslant\frac{D(S^2)}{\varepsilon^2}=\frac{2\sigma^4}{\varepsilon^2(n-1)}\to 0，$$

因此 $\lim\limits_{n\to\infty}P\{|S^2-\sigma^2|\geqslant\varepsilon\}=0$，故 S^2 为 σ^2 的相合估计.

定理 3.1.1　设 $T_{ni}=T_{ni}(\boldsymbol{X})$ 为参数 $g_i(\theta)$ 的相合估计，$i=1,2,\cdots,k$，记 $\boldsymbol{g}=(g_1(\theta),g_2(\theta),\cdots,g_k(\theta))^{\mathrm{T}}$，$\boldsymbol{T}_n=(T_{n1},T_{n2},\cdots,T_{nk})^{\mathrm{T}}$，若函数 $f(\boldsymbol{y})=f(y_1,y_2,\cdots,y_k)$ 在 $\boldsymbol{g}$ 处连续，则 $f(\boldsymbol{T}_n)$ 为 $f(\boldsymbol{g})$ 的相合估计.

证　因为 $f(\boldsymbol{y})$ 在 $\boldsymbol{g}=(g_1(\theta),g_2(\theta),\cdots,g_k(\theta))^{\mathrm{T}}$ 处连续，因此对 $\forall\delta>0$，$\exists\eta>0$，使得当 $|y_i-g_i(\theta)|<\eta$，$i=1,2,\cdots,k$ 时，总有

$$|f(\boldsymbol{y})-f(\boldsymbol{g})|<\delta.$$

由此可推知

$$P_\theta\{|f(\boldsymbol{T}_n)-f(\boldsymbol{g})|\geqslant\delta\}\leqslant P_\theta\left(\bigcup_{i=1}^{k}\{|T_{ni}-g_i(\theta)|\geqslant\eta\}\right).$$

又因为 $T_{ni}=T_{ni}(\boldsymbol{X})$ 为 $g_i(\theta)$ 的相合估计，因此对于 $\forall\varepsilon>0$ 和上述的 $\delta>0$，存在正整数 N，使得当 $n>N$ 时，有

$$P\{|T_{ni}-g_i(\theta)|\geqslant\delta\}<\frac{\varepsilon}{k},\quad i=1,2,\cdots,k.$$

故对上述的 $\varepsilon>0$，当 $n>N$ 时，有

$$P\{|f(\boldsymbol{T}_n)-f(\boldsymbol{g})|\geqslant\delta\}\leqslant P_\theta\left(\bigcup_{i=1}^{k}\{|T_{ni}-g_i(\theta)|\geqslant\eta\}\right)\leqslant\sum_{i=1}^{k}P\{|T_{ni}-g_i(\theta)|\geqslant\eta\}=\varepsilon,$$

结论得证.

3.1.6　渐近正态性

定义 3.1.7　设 $\hat{\theta}_n=\hat{\theta}_n(\boldsymbol{X})$ 是参数 θ 的一个估计量，若存在 $\sigma_n(\theta)>0$ 满足 $\lim\limits_{n\to\infty}\sqrt{n}\sigma_n(\theta)=\sigma(\theta)$，其中 $0<\sigma(\theta)<+\infty$，使得当 $n\to\infty$ 时，有

$$\frac{\hat{\theta}_n-\theta}{\sigma_n(\theta)}\xrightarrow{L}N(0,1),\tag{3.1.7}$$

则称 $\hat{\theta}_n$ 为 θ 的**渐近正态估计量**，记为 $\hat{\theta}_n\sim AN(\theta,\sigma_n^2(\theta))$，$\sigma_n^2(\theta)$ 称为 $\hat{\theta}_n$ 的**渐近方差**.

值得注意的是，定义 3.1.7 中的渐近方差 $\sigma_n^2(\theta)$ 是不唯一的. 由 Slutsky 定理可知，如果存在 $\tilde{\sigma}_n(\theta)$ 满足 $\lim\limits_{n\to\infty}\dfrac{\tilde{\sigma}_n(\theta)}{\sigma_n(\theta)}=1$，则 $\tilde{\sigma}_n^2(\theta)$ 也是 $\hat{\theta}_n$ 的渐近方差. 一般地，可取渐近方差 $\sigma_n^2(\theta)=D(\hat{\theta}_n)$.

探讨估计量的渐近正态性是现代统计分析的一个重要内容. 对于渐近正态估计 $\hat{\theta}_n$ 而言，当样本容量 n 比较大时，可以用渐近分布 $N(\theta,\sigma_n^2(\theta))$ 作为 $\hat{\theta}_n$ 的近似分布. 由于渐近正态估计可能有很多，往往利用渐近方差 $\sigma_n^2(\theta)$ 的大小评价其优劣，显然渐近方差越小，估计效果越好.

我们不加证明地给出如下结论.

定理 3.1.2　渐近正态估计一定是相合估计.

例 3.1.5　设 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 是来自总体 $X\sim b(m,p)$ 的一个样本，其中 $p\in(0,1)$ 为未知参数，试求 p 的渐近正态估计，并给出渐近方差.

解　由题意，$E(X)=mp$，$\mathrm{Var}(X)=mp(1-p)$，从而

$$E(\bar{X})=mp,\quad \mathrm{Var}(\bar{X})=\frac{mp(1-p)}{n},$$

由中心极限定理可知，当 $n\to\infty$ 时，有

$$\frac{\bar{X}-mp}{\sqrt{\dfrac{mp(1-p)}{n}}}\xrightarrow{L}N(0,1).$$

从而有

$$\frac{\dfrac{\bar{X}}{m}-p}{\sqrt{\dfrac{p(1-p)}{nm}}}\xrightarrow{L}N(0,1).$$

故 $\dfrac{\overline{X}}{m}$ 为未知参数 p 的渐近正态估计，渐近方差为 $\dfrac{p(1-p)}{nm}$.

3.2 矩 估 计

矩估计法是由英国著名统计学家 K.Pearson 于 1894 年提出来的，其理论依据是命题 2.1.1，即样本矩依概率收敛于总体矩，样本矩的连续函数依概率收敛于总体矩的连续函数.

设 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 是来自总体 X 的一个样本，以 μ_k 表示总体的 k 阶原点矩，以 A_k 表示样本的 k 阶原点矩，即

$$\mu_k=E(X^k),\quad A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k,\tag{3.2.1}$$

若参数 $\boldsymbol{\theta}=(\theta_1,\theta_2,\cdots,\theta_r)^{\mathrm{T}}$ 可以表示为总体前 m 阶矩的函数，即

$$\theta_j=\phi_j(\mu_1,\mu_2,\cdots,\mu_m),\quad j=1,2,\cdots,r,$$

则 $\hat{\theta}_j=\phi_j(A_1,A_2,\cdots,A_m)$ 称为参数 θ_j 的**矩估计量**.

设 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 为来自总体 X 的一组样本，X 的分布函数为 $F(x;\boldsymbol{\theta})$，其中 $\boldsymbol{\theta}=(\theta_1,\theta_2,\cdots,\theta_r)^{\mathrm{T}}$. 矩估计的具体做法为：

（1）求出总体 X 的前 m 阶矩

$$\mu_k=E(X^k)=g(\theta_1,\theta_2,\cdots,\theta_r),\quad k=1,2,\cdots,m;\tag{3.2.2}$$

（2）对方程组（3.2.2）进行求解，得到

$$\theta_j=\phi_j(\mu_1,\mu_2,\cdots,\mu_m),\quad j=1,2,\cdots,r;\tag{3.2.3}$$

（3）将式（3.2.3）中的 μ_k 分别换成 A_k，即可得到参数 $\boldsymbol{\theta}=(\theta_1,\theta_2,\cdots,\theta_m)^{\mathrm{T}}$ 的矩估计量 $\hat{\boldsymbol{\theta}}=(\hat{\theta}_1,\hat{\theta}_2,\cdots,\hat{\theta}_r)^{\mathrm{T}}$，其中 $A_k=\dfrac{1}{n}\sum_{i=1}^{n}X_i^k$ 为样本的 k 阶矩，$k=1,2,\cdots,m$.

注　选择矩估计时尽量选择低阶矩，若式（3.2.2）中的 m 个方程无法解出 θ_j 时，可以使用高于 m 阶的矩.

从矩估计的定义可以看出，矩估计的概率基础是大数定律，因此该估计方法是以大样本为应用前提的；由于矩估计没有用到总体的分布信息，因此本质上来讲，该方法是一种非参数估计方法.

例 3.2.1　$\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 是来自泊松分布 $X\sim P(\lambda)$ 的样本，因为 $E(X)=\lambda$，$\mathrm{Var}(X)=\lambda$，所以 $\overline{X}$ 和 $S_n^2=\dfrac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$ 都可以作为 λ 的矩估计. 由于 $\overline{X}$ 的阶数较低，一般地我们选择 $\overline{X}$ 作为 λ 的矩估计.

例 3.2.2　设总体 X 的密度函数为

$$f(x;\theta)=\begin{cases}(\theta+1)x^{\theta} & 0<x<1\\ 0 & \text{其他}\end{cases},$$

其中，$\theta>-1$ 为未知参数，$\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 为来自总体 X 的样本，试求参数 θ 的矩估计.

解　由于

$$\mu = E(X) = \int_{-\infty}^{+\infty} xf(x;\theta)\mathrm{d}x = \int_0^1 x(\theta+1)x^{\theta}\mathrm{d}x = \frac{\theta+1}{\theta+2},$$

因此有 $\theta = \dfrac{1}{1-\mu} - 2$，故参数 θ 的矩估计为 $\hat{\theta} = \dfrac{1}{1-\overline{X}} - 2$，其中 $\overline{X}$ 为样本均值.

例 3.2.3　设 $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)^{\mathrm{T}}$（$n > 2$）为来自二项分布 $b(k, p)$ 的样本，其中 k 和 p 未知，试求 k 和 p 的矩估计量.

解　由于

$$\mu_1 = E(X) = kp, \quad \mu_2 = E(X^2) = k^2p^2 + kp(1-p),$$

解得

$$p = 1 + \mu_1 - \frac{\mu_2}{\mu_1}, \quad k = \frac{\mu_1^2}{\mu_1 + \mu_1^2 - \mu_2},$$

从而 k 和 p 的矩估计为

$$\hat{p} = 1 + \overline{X} - \frac{A_2}{\overline{X}}, \quad \hat{k} = \frac{\overline{X}^2}{\overline{X} + \overline{X}^2 - A_2},$$

其中 $A_2 = \dfrac{1}{n}\sum_{i=1}^{n} X_i^2$.

为了说明矩估计法的估计效果，可以利用 R 软件从一个已知的二项分布 $b(k, p)$ 中产生随机数，然后利用矩估计给出未知参数的估计．R 代码及结果如下：

```
> set.seed(1)
> n=500;k=20;p=0.8
> x<-rbinom(n,k,p)
> A1<-mean(x)
> A2<-mean(x^2)
> p_esti<-1+A1-A2/A1
> k_esti<-A1^2/(A1+A1^2-A2)
> p_esti
[1] 0.8093514
> k_esti
[1] 19.76644
```

从运行结果可以看到，当抽样个数 n 较大时，矩估计的估计效果还是不错的.

3.3　极大似然估计

极大似然估计最早是由德国数学家 Gauss 于 1821 年针对正态分布提出的一种参数估计方法，之后由英国著名统计学家 R.A. Fisher 于 1922 年针对一般分布再次提出并研究了它的性质，使之成为一种普遍使用的点估计方法.

极大似然估计的基本思想是，概率大的事件比概率小的事件容易发生，概率最大的事件最容易发生.

3.3.1　极大似然估计的原理

若总体 X 为离散型，设 X 的分布列为 $P\{X=x\}=f(x;\boldsymbol{\theta})$，其中 $\boldsymbol{\theta}$ 为未知参数，$\boldsymbol{\theta}\in\Theta\subseteq R^r$，则样本 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 的联合分布列为 $\prod_{i=1}^{n}f(x_i;\boldsymbol{\theta})$，即事件 $\{\boldsymbol{X}=\boldsymbol{x}\}$ 发生的概率为

$$L(\boldsymbol{\theta};\boldsymbol{x})=L(\boldsymbol{\theta};x_1,\cdots,x_n)=\prod_{i=1}^{n}f(x_i;\boldsymbol{\theta}),\tag{3.3.1}$$

这里 $L(\theta;\boldsymbol{x})$ 称为样本的**似然函数**.

若总体 X 为连续型，设 X 的密度函数为 $f(x;\boldsymbol{\theta})$，$\boldsymbol{\theta}\in\Theta\subseteq R^r$，则样本 $\boldsymbol{X}$ 的联合密度函数为 $\prod_{i=1}^{n}f(x_i;\boldsymbol{\theta})$，随机点 $\boldsymbol{X}$ 落在 $\boldsymbol{x}$ 的邻域[①]内的概率为

$$\prod_{i=1}^{n}f(x_i;\boldsymbol{\theta})\mathrm{d}x_i=\prod_{i=1}^{n}f(x_i;\boldsymbol{\theta})\cdot\prod_{i=1}^{n}\mathrm{d}x_i\tag{3.3.2}$$

显然 $\prod_{i=1}^{n}\mathrm{d}x_i$ 与 $\boldsymbol{\theta}$ 无关，称 $L(\boldsymbol{\theta};\boldsymbol{x})=\prod_{i=1}^{n}f(x_i;\boldsymbol{\theta})$ 为样本的**似然函数**.

定义 3.3.1　设 X 服从参数型分布族 $\{f(x;\boldsymbol{\theta}):\boldsymbol{\theta}\in\Theta\subseteq R^r\}$，$\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 为来自总体 X 的一个样本，对于固定的 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$，若存在 $\boldsymbol{\theta}$ 的估计值 $\hat{\boldsymbol{\theta}}=\hat{\boldsymbol{\theta}}(\boldsymbol{x})=(\hat{\theta}_1(\boldsymbol{x}),\hat{\theta}_2(\boldsymbol{x}),\cdots,\hat{\theta}_r(\boldsymbol{x}))^{\mathrm{T}}$，使得似然函数达到最大，即

$$L(\hat{\boldsymbol{\theta}};\boldsymbol{x})=\max_{\boldsymbol{\theta}\in\Theta}L(\boldsymbol{\theta};\boldsymbol{x}),\tag{3.3.3}$$

则称 $\hat{\boldsymbol{\theta}}=\hat{\boldsymbol{\theta}}(\boldsymbol{x})$ 称为参数 θ 的**极大似然估计**或**最大似然估计**（Maximum Likelihood Estimate），简记为 MLE. 通常 $\hat{\boldsymbol{\theta}}=\hat{\boldsymbol{\theta}}(\boldsymbol{x})$ 称为参数 $\boldsymbol{\theta}$ 的**极大似然估计值**，$\hat{\boldsymbol{\theta}}=\hat{\boldsymbol{\theta}}(\boldsymbol{X})$ 称为参数 $\boldsymbol{\theta}$ 的**极大似然估计量**.

一般地，若 $f(x;\boldsymbol{\theta})$ 关于 $\boldsymbol{\theta}$ 可微，且 MLE 存在时，$\boldsymbol{\theta}$ 的极大似然估计通过求导数的方式求得，即若 $\hat{\boldsymbol{\theta}}=\hat{\boldsymbol{\theta}}(\boldsymbol{x})$ 是 Θ 的内点，则 $\hat{\boldsymbol{\theta}}(\boldsymbol{x})=(\hat{\theta}_1(\boldsymbol{x}),\hat{\theta}_2(\boldsymbol{x}),\cdots,\hat{\theta}_k(\boldsymbol{x}))^{\mathrm{T}}$ 是下列似然方程的解，

$$\frac{\partial L(\boldsymbol{\theta};\boldsymbol{x})}{\partial\theta_i}=0,\quad i=1,2,\cdots,k.\tag{3.3.4}$$

又因为 $L(\boldsymbol{\theta};\boldsymbol{x})$ 与 $\ln L(\boldsymbol{\theta};\boldsymbol{x})$ 在同一 $\boldsymbol{\theta}$ 处取到极值，因此 $\boldsymbol{\theta}$ 的 MLE $\hat{\boldsymbol{\theta}}=\hat{\boldsymbol{\theta}}(\boldsymbol{x})$ 也可从下述方程解得

$$\frac{\partial\ln L(\boldsymbol{\theta};\boldsymbol{x})}{\partial\theta_i}=0,\quad i=1,2,\cdots,k,\tag{3.3.5}$$

其中，$\ln L(\boldsymbol{\theta};\boldsymbol{x})$ 也称为样本的**对数似然函数**. 式（3.3.4）或式（3.3.5）称为**似然方程**.

例 3.3.1　设 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 为来自总体 $X\sim N(\mu,\sigma^2)$ 的一个样本，求 μ 和 σ^2 的极大似然估计量.

解　记 $\boldsymbol{\theta}=(\mu,\sigma^2)^{\mathrm{T}}$，总体 X 的密度函数为

$$f(x;\boldsymbol{\theta})=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\},$$

① 邻域指的是边长分别为 $\mathrm{d}x_1,\mathrm{d}x_2,\cdots,\mathrm{d}x_n$ 的 n 维立方体.

故似然函数为

$$L(\boldsymbol{\theta};\boldsymbol{x})=\prod_{i=1}^{n}f(x_i;\boldsymbol{\theta})=(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\},$$

取对数得

$$\ln L(\boldsymbol{\theta};\boldsymbol{x})=-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2,$$

令

$$\begin{cases}\dfrac{\partial l\ln L(\boldsymbol{\theta};\boldsymbol{x})}{\partial\mu}=\dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n}(x_i-\mu)=0\\[2ex]\dfrac{\partial\ln L(\boldsymbol{\theta};\boldsymbol{x})}{\partial\sigma^2}=-\dfrac{n}{2\sigma^2}+\dfrac{1}{2\sigma^4}\displaystyle\sum_{i=1}^{n}(x_i-\mu)^2=0\end{cases},$$

解得

$$\hat{\mu}=\bar{x},\quad \hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2.$$

例 3.3.2　设 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 为来自总体 $X\sim b(1,p)$ 的一个样本，其中 $p\in(0,1)$ 为未知参数，求 p 得极大似然估计值.

解　总体 X 的分布律为

$$P\{X=x\}=p^x(1-p)^{1-x},\quad x=0,1,$$

似然函数为

$$L(p;\boldsymbol{x})=\prod_{i=1}^{n}p^{x_i}(1-p)^{1-x_i}=p^{\sum\limits_{i=1}^{n}x_i}(1-p)^{n-\sum\limits_{i=1}^{n}x_i},\quad x_i=0,1,$$

取对数得

$$\ln L(p;\boldsymbol{x})=\sum_{i=1}^{n}x_i\ln p+\left(n-\sum_{i=1}^{n}x_i\right)\ln(1-p),$$

令

$$\frac{\mathrm{d}\ln L(p;\boldsymbol{x})}{\mathrm{d}p}=\frac{\sum\limits_{i=1}^{n}x_i}{p}-\frac{n-\sum\limits_{i=1}^{n}x_i}{1-p}=0,$$

可解得

$$\hat{p}=\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i.$$

当 $0<\bar{x}<1$ 时，容易验证 $\bar{x}$ 为 p 的 MLE；当 $\bar{x}=0$ 或 1 时，$\bar{x}\notin\Theta$，严格意义上 p 的 MLE 不存在，但由于 0 和 1 处于 Θ 的边界，因此在实际意义中，人们仍把 $\bar{x}$ 称为 p 的 MLE.

例 3.3.3　设总体 X 服从双参数指数分布，其密度函数为

$$f(x;\theta,c)=\begin{cases}\dfrac{1}{\theta}\mathrm{e}^{-\frac{x-c}{\theta}} & x\geqslant c\\ 0 & x<c\end{cases},$$

其中 $\theta>0$，$c>0$ 为未知参数，$\boldsymbol{x}=(x_1,x_2\cdots,x_n)^{\mathrm{T}}$ 为来自总体 X 的样本观测值，试求参数 θ 和 c 的 MLE.

解　记 $x_{(1)}=\min\{x_1,x_2,\cdots,x_n\}$，则似然函数为

$$L(\theta,c;\boldsymbol{x})=\prod_{i=1}^{n}f(x_i;\theta,c)=\begin{cases}\displaystyle\prod_{i=1}^{n}\frac{1}{\theta}\mathrm{e}^{-\frac{x_i-c}{\theta}} & x_{(1)}\geqslant c\\ 0 & x_{(1)}<c\end{cases},$$

则当 $x_{(1)}\geqslant c$ 时，

$$\ln L(\theta,c;\boldsymbol{x})=-n\ln\theta-\frac{1}{\theta}\sum_{i=1}^{n}x_i+\frac{nc}{\theta},$$

由于 $\dfrac{n}{\theta}>0$，因此 $\ln L(\theta,c;\boldsymbol{x})$ 为 c 的单调递增函数，故要使得 $\ln L(\theta,c;\boldsymbol{x})$ 达到最大，c 应该取最大值，从而 c 的最大似然估计值为 $\hat{c}=x_{(1)}$. 令

$$\frac{\partial\ln L(\theta,c;\boldsymbol{x})}{\partial\theta}=-\frac{n}{\theta}+\frac{1}{\theta^2}\sum_{i=1}^{n}x_i-\frac{nc}{\theta^2}=0,$$

解得 $\theta=\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}x_i-c=\overline{x}-c$，从而参数 θ 的最大似然估计值为 $\hat{\theta}=\overline{x}-x_{(1)}$.

在单参数情形下，我们可以调用 R 函数 optimize( )或 optimise( )来求参数的极大似然估计. 其调用格式为：

optimize(f = ，interval = ，...，lower = min(interval)，upper = max(interval)，
maximum = FALSE，tol = .Machine$double.eps^0.25)

其中，f 为似然函数；interval 为二维向量，表示参数 θ 的取值范围；lower 和 upper 分别为搜索区间的最小值和最大值，默认值即为 interval 的左端点和右端点；maximum 为逻辑值，maximum =TRUE 表示求最大值，默认值为 FALSE，表示求最小值；tol 用来给出求值的精度. optimize( )的返回值为极值点和极值.

以例 3.3.2 为例，若样本容量 $n=20$，$\displaystyle\sum_{i=1}^{n}x_i=13$，R 代码及结果如下：

```
> myfunction<-function(p,n,sum)    sum*log(p)+(n-sum)*log(1-p)   #对数似然函数
> optimize(myfunction,n=20,sum=13,c(0,1),maximum=TRUE)
$maximum
[1] 0.6500073
$objective
[1] -12.94893
```

其中$maximum 为最大值点，即极大似然估计值为 $\hat{\theta}=0.6500073\approx 0.65$；$objective 为在近似解处的对数似然函数值，这里为−12.94893.

若似然方程或对数似然方程没有显示解，对于一元方程可以用 R 软件中的 uniroot()函数求得其数值解．uniroot()函数的调用格式为

```
uniroot(f, interval, ...,lower = min(interval), upper = max(interval),
        f.lower = f(lower, ...), f.upper = f(upper, ...),
        extendInt = c("no", "yes", "downX", "upX"), check.conv = FALSE,
        tol = .Machine$double.eps^0.25, maxiter = 1000, trace = 0)
```

其中，f 为所求函数；interval 为二维向量，表示包含方程根的搜索区间； lower 和 upper 分别为求根区间的左、右端点（默认值为初始区间的左、右端点）；参数 extendInt =“TRUE”可以自动扩展参数搜索范围；tol 表示计算精度；maxiter 为最大迭代次数（默认值为 1000）．还是以例 3.3.2 为例，R 代码如下：

```
> f<-function(p,n,sum) sum/p-(n-sum)/(1-p)
> uniroot(f,n=20,sum=13,c(0,1))
```

运行结果为：

```
$root
    [1] 0.6499916
$f.root
    [1] 0.0007365165
$iter
    [1] 7
$init.it
    [1] NA
$estim.prec
    [1] 6.103516e-05
```

这里，$root 为方程的近似解，即极大似然估计值为 $\hat{\theta}=0.6499916\approx 0.65$；$f.root 为近似点处的函数值；$iter 为迭代次数；$estim.prec 为近似解绝对误差的估计值.

在多参数情形下，可以利用 optim( )或 nlm( )来求解未知参数的极大似然估计，其中 nlm()使用 Newton- Raphson 算法，而 optim( )有 5 种优化方法可供选择.

程序包 stats4 中的 mle( )函数也可以进行极大似然估计．其调用格式为

```
mle(minuslogl, start = formals(minuslogl), method = "BFGS", fixed = list(), ...)
```

其中，minuslogl 为用来计算的非负对数似然函数，start 为初始值，method 用于指定优化算法，默认值为 BFGS 算法，fixed 用于设定优化过程中保持固定的参数值.

以例 3.3.2 为例，R 代码如下：

```
> x<-c(0,1);y<-c(7,13)
> xx<-rep(x,y)
> nL<-function(prob) -sum(dbinom(xx,size=1,prob,log=TRUE))
> mle(nL, start = list(prob = 0.5))
```

运行结果为：

```
Call:
mle(minuslogl = nL, start = list(prob = 0.5))
Coefficients:
      prob
0.6499999
```

从运行结果可以看到，极大似然估计值为 $\hat{\theta}=0.6499999\approx 0.65$.

3.3.2 极大似然估计的性质

在实际问题中，有些时候需要给出未知参数的函数的极大似然估计，这时可以利用极大似然估计的不变性进行求解.

定理 3.3.1（极大似然估计的不变性） 设 $\hat{\boldsymbol{\theta}}=\hat{\boldsymbol{\theta}}(\boldsymbol{x})$ 为 $\boldsymbol{\theta}$ 的 MLE，$g(\boldsymbol{\theta})$ 为可测函数，则 $g(\hat{\boldsymbol{\theta}})$ 为 $g(\boldsymbol{\theta})$ 的 MLE.

证明见茆诗松等（2006），此处略. 利用极大似然估计的不变性可以很容易地给出参数函数的极大似然估计. 例如，例 3.3.1 给出 σ^2 的 MLE 为 $\hat{\sigma}^2=\dfrac{1}{n}\sum\limits_{i=1}^{n}(X_i-\bar{X})^2$，则 $\hat{\sigma}=\sqrt{\dfrac{1}{n}\sum\limits_{i=1}^{n}(X_i-\bar{X})^2}$ 为 σ 的 MLE .

定理 3.3.2 设 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 为来自总体 X 的一个样本，$T=T(\boldsymbol{X})$ 为参数 $\boldsymbol{\theta}$ 的充分统计量，则 $\boldsymbol{\theta}$ 的极大似然估计 $\hat{\boldsymbol{\theta}}=\hat{\boldsymbol{\theta}}(\boldsymbol{X})$ 可以表示为 T 的函数.

证 由于 $T=T(\boldsymbol{X})$ 为参数 $\boldsymbol{\theta}$ 的充分统计量，因此由因子分解定理可知，似然函数可以表示为

$$L(\boldsymbol{\theta};\boldsymbol{x})=g_{\theta}[T(\boldsymbol{x})]\cdot h(\boldsymbol{x}),$$

其中 $h(\boldsymbol{x})$ 与参数 $\boldsymbol{\theta}$ 无关，因此最大化 $L(\boldsymbol{\theta};\boldsymbol{x})$ 等价于最大化 $g_{\theta}[T(\boldsymbol{x})]$，因此 $\boldsymbol{\theta}$ 的极大似然估计 $\hat{\boldsymbol{\theta}}$ 可以表示为 T 的函数.

3.4 一致最小方差无偏估计

我们在 3.1.2 节给出了无偏估计的定义，然而在许多情形中，未知参数 θ 的无偏估计是不唯一的，这时一个自然的想法就是寻求一个无偏估计，使得它的方差达到最小，这就引出了一致最小方差无偏估计的概念.

3.4.1 一致最小方差无偏估计的概念

在讨论参数的无偏估计时，值得注意的是，在有些场合下，未知参数 $g(\theta)$ 的无偏估计是不存在的.

例 3.4.1 设 $X\sim b(n,\theta)$，其中 $\theta\in(0,1)$ 为未知参数，试证明当样本容量 $n=1$ 时，参数 $g(\theta)=\sin\theta$ 不存在无偏估计.

证　假若 $T=T(X_1)$ 是参数 $g(\theta)=\sin\theta$ 的无偏估计，则对 $\forall\theta\in(0,1)$，有

$$E_\theta(T)=\sum_{i=0}^{n}T(i)\binom{n}{i}\theta^i(1-\theta)^{n-i}=g(\theta)=\sin\theta,$$

显然 $\sum_{i=0}^{n}T(i)\binom{n}{i}\theta^i(1-\theta)^{n-i}$ 是关于 θ 的 n 阶多项式，它不可能在 $(0,1)$ 中处处等于一个超越函数 $\sin\theta$，因此 $g(\theta)=\sin\theta$ 不存在无偏估计.

为讨论方便，将存在无偏估计的参数称为**可估参数**. 以后讨论无偏估计时，总是对可估参数而言的. 设 $g(\theta)$ 为可估参数，把 $g(\theta)$ 的所有无偏估计组成的类记为 $\mathcal{U}_g$. 我们关心的问题是：在 $\mathcal{U}_g$ 中选取一个好的估计.

定义 3.4.1　设 $g(\theta)$ 为可估参数，$\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 为来自总体 X 的一个样本，$T(\boldsymbol{X})$ 为 $g(\theta)$ 的无偏估计，$\mathcal{U}_g$ 为 $g(\theta)$ 的所有无偏估计组成的类，对于 $\mathcal{U}_g$ 中的任意无偏估计 $\varphi(\boldsymbol{X})$，有

$$\mathrm{Var}_\theta(T(\boldsymbol{X}))\leqslant\mathrm{Var}_\theta(\varphi(\boldsymbol{X})),\quad\forall\theta\in\Theta,\tag{3.4.1}$$

则称 $T(\boldsymbol{X})$ 为 $g(\theta)$ 的**一致最小方差无偏估计**（Uniformly Minimum Variance Unbiased Estimate），简记为 UMVUE.

关于 UMVUE 的唯一性问题，我们不加证明地给出如下结论.

定理 3.4.1　参数 $g(\theta)$ 的 UMVUE 若存在则在几乎处处意义下是唯一的，即若 $T_1(\boldsymbol{X})$ 和 $T_2(\boldsymbol{X})$ 同为 $g(\theta)$ 的 UMVUE，则

$$P_\theta\{T_1(\boldsymbol{X})=T_2(\boldsymbol{X})\}=1,\quad\forall\theta\in\Theta.\tag{3.4.2}$$

由于参数 θ 的充分统计量包含了参数所有信息，因此猜想 $g(\theta)$ 的 UMVUE 是否可以仅仅在充分统计量的无偏估计类中寻找？换言之，若 $T=T(\boldsymbol{X})$ 是 θ 的充分统计量，$g(\theta)$ 的 UMVUE 能否表示为 T 的函数？

定理 3.4.2（Rao-Blackwell）　设 $S(\boldsymbol{X})$ 为 θ 的充分统计量，$\varphi(\boldsymbol{X})$ 为 $g(\theta)$ 的无偏估计，令 $T(\boldsymbol{X})=E[\varphi(\boldsymbol{X})\,|\,S(\boldsymbol{X})]$，则 $T(\boldsymbol{X})$ 也是 $g(\theta)$ 的无偏估计，且

$$\mathrm{Var}_\theta(T(\boldsymbol{X}))\leqslant\mathrm{Var}_\theta(\varphi(\boldsymbol{X})),\quad\forall\theta\in\Theta.\tag{3.4.3}$$

且等号成立当且仅当

$$P_\theta\{T(\boldsymbol{X})=\varphi(\boldsymbol{X})\}=1,\quad\forall\theta\in\Theta.\tag{3.4.4}$$

证　因为 $S(\boldsymbol{X})$ 是充分统计量，因此 $T(\boldsymbol{X})=E[\varphi(\boldsymbol{X})\,|\,S(\boldsymbol{X})]$ 与 θ 无关，故 $T(\boldsymbol{X})$ 为统计量，且

$$E_\theta[T(\boldsymbol{X})]=E_\theta\{E[\varphi(\boldsymbol{X})\,|\,S(\boldsymbol{X})]\}=E_\theta[\varphi(\boldsymbol{X})]=g(\theta),\quad\forall\theta\in\Theta,$$

因此 $T(\boldsymbol{X})$ 为 $g(\theta)$ 的无偏估计. 而

$$\begin{aligned}\mathrm{Var}_\theta[\varphi(\boldsymbol{X})]&=E_\theta[\varphi(\boldsymbol{X})-g(\theta)]^2=E_\theta[\varphi(\boldsymbol{X})-T(\boldsymbol{X})+T(\boldsymbol{X})-g(\theta)]^2\\&=E_\theta[\varphi(\boldsymbol{X})-T(\boldsymbol{X})]^2+\mathrm{Var}_\theta(T(\boldsymbol{X}))+2E_\theta\{[\varphi(\boldsymbol{X})-T(\boldsymbol{X})][T(\boldsymbol{X})-g(\theta)]\},\end{aligned}$$

而

$$\begin{aligned}E_\theta\{[\varphi(\boldsymbol{X})-T(\boldsymbol{X})][T(\boldsymbol{X})-g(\theta)]\}&=E_\theta\{E[(\varphi-T)(T-g)\,|\,S(\boldsymbol{X})]\}\\&=E_\theta\{(T-g)E[(\varphi-T)\,|\,S(\boldsymbol{X})]\}=E_\theta[(T-g)(T-T)]=0,\end{aligned}$$

因此对 $\forall\theta\in\Theta$，有

$$\mathrm{Var}_\theta[\varphi(\boldsymbol{X})] = E_\theta[\varphi(\boldsymbol{X}) - T(\boldsymbol{X})]^2 + \mathrm{Var}_\theta(T(\boldsymbol{X})) \geqslant \mathrm{Var}_\theta(T(\boldsymbol{X}))\,. \tag{3.4.5}$$

又因为式（3.4.5）中的等号成立当且仅当 $E_\theta[\varphi(\boldsymbol{X}) - T(\boldsymbol{X})]^2 = 0$，因此式（3.4.4）成立.

由定理 3.4.2 可知，寻求参数 $g(\theta)$ 的 UMVUE，只需在基于充分统计量的无偏估计类中讨论即可.

下面给出两种常用的 UMVUE 的求解方法.

3.4.2　零无偏估计法

为讨论方便，引入如下无偏估计类：

$$\mathcal{U}_0 = \{T : E_\theta[T(\boldsymbol{X})] = 0,\, E_\theta(T^2) < +\infty,\, \forall\, \theta \in \Theta\}\,, \tag{3.4.6}$$

即 $\mathcal{U}_0$ 表示 0 的具有二阶矩的无偏估计类.

定理 3.4.3　设 $g(\theta)$ 为可估参数，$T = T(\boldsymbol{X})$ 为 $g(\theta)$ 的无偏估计，且对 $\forall\, \theta \in \Theta$，有 $\mathrm{Var}_\theta(T(\boldsymbol{X})) < +\infty$，则 $T = T(\boldsymbol{X})$ 为 $g(\theta)$ 的 UMVUE 的充分必要条件是对于任意的 $\varphi = \varphi(\boldsymbol{X}) \in \mathcal{U}_0$，有

$$E_\theta(\varphi T) = \mathrm{Cov}(\varphi, T) = 0\,,\ \ \forall\, \theta \in \Theta\,. \tag{3.4.7}$$

证明参见王兆军和邹长亮的《数理统计教程》（2014）.

例 3.4.2　设 $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)^{\mathrm{T}}$ 是来自指数分布 $Exp(\lambda)$ 的样本，试求总体期望 $g(\lambda) = \dfrac{1}{\lambda}$ 的 UMVUE.

解　由于 $X_i \sim Exp(\lambda) = \Gamma(1, \lambda)$，由伽马分布的可加性可知，$T = \sum\limits_{i=1}^{n} X_i \sim \Gamma(n, \lambda)$，因此 T 的概率密度函数为

$$f(t;\lambda) = \frac{\lambda^n}{(n-1)!} t^{n-1} \mathrm{e}^{-\lambda t} I\{t > 0\}\,,$$

易证 $\dfrac{T}{n}$ 为 $g(\lambda)$ 的无偏估计. 由指数分布族的性质可知，T 为 λ 的充分统计量. 根据定理 3.4.2，$g(\lambda)$ 的 UMVUE 一定可以表示为 T 的函数. 设 $\varphi = \varphi(T)$ 为 0 的任一无偏估计，即有

$$0 = \int_0^{+\infty} \varphi(t) \frac{\lambda^n}{(n-1)!} t^{n-1} \mathrm{e}^{-\lambda t} \mathrm{d}t\,,\ \ \forall\, \lambda > 0\,,$$

即有

$$0 = \int_0^{+\infty} \varphi(t) t^{n-1} \mathrm{e}^{-\lambda t} \mathrm{d}t\,,\ \ \forall\, \lambda > 0\,,$$

等式两边关于 λ 求导数得

$$0 = \int_0^{+\infty} t\varphi(t) t^{n-1} \mathrm{e}^{-\lambda t} \mathrm{d}t\,,\ \ \forall\, \lambda > 0\,,$$

因此，对于 $\forall\, \lambda > 0$，有 $E_\theta(\varphi T) = 0$. 由定理 3.4.3 可知，$\dfrac{T}{n} = \dfrac{1}{n}\sum\limits_{i=1}^{n} X_i$ 为 $\dfrac{1}{\lambda}$ 的 UMVUE.

3.4.3　充分完备统计量法

设 $T=T(\boldsymbol{X})$ 是充分统计量，由定理 3.4.2 可知，若 $g(\theta)$ 的 UMVUE 存在，则它必为 T 的函数．若进一步知道 $g(\theta)$ 的无偏估计在几乎处处意义下是唯一的，则它一定为 $g(\theta)$ 的 UMVUE.

定理 3.4.4　设 $S(\boldsymbol{X})$ 为参数型分布族 $\{P_\theta,\theta\in\Theta\}$ 的充分完备统计量，$g(\theta)$ 为可估参数，$\varphi(\boldsymbol{X})$ 为 $g(\theta)$ 的一个无偏估计，且满足 $\mathrm{Var}_\theta(\varphi(\boldsymbol{X}))<+\infty$，$\forall\theta\in\Theta$，则

$$T(\boldsymbol{X})=E[\varphi(\boldsymbol{X})\,|\,S(\boldsymbol{X})] \tag{3.4.8}$$

是 $g(\theta)$ 的 UMVUE，且在几乎处处意义下是唯一的.

结合定理 3.4.2 和完备统计量的定义，容易证明定理 3.4.4 成立．定理 3.4.4 给出了一种寻求UMVUE的方法，即若 $S(\boldsymbol{X})$ 为 θ 的充分完备统计量，$h[S(\boldsymbol{X})]$ 为 $g(\theta)$ 的无偏估计，则 $h[S(\boldsymbol{X})]$ 必为 $g(\theta)$ 的 UMVUE.

例 3.4.3　设 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 为来自均匀分布 $U(0,\theta)$ 的一个样本，求 θ 的 UMVUE.

解　由因子分解定理可知，$T=X_{(n)}$ 是充分统计量，且概率密度函数为

$$p(t;\theta)=nt^{n-1}/\theta^n,\quad 0<t<\theta.$$

下证 $X_{(n)}$ 也是完备统计量．若函数 $\varphi(T)$ 是 0 的无偏估计，即

$$\int_0^\theta \varphi(t)\cdot\frac{nt^{n-1}}{\theta^n}\mathrm{d}t=0,\quad \forall\theta>0,$$

因此

$$\int_0^\theta \phi(t)\cdot t^{n-1}\mathrm{d}t=0,\quad \forall\theta>0,$$

等式两边对 θ 求导得，$\varphi(\theta)\theta^{n-1}=0$，从而 $\varphi(\theta)=0$，$\forall\theta>0$，故 $X_{(n)}$ 是完备统计量．又因为

$$E(X_{(n)})=\int_0^\theta t\cdot\frac{nt^{n-1}}{\theta^n}\mathrm{d}t=\frac{n}{n+1}\theta,$$

从而 $E\left(\dfrac{n+1}{n}X_{(n)}\right)=\theta$，故 $\dfrac{n+1}{n}X_{(n)}$ 为 θ 的 UMVUE .

3.5　Cramer-Rao 不等式

在前面的讨论中，我们的主要目的是在可估参数 $g(\theta)$ 的无偏估计类中寻找方差最小的估计．最小的方差到底有多大？本节我们将讨论可估参数 $g(\theta)$ 的无偏估计方差的下界.

3.5.1　C-R 正则分布族与 Fisher 信息

定义 3.5.1　若下列 5 个条件成立，参数分布族 $\{f(x;\boldsymbol{\theta}),\boldsymbol{\theta}\in\Theta\}$ 称为 **Cramer-Rao 正则分布族**或 **C-R 正则族**：

（1）Θ 为 R^r 上的开矩形；

（2）对于 $\forall\boldsymbol{\theta}\in\Theta$，$\dfrac{\partial\ln f(x;\boldsymbol{\theta})}{\partial\theta_i}$　$(i=1,2,\cdots,r)$ 都存在；

（3）支撑 $\mathcal{A}=\{x: f(x;\boldsymbol{\theta})>0\}$ 与 $\boldsymbol{\theta}$ 无关；

（4）对 $f(x;\boldsymbol{\theta})$ 的积分与微分可交换，即

$$\frac{\partial}{\partial\theta_i}\int f(x;\boldsymbol{\theta})\mathrm{d}x=\int\frac{\partial f(x;\boldsymbol{\theta})}{\partial\theta_i}\mathrm{d}x,\quad i=1,2,\cdots,r;$$

（5）对于 $\forall\boldsymbol{\theta}\in\Theta$，$E_{\boldsymbol{\theta}}\left(\dfrac{\partial\ln f(X;\boldsymbol{\theta})}{\partial\theta_i}\right)^2<+\infty$，$i=1,2,\cdots,r$.

常见的分布族大都属于 C-R 正则族，如指数型分布族是 C-R 正则族，但有些分布族不属于 C-R 正则族. 如均匀分布族 $U(\theta-1,\theta+1)$ 不属于 C-R 正则族，因为其支撑 $\mathcal{A}$ 与未知参数 θ 有关系.

定义 3.5.2　设 $\{f(x;\boldsymbol{\theta}),\boldsymbol{\theta}\in\Theta\subseteq R^r\}$ 为 C-R 正则族，记

$$\boldsymbol{S}_{\boldsymbol{\theta}}(x)=\left(\frac{\partial\ln f(x;\boldsymbol{\theta})}{\partial\theta_1},\frac{\partial\ln f(x;\boldsymbol{\theta})}{\partial\theta_2},\cdots,\frac{\partial\ln f(x;\boldsymbol{\theta})}{\partial\theta_r}\right)^{\mathrm{T}},\tag{3.5.1}$$

则有 $E_{\boldsymbol{\theta}}[\boldsymbol{S}_{\boldsymbol{\theta}}(X)]=0$，定义

$$I(\boldsymbol{\theta})=\mathrm{Var}_{\boldsymbol{\theta}}[\boldsymbol{S}_{\boldsymbol{\theta}}(X)]=E_{\boldsymbol{\theta}}[\boldsymbol{S}_{\boldsymbol{\theta}}(X)\boldsymbol{S}_{\boldsymbol{\theta}}^{\mathrm{T}}(X)],\tag{3.5.2}$$

则称 $I(\boldsymbol{\theta})$ 为该分布族的 **Fisher 信息矩阵**，简称 **Fisher 信息**；$r=1$ 时，$I(\theta)$ 称为 **Fisher 信息量**.

根据定义 3.5.2，当 $r=1$ 时，Fisher 信息量为

$$I(\theta)=\mathrm{Var}_{\theta}\left[\frac{\partial\ln f(x;\theta)}{\partial\theta}\right]=E_{\theta}\left[\frac{\partial\ln f(x;\theta)}{\partial\theta}\right]^2.\tag{3.5.3}$$

当 $r=2$ 时，Fisher 信息矩阵为

$$\boldsymbol{I}(\theta)=\begin{pmatrix}I_{11}&I_{12}\\I_{21}&I_{22}\end{pmatrix},\tag{3.5.4}$$

其中

$$I_{11}=E_{\boldsymbol{\theta}}\left[\frac{\partial\ln f(X;\boldsymbol{\theta})}{\partial\theta_1}\right]^2,$$

$$I_{12}=I_{21}=E_{\boldsymbol{\theta}}\left[\frac{\partial\ln f(X;\boldsymbol{\theta})}{\partial\theta_1}\frac{\partial\ln f(X;\boldsymbol{\theta})}{\partial\theta_2}\right],$$

$$I_{22}=E_{\boldsymbol{\theta}}\left[\frac{\partial\ln f(X;\boldsymbol{\theta})}{\partial\theta_2}\right]^2.$$

例 3.5.1　已知正态分布族 $\{N(\mu,\sigma^2),(\mu,\sigma^2)\in R\times R^+\}$ 为 Cramer-Rao 正则族，求其 Fisher 信息.

解　记 $\boldsymbol{\theta}=(\mu,\sigma^2)^{\mathrm{T}}$，则分布族的概率密度函数为

$$f(x;\boldsymbol{\theta})=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\},$$

取对数得

$$\ln f(x;\boldsymbol{\theta})=-\frac{1}{2}\ln\sigma^2-\frac{(x-\mu)^2}{2\sigma^2}-\ln\sqrt{2\pi},$$

因此

$$\frac{\partial\ln f(x;\boldsymbol{\theta})}{\partial\mu}=\frac{x-\mu}{\sigma^2},\quad \frac{\partial\ln f(x;\boldsymbol{\theta})}{\partial\sigma^2}=-\frac{1}{2\sigma^2}+\frac{(x-\mu)^2}{2\sigma^4},$$

故有

$$I_{11}=\mathrm{Var}_{\boldsymbol{\theta}}\left(\frac{X-\mu}{\sigma^2}\right)=\frac{1}{\sigma^4}\mathrm{Var}_{\boldsymbol{\theta}}(X)=\frac{1}{\sigma^2},$$

$$I_{22}=\mathrm{Var}_{\boldsymbol{\theta}}\left(-\frac{1}{2\sigma^2}+\frac{(X-\mu)^2}{2\sigma^4}\right)=\frac{1}{4\sigma^4}\mathrm{Var}_{\boldsymbol{\theta}}\left[\left(\frac{X-\mu}{\sigma}\right)^2\right]=\frac{2}{4\sigma^4}=\frac{1}{2\sigma^4},$$

$$I_{12}=E_{\boldsymbol{\theta}}\left[\left(\frac{X-\mu}{\sigma^2}\right)\left(-\frac{1}{2\sigma^2}+\frac{(X-\mu)^2}{2\sigma^4}\right)\right]=E_{\boldsymbol{\theta}}\left(-\frac{X-\mu}{2\sigma^4}+\frac{(X-\mu)^3}{2\sigma^6}\right)=0,$$

Fisher 信息阵为

$$I(\boldsymbol{\theta})=\begin{pmatrix}\dfrac{1}{\sigma^2} & 0\\ 0 & \dfrac{1}{2\sigma^4}\end{pmatrix}.$$

若进一步假定 $f(x;\boldsymbol{\theta})$ 对 $\boldsymbol{\theta}$ 存在二阶偏导，且积分与微分可交换次序，这时可以利用二阶偏导数求 Fisher 信息.

命题 3.5.1　设 $\{f(x;\boldsymbol{\theta}),\boldsymbol{\theta}\in\Theta\subseteq R^r\}$ 为 C-R 正则族，若 $\dfrac{\partial^2 f(x;\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}$ $(i,j=1,2,\cdots,r)$ 存在，则分布族的 Fisher 信息为 $I(\boldsymbol{\theta})=(I_{ij})_{r\times r}$，其中

$$I_{ij}=-E_{\boldsymbol{\theta}}\left(\frac{\partial^2\ln f(X;\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right).\tag{3.5.5}$$

例 3.5.2（续例 3.5.1）　求正态分布族 $\{N(\mu,\sigma^2),(\mu,\sigma^2)\in R\times R^+\}$ 的 Fisher 信息.

解　由例 3.5.1 可知，

$$\frac{\partial\ln f(x;\boldsymbol{\theta})}{\partial\mu}=\frac{x-\mu}{\sigma^2},\quad \frac{\partial\ln f(x;\boldsymbol{\theta})}{\partial\sigma^2}=-\frac{1}{2\sigma^2}+\frac{(x-\mu)^2}{2\sigma^4},$$

从而

$$\frac{\partial^2\ln f(x;\boldsymbol{\theta})}{\partial\mu^2}=\frac{-1}{\sigma^2},\quad \frac{\partial^2\ln f(x;\boldsymbol{\theta})}{\partial\mu\partial\sigma^2}=-\frac{x-\mu}{\sigma^4},\quad \frac{\partial^2\ln f(x;\boldsymbol{\theta})}{\partial(\sigma^2)^2}=\frac{1}{2\sigma^4}-\frac{(x-\mu)^2}{\sigma^6},$$

故

$$I_{11}=-E_{\boldsymbol{\theta}}\left[\frac{\partial^2\ln f(X;\boldsymbol{\theta})}{\partial\mu^2}\right]=\frac{1}{\sigma^2},\quad I_{12}=-E_{\boldsymbol{\theta}}\left[\frac{\partial^2\ln f(X;\boldsymbol{\theta})}{\partial\mu\partial\sigma^2}\right]=0,$$

$$I_{22}=-E_{\boldsymbol{\theta}}\left[\frac{\partial^2\ln f(X;\boldsymbol{\theta})}{\partial\sigma^2}\right]=\frac{1}{2\sigma^4}.$$

因此正态分布族的 Fisher 信息为

$$I(\boldsymbol{\theta})=\begin{pmatrix}\dfrac{1}{\sigma^2} & 0\\ 0 & \dfrac{1}{2\sigma^4}\end{pmatrix}.$$

命题 3.5.2　若 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 是来自总体 C-R 正则族 $\{f(x;\boldsymbol{\theta}),\boldsymbol{\theta}\in\Theta\subseteq R^r\}$ 的一个样本，记总体的 Fisher 信息为 $I_1(\boldsymbol{\theta})$，则样本 $\boldsymbol{X}$ 的 Fisher 信息为

$$I_n(\boldsymbol{\theta})=nI_1(\boldsymbol{\theta}). \tag{3.5.6}$$

证　样本 $\boldsymbol{X}$ 的概率密度函数为 $f(\boldsymbol{x};\boldsymbol{\theta})=\prod_{i=1}^{n}f(x_i;\boldsymbol{\theta})$，因此 $\ln f(\boldsymbol{x};\boldsymbol{\theta})=\sum_{i=1}^{n}\ln f(x_i;\boldsymbol{\theta})$，从而

$$\boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{x})=\frac{\partial}{\partial\boldsymbol{\theta}}\sum_{i=1}^{n}\ln f(x_i;\boldsymbol{\theta})=\sum_{i=1}^{n}\frac{\partial\ln f(x_i;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}},$$

故样本 $\boldsymbol{X}$ 的 Fisher 信息为

$$I_n(\boldsymbol{\theta})=\mathrm{Var}_{\boldsymbol{\theta}}[\boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{X})]=\sum_{i=1}^{n}\mathrm{Var}_{\boldsymbol{\theta}}\left[\frac{\partial\ln f(x_i;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right]=nI_1(\boldsymbol{\theta}). \tag{3.5.7}$$

3.5.2　统计量的 Fisher 信息

Fisher 信息是数理统计中的一个基本概念，很多统计结果都与 Fisher 信息有关．在实际问题中，一个常用的概念就是统计量的 Fisher 信息．

定义 3.5.3　设 $T=T(\boldsymbol{X})$ 是分布族 $\{f(x;\boldsymbol{\theta}),\boldsymbol{\theta}\in\Theta\}$ 上的统计量，$\{f_T(x;\boldsymbol{\theta}),\boldsymbol{\theta}\in\Theta\}$ 是 $T(\boldsymbol{X})$ 诱导的分布族，则分布族 $\{f_T(x;\boldsymbol{\theta}),\boldsymbol{\theta}\in\Theta\}$ 的 Fisher 信息称为统计量 $T(\boldsymbol{X})$ 的 Fisher 信息，记为 $I_T(\boldsymbol{\theta})$．

定理 3.5.1　设 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 是来自 C-R 正则族 $\{f(x;\boldsymbol{\theta}),\boldsymbol{\theta}\in\Theta\}$ 的一个样本，样本 $\boldsymbol{X}$ 的 Fisher 信息为 $I_n(\boldsymbol{\theta})$，又设统计量 $T=T(\boldsymbol{X})$ 的 Fisher 信息存在，记为 $I_T(\boldsymbol{\theta})$，则有

$$I_T(\boldsymbol{\theta})\leqslant I_n(\boldsymbol{\theta}),$$

且等号成立的充分必要条件为 $T=T(\boldsymbol{X})$ 为充分统计量．

定理 3.5.1 的证明参见韦博成的《参数统计教程》(2006)．该结论的直观含义非常明显，因为统计量 $T(\boldsymbol{X})$ 是对样本的加工、整理，统计量含有的参数 $\boldsymbol{\theta}$ 的信息不可能多于样本含有的参数信息．充分统计量包含了样本 $\boldsymbol{X}$ 中关于参数 $\boldsymbol{\theta}$ 的全部信息，因此其 Fisher 信息与样本 $\boldsymbol{X}$ 的 Fisher 信息相等．

推论 3.5.1　设 $Y=Y(\boldsymbol{X})$ 和 $T=T(\boldsymbol{X})$ 是 C-R 正则族 $\{f(x;\boldsymbol{\theta}),\boldsymbol{\theta}\in\Theta\}$ 上的两个统计量，其 Fisher 信息分别为 $I_Y(\boldsymbol{\theta})$ 和 $I_T(\boldsymbol{\theta})$，存在可测函数 $g(\cdot)$ 满足 $Y=g(T)$，则有

$$I_Y(\boldsymbol{\theta})\leqslant I_T(\boldsymbol{\theta}).$$

3.5.3　信息不等式与有效估计

信息不等式也称为 Cramer-Rao 不等式或 C-R 不等式，用 Fisher 信息表示无偏估计方差(协方差）的下界．

定理 3.5.2　设 $\{f(x;\theta),\theta\in\Theta\}$ 是单参数 Cramer-Rao 正则族，$g(\theta)$ 关于 θ 可导，$T=T(\boldsymbol{X})$