第七章　参数估计

总体是由总体分布来刻画的.在实际问题中我们根据问题本身的专业知识或以往的经验或适当的统计方法,有时可以判断总体分布的类型,但是总体分布的参数还是未知的,需要通过样本来估计.例如,为了研究人们的市场消费行为,我们要先搞清楚人们的收入状况.若假设某城市人均年收入服从正态分布 $N(\mu,\sigma^2)$,但参数 μ 和 σ^2 的具体值并不知道,需要通过样本来估计.又如,假定某城市在单位时间(譬如一个月)内交通事故发生次数服从泊松分布 $P(\lambda)$,其中的参数 λ 也是未知的,同样需要从样本来估计.通过样本来估计总体的参数,这称为参数估计,它是统计推断的一种重要形式.本章我们讨论参数估计的常用方法,估计的优良性以及若干重要总体的参数估计问题.

§7.1　矩估计

设有一个总体 X,为简单计,我们以 $f(x,\theta_1,\cdots,\theta_k)$ 记其概率密度函数或分布律.若总体分布为连续型的,它就是概率密度函数.若总体分布为离散型的,它就是分布律.$\theta_1,\theta_2,\cdots,\theta_k$ 为总体的 k 个未知参数.例如,对正态总体 $N(\mu,\sigma^2)$,它包含两个未知参数 μ 和 σ^2,它的概率密度函数为

$$f(x,\mu,\sigma^2)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(x-\mu)^2},\qquad -\infty<x<\infty.$$

若总体为二项分布 $B(n,p)$,那么根据二项分布的定义,我们知道 n 是试验次数,它是已知的.因此对二项分布只有一个未知参数 p,它的分布律为

$$f(x,p)=\binom{n}{x}p^x(1-p)^{n-x},\ x=0,1,2,\cdots,n.$$

为了估计总体参数 $\theta_1,\cdots,\theta_k$,我们就要从总体中抽出样本,记之为 $X_1,X_2,\cdots,X_n$.我们已经说过,这些样本都是独立同分布的,它们的公共分布就是总体分布 $f(x,\theta_1,\cdots,\theta_k)$.以 θ_1 的估计为例.为了估计 θ_1,需要构造适当的统计量 $\hat{\theta}_1(X_1,X_2,\cdots,X_n)$,它只依赖于样本.也就是说,一旦有了样本 $X_1,\cdots,X_n$,我们就可以算出统计量 $\hat{\theta}_1(X_1,X_2,\cdots,X_n)$ 的一个值,用来作为 θ_1 的估计值.我们称统计量 $\hat{\theta}_1(X_1,\cdots,X_n)$ 为 θ_1 的估计,简记为 $\hat{\theta}_1$.因为未知参数 θ_1

和估计 $\hat{\theta}_1$ 都是实轴上的点，所以这样的估计称为点估计.

矩估计是基于直观考虑而提出的，其方法比较简单. 对总体 X，它的 m 阶原点矩为

$$\alpha_m = E(X^m) = \int_{-\infty}^{+\infty} x^m f(x,\theta_1,\cdots,\theta_k)\mathrm{d}x.$$

若是离散型分布，这里的积分号应改为求和号. 而 m 阶样本原点矩为

$$A_m = \frac{1}{n}\sum_{i=1}^{n} X_i^m.$$

一般说来，α_m 是总体参数 $\theta_1,\cdots,\theta_k$ 的函数，记之为 $\alpha_m(\theta_1,\cdots,\theta_k)$. 因此，如果命总体的前 k 阶原点矩与同阶样本原点矩相等，就得到关于 $\theta_1,\cdots,\theta_k$ 的一个方程组

$$\alpha_m(\theta_1,\cdots,\theta_k) = A_m,\ m = 1,2,\cdots,k. \tag{7.1.1}$$

解这个方程组，其解记为

$$\hat{\theta}_i = \hat{\theta}_i(X_1,X_2,\cdots,X_n),\ i = 1,\cdots,k.$$

它们就可以做为 θ_i 的估计. 这样的估计叫做矩估计.

矩估计法的理论背景是：因为样本 $X_1,X_2,\cdots,X_n$ 是独立同分布的，于是 $X_1^m,X_2^m,\cdots,X_n^m$ 也是独立同分布的，因而 $E(X_1^m) = \cdots = E(X_n^m) = \alpha_m$. 按照大数定律，样本原点矩 A_m 作为 $X_1^m,X_2^m,\cdots,X_n^m$ 的算术平均值依概率收敛到均值 $\alpha_m = E(X_i^m)$，即

$$A_m = \frac{1}{n}\sum_{i=1}^{n} X_i^m \longrightarrow \alpha_m(\text{依概率}).$$

于是，对充分大的 n，近似地有 $\alpha_m(\theta_1,\cdots,\theta_k) \approx A_m$，将“近似等于号”换成“等号”就得到了(7.1.1).

设总体 X 的均值为 μ，方差为 σ^2，于是

$$\begin{cases} \alpha_1 = E(X) = \mu, \\ \alpha_2 = E(X^2) = \mathrm{Var}(X) + [E(X)]^2 = \sigma^2 + \mu^2. \end{cases}$$

对 $m = 1,2$，方程组(7.1.1)变为

$$\begin{cases} \mu = \bar{X}, \\ \sigma^2 + \mu^2 = \dfrac{1}{n}\sum_{i=1}^{n} X_i^2. \end{cases}$$

解这个方程组,得到 μ 和 σ^2 的矩估计

$$\begin{cases} \hat{\mu} = \bar{X}, \\ \hat{\sigma}^2 = \dfrac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2. \end{cases}$$

于是,我们得到如下结论:对一切均值为 μ,方差为 σ^2 的总体,不管总体的具体分布形式如何,μ 和 σ^2 的矩估计总是

$$\begin{cases} \hat{\mu} = \bar{X}, \\ \hat{\sigma}^2 = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2. \end{cases} \tag{7.1.2}$$

需要读者特别注意,方差的矩估计并不等于样本方差 S^2,而是等于 $\hat{\sigma}^2 = \dfrac{n-1}{n}S^2$.

例 7.1.1 对正态总体 $N(\mu,\sigma^2)$,因为 μ 和 σ^2 分别为总体均值和方差,所以,它们的矩估计为 $\hat{\mu} = \bar{X}$, $\hat{\sigma}^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/n$.

例 7.1.2 设 $X_1, X_2, \cdots, X_n$ 为从定义在 $[a,b]$ 上的均匀分布的总体抽取的样本,试导出 a 和 b 的矩估计.

根据 §4.1 和 §4.2 知,总体 X 的均值 $E(X) = (a+b)/2$ 和方差 $\mathrm{Var}(X) = (b-a)^2/12$,于是由(7.1.2)得

$$\begin{cases} \dfrac{b+a}{2} = \bar{X}, \\ \dfrac{(b-a)^2}{12} = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2. \end{cases}$$

由此方程组解得 a 和 b 的矩估计分别为

$$\hat{a} = \bar{X} - \sqrt{3\hat{\sigma}^2} = \bar{X} - \sqrt{3}\hat{\sigma},$$

$$\hat{b} = \bar{X} + \sqrt{3\hat{\sigma}^2} = \bar{X} + \sqrt{3}\hat{\sigma},$$

这里 $\hat{\sigma} = (\hat{\sigma}^2)^{1/2} = \left(\dfrac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^{1/2}$.

例 7.1.3 求事件发生概率 p 的矩估计

记事件 A 发生的概率 $P(A) = p$,定义随机变量

$$X = \begin{cases} 1, & \text{若在一次试验中事件 } A \text{ 发生}, \\ 0, & \text{若在一次试验中事件 } A \text{ 不发生}, \end{cases}$$

于是 $E(X) = p$,即我们所求的事件 A 发生的概率等于随机变量 X 的均值.现

有样本 $X_1,\cdots,X_n$，也就是说，我们做了 n 次试验，观测到

$$X_i=\begin{cases}1, & \text{若在第 } i \text{ 次试验中事件 } A \text{ 发生},\\ 0, & \text{若在第 } i \text{ 次试验中事件 } A \text{ 不发生}.\end{cases}$$

根据(7.1.2)，p 的矩估计为

$$\hat{p}=\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i,$$

注意，这里 $\sum_{i=1}^{n}X_i$ 是在 n 次试验中事件 A 出现的次数. 因而，$\bar{X}$ 是事件 A 出现的频率. 于是，我们的结论可叙述为：频率是概率的矩估计.

例 7.1.4　设总体为泊松分布 $P(\lambda)$，$X_1,X_2,\cdots,X_n$ 为从该总体抽取的样本. 因为 $E(X)=\lambda$，所以，据(7.1.2)中第一式，λ 的矩估计为 $\hat{\lambda}=\bar{X}$. 另一方面，λ 也是总体的方差，据(7.1.2)中第二式 λ 的矩估计为 $\hat{\lambda}=\hat{\sigma}^2$. 这样，一个参数 λ 就有了两个不同的矩估计. 在实际应用中，我们究竟采用哪一个呢？这除了要考虑在后面我们将要讨论的估计优良性的标准外，一般选用阶数较低的样本矩. 在本例中，$\bar{X}$ 是一阶样本原点矩，$\hat{\sigma}^2$ 是二阶样本中心矩，所以，我们采用 $\bar{X}$ 作为 λ 的矩估计.

§7.2　极大似然估计

设总体分布为 $f(x,\theta_1,\cdots,\theta_k)$，$X_1,\cdots,X_n$ 为从该总体抽出的样本. 因为 $X_1,\cdots,X_n$ 相互独立且同分布，于是，它们的联合密度函数为

$$L(x_1,\cdots,x_n;\theta_1,\cdots,\theta_k)=\prod_{i=1}^{n}f(x_i,\theta_1,\cdots,\theta_k).$$

这里 $\theta_1,\cdots,\theta_k$ 被看作固定但是未知的参数. 反过来，如果我们把 $x_1,\cdots,x_n$ 看成固定的，则 $L(x_1,\cdots,x_n;\theta_1,\cdots,\theta_k)$ 就是 $\theta_1,\cdots,\theta_k$ 的函数，这时我们把它称为似然函数.

假定现在我们已经观测到一组样本 $X_1,\cdots,X_n$，要去估计未知参数 $\theta_1,\cdots,\theta_k$. 一种直观的想法是，哪一组参数值使现在的样本 $X_1,\cdots,X_n$ 出现的可能性最大，哪一组参数可能就是真正的参数，我们就要用它作为参数的估计值. 这里，假定我们已知一组样本 $X_1,X_2,\cdots,X_n$. 如果对参数的两组不同的值 $\theta'_1,\cdots,\theta'_k$ 和 $\theta''_1,\cdots,\theta''_k$，似然函数有如下关系

$$L(x_1,\cdots,x_n;\theta'_1,\cdots,\theta'_k)>L(x_1,\cdots,x_n;\theta''_1,\cdots,\theta''_k),$$

那么，从 $L(x_1,\cdots,x_n;\theta_1,\cdots,\theta_k)$ 又是概率密度函数的角度来看，上式的意义就是参数 $\theta_1',\cdots,\theta_k'$ 使 $X_1,X_2,\cdots,X_n$ 出现的可能性比参数 $\theta_1'',\cdots,\theta_k''$ 使 $X_1,X_2,\cdots,X_n$ 出现的可能性大，当然参数 $\theta_1',\cdots,\theta_k'$ 比 $\theta_1'',\cdots,\theta_k''$ 更像是真正的参数. 这样的分析就导致了参数估计的一种方法，即用使似然函数达到最大值的点 $(\theta_1^*,\cdots,\theta_k^*)$，作为未知参数的估计，这就是所谓的极大似然估计.

现在我们讨论求极大似然估计的具体方法. 为简单起见，以下记 $L(\theta)=L(x_1,\cdots,x_n;\theta_1,\cdots,\theta_k)$. 求 θ 的极大似然估计就归结为求 $L(\theta)$ 的最大值点. 由于对数函数是单调增函数，所以

$$\log L(\theta)=\sum_{i=1}^{n}\log f(x_i;\theta_1,\cdots,\theta_k) \tag{7.2.1}$$

与 $L(\theta)$ 有相同的最大值点. 而在许多情况下，求 $\log L(\theta)$ 的最大值点比较简单，于是，我们就将求 $L(\theta)$ 的最大值点改为求 $\log L(\theta)$ 的最大值点. 对 $\log L(\theta)$ 关于 $\theta_1,\cdots,\theta_k$ 求导数，并命其等于零，得到方程组

$$\frac{\partial \log L(\theta)}{\partial \theta_i}=0,\quad i=1,\cdots,k \tag{7.2.2}$$

称为似然方程组. 解这个方程组，得到 $\log L(\theta)$ 的驻点，如果驻点是唯一的，又能验证它是一个极大值点，则它必是 $\log L(\theta)$，也就是 $L(\theta)$ 的最大值点，即为所求的极大似然估计. 许多常用的重要例子多属于这种情况. 然而在一些情况下，问题比较复杂，似然方程组的解可能不唯一，这时就需要进一步判定哪一个是最大值点.

还需要指出，若函数 $\log f(x,\theta_1,\cdots,\theta_k)$ 关于 $\theta_1,\cdots,\theta_k$ 的导数不存在时，我们就无法得到似然方程组(7.2.2)，这时就必须根据极大似然估计的定义直接去求 $L(\theta)$ 的最大值点.

在一些情况下，我们需要估计 $g(\theta_1,\cdots,\theta_k)$. 如果 $\theta_1^*,\cdots,\theta_k^*$ 分别是 $\theta_1,\cdots,\theta_k$ 的极大似然估计，则称 $g(\theta_1^*,\cdots,\theta_k^*)$ 为 $g(\theta_1,\cdots,\theta_k)$ 的极大似然估计.

下面我们举一些例子来说明求极大似然估计的方法.

例 7.2.1 设从正态总体 $N(\mu,\sigma^2)$ 抽出样本 $X_1,\cdots,X_n$，这里未知参数为 μ 和 σ^2(注意我们把 σ^2 看作一个参数). 似然函数为

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

它的对数为

$$\log L(\mu,\sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2,$$

似然方程组为

$$\begin{cases} \dfrac{\partial \log L(\mu,\sigma^2)}{\partial \mu} = \dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n}(x_i-\mu) = 0, \\ \dfrac{\partial \log L(\mu,\sigma^2)}{\partial \sigma^2} = -\dfrac{n}{2\sigma^2} + \dfrac{1}{2\sigma^4}\displaystyle\sum_{i=1}^{n}(x_i-\mu)^2 = 0, \end{cases}$$

由第一式解得

$$\mu^* = \bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i, \tag{7.2.3}$$

代入第二式得

$$\sigma^{*2} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2. \tag{7.2.4}$$

似然方程组有唯一解(μ^*,σ^{*2})，而且它一定是最大值点，这是因为当 $|\mu|\to\infty$ 或 $\sigma^2\to 0$ 或 ∞ 时，非负函数 $L(\mu,\sigma^2)\to 0$. 于是，μ 和 σ^2 的极大似然估计为

$$\mu^* = \bar{X},\quad \sigma^{*2} = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2. \tag{7.2.5}$$

这里，我们用大写字母表示所有涉及的样本，因为极大似然估计 μ^* 和 σ^{*2} 都是统计量，离开了具体的一次试验或观测，它们都是随机的.

例 7.2.2　设总体 X 服从参数为 λ 的泊松分布，它的分布律为

$$P(X=x) = \frac{\mathrm{e}^{-\lambda}\lambda^x}{x!},\quad x = 0,1,2\cdots.$$

有了样本 $X_1,\cdots,X_n$ 之后，参数 λ 的似然函数为

$$L(\lambda) = \prod_{i=1}^{n}\frac{\mathrm{e}^{-\lambda}\lambda^{x_i}}{x_i!} = \mathrm{e}^{-n\lambda}\cdot\frac{\lambda^{\sum_{i=1}^{n}x_i}}{\prod_{i=1}^{n}x_i!},$$

似然方程为

$$\frac{\partial \log L(\lambda)}{\partial \lambda} = -n + \frac{1}{\lambda}\sum_{i=1}^{n}x_i = 0,$$

解得

$$\lambda^* = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

因为 $\log L(\lambda)$ 的二阶导数总是负值，可见，似然函数在 λ^* 处达到最大值. 所以，$\lambda^* = \bar{X}$ 是 λ 的极大似然估计.

例 7.2.3 求事件发生概率 p 的极大似然估计.

设事件 A 发生的概率 $P(A) = p$，定义随机变量

$$X = \begin{cases} 1, & \text{若在一次试验中事件 } A \text{ 发生}, \\ 0, & \text{若在一次试验中事件 } A \text{ 不发生}. \end{cases}$$

则 $X \sim B(1,p)$，它的分布律为

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

现有样本 $X_1, X_2, \cdots, X_n$，故似然函数为

$$L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i},$$

它的对数为

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

似然方程为

$$\frac{\mathrm{d}\log L(p)}{\mathrm{d}p} = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{\left(n - \sum_{i=1}^{n} x_i\right)}{1-p} = 0.$$

设方程的解

$$p^* = \bar{X}.$$

注意到 $\sum_{i=1}^{n} x_i \leqslant n$，很容易验证，$\log L(p)$ 的二阶导数在 $\bar{x}$ 处取负值，于是 $\bar{x}$ 是 $\log L(p)$ 的最大值点. 因而 $\bar{X}$ 是 p 的极大似然估计. 因为 $\bar{X}$ 是事件 A 在 n 次试验中出现的频率，于是，我们得到如下结论：频率是概率的极大似然估计.

结合例 7.1.3，我们看到，频率既是概率的矩估计，又是概率的极大似然估计，这就为我们在各种具体场合用频率去估计概率提供了理论依据. 例如，当我们要估计一种产品的合格率 p 时，就可以随机抽取这种产品 N 件进行检

查，若发现其中有 n 件合格品，那么 $\hat{p} = n/N$ 就是该产品合格率的矩估计和极大似然估计. 又如，据记载，1982 年在北京某家医院出生的 1449 名婴儿中有男性 754 人. 那么男婴出生率的估计就是 $\hat{p} = 754/1449 = 52.6\%$，这个数字与从遗传学原理算出的男婴出生率 $22/43 \approx 51.2\%$ 很接近.

例 7.2.4　设总体 X 为 $[a,b]$ 上的均匀分布，求 a,b 的极大似然估计.

X 的概率密度函数为

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & \text{当 } a \leqslant x \leqslant b \text{ 时,} \\ 0, & \text{其它.} \end{cases}$$

对样本 $X_1, X_2, \cdots, X_n$，

$$L(a,b) = \begin{cases} \dfrac{1}{(b-a)^n}, & \text{若 } a \leqslant x_i \leqslant b, i = 1, \cdots, n, \\ 0, & \text{其它.} \end{cases}$$

很显然，$L(a,b)$ 作为 a 和 b 的二元函数是不连续的. 这时我们不能用似然方程组 (7.2.2) 来极大似然估计，而必须从极大似然估计的定义出发，求 $L(a,b)$ 的最大值. 为使 $L(a,b)$ 达到最大，$b-a$ 应该尽量地小，但 b 又不能小于 $\max\{x_1, \cdots, x_n\}$，否则，$L(a,b) = 0$. 类似地，a 不能大过 $\min\{x_1, \cdots, x_n\}$. 因此，a 和 b 的极大似然估计为

$$a^* = \min\{X_1, \cdots, X_n\},$$

$$b^* = \max\{X_1, \cdots, X_n\}.$$

到现在为止，我们以正态分布，泊松分布，均匀分布的参数以及事件发生的概率的估计为例子讨论了矩估计和极大似然估计. 在我们所举的例子中，除了均匀分布外，两种估计都是一致的. 矩估计的优点是简单，只需知道总体的矩，总体的分布形式不必知道. 而极大似然估计则必须知道总体分布形式，并且在一般情况下，似然方程组的求解较复杂，往往需要在计算机上通过迭代运算才能计算出其近似解.

§7.3　估计量的优良性准则

从前面两节的讨论中我们看到，有时候同一个参数可以有几种不同的估计，这时就存在采用哪一个估计的问题. 另一方面，对一个参数，用矩法和极大似然法即使得到同一种估计，也存在一个衡量这个估计优劣的问题. 本节我们

就讨论评价一个估计的标准问题.

一、无偏性

假设总体分布的参数为 θ,设 $\hat{\theta}(X_1,X_2,\cdots,X_n)$ 是 θ 的一个估计,它是一个统计量.对于不同的样本 $X_1,X_2,\cdots,X_n$,估计 $\hat{\theta}$ 取不同的值.如果 $\hat{\theta}(X_1,X_2,\cdots,X_n)$ 的均值等于未知参数 θ,即

$$E[\hat{\theta}(X_1,X_2,\cdots,X_n)]=\theta,\text{对一切可能的 }\theta\text{ 成立,} \tag{7.3.1}$$

则称 $\hat{\theta}$ 为 θ 的无偏估计.无偏性的意义是,用一个估计量 $\hat{\theta}(X_1,X_2,\cdots,X_n)$ 去估计未知参数 θ,有时候可能偏高,有时候可能偏低,但是平均来说它等于未知参数 θ.

在(7.3.1)式中,"一切可能的 θ"是指每个具体参数估计问题中,参数 θ 取值范围内的一切可能的值.例如,若 θ 是正态总体 $N(\mu,\sigma^2)$ 的均值 μ,那么,它的一切可能取值范围是 $(-\infty,\infty)$.若 θ 为方差 σ^2,则它的取值范围是 $(0,\infty)$.我们之所以要求(7.3.1)式对一切可能的 θ 都成立,是因为在参数估计问题中,我们并不知道参数的真值.所以,当我们要求一个估计量具有无偏性时,自然要求它在参数的一切可能取值范围内处处都是无偏的.

例 7.3.1 设 $X_1,X_2,\cdots,X_n$ 为抽自均值为 μ 的总体的样本,考虑 μ 的估计量

$$\hat{\mu}_1=X_1,$$

$$\hat{\mu}_2=\frac{X_1+X_2}{2},$$

$$\hat{\mu}_3=\frac{X_1+X_2+X_{n-1}+X_n}{4}\quad(\text{假设 }n\geqslant 4).$$

因为 $E(X_i)=\mu$,容易验证,$E(\hat{\mu}_i)=\mu$,$i=1,2,3$,所以 $\hat{\mu}_1,\hat{\mu}_2$ 和 $\hat{\mu}_3$ 都是 μ 的无偏估计.但是

$$\hat{\mu}_4=2X_1,$$

$$\hat{\mu}_5=\frac{X_1+X_2}{3}$$

都不是 μ 的无偏估计.

定理 7.3.1 设总体均值为 μ,方差为 σ^2,$X_1,X_2,\cdots,X_n$ 为来自该总体的样本,则

(1) $E(\bar{X}) = \mu$,

(2) $E(S^2) = \sigma^2$,

这里 $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$ 为样本方差. 即样本均值与样本方差分别为总体均值和总体方差的无偏估计.

证明　(1) 因为 $X_1, X_2, \cdots, X_n$ 是同分布的,于是 $E(X_i) = \mu$,故有

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

(2) 因为

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \bar{X})^2 &= \sum_{i=1}^{n}X_i^2 - 2(\sum_{i=1}^{n}X_i)\bar{X} + n\bar{X}^2 \\ &= \sum_{i=1}^{n}X_i^2 - n\bar{X}^2,\end{aligned}$$

注意到

$$E(X_i^2) = \mathrm{Var}(X_i) + [E(X_i)]^2 = \sigma^2 + \mu^2,$$

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

于是

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

定理证毕.

在前两节中,我们曾经用矩法和极大似然法求得 σ^2 的估计,两者是相同的,皆为

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

很明显,它不是 σ^2 的无偏估计. 这就是我们为什么把 $\hat{\sigma}^2$ 的分母 n 修正为 $n-1$ 获得样本方差 S^2 的理由.

若 $\hat{\theta}$ 为 θ 的一个估计,$g(\theta)$ 为 θ 的一个实值函数,通常我们总是用 $g(\hat{\theta})$ 去估计 $g(\theta)$. 但是,需要注意的是,即便 $E(\hat{\theta}) = \theta$,也不一定有 $E[g(\hat{\theta})] = g(\theta)$. 也就是说,由 $\hat{\theta}$ 是 θ 的无偏估计,不能断言 $g(\hat{\theta})$ 是 $g(\theta)$ 的无偏估计.

例 7.3.2 样本标准差 S 不是总体标准差 σ 的无偏估计.

事实上,由于 $\sigma^2 = E(S^2) = \mathrm{Var}(S) + [E(S)]^2$,并注意到方差总是非负的,即 $\mathrm{Var}(S) \geqslant 0$,故有 $\sigma^2 \geqslant [E(S)]^2$,于是我们得到

$$E(S) \leqslant \sigma. \tag{7.3.2}$$

这表明,虽然样本方差是总体方差的无偏估计,但是它的平方根即样本标准差并不是总体标准差的无偏估计.(7.3.2) 式表明用样本标准差去估计总体标准差,平均来说是偏低的.

二、均方误差准则

用估计量 $\hat{\theta}$ 去估计 θ,其误差为 $\hat{\theta} - \theta$,它随样本 $X_1, X_2, \cdots, X_n$ 的值而定,也是随机的,即 $\hat{\theta} - \theta$ 是随机变量.我们对它求均值,为了防止求均值时正误差和负误差相互抵消,我们先将其平方再求均值,并将其称为均方误差,记为 $\mathrm{MSE}(\hat{\theta})$,即

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

这个量越小,表示用 $\hat{\theta}$ 去估计 θ 时平均误差就越小,因而也就越优.均方误差能够分解成两部分

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + (E\hat{\theta} - \theta)^2, \tag{7.3.3}$$

这个式子的证明是很容易的.事实上

$$\begin{aligned}\mathrm{MSE}(\hat{\theta}) &= E(\hat{\theta} - \theta)^2 = E[(\hat{\theta} - E\hat{\theta}) + (E\hat{\theta} - \theta)]^2 \\ &= E(\hat{\theta} - E\hat{\theta})^2 + 2E(\hat{\theta} - E\hat{\theta})(E\hat{\theta} - \theta) + (E\hat{\theta} - \theta)^2 \\ &= \mathrm{Var}(\hat{\theta}) + (E\hat{\theta} - \theta)^2.\end{aligned}$$

(7.3.3) 式表明,均方误差由两部分构成.第一部分是估计量 $\hat{\theta}$ 的方差,第二部分是估计量的偏差 $E\hat{\theta} - \theta$ 的平方.如果一个估计量是无偏的,那么这第二部分就是零,这时

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}). \tag{7.3.4}$$

这就是说,如果限定在无偏估计里,则均方误差准则就变成了方差准则.这时两个估计中哪一个估计的方差小,哪一个估计就比较优.

例 7.3.3 设 $X_1, X_2, \cdots, X_n$ 抽自均值为 μ 的总体,考虑 μ 的如下两个估计

$$\hat{\mu} = \bar{X},$$

$$\hat{\mu}_{(-i)} = \sum_{j \neq i} X_j \Big/ (n-1),$$

这里 $\hat{\mu}_{(-i)}$ 表示去掉第 i 个样本 X_i 后，对其余 $n-1$ 个样本所求的样本均值. 显然，$\hat{\mu}$ 和 $\hat{\mu}_{(-i)}$ 都是 μ 的无偏估计，但是，因为 $X_1, X_2, \cdots, X_n$ 都是独立同分布的，且 $\mathrm{Var}(X_i) = \sigma^2$，于是

$$\mathrm{Var}(\hat{\mu}) = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{\sigma^2}{n},$$

$$\mathrm{Var}(\hat{\mu}_{(-i)}) = \frac{\sigma^2}{n-1}.$$

可见，$\hat{\mu} = \bar{X}$ 比 $\hat{\mu}_{(-i)}$ 有较小的方差，因而 $\bar{X}$ 优于 $\hat{\mu}_{(-i)}$. 这表明，当我们用样本均值去估计总体均值时，使用全体样本总比不使用全体样本要好.

§7.4　正态总体的区间估计(一)

在日常生活中，当我们估计一个未知量的时候，通常采用两种方法. 一种方法是用一个数，也就是用实轴上的一个点去估计，我们称它为点估计. 前面讨论的矩估计和极大似然估计都属于这种情况. 另一种方法是采用一个区间去估计未知量，例如，估计某人的身高在 170 厘米到 180 厘米之间；明天北京的最高气温在 30 ～ 32℃ 之间等等. 这类估计称为区间估计.

不难看出，区间估计的长度度量了该区间估计的精度. 区间估计的长度愈长，它的精度也就愈低. 例如：估计某人的身高. 甲估计他是在 170 ～ 180 厘米之间，而乙估计他是在 150 ～ 190 厘米之间，显然，甲的区间估计较乙的短，因而精度较高. 但是，这个区间短，包含该人真正身高的可能性即概率就小. 我们把这个概率称为区间估计的可靠度. 那么，乙的区间估计的长度长，精度差，但可靠度比甲的大. 由此可见，在区间估计中，精度(用区间估计的长度来度量)和可靠度(用估计的区间包含未知量的概率来度量)是相互矛盾着的. 在实际问题中，我们总是在保证可靠度的条件下，尽可能地提高精度. 下面我们来讨论如何构造未知参数的区间估计.

在统计文献中，将可靠度称为“置信系数”. 区间估计也常常称为“置信区间”.

定义 7.4.1　设 $X_1, X_2, \cdots, X_n$ 为从总体中抽出的样本，θ 为总体中未知参数. 记 $\theta_1 = \theta_1(X_1, X_2, \cdots, X_n)$，$\theta_2 = \theta_2(X_1, X_2, \cdots, X_n)$ 为两个统计量，对

给定的 $\alpha(0<\alpha<1)$,若

$$P\{\theta\in[\hat{\theta}_1,\hat{\theta}_2]\}=1-\alpha, \tag{7.4.1}$$

则称区间$[\hat{\theta}_1,\hat{\theta}_2]$为 θ 的置信系数为 $1-\alpha$ 的置信区间.

需要特别强调的是,置信区间$[\hat{\theta}_1,\hat{\theta}_2]$是一个随机区间,对一个给定的样本 $X_1,X_2,\cdots,X_n$,这个区间可能包含未知参数 θ,也可能不包含.但(7.4.1)式表明,对置信系数 $1-\alpha$ 的置信区间,它包含未知参数的概率是 $1-\alpha$.一般在应用上,取 $\alpha=0.05$ 的最多,这时置信系数 $1-\alpha=0.95$,那么置信区间包含未知参数的概率就是 95%.当然也可以取 $\alpha=0.01,0.10$ 等等.

现在我们来讨论正态总体参数的区间估计.

设 $X_1,X_2,\cdots,X_n$ 为来自正态总体$N(\mu,\sigma^2)$的样本,σ^2 已知,求均值 μ 的置信系数为 $1-\alpha$ 的置信区间.根据基本定理(见定理6.4.1),$\bar{X}\sim N(\mu,\frac{\sigma^2}{n})$,于是

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1). \tag{7.4.2}$$

记 $\Phi(x)$ 为 $N(0,1)$ 的分布函数,Z_α 为其上α 分位点,即 $\Phi(Z_\alpha)=1-\alpha$.于是

$$P\left\{\left|\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right|\leqslant Z_{\alpha/2}\right\}=1-\alpha, \tag{7.4.3}$$

等价地

$$P\left\{\bar{X}-\frac{\sigma}{\sqrt{n}}Z_{\alpha/2}\leqslant\mu\leqslant\bar{X}+\frac{\sigma}{\sqrt{n}}Z_{\alpha/2}\right\}=1-\alpha,$$

这样,我们就得到了 μ 的置信系数为 $1-\alpha$ 的置信区间

$$\left[\bar{X}-\frac{\sigma}{\sqrt{n}}Z_{\alpha/2},\ \bar{X}+\frac{\sigma}{\sqrt{n}}Z_{\alpha/2}\right]. \tag{7.4.4}$$

这个区间估计的长度为 $2\sigma Z_{\alpha/2}/\sqrt{n}$,它刻画了此区间估计的精度.从这个例子可以看出:

(1) 置信系数愈大,α 就愈小,因而 $Z_{\alpha/2}$ 就越大,这时区间估计的长度越长,精确度就越小.

(2) 样本大小 n 越大,区间估计的长度越短,因而精度也就越高.这是在情理中的事,样本个数增加,就意味着从样本中获得的关于 μ 的信息增加了,自然应该构造出比较短的区间估计.

例 7.4.1　某工厂生产的零件长度 X 被认为服从 $N(\mu,0.04)$，现从该产品中随机抽取 6 个，其长度的测量值如下(单位：毫米)

14.6，15.1，14.9，14.8，15.2，15.1.

试求该零件长度的置信系数为 0.95 的区间估计.

解　$\alpha = 0.05, n = 6, Z_{\alpha/2} = Z_{0.025} = 1.96, \sigma^2 = 0.04$，因而

$$\frac{\sigma}{\sqrt{n}} Z_{\alpha/2} = \frac{\sqrt{0.04}}{\sqrt{6}} 1.96 = 0.16,$$

$$\bar{X} = 14.95.$$

所以，我们得到该零件长度的置信系数为 95% 的置信区间为 $[14.79, 15.11]$.

现在我们对区间 $[14.79, 15.11]$ 包含 μ 的置信系数为 0.95 这句话作一些解释. 因为现在的区间 $[14.79, 15.11]$ 是固定的，不再是随机区间，它要么包含 μ，要么不包含 μ，两者必居其一，因此，从字面上看置信系数已没有实际意义. 这里的置信系数 0.95 是指，如果我们把上述抽样多次重复，构造出很多个这样的区间，它们包含 μ 的频率大约是 95%. 因此，置信系数实际上是对构造置信区间的这种方法的可靠程度的整体评价.

上面讨论的是正态总体 $N(\mu,\sigma^2)$，方差 σ^2 已知的情形. 但在应用上，σ^2 往往是未知的，它是通过样本方差

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

来估计的，这时，依据基本定理(定理 6.4.1)，我们有

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}, \tag{7.4.5}$$

和(7.4.2)相比，我们是用 S 代替了 σ，用 t_{n-1} 代替了 $N(0,1)$，于是，不难得到 μ 的置信系数为 $1-\alpha$ 的置信区间为

$$\left[\bar{X} - \frac{S}{\sqrt{n}} t_{n-1}\left(\frac{\alpha}{2}\right), \bar{X} + \frac{S}{\sqrt{n}} t_{n-1}\left(\frac{\alpha}{2}\right)\right], \tag{7.4.6}$$

这里 $t_{n-1}(\alpha)$ 表示自由度为 $n-1$ 的 t 分布的上 α 分位点.

例 7.4.2　为了估计一件物体的重量 μ，将其称了 10 次，得到的重量(单位：千克)为

10.1，10，9.8，10.5，9.7，10.1，9.9，10.2，10.3，9.9

假设所称出的物体重量都服从 $N(\mu,\sigma^2)$，求该物体重量 μ 的置信系数为 0.95 的置信区间.

解 $\alpha = 0.05, n = 10, t_9(0.025) = 2.2622, \bar{X} = 10.05,$

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

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

$$= \frac{1}{9}\left[1010.55 - \frac{1}{10}\times 10100.25\right]$$

$$= \frac{0.525}{9} = 0.0583,$$

故 $S = 0.24$.

将这些数据代入(7.4.6)式得到 μ 的置信系数为 0.95 的置信区间为 $[9.87, 10.22]$.

通过上面的例子,我们可以把构造未知参数 θ 的置信区间的方法归纳为

(1) 寻找样本 $X_1, \cdots, X_n$ 和未知参数 θ 的一个函数 $g(X_1, \cdots, X_n; \theta)$,其分布完全已知,且这个分布与参数 θ 无关(如:(7.4.2)式的 $(\bar{X} - \mu)\Big/(\sigma/\sqrt{n})$ 和(7.4.5)的 $(\bar{X} - \mu)\Big/(S/\sqrt{n})$ 就是这样的一个函数 $g(X_1, \cdots, X_n; \mu)$,它们的分布分别为 $N(0,1)$ 和 t_{n-1},与参数 μ 无关).

(2) 对给定的置信系数 $1-\alpha$,根据 $g(X_1, \cdots, X_n; \theta)$ 的分布函数,确定出 a 和 b 使

$$P\{a \leqslant g(X_1, \cdots, X_n; \theta) \leqslant b\} = 1 - \alpha.$$

(3) 解不等式 $a \leqslant g(X_1, \cdots, X_n; \theta) \leqslant b$,得到 $\hat{\theta}_1 \leqslant \theta \leqslant \hat{\theta}_2$,此即为 θ 的置信系数为 $1-\alpha$ 的置信区间.

现在我们按照上述步骤来构造正态总体 $N(\mu, \sigma^2)$ 中,σ^2 的置信区间.

(1) 由基本定理(定理 6.4.1)知

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

这里 $\frac{(n-1)S^2}{\sigma^2}$ 就是我们要寻找的函数 $g(X_1, \cdots, X_n; \sigma^2)$,且它的分布 χ^2_{n-1} 与 σ^2 无关.

(2) 对给定的 $1-\alpha$,我们取

$$a = \chi^2_{n-1}(1-\frac{\alpha}{2}),\quad b = \chi^2_{n-1}(\frac{\alpha}{2}),$$

它们是χ^2_{n-1}分布的两个分位点,如图 7.4.1 所示.则有

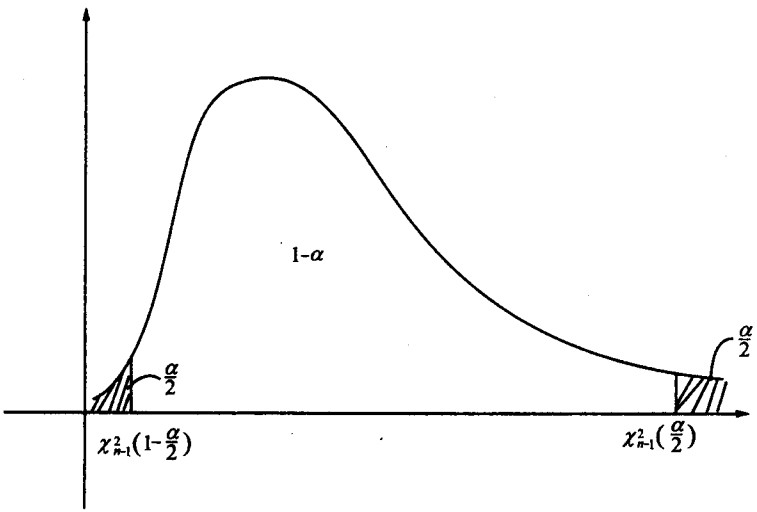

图 7.4.1　χ^2_{n-1}分布的分位点

$$P\left\{\chi^2_{n-1}\left(1-\frac{\alpha}{2}\right)\leqslant\frac{(n-1)S^2}{\sigma^2}\leqslant\chi^2_{n-1}\left(\frac{\alpha}{2}\right)\right\}=1-\alpha.$$

(3) 由$\chi^2_{n-1}\left(1-\frac{\alpha}{2}\right)\leqslant\frac{(n-1)S^2}{\sigma^2}\leqslant\chi^2_{n-1}\left(\frac{\alpha}{2}\right)$解得

$$\theta_1=\frac{(n-1)S^2}{\chi^2_{n-1}\left(\frac{\alpha}{2}\right)},\ \theta_2=\frac{(n-1)S^2}{\chi^2_{n-1}\left(1-\frac{\alpha}{2}\right)},$$

故 σ^2 的置信系数为 $1-\alpha$ 的置信区间为

$$\left[\frac{(n-1)S^2}{\chi^2_{n-1}\left(\frac{\alpha}{2}\right)},\frac{(n-1)S^2}{\chi^2_{n-1}\left(1-\frac{\alpha}{2}\right)}\right].\tag{7.4.7}$$

例 7.4.3(续例 7.4.2)　求 σ^2 的置信系数为 0.95 的置信区间

解　$\alpha=0.05,\ n=10,\ \chi^2_{n-1}\left(\frac{\alpha}{2}\right)=\chi^2_9(0.025)=19.023,$

$$\chi^2_{n-1}\left(1-\frac{\alpha}{2}\right)=\chi^2_9(0.975)=2.70,$$

$$S^2=0.0583.$$

代入(7.4.7)得,σ^2 的置信系数为 0.95 的置信区间为[0.028,0.194].

§7.5 正态总体的区间估计(二)

两个正态总体的区间估计问题,在实际应用中经常会遇到.例如,如果要考察一项新技术对提高产品的某项质量指标的作用,我们把实施新技术前产品的质量指标看成一个正态总体 $N(\mu_1,\sigma_1^2)$,而把实施新技术后产品质量指标看成另一个正态总体 $N(\mu_2,\sigma_2^2)$.于是,评价此新技术的效果问题,就归结为研究两个正态总体均值之差 $\mu_1-\mu_2$ 的问题.又如,当我们要比较甲、乙两厂生产某种药物的治疗效果时,可以把两个厂的药效分别看成服从正态分布的两个总体,那么,两厂药效的差异,也就是两个正态总体均值的差异.从而,评价两厂生产的药物的效果,就归结为研究对应的两个正态总体的均值之差.类似的例子可以举出很多,说明两个正态总体的区间估计问题,有很广泛的实际意义.本节我们讨论如何构造两个正态总体均值之差的区间估计.

设 $X_1,\cdots,X_m$ 是抽自正态总体$N(\mu_1,\sigma_1^2)$ 的样本,$Y_1,\cdots,Y_n$ 是抽自正态总体 $N(\mu_2,\sigma_2^2)$ 的样本,$\bar{X}$ 和 $\bar{Y}$ 分别为它们的样本均值,样本方差分别为

$$S_1^2=\frac{1}{m-1}\sum_{i=1}^{m}(X_i-\bar{X})^2,$$

$$S_2^2=\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2.$$

定理 7.5.1 (1)

$$\bar{X}-\bar{Y}\sim N\left(\mu_1-\mu_2,\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}\right),\tag{7.5.1}$$

或

$$\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}}\sim N(0,1).\tag{7.5.2}$$

(2) 设 $\sigma_1^2=\sigma_2^2=\sigma^2$,但 σ^2 未知,则

$$\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S\sqrt{\frac{1}{m}+\frac{1}{n}}}\sim t_{m+n-2},\tag{7.5.3}$$

这里

$$S^2 = \frac{(m-1)S_1^2 + (n-1)S_2^2}{m+n-2} = \frac{m-1}{m+n-2}S_1^2 + \frac{n-1}{m+n-2}S_2^2 \tag{7.5.4}$$

是 S_1^2 和 S_2^2 的加权平均.

证明　(1) 由基本定理(见定理 6.4.1) 知

$$\overline{X} \sim N(\mu_1, \frac{\sigma_1^2}{m}),\ \overline{Y} \sim N(\mu_2, \frac{\sigma_2^2}{n}).$$

因为来自两个不同总体的样本总是相互独立的,于是 $\overline{X}$ 与 $\overline{Y}$ 也是相互独立的,因而(7.5.1) 成立.

(2) 当 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 时,S_1^2 和 S_2^2 都是 σ^2 的估计,由基本定理得

$$\frac{(m-1)S_1^2}{\sigma^2} \sim \chi^2_{m-1},\ \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1},$$

且相互独立,根据 χ^2 分布的可加性,我们有

$$\frac{(m-1)S_1^2 + (n-1)S_2^2}{\sigma^2} \sim \chi^2_{m+n-2}. \tag{7.5.5}$$

另一方面,当 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 时,(7.5.2) 变为

$$\frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim N(0,1) \tag{7.5.6}$$

由(7.5.5) 和(7.5.6) 以及 t 分布的定义,便得到(7.5.3),定理证毕.

利用这个定理,我们可以得到 $\mu_1 - \mu_2$ 的置信系数为 $1-\alpha$ 的置信区间.

(1) 当 σ_1^2 和 σ_2^2 皆已知时,由(7.5.2) 得

$$P\left\{\left|\frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}\right| \leqslant Z_{\alpha/2}\right\} = 1 - \alpha,$$

于是所求的置信区间为

$$\left[\overline{X} - \overline{Y} \pm Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}\right]. \tag{7.5.7}$$

(2) 当 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 但未知时,由(7.5.3) 得

$$P\left\{\left|\frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{m} + \frac{1}{n}}}\right| \leqslant t_{m+n-2}\left(\frac{\alpha}{2}\right)\right\} = 1 - \alpha,$$

于是所求的置信区间为

$$\left[\bar{X}-\bar{Y}\pm t_{m+n-2}\left(\frac{\alpha}{2}\right)S\sqrt{\frac{1}{m}+\frac{1}{n}}\right]. \tag{7.5.8}$$

例 7.5.1 欲比较甲、乙两种棉花品种的优劣.现假设用它们纺出的棉纱强度分别服从 $N(\mu_1,2.18^2)$ 和 $N(\mu_2,1.76^2)$,试验者从这两种棉纱中分别抽取样本 $X_1,\cdots,X_{200}$ 和 $Y_1,\cdots,Y_{100}$.其均值 $\bar{X}=5.32,\bar{Y}=5.76$.试给出 $\mu_1-\mu_2$ 的置信系数为 0.95 的区间估计.

解 由 $\sigma_1^2=2.18^2,\sigma_2^2=1.76^2,m=200,n=100,Z_{\alpha/2}=Z_{0.025}=1.96$,代入(7.5.7)得到所求区间估计为$[-0.899,0.019]$.

例 7.5.2 某公司利用两条自动化流水线灌装矿泉水.现从生产线上随机抽取样本 $X_1,\cdots,X_{12}$ 和 $Y_1,\cdots,Y_{17}$,它们是每瓶矿泉水的体积(毫升).算得样本均值 $\bar{X}=501.1$ 和 $\bar{Y}=499.7$.样本方差 $S_1^2=2.4,S_2^2=4.7$.假设这两条流水线所装的矿泉水的体积都服从正态分布,分别为 $N(\mu_1,\sigma^2)$ 和 $N(\mu_2,\sigma^2)$.给定置信系数 0.95,试求 $\mu_1-\mu_2$ 的区间估计.

解 由(7.5.4)式算出

$$S^2=\frac{(m-1)S_1^2+(n-1)S_2^2}{m+n-2}=\frac{11\times 2.4+16\times 4.7}{12+17-2}=3.763,$$

于是 $S=1.94$,查 t 分布表 $t_{m+n-2}\left(\frac{\alpha}{2}\right)=t_{27}(0.025)=2.05$,由(7.5.8)算得所求区间估计为$[-0.101,2.901]$.

在这两个例子中,$\mu_1-\mu_2$ 的区间估计包含了零,也就是说,μ_1 可能大于 μ_2,也可能小于 μ_2,这时我们就认为 μ_1 与 μ_2 并没有显著差异.

§7.6 非正态总体的区间估计

前面两节我们讨论了正态总体参数的区间估计.但是在实际应用中,我们有时不能判断手中的数据是否服从正态分布或者有足够理由认为它们不服从正态分布,这时,只要样本大小 n 比较大,总体均值 μ 的置信区间仍可用正态总体情形的公式(7.4.4),所不同的是这时的置信区间是近似的.这是求一般总体均值的一种简单有效的方法,其理论依据是中心极限定理,它要求样本大小 n 比较大,因此,这个方法称为大样本方法.

设总体均值为 μ,方差为 σ^2,$X_1,X_2,\cdots,X_n$ 为来自该总体的样本.因为这些样本是独立同分布的,根据中心极限定理,对充分大的 n,下式近似成立

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \sim N(0,1), \tag{7.6.1}$$

因而,近似地有

$$P\left\{\left|\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right| \leqslant Z_{\alpha/2}\right\} \approx 1-\alpha.$$

于是我们得到 μ 的置信系数约为 $1-\alpha$ 的置信区间

$$[\bar{X} - Z_{\alpha/2}\sigma/\sqrt{n},\ \bar{X} + Z_{\alpha/2}\sigma/\sqrt{n}], \tag{7.6.2}$$

形式上,这个置信区间和(7.4.4)完全一样,所不同的是这里的置信系数是近似的.若 σ 未知,用 σ 的一个估计,例如 S,来代替得

$$[\bar{X} - Z_{\alpha/2}S/\sqrt{n},\ \bar{X} + Z_{\alpha/2}S/\sqrt{n}], \tag{7.6.3}$$

只要 n 很大,(7.6.3)所提供的置信区间在应用上还是令人满意的.那么 n 究竟应该是多大呢?很明显,对相同的 n,(7.6.3)所给出置信区间的近似程度随总体分布与正态分布接近程度而变化,因此,从理论上很难给出 n 的一个界限,但许多应用实践表明,当 $n \geqslant 30$ 时,近似程度还是可以接受的.

例 7.6.1　某公司欲估计自己生产的电池寿命.现从其产品中随机抽取50只电池做寿命试验.这些电池的寿命的平均值 $\bar{X} = 2.266$(单位:100小时),$S = 1.935$.求该公司生产的电池平均寿命的置信系数为95% 的置信区间.

解　查正态分布表得 $Z_{\alpha/2} = Z_{0.025} = 1.96$,由公式(7.6.3)得到

$$\left[2.266 \pm 1.96 \times \frac{1.935}{\sqrt{50}}\right],$$

经简单计算上式化为[1.730,2.802].于是,我们有如下近似结论:该公司电池的平均寿命的置信系数约为95% 的置信区间为[1.730,2.802].

为了进一步说明前面引进的大样本方法,下面讨论二项分布和泊松分布的参数的区间估计.

一、二项分布

假设事件 A 在一次试验中发生的概率为 p,现在做了 n 次试验.以 Y_n 记事件 A 发生的次数,则 $Y_n \sim B(n,p)$,依中心极限定理,对充分大的 n,近似地有

$$\frac{Y_n - np}{\sqrt{np(1-p)}} \sim N(0,1). \tag{7.6.4}$$

(7.6.4)式是(7.6.1)式的一种特殊情形．读者从§5.2可以明白这一点．事实上，若记

$$X_i = \begin{cases} 1, & \text{若事件 } A \text{ 在第 } i \text{ 次试验中发生,} \\ 0, & \text{不然,} \end{cases}$$

这些 $X_i, i = 1,2,\cdots,n$ 独立同分布，$E(X_i) = p, \mathrm{Var}(X_i) = p(1-p)$，且 $Y_n = \sum_{i=1}^{n} X_i$．将这些量代入(7.6.1)式就得到(7.6.4)．对现在情形(7.6.3)变为

$$\left[\hat{p} - Z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n},\quad \hat{p} + Z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\right], \tag{7.6.5}$$

这里 $\hat{p} = Y_n/n = \sum_{i=1}^{n} X_i/n$．这就是二项分布参数 p 的置信系数约为 $1-\alpha$ 的置信区间．

例 7.6.2 商品检验部门随机抽查了某公司生产的产品 100 件，发现其中合格产品 84 件，试求该产品合格率的置信系数为 0.95 的置信区间．

解 $n = 100, Y_n = 84, \hat{p} = Y_n/n = 0.84, Z_{\alpha/2} = Z_{0.025} = 1.96$，(7.6.5)变为

$$\left[0.84 \pm 1.96\sqrt{\frac{0.84(1-0.84)}{100}}\right] = [0.77, 0.91],$$

即该产品合格率的置信系数约为 95% 的置信区间为[0.77,0.91]．

例 7.6.3 在环境保护问题中，饮水质量研究占有重要地位，其中一项工作是检查在饮水中是否存在各种类型的微生物．假设在随机抽取的 100 份一定容积的水样品中有 20 份样品含有某种微生物．试求同样容积的这种水含有这种微生物的概率 p 的置信区间，置信系数为 0.90．

解 $n = 100, Y_n = 20, \hat{p} = 0.20, Z_{\alpha/2} = Z_{0.05} = 1.645$，应用(7.6.5)式得

$$\left[0.20 \pm 1.645\sqrt{\frac{0.2 \times (1-0.2)}{100}}\right] = [0.134,\ \ 0.266],$$

即[0.134, 0.266]是概率 p 的置信系数约为 0.90 的置信区间．

二、泊松分布

设 $X_1, X_2, \cdots, X_n$ 为抽自具有泊松分布 $P(\lambda)$ 的总体的样本，因为 $E(X_i) = \lambda$，$\mathrm{Var}(X_i) = \lambda$，应用(7.6.3)，并用 $\bar{X}$ 去估计 λ，得到参数 λ 的置信系数约为 $1-\alpha$ 的置信区间

$$[\bar{X} - Z_{\alpha/2}\sqrt{\bar{X}/n},\quad \bar{X} + Z_{\alpha/2}\sqrt{\bar{X}/n}]. \tag{7.6.6}$$

例 7.6.4　公共汽车站在一单位时间内(如半小时或1小时或一天等)到达的乘客数服从泊松分布 $P(\lambda)$，对不同的车站，所不同的仅仅是参数 λ 的取值不同. 现对一城市某一公共汽车站进行了100个单位时间的调查. 这里单位时间是20分钟．计算得到每20分钟内来到该车站的乘客数平均值 $\bar{X} = 15.2$ 人. 试求参数 λ 的置信系数为95% 的置信区间.

解　$n = 100, \alpha = 0.05, Z_{\alpha/2} = Z_{0.025} = 1.96, \bar{X} = 15.2$，应用(7.6.6)式得

$$\left[\bar{X} \pm Z_{\alpha/2}\sqrt{\bar{X}/n}\right] = \left[15.2 \pm 1.96\sqrt{15.2/100}\right] = [14.44,\ 15.96],$$

即[14.44,15.96]为参数 λ 的置信系数约为95% 的置信区间.

习　题　七

7.1　设 $X_1, X_2, \cdots, X_n$ 为抽自二项分布 $B(m,p)$ 的样本，试求 p 的矩估计和极大似然估计.

注：这个问题的实际背景是很丰富的. 例如，生物医学方面的学者要研究某种物质致癌性质，往往用小白鼠做试验. 假定把50只小白鼠随机地分成10组，每组5只. 对每只小白鼠注射该物质，经过一段时间后，观察每组小白鼠患癌的个数，得到 $X_1, X_2, \cdots, X_{10}$，则 $X_i \sim B(5,p), i = 1,2,\cdots,10$. 这里 p 就是这种物质致癌的概率.

7.2　设总体为指数分布，其概率密度函数为

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{若 } x > 0, \\ 0, & \text{若 } x \leqslant 0. \end{cases}$$

求参数 λ 的矩估计和极大似然估计.

7.3　设总体为 $[0,\theta]$ 上的均匀分布，求参数 θ 的矩估计和极大似然估计.

7.4　设总体为 $[\theta, 2\theta]$ 上的均匀分布，求参数 θ 的矩估计和极大似然估计.

7.5　假设 $X_1, X_2, \cdots, X_n$ 为来自正态总体 $N(\mu,\sigma^2)$ 的样本，其中 μ 已知，求 σ^2 的极大似然估计.

7.6　设总体为指数分布，其概率密度函数为

$$f(x) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}}, & \text{当 } x > 0 \text{ 时}, \\ 0, & \text{其它}. \end{cases}$$

从该总体中抽出样本 X_1, X_2 和 X_3，考虑 θ 的如下四种估计

$$\hat{\theta}_1 = X_1,$$

$$\hat{\theta}_2 = (X_1 + X_2)/2,$$

$$\hat{\theta}_3 = (X_1 + 2X_2)/3,$$

$$\hat{\theta}_4 = \bar{X},$$

(1) 这四个估计中，哪些是 θ 的无偏估计？

(2) 试比较这些估计的方差.

7.7 一个电子线路上电压表的读数 X 服从 $[\theta, \theta+1]$ 上的均匀分布，其中 θ 是该线路上电压的真值，但它是未知的，假设 $X_1, \cdots, X_n$ 是此电压表上读数的一组样本，

(1) 证明样本均值 $\bar{X}$ 不是 θ 的无偏估计.

(2) 求 θ 的矩估计，证明它是 θ 的无偏估计.

7.8 设 $\hat{\theta}_1$ 和 $\hat{\theta}_2$ 都是 θ 的无偏估计，且 $\mathrm{Var}(\hat{\theta}_1) = \sigma_1^2$，$\mathrm{Var}(\hat{\theta}_2) = \sigma_2^2$，构造一个新无偏估计

$$\hat{\theta} = c\hat{\theta}_1 + (1-c)\hat{\theta}_2, \quad 0 \leqslant c \leqslant 1.$$

如果 $\hat{\theta}_1$ 和 $\hat{\theta}_2$ 相互独立，确定 c 使得 $\mathrm{Var}(\hat{\theta})$ 达到最小.

7.9 从工厂产品库中随机抽取 16 只零件，测得它们的长度(单位：厘米)为

2.14 2.10 2.13 2.15 2.13 2.12 2.13 2.10

2.15 2.12 2.14 2.10 2.13 2.11 2.14 2.11

假设零件长度分布为 $N(\mu, \sigma^2)$，试分如下两种情况求 μ 的置信系数为 0.90 的区间估计. (1) $\sigma^2 = 0.01^2$，(2) σ^2 未知.

7.10 甲、乙两组生产同种导线，现从甲组生产的导线中随机抽取 4 根，从乙组生产的导线中随机抽取 5 根，它们的电阻值(单位：欧姆)分别为

甲组： 0.143 0.142 0.143 0.137

乙组： 0.140 0.142 0.136 0.138 0.140

假设两组电阻值分别服从正态分布 $N(\mu_1, \sigma^2)$ 和 $N(\mu_2, \sigma^2)$，σ^2 未知. 试求 $\mu_1 - \mu_2$ 的置信系数为 0.95 的区间估计.

7.11 某市随机抽取 1000 个家庭，调查知道其中有 228 家拥有彩色电视机. 试据此数据对该市拥有彩色电视机家庭比例 p 作出区间估计. 取置信系数为 0.95.

7.12 根据实际经验可以认为，任一地区单位时间内(如一天或一个月或一年)火灾发生次数服从泊松分布 $P(\lambda)$，若以月为单位，从公安局记录得知，某城市过去 120 个月(即 10 年)火灾发生月平均次数为 7.5 次. 试求该城市火灾月平均次数 λ 的置信系数为 0.95 的置信区间.