第三节　参数的区间估计

一、　区间估计的基本概念

二、　典型例题

三、　小结

对于一个未知参数θ, 利用参数的点估计法可给出θ的点估计量. 当样本观测值给定后, 可进一步得到θ的点估计值. 人们不以得到点估计值为满足, 还要求知道它与θ的真值有无误差, 误差是多少.

即人们希望估计出一个范围, 并希望知道这个范围包含参数θ真值的可信程度. 为此引入估计的另一种形式---区间估计.

在区间估计理论中, 被广泛接受的一种观点是
置信区间**(confidence interval)**, 它由奈曼**(Neyman)**于**1934**年引入.

一、区间估计的基本概念

1. 置信区间的定义

设总体 X 的分布函数 $F(x;\theta)$ 含有一个未知参数 θ, 对于给定值 α $(0<\alpha<1)$, 若由样本 $X_1, X_2, \cdots,$ X_n 确定的两个统计量

$\theta_1=\theta_1(X_1, X_2, \cdots, X_n)$和 $\theta_2=\theta_2(X_1, X_2, \cdots, X_n)$ 满足

$$P\{\theta_1(X_1, X_2, \cdots, X_n)<\theta<\theta_2(X_1, X_2, \cdots, X_n)\}=1-\alpha,$$

则称$1-\alpha$为置信度或置信水平,

称随机区间 (θ_1, θ_2)是θ的置信度（置信水平） 为$1-\alpha$的置信区间, θ_1 和 θ_2分别称为置信度（置信 水平）为$1-\alpha$的双侧置信区间的置信下限和置信上限 。

关于定义的说明

被估计的参数 θ虽然未知，但它是一个常数，没有随机性，而区间(θ_1, θ_2)是随机的.

因此定义中如下表达式

$$P\{\theta_1(X_1, X_2, \cdots, X_n) < \theta < \theta_2(X_1, X_2, \cdots, X_n)\} = 1 - \alpha$$

的本质是：

随机区间(θ_1, θ_2)以$1-\alpha$的概率包含着参数 θ的真值，而不能说参数 θ以$1-\alpha$的概率落入某区间 (θ_1, θ_2).

例

设总体$X \sim N(\mu,4)$, μ未知，$X_1,\cdots,X_4$是一个样本，则$\overline{X} \sim N(\mu,1)$.

$$P\{\overline{X}-2<\mu<\overline{X}+2\}=P\{|\overline{X}-\mu|<2\}=2\Phi(2)-1=0.9544.$$

$\Rightarrow (\overline{X}-2,\overline{X}+2)$是$\mu$的置信度为0.95的置信区间.

若$\mu=0.5$,当$\overline{x}$分别为3,2,1时，对应区间分别为

(1,5),

(0,4), √

(-1,3), √

对于一个具体的区间，或者包含真值，或者不包含真值，无概率可言。

另外定义中的表达式

$$P\{\theta_1(X_1, X_2, \cdots, X_n) < \theta < \theta_2(X_1, X_2, \cdots, X_n)\} = 1-\alpha$$

还可以描述为：

若反复抽样多次(各次得到的样本容量相等,都是n)

每个样本值确定一个区 间(θ_1, θ_2),

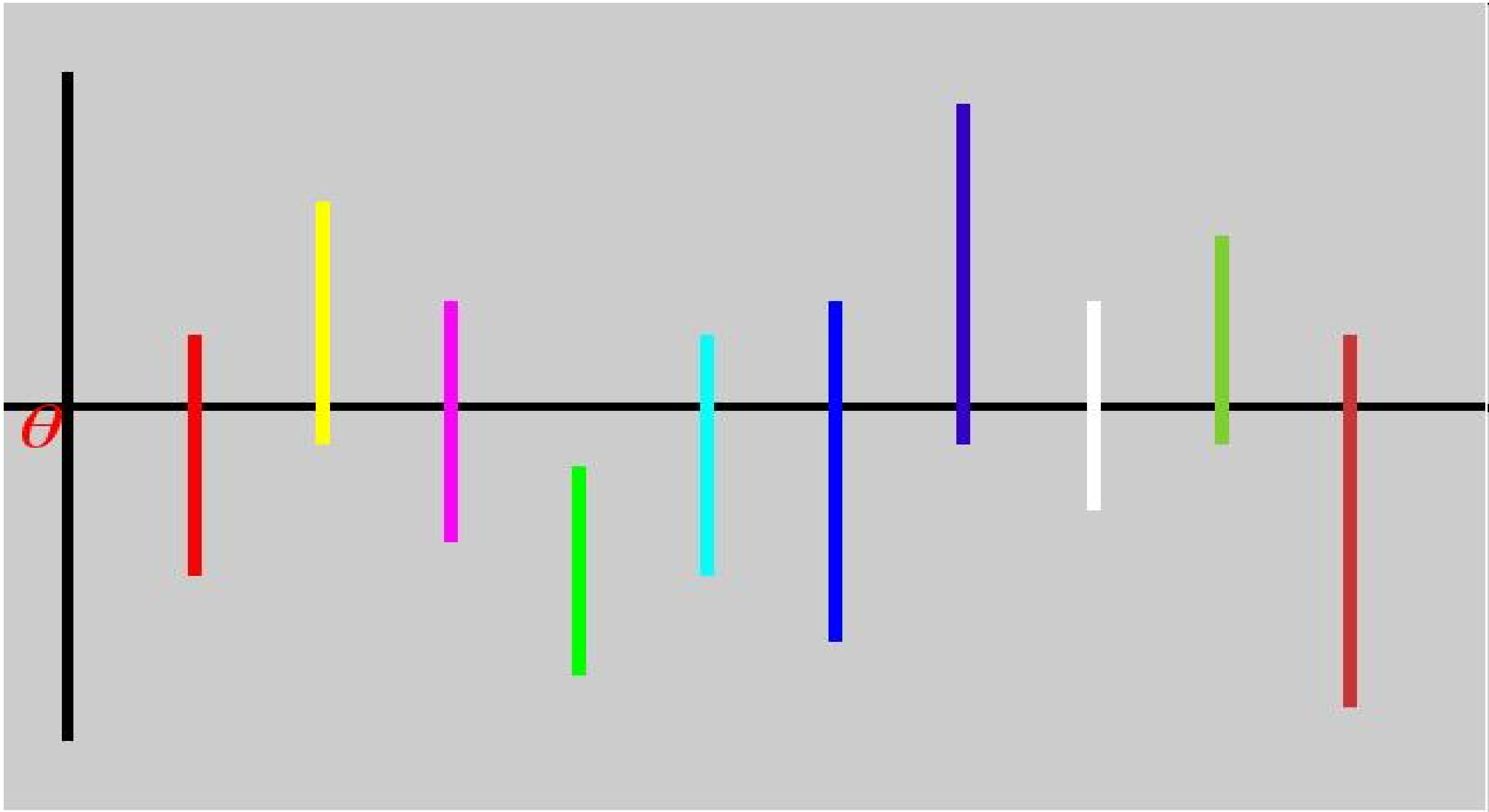
每个这样的区间或包含θ的真值或不包含θ的真值,

按**伯努利大数定理**, 在这样多的区间中,

包含θ真值的约占$100(1-\alpha)\%$,不包含的约占$100\alpha\%$.

例如　若 $\alpha = 0.01$, 反复抽样 1000 次,

则得到的 1000 个区间中不包含 θ 真值的约为10个.



置信区间的长度$\theta_2-\theta_1$反映了估计精度, $\theta_2-\theta_1$越小, 估计精度越高.

α反映了估计的可靠度, α越小, 1- α越大, 估计的可靠度越高, 同时, $\theta_2-\theta_1$往往增大, 因而估计精度降低.

一般是在保证可靠度的条件下尽可能提高精度。

α确定后, 置信区间选取方法不唯一, 常选最小的一个.

2. 求置信区间的一般步骤(共3步)

(1) 寻求一个样本 $X_1, X_2, \cdots, X_n$ 的函数：

$$Z = Z(X_1, X_2, \cdots, X_n; \theta)$$

其中仅包含待估参数 θ, 并且 Z 的分布已知且不依赖于任何未知参 数 (包括 θ).

(2) 对于给定的置信度 $1-\alpha$, 定 出两个常数 a, b, 使 $P\{a < Z(X_1, X_2, \cdots, X_n; \theta) < b\} = 1-\alpha$.

（3）若能从$a < Z(X_1, X_2, \cdots, X_n; \theta) < b$得到等价的不等式$\theta_1 < \theta < \theta_2$, 其中 $\theta_1 = \theta_1(X_1, X_2, \cdots, X_n)$, $\theta_2 = \theta_2(X_1, X_2, \cdots, X_n)$都是统计量，那么$(\theta_1，\theta_2)$就是$\theta$的一个置信度为$1-\alpha$的置信区间。

样本容量 n 固定，置信水平 $1-\alpha$ 增大，置信区间长度增大，可信程度增大，区间估计精度降低．

置信水平 $1-\alpha$ 固定，样本容量 n 增大，置信区间长度减小，可信程度不变，区间估计精度提高．

二、典型例题

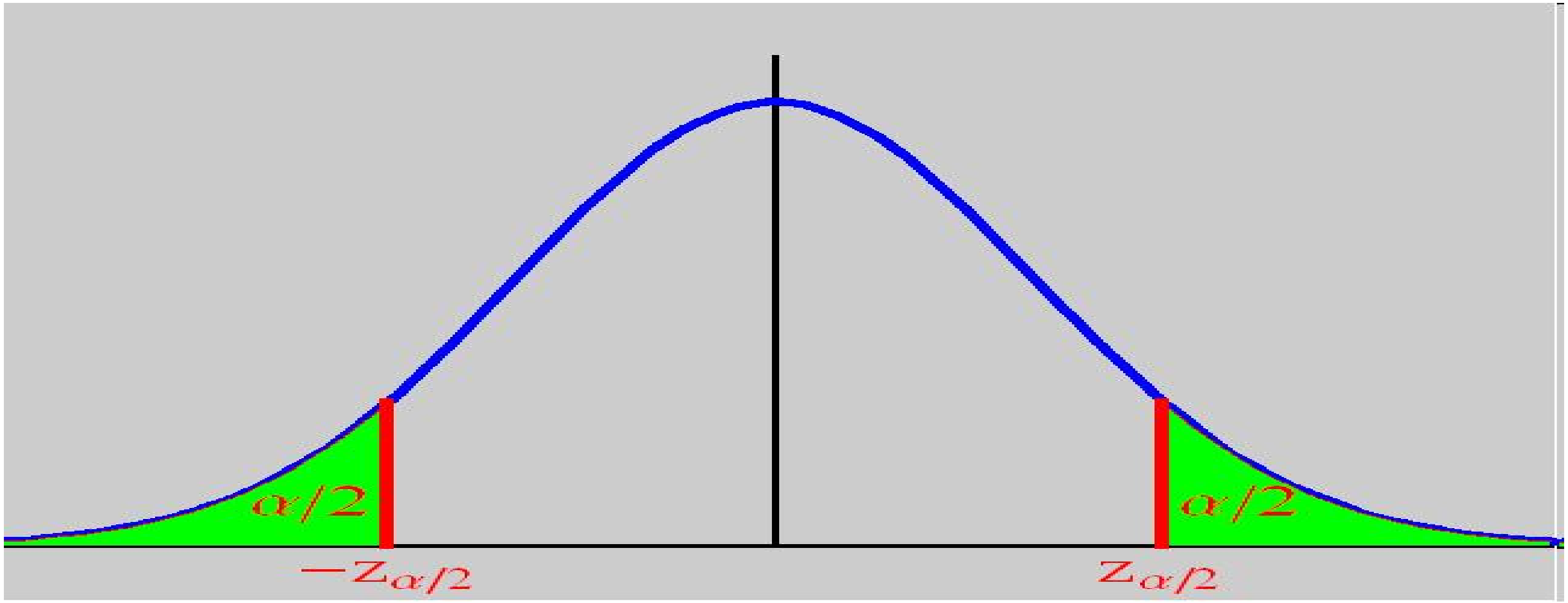
例1 设 $X_1, X_2, \cdots, X_n$ 是来自正态总体 $N(\mu, \sigma^2)$ 的样本,其中 σ^2 为已知,μ 为未知,求 μ 的置信水平为 $1-\alpha$ 的置信区间.

解 因为 $\overline{X}$ 是 μ 的无偏估计,

$$且\ U = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1),$$

$\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ 是不依赖于任何未知参 数的,

由标准正态分布的上 α 分位点的定义知



$$P\left\{\left|\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right|<u_{\alpha/2}\right\}=1-\alpha,$$

即 $$P\left\{\overline{X}-\frac{\sigma}{\sqrt{n}}u_{\alpha/2}<\mu<\overline{X}+\frac{\sigma}{\sqrt{n}}u_{\alpha/2}\right\}=1-\alpha,$$

于是得 μ 的一个置信水平为 $1-\alpha$ 的置信区间

$$\left(\overline{X}-\frac{\sigma}{\sqrt{n}}u_{\alpha/2},\ \overline{X}+\frac{\sigma}{\sqrt{n}}u_{\alpha/2}\right).$$

这样的置信区间常写成 $\left(\overline{X}\pm\frac{\sigma}{\sqrt{n}}u_{\alpha/2}\right)$.

其置信区间的长度为 $2\times\frac{\sigma}{\sqrt{n}}u_{\alpha/2}$.

注意：置信水平为 $1-\alpha$ 的置信区间是不唯一的.

如果在例 1中取 $n = 16$, $\sigma = 1$, $\alpha = 0.05$,

查表可得 $u_{\alpha/2} = u_{0.025} = 1.96$,

得一个置信水平为 0.95的置信区间 $\left(\overline{X} \pm \frac{1}{\sqrt{16}} \times 1.96\right)$.

由一个样本值算得样本均值的观察值 $\overline{x} = 5.20$,

则置信区间为 (5.20 ± 0.49), 即 $(4.71,\ 5.69)$.

在例 1中如果给定 $\alpha = 0.05$,

则又有 $P\left\{-u_{0.04} < \dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} < u_{0.01}\right\} = 0.95,$

即 $P\{\bar{X} - \dfrac{\sigma}{\sqrt{n}} u_{0.01} < \mu < \bar{X} + \dfrac{\sigma}{\sqrt{n}} u_{0.04}\} = 0.95,$

故 $\left(\bar{X} - \dfrac{\sigma}{\sqrt{n}} u_{0.01},\ \bar{X} + \dfrac{\sigma}{\sqrt{n}} u_{0.04}\right)$ 也是 μ 的置信水平为0.95的置信区间.

其置信区间的长度为 $\dfrac{\sigma}{\sqrt{n}}(u_{0.04} + u_{0.01})$.

比较两个置信区间的长度

$$L_1 = 2 \times \frac{\sigma}{\sqrt{n}} u_{0.025} = 3.92 \times \frac{\sigma}{\sqrt{n}},$$

$$L_2 = \frac{\sigma}{\sqrt{n}}(u_{0.04} + u_{0.01}) = (1.75 + 2.33) \times \frac{\sigma}{\sqrt{n}} = 4.08 \times \frac{\sigma}{\sqrt{n}},$$

显然 $L_1 < L_2$. 置信区间短表示估计的精度高.

说明: 对于概率密度的图形是单峰且关于纵坐标轴对称的情况, 易证取 a 和 b 关于原点对称时,能使置信区间长度最小.

三、小结

点估计不能反映估计的精度, 故而本节引入了区间估计.

置信区间是一个随机区间 (θ_1, θ_2), 它覆盖未知参数具有预先给定的概率（置信水平）, 即对于任意的 $\theta \in \Theta$, 有 $P\{\theta_1 < \theta < \theta_2\} \geq 1 - \alpha$.

求置信区间的一般步骤(分三步).