软件工程统计方法

参数估计

陈振宇

南京大学软件学院

Email:zychen@software.nju.edu.cn

Homepage:software.nju.edu.cn/zychen

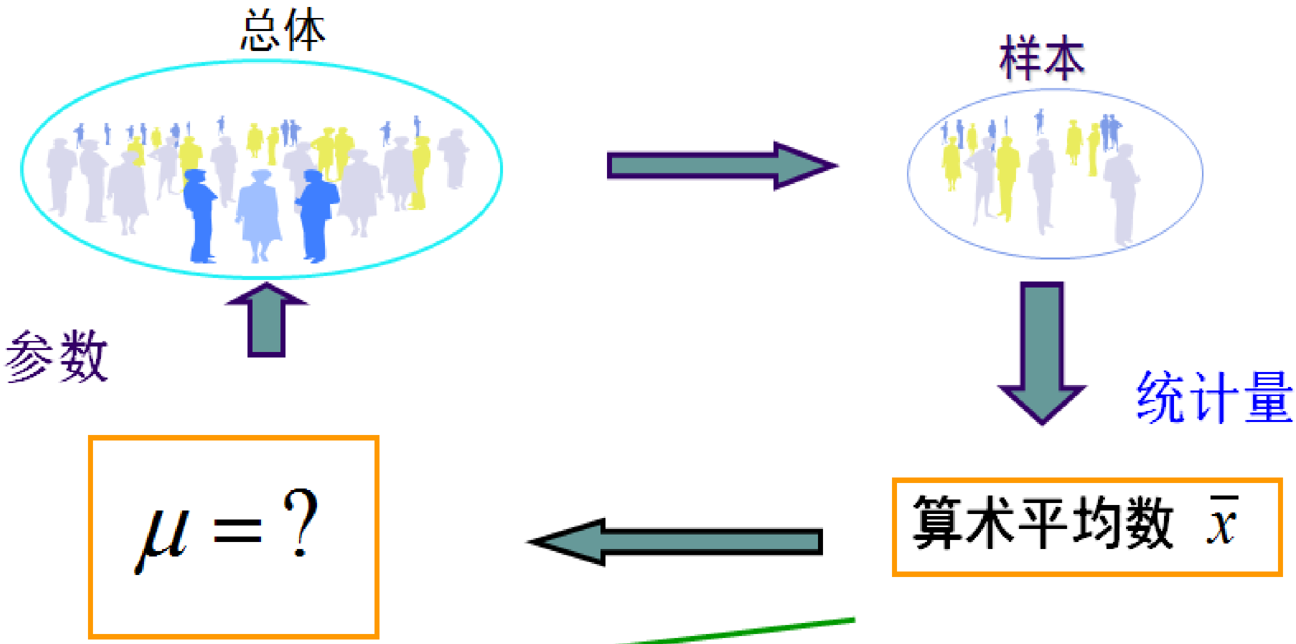
内 容 提 纲

- 统计学导论
- 描述统计
- 概率计算基础
- 随机变量及其分布
- 统计量及其抽样分布
- 参数估计
- 参数假设检验
- 非参数假设检验
- 方差分析
- 回归分析

1 本节内容

- 点估计
  - 矩估计法
  - 极大似然估计法
  - 截尾样本
  - 估计量的评选标准
- 区间估计
  - 区间估计的基本概念
  - 正态总体均值与方差的区间估计
  - (0-1)分布参数的区间估计
  - 单侧置信区间

2 参数估计



用来推断总体参数的统计量称为**估计量**（estimator),其取值称为**估计值**（estimate）。 同一个参数可以有多个不同的估计量。参数是唯一的，但**估计量（统计量）是随机变量**，取值是不确定的。

参 数 估 计

参数估计可以抽象如下的数学模型：总体的概率分布$F(x;\theta)$含有未知参数θ, 从总体中随机抽样, 得样本值$x_1,\cdots,x_n$去获得对未知参数θ的了解。

- 包括估计问题和检验问题。
- 假如估计的对象是参数(即总体分布已知), 所以称为参数估计。
- 参数估计分为两种基本形式：点估计和区间估计。

- 估计量：用于估计总体参数的随机变量。
  - 如样本均值，样本比例、样本方差等。
  - 例如: 样本均值就是总体均值μ的一个估计量。
- 通常参数用θ表示, 估计量用$\hat{\theta}$表示。
- 估计值：估计参数时计算出来的统计量的具体值。
  如果样本均值$\overline{x}=10$, 则10就是μ的估计值。

3 点估计

- 设总体的分布函数$F(x;\theta)$的形式已知, θ是待估参数;
- $X_1,\cdots,X_n$是X的一个样本, $x_1,\cdots,x_n$是相应的样本值;
- 构造一个适合的统计量$\hat{\theta}(X_1,\cdots,X_n))$, 它的观察值$\hat{\theta}(x_1,\cdots,x_n))$作为未知参数$\theta$的近似值;
- 估计量: $\hat{\theta}(X_1,\cdots,X_n))$, 估计值: $\hat{\theta}(x_1,\cdots,x_n))$, 或统称为估计。

用来推断总体参数的统计量称为估计量(Estimator), 其取值称为估计值(Estimate)。同一个参数可以有多个不同的估计量。参数是唯一的, 但估计量(统计量)是随机变量, 取值是不确定的。

英国统计学家皮尔逊(K.Pearson) 1900年提出一种估计方法，它源于替换原理，用从样本观测得到的频率去估计相应的总体的概率，比如用样本次品率作为总体次品率的估计。频率替代从某种意义上可以看成是用样本均值估计总体期望的估计方法。

例 子

Example 1 某炸药厂一天发生火灾的次数X服从泊松分布$\pi(\lambda)$, 根据以下抽样估计λ:

着火次数k	0	1	2	3	4	5	6	
发生k次火灾的天数	75	90	54	22	6	2	1	$\sum = 250$

$$\overline{x} = \frac{1}{250}[0 \times 75 + 1 \times 90 + 2 \times 54 + \cdots + 6 \times 1] = 1.22$$

所以$E(X) = \lambda$的估计为1.22.

再根据泊松分布的概率分布律:

$$P\{X = k\} = \frac{1.22^k e^{-1.22}}{k!}$$

可求得总体分布情况.

4 矩估计法

以样本矩估计相应总体矩来求出估计量的方法, 称为矩估计法, 又称数字特征法.

设X为连续型随机变量, 其概率密度为$f(x;\theta_1,\cdots,\theta_k)$, $X_1,\cdots,X_n$是来自X的样本, 假设总体X的前k阶矩

$$\mu_l = E(X^l) = \int_{-\infty}^{+\infty} x^l f(x;\theta_1,\cdots,\theta_k)dx$$

$(l=1,\cdots,k)$存在, 它们是$\theta_1,\cdots,\theta_k$的函数. 其样本$X_1,\cdots,X_n$的前k阶原点矩

$$A_l = \frac{1}{n}\sum_{i=1}^{n} X_i^l$$

依概率收敛于总体矩μ_l. 样本矩的连续函数依概率收敛于总体矩的连续函数. 以样本矩的连续函数值作为总体矩的连续函数估计值. 这种方法称为称为矩估计法.

均匀分布的矩估计

Example 2 设$X \sim \mathbb{U}(a,b)$, a,b未知. $1,2,3,4,5$是这个均匀总体的一个样本, 试求a,b的矩估计值.

均匀分布的矩估计

Example 3 设$X \sim \mathbb{U}(a,b)$, a,b未知. $X_1,\cdots,X_n$是来自X的样本, 试求a,b的矩估计量.

解:

$$\mu_1 = E(X) = (a+b)/2$$

$$\mu_2 = E(X^2) = D(X) + [E(X)]^2 = (b-a)^2/12 + (a+b)^2/4$$

$$\begin{cases} a+b &= 2\mu_1 \\ b-a &= \sqrt{12(\mu_2-\mu_1^2)} \end{cases}$$

解方程组得

$$a = \mu_1 - \sqrt{3(\mu_2-\mu_1^2)}, b = \mu_1 + \sqrt{3(\mu_2-\mu_1^2)}$$

分别以A_1, A_2代替μ_1,μ_2, 得到a,b的矩估计量为

$$a = A_1 - \sqrt{3(A_2-A_1^2)}, b = A_1 + \sqrt{3(A_2-A_1^2)}$$

正态分布的矩估计

Example 4 设$X \sim \mathbb{N}(\mu, \sigma^2)$, μ,σ^2未知. $1, 2, 3, 4, 5$是这个正态总体的一个样本, 试求μ,σ^2的矩估计值.

伯努利分布的矩估计

Example 5 设$X \sim \mathbb{B}(1,p)$, p未知. $1,0,1,1,0$ 是这个二项总体的一个样本, 试求p的矩估计值.

Example 6 设$X \sim \mathbb{B}(1,p)$, p未知. $x_1, x_2, \cdots, x_n$ 是这个二项总体的一个样本, 试求p的矩估计值.

数学期望与方差的矩估计

Example 7 设X的期望μ和方差σ^2都存在，且有$\sigma^2 > 0$. μ, σ^2均为未知. $X_1, \cdots, X_n$是来自X的样本. 试求μ, σ^2的矩估计量.

解:

$$\mu_1 = E(X) = \mu$$

$$\mu_2 = E(X^2) = D(X) + [E(X)]^2 = \sigma^2 + \mu^2$$

解方程组得

$$\mu = \mu_1, \sigma^2 = \mu_2 - \mu_1^2$$

分别以A_1, A_2代替μ_1, μ_2, 得到μ, σ^2的矩估计量为

$$\mu = A_1 = \overline{X}, \sigma^2 = A_2 - A_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

非经典分布的矩估计

Example 8 设总体X的概率密度为: $f(x)=\frac{1}{\theta}e^{-(x-\mu)/\theta}\ x\geq\mu$, 其他$f(x)=0$. 其中$\theta>0$, $x_1,\cdots,x_n$为X的样本, 求θ,μ的矩估计值.

解:

$$E(X)=\int_{-\infty}^{+\infty}xf(x)dx=\int_{\mu}^{+\infty}x\frac{1}{\theta}e^{-(x-\mu)/\theta}dx=\mu+\theta$$

$$E(X^2)=\int_{-\infty}^{+\infty}x^2f(x)dx=\mu^2+2\theta\int_{\mu}^{+\infty}\frac{x}{\theta}e^{-(x-\mu)/\theta}dx=\mu^2+2\mu\theta+2\theta^2$$

$$D(X)=E(X^2)-E^2(X)=\theta^2$$

$$\begin{cases}E(X)=\overline{X}=\mu+\theta\\ D(X)=\frac{1}{n}\sum_{i=1}^n(X_i-\overline{X})^2=\theta^2\end{cases}\Rightarrow\begin{cases}\hat{\theta}=\sqrt{\frac{1}{n}\sum_{i=1}^n(X_i-\overline{X})^2}\\ \hat{\mu}=\overline{X}-\sqrt{\frac{1}{n}\sum_{i=1}^n(X_i-\overline{X})^2}\end{cases}$$

均匀分布的矩估计

Example 9 设$X \sim \mathbb{U}(0,b)$，其中b是未知参数，$1,2,9$ 是来自总体的一个样本，求b的矩估计值。

5 极大似然估计法

假设有某位同学与一位老战士一同进行实弹射击, 两人同打一个靶子, 每人各打一发, 结果仅中一发, 试问认为这一发是谁打中的较为合理?

在已知总体X概率分布时, 对总体进行n次观测, 得到一个样本, 选择概率最大的θ值$\hat{\theta}$ 作为未知参数θ的真值估计是最合理的.

1821年首先由德国数学家高斯(C.F.Gauss)提出极大似然估计(maximum likelihood estimation); 英国统计学家R.A.Fisher在1922年的论文《数理统计学的数学基础》(On the mathematical foundations of theoretical statistics)再次提出了这个思想，并且首先研究了这种方法的性质。

例 子

Example 10 假设在一个罐中放着许多白球和黑球, 并假定已经知道两种球的数目之比是*1:3*, 但不知道哪种颜色的球多, 如果用有放回抽样法从罐中任意抽取n个球, 则其中黑球的个数为x的概率为

$$P(x;p) = C_n^x p^x q^{1-x}, q = 1 - p$$

假定n=*3*, 求p.

解:由题意可得$p = 1/4$或$p = 3/4$, 先算两者的概率.

X	0	1	2	3
P(x;3/4)	1/64	9/64	27/64	27/64
P(x;1/4)	27/64	27/64	9/64	1/64

则定义估计量如下:

$$\hat{p}(x) = 1/4, \text{当} x = 0, 1$$

$$\hat{p}(x) = 3/4, \text{当} x = 2, 3$$

即我们选择概率较大的那个参数估计, 这就是极大似然估计的基本思想.

极大似然估计法步骤

Definition 1 以总体X的样本的似然函数$L(\theta)$的解$\theta(X_1,\cdots,X_n)$来估计参数真值的方法. 解方程

$$\frac{dL}{d\theta}=0\ \text{或}\ \frac{d\ln L}{d\theta}=0$$

得$\hat{\theta}=\hat{\theta}(X_1,\cdots,X_n)$称为$\theta$的极大似然估计值.

(1)X是离散型随机变量, $P\{X=x\}=P(x;\theta)$, 则似然函数为

$$L(\theta)=\prod_{i=1}^{n}P(x;\theta)$$

(2) (1)X是连续型随机变量, 概率密度为$f(x;\theta)$, 则似然函数为

$$L(\theta)=\prod_{i=1}^{n}f(x;\theta)$$

有k个参数时, 似然函数$L(\theta_1,\cdots,\theta_k)$为多元函数, 解$L(\theta_1,\cdots,\theta_k)$或$\ln L(\theta_1,\cdots,\theta_k)$的偏导数方程组即可得$\hat{\theta}_i, i=1,\cdots,k$.

二项分布的极大似然估计

Example 11 设$X \sim \mathbb{B}(n, p)$, p未知. $1, 0, 1, 1, 0$ 是这个二项总体的一个样本, 试求p的极大似然估计值.

例 子

Example 12 设$X \sim \mathbb{B}(n,p)$. $x_1,\cdots,x_n$是来自X的一个样本, 对应样本值为$x_1,\cdots,x_n$. 试求参数p的极大似然估计量.

解: X的分布律为

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

故似然函数为

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

而

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

令

$$\frac{d}{dp}\ln L(p)=\frac{\sum_{i=1}^{n}x_i}{p}-\frac{n-\sum_{i=1}^{n}x_i}{1-p}=0$$

解得p的最大似然估计值$\hat{p}=\frac{1}{n}\sum_{i=1}^{n}x_i=\overline{x}$.
p的最大似然估计量为$\hat{p}=\frac{1}{n}\sum_{i=1}^{n}X_i=\overline{X}$.
这一估计量跟矩估计量相同.

均匀分布的极大似然估计

Example 13 设$X \sim \mathbb{U}(0,b)$, 其中b是未知参数, $1,2,9$ 是来自总体的一个样本，求b的极大似然估计值。

Example 14 设$X \sim \mathbb{U}(a,b)$, 其中a,b是未知参数, $x_1,\cdots,x_n$ 是来自总体的一个样本，求a,b的极大似然估计值。

例 子

Example 15 设$X \sim N(\mu, \sigma^2)$, μ, σ^2为未知参数, $X_1, \cdots, X_n$是来自X的一个样本, 对应样本值为$x_1, \cdots, x_n$. 试求参数μ, σ^2的极大似然估计量.

解: X的概率密度为

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

故似然函数为

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

而

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

例 子

令

$$\frac{\partial}{\partial \mu}\ln L = \frac{1}{\sigma^2}[\sum_{i=1}^{n} x_i - n\mu] = 0$$

$$\frac{\partial}{\partial \sigma^2}\ln L = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

解得μ, σ^2的极大似然估计值

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}$$

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

这个极大似然估计量跟矩估计量相同.

指数分布的极大似然估计

Example 16 设$X \sim \mathbb{E}(\theta)$, 其中$\theta$是未知参数, $x_1, \cdots, x_n$ 是来自总体的一个样本，求θ的极大似然估计值。

$$f(x) = \frac{1}{\theta}e^{-\frac{x}{\theta}}, L(\theta) = \prod_{i=1}^{n} \frac{1}{\theta}e^{-\frac{x_i}{\theta}}$$

$$\ln L(\theta) = \sum_{i=1}^{n}(\ln\frac{1}{\theta} - \frac{x_i}{\theta}) = n\ln\frac{1}{\theta} - \frac{1}{\theta}\sum_{i=1}^{n} x_i$$

$$\frac{\partial \ln L(\theta)}{\partial(\frac{1}{\theta})} = n\theta - \sum_{i=1}^{n} x_i$$

$$n\theta - \sum_{i=1}^{n} x_i = 0, \theta = \frac{1}{n}\sum_{i=1}^{n} x_i$$

例子

Example 17 设总体X的概率分布律为:

X	*0*	*1*	*2*	*3*
p_k	p^2	$2p(1-p)$	p^2	$1-2p$

其中p是未知参数. 利用如下样本值: $1,3,0,2,3,3,1,3$. 求*(1)* p的矩估计值; *(2)* p的极大似然估计值.

例子

Example 18 设总体X的概率分布律为:

X	0	1	2	3
p_k	p^2	$2p(1-p)$	p^2	$1-2p$

其中p是未知参数. 利用如下样本值: $1,3,0,2,3,3,1,3$. 求(1) p的矩估计值; (2) p的极大似然估计值.

解: (1)

$$\overline{x} = 16/8 = 2$$

$$E(X) = 0*p^2 + 1*2p(1-p) + 2*p^2 + 3*(1-2p) = 3-4p = 2, p = 1/4$$

(2)

$$L(p) = P(X=0)P(X=1)^2P(X=2)P(X=3)^4 = 4p^6(1-p)^2(1-2p)^4$$

$$\ln(L(p)) = \ln 4 + 6\ln p + 2\ln(1-p) + 4\ln(1-2p)$$

$$\ln(L(p))' = \frac{6}{p} - \frac{2}{1-p} - \frac{8}{1-2p} \Rightarrow 12p^2 - 14p + 3 = 0$$

$$p = \frac{7 \pm \sqrt{13}}{12}$$

因为$p < 1/2$, 所以$p = \frac{7-\sqrt{13}}{12}$.

例 子

Example 19 设总体X的概率密度为: $f(x)=\frac{1}{\theta}e^{-(x-\mu)/\theta}\ x\geq\mu$, 其他$f(x)=0$. 其中$\theta>0$, $x_1,\cdots,x_n$为X的样本, 求θ,μ的极大似然估计值.

解:

$$L(\theta,\mu)=\prod_{i=1}^{n}\frac{1}{\theta}e^{-(x_i-\mu)/\theta}=\frac{1}{\theta^n}e^{-\frac{1}{\theta}\sum_{i=1}^{n}(x_i-\mu)},x_i\geq\mu$$

注意: 此时不能直接求导解估计值, 因为
(1) $x_i\geq\mu$, 故μ的取值最大不能超过$min(x_1,\cdots,x_n)$.
(2) $L(\theta,\mu)$是μ的增函数, 即μ取得最大时, L达到最大.
所以

$$\hat{\mu}=min(X_1,\cdots,X_n)$$

$$\ln L(\theta)=-n\ln\theta-\frac{1}{\theta}\sum_{i=1}^{n}(X_i-\hat{\mu})$$

$$\frac{d\ln L(\theta)}{d\theta}=-\frac{n}{\theta}+\frac{1}{\theta^2}\sum_{i=1}^{n}(X_i-\hat{\mu})=0$$

则$\hat{\theta}=\overline{X}-min(X_i)$.

6 矩估计与极大似然估计

矩估计法和极大似然估计法的比较

- 矩估计法采用样本矩替换总体矩来估计参数，相当于使用了分布函数的部分信息；
- 极大似然估计法采用似然函数来求得参数的估计，理论上相当于使用了分布函数的全部信息；
- 在已知总体分布的前提下，采用极大似然估计法的理由更充分，而在总体分布函数未知但有关的总体矩已知的情况下，采用矩估计法更合适。

极大似然估计的不变性： 如果$\hat{\theta}$是θ的极大似然估计，则对任意函数$g(\theta)$，其极大似然估计为$g(\hat{\theta})$。

7 截尾样本

产品寿命X是一个随机变量, 对X进行统计推断, 通常可以采取以下的试验方法:

- 完全样本: 随机抽取的n个产品在时间$t=0$同时投入试验, 直到每个产品失效, 得到样本$0 \le t_1 \le \cdots \le t_n$.
- 定时截尾样本: 随机抽取的n个产品在时间$t=0$同时投入试验, 直到规定截尾时间t_0时停止, 停止是有$m \le n$个产品失效, 得到样本得到样本$t_1, \cdots, t_m$ $(0 \le t_1 \le \cdots \le t_m \le t_0)$.
- 定数截尾样本: 随机抽取的n个产品在时间$t=0$同时投入试验, 直到规定$m \le n$(m称为截尾数)个产品失效时停止, 得到样本$t_1, \cdots, t_m (0 \le t_1 \le \cdots \le t_m)$.

截 尾 样 本

设寿命分布为

$$f(t) = \frac{1}{\theta}e^{-t/\theta}, t \geq 0, f(t) = 0, t < 0$$

求θ基于截尾样本的极大似然估计值.

- 定时截尾样本:

$$L(\theta) = \frac{m}{\theta}e^{-\frac{1}{\theta}[t_1+\cdots+t_m+(n-m)t_0]}$$

令$\frac{dL}{d\theta} = 0$, 解得$\hat{\theta} = \frac{s(t_0)}{m}$, $s(t_0) = t_1 + \cdots + t_m + (n-m)t_0$.

- 定数截尾样本:

$$L(\theta) = \frac{m}{\theta}e^{-\frac{1}{\theta}[t_1+\cdots+t_m+(n-m)t_m]}$$

令$\frac{dL}{d\theta} = 0$, 解得$\hat{\theta} = \frac{s(t_m)}{m}$, $s(t_m) = t_1 + \cdots + t_m + (n-m)t_m$.

例 子

Example 20 设某产品寿命服从指数分布$X \sim \mathcal{E}(\theta)$, θ未知. 随机取*50*只电池投入寿命试验, 规定试验进行到其中*15*只失效时结束试验, 得到失效时间*(*小时*)*为

$$115, 119, 131, 138, 142, 147, 148, 155$$

$$158, 159, 163, 166, 167, 170, 172$$

试求θ的最大似然估计.

解: $n = 50$, $m = 15$, $s(t_{15}) = 115 + \cdots + 172 + (50 - 15) \times 172 = 8270$, 得$\theta$得最大似然估计为

$$\hat{\theta} = \frac{8270}{15} = 551.33$$

8 估计量的评估

原则上任何统计量都可以作为未知参数的估计量, 那么哪一个估计量好?

- 无偏性
- 有效性
- 一致性

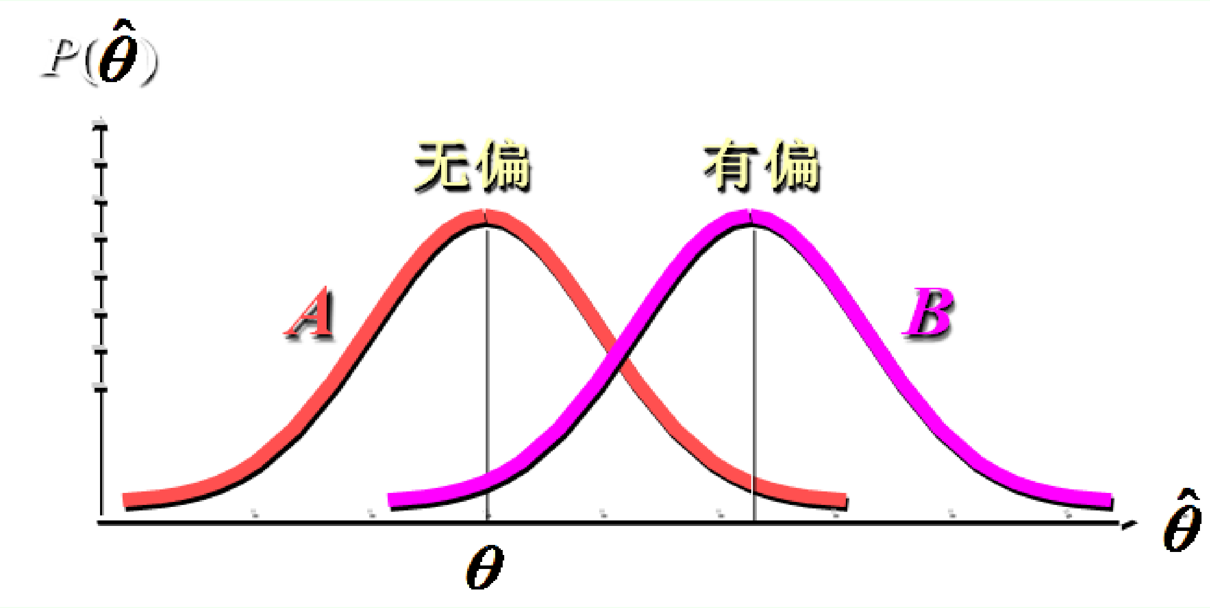
无 偏 性

Definition 2 (无偏性) 设$\hat{\theta}(X_1, \cdots, X_n)$, $E(\hat{\theta})$存在且

$$E(\hat{\theta}) = \theta$$

则称$E(\hat{\theta})$ 为θ的无偏估计.

若$E(\hat{\theta}) \neq \theta$, 则称$|E(\hat{\theta}) - \theta|$为估计值$\hat{\theta}$的偏差, 即以$\hat{\theta}$为$\theta$估计的系统误差.
若$E(\hat{\theta}) = a\theta + b$, 则$\frac{1}{a}(\hat{\theta} - b)$是$\theta$的无偏估计.



无 偏 估 计 量

Theorem 1 设X(不管什么分布)的均值为μ，方差为σ^2，则$\overline{X}$，S^2分别为μ和σ^2的无偏估计, 即$E(\overline{X})=\mu, E(S^2)=\sigma^2$.

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

$$D(\overline{X})=D(\frac{1}{n}\sum_{i=1}^{n}X_i)=\frac{1}{n^2}\sum_{i=1}^{n}D(X_i)=\frac{\sigma^2}{n}$$

$$E(\overline{X}^2)=D(\overline{X})+E(\overline{X})^2=\frac{\sigma^2}{n}+\mu^2$$

$$E(X_i^2)=D(X_i)+E(X_i)^2=\sigma^2+\mu^2$$

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

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

无 偏 估 计 量

Theorem 2 设总体X的k阶矩$\mu_k = E(X^k), k \geq 1$存在, 试证明无论总体服从什么分布, k阶样本矩$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$是$k$阶总体矩$\mu_k$的无偏估计量.

证: X_i与X同分布, 故有

$$E(X_i^k) = E(X^k) = \mu_k, i = 1, \cdots, n$$

则有

$$E(A_k) = \frac{1}{n}\sum_{i=1}^{n} E(X_i^k) = \mu_k$$

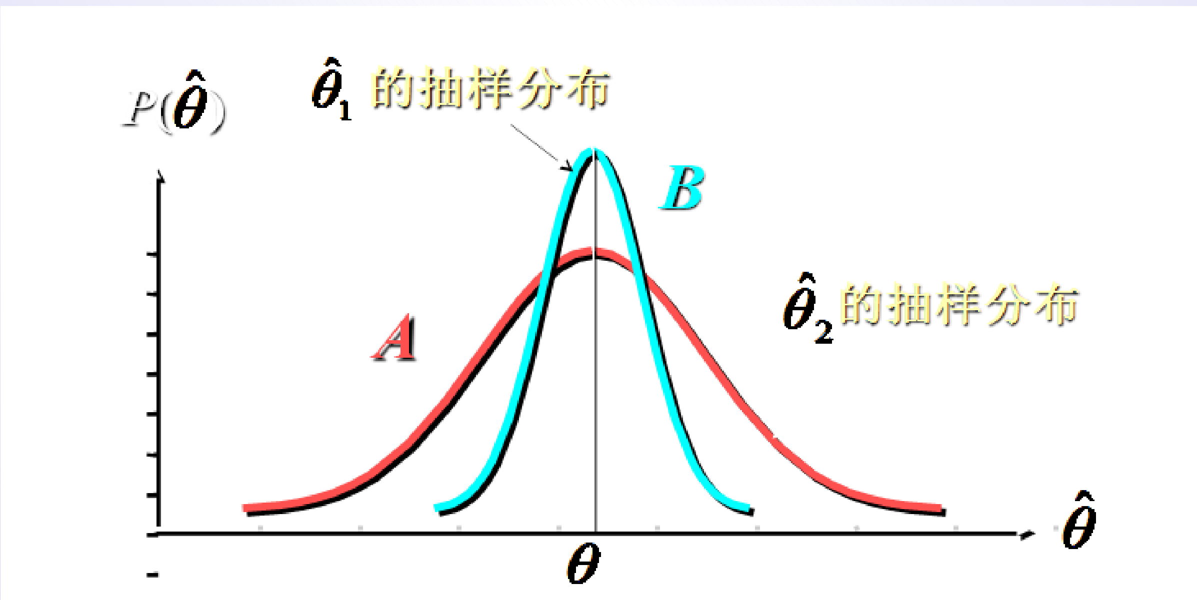
有效性

如何比较两个无偏估计$\hat{\theta}_1$和$\hat{\theta}_2$?

Definition 3 (有效性) 设$\hat{\theta}_1$和$\hat{\theta}_2$为θ的两个无偏估计, 若对于任意θ, 有

$$D(\hat{\theta}_1) \le D(\hat{\theta}_2)$$

且至少存在某一个θ使得不等号严格成立, 则称$\hat{\theta}_1$和$\hat{\theta}_2$有效.

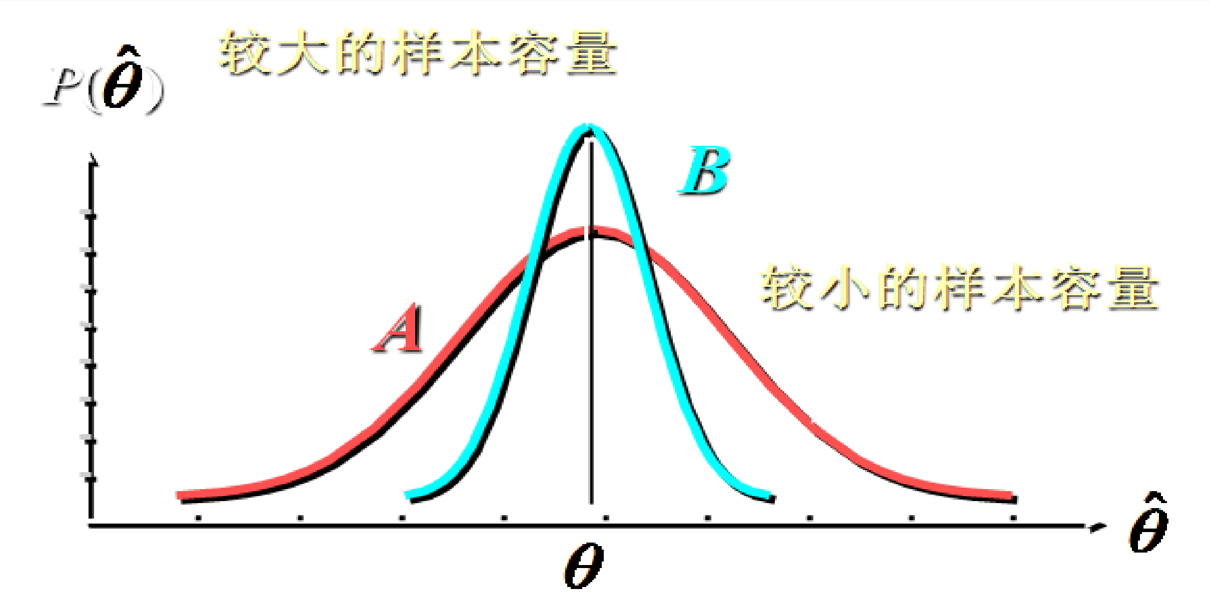


一致性

Definition 4 (一致性) 对于任意$\epsilon > 0$, 有

$$\lim_{n\to\infty} P\{|\hat{\theta} - \theta| < \epsilon\} = 1$$

则称$\hat{\theta}$是θ的一致估计量.



区 间 估 计

点估计的缺点：无法给出估计值接近总体参数程度的信息。

- 虽然在重复抽样条件下，点估计的均值可望等于总体真值，但由于样本是随机的，抽出一个具体的样本得到的估计值很可能不同于总体真值。
- 一个点估计量的可靠性是由它的抽样标准误差来衡量的，一个具体的点估计值无法给出估计的可靠性的度量。

区 间 估 计

某商场欲购买A、B两种牌号的灯泡，B牌号的灯泡的质量比A牌号的灯泡好一些，但价格也贵一些；商场经理希望知道：两种牌号的灯泡的平均使用寿命；两种牌号的灯泡的平均使用寿命的差别；现在随机选取A牌号的灯泡8只，B牌号的灯泡10只，测得使用寿命如下（单位：小时）：

- A牌号：950,1000,1100,900,1200,1050,1150,980；
- B牌号：1200,1350,1450,1100,1500,1000,1400,1300,1250,1150；

试问：A、B两种牌号的灯泡寿命长？还是差不多？

区 间 估 计

设总体X的分布$F(x;\theta)$中含有未知参数θ, 若存在样本的两个函数(统计量) $\underline{\theta}(X_1,\cdots,X_n)$和$\overline{\theta}(X_1,\cdots,X_n)$, 使对于给定的$\alpha(0<\alpha<1)$, 有

$$P\{\underline{\theta}<\theta<\overline{\theta}\}=1-\alpha$$

则随机区间$(\underline{\theta},\overline{\theta})$称为参数$\theta$的置信水平为$1-\alpha$的双侧置信区间. 若有

$$P\{\underline{\theta}<\theta\}=1-\alpha$$

则随机区间$(\underline{\theta},+\infty)$称为参数$\theta$的置信水平为$1-\alpha$的单侧置信区间.

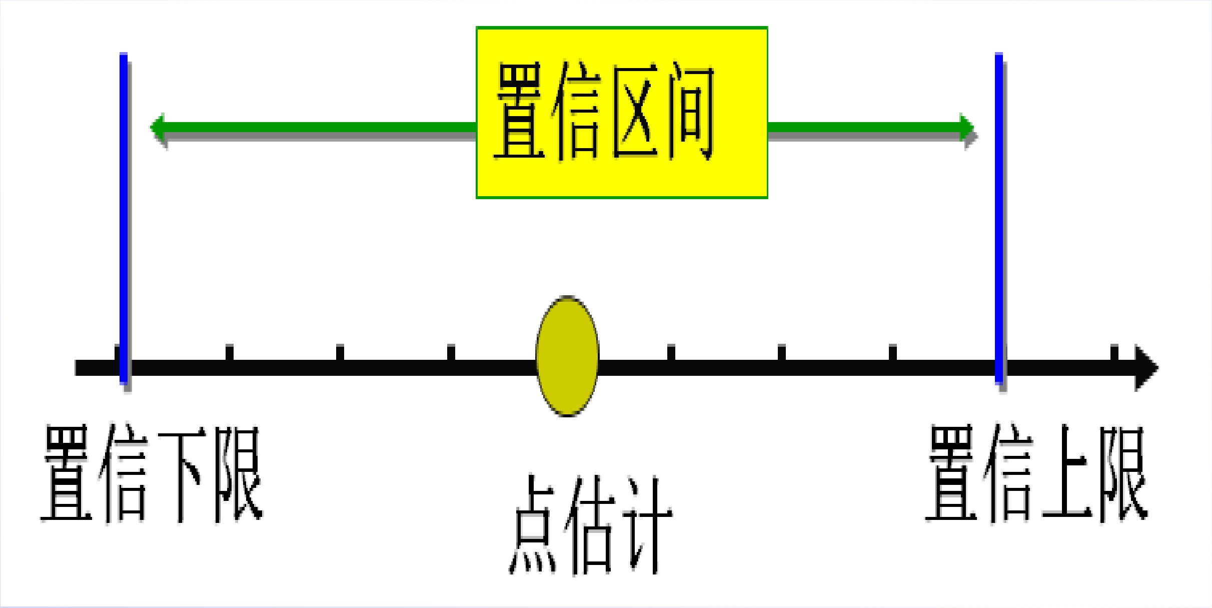
$$P\{\theta<\overline{\theta}\}=1-\alpha$$

则随机区间$(-\infty,\overline{\theta})$称为参数$\theta$的置信水平为$1-\alpha$的单侧置信区间.

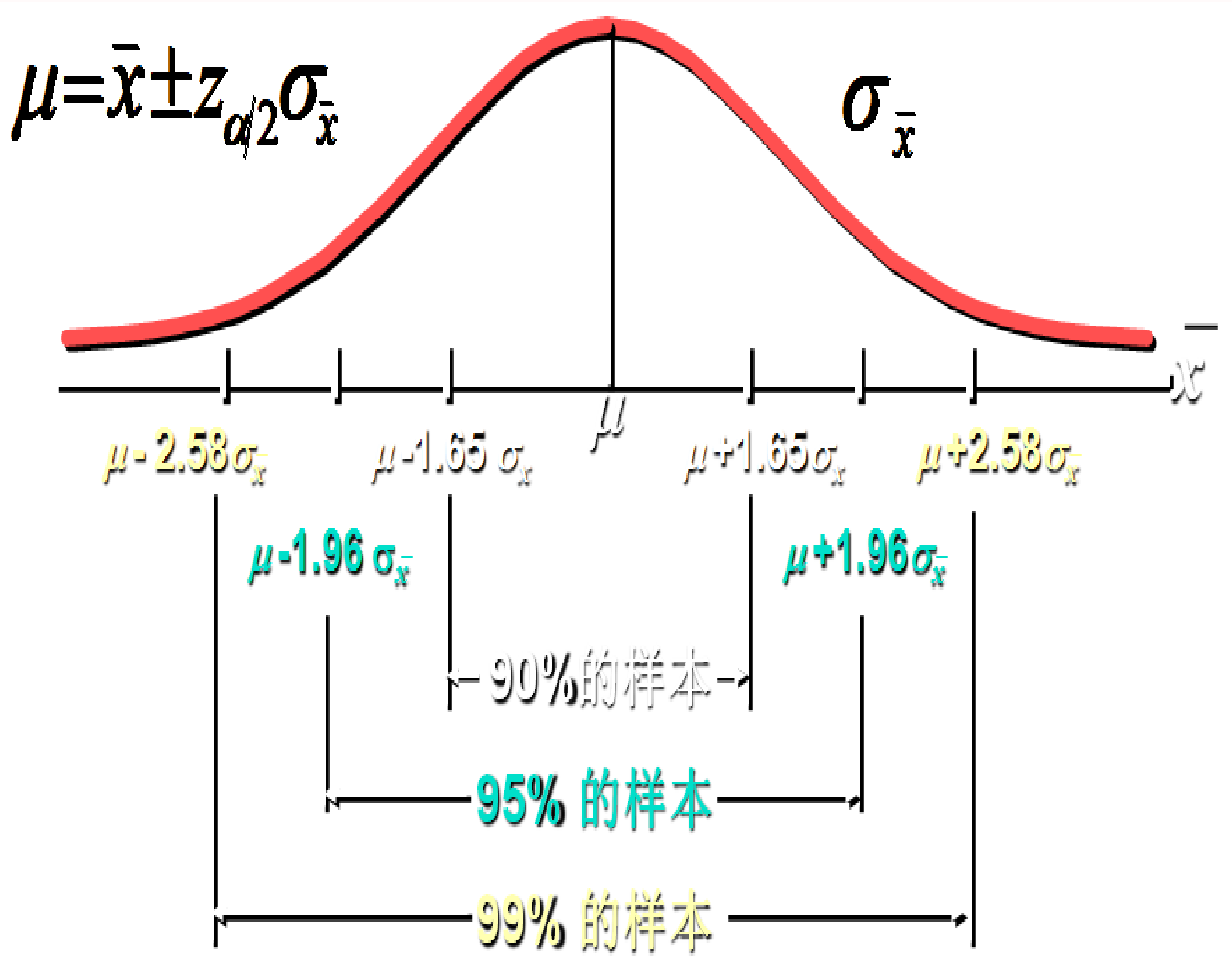
由样本统计量所构造的总体参数的估计区间称为置信区间统计学家在某种程度上确信这个区间会包含真正的总体参数，所以给它取名为置信区间

区 间 估 计

当X是连续型随机变量, 对于给定的α, 可以通过$P\{\underline{\theta} < \theta < \overline{\theta}\} = 1-\alpha$求出置信区间. 当$X$是离散型随机变量, 对于给定的$\alpha$, 不一定找到$(\underline{\theta}, \overline{\theta})$恰好$P\{\underline{\theta} < \theta < \overline{\theta}\} = 1-\alpha$, 一般找到概率不小于(但尽可能接近)$1-\alpha$的区间即可。

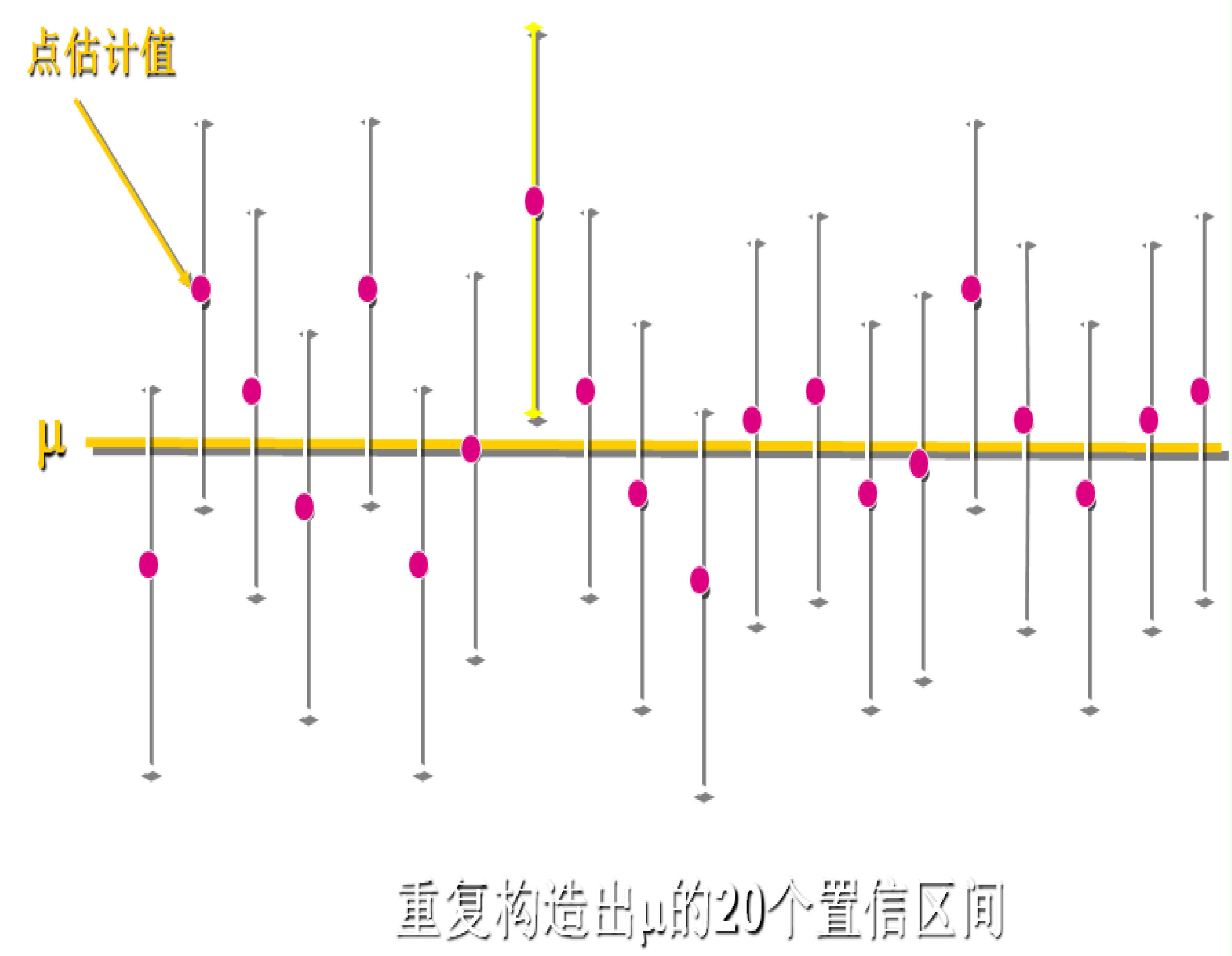


区间估计



$\mu = \bar{x} \pm z_{\alpha/2}\sigma_{\bar{x}}$
$\sigma_{\bar{x}}$
$\bar{x}$
μ
$\mu - 2.58\sigma_{\bar{x}}$
$\mu - 1.65\ \sigma_{\bar{x}}$
$\mu + 1.65\sigma_{\bar{x}}$
$\mu + 2.58\sigma_{\bar{x}}$
$\mu - 1.96\ \sigma_{\bar{x}}$
$\mu + 1.96\sigma_{\bar{x}}$
90%的样本
95% 的样本
99% 的样本

区间估计



置信水平

- 将构造置信区间的步骤重复很多次，置信区间包含总体参数真值的次数所占的比例称为置信水平
- 置信水平为(1-α), 其中α表示总体参数未在区间内的比例。
- 常用的置信水平值有99%, 95%, 90% 相应的α为0.01, 0.05, 0.10。
- 用一个具体的样本所构造的区间是一个特定的区间,我们无法知道这个样本所产生的区间是否包含总体参数的真值
  - 希望这个区间是大量包含总体参数真值的区间中的一个, 但它也可能是少数几个不包含参数真值的区间中的一个。
  - 总体参数以一定的概率落在这一区间的表述是错误的

例 子

我们先以一个例子来说明问题

Example 21 某车间用一台包装机包装葡萄糖, 包得的袋装糖重是一个随机变量, 它服从正态分布, 当机器正常时, 其均值为$\mu_0 = 0.5(kg)$, 标准差是$\sigma_0 = 0.015(kg)$. 某日开工后为检验包装机是否正常, 随机地抽取它所包装的9袋糖, 称重为

$$0.497, 0.506, 0.518, 0.524, 0.498, 0.511, 0.520, 0.515, 0.512$$

试估计包装重量的区间。

例 子

Example 22 设总体$X \sim N(\mu, \sigma^2)$, σ^2已知, μ未知, $X_1, \cdots, X_n$是来自X的样本, 求μ的置信水平为$1-\alpha$的置信区间。

解: 我们知道$\overline{X}$是μ的无偏估计. 且有

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

按标准正态分布上分位点的定义, 有

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

$$P\{\overline{X}-\sigma/\sqrt{n}z_{\alpha/2} < \mu < \overline{X}+\sigma/\sqrt{n}z_{\alpha/2}\} = 1-\alpha$$

所以置信区间为

$$(\overline{X} \pm \frac{\sigma}{\sqrt{n}}z_{\alpha/2})$$

如果取$\alpha = 0.05$, 即$1-\alpha = 0.95$, $\overline{x} = 5.20$, $\sigma = 1$, $n = 16$, 查表得$z_{\alpha/2} = z_{0.025} = 1.96$. 于是我们得到一个置信水平为0.95的置信区间为(5.20 ± 0.49).

例 子

然而, 置信水平为$1-\alpha$的置信区间并不是唯一的. 如$\alpha=0.05$, 则有

$$P\{-z_{0.04}<\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}<z_{0.01}\}=0.95$$

$$P\{\overline{X}-\frac{\sigma}{\sqrt{n}}z_{0.01}<\mu<\overline{X}+\frac{\sigma}{\sqrt{n}}z_{0.04}\}=0.95$$

$$(\overline{X}-\frac{\sigma}{\sqrt{n}}z_{0.01},\overline{X}+\frac{\sigma}{\sqrt{n}}z_{0.04})$$

这个区间长度为

$$\frac{\sigma}{\sqrt{n}}(z_{0.04}+z_{0.01})=4.08*\frac{\sigma}{\sqrt{n}}$$

而均分的区间长度为

$$2*\frac{\sigma}{\sqrt{n}}(z_{0.025})=3.92*\frac{\sigma}{\sqrt{n}}$$

置信区间的长度越短精度越高.

区间估计步骤

1. 寻找一个样本$X_1, \cdots, X_n$的函数:

$$W = W(X_1, \cdots, X_n; \theta)$$

W的分布已知, 且不含任何参数;

2. 对于给定的置信水平$1-\alpha$, 得到a, b, 使得

$$P\{a < W < b\} \geq 1-\alpha$$

3. 由$a < W < b$得到等价的不等式$\underline{\theta} < \theta < \overline{\theta}$, 其中$\underline{\theta}, \overline{\theta}$都是统计量, 那么$(\underline{\theta}, \overline{\theta})$就是$\theta$的一个置信水平为$1-\alpha$的置信区间.

9 大样本均值估计

- 参数估计条件(满足其一):
  - 总体服从正态分布, 且方差(σ^2)已知。
  - 总体不是正态分布, 大样本(经验为$n \geq 30$)可由正态分布来近似。
- 使用正态分布统计量Z

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

- 总体均值μ在1-α置信水平下的置信区间为:

  σ已知

$$\overline{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

  σ未知

$$\overline{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

例 子

Example 23 一家食品生产企业以生产袋装食品为主，为对产量质量进行监测，企业质检部门经常要进行抽检，以分析每袋重量是否符合要求。现从某天生产的一批食品中随机抽取了*25*袋，测得每袋重量如下表所示。

112.5	*108.8*	*115.6*	*100.0*	*123.5*
102.0	*101.6*	*102.2*	*116.6*	*95.4*
97.8	*108.6*	*105.0*	*136.8*	*102.8*
101.5	*98.4*	*93.3*	*101.0*	*103.0*
102.0	*100.5*	*102.6*	*107.5*	*95.0*

试以95%为置信水平建立重量均值的置信区间。

已知$X \sim N(\mu, 10^2)$, $n = 25$, $\alpha = 0.05$, $z_{\alpha/2} = 1.96$。根据样本数据计算得: $\overline{x} = 105.36$。由于是正态总体，且方差$\sigma^2 = 10^2$已知。总体均值μ在1-α置信水平下的置信区间为

$$105.36 \pm 1.96 * 10/\sqrt{25} = 105.36 \pm 3.92 = (101.44, 109.28)$$

思考：若方差未知，如何求解？

10 小样本均值估计

- 参数估计条件:
  - 总体服从正态分布, 但方差(σ^2)未知, 且小样本($n < 30$)。
- 使用正态分布统计量Z

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}} \sim t(n-1)$$

- 总体均值μ在1-α置信水平下的置信区间为
  σ未知

$$\overline{x} \pm t(n-1)_{\alpha/2}\frac{s}{\sqrt{n}}$$

例子

Example 24 某批量产品随机取*16*个, 测得其质量的样本均值和方差分别为$\overline{x} = 503.75$, $s = 6.2022$. 设其质量近似服从正态分布, 试求总体均值μ的置信水平为*0.95*的置信区间.

查表得$t_{0.025}(15) = 2.1315$, 则置信区间为

$$(503.75 \pm \frac{6.2022}{\sqrt{16}} * 2.1315)$$

即

$$(500.4, 507.1)$$

也就是说这批产品质量在500.4与507.1之间的可信度为95%.

11 方差估计

方差σ^2的置信区间:

- 总体服从正态分布
- 通常μ未知时
- s^2是σ^2的无偏估计, 则

$$\frac{(n-1)s^2}{\sigma^2} \sim \mathcal{X}^2(n-1)$$

- 则σ^2的置信水平为$1-\alpha$的置信区间为

$$P\{\mathcal{X}^2_{1-\alpha/2}(n-1) < \frac{(n-1)s^2}{\sigma^2} < \mathcal{X}^2_{\alpha/2}(n-1)\}$$

$$(\frac{(n-1)s^2}{\mathcal{X}^2_{\alpha/2}(n-1)}, \frac{(n-1)s^2}{\mathcal{X}^2_{1-\alpha/2}(n-1)})$$

例子

Example 25 一家食品生产企业以生产袋装食品为主, 为对产量质量进行监测, 企业质检部门经常要进行抽检, 以分析每袋重量是否符合要求。现从某天生产的一批食品中随机抽取了*25*袋, 测得每袋重量如下表所示。

112.5	*108.8*	*115.6*	*100.0*	*123.5*
102.0	*101.6*	*102.2*	*116.6*	*95.4*
97.8	*108.6*	*105.0*	*136.8*	*102.8*
101.5	*98.4*	*93.3*	*101.0*	*103.0*
102.0	*100.5*	*102.6*	*107.5*	*95.0*

试以95%为置信水平建立重量方差的置信区间。

$n=25$, 根据样本数据计算得: $s^2=93.21$。$\alpha=0.05$查表得

$$\chi^2_{0.025}(24)=39.3641, \chi^2_{0.975}(24)=12.4011$$

由于是正态总体, 方差σ^2置信水平为95%的置信区间为

$$(\frac{24*93.21}{39.3641}, \frac{24*93.21}{12.4011})=(56.83, 180.39)$$

12 大样本均值差估计

- 参数估计条件(都要满足):
  - 两个样本是独立的随机样本。
  - 两总体服从正态分布且方差(σ_1^2,σ_2^2)已知, 否则要求大样本($n_1 \geq 30$, $n_2 \geq 30$)。
- 使用正态分布统计量Z

$$Z = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/n_1 + \sigma^2/n_2)}} \sim N(0,1)$$

- 总体均值差$\mu_1 - \mu_2$在1-α置信水平下的置信区间为:
  方差(σ_1^2,σ_2^2)已知

$$(\overline{X_1} - \overline{X_2} \pm Z_{\alpha/2}\sqrt{(\sigma_1^2/n_1 + \sigma_2^2/n_2)}$$

方差(σ_1^2,σ_2^2)未知

$$(\overline{X_1} - \overline{X_2} \pm Z_{\alpha/2}\sqrt{(s_1^2/n_1 + s_2^2/n_2)}$$

例 子

Example 26 某地区教育管理部门想估计两所中学的学生高考时的英语平均分数之差，为此在两所中学独立抽取两个随机样本, 有关数据如下表所示。

中学A	$n_1 = 46$	$\overline{x_1} = 86$	$s_1 = 5.8$
中学B	$n_1 = 33$	$\overline{x_1} = 78$	$s_1 = 7.2$

试建立两所中学高考英语平均分数之差95%的置信区间。

解: 两个总体均值之差在1-α置信水平下的置信区间为

$$(\overline{X_1} - \overline{X_2} \pm Z_{\alpha/2}\sqrt{s_1^2/n_1 + s_2^2/n_2})$$

$$= (86 - 78 \pm 1.96\sqrt{5.8^2/46 + 7.2^2/33})$$
$$= 8 \pm 2.97 = (5.03, 10.97)$$

13 小样本均值差估计

- 参数估计条件(都要满足)：
  - 两个样本是独立的随机样本。
  - 两个总体都服从正态分布。
  - 两个总体方差未知但相等。
  - 两个独立的小样本。
- 总体均值差$\mu_1-\mu_2$在1-α置信水平下的置信区间为：

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S_w\sqrt{1/n_1+1/n_2}}\sim t(n_1+n_2-2)$$

$$(\overline{X}-\overline{Y}\pm t_{\alpha/2}(n_1+n_2-2)S_w/\sqrt{1/n_1+1/n_2})$$

其中$S_w=\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$

例 子

Example 27 比较两种型号子弹的枪口速度, 设两总体近似服从正态分布且方差相等. 对于甲, 随机抽*10*发, 测得$\overline{x}_1 = 500$, $s_1 = 1.10$; 对于乙, 随机取*20*发, 测得测得$\overline{x}_1 = 496$, $s_1 = 1.20$. 求两总体差$\mu_1 - \mu_2$置信水平为*0.95*的置信区间.

解:$n_1 = 10, n_2 = 20, t_{0.025}(28) = 2.0484,$

$$s_w = \sqrt{(9 * 1.10^2 + 19 * 1.20^2)/28} = 1.1688$$

则$\mu_1 - \mu_2$的一个置信水平为0.95的置信区间为

$$(\overline{x}_1 - \overline{x}_2 \pm s_w \times t_{0.025}(28)\sqrt{1/10 + 1/20}) = (4 \pm 0.93) = (3.07, 4.93)$$

置信区间下限大于零, 因此在0.95可信度下认为μ_1比μ_2大.

14 方差比估计

- 方差比σ_1^2/σ_2^2的置信区间.
  - (1) μ_1,μ_2已知时, σ_1^2/σ_2^2的置信水平为$1-\alpha$的置信区间为

$$\left(\frac{n_2\sum_{i=1}^{n_1}(X_i-\mu_1)^2}{n_1F_{\alpha/2}(n_1,n_2)\sum_{i=1}^{n_2}(Y_i-\mu_2)^2},\frac{n_2\sum_{i=1}^{n_1}(X_i-\mu_1)^2}{n_1F_{1-\alpha/2}(n_1,n_2)\sum_{i=1}^{n_2}(Y_i-\mu_2)^2}\right)$$

  - (2) μ_1,μ_2未知时,

$$\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2}\sim F(n_1-1,n_2-1)$$

σ_1^2/σ_2^2的置信水平为$1-\alpha$的置信区间为

$$\left(\frac{S_1^2}{S_2^2F_{\alpha/2}(n_1-1,n_2-1)},\frac{S_1^2}{S_2^2F_{1-\alpha/2}(n_1-1,n_2-1)}\right)$$

例 子

Example 28 比较两种方法生产产品. 方法*A*生产*18*个, $s_1^2 = 0.34$; 方法*B*生产*13*个, $s_2^2 = 0.29$. 两产品近似服从正态分布但μ_i, σ_i未知. 试求方差比σ_1^2/σ_2^2的置信水平为*0.90*的置信区间.

解:

$$F_{\alpha/2}(n_1-1, n_2-1) = F_{0.05}(17,12) = 2.59$$

$$F_{1-\alpha/2}(n_1-1, n_2-1) = F_{0.95}(17,12) = \frac{1}{F_{0.05}(12,17)} = \frac{1}{2.38}$$

则置信区间为

$$(\frac{0.34}{0.29} \times \frac{1}{2.59}, \frac{0.34}{0.29} \times 2.38) = (0.45, 2.79)$$

区间包含1, 认为两者没有显著差别.

15 0-1分布的置信区间

已知0-1分布$X \sim b(1,p)$的均值和方差分别为

$$\mu = p, \sigma^2 = p(1-p)$$

设$X_1, \cdots, X_n$为样本, 若样本容量较大($n \geq 50$), 则有中心极限定理可知

$$\frac{\sum_{i=1}^{n} X_i - np}{\sqrt{np(1-p)}} = \frac{n\overline{X} - np}{\sqrt{np(1-p)}} \sim N(0,1)$$

则有

$$P\{-z_{\alpha/2} < \frac{n\overline{X} - np}{\sqrt{np(1-p)}} < z_{\alpha/2}\} \approx 1 - \alpha$$

其中不等式

$$-z_{\alpha/2} < \frac{n\overline{X} - np}{\sqrt{np(1-p)}} < z_{\alpha/2}$$

等价于

$$(n + z_{\alpha/2}^2)p^2 - (2n\overline{X} + z_{\alpha/2}^2)p + n\overline{X}^2 < 0$$

则置信区间为

$$(\frac{-b - \sqrt{b^2 - 4ac}}{2a}, \frac{-b + \sqrt{b^2 - 4ac}}{2a})$$

其中$a = (n + z_{\alpha/2}^2)$, $b = -(2n\overline{X} + z_{\alpha/2}^2)$, $c = n\overline{X}^2$.

例 子

Example 29 一大批产品的*100*个样品中合格个数为*60*, 求合格率p置信水平为*0.95*的置信区间.

解: $\alpha = 0.05, n = 100, \overline{x} = 0.60, z_{\alpha/2} = z_{0.025} = 1.96.$ 则

$$a = (n + z_{\alpha/2}^2) = 103.84$$

$$b = -(2n\overline{X} + z_{\alpha/2}^2) = -123.84$$

$$c = n\overline{X}^2 = 36$$

所以置信区间为

$$(0.50, 0.69)$$

16 单侧置信区间

在一些估计中,我们只关心某一侧的区间. 例如

$$P\{\mu > \underline{\mu}\} = 1 - \alpha$$

或者

$$P\{\mu < \overline{\mu}\} = 1 - \alpha$$

其中$\underline{\mu}$和$\overline{\mu}$分别称为单侧置信区间的下限和上限. 例如在正态总体估计中(μ, σ均未知), 则

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

$$P\{\frac{\overline{X} - \mu}{S/\sqrt{n}} < t_\alpha(n-1)\} = 1 - \alpha$$

所以单侧置信区间为

$$(\overline{X} - \frac{S}{\sqrt{n}} t_\alpha(n-1), \infty)$$

即

$$\underline{\mu} = \overline{X} - \frac{S}{\sqrt{n}} t_\alpha(n-1)$$

例 子

Example 30 某产品寿命近似服从正态分布, 随机取5只, 测得$\overline{x} = 1160$, $\sigma^2 = 9950$, 求单侧置信区间下限.

解: $t_\alpha(n-1) = t_{0.05}(4) = 2.1318$

$$\underline{\mu} = \overline{x} - \frac{s}{\sqrt{n}} t_\alpha(n-1) = 1065$$

内 容 回 顾

- 点估计
  - 矩估计法
  - 最大似然估计法
  - 估计量的评选标准
- 区间估计
  - 区间估计的基本概念
  - 正态总体均值与方差的区间估计
  - (0-1)分布参数的区间估计
  - 单侧置信区间