则称 θ_1 比 θ_2 有效.

例 10.11　设 $X_1, X_2, \cdots, X_n$ 是来自总体 X 的样本, 且 X 的概率密度为

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

令 $Z = \min\{X_1, X_2, \cdots, X_n\}$, 证明: 当 $n > 1$ 时 $\bar{X} = \sum_{i=1}^n X_i/n$ 比 nZ 有效.

证明　根据独立性有

$$\mathrm{Var}(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(X_i) = \frac{\theta^2}{n}.$$

根据例 10.10 可知随机变量 Z 的概率密度为

$$f(z) = \begin{cases} 0 & z < 0 \\ \frac{n}{\theta}e^{-\frac{nz}{\theta}} & z \geqslant 0 \end{cases}$$

从而得到

$$\mathrm{Var}(nZ) = n^2\mathrm{Var}(Z) = n^2\frac{\theta^2}{n^2} = \theta^2,$$

因此当 $n \geqslant 1$ 时有 $\mathrm{Var}(\bar{X}) \leqslant \mathrm{Var}(nZ)$ 成立, 故 $\bar{X}$ 比 nZ 有效.

例 10.12　设 $X_1, X_2, \cdots, X_n$ 是总体 X 的样本, 且 $E(X) = \mu$ 和 $\mathrm{Var}(X) = \sigma^2$. 设常数 $c_1, c_2, \cdots, c_n \geqslant 0$ 满足 $\sum_{i=1}^n c_i = 1, c_i \neq 1/n$, 求证: $\bar{X}$ 比 $\sum_{i=1}^n c_iX_i$ 有效.

证明　根据样本的独立同分布条件有

$$E\left[\bar{X}\right] = \mu \quad 和 \quad \mathrm{Var}\left(\bar{X}\right) = \sigma^2/n.$$

根据期望的性质有 $E[\sum_{i=1}^n c_iX_i] = \mu$, 进一步有

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

这里利用不等式 $\sum_{i=1}^n c_i^2/n \geqslant (\sum_{i=1}^n c_i/n)^2 = 1/n^2$, 所以有 $\mathrm{Var}(\sum_{i=1}^n c_iX_i) \geqslant \mathrm{Var}(\bar{X})$.

下面定义有效统计量:

定理 10.1 (Rao-Crammer 不等式)　设随机变量 X 的概率密度为 $f(x;\theta)$ 或分布函数为 $F(x;\theta)$, 令

$$\mathrm{Var}_0(\theta) = \frac{1}{nE\left[\left(\frac{\partial \ln f(X;\theta)}{\partial\theta}\right)^2\right]} \quad 或 \quad \mathrm{Var}_0(\theta) = \frac{1}{nE\left[\left(\frac{\partial \ln F(X;\theta)}{\partial\theta}\right)^2\right]},$$

对任意的无偏估计量 $\hat{\theta}$ 有

$$\operatorname{Var}(\hat{\theta}) \geqslant \operatorname{Var}_0(\theta),$$

称 $\operatorname{Var}_0(\theta)$ 为估计量 $\hat{\theta}$ 方差的下界. 当 $\operatorname{Var}(\hat{\theta}) = \operatorname{Var}_0(\theta)$ 时称 $\hat{\theta}$ 为达到方差下界的无偏估计量, 此时 $\hat{\theta}$ 为最有效估计量, 简称有效估计量.

例 10.13 设 $X_1, X_2, \cdots, X_n$ 为总体 X 的样本, 令总体 X 的密度函数为

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}} & x > 0 \\ 0 & x \leqslant 0, \end{cases}$$

证明: θ 的最大似然估计为有效估计量.

解 首先计算对数似然函数

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

进一步得到统计量的方差

$$\operatorname{Var}(\hat{\theta}) = \operatorname{Var}\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right) = \frac{\theta^2}{n}.$$

同时考察

$$\ln f(X;\theta) = -\ln \theta - \frac{X}{\theta}, \quad \frac{\partial \ln f(X;\theta)}{\partial \theta} = -\frac{1}{\theta} + \frac{X}{\theta^2}$$

所以

$$E\left[\frac{\partial \ln f(X;\theta)}{\partial \theta}\right]^2 = E\left[\left(-\frac{1}{\theta} + \frac{X}{\theta^2}\right)^2\right] = \frac{1}{\theta^4} E[(X - E[X])^2] = \frac{1}{\theta^2},$$

从而得到 $\operatorname{Var}_0(X) = \theta^2/n = \operatorname{Var}(\hat{\theta})$, 因此 θ 的最大似然估计是有效估计量.

10.2.1.2 一致性

定义 10.3 设 $\hat{\theta}_n = \hat{\theta}_n(X_1, X_2, \cdots, X_n)$ 是 θ 的一个估计量, 若当 $n \to \infty$ 时有 $\hat{\theta}_n \xrightarrow{P} \theta$ 成立, 即对任意 $\epsilon > 0$ 有

$$\lim_{n \to 0} \Pr[|\hat{\theta}_n - \theta| > \epsilon] = 0,$$

则称 $\hat{\theta}_n$ 为 θ 的一致估计量.

估计量的一致性刻画了在足够多样本情形下估计量 $\hat{\theta}$ 能有效逼近真实值 θ, 一致性是对估计的基本要求, 不满足一致性的估计量一般不予考虑. 下面给出满足一致性的充分条件:

定理 10.2 设 $\hat{\theta}_n = \hat{\theta}_n(X_1, X_2, \cdots, X_n)$ 是 θ 的一个估计量, 若满足以下两个条件:

$$\lim_{n \to \infty} E\left[\hat{\theta}_n\right] = \theta \quad 和 \quad \lim_{n \to \infty} \operatorname{Var}\left(\hat{\theta}_n\right) = 0,$$

则 $\hat{\theta}_n$ 为 θ 的一致估计量.

证明 根据 $\lim_{n\to\infty} E[\hat{\theta}_n] = \theta$ 知道对任意 $\epsilon > 0$, 存在一个N_0, 当 $n \geqslant N_0$ 有 $|E[\hat{\theta}_n] - \theta| \leqslant \theta/2$, 于是有

$$\lim_{n\to\infty} \Pr\left[|E[\hat{\theta}_n] - \theta| > \epsilon/2\right] = 0.$$

根据 Chebyshev 不等式有

$$\lim_{n\to 0} \Pr\left[\left|\hat{\theta}_n - E[\hat{\theta}_n]\right| > \epsilon/2\right] \leqslant \lim_{n\to 0} \frac{4}{\epsilon}\mathrm{Var}(\hat{\theta}_n) = 0$$

再根据

$$\Pr\left[|\hat{\theta}_n - \theta| > \epsilon\right] \leqslant \Pr\left[\left|\hat{\theta}_n - E[\hat{\theta}_n]\right| > \epsilon/2\right] + \Pr\left[\left|E[\hat{\theta}_n] - \theta\right| > \epsilon/2\right]$$

完成证明.

定理 10.3 设 $\hat{\theta}_{n_1}, \hat{\theta}_{n_2}, \cdots, \hat{\theta}_{n_k}$ 分别为 $\theta_1, \theta_2, \cdots, \theta_k$ 满足一致性的估计量, 对连续函数 $g\colon \mathbb{R}^n \to \mathbb{R}$, 有函数 $\hat{\eta}_n = g(\hat{\theta}_{n_1}, \hat{\theta}_{n_2}, \cdots, \hat{\theta}_{n_k})$ 是 $\eta = g(\theta_1, \theta_2, \cdots, \theta_k)$ 满足一致性的估计量.

根据大数定理可知样本的 k 阶矩是总体 k 阶矩的一致估计量. 矩估计法得到的估计量一般是一致估计量. 最大似然估计量在一定条件下是一致性估计量.

例 10.14 设$X_1, X_2, \cdots, X_n$是来自总体X的样本, 以及总体 X 的密度函数为

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

则样本均值 $X_n = \sum_{i=1}^n X_i/n$ 为 θ 的无偏、有效、一致估计量.

由前面的例子可知估计的无偏性和有效性, 一致性可根据 $E[X_n] = \theta$ 以及

$$\lim_{n\to\infty} Var(\bar{X}) = \lim_{n\to\infty} \frac{\theta^2}{n} = 0.$$

例 10.15 设 $X_1, X_2, \cdots, X_n$ 是来自总体 $X \sim U(0, \theta)$ 的样本, 证明: θ 的最大似然估计量是一致估计量.

证明 根据前面的例题可知 θ 的最大似然估计为 $\hat{\theta}_n = \max(X_1, X_2, \cdots, X_n)$. 设随机变量 $Z = \max(X_1, X_2, \cdots, X_n)$, 则由 Z 的分布函数

$$F_Z(z) = \Pr[Z \leqslant z] = \Pr[\max(X_1, X_2, \cdots, X_n) \leqslant z] = \prod_{i=1}^n \Pr[X_i \leqslant z] = \begin{cases} 1 & z > \theta \\ (\frac{z}{\theta})^n & z \in [0, \theta] \\ 0 & z < 0. \end{cases}$$

由此得到当 $z \in [0,\theta]$ 时随机变量 Z 的密度函数 $f_Z(z) = nz^{n-1}/\theta^n$, 进一步有

$$E\left[\hat{\theta}_n\right] = E[Z] = \int_0^\theta \frac{nz^n}{\theta^n}dz = \frac{n}{n+1}\theta,$$

因此 $\hat{\theta}$ 是 θ 的有偏估计. 另一方面有

$$E\left[Z^2\right] = \int_0^\theta \frac{nz^{n+1}}{\theta^n}dz = \frac{n}{n+2}\theta^2,$$

从而得到

$$\mathrm{Var}\left(\hat{\theta}_n\right) = \mathrm{Var}(Z) = E[Z^2] - (E[Z])^2 = \frac{n}{n+2}\theta^2 - (\frac{n\theta}{n+1})^2 = \frac{n}{(n+1)^2(n+2)}\theta^2,$$

于是有

$$\lim_{n\to\infty} E[\hat{\theta}_n] = \theta \quad \text{和} \quad \lim_{n\to\infty} \mathrm{Var}(\hat{\theta}_n) = 0,$$

由此可得 $\hat{\theta}$ 是 θ 的有偏、但一致估计量.

10.3 区间估计

区间估计问题: 设 $X_1, X_2, \cdots, X_n$ 是来自总体 X 的样本, θ 为总体 X 的分布函数 $F(x,\theta)$ 的未知参数, 根据样本估计 θ 的范围 $(\hat{\theta}_1, \hat{\theta}_2)$, 其中 $\hat{\theta}_1 = \hat{\theta}_1(X_1, X_2, \cdots, X_n)$ 和 $\hat{\theta}_2 = \hat{\theta}_2(X_1, X_2, \cdots, X_n)$, 使得以较大的概率保证有 $\theta \in (\hat{\theta}_1, \hat{\theta}_2)$ 成立. 具体而言, 对任意给定 $\alpha \in (0,1)$, 有

$$\Pr\left[\hat{\theta}_1(X_1, X_2, \cdots, X_n) < \theta < \hat{\theta}_2(X_1, X_2, \cdots, X_n)\right] \geqslant 1 - \alpha.$$

定义 10.4 (置信区间与置信度) 设 $X_1, X_2 \cdots, X_n$ 是来自总体 X 的样本, 总体 X 的分布函数含未知参数 θ, 找出统计量 $\hat{\theta}_1 = \hat{\theta}_1(X_1, X_2, \cdots, X_n)$ 和 $\hat{\theta}_2 = \hat{\theta}_2(X_1, X_2, \cdots, X_n)$ $(\hat{\theta}_1 < \hat{\theta}_2)$, 使得

$$\Pr\left[\hat{\theta}_1 < \theta < \hat{\theta}_2\right] \geqslant 1 - \alpha$$

成立, 则称 $1-\alpha$ 为置信度, $[\hat{\theta}_1, \hat{\theta}_2]$ 为 θ 的置信度为 $1-\alpha$ 的置信区间.

注意: 置信区间 $[\hat{\theta}_1, \hat{\theta}_2]$ 是随机区间, $1-\alpha$ 为该区间包含 θ 的概率/可靠程度. 若 $\alpha = 0.05$, 则置信度为 95%. 通常采用 95% 的置信度, 有时也可 99% 或 90%等. 说明:

i) $\hat{\theta}_2 - \hat{\theta}_1$ 反映了估计精度, 长度越小精度越大.

ii) α 反映了估计的可靠度, α 越小可靠度越高.

iii) 给定 α, 区间 $[\hat{\theta}_1, \hat{\theta}_2]$ 的选取并不唯一确定, 通常选长度最小的一个区间.

置信区间的求解方法: **枢轴变量法**.

1) 先找一样本函数 $W(X_1, X_2, \cdots, X_n; \theta)$ 包含待估参数 θ, 但不含其它参数, 函数 W 的分布已知, 称 W 为枢轴变量.

2) 给定置信度 $1-\alpha$, 根据 W 的分布找出临界值 a 和 b, 使得 $\Pr[a < W < b] = 1-\alpha$ 成立.

3) 根据 $a < W < b$ 解出 $\hat{\theta}_1 < \theta < \hat{\theta}_2$, 则 $(\hat{\theta}_1, \hat{\theta}_2)$ 为 θ 的置信度为 $1-\alpha$ 的置信区间.

10.3.1 正态总体, 方差已知, 求期望的区间估计

设 $X_1, X_2, \cdots, X_n$ 是来自总体 $X \sim \mathcal{N}(\mu, \sigma^2)$ 的样本, 若方差 σ^2 已知. 给定 $\alpha \in (0,1)$, 确定置信度为 $1-\alpha$ 下 μ 的置信区间 $[\hat{\theta}_1, \hat{\theta}_2]$. 令样本均值为 $\bar{X} = \sum_{i=1}^n X_i/n$, 根据正太分布的性质找出枢轴变量:

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

给定置信度 $1-\alpha$, 找出临界值 a 和 b 使得

$$\Pr[a < W < b] = 1-\alpha.$$

根据正态分布的性质、对称性和上分位点可知

$$\Pr[W \geqslant \mu_{\alpha/2}] = 1-\alpha/2 \quad 和 \quad \Pr[W \leqslant -\mu_{\alpha/2}] = 1-\alpha/2.$$

求解可得 $a = -\mu_{\alpha/2}$ 和 $b = \mu_{\alpha/2}$. 于是有

$$\Pr[-\mu_{\alpha/2} < W < \mu_{\alpha/2}] = 1-\alpha.$$

根据 $W = (\bar{X}-\mu)/(\sigma/\sqrt{n})$ 可得

$$\Pr\left[\bar{X} - \frac{\sigma}{\sqrt{n}}\mu_{\alpha/2} < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}}\mu_{\alpha/2}\right] = 1-\alpha.$$

例 10.16 某地区儿童身高服从正态分布, 现随机抽查 9 人, 高度分别为 $115, 120, 131, 115, 109, 115, 115, 105, 110$, 已知 $\sigma^2 = 7$ 和置信度为 95%, 求期望 μ 的置信区间 $(\mu_{0.025} = 1.96)$.

10.3.2 正态总体, 方差未知, 求期望的区间估计

设 $X_1, X_2, \cdots, X_n$ 是来自总体 $X \sim \mathcal{N}(\mu, \sigma^2)$ 的样本, 若方差 σ^2 未知, 考虑期望 μ 的置信度为 $1-\alpha$ 的置信区间. 设 $\bar{X} = \sum_{i=1}^n X_i/n$ 和 $S^2 = \sum_{i=1}^n (X_i-\bar{X})^2/(n-1)$, 根据正态总体抽样定理可知:

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

由此设枢轴变量

$$W = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1).$$

给定置信度 $1-\alpha$, 设临界值 a 和 b 满足

$$\Pr[a \leqslant W \leqslant b] = 1-\alpha \quad \Rightarrow \quad b = t_{\alpha/2}(n-1),\ a = -t_{\alpha/2}(n-1).$$

整理可得

$$\Pr\left[\bar{X} - \frac{S}{\sqrt{n}} t_{\alpha/2}(n-1) < \mu < \bar{X} + \frac{S}{\sqrt{n}} t_{\alpha/2}(n-1)\right] = 1-\alpha.$$

10.3.3 正态总体, 求方差 σ^2 的置信区间

设 $X_1, X_2, \cdots, X_n$ 是来自总体 $X \sim \mathcal{N}(\mu, \sigma^2)$ 的样本, 考虑方差 σ^2 的置信度为 $1-\alpha$ 的置信区间. 设修正样本方差 $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$, 根据正态总体抽样定理有

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

由此设枢轴变量 $W = (n-1)S^2/\sigma^2$, 设临界值 a 和 b 满足

$$\Pr[a \leqslant W \leqslant b] = 1-\alpha.$$

根据 χ^2 分布的不对称性, 采用概率对称的区间

$$\Pr[W \leqslant a] = \Pr[\geqslant b] = \alpha/2 \quad \Rightarrow \quad b = \chi^2_{\alpha/2}(n-1), \quad a = \chi^2_{1-\alpha/2}(n-1).$$

根据枢轴变量 $W = (n-1)S^2/\sigma^2$ 可得

$$\Pr\left[\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}\right] = 1-\alpha.$$

10.3.4 双正态总体情形

设 $X_1, X_2, \cdots, X_n$ 是来自总体 $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ 的样本, 设 $Y_1, Y_2, \cdots, Y_m$ 是总体 $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$的样本, 令

$$\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i, \quad \bar{Y} = \frac{1}{m}\sum_{i=1}^m Y_i, \quad S_1^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2, \quad S_2^2 = \frac{1}{m-1}\sum_{i=1}^m (Y_i - \bar{Y})^2.$$

考虑 $\mu_1 - \mu_2$ 和 σ_1^2/σ_2^2 的置信度为 $1-\alpha$ 的区间估计.

1) 已知方差 σ_1^2 和 σ_2^2, 求 $\mu_1 - \mu_2$ 的置信区间. 根据正态分布的性质有

$$\bar{X} \sim \mathcal{N}(\mu_1, \frac{\sigma_1^2}{n}), \quad \bar{Y} \sim \mathcal{N}(\mu_2, \frac{\sigma_2^2}{m}) \quad \bar{X} - \bar{Y} \sim \mathcal{N}(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}),$$

进一步有

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

于是设枢轴变量

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

求解置信区间

$$\Pr\left[\bar{X}-\bar{Y}-\mu_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}<\mu_1-\mu_2<\bar{X}-\bar{Y}+\mu_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}\right]=1-\alpha.$$

2) 若 σ_1^2 和 σ_2^2未知, 但已知 $\sigma_1^2=\sigma_2^2$, 设

$$S_W=\frac{(n-1)S_1^2+(m-1)S_2^2}{n+m-2},$$

则考虑枢轴变量

$$W=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_W\sqrt{\frac{1}{n}+\frac{1}{m}}}\sim t(n+m-2).$$

于是有

$$\Pr\left[-t_{\alpha/2}(n+m-2)<\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_W\sqrt{\frac{1}{n}+\frac{1}{m}}}<t_{\alpha/2}(n+m-2)\right]=1-\alpha.$$

3) 求方差比 σ_1^2/σ_2^2 的置信度为 $1-\alpha$ 的置信区间. 设枢轴变量

$$W=\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}\sim F(n-1,m-1),$$

根据 F 分布的不对称性, 采用概率对称的区间

$$\Pr[W\leqslant a]=\Pr[W\geqslant b]=\alpha/2\quad\Rightarrow\quad b=F_{\frac{\alpha}{2}}(n-1,m-1),\quad a=F_{1-\alpha/2}(n-1,m-1).$$

由此可得置信区间

$$\Pr\left[\frac{S_1^2}{S_2^2}\frac{1}{F_{\alpha/2}(n-1,m-1)}<\frac{\sigma_1^2}{\sigma_2^2}<\frac{S_1^2}{S_2^2}\frac{1}{F_{1-\alpha/2}(n-1,m-1)}\right]=1-\alpha.$$

10.3.5 单侧置信区间

对某些实际问题, 我们往往只关心置信区间的上限或下限, 例如, 次品率只关心上限, 产品的寿命只关心下限, 由此引入单侧置信区间及其估计.

定义 10.5 (单侧置信区间) 给定 $\alpha \in (0,1)$, 若样本 $X_1,\cdots,X_n$ 的统计量 $\hat{\theta}_1 = \hat{\theta}_1(X_1,X_2,\cdots,X_n)$ 满足

$$\Pr\left[\theta > \hat{\theta}_1\right] \geqslant 1-\alpha,$$

则称 $(\hat{\theta}_1,+\infty)$ 为 θ 的置信度为 $1-\alpha$ 的单侧置信区间, $\hat{\theta}_1$ 称为单侧置信下限.

同理定义单侧置信上限. 对正态总体, 可以将相关置信区间的估计都扩展到单侧置信估计, 枢轴变量的定理类似, 我们将不再重复讨论, 下面仅举两个实例:

例 10.17 设 $X_1,X_2,\cdots,X_n$ 是来自总体 $X\sim\mathcal{N}(\mu,\sigma^2)$ 的样本, 若方差 σ^2 已知, 求 μ 的置信度为 $1-\alpha$ 的单侧置信下限和上限.

解 设样本均值 $\bar{X}=\sum_{i=1}^n X_i/n$, 根据正态分布的性质考虑枢轴变量

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

于是有

$$\Pr\left[\frac{\bar{X}-\mu}{S/\sqrt{n}}<\mu_\alpha\right]=1-\alpha,\quad \Pr\left[\frac{\bar{X}-\mu}{S/\sqrt{n}}>-\mu_\alpha\right]=1-\alpha,$$

整理计算完成估计.

例 10.18 从一批出厂的灯泡中随机抽取 10 盏灯泡, 测试其寿命分别为: 1000, 1500, 1250, 1050, 950, 1000, 1150, 1050, 950, 1000, (单位: 小时). 假设这批灯泡的寿命服从正态分布, 求这批灯泡平均寿命的置信度为 95% 的单侧置信下限.

解 首先计算样本均值和样本修正方差分别为

$$\bar{X}=\sum_{i=1}^{10}X_i/10=1090 \quad 和 \quad S^2=\sum_{i=1}^{10}(X_i-\bar{X})^2/9=8800/3.$$

根据正态分布的性质考虑枢轴变量

$$W=\frac{\bar{X}-\mu}{S/3}\sim t(9),$$

于是有

$$\Pr\left[\frac{\bar{X}-\mu}{S/3}<t_{0.05}(9)\right]=0.95,$$

查表 $t_{0.05}(9)=1.833$ 可得

$$\mu>\bar{X}-t_{0.05}(9)S/3=1090-\sqrt{8800/3}\times1.833/3>1056.$$

10.3.6 非正态分布的区间估计

设 $X_1,X_2,\cdots,X_n$ 是来自总体 X 的样本, 若总体 X 的分布未知或非正态分布, 我们可以给出总体期望 $\mu=E[X]$ 的区间估计, 方法分为两种: 利用 Concentration 不等式和中心极限定理.

(1) 首先考虑 Concentration 不等式, 若总体 $X \in [a,b]$, 设 $\bar{X} = \sum_{i=1}^n X_i/n$, 根据 Concentration 不等式有

$$\Pr\left[|\mu - \bar{X}| \geqslant \epsilon\right] \leqslant 2\exp(-2n\epsilon^2/(b-a)^2).$$

令 $\alpha = 2\exp(-2n\epsilon^2/(b-a)^2)$ 求解 $\epsilon = \sqrt{(b-a)^2\ln(2/\alpha)/n}$, 于是有

$$\Pr\left[\bar{X} - \sqrt{(b-a)^2\ln(2/\alpha)/n} < \mu < \bar{X} + \sqrt{(b-a)^2\ln(2/\alpha)/n}\right] > 1-\alpha.$$

可基于其它 Concentration 不等式给出类似的置信区间估计, 以及其它 sub-Gaussian 型随机变量的期望的置信区间估计.

(2) 利用中心极限定理, 求枢轴变量的近似分布, 再给出置信区间估计. 设总体 X 的期望 $E(X) = \mu$, 方差 $\mathrm{Var}(X) = \sigma^2$, 利用中心极限定理设枢轴变量

$$W = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

枢轴变量 W 的分布近似于标准正态分布 $\mathcal{N}(0,1)$. 当方差 σ^2 已知时有

$$\Pr\left[-\mu_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \mu_{\alpha/2}\right] \approx 1-\alpha.$$

当方差 σ^2 未时, 用修正样本方差 $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$ 代替方差 σ^2, 于是有

$$\Pr\left[-\mu_{\alpha/2} < \frac{\bar{X} - \mu}{S^2/\sqrt{n}} < \mu_{\alpha/2}\right] \approx 1-\alpha.$$

例 10.19 设 $X_1, X_2, \cdots, X_n$ 是来自总体 $X \sim \mathrm{Ber}(p)$ 的样本, 求 p 的置信度为 $1-\alpha$ 的区间估计.

解 根据 Bernoulli 分布的性质有 $X_i \in \{0,1\}$ 以及 $p = E[X]$, 根据 Chernoff 不等式有

$$\Pr[|\bar{X} - p| > \epsilon p] \leqslant 2\exp(-np\epsilon^2/3),$$

设 $\alpha = 2\exp(-np\epsilon^2/3)$, 于是有

$$\Pr\left[\bar{X} - \sqrt{3p\ln(2/\alpha)/n} < p < \bar{X} + \sqrt{3p\ln(2/\alpha)/n}\right] \geqslant 1-\alpha,$$

最后求解 p 的置信区间.

方法二: 根据 Bernoulli 分布的性质有 $E[X] = p$ 和 $\mathrm{Var}(X) = p(1-p)$, 设枢轴变量

$$W = \frac{n\bar{X} - np}{\sqrt{np(1-p)}}$$

根据中心极限定理可知 W 近似于标准正态分布 $\mathcal{N}(0,1)$. 于是有

$$\Pr\left[-\mu_{\alpha/2} < \frac{n\bar{X} - np}{\sqrt{np(1-p)}} < \mu_{\alpha/2}\right] \approx 1 - \alpha.$$

最后求解 p 的近似置信区间.