9 参数估计

设总体 X 的分布函数为 $F(X,\theta)$, 其中 θ 为未知参数(也可为向量). 现从总体中抽取一样本 $X_1,X_2,\cdots,X_n$, 如何依据样本估计参数 θ, 或 θ 的函数 $g(\theta)$, 此类问题称为参数估计问题. 内容包括: 点估计, 估计量标准, 区间估计.

9.1 点估计

9.1.1 矩估计法

总体 X 的 k 阶矩: $a_k = E[X^k]$　　　　样本 k 阶矩: $A_k = \frac{1}{n}\sum_{i=1}^n X_i^k$

用相应的样本矩去估计总体矩, 求解参数 θ 的方法称为 **矩估计法**. 矩估计法的理论基础是大数定理: $X_1,X_2,\cdots,X_n$ 为 i.i.d. 的随机变量, 若 $E(X)=\mu$, 则当 $n\to\infty$ 时有

$$\frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{P} \mu.$$

推论: 若 $E[X^k]=a_k$ 存在, 则当 $n\to\infty$ 时有

$$\frac{1}{n}\sum_{i=1}^n X_i^k \xrightarrow{P} a_k = E[X^k].$$

还可利用中心距进行估计:

总体 X 的 k 阶中心矩: $b_k = E[(X-E(X))^k]$　　　　样本 k 阶中心矩: $B_k = \frac{1}{n}\sum_{i=1}^n (X_i-\bar{X})^k$

矩估计方法: 总体 X 的分布函数 F 包含 m 个未知参数 $\theta_1,\theta_2,\cdots,\theta_m$,

1) 求总体 X 的 k 阶矩: $a_k = a_k(\theta_1,\theta_2,\cdots,\theta_m) = E[X^k],\ k\in[m]$ (a_k 一般为 $\theta_1,\theta_2,\cdots,\theta_m$ 的函数).
2) 计算样本的 k 阶矩: $A_k = \frac{1}{n}\sum_{i=1}^n X_i^k$.
3) 令样本矩等于总体矩 $A^k = a^k = a^k(\theta_1,\theta_2,\cdots,\theta_m)$ $(k=1,2,\cdots,m)$, 得到 m 个关于 $\theta_1,\theta_2,\cdots,\theta_m$ 的方程组.
4) 求解方程组得到估计量 $\widehat{\theta}_1,\widehat{\theta}_2,\cdots,\widehat{\theta}_n$.

例9.1. 设总体 X 的概率密度函数

$$f(x) = \begin{cases} (\alpha+1)x^\alpha & x\in(0,1) \\ 0 & \text{其它}, \end{cases}$$

设 $X_1,X_2,\cdots,X_n$ 是来自总体 X 的样本, 求参数 α 的矩估计.

解. 首先计算总体 X 的期望

$$E[X] = \int_{-\infty}^{+\infty} xf(x)dx = \int_0^1 x(\alpha+1)x^{\alpha+1}dx = \frac{\alpha+1}{\alpha+2}.$$

样本 X 的均值 $\bar{X}=\sum_{i=1}^{n}X_i/n$. 样本矩等于总体矩有

$$E(X)=\frac{\alpha+1}{\alpha+2}=\bar{X},$$

求解可得 $\alpha=(2\bar{X}-1)/(1-\bar{X})$. □

例9.2. 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

$$f(x)=\begin{cases}\frac{1}{\theta}e^{-\frac{x-\mu}{\theta}} & x\geq\mu\\ 0 & \text{其它},\end{cases}$$

其中 $\theta>0$, 求 μ 和 θ 的矩估计.

解. 设随机变量 $Y=X-\mu$, 则 Y 服从参数为 $1/\theta$ 的指数分布, 有

$$E(Y)=\theta \quad \text{和} \quad \mathrm{Var}(Y)=\theta^2.$$

由此可得 $E(X)=\mu+\theta$ 和 $\mathrm{Var}(X)=\theta^2$. 计算对应的样本矩

$$A_1=\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i, \qquad B_2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2.$$

求解方程组

$$\mu+\theta=A_1 \quad \text{和} \quad \theta^2=B_2,$$

解得 $\mu=\bar{X}-\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2/n}$ 和 $\theta=\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2/n}$. □

课堂习题:

- 求正态总体 $\mathcal{N}(\mu,\sigma^2)$ 的 μ,σ^2 的矩估计法.
- 求总体 $X\sim\mathcal{U}(a,b)$ 中 a,b 的矩估计法.

9.1.2 最大似然估计法

设 $X_1,X_2,\cdots,X_n$ 是来自总体 X 的一个样本. 若总体 X 为离散型随机变量, 其分布列为 $\Pr(X=x)=\Pr(X=x;\theta)$, 则样本 $X_1,X_2,\cdots,X_n$ 的分布列为

$$L(\theta)=L(x_1,x_2,\cdots,x_n;\theta)=\prod_{i=1}^{n}\Pr(x_i;\theta).$$

这里 $L(\theta)$ 表示样本 $X_1=x_1,X_2=x_2,\cdots,X_n=x_n$ 发生的概率.

若总体 X 为连续型随机变量, 其概率密度为 $f(x;\theta)$, 则 $X_1=x_1,X_2=x_2,\cdots,X_n=x_n$ 的联合概率密度为

$$L(\theta)=L(x_1,x_2,\cdots,x_n;\theta)=\prod_{i=1}^{n}f(x_i;\theta).$$

根据概率密度定义可知 $L(\theta)$ 越大, 样本 $(X_1, X_2, \cdots, X_n)$ 落入 $(x_1, x_2, \cdots, x_n)$ 的邻域内概率越大.

综合上述离散和连续两种随机变量, 统称 $L(\theta)$ 为样本 $X_1, X_2, \cdots, X_n$ 的似然函数, 可以发现 $L(\theta)$ 是 θ 的函数, 若

$$\hat{\theta} = \arg\max_{\theta} L(x_1, x_2, \cdots, x_m; \theta),$$

则称 $\hat{\theta}$ 为 θ 的最大似然估计量. 直觉而言: 最大似然估计量 $\hat{\theta}$ 是使观测值 $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$ 出现的概率最大.

求解最大似然估计量的步骤如下:

i) 计算对数似然函数 $\log(L(x_1, x_2, \cdots, x_m; \theta)$;

ii) 求对数似然函数中参数 θ 的一阶偏导, 令其等于零;

iii) 求解方程组得到最大似然估计量 $\hat{\theta}$.

例9.3. 设 $X_1, X_2, \cdots, X_n$ 是取自总体 $X \sim B(1, p)$ 的样本, 求参数 p 的最大似然估计.

解. 首先计算似然函数

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

从而得到对数似然函数

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

求一阶偏导并令其为零可得

$$\frac{\partial \ln L(p)}{\partial p} = \frac{1}{p}\sum_{i=1}^{n} X_i - \frac{1}{1-p}\left(n - \sum_{i=1}^{n} X_i\right) = 0.$$

由此求解 $p = \sum_{i=1}^{n} X_i / n = \bar{X}$. [验证矩估计法] □

下面讨论 **最大似然估计不可变性**

性质9.1. 设 $\mu(\theta)$ 为 θ 的函数, 且存在反函数 $\theta = \theta(\mu)$. 若 $\hat{\theta}$ 是 θ 的最大似然估计, 则 $\hat{\mu} = \mu(\hat{\theta})$ 是 μ 的最大似然估计.

例9.4. 设 $X_1, X_2, \cdots, X_n$ 为总体 $X \sim \mathcal{N}(\mu, \sigma^2)$ 的样本, 求 μ 和 $\sigma > 0$ 的最大似然估计.

解. 根据高斯分布知 X 的概率密度为 $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. 样本 $X_1, X_2, \cdots, X_n$ 的似然函数为

$$L(\mu, \sigma) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{2\sigma^2}\right).$$

其对数似然函数为 $\ln L(\mu, \sigma) = -n \ln(2\pi)^{1/2} - n \ln \sigma - \sum_{i=1}^{n} (X_i - \mu)^2 / 2\sigma^2$. 对参数 μ 求导计算可得

$$\frac{\partial \ln L(\mu, \sigma)}{\partial \mu} = \sum_{i=1}^{n} (X_i - \mu) = 0 \quad \Longrightarrow \quad \mu = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X},$$

对 σ 求导计算可得

$$\frac{\partial \ln L(\mu,\sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(X_i-\mu)^2 = 0 \quad \Longrightarrow \quad \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2}.$$

□

根据最大似然估计的不变性可知方差 σ^2 的最大似然估计为 $\sigma^2 = \sum_{i=1}^{n}(X_i-\bar{X})^2/n$. 下面进行验证最大似然估计的不变性: 设 $X_1, X_2, \cdots, X_n$ 为总体 $X \sim \mathcal{N}(\mu,\nu)$ 的样本, 求 μ 和 ν 的最大似然估计. 根据题意可知样本 $X_1, X_2, \cdots, X_n$ 的对数似然函数为

$$\ln L(\mu,\nu) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\nu - \sum_{i=1}^{n}\frac{(X_i-\mu)^2}{2\nu}.$$

对参数 μ 求偏导计算其最大似然估计 $\mu = \sum_{i=1}^{n} X_i/n = \bar{X}$, 对 ν 求偏导计算可得

$$\frac{\partial \ln L(\mu,\nu)}{\partial \nu} = -\frac{n}{2\nu} + \frac{1}{2\nu^2}\sum_{i=1}^{n}(X_i-\mu)^2 = 0 \quad \Longrightarrow \quad \nu = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2,$$

从而完成验证.

例9.5. 设总体 X 的密度函数为

$$f(x) = \begin{cases} (\alpha+1)x^{\alpha} & x \in (0,1) \\ 0 & \text{其它} \end{cases}$$

设 $X_1, X_2, \cdots, X_n$ 是总体 X 的样本, 求 α 的最大似然估计.

解. 首先得到似然函数为

$$L(\alpha) = (\alpha+1)^n \prod_{i=1}^{n} X_i^{\alpha} = (\alpha+1)^n (X_1 X_2 \cdots X_n)^{\alpha},$$

以及其对数似然函数 $\ln L(\alpha) = n\ln(\alpha+1) + \alpha\ln(X_1X_2\cdots X_n)$. 求导并令偏导为零有

$$\frac{\partial \ln L(\alpha)}{\partial \alpha} = \frac{n}{\alpha+1} + \ln(X_1X_2\cdots X_n) = 0,$$

求解得

$$\alpha = \frac{-n}{\sum_{i=1}^{n}\ln(X_i)} - 1 = \frac{-1}{\frac{1}{n}\sum_{i=1}^{n}\ln(X_i)} - 1.$$

□

对上例, 矩估计值为 $\alpha = (2\bar{X}-1)/(1-\bar{X})$, 因此矩估计值与最大似然估计值可能不同.

例9.6. 设 $X_1, X_2, \cdots, X_n$ 是总体 $X \sim \mathcal{U}(a,b)$ 的样本, 求 a 和 b 的最大似然估计.

解. 当 $x \in [a,b]$ 时, 总体 X 的概率密度为 $f(x) = 1/(b-a)$, 其它情况为零, 因此似然函数为

$$L(a,b) = \begin{cases} \frac{1}{(b-a)^n} & a \le X_1, X_2, \cdots, X_n \le b \\ 0 & \text{其它} \end{cases}$$

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

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

由此可得 $\hat{\theta}$ 是 θ 的有偏、但一致估计量. □

9.3 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

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

定义9.4 (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*

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

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 $\Pr[a < \mu < b] = 1 - \alpha$ 成立.

3) 根据 $a < \mu < b$ 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.

9.3.1 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

$$\Pr[W \geq \mu_{\alpha/2}] = 1-\alpha/2 \quad \text{和} \quad \Pr[W \leq -\mu_{\alpha/2}] = 1-\alpha/2.$$

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

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

课题练习. *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 ($\mu_{0.025} = 1.96$).*

9.3.2 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

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

整理可得

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

9.3.3 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

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

根据 χ^2 分布的不对称性, 采用概率对称的区间

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

根据枢轴变量 $W = (n-1)S^2/\sigma^2$ 可得

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

9.3.4 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

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

由此可得置信区间

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

9.3.5 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

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

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9.16. 设 $X_1, X_2, \cdots, X_n$ 是来自总体 $X \sim Ber(p)$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

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

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

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

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. □

9.3.6 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9.5 (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

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

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