数理统计笔记

强基数学 002 吴天阳

1 置信区间估计

定义 1.1 (置信区间). 令 X_1, \dots, X_n 是来自 $f(x;\theta)$ 的随机变量, $T_1 = T_1(X_1, \dots, X_n)$, $T_2 = T_2(X_1, \dots, X_n)$ 是两个关于样本的统计量,且 $T_1 \leqslant T_2$,满足 $P[T_1 < \tau(\theta) < T_2] = \gamma$,其中 γ 与 θ 无关,则称 $[T_1, T_2]$ 为随机区间,也是 θ 的 100γ 置信度的置信区间,其中 γ 称为置信系数, T_1 与 T_2 分别称为置信下限和置信上限.

注 1. 若 $\tau(\theta)$ 在随机区间中是严格单调的,则 $P(\tau(T_1) < \tau(\theta) < \tau(T_2)) = P(T_1 < \theta < T_2) = \gamma$. 这告诉我们可以将许多问题转化为求解 θ 置信区间.

定义 1.2 (枢轴量). 设 X_1, \dots, X_n 是来自 $f(x; \theta)$ 的随机变量,令 $Q = q(x_1, \dots, x_n; \theta)$,若 Q 的分布与 θ 无关,则称 Q 为枢轴量.

命题 1.3 (枢轴量方法). 给定置信系数 $0 < \gamma < 1$,则 $\exists q_1, q_2$ 使得 $P(q_1 < Q < q_2)$,则 可通过 $q_1 < q(x_1, \dots, x_n; \theta) < q_2$ 反解出 $T_1(x_1, \dots, x_n) < \tau(\theta) < T_2(x_1, \dots, x_n)$,于是 $P(T_1 < \tau(\theta) < T_2) = \gamma$. 于是得到 $\tau(\theta)$ 的置信区间为 (T_1, T_2) .

注 1. q_1, q_2 与 θ 无关.

注 2. 对于固定的 γ , q_1, q_2 的取值不唯一, 置信区间的期望长度 $\mathbb{E}(T_2 - T_1)$ 越小越好.

注 3. 若 Q 为 $\mu = 0$ 的对称分布, 如标准正态分布, t 分布, 取对称分布 $q_2 = -q_1 = z_{\frac{1+\gamma}{2}}$; 若为不对称分布, 通常取等尾分布 $q_2 = z_{\frac{1+\gamma}{2}}$, $q_1 = z_{\frac{1-\gamma}{2}}$, 其中 z_x 为 Q 的 x 分位数; 若为不对称分布, 还可通过转化为最优化极小化问题求解.

1.1 正态分布参数的置信区间

1.1.1 单参数估计

下面讨论**来自正态分布的样本**对其中未知参数的置信区间估计,分为以下三种:(括号内为枢轴量满足的分布)

- (1) μ 未知, σ^2 已知. (标准正态分布)
- (2) μ 已知, σ^2 未知. ($\chi^2(n)$ 分布)
- (3) μ 未知, σ^2 未知. (μ 对应 t(n) 分布, σ^2 对应 $\chi^2(n-1)$ 分布)

例 1.1. $X_1, \dots, X_n \stackrel{iid}{\sim} N(\theta, \sigma^2)$, 其中 σ^2 已知, 求 θ 的置信区间.

解答. 由于 $\bar{X} \sim N(\theta, \sigma^2/n)$,则 $\frac{\bar{X} - \theta}{\sigma/\sqrt{n}} \sim N(0, 1)$,设 $q_1, q_2 \in \mathbb{R}, q_1 < q_2$ 满足

$$\gamma = P\left(q_1 < \frac{\bar{X} - \theta}{\sigma/\sqrt{n}} < q_2\right) = P\left(\bar{X} - \frac{\sigma}{\sqrt{n}}q_2 < \theta < \bar{X} - \frac{\sigma}{\sqrt{n}}q_1\right)$$

由于 N(0,1) 为对称分布,取 $q_2=-q_1=z_{\frac{1+\gamma}{2}}$, θ 的置信系数为 γ 的区间为

$$\left(\bar{X} - \frac{\sigma}{\sqrt{n}} z_{\frac{1+\gamma}{2}} < \theta < \bar{X} + \frac{\sigma}{\sqrt{n}} z_{\frac{1+\gamma}{2}}\right)$$

例 1.2. $X_1, \dots, X_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$, 其中 μ 已知, 求 σ^2 的置信区间.

解答. 观察: 若直接使用上题的枢轴量,由于 $q_1 < 0$,通过计算发现,只能获得 σ^2 的单边置信下限,考虑构造值域非负的枢轴量,例如 Γ, χ^2 .

由于
$$\frac{X_i - \mu}{\sigma} \sim N(0, 1)$$
 则 $\sum_{i=1}^{N} \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(n)$. 设 $q_1, q_2 \in [0, \infty), \ q_1 < q_2$ 满足

$$\gamma = P\left(q_1 < \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{\sigma^2} < q_2\right) = P\left(\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{q_2} < \sigma^2 < \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{q_1}\right)$$

由于 χ^2 不是对称分布,考虑等尾分布: $q_1=\chi^2_{\frac{1-\gamma}{2}}(n),\ q_2=\chi^2_{\frac{1+\gamma}{2}}(n)$,于是 σ^2 的置信度为 γ 的区间为

$$\left(\frac{\sum_{i=1}^{N}(X_{i}-\mu)^{2}}{X_{\frac{1+\gamma}{2}}^{2}(n)}, \frac{\sum_{i=1}^{N}(X_{i}-\mu)^{2}}{X_{\frac{1-\gamma}{2}}^{2}(n)}\right)$$

例 1.3. $X_1, \dots, X_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$, 其中 μ, σ^2 均未知, 分别求 μ, σ^2 的置信区间.

解答. (i) 求解 μ 的置信区间.

观察:考虑使用**例**I.I的做法,由于分布处含有未知的 σ^2 ,会使得上下限中含有 σ^2 ,考虑利用随机变量相除将 σ^2 消去且满足某种分布,于是我们引入了 t 分布:若 $U \sim N(0,1), V \sim \chi^2(n)$,则 $\frac{U}{\sqrt{V/n}} \sim t(n)$.

由于
$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$
, $\sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma}\right)^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$, 则
$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \frac{\sigma\sqrt{n-1}}{\sqrt{n-1}S} = \sqrt{n} \frac{\bar{X} - \mu}{S} \sim t(n-1)$$

设 $q_1, q_2 \in \mathbb{R}, q_1 < q_2$ 满足

$$\gamma = P\left(q_1 < \sqrt{n}\frac{\bar{X} - \mu}{S} < q_2\right) = P\left(\bar{X} - \frac{S}{\sqrt{n}}q_2 < \mu < \bar{X} - \frac{S}{\sqrt{n}}q_1\right)$$

其中 $S^2 = \frac{1}{n-1} \sum_{i=1}^N (X_i - \bar{X})^2$. 由于 t 为对称分布,取 $q_2 = -q_1 = t_{\frac{1+\gamma}{2}}(n-1)$,则 μ 的置信度为 γ 的区间为

$$\left(\bar{X} - \frac{S}{\sqrt{t}} t_{\frac{1+\gamma}{2}}(n-1), \bar{X} + \frac{S}{\sqrt{n}} t_{\frac{1+\gamma}{2}}(n-1)\right)$$

(ii) 求 σ^2 的置信区间. 考虑使用例1.2的做法,将 μ 用 \bar{X} 代替则

$$\sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

类似例1.2可得 σ^2 的置信度为 γ 的区间为

$$\left(\frac{(n-1)S^2}{\chi^2_{\frac{1+\gamma}{2}}(n-1)}, \frac{(n-1)S^2}{\chi^2_{\frac{1-\gamma}{2}}(n-1)}\right)$$

1.1.2 双参数估计

我们只关心两个独立的正态分布的**均值之差**与**方差之比**的区间估计,令 X_1, \dots, X_n 是来自 $N(\mu_1, \sigma_1^2)$ 的随机样本, Y_1, \dots, Y_m 是来自 $N(\mu_2, \sigma_2^2)$ 的随机样本,考虑 $\mu_2 - \mu_1$ 的置信区间,主要分为以下三种情况(括号内为枢轴量满足的分布)

- (1) σ_1^2, σ_2^2 已知. (标准正态分布)
- (2) $\sigma_1^2 = \sigma_2^2 = \sigma^2$, σ 未知. (t(n+m-2) 分布)
- (3) σ_1^2, σ_2^2 未知. (**样本量大**利用大数定律构造标准正态分布,**样本量小**记结论构造 t(n+m-2) 分布)

不难发现上述构造的枢轴量都是对称分布,所以最终求出来的置信区间也都是对称的.

例 1.4. 设 $X_1, \dots, X_n \stackrel{iid}{\sim} N(\mu_1, \sigma_1^2), Y_1, \dots, Y_m \stackrel{iid}{\sim} N(\mu_2, \sigma_2^2), \ \sigma_1^2, \sigma_2^2$ 已知,求 $\mu_2 - \mu_1$ 的区间估计.

解答. 由于 $\bar{Y} - \bar{X} \sim N\left(\mu_2 - \mu_1, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)$,则 $\frac{\bar{Y} - \bar{X} - (\mu_2 - \mu_1)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim N(0, 1)$,取对称分布点 $q_1 = -q_2 = z_{\frac{1+\gamma}{2}}$,则 $\mu_2 - \mu_1$ 的置信度为 μ 的区间为

$$(Q_1, Q_2), \quad Q_1, Q_2 = \bar{Y} - \bar{X} \pm z_{\frac{1+\gamma}{2}} \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}$$

例 1.5. 设 $X_1, \dots, X_n \stackrel{iid}{\sim} N(\mu_1, \sigma_1^2), Y_1, \dots, Y_m \stackrel{iid}{\sim} N(\mu_2, \sigma_2^2), \ \sigma_1 = \sigma_2 = \sigma, \ \sigma 未知,求 <math>\mu_2 - \mu_1$ 的区间估计.

解答. 观察:由于分母中有未知参数 σ ,类似**例**I.2的方法,通过除以 χ^2 分布消去 σ 即可得到枢轴量.

得到枢轴重.
由于
$$U = \frac{\bar{Y} - \bar{X} - (\mu_2 - \mu_1)}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}} \sim N(0, 1), \frac{(n-1)S_1^2}{\sigma^2} \sim \chi^2(n-1), \frac{(m-1)S_2^2}{\sigma^2} \sim \chi^2(m-1),$$
于是 $V = \frac{(n-1)S_1^2 + (m-1)S_2^2}{\sigma^2} \sim \chi^2(n+m-2),$ 则

$$\frac{U}{\sqrt{V/(n+m-2)}} = \frac{\bar{Y} - \bar{X} - (\mu_2 - \mu_1)}{S_p \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t(n+m-2)$$

! 置信区间估计 4

其中 $S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}$. 取对称分布点 $q_1 = -q_2 = t_{\frac{1+\gamma}{2}}(n+m-2)$ 可得 $\mu_2 - \mu_1$ 置信度为 γ 的置信区间为

$$(Q_1, Q_2), \quad Q_1, Q_2 = \bar{Y} - \bar{X} \pm t_{\frac{1+\gamma}{2}}(n+m-2)S_p\sqrt{\frac{1}{n} + \frac{1}{m}}$$

注 1. 当 $\sigma_1 = \lambda \sigma_2$ 时,也可转化为上述问题,最终得到的区间估计为

$$(Q_1,Q_2), \quad Q_1,Q_2 = \bar{Y} - \bar{X} \pm t_{\frac{1+\gamma}{2}}(m+n-2)S\sqrt{\frac{1}{n} + \frac{\lambda}{m}}, \ \not \exists \dot + \dot F^2 = \frac{(n-1)S_1^2 + \frac{m-1}{\lambda}S_2^2}{m+n-2}$$

例 1.6. 设 $X_1, \dots, X_n \stackrel{iid}{\sim} N(\mu_1, \sigma_1^2), Y_1, \dots, Y_m \stackrel{iid}{\sim} N(\mu_2, \sigma_2^2), \ \sigma_1 \neq \sigma_2, \ \sigma_1, \sigma_2$ 均未知,求 $\mu_2 - \mu_1$ 的区间估计.

解答. (i) 当样本量 n, m 非常大(一般为 $\geqslant 30$)时,可以认为 $S_1^2 \stackrel{p}{\to} \sigma_1^2, S_2^2 \stackrel{p}{\to} \sigma_2^2$,则 $\frac{\bar{Y} - \bar{X} - (\mu_2 - \mu_1)}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{n}}}$ 近似服从 N(0,1),则 $\mu_2 - \mu_1$ 置信度为 γ 的区间为

$$(Q_1, Q_2), \quad Q_1, Q_2 = \bar{Y} - \bar{X} \pm z_{\frac{1+\gamma}{2}} \sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}$$

$$(ii) \, \diamondsuit \, S_k^2 = \frac{S_1^2}{n} + \frac{S_2^2}{m} \, \text{则} \, T = \frac{\bar{Y} - \bar{X} - (\mu_2 - \mu_1)}{S_k} \, \text{近似服从} \, t(r) \, , 其中 \, r = \frac{S_k^4}{\frac{S_1^4}{n^2(n-1)} + \frac{S_2^4}{m^2(m-1)}} \, ,$$

则 $\mu_2 - \mu_1$ 置信度为 γ 的区间为

$$(Q_1, Q_2), \quad Q_1, Q_2 = \bar{Y} - \bar{X} \pm t_{\frac{1+\gamma}{2}}(r)S_k$$

例 1.7 (联合正态分布). 令 $(X_1,Y_1),\cdots,(X_n,Y_n)\stackrel{iid}{\sim} N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix},\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$ 服从二维正态分布,其中 $\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$,求 $\mu_2 - \mu_1$ 的区间估计.

解答. 令 $D_i = Y_i - X_i$,则 $D_i \sim N(\mu_2 - \mu_1, \sigma_D^2)$, $\sigma_D^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$,转化为 μ , σ 都未知,求 μ 的单参数区间估计问题(转化为 t 分布),由例1.3(i) 可得, $\mu_2 - \mu_1$ 的置信度为 γ 的区间为

$$(Q_1, Q_2), \quad Q_1, Q_2 = \bar{D} \pm \frac{S_D}{\sqrt{n}} t_{\frac{1+\gamma}{2}} (n-1)$$

例 1.8 (方差之比). 设 $X_1, \dots, X_n \stackrel{iid}{\sim} N(\mu_1, \sigma_1^2), Y_1, \dots, Y_m \stackrel{iid}{\sim} N(\mu_2, \sigma_2^2), \ \sigma_1 \neq \sigma_2, \ \mu_1, \mu_2$ 均未知,求 σ_2^2/σ_1^2 的区间估计.

解答. 由于
$$T_1 = \frac{\sum_{i=1}^n (X_i - \mu_1)^2}{\sigma_1^2} \sim \chi^2(n), \ T_2 = \frac{\sum_{j=1}^m (Y_j - \mu_2)^2}{\sigma_2^2} \sim \chi^2(m), \ \mathcal{M}$$

$$F = \frac{T_1/n}{T_2/m} = \frac{\sigma_2^2}{\sigma_1^2} \frac{\sum_{i=1}^n (X_i - \mu_1)^2}{\sum_{i=1}^m (Y_i - \mu_2)^2} \sim F(n, m)$$

由于 F 分布不是对称的,取等尾分布 $q_1=F_{\frac{1-\gamma}{2}}(n,m), q_2=F_{\frac{1+\gamma}{2}}(n,m)$ 则 σ_2^2/σ_1^2 置信度为 γ 的区间为

$$\left(\frac{\sum_{j=1}^{m}(Y_j - \mu_2)^2/m}{\sum_{i=1}^{n}(X_i - \mu_1)^2/n}F_{\frac{1-\gamma}{2}}(n,m), \frac{\sum_{j=1}^{m}(Y_j - \mu_2)^2/m}{\sum_{i=1}^{n}(X_i - \mu_1)^2/n}F_{\frac{1+\gamma}{2}}(n,m)\right)$$

注 1. μ_1, μ_2 未知,则问题转化为类似 μ, σ 都未知求 σ 分布,见例1.3(ii),利用 \bar{X}, \bar{Y} 分别替换 μ_1, μ_2 ,将自由度 -1 即可, σ_2^2/σ_1^2 的置信度为 γ 的区间为

$$\left(\frac{S_2^2}{S_1^2}F_{\frac{1-\gamma}{2}}(n-1,m-1), \frac{S_2^2}{S_1^2}F_{\frac{1+\gamma}{2}}(n-1,m-1)\right)$$

1.2 一般的枢轴量方法

命题 1.4 (分布函数的枢轴量方法). 设 X_1, \cdots, X_n 是来自 $f(x;\theta)$ 的随机样本,将 $F(X;\theta)$ 视为随机变量,则 $F(X;\theta) \sim I(0,1)$,于是 $-\log F(X;\theta) \sim \Gamma(1,1)$,则 $-\sum_{i=1}^n \log F(X_i;\theta) \sim \Gamma(n,1)$ 为枢轴量.

注1. 上述方法在分布函数便于计算时可以使用.

注 2. 一般我们期望转化为关于 χ^2 分布的枢轴量 (便于查表), 于是

$$-2\sum_{i=1}^{n} \log F(X_i; \theta) \sim \Gamma(n, 1/2) = \chi^2(2n)$$

再利用 $\chi^2(2n)$ 等尾概率求出置信区间.

注 3. $F(x;\theta) \sim I(0,1)$ 是因为 $P(F(x;\theta) \leq y) = P(x \leq F^{-1}(y)) = F(F^{-1}(y)) = y$.

注 4. 令 $z=-\log Y$ 且 $Y\sim I(0,1)$,则 $Z\in (0,\infty)$ 且 $f_Z(z)=f_Y(\mathrm{e}^{-z})\mathrm{e}^{-z}=\mathrm{e}^{-z}\sim \Gamma(1,1)$.