基础数理统计

第七章 CDF 和统计泛函的估计

7.1 经验分布函数

经验分布函数

定义 1 (经验分布函数)

令 $X_1, X_2, \ldots, X_n \sim F$ 为 $i.i.d$ 样本，经验分布函数 $\widehat{F}_n$ 指在每一个数据点 X_i 上的概率密度为 $1/n$ 的 CDF，用公式表示为

$$\widehat{F}_n(x) = \frac{\sum_{i=1}^{n} I(X_i \leq x)}{n},$$

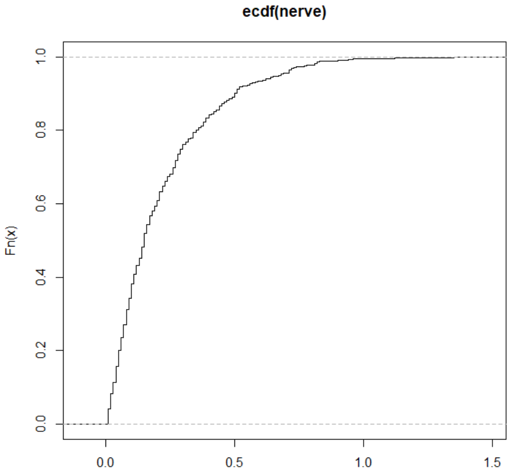
其中

$$I(X_i \leq x) = \begin{cases} 1 & X_i \leq x \\ 0 & X_i > x \end{cases}$$

例子

例 1

799 个神经数据（*library(ACSWR)→data(nerve)*）的经验分布函数如下：



逐点性质

定理 1

在任意固定点 x 有:

- 无偏性: $E(\widehat{F}_n(x)) = F(x)$.
- $V(\widehat{F}_n(x)) = \dfrac{F(x)(1-F(x))}{n}$.
- $MSE = \dfrac{F(x)(1-F(x))}{n} \to 0$.
- $\widehat{F}_n(x) \xrightarrow{P} F(x)$.
- 如果 $F(x) \in (0,1)$,

$$\sqrt{n}(\widehat{F}_n(x) - F(x)) \rightsquigarrow N(0, F(x)(1-F(x))).$$

全局性质

定理 2 (Gilvenko-Cantelli 定理)

$$\sup_x |\widehat{F}_n(x) - F(x)| \xrightarrow{P} 0.$$

定理 3 (Dvoretzky–Kiefer–Wolfowitz (DKW) 不等式)

令 $X_1, X_2, \ldots, X_n \sim F$，则对任意 $\epsilon > 0$，有

$$P(\sup_x |\widehat{F}_n(x) - F(x)| > \epsilon) \leq 2\exp(-2n\epsilon^2).$$

分布函数的置信区间

定义 2 (F 的非参数 $1-\alpha$ 置信带)

定义

$$L(x) = \max\{\widehat{F}_n(x) - \epsilon_n, 0\},\ \ R(x) = \min\{\widehat{F}_n(x) + \epsilon_n, 1\},$$

其中 $\epsilon_n = \sqrt{\log(2/\alpha)/2n}$, 我们有

$$P(L(x) \le F(x) \le R(x),\ \forall\ x) \ge 1-\alpha.$$

7.2 统计泛函

分布函数的参数

统计泛函 $T(F)$ 是分布函数 F 的任意函数，例如

- $T(F) = F(c) = \int I(x \leq c) dF(x)$
- $T(F) = F^{-1}(p)$
- $T(F) = \int x dF(x)$

嵌入式估计量

定义 3

$\theta = T(F)$ 的嵌入式估计量定义为

$$\widehat{\theta}_n = T(\widehat{F}_n).$$

定义 4

如果对函数 $r(x)$ 有 $T(F) = \int r(x)dF(x)$，则称 T 为线性泛函。

说明：T 满足 $T(aF + bG) = aT(F) + bT(G)$。

嵌入式估计量

定义 5

线性泛函 $T(F)=\int r(x)dF(x)$ 的嵌入式估计量为

$$T(\widehat{F}_n)=\int r(x)d\widehat{F}_n(x)=\frac{1}{n}\sum_{i=1}^{n}r(X_i).$$

说明：$T(F)$ 的近似 $1-\alpha$ 的置信区间为 $T(\widehat{F}_n)\pm z_{\alpha/2}\widehat{\mathrm{se}}$(基于正态的置信区间)，这里 $\widehat{\mathrm{se}}$ 是 $T(\widehat{F}_n)$ 的标准误差的估计。

常见例子

例 2

- 期望:

$$\mu = \int x dF(x), \ \ \widehat{\mu} = \int x d\widehat{F}_n(x) = \frac{1}{n}\sum_{k=1}^{n} X_k,$$

- 方差:

$$\begin{aligned}\sigma^2 = T(F) &= \int x^2 dF(x) - (\int x dF(x))^2;\\ \widehat{\sigma}^2 = T(F_n) &= \frac{1}{n}\sum_{k=1}^{n} X_k^2 - (\frac{1}{n}\sum_{k=1}^{n} X_k)^2.\end{aligned}$$

- 偏度（分布偏离对称的程度）:

$$\kappa = \frac{E(X-\mu)^3}{\sigma^3} = \frac{\int (x-\mu)^3 dF(x)}{[\int (x-\mu)^2 dF(x)]^{3/2}},$$

常见例子–续

例 3 (相关系数)

令 $Z=(X,Y)$, $\rho=T(F)=E(X-\mu_X)(Y-\mu_Y)/(\sigma_X\sigma_Y)$ 表示 X 和 Y 的相关系数，其中 $F(x,y)$ 是二元函数，可记为

$$T(F)=a(T_1(F),T_2(F),T_3(F),T_4(F),T_5(F)),$$

其中

$$\begin{aligned}
T_1(F) &= \int x dF(z),\ T_2(F)=\int y dF(z),\\
T_3(F) &= \int xy dF(z),\\
T_4(F) &= \int x^2 dF(z),\ \ T_5(F)=\int y^2 dF(z).
\end{aligned}$$

常见例子–续

则

$$\widehat{\rho} = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2}}$$

(样本相关系数)

常见例子–续

例 4

$X_{i1}, X_{i2}, \ldots, X_{in_i}$ 是来自分布 $F_i (i=1,2)$ 的简单随机样本，它们相互独立，记 F_1, F_2 的总体均值分别为 μ_1, μ_2，总体标准差分别为 σ_1, σ_2，求 $\mu_1 - \mu_2$ 的置信区间。

解： μ_i 的嵌入式估计为 $\bar{X}_i = n_i^{-1} \sum_{j=1}^{n_i} X_{ij}$，且 $\widehat{\mathrm{se}}(\bar{X}_i) = \widehat{\sigma}_i / \sqrt{n_i}$，这里 $\widehat{\sigma}_i$ 是 σ_i 的嵌入式估计 $\widehat{\sigma}_i = \sqrt{\sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 / n_i}$。故 $\mu_1 - \mu_2$ 的近似 $1-\alpha$ 置信区间为

$$\left((\bar{X}_1 - \bar{X}_2) - z_{\alpha/2} \sqrt{\widehat{\sigma}_1^2 / n_1 + \widehat{\sigma}_2^2 / n_2}, (\bar{X}_1 - \bar{X}_2) + z_{\alpha/2} \sqrt{\widehat{\sigma}_1^2 / n_1 + \widehat{\sigma}_2^2 / n_2} \right).$$

例 5 (分位数)

p 分位数为 $T(F) = F^{-1}(p)$，定义 $\widehat{F}_n^{-1}(p) = \inf\{x : \widehat{F}_n(x) \geq p\}$，称 $T(\widehat{F}_n) = \widehat{F}_n^{-1}(p)$ 为第 p 样本分位数。

作业：2，5，6，9