Chapter 7 - Estimating CDF and Statistical functional

Section 7.1 - The Empirical Distribution Function

$X_1, \dots, X_n \overset{iid}{\sim} F$

Definition 7.1: The empirical distribution function $\hat{F}_n$ is the CDF that puts mass $\frac{1}{n}$ at each data point X_i. Formally,

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

Where $I(X_i \le x) = \begin{cases} 1, & \text{if } X_i \geq x, \\ 0, & \text{if } X_i < x. \end{cases}$

Theorem 7.3: At any fixed value of x,

- $\hat{F}_n(x) \xrightarrow{P} F(x)$
- $E(\hat{F}_n(x)) = F(x)$
- $Var(\hat{F}_n(x)) = \frac{F(x)(1-F(x))}{n}$

$$MSE = \frac{F(x)(1-F(x))}{n} \longrightarrow 0$$

Implement it //

Theorem 7.4: (The Glivenko-Cantelli Theorem) Let $X_1, \dots, X_n \overset{iid}{\sim} F$. Then

$$\sup_x |\hat{F}_n(x) - F(x)| \xrightarrow{a.s.} 0.$$

Theorem 7.5: (The Dvoretzky-Kiefer-Wolfowitz (DKW) Inequality)
Let $X_1, \dots, X_n \overset{iid}{\sim} F$, then for any $\epsilon > 0$, we have

$$P\left(\sup_x |\hat{F}_n(x) - F(x)| > \epsilon\right) \leq 2e^{-2n\epsilon^2}$$

$$P\left(\sup_x |\hat{F}_n(x) - F(x)| > \epsilon\right) + P\left(\sup_x |\hat{F}_n(x) - F(x)| \le \epsilon\right) = 1$$

$$P\left(\sup_x |\hat{F}_n(x) - F(x)| \le \epsilon\right) = 1 - P\left(\sup_x |\hat{F}_n(x) - F(x)| > \epsilon\right)$$

$$\geq 1 - 2e^{-2n\epsilon^2}$$

$$\left[\sup_x |\hat{F}_n(x) - F(x)| \le \epsilon\right] \subset \left[|\hat{F}_n(x) - F(x)| \le \epsilon\right]$$

$$\underbrace{1 - 2e^{-2n\epsilon^2}}_{1-\alpha} \leq P\left(\sup_x |\hat{F}_n(x) - F(x)| \le \epsilon\right) \leq P\left(|\hat{F}_n(x) - F(x)| \le \epsilon\right)$$

$$1-\alpha = 1-2^{-2n\varepsilon^2} \Rightarrow 2^{-2n\varepsilon^2} = \alpha \Rightarrow -2n\varepsilon^2 = \log_2(\alpha)$$
$$\Rightarrow \varepsilon^2 = \frac{\log_2(\alpha)}{-2n} \Rightarrow \varepsilon_n = \sqrt{\frac{\log_2(\alpha)}{-2n}}$$

For α and fixed, we have:

$$-\varepsilon_n + \hat{F}_n(x) \leq F(x) \leq \hat{F}_n(x) + \varepsilon_n \text{ where } \varepsilon_n = \sqrt{\frac{\log_2(\alpha)}{-2\cdot n}}$$

$$\Rightarrow IC(F(x); 1-\alpha) = \left(\underbrace{\max\left(\sqrt{\frac{\log_2(\alpha)}{-2\cdot n}} + \hat{F}_n(x) \right)}_{L(x)}; \underbrace{\min\left(\sqrt{\frac{\log_2(\alpha)}{-2\cdot n}}; 1 \right)}_{U(x)} \right)$$

$$P\left(L(x) \leq F(x) \leq U(x)\right) \geq 1-\alpha.$$

Implement: $\begin{cases} * \text{ Sample (random)} \\ * \text{ real data} \end{cases}$

Section 7.2 - Statistical Functionals

- Statistical functional (ST): $T(F)$; $\mu = \int x \, dF(x)$; $\sigma^2 = \int (x-\mu)^2 \, dF(x)$.

Definition 7.7: The plug-in estimator of $\theta = T(F)$ is defined by

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

In another words, just plug-in $\hat{F}_n$ for the unknown F.

Definition 7.8: If $T(F) = \int r(x) dF(x)$ for some $r(x)$ then T is called a linear functional.

$$\left(T(aF + bG) = aT(F) + bT(G)\right)$$

Theorem 7.9: The plug-in estimator for linear functional $T(F) = \int r(x) dF(x)$ is:

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

Example 7.10: (The mean) $\mu = T(F) \approx \int x \, d\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} X_i = \hat{\mu}$.

Example 7.11: (The Variance) $\sigma^2 = T(F) = \int (x-\mu)^2 dF(x) \approx \int (x-\mu)^2 d\hat{F}_n(x)$

$$= \frac{1}{n}\sum_{i=1}^{n} (X_i - \mu)^2$$

Example 7.12: (The Skewness) -

$$K = \frac{\int (x-\mu)^3 \, dF(x)}{\left\{\int (x-\mu)^2 \, dF(x)\right\}^{3/2}} \approx \frac{\int (x-\mu)^3 \, d\hat{F}_n(x)}{\left\{\int (x-\mu)^2 \, d\hat{F}_n(x)\right\}^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\hat{\mu})}{\left\{\frac{1}{n}\sum_{i=1}^{n}(X_i-\hat{\mu})^2\right\}^{3/2}}$$

Example 7.14: (Quantiles). $T(F) = F^{-}(p) \Rightarrow T(F) \approx \hat{F}_n^{-}(p)$,

$$\hat{F}_n^{-}(p) = \inf\{x \mid \hat{F}_n(x) \geq p\}.$$