ECDF and Plug-In Estimator

Empirical Cumulative Distribution Function

Consider simple sample $X_1, \dots, X_n \sim F(x)$, where $F(x)$ is unknown and we'd like to come up with an estimate $\widehat{F}(x)$ of it. We would like that estimate to be **unbiased** and **consistent**.

Consider an estimate that is called the **empirical cumulative distribution function** (ECDF):

$$\widehat{F}(x) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{I}\mathrm{nd}\{X_n \leqslant x\}$$

where indicator function $\mathbb{I}\mathrm{nd}\{A\} = 1$ if event A is realized and 0 otherwise.

Is this estimate unbiased and consistent?

- $\mathbb{E}\left[\widehat{F}(x)\right] = F(x)$?
- $\widehat{F}(x) \xrightarrow{P} F(x)$?

ECDF properties

$$\widehat{F}(x) = \frac{1}{n}\sum_{k=1}^{n} \underbrace{\mathbb{I}\mathrm{nd}\{X_n \leqslant x\}}_{\xi_k}$$

What is the distribution of ξ_k?

By definition, $\xi_k \sim Be(F(x))$ and, consequently, $n\widehat{F}(x) \sim Bin(n, F(x))$. What is the expected value and variance of binomial random variable?

- $\mathbb{E}\left[\widehat{F}(x)\right] = \frac{1}{n} nF(x) = F(x)$ so ECDF is unbiased
- $\mathbb{V}\mathrm{ar}\left(\widehat{F}(x)\right) = \frac{1}{n^2} nF(x)(1-F(x)) \leqslant \frac{1}{4n} \to 0$ so ECDF is consistent

ECDF convergence

We can estimate the speed of convergence from CLT:

$$\sqrt{n}\frac{\frac{1}{n}\sum_{k=1}^{n}\xi_k - \mathbb{E}\left[\xi_k\right]}{\sqrt{\mathbb{V}\mathrm{ar}\left(\xi_k\right)}} \xrightarrow{d} \eta \sim \mathcal{N}(0,1)$$

$$\sqrt{n}\left(\frac{1}{n}\sum_{k=1}^{n}\xi_k - \mathbb{E}\left[\xi_k\right]\right) \xrightarrow{d} \sqrt{\mathbb{V}\mathrm{ar}\left(\xi_k\right)}\eta \sim \mathcal{N}(0, \mathbb{V}\mathrm{ar}\left(\xi_k\right))$$

$$\sqrt{n}\left(\widehat{F}(x) - F(x)\right) \xrightarrow{d} \mathcal{N}(0, F(x)(1-F(x)))$$

ECDF uniform convergence

Glivenko-Cantelli theorem:

$$\sup_x \left|F(x) - \widehat{F}(x)\right| \xrightarrow{a.s.} 0$$

But how fast? **Kolmogorov's theorem**: If $F(x)$ is continuous, then

$$D_n = \sqrt{n}\sup_x \left|F(x) - \widehat{F}(x)\right| \xrightarrow{d} \eta \sim K$$

$$\mathbb{P}(D_n \leqslant z) \to \sum_{k=-\infty}^{\infty}(-1)^k e^{-2k^2z^2}$$

$$\mathbb{P}(D_n > z) \leqslant 2e^{-2nz^2}$$

ECDF uniform covergence

If $F(x)$ is continuous, then $\sup_x \left|F(x) - \widehat{F}(x)\right|$ does not depend on $F(\cdot)$.

Proof

$$\sup_x \left|\widehat{F}(x) - F(x)\right| = \sup_x \left|\frac{1}{n}\sum_{k=1}^{n}\mathbb{I}\mathrm{nd}\{X_i \leqslant x\} - F(x)\right|$$

$$= \sup_x \left|\frac{1}{n}\sum_{k=1}^{n}\mathbb{I}\mathrm{nd}\{F(X_i) \leqslant F(x)\} - \underbrace{F(x)}_{z\in[0,1]}\right|$$

$$= \sup_z \left|\frac{1}{n}\sum_{k=1}^{n}\mathbb{I}\mathrm{nd}\{\underbrace{F(X_i)}_{?} \leqslant z\} - z\right|$$

$$= \sup_z \left|\frac{1}{n}\sum_{k=1}^{n}\mathbb{I}\mathrm{nd}\{U \leqslant z\} - z\right|$$

Estimating functionals

A statistical functional $T(F)$ is any functional of CDF F.

A **linear functional** $T(F)$ can be written as:

$$T(F) = \int r(x)dF(x)$$

A **plug-in** estimator $\widehat{T}$ of $T(F)$ is

$$\widehat{T} = T\left(\widehat{F}\right)$$

A plug-in estimator $\widehat{T}$ of **linear** $T(F)$ is

$$\widehat{T} = \int r(x)d\widehat{F}(x) = \frac{1}{n}\sum_{k=1}^{n} r(X_k)$$

Estimating mean

Mean functional is:

$$\mu(F) = \int x dF(x)$$

So $r(x) = x$, and the plug-in estimator is

$$\widehat{\mu} = \frac{1}{n}\sum_{k=1}^{n} X_k = \overline{X}$$

Estimating standard deviation

Standard deviation functional is

$$s(F) = \int (x-\mu)^2 dF(x)$$

So $r(x) = (x-\mu)^2$ and the plug-in estimator is

$$\widehat{\sigma_1^2} = \frac{1}{n}\sum_{k=1}^{n}(X_k - \mu)^2$$

Estimating standard deviation without the mean

$$\widehat{\sigma_1^2} = \frac{1}{n}\sum_{k=1}^{n}(X_k - \mu)^2$$

$$\widehat{\sigma_2^2} = \frac{1}{n-1}\sum_{k=1}^{n}\left(X_k - \overline{X}\right)^2 = s^2$$

Properties of normal distribution

Consider $X_1, \dots, X_n \sim \mathcal{N}(m, \sigma^2)$. Then, $\widehat{\mu}$, σ_1^2 and σ_2^2 are unbiased consistent estimates.

Proof 1

If $X_1, \dots, X_n \sim \mathcal{N}(m, \sigma^2)$, then $\frac{1}{n}\sum_{k=1}^{n} X_k \sim$?

$$\frac{1}{n}\sum_{k=1}^{n} X_k \sim \mathcal{N}(m, \frac{1}{n}\sigma^2)$$

So it is unbiased and consistent.

Proof 2

$$\widehat{\sigma_1^2} = \frac{1}{n}\sum_{k=1}^{n}(X_k - m)^2$$

$$\mathbb{E}\left[\widehat{\sigma_1^2}\right] = \mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n}(X_k - m)^2\right] = \frac{\sigma^2}{n}\mathbb{E}\left[\sum_{k=1}^{n}\left(\frac{X_k - m}{\sigma}\right)^2\right]$$

$$\eta = \sum_{k=1}^{n}\left(\frac{X_k - m}{\sigma}\right)^2 \sim ?$$

$$\eta = \sum_{k=1}^{n}\left(\frac{X_k - m}{\sigma}\right)^2 \sim \chi^2(n)$$

$$\mathbb{E}[\eta] = ?$$

$$\mathbb{V}\mathrm{ar}\left(\eta\right) = ?$$

Proof 2

$$\eta = \sum_{k=1}^{n}\left(\frac{X_k - m}{\sigma}\right)^2 \sim \chi^2(n)$$

$$\mathbb{E}[\eta] = n$$

$$\mathbb{V}\mathrm{ar}\left(\eta\right) = 2n$$

$$\mathbb{E}\left[\widehat{\sigma_1^2}\right] = \frac{\sigma^2}{n}\mathbb{E}\left[\eta\right] = \frac{\sigma^2}{n}n$$

$$\mathbb{V}\mathrm{ar}\left(\widehat{\sigma_1^2}\right) = \frac{\sigma^4}{n^2}\mathbb{V}\mathrm{ar}\left(\eta\right) = \frac{\sigma^4}{n^2}2n$$