STAT 403

4/1~~8~~11/25

Lecture 5

EDF is a sample mean

Sampling

MC : computing $E[f]$

HW 1 TB posted

L11, slides L11

Lecture Notes I – CDF and EDF

Marina Meila
mmp@stat.washington.edu

Department of Statistics
University of Washington

April 2025

CDF: Cumulative Distribution Function

Statistics and Motivation of Resampling Methods

EDF: Empirical Distribution Function

Properties of the EDF

Inverse of a CDF and sampling

Applications of EDF: testing if data come from known distribution

Lab 2

Reading: Lectures 0, 1, Lab 2

EDF: Empirical Distribution Function

Recall Given a value x_0, $F(x_0) = P(X_i \le x_0)$ for any $i = 1, \cdots, n$.

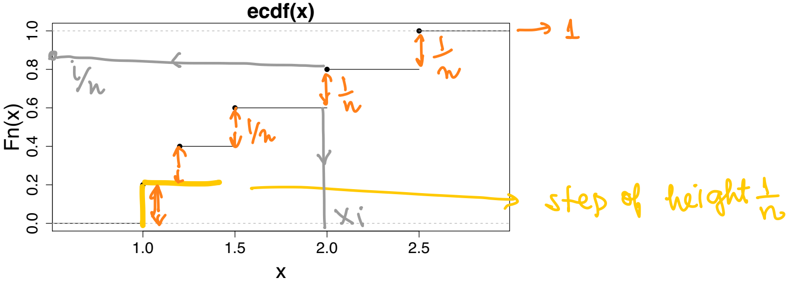
- Namely, $F(x_0)$ is the probability of the event $\{X_i \le x_0\}$.

Idea Use $F_n(x_0)$ as the estimator of $F(x_0)$.

$$\hat{F}_n(x_0) = \frac{\text{number of } X_i \le x_0}{\text{total number of observations}} = \frac{\sum_{i=1}^n I(X_i \le x_0)}{n} = \frac{1}{n}\sum_{i=1}^n I(X_i \le x_0)$$

- Hence $\hat{F}_n(x)$ (as a function) is estimator for $F(x)$ (as a function)
- We call $\hat{F}_n(x)$, **empirical distribution function (EDF)**.

Example EDF of 5 observations $1, 1.2, 1.5, 2, 2.5$



There are 5 jumps, each located at the position of an observation. Moreover, the height of each jump is the same: $\frac{1}{5}$.

CDF is an average

fix x

think of $F(x)$ CDF

$\hat{F}_n(x)$ EDF

- Properties of $Y_i = I(X_i \le x)$

$$Y_i = \begin{cases} 1, & \text{if } X_i \le x \\ 0, & \text{if } X_i > x \end{cases} .$$

for any X_i because $X_i \sim$ iid

- Hence, for some fixed x, $Y_i \sim \text{Ber}(F(x))$.
  Proof $p = P(Y_i = 1) = P(X_i \le x) = F(x)$.
- Then,

$$\mathbb{E}(I(X_i \le x)) = \mathbb{E}(Y_i) = F(x)$$
$$\text{Var}(I(X_i \le x)) = \text{Var}(Y_i) = F(x)(1 - F(x))$$

for a given x.

$I(\text{event}) = \begin{cases} 1 & \text{if event true} \\ 0 & \text{otherwise} \end{cases}$

EDF is an average

x fixed

$$\boxed{\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n I(X_i \le x) = \frac{1}{n}\sum_{i=1}^n Y_i.} \longrightarrow \text{estimator for } F(x)$$

Then

- $\mathbb{E}\left(\hat{F}_n(x)\right) = \mathbb{E}(I(X_1 \le x)) = F(x)$ **Bias**$= 0 = F(x) - E[\hat{F}_n(x)]$
- $\text{Var}\left(\hat{F}_n(x)\right) = \frac{\sum_{i=1}^n \text{Var}(Y_i)}{n^2} = \frac{F(x)(1-F(x))}{n}$. $\longrightarrow$ **variance converges to** 0 when $n \to \infty$.
- Hence, for a given x, $\hat{F}_n(x) \xrightarrow{P} F(x)$. i.e., $\hat{F}_n(x)$ is a consistent estimator of $F(x)$.

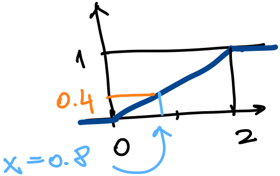
$$\text{Var}\,\hat{F}_n(x) = \frac{1}{n^2}\sum_{i=1}^n \text{Var}(Y_i) = \frac{\text{Var}(Y_i)}{n}$$

EDF is asymptotically normal

fixed x

Theorem

For a given x, $\sqrt{n}\left(\hat{F}_n(x) - F(x)\right) \xrightarrow{D} N(0, F(x)(1 - F(x)))$.



Example $n = 100$ samples from uniform distribution over $[0, 2]$

- $\mathbb{E}\left(\hat{F}_n(0.8)\right) = F(0.8) = P(x \leq 0.8) = \int_0^{0.8} \frac{1}{2} dx = 0.4$.
- $\mathrm{Var}\left(\hat{F}_n(0.8)\right) = \frac{F(0.8)(1-F(0.8))}{100} = \frac{0.4 \times 0.6}{100} = 2.4 \times 10^{-3}$.

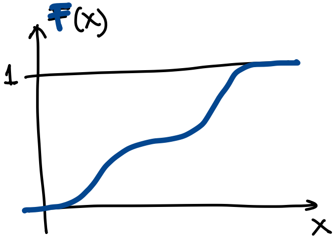
Theorem (Uniform convergence (proof not elementary))

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

Inverse of a CDF and sampling

- Let X be a continuous random variable with CDF $F(x)$.
- Let U be a uniform distribution over $[0, 1]$.

- We define a new random variable $W = F^{-1}(U)$



$$\begin{aligned} F_W(w) &= P(W \leq w) \\ &= P(F^{-1}(U) \leq w) \\ &= P(U \leq F(w)) \\ &= \int_0^{F(w)} 1 \; dx = F(w) - 0 = F(w). \end{aligned}$$

Algorithm for sampling from F

Input F (the CDF of P we want to sample from)

1. Sample $u \sim \mathrm{Uniform}[0, 1]$

Output $x = F^{-1}(u)$

Example Sampling from Exp(λ)

$$F(x) = 1 - e^{-\lambda x} \quad \text{when } x \geq 0.$$

$$F^{-1}(u) = \frac{-1}{\lambda} \log(1 - u).$$

So the random variable $W = F^{-1}(U) = \frac{-1}{\lambda} \log(1 - U)$ will be an Exp(λ) random variable.

called "SIMULATION"

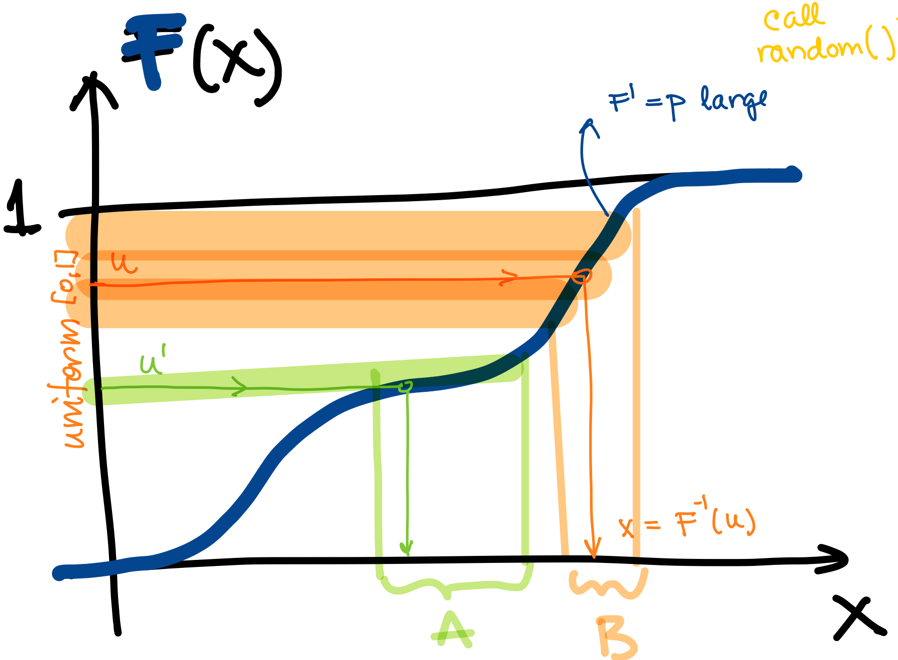
Sampling from F

call random() → 1. sample $u \sim unif[0,1]$

2. $x = F^{-1}(u)$

Output x

- will be repeated n times
- all sampling uses it (internally)
- for complicated models



Ex: Prove Alg is correct

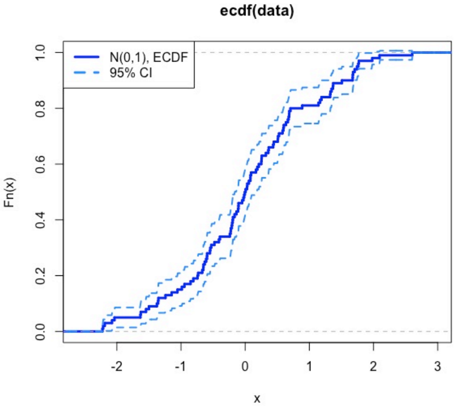
$p(x) = F'(x)$

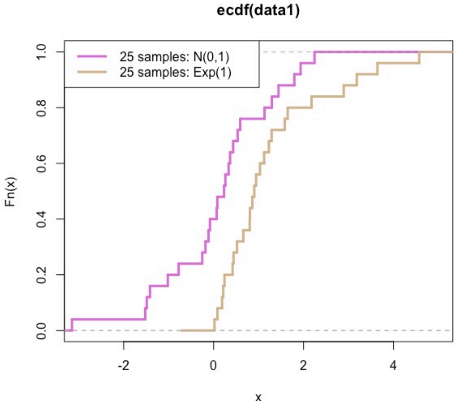
Statistical tests

Application #1 of $\hat{F}_n$

data

- Given sample $X_1, \cdots, X_n \sim \text{i.i.d.} P^{\text{unk}}$
- Question Is $P^{\text{unk}} =$ some P_0? (e.g. normal) $\equiv F_0$ Model — **goodness of fit test**
- Question Given also $X_1', \cdots X_n' \sim P'^{\text{unk}}$, is $P^{\text{unk}} = P'^{\text{unk}}$ true? — **two-sample test**


ecdf(data)
N(0,1), ECDF
95% CI
Fn(x)
x


ecdf(data1)
25 samples: N(0,1)
25 samples: Exp(1)
Fn(x)
x

Lecture Notes II – Monte Carlo simulation

Marina Meila
mmp@stat.washington.edu

Department of Statistics
University of Washington

April 2025

$X_{1:n} \sim F^{unk}$ data

"collect data"

$X_{1:N} \sim F$ known

simulation

MC: Calculating the expectations of a function by sampling

MC for computing an integral

MC for estimating a probability

MC for estimating a distribution

MC: Calculating the expectation of a function by sampling

Given a function $f(x)$ and a distribution F known (and its density $p(x) = F'(x)$).

- Let $\theta = \mathbb{E}[f(X)]$ be the parameter of interest

$$\text{wanted} \rightarrow \theta = \mathbb{E}[f(X)] \equiv \mu_f \equiv \int_{-\infty}^{\infty} f(x)p(x)dx.$$

Idea Estimate θ by sample average $\hat{\theta}_N$ $\leftarrow f(x) = 1$

1. Sample $X_{1:N} \sim F$
2. $\boxed{\hat{\theta}_N = \frac{1}{N}\sum_{i=1}^{N} f(X_i)}$

Note Here we don't collect data, we sample from a **known** F

Example $f(x) = x$, $F = \exp(\lambda = 0.9)$ $\theta = \mathbb{E}[X] = \mu$, $\hat{\theta} = \hat{\mu} = \bar{X}$ sample mean

Mean and variance of $\hat{\theta}_N$

- $\mathbb{E}[\hat{\theta}_N] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} f(X_i)\right] = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[f(X_i)\right] = \boxed{\mu_f}$ unbiased
- $\mathrm{Var}\,\hat{\theta}_N = \frac{1}{N}\mathrm{Var}\, f(X_1) = \frac{1}{N}\left(\boxed{\int f^2(x)p(x)dx} - \mu_f^2\right)$ $(*)$