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

Reading: Lectures 0, 1, Lab 2

The CDF

When $x \in S \subseteq (-\infty, \infty)$.

$$F(x) = P[X \leq x] = P(-\infty, x] = \int_{-\infty}^{x} p(u)du \quad (1)$$

Here are some properties of $F(x)$:

- (probability) $0 \leq F(x) \leq 1$.
- (monotonicity) $F(x) \leq F(y)$ for every $x \leq y$.
- (right-continuity) $\lim_{x \to y^+} F(x) = F(y)$, where $y^+ = \lim_{\epsilon > 0, \epsilon \to 0} y + \epsilon$.
- $\lim_{x \to -\infty} F(x) = F(-\infty) = 0$.
- $\lim_{x \to +\infty} F(x) = F(\infty) = 1$.
- $P(X = x) = F(x) - F(x^-)$, where $x^- = \lim_{\epsilon < 0, \epsilon \to 0} x + \epsilon$.

Examples of CDF's

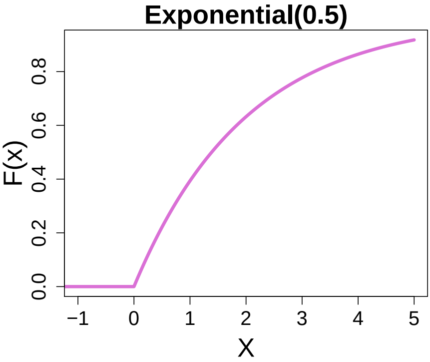
Example Uniform random variable over $[0,1]$

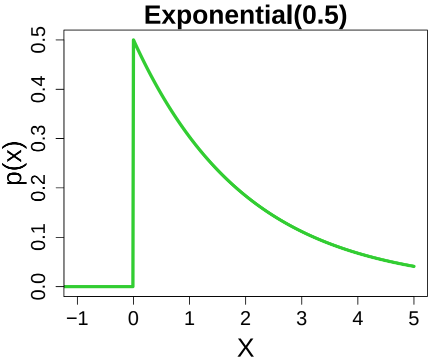
$$F(x) = \int_0^x 1 \; du = x$$

when $x \in [0,1]$ and $F(x) = 0$ if $x < 0$ and $F(x) = 1$ if $x > 1$.
Example Exponential random variable with parameter λ

$$F(x) = \int_0^x \lambda e^{-\lambda u} du = 1 - e^{-\lambda x}$$

when $x \geq 0$ and $F(x) = 0$ if $x < 0$. The following provides the CDF (left) and PDF (right) of an exponential random variable with $\lambda = 0.5$:





Statistics

Given a sample $X_1, \cdots, X_n$ (not necessarily an IID sample), a **statistic** $S_n = S(X_1, \cdots, X_n)$ is a function of the sample.

- Sample mean (average): $S(X_1, \cdots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i$.
- Sample maximum: $S(X_1, \cdots, X_n) = \max\{X_1, \cdots, X_n\}$.
- Sample range: $S(X_1, \cdots, X_n) = \max\{X_1, \cdots, X_n\} - \min\{X_1, \cdots, X_n\}$.
- Sample variance: $S(X_1, \cdots, X_n) = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2, \quad \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$.

...and more statistics

- Number of observations above a threshold t: $S(X_1, \cdots, X_n) = \sum_{i=1}^{n} I(X_i > t)$.

- Rank of the first observation (X_1): $S(X_1, \cdots, X_n) = 1 + \sum_{i=2}^{n} I(X_i > X_1)$.
  - If X_1 is the largest number, then $S(X_1, \cdots, X_n) = 1$;
  - if X_1 is the smallest number, then $S(X_1, \cdots, X_n) = n$.

- Sample second moment: $S(X_1, \cdots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i^2$.
  (The sample second moment is a consistent estimator of $E(X_i^2)$.)

Statistics S_n are determined by CDF

- a statistic $S_n = S(X_1, \cdots, X_n)$, it is a random variable
- because S_n is a function of the input data points $X_1, \cdots, X_n$, the distribution of S_n is completely determined by the joint CDF of $X_1, \cdots, X_n$.
- $F_{S_n}(x)$ is determined by $F_{X_1,\cdots,X_n}(x_1, \cdots, x_n)$
- and $F_{X_1,\cdots,X_n}(x_1, \cdots, x_n)$ is determined by $F(x)$ and n

- Therefore, F is sufficient to study the randomness of any statistic S_n.

Example Sample average for Normal(μ, σ^2)

- Assume $X_1, \cdots, X_n \sim N(0,1)$, let $S_n = \frac{1}{n}\sum_{i=1}^{n} X_i$
- Then $S_n \sim N(0, 1/n)$.

- if $X_{1:n} \sim N(1,4)$, then $S_n \sim N(1, 4/n)$.

Problem In practice the CDF F is unknown. How to estimate F from sample $X_1, \cdots, X_n$?

EDF: Empirical Distribution Function

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

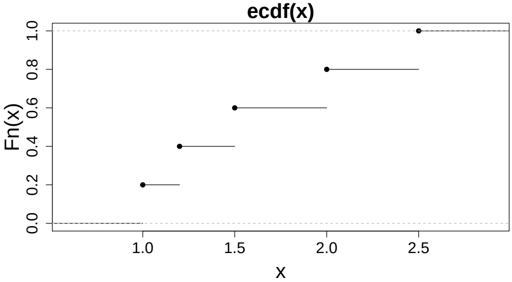
- Namely, $F(x_0)$ is the probability of the event $\{X_i \leq x_0\}$.

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

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

- 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)**.

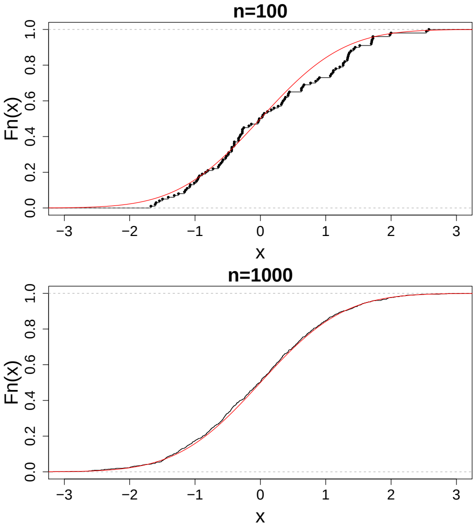
Example EDF of 5 observations $1, 1.2, 1.5, 2, 2.5$



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

EDF for larger n

Example EDF versus CDF for $n = 100$, 1000 random points from $N(0, 1)$



red=true CDF

CDF is an average

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

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

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

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

  for a given x.

EDF is an average

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

Then

- $\mathbb{E}\left(\hat{F}_n(x)\right) = \mathbb{E}(I(X_1 \le x)) = F(x)$ **Bias**$= 0$
- $\mathrm{Var}\left(\hat{F}_n(x)\right) = \frac{\sum_{i=1}^n \mathrm{Var}(Y_i)}{n^2} = \frac{F(x)(1-F(x))}{n}$. **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 $F(x)$.

EDF is asymptotically normal

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 $[0, 2]$

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

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.

Uniformization

Example Uniformization

- Let X be a r.v. with CDF F
- Let $V = F(X)$ another r.v.

- The CDF of V

$$F_V(v) = P(V \leq v) = P(F(X) \leq v) = P(X \leq F^{-1}(v)) = F(F^{-1}(v)) = v \text{ for any } v \in [0, 1].$$

- Therefore, V is actually a uniform random variable over $[0, 1]$!

Statistical tests

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

Does sample come from known distribution P_0?

- Let F_0 be the CDF of P_0

1. **KS test (Kolmogorov–Smirnov test)**[1],

$$T_{KS} = \sup |\hat{F}_n(x) - F_0(x)|.$$

2. **Cramér–von Mises test**[2],

$$T_{CM} = \int \left(\hat{F}_n(x) - F_0(x)\right)^2 dF_0(x).$$

3. **Anderson-Darling test**[3] and the test statistic is

$$T_{AD} = n \int \frac{\left(\hat{F}_n(x) - F_0(x)\right)^2}{F_0(x)(1 - F_0(x))} dF_0(x).$$

- Reject the null hypothesis ($H_0 : X_1, \cdots, X_n \sim F_0$) when the test statistic is greater than some threshold depending on the significance level α.

[1] https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

[2] https://en.wikipedia.org/wiki/Cram%C3%A9r%E2%80%93von_Mises_criterion

[3] https://en.wikipedia.org/wiki/Anderson%E2%80%93Darling_test