STAT 425: Introduction to Nonparametric Statistics **Winter 2026**

Lecture 2: CDF and EDF

Instructor: Yen-Chi Chen

2.1 CDF: Cumulative Distribution Function

For a random variable X, its CDF $F(x)$ contains all the probability structures of X. 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 $x^- = \lim_{\epsilon < 0, \epsilon \to 0} x + \epsilon$.

Example. For a uniform random variable over $[0, 1]$, its CDF

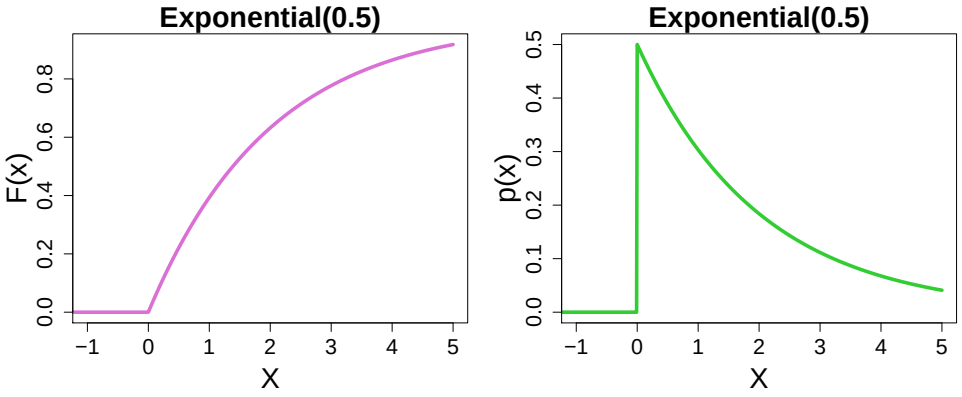
$$F(x) = \int_0^x 1 \, du = x$$

when $x \in [0, 1]$ and $F(x) = 0$ if $x < 0$ and $F(x) = 1$ if $x > 1$.

Example. For an exponential random variable with parameter λ, its CDF

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

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 $\lambda = 0.5$:



Why do we care about the CDF? Why not just use the PDF or PMF? CDF is the actual quantity that defines the probability structure of a random variable. The PDF exists only when the RV is continuous and the PMF exists when the RV is discrete. But CDF always exists – it is a unified quantity regardless of the RV being continuous or discrete. Moreover, there are cases where the neither PDF nor PMF exist.

Example. X is a random variable such that with a probability of 0.5, it is from a uniform distribution over the interval $[0, 1]$ and with a probability of 0.5 it takes a fixed value 0.5. Such X does not have a PDF nor a PMF but its CDF still exists (think about what does its CDF look like).

In the two-sample test, the P_X and P_Y in the hypothesis $H_0 : P_X = P_Y$ are actually the CDF of the sample of X and the CDF of the sample of Y. Essentially, the two-sample test is to determine if the two CDFs are the same or not.

2.2 EDF: Empirical Distribution Function

Let first look at the function $F(x)$ more closely. Given a value x_0,

$$F(x_0) = P(X_i \leq x_0)$$

for every $i = 1, \cdots, n$. Namely, $F(x_0)$ is the probability of the event $\{X_i \leq x_0\}$.

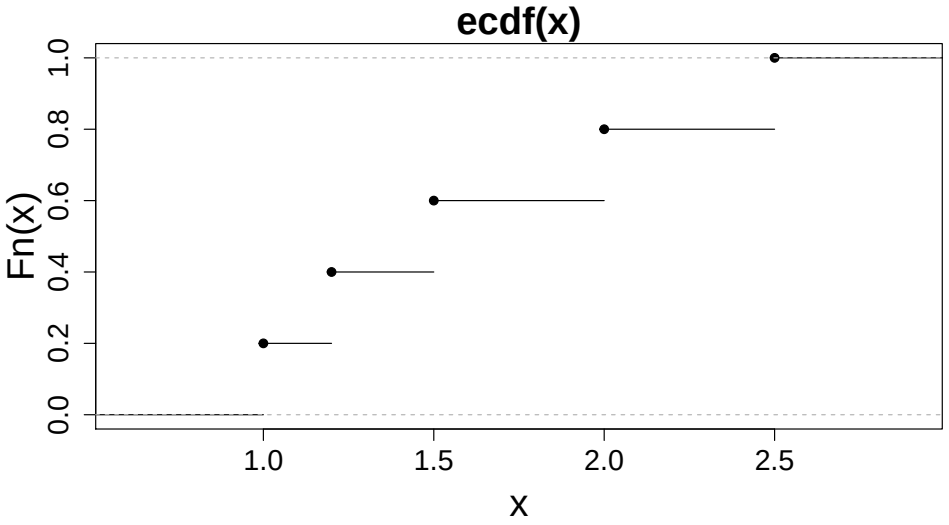
A natural estimator of a probability of an event is *the ratio of such an event in our sample*. Thus, we use

$$\widehat{F}_n(x_0) = \frac{\text{number of } X_i \leq x_0}{\text{total number of observations}} = \frac{\sum_{i=1}^n I(X_i \leq x_0)}{n} = \frac{1}{n}\sum_{i=1}^n I(X_i \leq x_0) \tag{2.1}$$

as the estimator of $F(x_0)$.

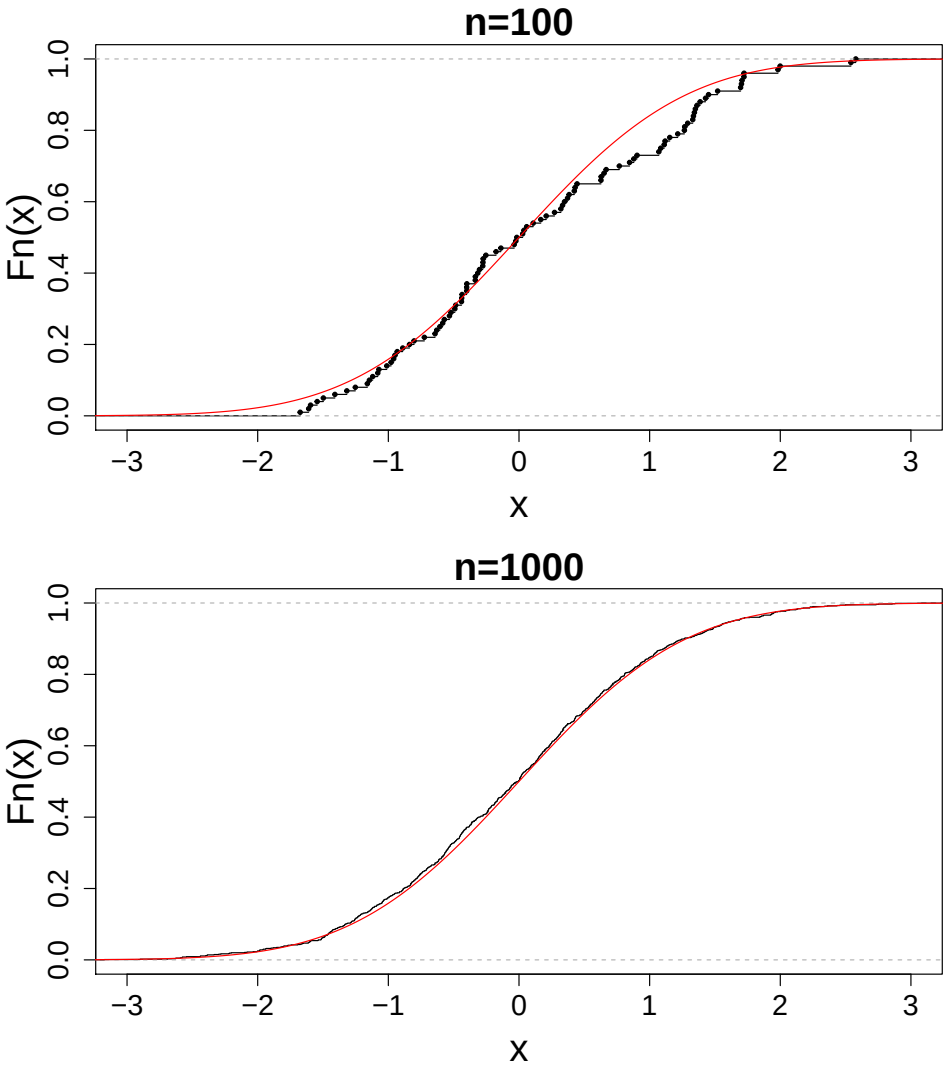
For every x_0, we can use such a quantity as an estimator, so the estimator of the CDF, $F(x)$, is $\widehat{F}_n(x)$. This estimator, $\widehat{F}_n(x)$, is called the *empirical distribution function (EDF)*.

Example. Here is an example of the EDF of 5 observations of $1, 1.2, 1.5, 2, 2.5$:



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

Example. While the previous example might not be look like an idealized CDF, the following provides a case of EDF versus CDF where we generate $n = 100, 1000$ random points from the standard normal $N(0, 1)$:



The red curve indicates the true CDF of the standard normal. Here you can see that when the sample size is large, the EDF is pretty close to the true CDF.

2.2.1 Property of EDF

Because EDF is the average of $I(X_i \leq x)$, we now study the property of $I(X_i \leq x)$ first. For simplicity, let $Y_i = I(X_i \leq x)$. What is the random variable Y_i?

Here is the breakdown of Y_i:

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

So Y_i only takes value 0 and 1–so it is actually a Bernoulli random variable! We know that a Bernoulli random variable has a parameter p that determines the probability of outputing 1. What is the parameter

p for Y_i?

$$p = P(Y_i = 1) = P(X_i \le x) = F(x).$$

Therefore, for a given x,

$$Y_i \sim \mathsf{Ber}(F(x)).$$

This implies

$$\begin{aligned}\mathbb{E}(I(X_i \le x)) &= \mathbb{E}(Y_i) = F(x)\\ \mathsf{Var}(I(X_i \le x)) &= \mathsf{Var}(Y_i) = F(x)(1-F(x))\end{aligned}$$

for a given x.

Now what about $\widehat{F}_n(x)$? Recall that $\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n I(X_i \le x) = \frac{1}{n}\sum_{i=1}^n Y_i$. Then

$$\begin{aligned}\mathbb{E}\left(\widehat{F}_n(x)\right) &= \mathbb{E}(I(X_1 \le x)) = F(x)\\ \mathsf{Var}\left(\widehat{F}_n(x)\right) &= \frac{\sum_{i=1}^n \mathsf{Var}(Y_i)}{n^2} = \frac{F(x)(1-F(x))}{n}.\end{aligned}$$

What does this tell us about using $\widehat{F}_n(x)$ as an estimator of $F(x)$?

First, at each x, $\widehat{F}_n(x)$ is an *unbiased estimator* of $F(x)$:

$$\textbf{bias}\left(\widehat{F}_n(x)\right) = \mathbb{E}\left(\widehat{F}_n(x)\right) - F(x) = 0.$$

Second, the *variance converges to* 0 when $n \to \infty$. By Lemma 0.3, this implies that for a given x,

$$\widehat{F}_n(x) \xrightarrow{P} F(x).$$

i.e., $\widehat{F}_n(x)$ is a *consistent* estimator of $F(x)$.

In addition to the above properties, the EDF also have the following interesting feature: for a given x,

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

Namely, $\widehat{F}_n(x)$ is asymptotically normally distributed around $F(x)$ with variance $F(x)(1-F(x))$.

Example. Assume $X_1, \cdots, X_{100} \sim F$, where F is a uniform distribution over $[0, 2]$. Questions:

- What will be the expectation of $\widehat{F}_n(0.8)$?

$$\longrightarrow \mathbb{E}\left(\widehat{F}_n(0.8)\right) = F(0.8) = P(x \le 0.8) = \int_0^{0.8} \frac{1}{2} dx = 0.4.$$

- What will be the variance of $\widehat{F}_n(0.8)$?

$$\longrightarrow \mathsf{Var}\left(\widehat{F}_n(0.8)\right) = \frac{F(0.8)(1-F(0.8))}{100} = \frac{0.4 \times 0.6}{100} = 2.4 \times 10^{-3}.$$

Remark. The above analysis shows that for a given x,

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

This is related to the pointwise convergence in mathematical analysis (you may have learned this in STAT 300). We can extend this result to a uniform sense:

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

However, deriving such a uniform convergence requires more involved probability tools so we will not cover it here. But an important fact is that such a uniform convergence in probability can be established under some conditions.

Question to think: Think about how to construct a 95% confidence interval of $F(x)$ for a given x.

2.3 Inverse of a CDF

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

$$W = F^{-1}(U),$$

where F^{-1} is the inverse of the CDF. What will this random variable W be?

To understand W, we examine its CDF F_W:

$$F_W(w) = P(W \le w) = P(F^{-1}(U) \le w) = P(U \le F(w)) = \int_0^{F(w)} 1 \, dx = F(w) - 0 = F(w).$$

Thus, $F_W(w) = F(w)$ for every w, which implies that the random variable W has *the same* CDF as the random variable X! So this leads a simple way to *generate* a random variable from F as long as we know F^{-1}. We first generate a random variable U from a uniform distribution over $[0, 1]$. And then we feed the generated value into the function F^{-1}. The resulting random number, $F^{-1}(U)$, has a CDF being F.

This interesting fact also leads to the following result. Consider a random variable $V = F(X)$, where F is the CDF of X. Then the CDF of V

$$F_V(v) = P(V \le v) = P(F(X) \le v) = P(X \le F^{-1}(v)) = F(F^{-1}(v)) = v$$

for any $v \in [0, 1]$. Therefore, V is actually a uniform random variable over $[0, 1]$.

Example. Here is a method of generating a random variable X from $\mathsf{Exp}(\lambda)$ from a uniform random variable over $[0, 1]$. We have already calculated that for an $\mathsf{Exp}(\lambda)$, the CDF

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

when $x \ge 0$. Thus, $F^{-1}(u)$ will be

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

So the random variable

$$W = F^{-1}(U) = \frac{-1}{\lambda} \log(1 - U)$$

will be an $\mathsf{Exp}(\lambda)$ random variable.

2.4 Applications in Testings

Going back to the two-sample test problem, because $H_0 : P_X = P_Y$ is essentially testing $H_0 : F_X = F_Y$, we can use EDF to estimate both F_X and F_Y and then carry out the test. Before we describe the two-sample test, we first study a simpler case – one-sample test problem – testing against a known distribution. This problem is also known as the goodness-of-fit test. Later we will discuss how to generalize to two-sample test.

2.4.1 Goodness-of-fit Test

Sometimes we have only one sample and the goal is to determine if this sample is from a known distribution. Say we want to test if a collection of values are from a normal distribution. Then we can use the one-sample test or goodness-of-fit test approach.

Assume that we want to test if $X_1, \cdots, X_n$ are from an known distribution F_0 (goodness-of-fit test), i.e.,

$$H_0 : X_1, \cdots, X_n \sim F_0.$$

There are three approaches for testing if they are from F_0:

- The first one is called *KS test (Kolmogorov–Smirnov test)*[1], where the test statistic is the KS-statistic

  $$T_{KS} = \sqrt{n} \sup_x |\widehat{F}_n(x) - F_0(x)|.$$

  The idea is: when H_0 is correct, $\widehat{F}_n$ should converge to F_0 for every point[2]. Thus, we just choose the largest deviation between the EDF and the CDF suggested by H_0. Note that there is a known limiting distribution of the test statistic T_{KS} called the Kolmogorov distribution so the test can be carried out very quickly using many statistical softwares.

- The second approach is the *Cramér–von Mises test*[3], which uses the Cramér–von Mises statistic as the test statistic

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

  If H_0 is correct, this quantity should be around 0 and if we do not scale it by n, it should converge to 0. Such a test statistic also has a known limiting distribution so the actual test is to compare to the limiting distribution. Note that the integrated squared difference between two functions is also known as the $L_2(P_0)$ distance for functions.

- The third approach is the *Anderson-Darling test*[4] and the test statistic is

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

  In a sense, this test statistic is just a weighted version of the statistic used in *Cramér–von Mises test*. The weight comes from the fact that under H_0, the variance of $\sqrt{n}\widehat{F}_n(x)$ is $F_0(x)(1 - F_0(x))$. Thus, the weight is to balance out the variance at each x before integrating them.

The p-values will be computed based on the limiting distribution and the observed test statistic.

[1] https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test
[2] There are assumptions for such a uniform convergence but it is generally true so we omit the details.
[3] https://en.wikipedia.org/wiki/Cram%C3%A9r%E2%80%93von_Mises_criterion
[4] https://en.wikipedia.org/wiki/Anderson%E2%80%93Darling_test

2.4.2 Two-sample Test

Going back to the two-sample test problem. Let $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_m$ be the two samples we have. Let $\widehat{F}_X$ and $\widehat{F}_Y$ denote the EDFs of the first and the second samples, respectively. In addition, we define the EDF by treating both samples as the same sample:

$$\widehat{H}(t) = \frac{n\widehat{F}_X(t) + m\widehat{F}_Y(t)}{n+m} = \frac{1}{n+m}\left(\sum_{i=1}^{n} I(X_i \le t) + \sum_{j=1}^{m} I(Y_j \le t)\right).$$

- The *KS test (Kolmogorov–Smirnov test)* will be using

$$T_{KS} = \sqrt{\frac{nm}{n+m}} \sup_x |\widehat{F}_X(t) - \widehat{F}_Y(t)|.$$

- The *Cramér–von Mises test* is based on

$$T_{CM} = \frac{nm}{n+m} \int \left(\widehat{F}_X(t) - \widehat{F}_Y(t)\right)^2 d\widehat{H}(t).$$

- The *Anderson-Darling test* is using

$$T_{AD} = \frac{nm}{n+m} \int \frac{\left(\widehat{F}_X(t) - \widehat{F}_Y(t)\right)^2}{\widehat{H}(t)(1-\widehat{H}(t))} d\widehat{H}(t).$$