STAT/Q SCI 403: Introduction to Resampling Methods **Spring 2017**

Lecture 1: CDF and EDF

Instructor: Yen-Chi Chen

1.1 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

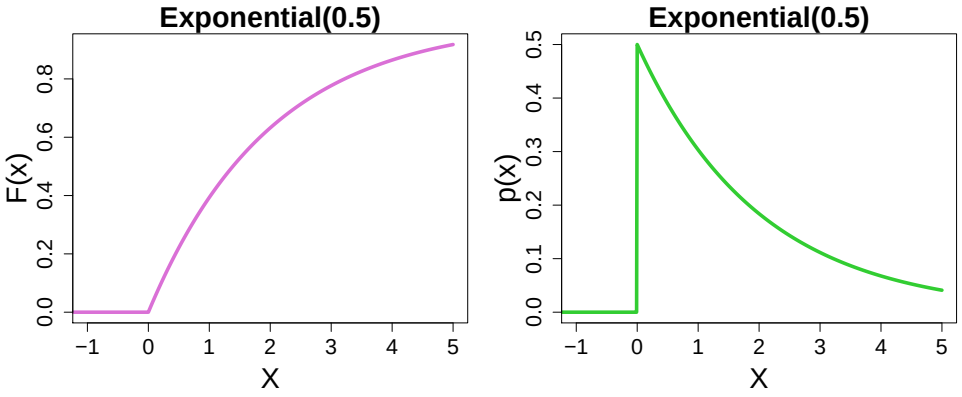
$$F(x) = \int_0^x 1 \, du = x$$

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 $\lambda = 0.5$:



1.2 Statistics and Motivation of Resampling Methods

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.

Here are some common examples of a statistic:

- 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:
$$S(X_1, \cdots, X_n) = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2, \quad \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Here are some useful statistics but might not be so common as the previous few examples:

- 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 $E(X_i^2)$.

Now we assume that our sample $X_1, \cdots, X_n$ is generated from a sampling distribution. Then the distribution of these n numbers is determined by the joint CDF $F_{X_1,\cdots,X_n}(x_1, \cdots, x_n)$. In the IID case (or sometimes we call it a random sample), the joint CDF is the product of the individual CDF's (and they are all the same because of being *identically distributed*). Thus, in the IID case, the individual CDF $F(x) = F_{X_1}(x)$ and the sample size n determines the entire joint CDF.

For a statistic $S_n = S(X_1, \cdots, X_n)$, it is a random variable when the sample is random. 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 $X_1, \cdots, X_n$. Let $F_{S_n}(x)$ be the CDF of S_n. Then $F_{S_n}(x)$ is determined by $F_{X_1,\cdots,X_n}(x_1, \cdots, x_n)$, which under the IID case, is determined by $F(x)$ and n (sample size).

Thus, when $X_1, \cdots, X_n \sim F$,

$$(F(x), n) \overset{determine}{\longrightarrow} F_{X_1,\cdots,X_n}(x_1, \cdots, x_n) \overset{determine}{\longrightarrow} F_{S_n}(x). \tag{1.1}$$

Mathematically speaking, there is a map $\Psi : \mathcal{F} \times \mathbb{N} \mapsto \mathcal{F}$ such that

$$F_{S_n} = \Psi(F, n), \tag{1.2}$$

where $\mathcal{F}$ is a collection of all possible CDF's.

Example. Assume $X_1, \cdots, X_n \sim N(0, 1)$. Let $S_n = \frac{1}{n}\sum_{i=1}^n X_i$ be the sample average. Then the CDF of S_n is the CDF of $N(0, 1/n)$ by the property of a normal distribution. In this case, F is the CDF of $N(0, 1)$. Now if we change the sampling distribution from $N(0, 1)$ to $N(1, 4)$, then the sample average S_n has a CDF of $N(1, 4/n)$. Here you see that the CDF of the sample average, a statistic, changes when the sampling distribution F changes (and the CDF of S_n is clearly dependent on the sample size n). This is what equations (1.1) and (1.2) refer to.

Therefore, a key conclusion is:

> Given F and the sample size n, the distribution of any statistic from the random sample $X_1, \cdots, X_n$ is determined.

Even if we cannot analytically write down the function $F_{S_n}(x)$, as long as we can sample from F, we can generate many sets of size-n random samples and compute S_n of each random sample and find out the distribution of F_{S_n}.

Here you see that the CDF F is very important in analyzing the distribution of any statistic. However, in practice the CDF F is unknown to us; all we have is the random sample $X_1, \cdots, X_n$. So here comes the question:

> Given a random sample $X_1, \cdots, X_n$, how can we estimate F?

1.3 EDF: Empirical Distribution Function

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

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

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

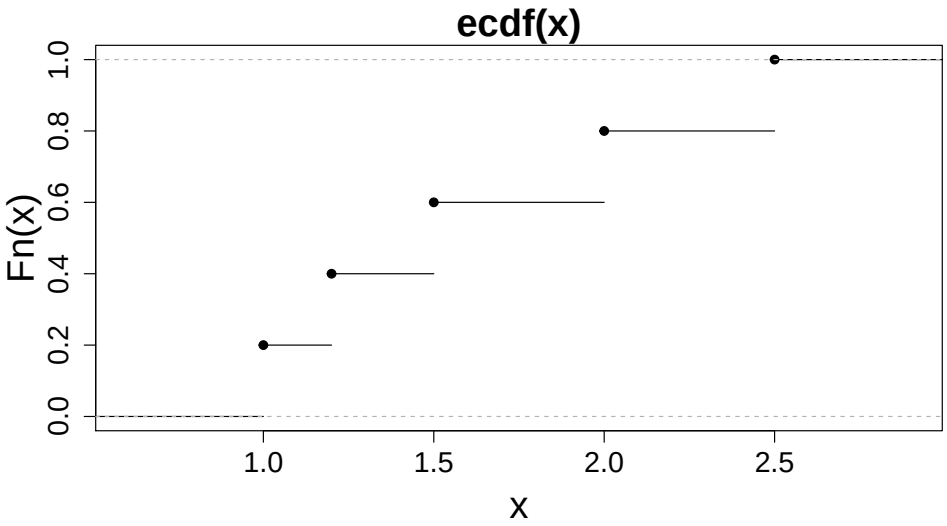
A natural estimator of a probability of an event is *the ratio of such an event in our sample*. Thus, we use

$$\widehat{F}_n(x_0) = \frac{\text{number of } X_i \le x_0}{\text{total number of observations}} = \frac{\sum_{i=1}^n I(X_i \le x_0)}{n} = \frac{1}{n}\sum_{i=1}^n I(X_i \le x_0) \tag{1.3}$$

as the estimator of $F(x_0)$.

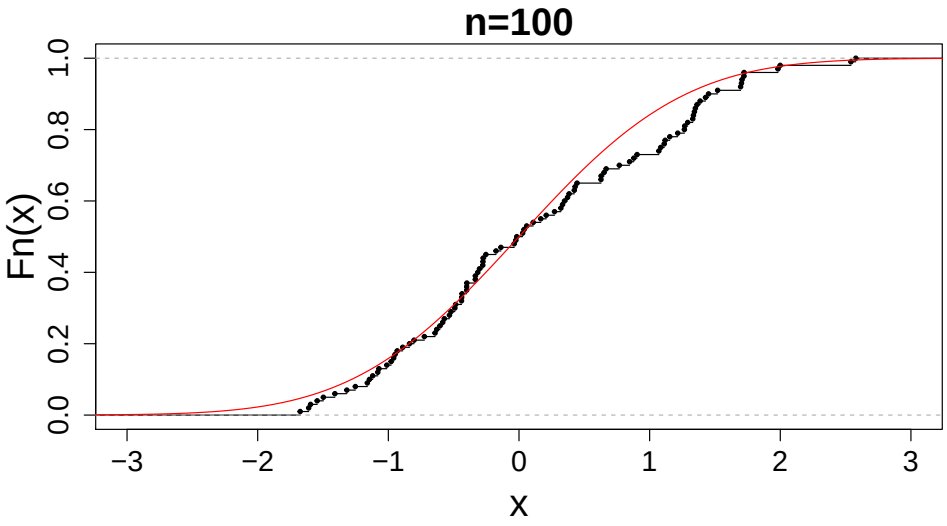
For every x_0, we can use such a quantity as an estimator, so the estimator of the CDF, $F(x)$, is $\widehat{F}_n(x)$. This estimator, $\widehat{F}_n(x)$, is called the *empirical distribution function (EDF)*.

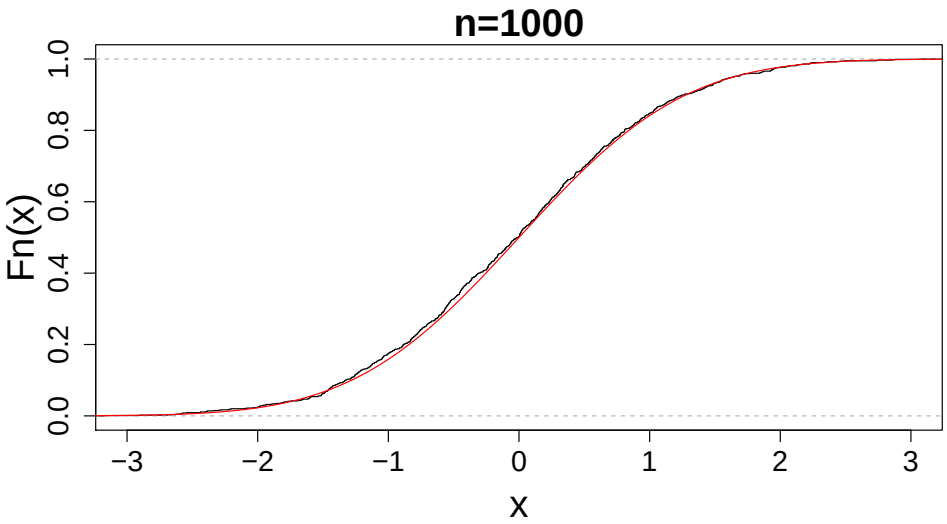
Example. Here is an example of the EDF of 5 observations of $1, 1.2, 1.5, 2, 2.5$:



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.

1.3.1 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?

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

Therefore, for a given x,

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

This implies

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

for a given x.

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

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

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.

♦ : The EDF can be used to test if the sample is from a known distribution or two samples are from the same distribution. The former is called the *goodness-of-fit test* and the latter is called the *two-sample test*. Assume that we want to test if $X_1, \cdots, X_n$ are from an known distribution F_0 (goodness-of-fit test). There are three common approaches to carry out this test. The first one is called the *KS test (Kolmogorov–Smirnov test)*[1], where the test statistic is the KS-statistic

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

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

The second approach is the *Cramér–von Mises test*[2], which uses the Cramér–von Mises statistic as the test statistic

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

The third approach is the *Anderson-Darling test*[3] and the test statistic is

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

We 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. Note that here we present the test statistics for the goodness-of-fit test, there are corresponding two-sample test version of each of them.

1.4 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] https://en.wikipedia.org/wiki/Cram%C3%A9r%E2%80%93von_Mises_criterion
[3] https://en.wikipedia.org/wiki/Anderson%E2%80%93Darling_test