Nonparametric Statistics

Ch.6 Cumulative distribution function

Motivation

❖ Definition of Cumulative distribution function (CDF) :

$$F(x) = P(X \leq x) \quad \forall x \in R$$

❖ The CDF plays a central role in statistical inferences. It contains all information about the random variable X. We can express some well-known quantities as functions of the CDF.

$$E(X) = \int x \, dF(x)$$

$$Var(X) = \int \left(x - E(X)\right)^2 dF(x).$$

$$p^{th} \ quantile = F^{-1}(p) \quad when \ F \ is \ strictly \ increasing.$$

❖ Therefore, estimating the CDF is a first step towards solving more important problems.

Properties of CDF

- $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to +\infty} F(x) = 1$

- If $f(x)$ is a probability density function of X,

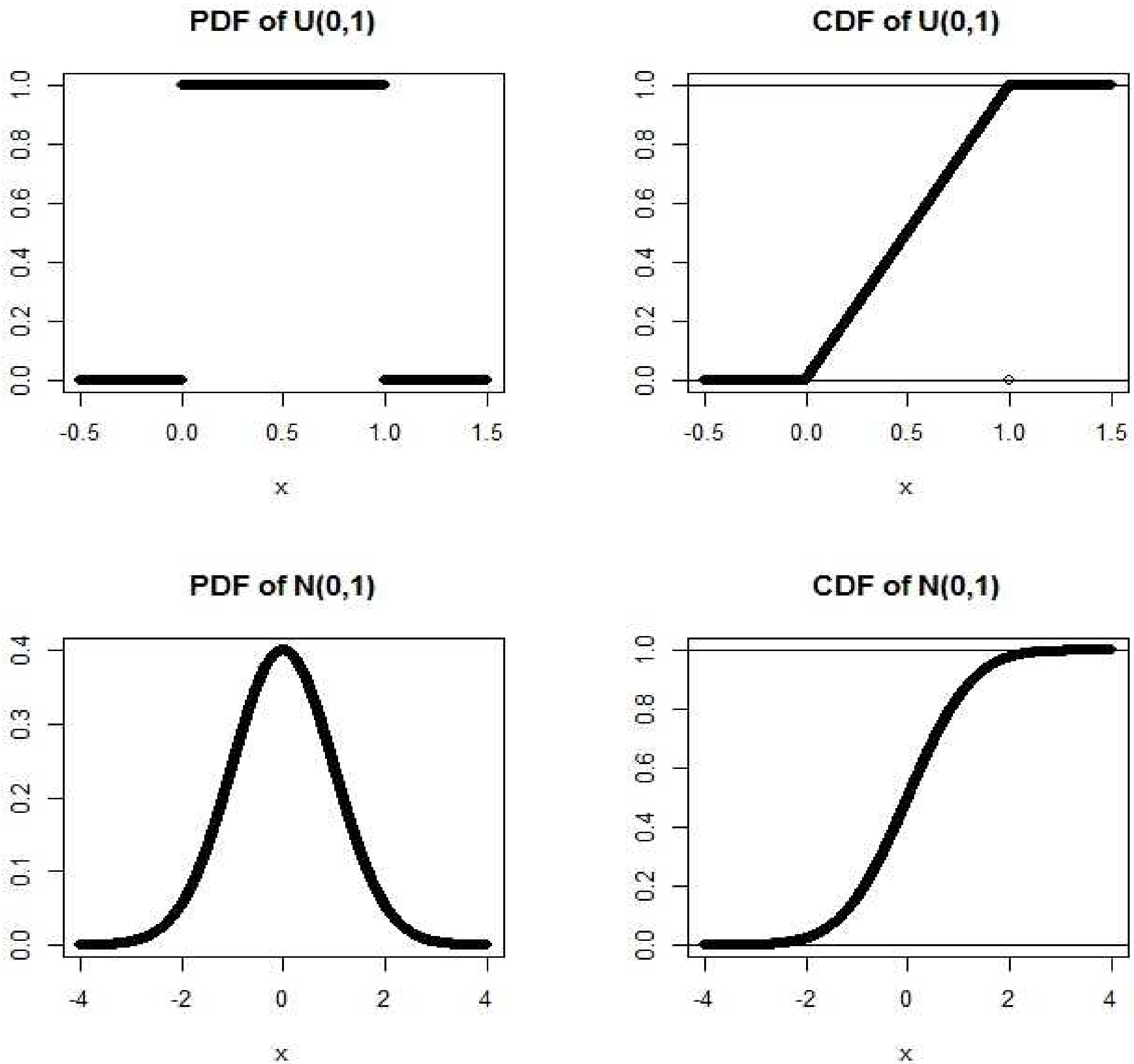
$$F(x) = \int_{-\infty}^{x} f(x)\ dx \quad : continuous \quad case$$

$$F(x) = \sum_{k \le x} f(k) \quad : discrete \quad case$$

- If F is continuous, then the random variable X is a continuous random variable.

- F is non-decreasing and right-continuous.

PDF & CDF


PDF of U(0,1)
CDF of U(0,1)
PDF of N(0,1)
CDF of N(0,1)
x
-0.5 0.0 0.5 1.0 1.5
0.0 0.2 0.4 0.6 0.8 1.0
-4 -2 0 2 4
0.0 0.1 0.2 0.3 0.4

Empirical distribution function (EDF)

❖ The EDF is an estimator of the CDF.

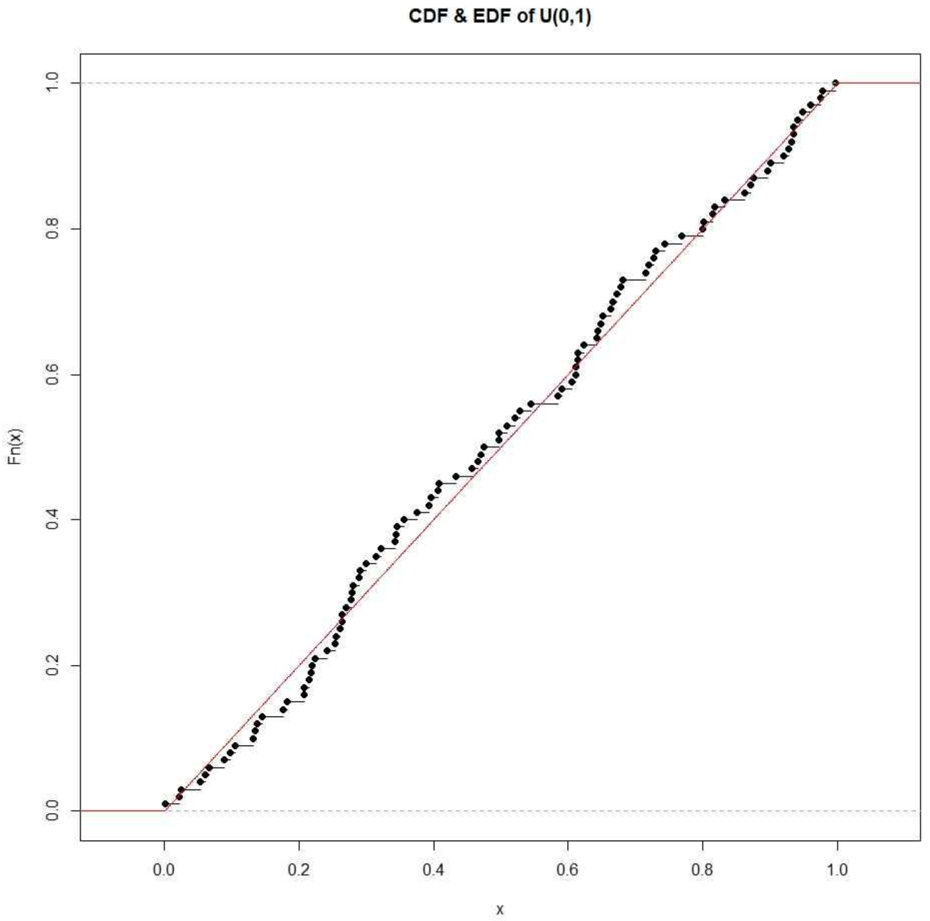
❖ Definition of the EDF : Let $X_1, \ldots, X_n$ be a random sample of size n from the distribution F.

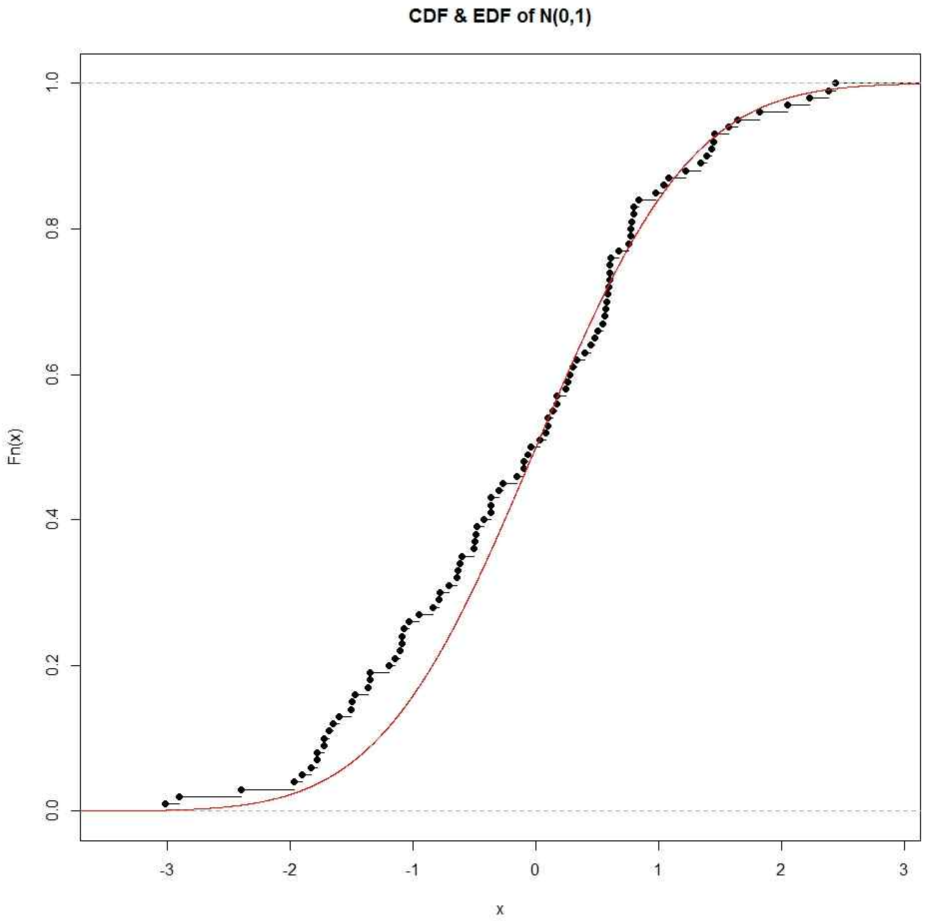
$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$$

❖ This satisfies $\lim_{x \to -\infty} \hat{F}_n(x) = 0$, $\lim_{x \to +\infty} \hat{F}_n(x) = 1.$

❖ $\hat{F}_n$ is the CDF that puts mass $1/n$ at each data point X_i . Note that $\hat{F}_n$ is a right-continuous step function having jumps on X_i's.

Empirical distribution function (EDF)





Properties of EDF

- (Consistency) Note that, for each fixed x,

$$E\left(\hat{F}_n(x)\right) = F(x) \quad \& \quad Var\left(\hat{F}_n(x)\right) = \frac{F(x)\left(1 - F(x)\right)}{n}.$$

By Chebyshev's inequality,

$$P\left(\left|\hat{F}_n(x) - F(x)\right| \geq k\right) \leq \frac{F(x)\left(1 - F(x)\right)}{nk^2} \to 0 \quad for \;\; every \quad k > 0,$$

which implies that $\hat{F}_n(x) \to F(x)$ in probability.

Properties of EDF

- (Uniform convergence : Glivenko-Cantelli theorem)

$$P\left(\sup_x |\hat{F}_n(x) - F(x)\Big| \to 0\right) = 1.$$

This is a much stronger result than the previous one.

- (Dvoretzky-Kiefer-Wolfowitz (DKW) inequality)

$$P\left(\sup_x |\hat{F}_n(x) - F(x)\Big| > k\right) \leq 2\exp(-2nk^2) \quad for \;\; any \quad k > 0.$$

This can be used to construct a confidence band for the CDF.

Example

Ex1] A random sample of size 8 yields the number of times people swam in the past month. Calculate the EDF.

$$1 \quad 0 \quad 6 \quad 2 \quad 4 \quad 2 \quad 6 \quad 7$$

At first, we make the data ordered.

$$0 \quad 1 \quad 2 \quad 2 \quad 4 \quad 6 \quad 6 \quad 7$$

Then,

$$\hat{F}_n(x) = 0, \qquad x < 0$$

$$\hat{F}_n(x) = \frac{1}{8}, \qquad 0 \le x < 1$$

$$\hat{F}_n(x) = \frac{1}{8} + \frac{1}{8} = \frac{1}{4}, 1 \le x < 2$$

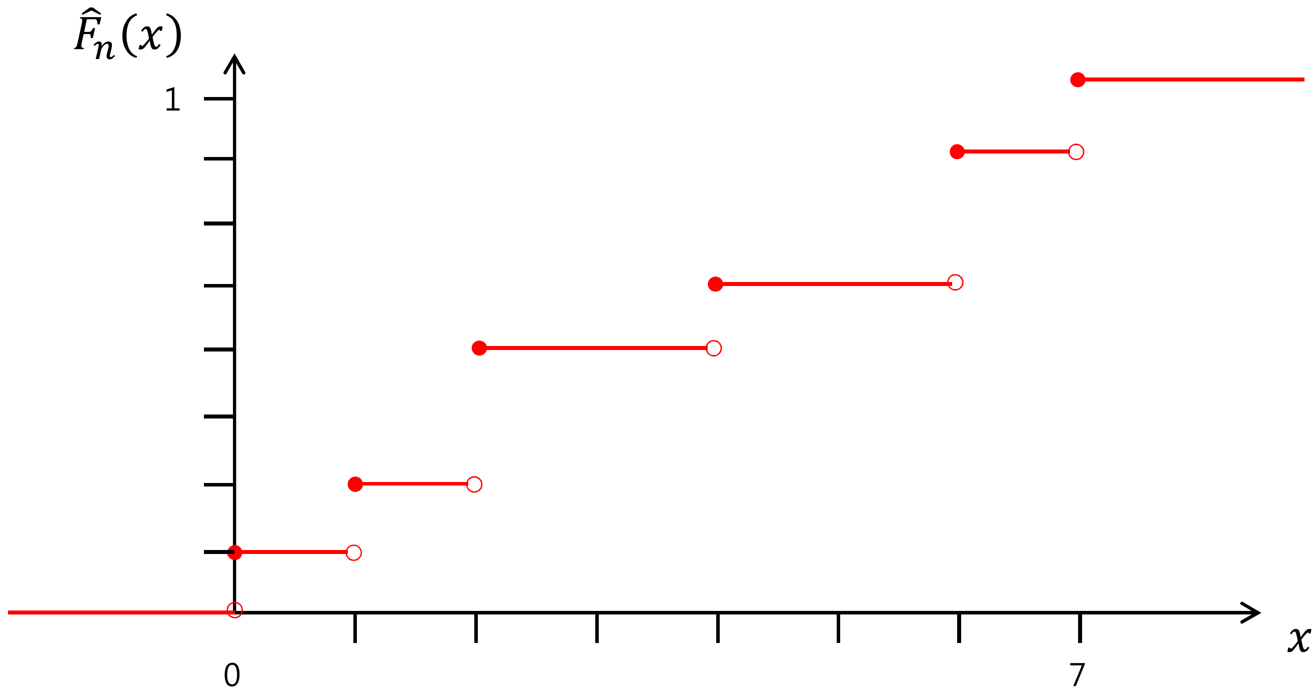
$$\hat{F}_n(x) = \frac{1}{8} + \frac{1}{8} + \frac{2}{8} = \frac{1}{2}, \qquad 2 \le x < 4$$

Example

$$\hat{F}_n(x) = \frac{1}{8} + \frac{1}{8} + \frac{2}{8} + \frac{1}{8} = \frac{5}{8}, \qquad 4 \le x < 6$$

$$\hat{F}_n(x) = \frac{1}{8} + \frac{1}{8} + \frac{2}{8} + \frac{1}{8} + \frac{2}{8} = \frac{7}{8}, \qquad 6 \le x < 7$$

$$\hat{F}_n(x) = \frac{1}{8} + \frac{1}{8} + \frac{2}{8} + \frac{1}{8} + \frac{2}{8} + \frac{1}{8} = 1, \qquad 7 \le x$$



Confidence sets

- Note that, $n\hat{F}_n(x) \sim B(n, F(x))$ for fixed x. Therefore, a pointwise confidence interval for $F(x)$ is given by

$$\hat{F}_n(x) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{F}_n(x)\left(1 - \hat{F}_n(x)\right)}{n}}$$

- However, we are mainly interested in estimating the whole function F. Therefore, we need to consider a functional version of confidence interval. We call it "confidence band".

Confidence sets

- From DKW inequality, if we set $k = \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}$,

$$P\left(\sup_x |\hat{F}_n(x) - F(x)\right| \leq k\Big) \geq 1 - \alpha.$$

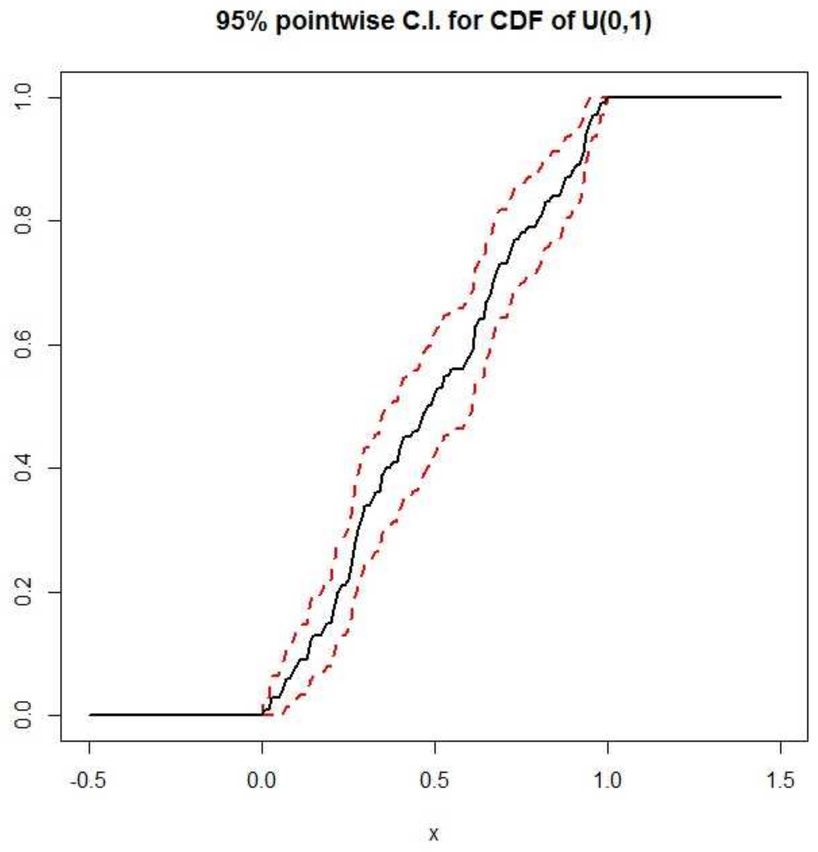
Therefore, for

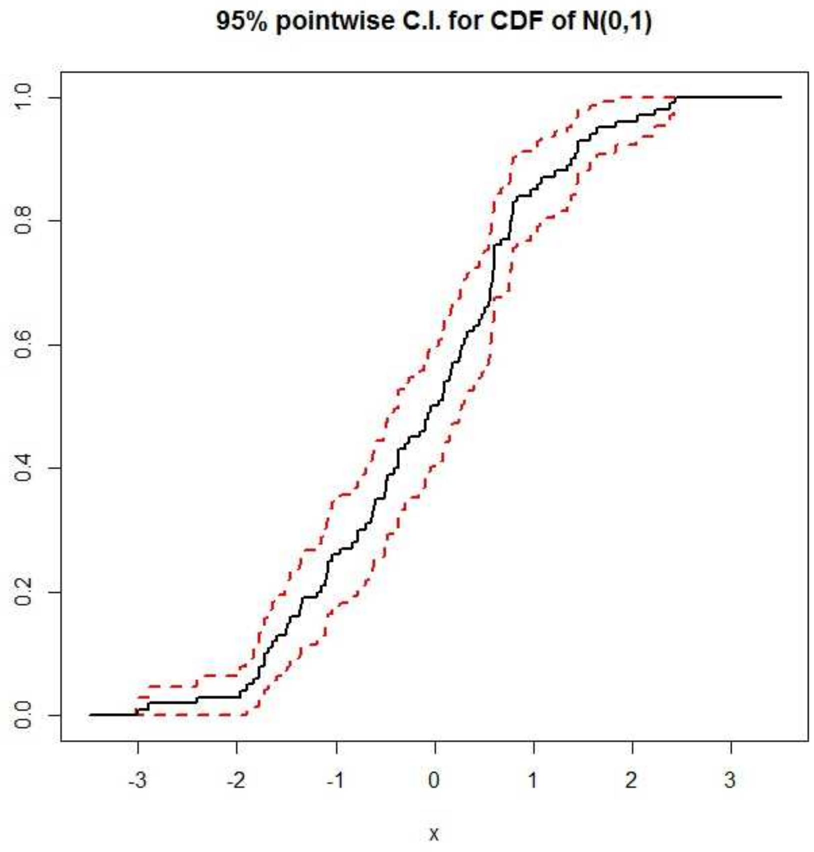
$$L(x) = \max\{\hat{F}_n(x) - k, 0\}$$

$$U(x) = \min\{\hat{F}_n(x) + k, 1\},$$

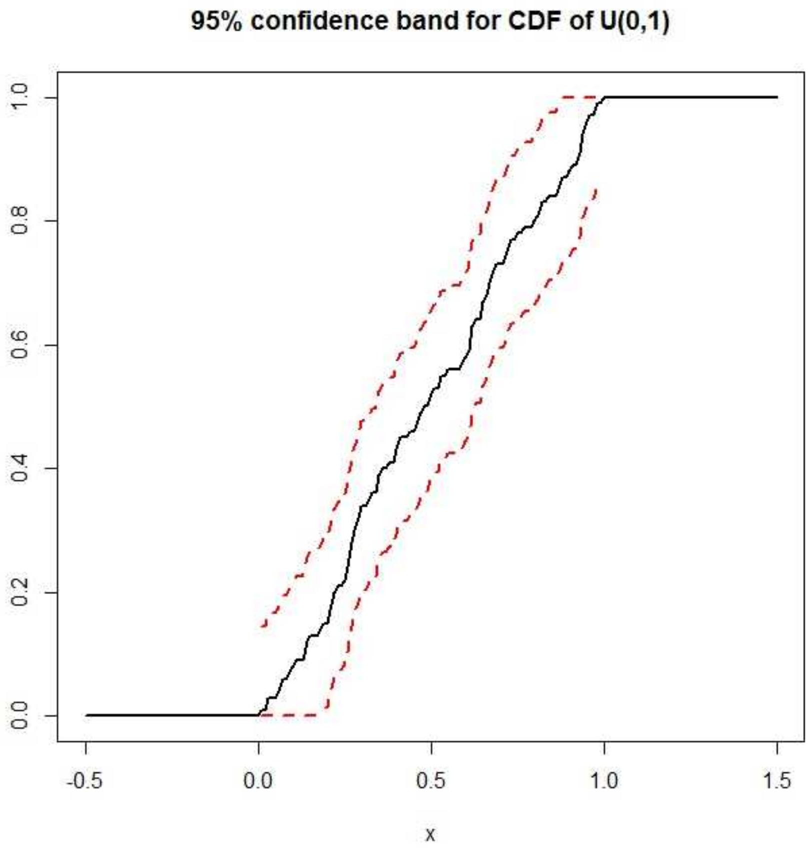
$$P(L(x) \leq F(x) \leq U(x) \quad \forall x) \geq 1 - \alpha.$$

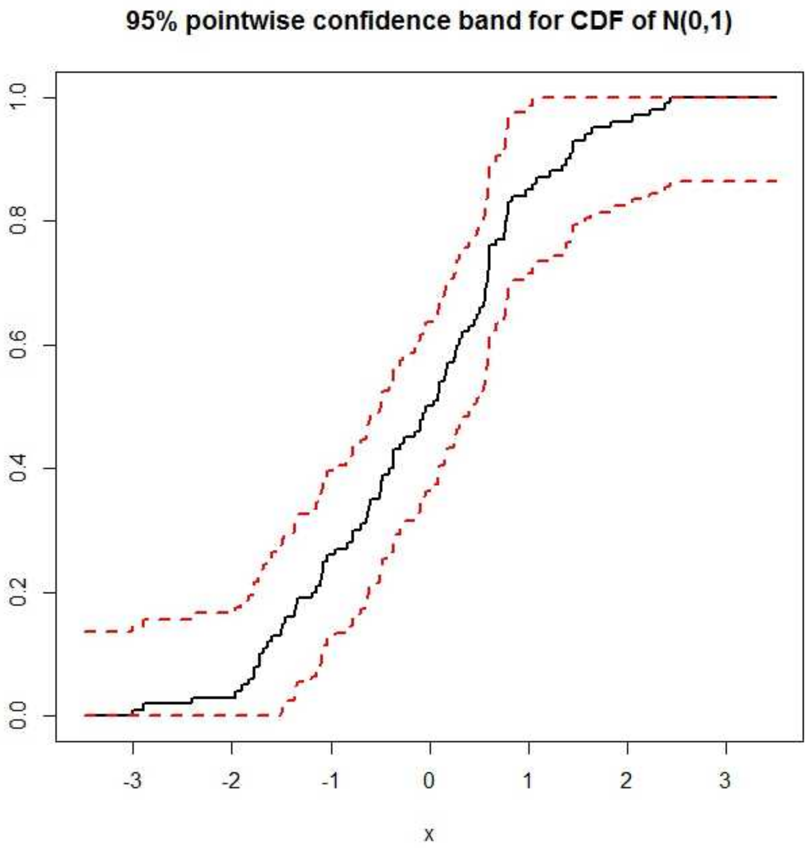
Pointwise confidence interval



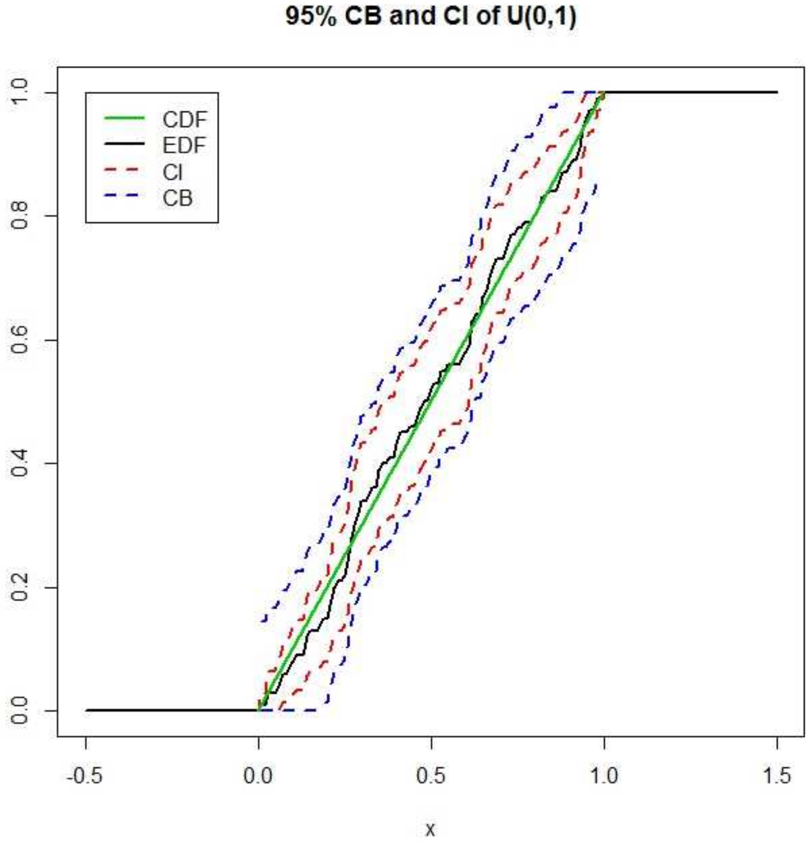


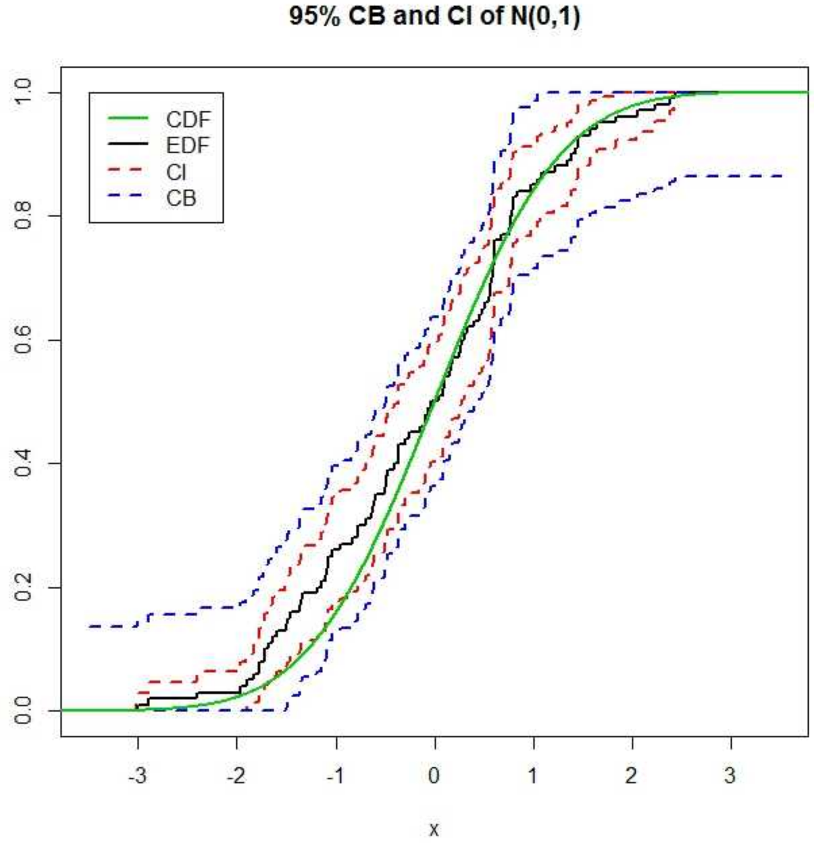
Confidence band





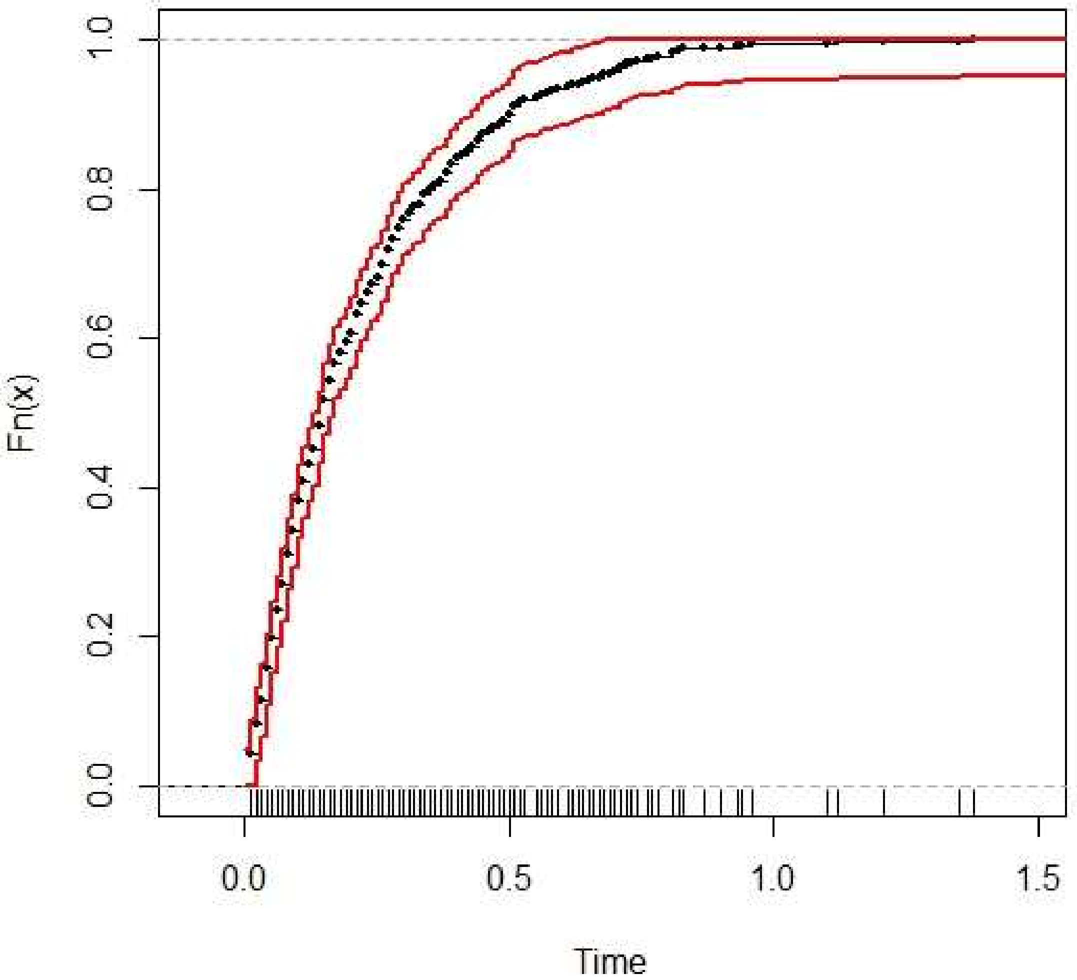
Confidence band vs Confidence interval





Real data example

- Nerve data (Cox and Lewis (1966)) : 799 waiting times between successive pulses along a nerve fiber.



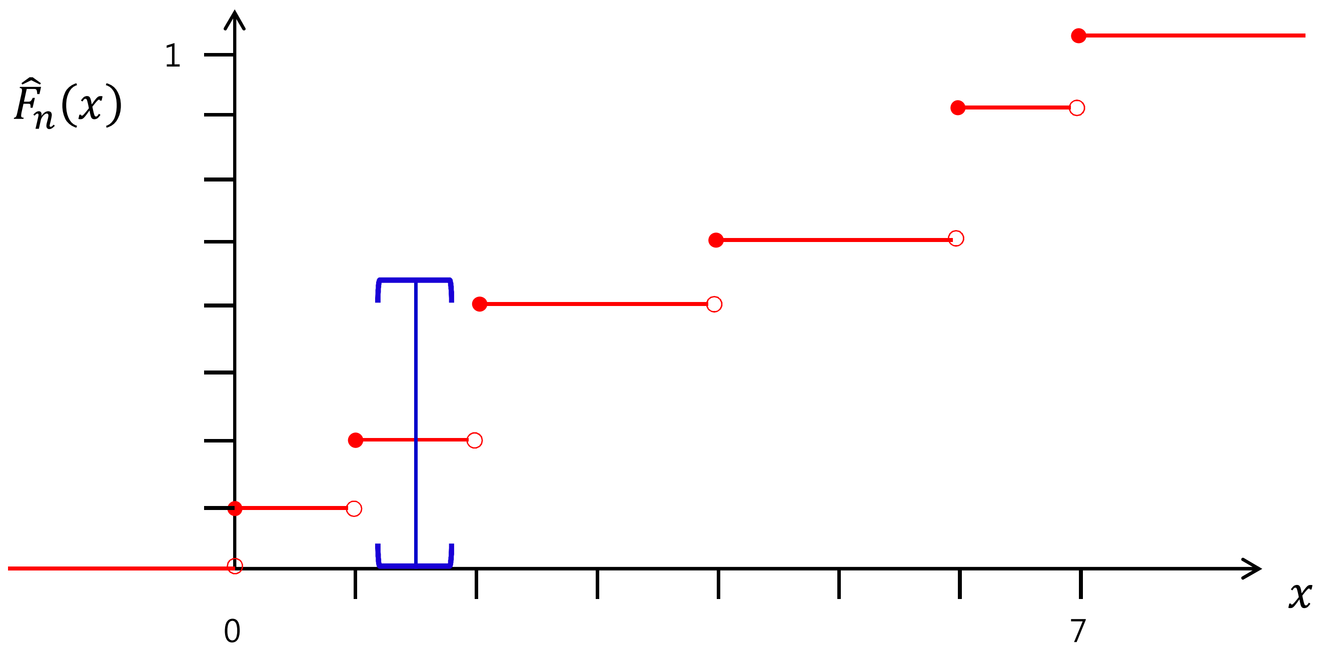
EDF
& confidence
band

Example

Ex1:revisited] What is a $100 * (1-\alpha)\%$ pointwise confidence interval for $F(1.5)$?

$$\hat{F}_n(1.5) \pm z_{0.025}\sqrt{\frac{\hat{F}_n(1.5)\left(1-\hat{F}_n(1.8)\right)}{8}} = 0.25 \pm 1.96 \times \sqrt{0.25 * \frac{0.75}{8}}$$

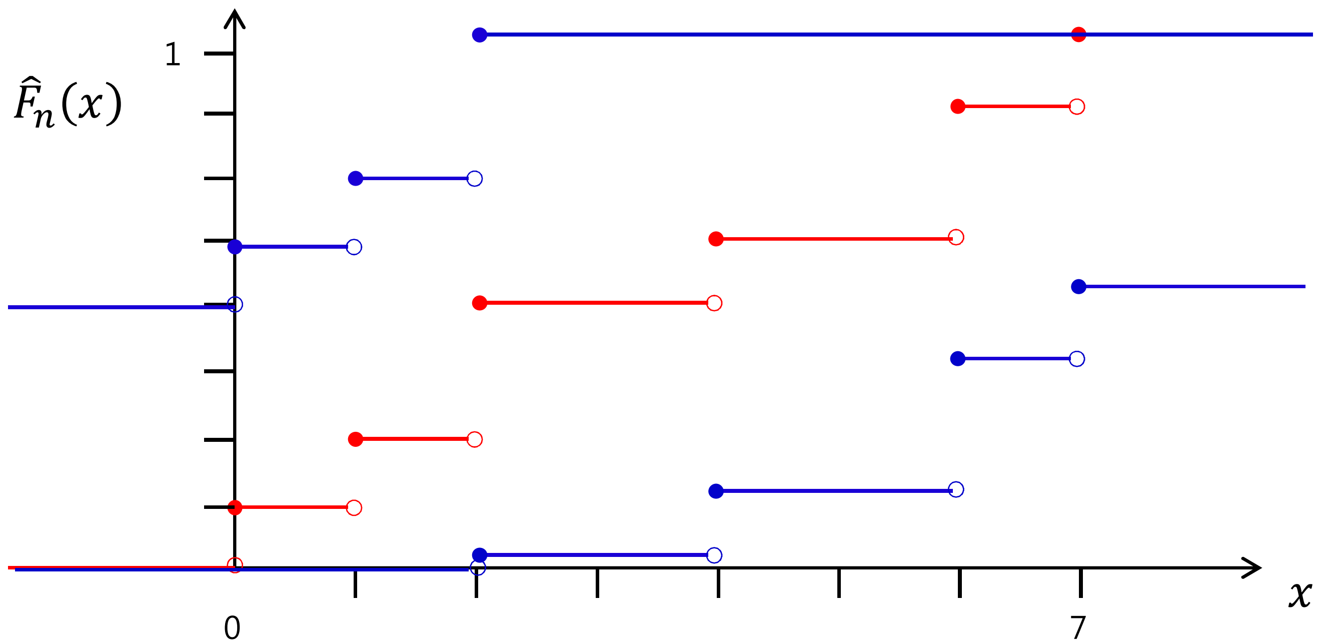
$$= (-0.0501{,}0.5501)$$



Example

Ex1:revisited] What is a $100 * (1-\alpha)\%$ confidence band for F?

$$k = \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}} = \sqrt{\frac{1}{16}\log\frac{2}{0.05}} = 0.4802$$



Estimation based on the EDF

- (Sample mean & variance)

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \int x \, d\hat{F}_n(x) \quad \& \quad \frac{n-1}{n} \cdot S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \int (x - \bar{X})^2 \, d\hat{F}_n(x)$$

- (Sample quantile)

$$p^{th} \text{ sample quantile} = \hat{F}_n^{-1}(p) = X_{([np])}$$

where $F^{-1}(p) = \inf\{x : F(x) \geq p\}$ and $[np]$ is the smallest integer equal to or larger than np.

Goodness of fit (GoF) test

- Sometimes, we want to test if data comes from a specific distribution. Gof tests are designed for this purpose.

- Essentially, Gof test statistics try to figure out how far the EDF based on data is from a specified distribution. There are many different ways to measure the distance between the EDF and the specified distribution.

- Hypotheses (two sided)

$$H_0: F(x) = F_0(x) \quad vs \quad H_1: F(x) \neq F_0(x) \quad \forall x$$

One-sided test is also possible. If $F(x) > F_0(x)$, it means that F is stochastically larger than F_0.

Goodness of fit (GoF) test

- (Kolmogorov-Smirnov statistic)

$$D_n = \sup_x |\hat{F}_n(x) - F_0(x)|$$

- (Cramer-von Mises statistic)

$$T_n = n\int \left(\hat{F}_n(x) - F_0(x)\right)^2 dF_0(x)$$

- (Anderson-Darling statistic)

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

Goodness of fit (GoF) test

- (Pearson chi-square statistic)

  For this test, we need to first make binning of data. Let $(-\infty, x_1], (x_1, x_2], \dots, (x_{k-2}, x_{k-1}], (x_{k-1}, \infty)$ be a partition of the Real line and $O_1, \dots, O_k$ be the numbers of data belonging to each bin. Then, the test statistic is defined by

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where $E_i = F(x_i) - F(x_{i-1})$. This follows $\chi^2(k-1)$ under the null hypothesis.

Goodness of fit (GoF) test

- The aforementioned test statistics are for general distributions. However, those are particularly useful to test for normality.

- There are another test statistics designed only for normality. (ex] Shapiro–Wilk test statistic, Jarque–Bera test and so on.

- Two sample goodness of fit tests also exist. For example,

$$D_{m,n} = \sup_{x} |\hat{F}_{1,m}(x) - \hat{F}_{2,n}(x)|$$

where $\hat{F}_{1,m}(x)$ and $\hat{F}_{2,n}(x)$ are the EDFs based on each sample.

Example

Ex1:revisited] Calculate the Kolmogorov-Smirnov statistic when we want to test if the number of times is uniformly distributed on [0,8].

The CDF of $U[0,8]$ is $F_0(x) = \frac{1}{8}x\, I(0 \le x \le 8) + I(x > 8)$.

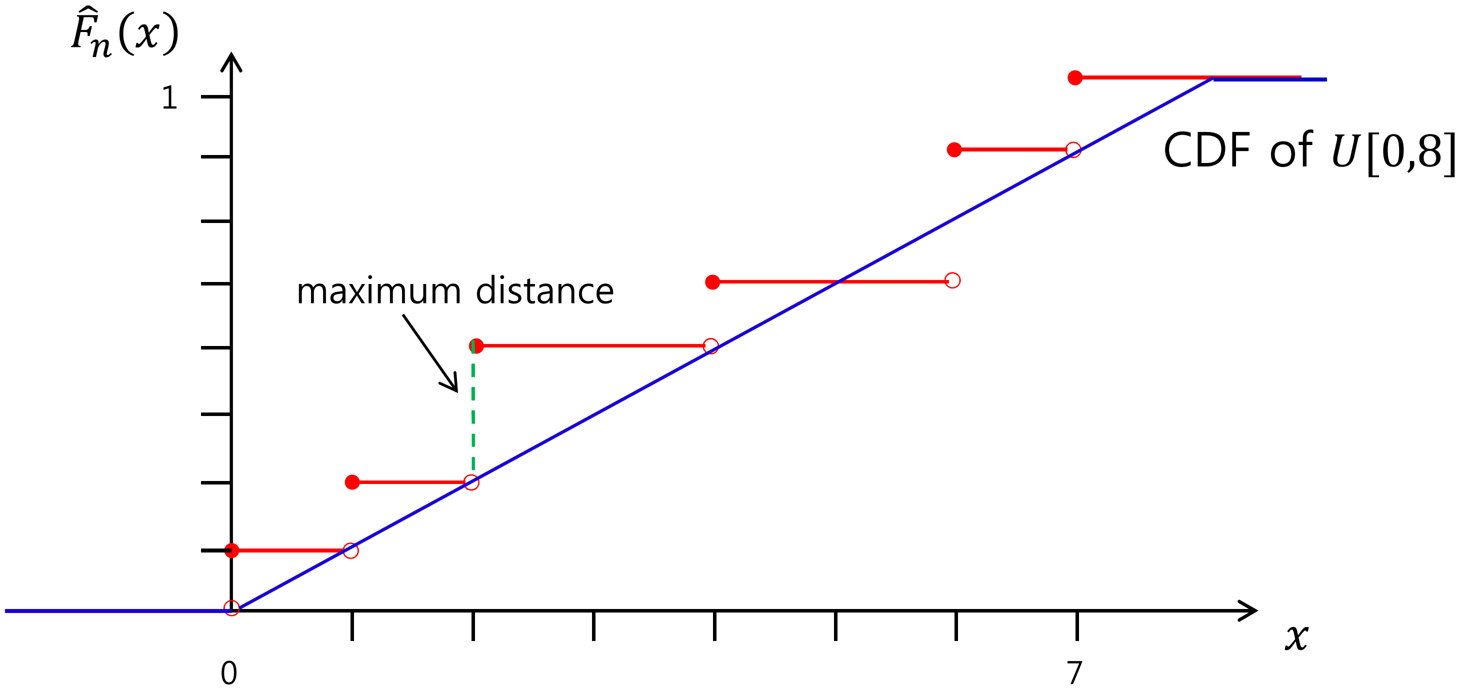
We want to test

$$H_0: F(x) = F_0(x) \quad vs \quad H_1: F(x) \neq F_0(x) \quad \forall x$$

where F is the true CDF.

Example

From the figure, we can see that the observed test statistic is 0.25



Example

Ex1:revisited] Perform the Pearson's chi-square test to see if the number of times is uniformly distributed on [0,8] at the significance level 0.05.

(1) Binning

We divide [0,8] into 4 intervals. $[0,2), [2,4), [4,6), [6,8]$

(2) Frequency table

Bin	$[0,2)$	$[2,4)$	$[4,6)$	$[6,8]$
O	2	2	1	3
E	2	2	2	2

(3) Observed test statistic

$$\frac{(2-2)^2}{2} + \frac{(2-2)^2}{2} + \frac{(1-2)^2}{2} + \frac{(3-2)^2}{2} = 1$$

(4) Critical region

$$\chi^2_{0.05}(4-1) = 7.81$$

We do not reject the null hypothesis. No clear evidence that the true CDF is not U[0,8].