$X \sim F$ $\quad F(x) = P(X \leq x) \quad \forall \; x \in \mathbb{R}$

STAT 824 sp 2023 Lec 01 slides

Estimating a cdf

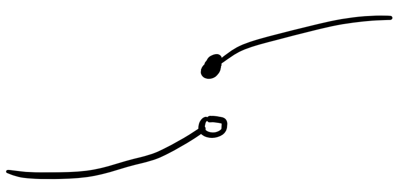
i) Nondecreasing

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

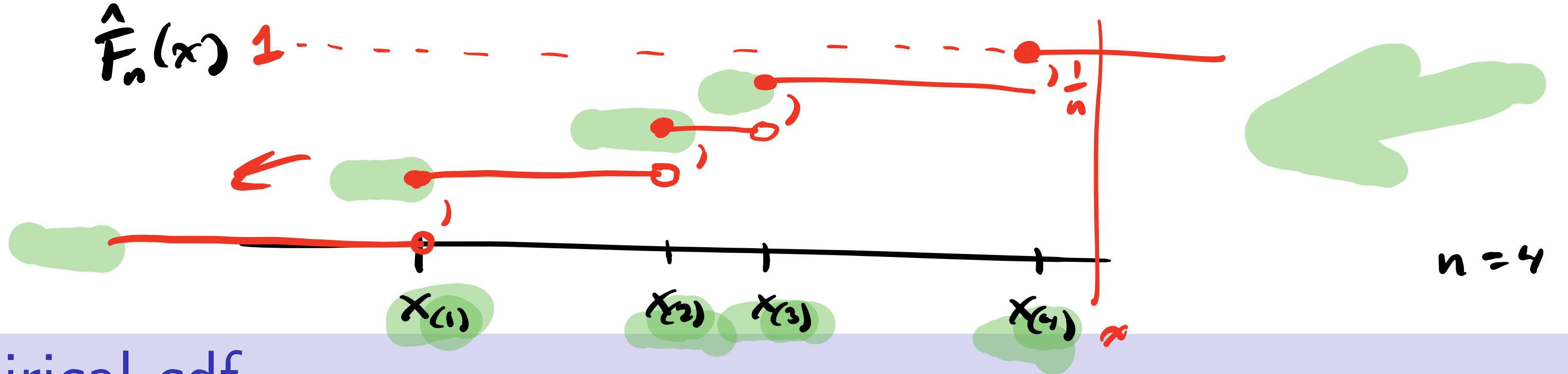
iii) Right-continuous

Karl B. Gregory

University of South Carolina



These slides are an instructional aid; their sole purpose is to display, during the lecture, definitions, plots, results, etc. which take too much time to write by hand on the blackboard. They are not intended to explain or expound on any material.



Empirical cdf

The empirical cdf of a set of values $X_1, \dots, X_n \in \mathbb{R}$ is given by

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i \le x) \quad \text{for all } x \in \mathbb{R}.$$

$\hat{F}_n(x)$ = proportion of X_is $\le x$

Discuss: Is this a legitimate cdf? (Three properties).

i) nondecreasing

ii) ✓

iii) ✓

Recall: $X_1, \dots, X_n$ iid w/ mean μ, var σ^2; $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{D} N(0, \sigma^2)$ [CLT].

$$\hat{F}_n(x_0) = \frac{1}{n} \sum_{i=1}^n \underbrace{\mathbb{1}(X_i \le x_0)}_{W_i} = \frac{1}{n} \sum_{i=1}^n W_i,$$

$W_i \sim \text{Bernoulli}(F(x_0))$

$\mathbb{E} W_i = F(x_0)$ $\quad$ $\text{Var}\, W_i = F(x_0)[1 - F(x_0)]$

Central limit result for empirical cdf at a point

If $X_1, \dots, X_n$ is a rs from a distribution with cdf F, then for any $x_0 \in \mathbb{R}$ we have

$\mu = \mathbb{E} W_i$

$$\sqrt{n}(\hat{F}_n(x_0) - F(x_0)) \to \text{Normal}(0, \underbrace{F(x_0)[1 - F(x_0)]}_{\sigma^2 = \text{Var}\, W_i}) \text{ in distribution}$$

This is a mean.

as $n \to \infty$.

Exercise:

1. Prove the above result. ✓
2. Use the result to construct an asymptotic $(1-\alpha)100\%$ CI for $F(x_0)$. ✓

$$\frac{\sqrt{n}\,(\hat{F}_n(x_0) - F(x_0))}{\sqrt{F(x_0)[1 - F(x_0)]}} \xrightarrow{D} N(0,1)$$

$$\lim_{n\to\infty} P\left(-z_{\alpha/2} < \frac{\sqrt{n}\left(\hat{F}_n(x_0) - F(x_0)\right)}{\sqrt{F(x_0)[1-F(x_0)]}} < z_{\alpha/2} \right) = 1-\alpha$$

Since

$$\hat{F}_n(x_0) \xrightarrow{p} F(x_0) \qquad \left[\quad \bar{X}_n \xrightarrow{p} \mu \quad \text{by} \quad \text{WLLN} \right]$$

$$\frac{\sqrt{n}\left(\hat{F}_n(x_0) - F(x_0)\right)}{\sqrt{\hat{F}_n(x_0)[1-\hat{F}_n(x_0)]}} \xrightarrow{D} N(0,1) \quad \text{, by Slutzky's theorem.}$$

$$\lim_{n\to\infty} P\left(-z_{\alpha/2} < \frac{\sqrt{n}\left(\hat{F}_n(x_0) - F(x_0)\right)}{\sqrt{\hat{F}(x_0)[1-\hat{F}(x_0)]}} < z_{\alpha/2} \right) = 1-\alpha$$

Rearrange to get $F(x_0)$ alone in middle

$$\lim_{n\to\infty} P\left(\hat{F}_n(x) - z_{\alpha/2}\sqrt{\frac{\hat{F}_n(x_0)[1-\hat{F}_n(x_0)]}{n}} < F(x) < \hat{F}_n(x_0) + z_{\alpha/2}\sqrt{\frac{\hat{F}_n(x_0)[1-\hat{F}_n(x_0)]}{n}} \right) = 1-\alpha$$

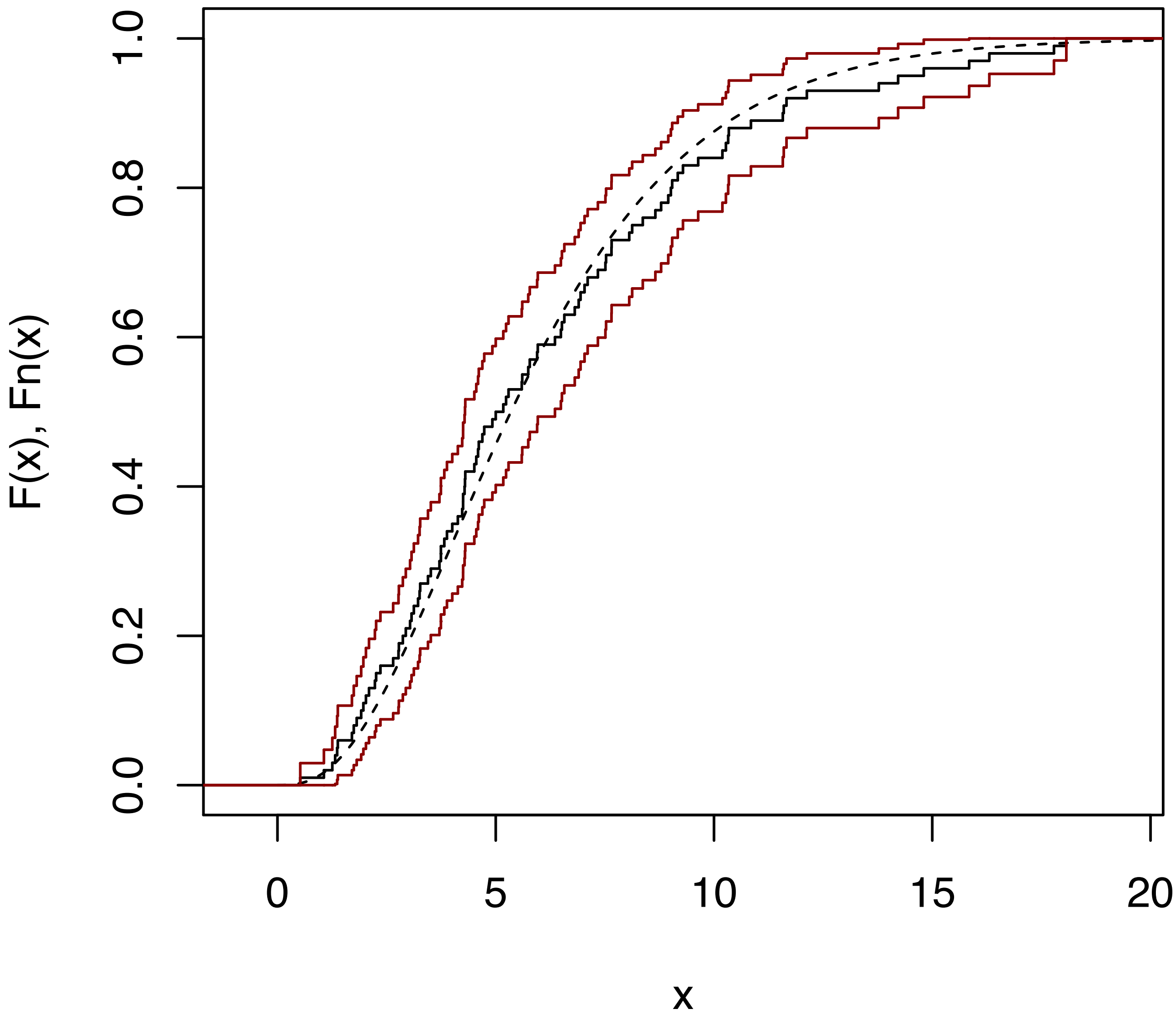
So a $(1-\alpha)100\%$ C.I. for $F(x_0)$ is

$$\hat{F}_n(x_0) \pm z_{\alpha/2}\sqrt{\frac{\hat{F}_n(x_0)\left[1-\hat{F}_n(x_0)\right]}{n}}.$$

Exercise: Generate some data $X_1, \ldots, X_n$ and make a plot with

1. the empirical cdf.
2. the true cdf.
3. pointwise confidence intervals at each of the values $X_1, \ldots, X_n$.

Can plot nicely with the `stepfun` function in R.



F(x), Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
x

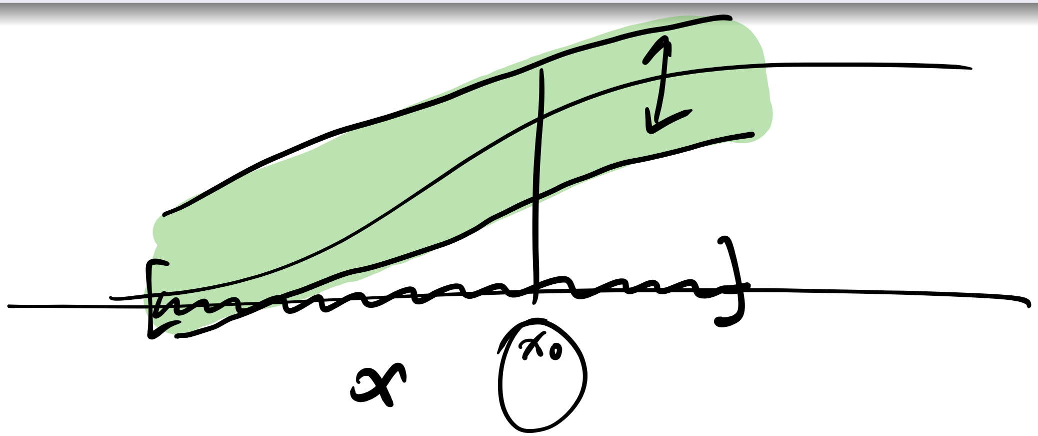
Pointwise CIs versus confidence bands for a function

1. A $(1-\alpha) \times 100\%$ CI for F at a point x_0 is an interval $[L(x_0), U(x_0)]$ such that

$$P(L(x_0) \leq F(x_0) \leq U(x_0)) \geq 1 - \alpha.$$

2. A $(1-\alpha) \times 100\%$ *confidence band* for F over an interval $\mathcal{X}$ is a region $\{(x, y) : L(x) \leq y \leq U(x), x \in \mathcal{X}\}$ such that

$$P(L(x) \leq F(x) \leq U(x) \text{ for all } x \in \mathcal{X}) \geq 1 - \alpha.$$



Dvoretzky-Kiefer-Wolfowitz inequality

If $X_1, \ldots, X_n$ is a rs from a distribution with cdf F, then

$$P\left(\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \leq \varepsilon\right) \geq 1 - 2e^{-2n\varepsilon^2}$$

Exercise:

1. Use the DKW result to construct a $(1-\alpha) \times 100\%$ confidence band for F.
2. Add the bans to the plot with the pointwise CIs.

$$P\left(\sup_{x\in\mathbb{R}} \left|\hat{F}_n(x) - F(x)\right| \le \varepsilon\right)$$

$$= P\left(|\hat{F}_n(x) - F(x)| \le \varepsilon \quad \text{for all } x \in \mathbb{R}\right)$$

$$= P\left(-\varepsilon \le \hat{F}_n(x) - F(x) \le \varepsilon \quad \text{for all } x \in \mathbb{R}\right)$$

$$= P\left(\hat{F}_n(x) - \varepsilon \le F(x) \le \hat{F}_n(x) + \varepsilon \quad \text{for all } x\in\mathbb{R}\right)$$

$$\ge 1 - 2e^{-2n\varepsilon^2}$$

How to make a $(1-\alpha)*100\%$ C. band?

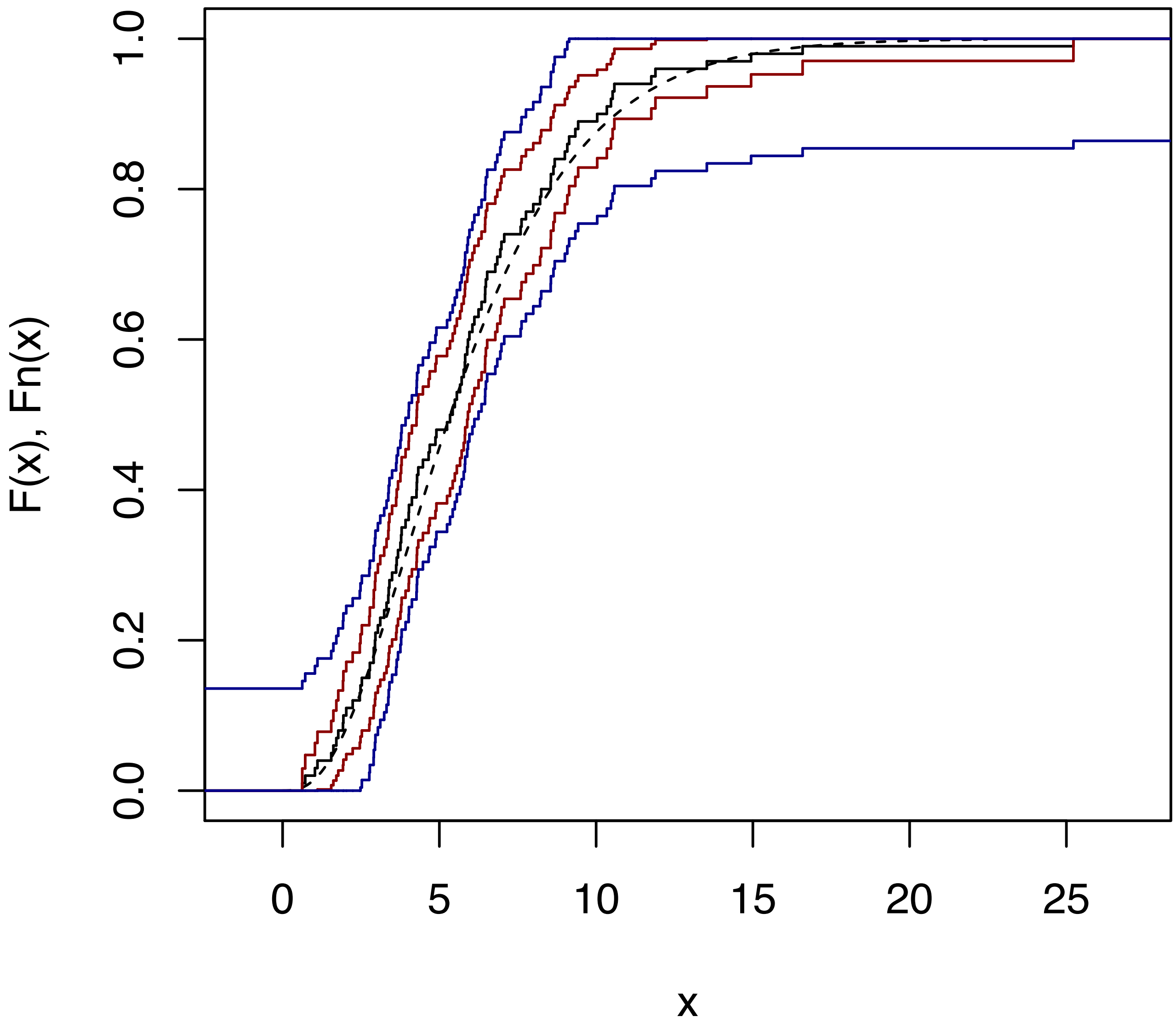
Set $\quad 1-\alpha = 1 - 2e^{-2n\varepsilon^2}$

$$\iff \frac{\alpha}{2} = e^{-2n\varepsilon^2}$$

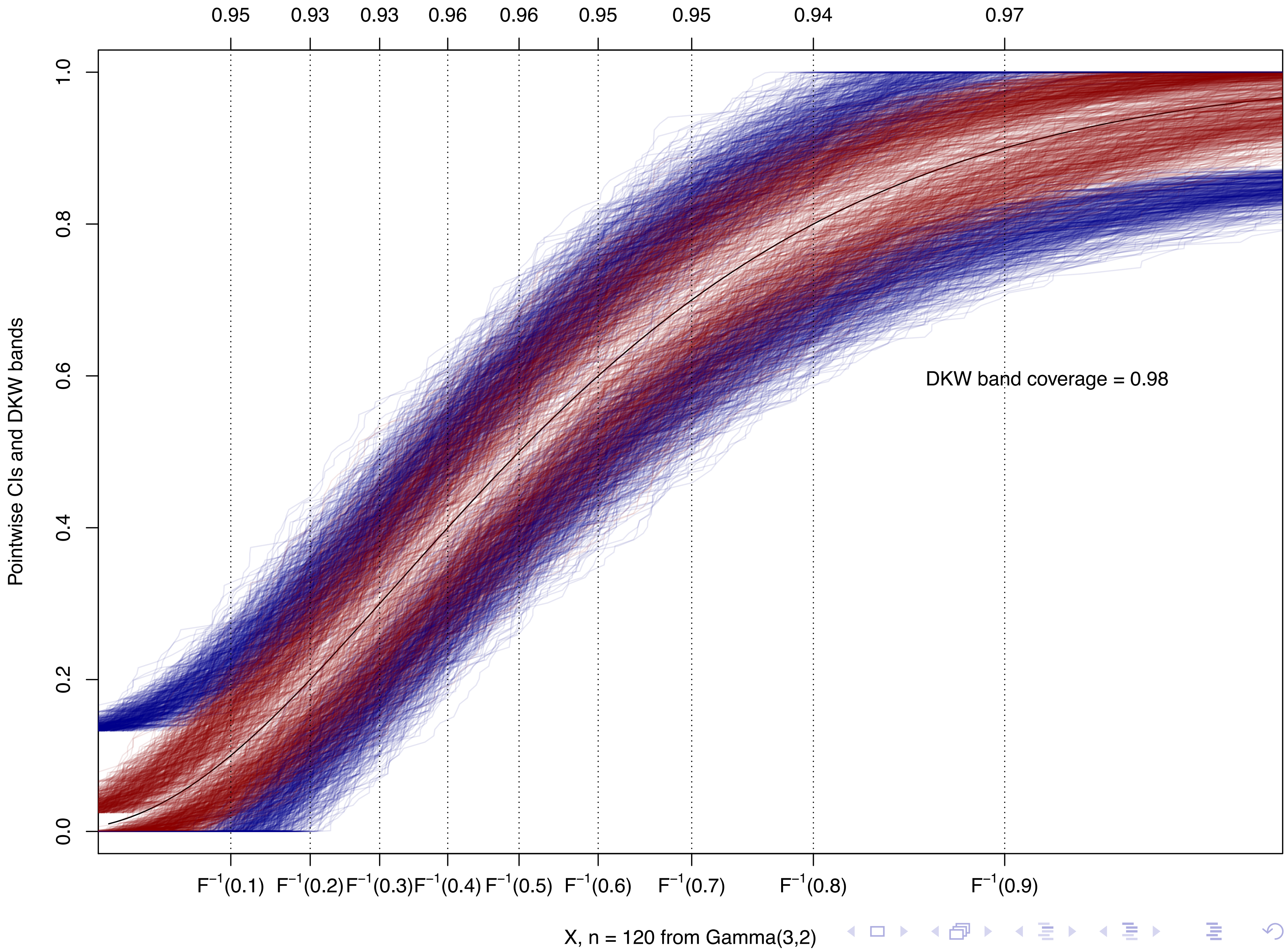
$$\iff \log\left(\frac{\alpha}{2}\right) = -2n\varepsilon^2$$

$$\iff \sqrt{\frac{1}{n}\frac{1}{2}\log\left(\frac{2}{\alpha}\right)} = \varepsilon$$

$(1-\alpha)*100\%$ C.B. is $\quad \hat{F}_n(x) \pm \frac{1}{\sqrt{n}}\sqrt{\frac{1}{2}\log\left(\frac{2}{\alpha}\right)}$



F(x), Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
x


0.95
0.93
0.93
0.96
0.96
0.95
0.95
0.94
0.97
Pointwise CIs and DKW bands
0.0
0.2
0.4
0.6
0.8
1.0
DKW band coverage = 0.98
$F^{-1}(0.1)$
$F^{-1}(0.2)$
$F^{-1}(0.3)$
$F^{-1}(0.4)$
$F^{-1}(0.5)$
$F^{-1}(0.6)$
$F^{-1}(0.7)$
$F^{-1}(0.8)$
$F^{-1}(0.9)$
X, n = 120 from Gamma(3,2)

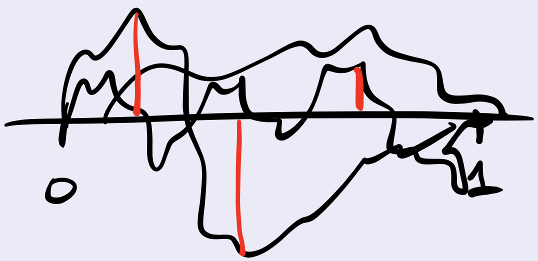
$\sqrt{n}(\bar{x}_n - \mu) \xrightarrow{D} N(0, \sigma^2)$

Kolmogorov-Smirnov-Donsker

If $X_1, \dots, X_n$ is a rs from a distribution with *continuous* cdf F, then

1. $$\sqrt{n} \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \to \sup_{t \in [0,1]} |B_0(t)| \quad \text{in distribution}$$

   as $n \to \infty$, where B_0 is a Brownian bridge.



2. $$P\left(\sup_{t \in [0,1]} |B_0(t)| \leq x \right) = 1 - 2 \sum_{i=1}^{\infty} (-1)^{i+1} \exp(-2i^2 x^2) \quad \text{for all } x \in \mathbb{R}.$$

   1.365

Discuss: How to build confidence bands with above.

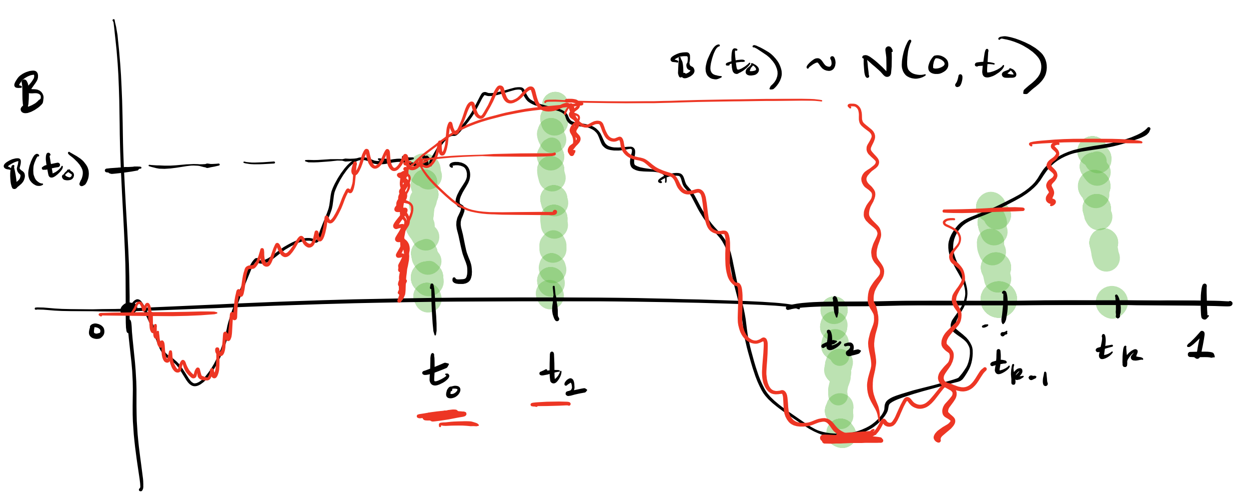
B_α = upper α quantile of the rv $\sup_{t \in [0,1]} |B_0(t)|$.

$$\lim_{n\to\infty} P\left(\sqrt{n} \sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| \le B_\alpha \right) = 1-\alpha$$

$$\Rightarrow \lim_{n\to\infty} P\left(\hat{F}_n(x) - \frac{1}{\sqrt{n}} B_\alpha \le F(x) \le \hat{F}_n(x) + \frac{1}{\sqrt{n}} B_\alpha \text{ for all } x\in\mathbb{R} \right) = 1-\alpha$$

So a $(1-\alpha)\cdot 100\%$ C.B. for $F(x)$, $x\in\mathbb{R}$ is

$$\hat{F}_n(x) \pm \frac{1}{\sqrt{n}} B_\alpha, \quad x\in\mathbb{R}.$$



random function

Wiener process or standard Brownian motion

A *Wiener process* B is a rf in the space $\boxed{C[0,1]}$ of cont. fns on $[0,1]$ which satisfies

1. $B(0) = 0$ with probability 1.
2. $B(t) \sim \text{Normal}(0, t)$, for $t \in (0, 1]$.
3. For $0 \leq t_0 \leq t_1 \leq \cdots \leq t_k \leq 1$, the increments

$$B(t_0) - B(0), \ldots, B(t_k) - B(t_{k-1})$$

are mutually independent.

Is also called *standard Brownian motion (SBM)*.

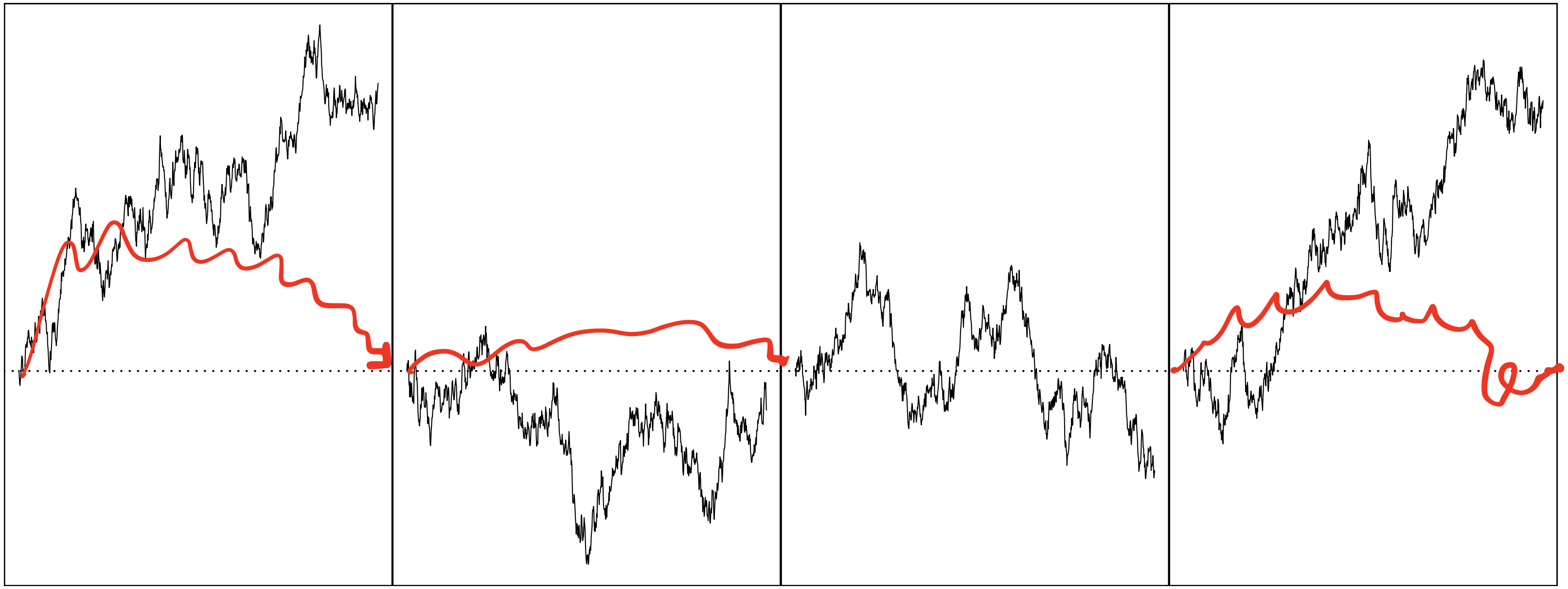
How to generate a standard Brownian motion

For each $n \geq 1$, let

$$B_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor tn \rfloor} Z_i, \quad Z_1, \ldots, Z_n \overset{\text{ind}}{\sim} \text{Normal}(0,1).$$

Then B_n converges to B by a functional CLT called Donsker's Theorem.

Exercise: Generate some (approximate) realizations of SBM and plot them.



$t=0$

$B_0(0) = B(0)$

$t=1$

$B_0(1) = B(1) - 1\,B(1) = 0$

Brownian bridge

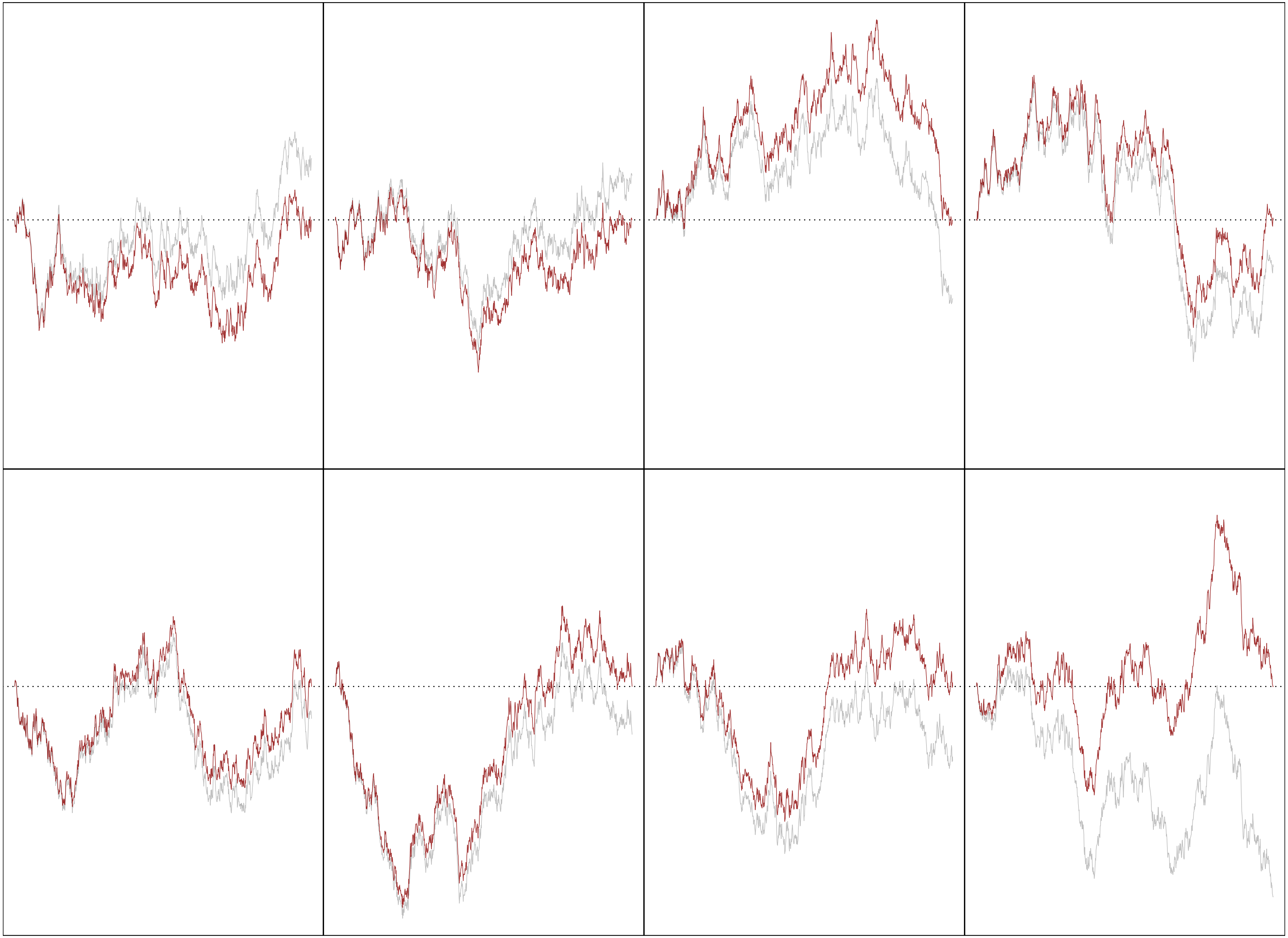
A *Brownian bridge* is the random function in $C[0,1]$ given by

$$B_0(t) = B(t) - tB(1),$$

where B is a standard Brownian motion.

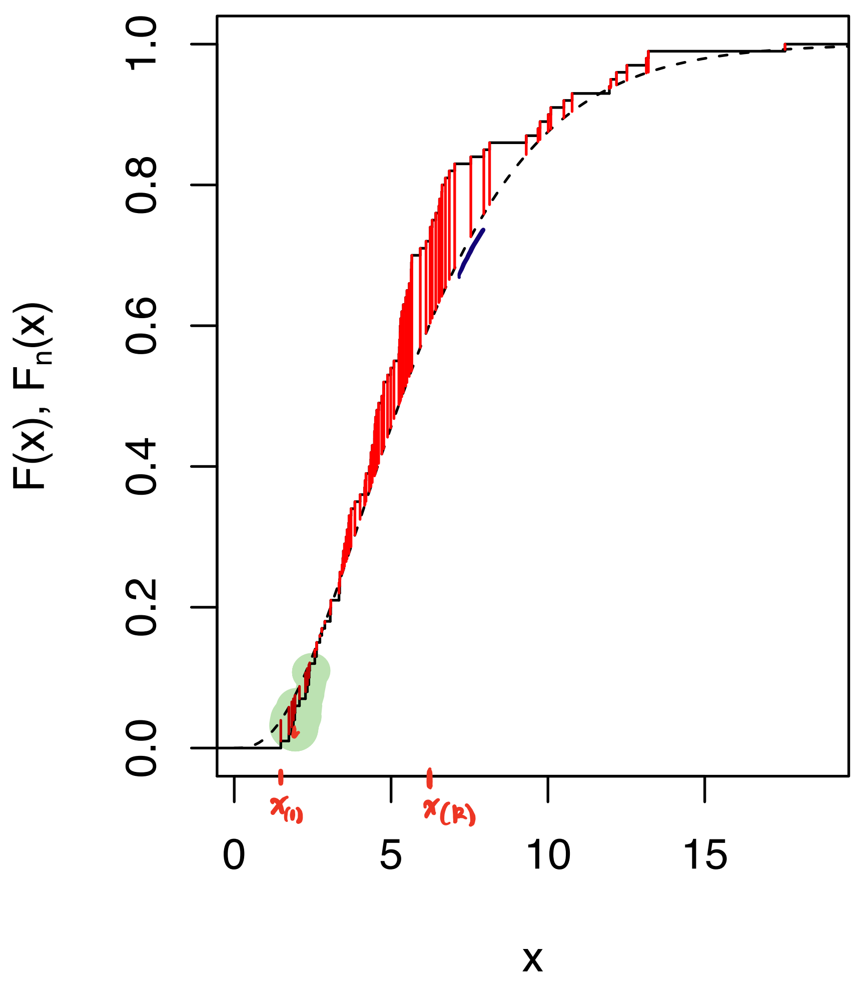
The "bridge" begins and ends at 0.

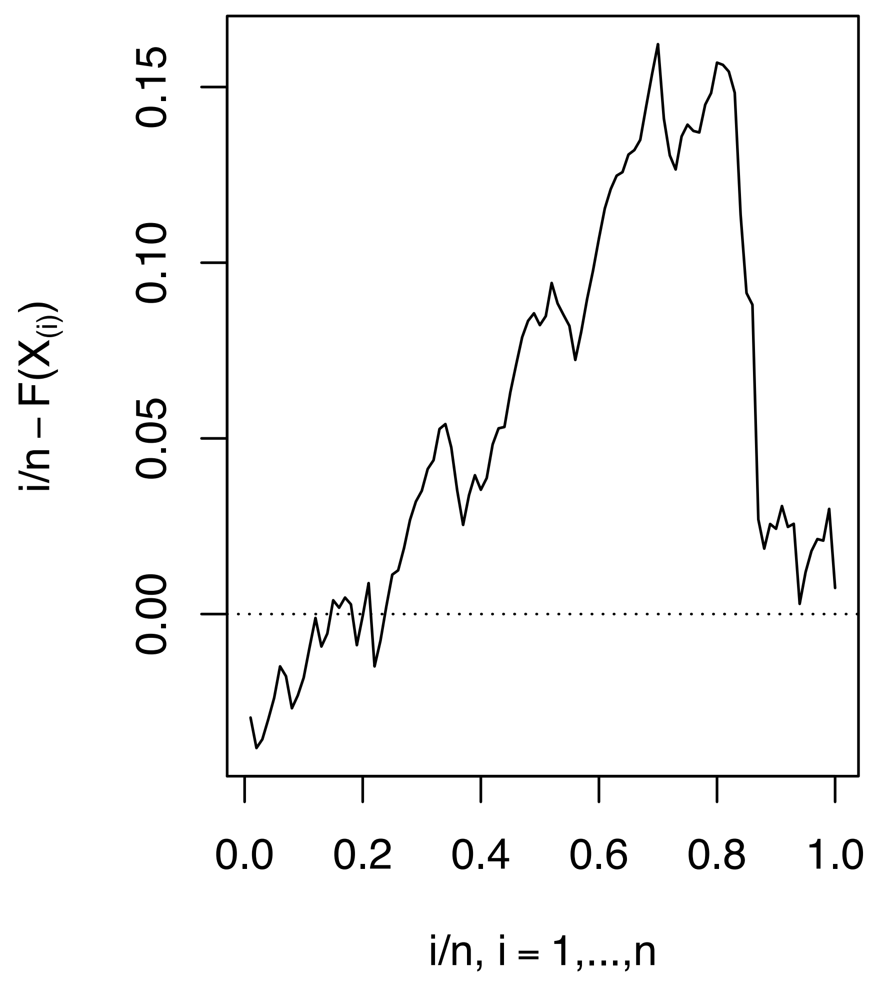
Exercise: Generate some (approximate) realizations of the Brownian bridge.



$\sup_{x \in \mathbb{R}} \sqrt{n}\,|\hat{F}_n(x) - F(x)|$

Basically, $\sqrt{n}[\hat{F}_n(X_{(i)}) - F(X_{(i)})]$, $i = 1, \dots, n$, acts like a Br. bridge for large n.





$B_{0.05}$ upper 0.05 quantile.

Exercise:

1. Run a simulation to get the 0.95 quantile of $\sup_{t\in[0,1]}|B_0(t)|$.
2. Check accuracy using the cdf of $\sup_{t\in[0,1]}|B_0(t)|$.
3. Compute $\sqrt{[\log(2/0.05)]/2}$.
4. Discuss.

The DKW bound is quite good.

DWK: $\hat{F}_n(x) \pm \frac{1}{\sqrt{n}} \underbrace{\sqrt{\frac{1}{2}\log\left(\frac{2}{0.05}\right)}}_{1.358}$

True asymptotic $(1-\alpha)100\%$ C. Band $\hat{F}_n(x) \pm \frac{1}{\sqrt{n}} \underbrace{B_{0.05}}_{1.342}$

Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be ind. rs with cdfs F and G, resp. Consider

$$H_0\colon F = G \text{ versus } H_1\colon F \neq G.$$

Two-sample Kolmogorov-Smirnov test

If $F = G$ the statistic

$$D_{nm} = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - \hat{G}_m(x)|$$

satisfies

$$P(\sqrt{mn/(m+n)}D_{nm} \leq x) \to 1 - 2\sum_{i=1}^{\infty}(-1)^{i+1}e^{-2i^2x^2}$$

as $n, m \to \infty$.

Compute D_{nm} as

$$D_{nm} = \max_{1 \leq i \leq n}[i/n - \hat{G}_m(X_{(i)})] \vee \max_{1 \leq j \leq m}[j/m - \hat{F}_n(Y_{(j)})].$$