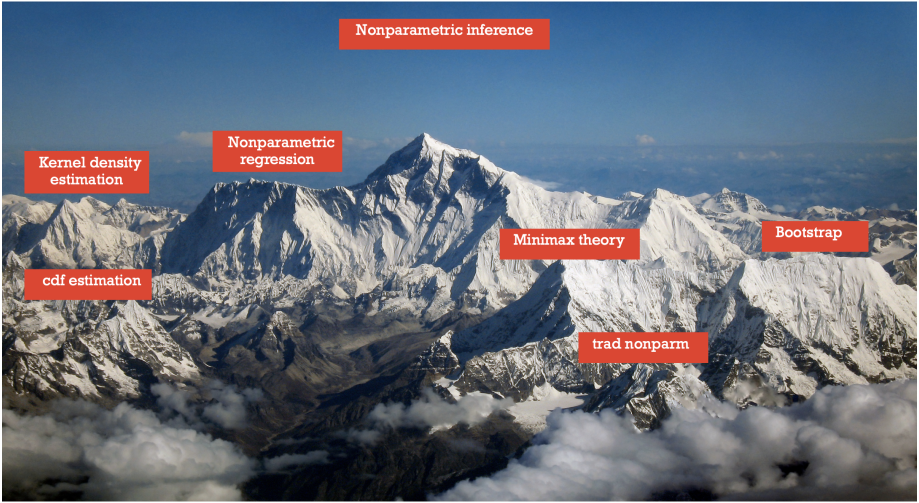
STAT 824 sp 2025 Lec 01 slides

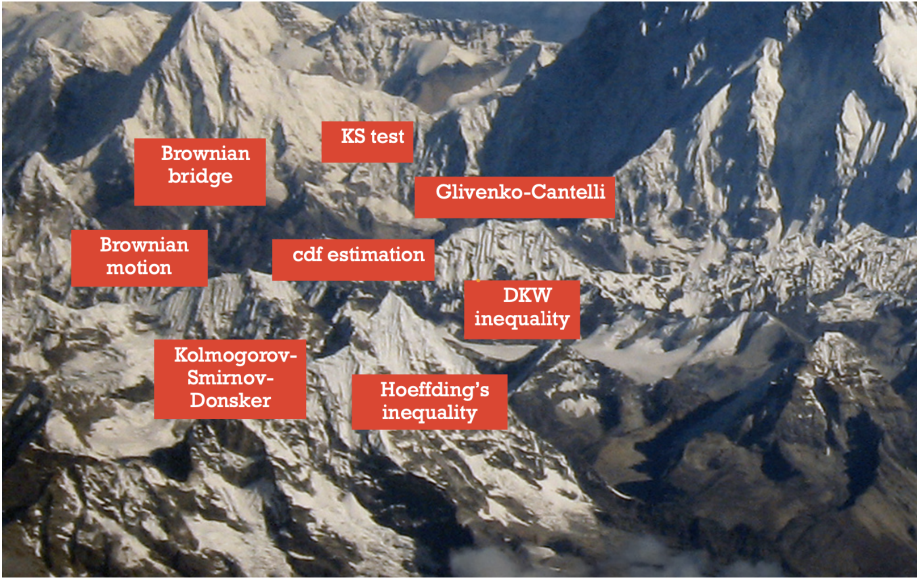
Estimating a cdf

Karl 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.


Nonparametric inference
Nonparametric regression
Kernel density estimation
Minimax theory
Bootstrap
cdf estimation
trad nonparm


KS test
Brownian bridge
Glivenko-Cantelli
Brownian motion
cdf estimation
DKW inequality
Kolmogorov-Smirnov-Donsker
Hoeffding's inequality

Empirical cdf

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

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

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

Glivenko-Cantelli Theorem

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

$$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \to 1$$

almost surely as $n \to \infty$.

Covered in STAT 810 and STAT 811.

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 each $x \in \mathbb{R}$ we have

$$\sqrt{n}(\hat{F}_n(x) - F(x)) \to \text{Normal}(0, F(x)[1 - F(x)]) \text{ in distribution}$$

as $n \to \infty$.

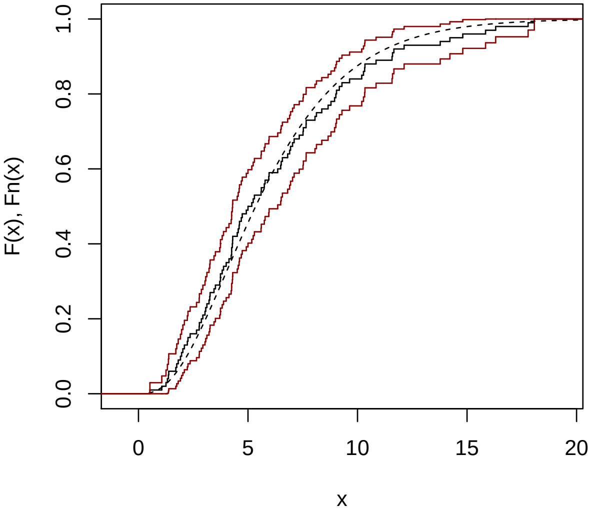
Exercise:

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

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

A $(1-\alpha) \times 100\%$

1. *confidence interval* for F at a point x is an interval $[L(x), U(x)]$ such that

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

2. *confidence band* for F over an interval $[a, b]$ is a region $\{(x, y) : L(x) \leq y \leq U(x), x \in [a, b]\}$ such that

$$P(L(x) \leq F(x) \leq U(x) \text{ for all } x \in [a, b]) \geq 1 - \alpha.$$

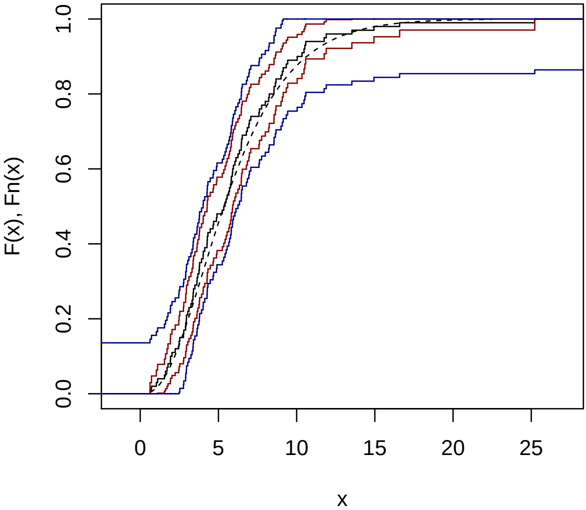
Dvoretzky-Kiefer-Wolfowitz inequality

If $X_1, \dots, X_n$ is a rs from a distribution with cdf F, then for any $\varepsilon > 0$ we have

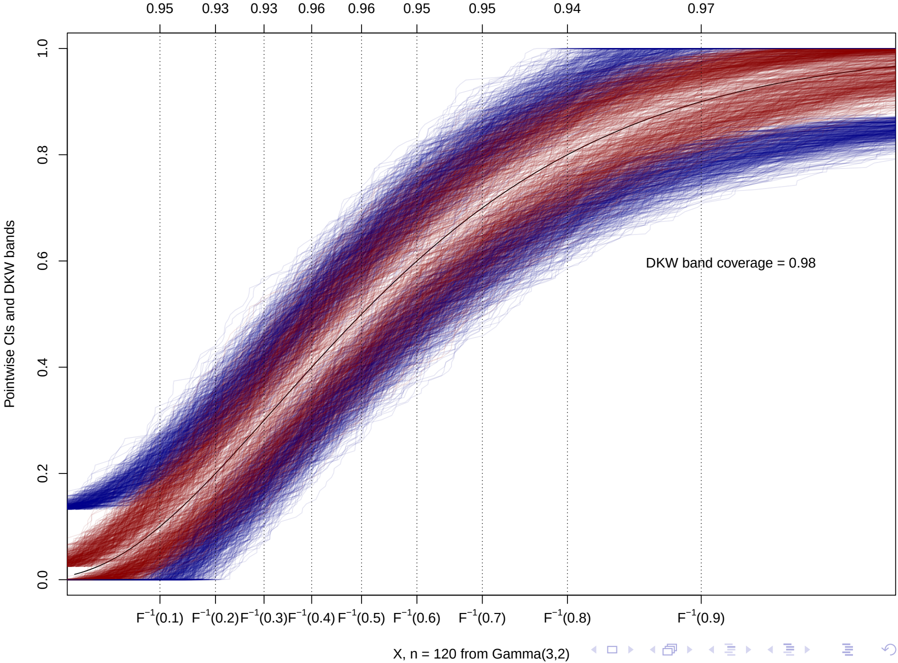
$$P\left(\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| \leq \varepsilon\right) \geq 1 - 2e^{-2n\varepsilon^2}.$$

Exercise:

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



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)

Hoeffding's inequality can give us a weaker inequality resembling the DKW.

Hoeffding's inequality

Let $Y_1, \dots, Y_n$ be independent zero-mean rvs such that $Y_i \in [a_i, b_i]$, $i = 1, \dots, n$. Then for any $\varepsilon > 0$ we have

$$P\Big(\sum_{i=1}^n Y_i \geq \varepsilon\Big) \leq \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n (a_i - b_i)^2}\right).$$

Exercise:

1. For $Y \in [a, b]$ with zero mean, show that $\log \mathbb{E}e^{tY} \leq t^2(b-a)^2/8$ for all t.
2. Prove Hoeffding's inequality.
3. Show that Hoeffding's gives $P(|\hat{F}_n(x) - F(x)| \leq \varepsilon) \geq 1 - 2e^{-2n\varepsilon^2}$ for each x.

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

Discuss: How to build confidence bands with above.

Wiener process or standard Brownian motion

A *Wiener process* B is a rf in the space $C[0,1]$ 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.

This is also called *standard Brownian motion (SBM)*.

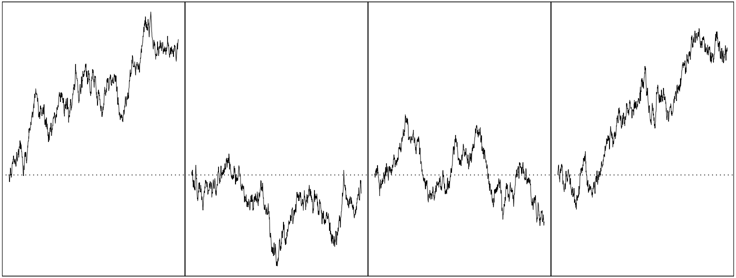
Generate an approximation to a standard Brownian motion

For each $n \geq 1$, let

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

Then B_n converges to B as $n \to \infty$ by a functional CLT called Donsker's Theorem.

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



Brownian bridge

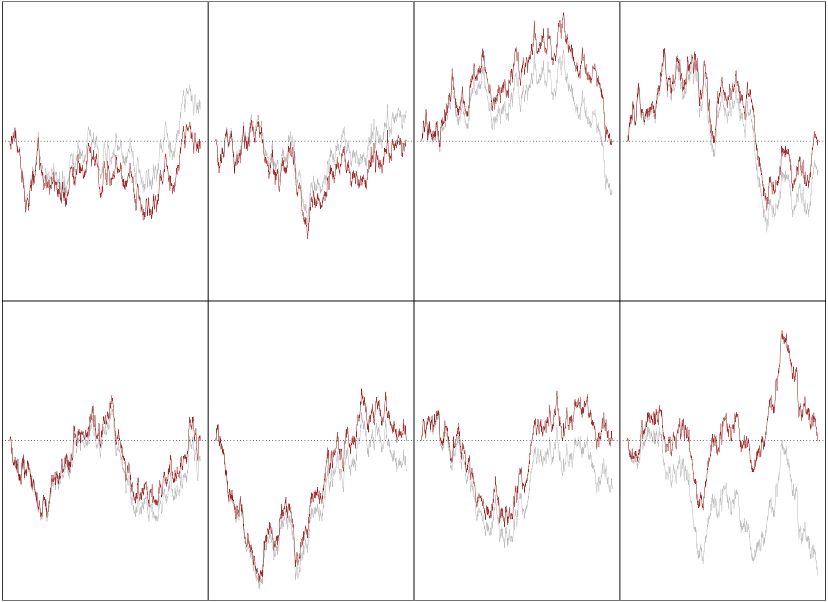
A *Brownian bridge* is the random function in $C[0, 1]$ given by

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

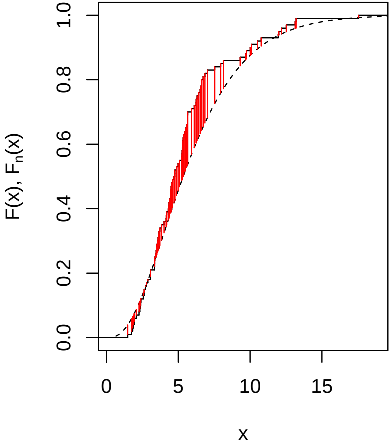
where B is a standard Brownian motion.

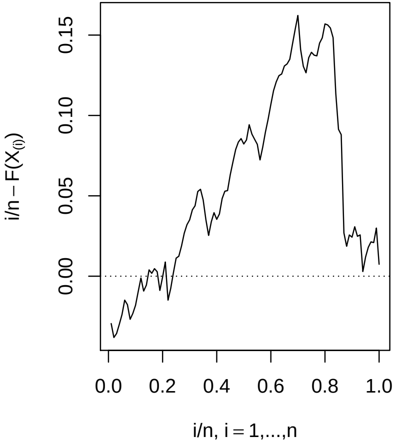
The "bridge" begins and ends at 0.

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



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





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.

Let $X_1, \dots, X_n$ and $Y_1, \dots, Y_m$ 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

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