The Empirical Distribution

STAT 510

Setup

$$X_1, X_2, \ldots X_n \sim F$$

- How to estimate F?
- How to estimate properties of F?

DEFN THE EMPIRICAL DISTRIBUTION FUNCTION, $\hat{F}_n$,

IS THE CDF THAT PUTS MASS $\frac{1}{n}$ AT EACH POINT X_i:

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

WHERE

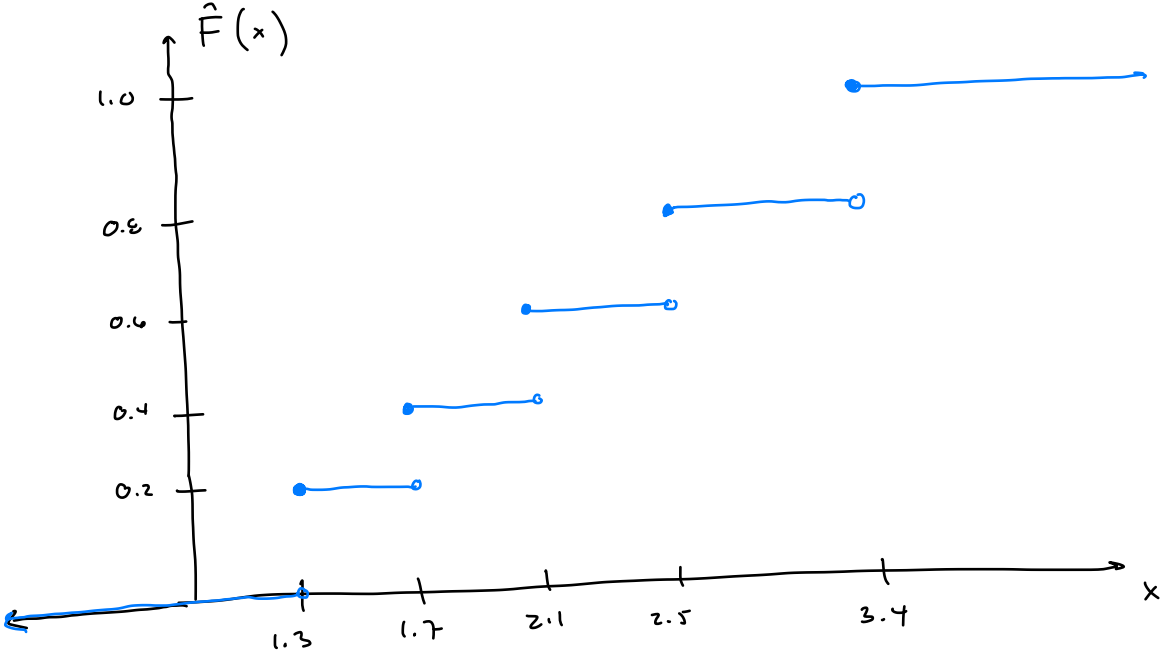
$$I(X_i \leq x) = \begin{cases} 1 & \text{IF } X_i \leq x \\ 0 & \text{IF } X_i > x \end{cases}$$

Example

Given data: 2.1, 1.3, 3.4, 2.5, 1.7

Sorted: 1.3, 1.7, 2.1, 2.5, 3.4

$$\hat{F}_n(x) = \begin{cases} 0 & x < 1.3 \\ 0.2 & 1.3 \le x < 1.7 \\ 0.4 & 1.7 \le x < 2.1 \\ 0.6 & 2.1 \le x < 2.5 \\ 0.8 & 2.5 \le x < 3.4 \\ 1 & x \ge 3.4 \end{cases}$$


$\hat{F}(x)$
1.0
0.8
0.6
0.4
0.2
1.3
1.7
2.1
2.5
3.4
x

Thm

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

For any fixed x

$$\mathbb{E}\left[\hat{F}_n(x)\right] = F(x)$$

$$\mathbb{V}\left[\hat{F}_n(x)\right] = \frac{F(x)(1-F(x))}{n}$$

$$\text{mse} = \frac{F(x)(1-F(x))}{n} \to 0$$

$$\hat{F}_n(x) \xrightarrow{P} F(x)$$

THM GLIVENKO - CANTELLI THEOREM

$$X_1, \ldots, X_n \sim F$$

THEN

$$\sup_x \left| \hat{F}_n(x) - F(x) \right| \xrightarrow{P} 0$$

Thm DKW INEQUALITY

$X_1, \ldots, X_n \sim F$

FOR ANY $\varepsilon > 0$

$$P\left(\sup_x \left| \hat{F}_n(x) - F(x) \right| > \varepsilon \right) \leq 2 e^{-2n\varepsilon^2}$$

Nonparametric $1-\alpha$ Confidence Band For F

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

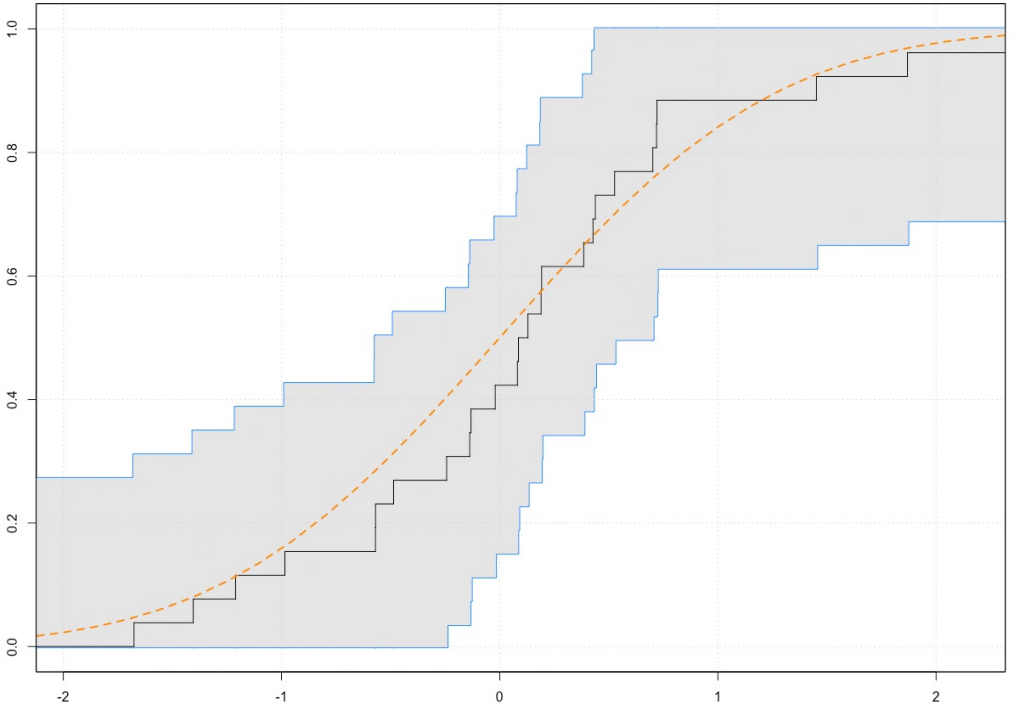
$$L(x) = \max\left\{\hat{F}_n(x) - \varepsilon_n,\ 0\right\}$$

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

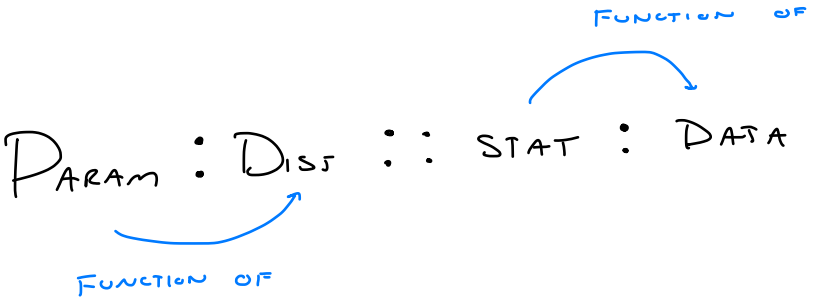
Then DKW gives

$$P\left(L(x) \le F(x) \le U(x) \text{ for all } x\right) \ge 1-\alpha.$$

$X_i \sim N(0,1) \quad n = 25$



IDEA



Statistical Functionals

Any function $T(F)$ of F.

- $\mu = \int x \, dF(x)$ MEAN
- $\sigma^2 = \int (x-\mu)^2 \, dF(x)$ VARIANCE
- $m = F^{-1}(1/2)$ MEDIAN

DEFN PLUG-IN ESTIMATOR OF $\theta = T(F)$

$$\hat{\theta}_n = T(\hat{F}_n)$$

"PLUG-IN" $\hat{F}_n$ FOR F!

DEFN

IF $T(F) = \int r(x)\, dF(x)$ FOR SOME $r(x)$

THEN T IS A LINEAR FUNCTIONAL.

$$T(aF + bG) = aT(F) + bT(G)$$

Thm

$$T(\hat{F}_n) = \int r(x)\, d\hat{F}_n(x) = \sum_{i=1}^{n} r(x_i) f(x_i) = \frac{1}{n} \sum_{i=1}^{n} r(x_i)$$

DEFN

$\tilde{F}_n$ DISCRETE

EACH POINT HAS PROB $1/n$

IF $T(\hat{F}_n) \approx N\left(T(F), \hat{se}^2\right)$

THEN A 95% "NORMAL" CI FOR $T(F)$ IS APPROX

$$T(\hat{F}_n) \pm 2\,\hat{se}$$

Example MEAN

Let $\mu = T(F) = \int x \, dF(x)$

$$\hat{\mu} = T(\hat{F}_n) = \int x \, d\hat{F}_n(x) = \frac{1}{n}\sum x_i = \bar{x} \quad \checkmark$$

We often know $\mathbb{V}[\bar{x}] = \sigma^2/n$

Example — Variance

Let $\sigma^2 = T(F) = \mathbb{V}[X] = \int (x-\mu)^2 \, dF(x)$

$$= \int x^2 \, dF(x) - \left(\int x \, dF(x) \right)^2$$

$$\hat{\sigma}^2 = \int x^2 \, d\hat{F}_n(x) - \left(\int x \, d\hat{F}_n(x) \right)^2$$

$$= \frac{1}{n} \sum x_i^2 - \left(\frac{1}{n} \sum x_i \right) = \frac{1}{n} \sum (x_i - \bar{x})^2$$

$\mathbb{V}[\hat{\sigma}^2] = ???$

$\neq \frac{1}{n-1} \sum (x_i - \bar{x})^2 = S_n^2$