STA355 Lecture 3

3 tasks in statistical inference.

- (Point) estimation.

Data $X \longrightarrow \hat{\theta}_n(X)$ (indp of underlying θ).

Want $\hat{\theta}_n$ to be close to θ.

(Extension: estimate a functional of the underlying param)

$T(\theta)$, $\hat{T}_n$

Criteria of evaluation:

$$MSE(\hat{\theta}_n) = \mathbb{E}_\theta\left[\|\hat{\theta}_n - \theta\|^2\right]$$ (for some norm $\|\cdot\|$).

- expectation w.r.t. model $\mathbb{P}_\theta$.
- Randomness comes only from $\hat{\theta}_n$ (depending on X, and potential randomness from algorithm).

Consider Euclidean norm.

$$MSE(\hat{\theta}_n) = \underbrace{\|\mathbb{E}_\theta[\hat{\theta}_n] - \theta\|_2^2}_{\text{bias}(\hat{\theta}_n)^2} + \text{var}_\theta(\hat{\theta}_n)$$

$$\text{Var}_\theta(Z) = \mathbb{E}_\theta[Z^2] - (\mathbb{E}_\theta[Z])^2.$$

We say $\hat{\theta}_n$ is unbiased when $\mathrm{bias}(\hat{\theta}_n) = 0 \quad \forall \theta$.

e.g. $X_1, X_2, \cdots, X_n \overset{iid}{\sim} \mathbb{P}$. $\quad \mathbb{E}_{\mathbb{P}}[|X_1|^2] < +\infty$

$\mathbb{P}$ defined on $\mathbb{R}$.

$\mu(\mathbb{P}) := \mathbb{E}_{\mathbb{P}}[X_1]$.

$\sigma^2(\mathbb{P}) := \mathrm{var}_{\mathbb{P}}(X_1)$.

$\hat{\mu}_n := \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)^2$

are unbiased estimators of $\mu(\mathbb{P})$ and $\sigma^2(\mathbb{P})$.

$\tilde{\mu}_n = X_1$ is also unbiased.

Def. "Consistency". $\quad X_1, X_2, \cdots X_n \cdots \overset{iid}{\sim} \mathbb{P}_\theta : \theta \in \Theta$

$\forall n$, $\hat{\theta}_n$ computed on $(X_1, X_2, \cdots, X_n)$

we say $\hat{\theta}_n$ is consistent estimator for θ

if $\quad \hat{\theta}_n \xrightarrow{P} \theta$.

Def. We say $\hat{\theta}_n$ is asymptotically normal if

$$\frac{\hat{\theta}_n - \theta}{\sqrt{\mathrm{var}_\theta(\hat{\theta}_n)}} \xrightarrow{d} N(0,1) \quad (\text{under } P_\theta)$$

→ generally unknown, but can be estimated.

Task 2: confidence set (interval)

For $\alpha \in (0,1)$, $(1-\alpha)$ confidence set for parameter θ

we say C_n

if $\mathbb{P}_\theta(\underbrace{\theta}_{\text{Deterministic but unknown}} \in \underbrace{C_n}_{\text{is a random set that depends on data}}) \geq 1-\alpha \quad \forall \theta \in \Theta$.

e.g. $X_1, X_2, \cdots X_n \overset{iid}{\sim} Ber(\theta) \qquad \theta \in [0,1]$.

CI for true parameter θ.?

One (not so good) solution: Chebyshev ineq.

under $\mathbb{P}_\theta$, $\quad \hat{\theta}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ satisfies

$$\mathbb{E}_\theta\left[|\hat{\theta}_n-\theta|^2\right] = \frac{Var_\theta(X_1)}{n} = \frac{\theta(1-\theta)}{n} \leq \frac{1}{4n}$$

By Chebyshev,

$$\mathbb{P}_\theta\left(|\hat{\theta}_n-\theta| \geq t\right) \leq \frac{1}{4nt^2} \leq \alpha.$$

Solve for t, $\quad C_n = \left(\hat{\theta}_n - \frac{1}{2\sqrt{n\alpha}},\ \hat{\theta}_n + \frac{1}{2\sqrt{n\alpha}}\right).$

$$\mathbb{P}_\theta\left(\theta \in C_n\right) \geq 1-\alpha \quad \left(\begin{matrix}\forall n\\ \forall \theta\end{matrix}\right)$$

(Usually, we want C_n to be as small as possible),

Is C_n a good one?

- $1/\sqrt{n}$ dependence on n
- bad dependence on α.

Detour: Hoeffding inequality:

$X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathbb{P}$, $\quad X_i \in [a,b]$ a.s. $a<b$, $\quad \forall t>0$

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}[X_1]\right| \geq t\right) \leq 2\exp\left(-\frac{2nt^2}{(b-a)^2}\right).$$

Better CI via Hoeffding's ineq. $(a=0, b=1)$

Want: $\mathbb{P}_\theta\left(|\hat{\theta}_n-\theta| \geq t\right) \leq \alpha$

Solve for $t = \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)}$.

We get $(1-\alpha)$ CI :

$$\left[\hat{\theta}_n - \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)},\ \hat{\theta}_n + \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)}\right].$$

CI via asymptotic normality.

Assuming $\dfrac{\hat{\theta}_n - \theta}{\hat{se}_n} \xrightarrow{d} N(0,1)$.

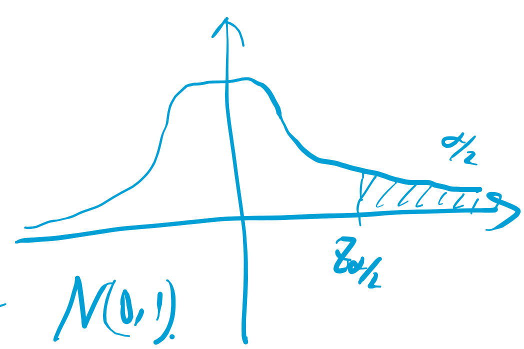
Assumed to be computable from data.

CI construction :

$$CI_n = \left(\hat{\theta}_n - z_{\alpha/2}\,\hat{se}_n,\ \hat{\theta}_n + z_{\alpha/2}\,\hat{se}_n\right).$$

Thm : under above setup. $\forall \theta$

$$\mathbb{P}_\theta(\theta \in CI_n) \to 1-\alpha$$



$z_{\alpha/2}$: quantile of standard normal.

$z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$. Φ is cdf of $N(0,1)$.

Asymptotic CI, valid under $n \to +\infty$.

Proof. $Z_n = \dfrac{\hat{\theta}_n - \theta}{\hat{se}_n}$, we have $Z_n \xrightarrow{d} N(0,1)$.

So we have cdf convergence.

$$\mathbb{P}_\theta(\theta \in C_n) = \mathbb{P}\left(-z_{\alpha/2} \leq \frac{\hat{\theta}_n - \theta}{\hat{se}_n} \leq z_{\alpha/2}\right)$$

$$= \mathbb{P}(Z_n \leq z_{\alpha/2}) - \mathbb{P}(Z_n < -z_{\alpha/2})$$

$$\longrightarrow \Phi(z_{\alpha/2}) - \Phi(-z_{\alpha/2}) = 1-\alpha.$$

Comparison in Bernoulli example

$$C_n = \left[\hat{\theta}_n - \sqrt{\frac{1}{2n}\log(2/\alpha)},\ \hat{\theta}_n + \sqrt{\frac{1}{2n}\log(2/\alpha)}\right].$$

$$C_n' = \left[\hat{\theta}_n - \hat{se}_n z_{\alpha/2},\ \hat{\theta}_n + \hat{se}_n z_{\alpha/2}\right].$$

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{d} N(0, \theta(1-\theta)) \quad \text{by CLT}$$

$se_n = \sqrt{\dfrac{\theta(1-\theta)}{n}}$, estimated by data $\hat{se}_n = \sqrt{\dfrac{\hat{\theta}_n(1-\hat{\theta}_n)}{n}}$

(same asymptotic normality by Slutzky).

$$\frac{\hat{\theta}_n - \theta}{se_n} \cdot \frac{se_n}{\hat{se}_n} \xrightarrow{d} N(0,1).$$

length of $C_n = \sqrt{\frac{2}{n}\log\left(\frac{2}{\alpha}\right)}$.

length of $C_n' = 2 z_{\alpha/2} \cdot \widehat{se}_n = 2 z_{\alpha/2} \sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n}} \approx \sqrt{\frac{\theta(1-\theta)}{n}}$

of order $\sqrt{\log(1/\alpha)}$.

when θ close to 0 or 1, C_n' shorter than C_n.

but you paid finite-sample validity.

Hypothesis testing.

Null hypothesis $H_0: \theta \in \Theta_0$

Alternative hypothesis $H_1: \theta \in \Theta_1$

Assume that $\Theta_0 \cap \Theta_1 = \emptyset$.

Goal: output a decision $T = T(X) \in \{0, 1\}$

to decide which set θ lives in.

Type I error: $\mathbb{P}_\theta(T(X)=1)$ for $\theta \in \Theta_0$

Type II error: $\mathbb{P}_\theta(T(X)=0)$ for $\theta \in \Theta_1$.

Testing is asymmetric: level-α test:

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(T(X)=1) \le \alpha$$

(and also asymptotic version)

Problem: given $X_1, X_2, \cdots, X_n \sim P$ with cdf F on $\mathbb{R}$.

Can we estimate F accurately.

Solution: empirical cdf.

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

Fact. $\cdot\ \mathbb{E}[\hat{F}_n(x)] = F(x)$ (unbiased)

$\cdot\ \mathrm{var}(\hat{F}_n(x)) = \dfrac{F(x)(1-F(x))}{n} = MSE(\hat{F}_n(x))$

$\cdot\ \sqrt{n}(\hat{F}_n(x) - F(x)) \xrightarrow{d} N(0, F(x)(1-F(x)))$ $(\forall x \in \mathbb{R})$.

Indeed, $\left(\sqrt{n}(\hat{F}_n(x) - F(x)) : x \in \mathbb{R}\right) \xrightarrow{d}$ Gaussian Process.

Thm. (Glivenko-Cantelli)

$$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \xrightarrow{P} 0.$$

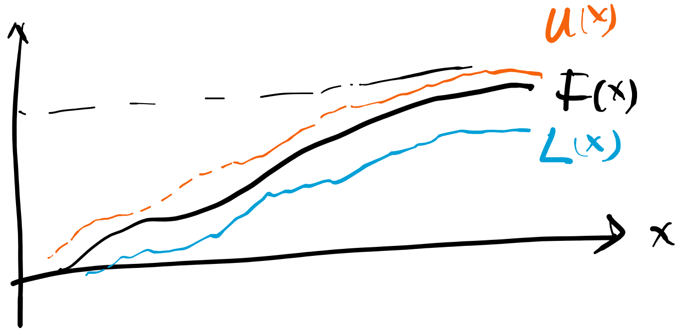
Thm (DKW inequality)

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

(same bound as Hoeffding, but here we can take supremum of $x\in\mathbb{R}$).

Confidence set for nonparametric estimation:

"confidence band"



$$\mathbb{P}\left(\forall x\in\mathbb{R},\ L(x)\le F(x)\le U(x)\right)\ge 1-\alpha.$$

By DKW ineq. construct

$$L_n(x) := \hat{F}_n(x) - \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)}$$

$$U_n(x) = \hat{F}_n(x) + \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)}.$$

(L_n, U_n) is $(1-\alpha)$ confidence band for F.

How about functionals of cdf

$T(F)$ functional.

e.g. $T(F) = \int_{\mathbb{R}} x \, dF(x) = \mathbb{E}_{X \sim F}[X]$.

e.g. $\mathrm{var}(F) = \int_{\mathbb{R}} x^2 \, dF(x) - \left(\int_{\mathbb{R}} x \, dF(x) \right)^2$

e.g. $\mathrm{median}(F) = \inf \{ z \in \mathbb{R} : F(z) \geq 1/2 \}$.

Plugin principle: estimate $T(F)$ by $T(\hat{F}_n)$.

"empirical estimator"!

- Linear functionals.

We consider $T(F) = \int r(x) \, dF(x)$

for some function r

(in functional analysis, linear functionals are broader than this.

but let's first focus on this)

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

$$= \frac{1}{n} \sum_{i=1}^{n} r(X_i)$$

Fact. $\mathbb{E}[T(\hat{F}_n)] = T(F)$

$$\mathrm{var}(T(\hat{F}_n)) = \frac{1}{n} \mathrm{var}(r(X_1)).$$

$$\sqrt{n}\left(T(\hat{F}_n) - T(F)\right) \xrightarrow{d} \mathcal{N}\left(0,\ \mathrm{var}(r(X_1))\right)$$

eg. $\mu = \mathbb{E}[X]$ $\quad r(x) = x.$

eg. $\sigma^2 = \mathrm{var}(X) = \mathbb{E}[X^2] - \mu^2$ $\quad r(x) = x^2$

then subtract $\hat{\mu}_n^2$.

eg. "Skewness"

$$\kappa = \frac{\mathbb{E}[(X-\mu)^3]}{\sigma^3} \qquad r(x) = (x-\mu)^3.$$

empirical estimator:

$$\hat{\kappa}_n = \frac{1}{\hat{\sigma}_n^3} \cdot \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)^3$$

eg. of nonlinear functionals.

- Quantile (assuming F is strictly increasing)

For $0 < p < 1$, $\quad T(F) = F^{-1}(p) = \inf\{x : F(x) \geq p\}$

$$T_p(\hat{F}_n) = \inf\{x : \hat{F}_n(x) \geq p\}$$

"sample quantile"

e.g. $X \sim \mathrm{Unif}([0,1])$. $\quad F(x) = x$.

$$\forall p \in (0,1),\ \mathbb{P}\left(T_p(\hat{F}_n) \le p - \varepsilon\right) = \mathbb{P}\left(\hat{F}_n(p-\varepsilon) \ge p\right).$$

$$\le \exp(-2n\varepsilon^2)$$

$$\mathbb{P}\left(T_p(\hat{F}_n) \ge p + \varepsilon\right) \le \exp(-2n\varepsilon^2).$$

So we get $\quad \left|T_p(\hat{F}_n) - T_p(F)\right| \le \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)}$

w.p. $1-\alpha$.

This holds true in general.

e.g. for F strictly increasing with uniform lower bound on $F'(=\text{pdf})$ with some interval.

e.g. $p > 1 - \frac{1}{n}$. $\quad T_p(\hat{F}_n) = \max\limits_{1 \le i \le n} X_i$, $\quad \mathbb{E}\left[T_p(\hat{F}_n)\right] \approx F^{-1}\left(1-\frac{1}{n}\right)$

while $T_p(F)$ can be much larger.

e.g. $f(x_0) = F'(x_0) =: T(F)$.

$x_0 \in \mathbb{R}$ $\quad T(\hat{F}_n) = \hat{F}_n'(x_0) = 0 \quad$ (a.s.)