Ref :

[2004][L. Wasserman] All of Statistics.

Chap 6 Models, Statistical Inference and Learning.

6.1 Intro.

Q: Given a sample $X_1, \cdots, X_n \sim F$, how do we infer F ?

6.2 Parametric and Nonparametric Model.

Def

A $\boxed{\text{statistical model}}$ $\boxed{\mathfrak{F}}$ is a set of distributions (or densities or regression fun.)

A $\boxed{\text{parametric model}}$ is a model $\mathfrak{F}$ parametrizable by finitely many parameters.

Notation : (parametric model)

$\mathfrak{F} = \{ f(x;\theta) \mid \theta \in \Theta \}$.

$\boxed{\Theta}$ is called the $\boxed{\text{parameter space}}$.

Def

A parameter that we are not interested in is called a $\boxed{\text{nuisance parameter}}$.

A $\boxed{\text{nonparametric model}}$ is a model not parametrizable by finitely many parameters.

Def

$\boxed{\mathfrak{F}_{SOB}} := \{ f \mid \int (f''(x))^2 dx < \infty \}$ is called the $\boxed{\text{Sobolev space}}$. (Intuitively, functions not "too wiggly").

Def

Any function $\boxed{T(F)}$ of a CDF F is called a $\boxed{\text{statistical functional}}$.

e.g. mean and variance are statistical functionals.

Def

When our data look like $(X_1, Y_1), \cdots, (X_n, Y_n)$,

X is called $\boxed{\text{predictor}}$ / $\boxed{\text{regressor}}$ / $\boxed{\text{feature}}$ / $\boxed{\text{indep. variable}}$.

Y is called $\boxed{\text{outcome}}$ / $\boxed{\text{response variable}}$ / $\boxed{\text{dep. variable}}$.

$\boxed{r(x)} := \mathbb{E}[Y \mid X = x]$ is called the $\boxed{\text{regression function}}$. i.e. $r = \mathbb{E}[Y|X]$.

The goal of predicting the Y-value based on the X-value is called $\boxed{\text{prediction}}$.

If Y is discrete, it's called $\boxed{\text{classification}}$.

If Y is cont. (i.e. $\mathbb{R}$-values), then it's called $\boxed{\text{regression}}$ or $\boxed{\text{curve estimation}}$.

Rmk :

$Y = r(X) + \varepsilon$, w/ $\mathbb{E}[\varepsilon] = 0$.

Notation : pdf

$\mathfrak{F} = \{ f(x;\theta) \mid \theta \in \Theta \}$ a parametric model.

$\boxed{P_\theta(X \in A)} := \int_A f(x;\theta)\, dx$.

$\boxed{\mathbb{E}_\theta(r(X))} := \int r(x) f(x;\theta)\, dx$.

$\boxed{\mathbb{V}_\theta}$: variance.

i.e. the subindex θ indicates the para. instead of averaging over θ.

6.3 Fundamental Concepts in Inference.

6.3.1 Pt Estimation.

$\boxed{\text{Point estimation}}$ = providing a single "best guess" of some quantity of interest.

Quantity of interest could be ① a parameter ② a CDF ③ a pdf ④ a regression fun. ⑤ a prediction.

Def

$X_1, \cdots, X_n$: iid from a CDF F.

A $\boxed{\text{point estimator}}$ $\boxed{\hat{\theta}_n}$ of a parameter θ (fixed but unknown) is a function of $X_1, \cdots, X_n$. i.e. $\hat{\theta}_n = g(X_1, \cdots, X_n)$, for some g.

The $\boxed{\text{bias}}$ of $\hat{\theta}_n$ is $\boxed{\text{bias}(\hat{\theta}_n)} = \mathbb{E}_\theta(\hat{\theta}_n) - \theta$.

$\hat{\theta}_n$ is called $\boxed{\text{unbiased}}$ if $\text{bias}(\hat{\theta}_n) = 0$.

Rmk :

Unbiasedness used to receive much attention, but is considered less important these days.

Def

$\hat{\theta}_n$ is $\boxed{\text{consistent}}$ if $\hat{\theta}_n \xrightarrow{P} \theta$.

Def

The dist. of $\hat{\theta}_n$ is called the $\boxed{\text{sampling distribution}}$.

The standard deviation of $\hat{\theta}_n$ is called the $\boxed{\text{standard error}}$, denoted $\boxed{se}$. i.e. $se = \sqrt{\mathrm{Var}_\theta(\hat{\theta}_n)}$.

Rmk :

$\because F$ is unknown $\therefore se$ is also unknown, but we can sometimes estimate it.

The estimated standard error is denoted $\boxed{\hat{se}}$.

Def

The $\boxed{\text{mean squared error}}$ of $\hat{\theta}_n$, denoted $\boxed{MSE}$, is

$MSE = \mathbb{E}_\theta[(\hat{\theta}_n - \theta)^2]$.

Thm 6.9

$$MSE = bias^2(\hat{\theta}_n) + Var_\theta(\hat{\theta}_n) = bias^2 + se^2.$$

Thm 6.10

$bias \to 0$ and $se \to 0$ as $n \to \infty$.

Then $\hat{\theta}_n$ is consistent.

i.e. $\hat{\theta}_n \xrightarrow{P} \theta$.

<Pf>

By Thm 6.9, $\hat{\theta}_n \xrightarrow{qm} \theta$. $\Rightarrow$ $\hat{\theta}_n \xrightarrow{P} \theta$. (Thm 5.4). #

Def

An estimator $\hat{\theta}_n$ is $\boxed{\text{asymptotically Normal}}$ if

$$\frac{\hat{\theta}_n - \theta}{se} \rightsquigarrow N(0,1), \text{ written } \boxed{\hat{\theta}_n \approx N(\theta, se^2)}.$$

6.3.2 Confidence Sets.

Def

A $\boxed{(1-\alpha) \text{ confidence interval}}$ for a parameter θ is an interval $C_n = (a,b)$, w/ $a = a(X_1, \cdots, X_n)$ and $b = b(X_1, \cdots, X_n)$ s.t.

$$P_\theta(\theta \in C_n) \geq 1-\alpha, \ \forall \theta \in \Theta.$$

$(1-\alpha)$ is called the $\boxed{\text{coverage}}$ of the confidence interval.

Rmk

① When θ is a vector, we use the terminology $\boxed{\text{confidence set}}$ instead.

② Look at the intuitive (better) interpretation of confidence interval on P95.

Thm 6.16 (Normal-based confidence interval)

$\hat{\theta}_n \approx N(\theta, \hat{se}^2)$.

Φ: CDF of $N(0,1)$.

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

$C_n := (\hat{\theta}_n - z_{\alpha/2}\hat{se}, \hat{\theta}_n + z_{\alpha/2}\hat{se})$.

Then $P_\theta(\theta \in C_n) \to 1-\alpha$, as $n \to \infty$.

6.3.3 Hypothesis Testing.

$\boxed{\text{Hypothesis testing}}$ = starting w/ a default theory, called $\boxed{\text{null hypothesis}}$, and "test" if the data provide sufficient evidence to reject the theory. If not, "retain" the null hypothesis.

Rmk

① The null hypothesis is usually denoted $\boxed{H_0}$.

② We usually have a "plan B" for H_0, called the $\boxed{\text{alternative hypothesis}}$ and denoted $\boxed{H_1}$, which we retain when H_0 is rejected.

Def

① An interval C_n is called a $\boxed{\text{pointwise asymptotic}}$ $\boxed{(1-\alpha) \text{ confidence interval}}$ if

$$\lim_{n\to\infty} P_\theta(\theta \in C_n) \geq 1-\alpha, \ \forall \theta \in \Theta.$$

② An interval C_n is called a $\boxed{\text{uniform asymptotic}}$ $\boxed{(1-\alpha) \text{ confidence interval}}$ if

$$\lim_{n\to\infty} \inf_{\theta \in \Theta} P_\theta(\theta \in C_n) \geq 1-\alpha.$$

Rmk

The interval in Thm 6.16 is a ptwise asymptotic confidence interval.

Chap 7 Estimating the CDF and Statistical Functionals.

7.1 The Empirical Distribution Function.

Def

$X_1, \cdots, X_n \sim F$ iid, where F: CDF on $\mathbb{R}$.

The $\boxed{\text{empirical distribution function}}$ $\boxed{\hat{F}_n}$ is the CDF putting $1/n$ mass on each X_i.

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

where $$I(X_i \leq x) = \begin{cases} 1, & \text{if } X_i \leq x \\ 0, & \text{if } X_i > x. \end{cases}$$

Thm 7.3 Fix $x \in \mathbb{R}$. Then

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

$Var[\hat{F}_n(x)] = F(x)(1-F(x))/n = MSE \to 0$ as $n \to \infty$.

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

$\boxed{\text{Thm 7.4 (The Glivenko-Cantelli Thm)}}$

$X_1, \cdots, X_n \sim F$ iid.

Then $\sup_x |\hat{F}_n(x) - F(x)| \xrightarrow{P} 0$, as $n \to \infty$.

$\boxed{\text{Thm 7.5 (The Dvoretzky-Kiefer-Wolfowitz (DKW) inequality)}}$

$X_1, \cdots, X_n \sim F$ iid.

Then, $\forall \varepsilon > 0$, as $n \to \infty$,

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

From DKW inequality, we can construct:

$\boxed{\text{Prop (Nonparametric } 1-\alpha \text{ confidence band for } F)}$

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

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

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

Then, $\forall$ CDF F,

$$P(L(x) \le F(x) \le U(x)) \ge 1-\alpha.$$

7.2 Statistical Functionals.

Recall that a statistical functional $T(F)$ is any function of F. (F: CDF).

Def (plug-in method)

The $\boxed{\text{plug-in estimator}}$ of $\theta = T(F)$ is

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

i.e. plug in the $\hat{F}_n$ for the unknown F.

Def

If $T(F) = \int r(x)\, dF(x)$ for some fun. $r(x)$, then T is called a $\boxed{\text{linear functional}}$.

Rmk:

If T is a linear functional, then $T(aF + bG) = aT(F) + bT(G)$. (Thus the name).

$\boxed{\text{Thm 7.9}}$

The plug-in estimator of a linear functional $T(F) = \int r(x)\, dF(x)$ is

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

Rmk:

A general method for finding the standard error $\widehat{se}$ of $T(\hat{F}_n)$ will be introduced in the next chap. St P3

Def

The $\boxed{\text{skewness}}$ of a random variable X w/ mean μ and variance σ^2 is

$$\boxed{K} = \frac{\mathbb{E}[(X-\mu)^3]}{\sigma^3} = \frac{\int (x-\mu)^3\, dF(x)}{\left(\int (x-\mu)^2\, dF(x)\right)^{3/2}}.$$

Thus, the plug-in estimator is

$$\hat{K} = \frac{\frac{1}{n}\sum_i (X_i - \hat{\mu})^3}{\hat{\sigma}^3}.$$

Def

F: a CDF.

$p \in (0,1)$.

The $\boxed{\text{pth quantile}}$ is, denoted $\boxed{F^{-1}(p)}$,

$$\inf\{x \mid F(x) \ge p\}.$$

Thus, the plug-in estimator is

$$\hat{F}_n^{-1}(p) = \inf\{x \mid F_n(x) \ge p\},$$

called the $\boxed{\text{pth sample quantile}}$.

7.3 Bibliographic Remarks.

For further reading, look at "empirical process".

Supplements:

Ref: All of Nonpara. Stat. S2.4.

Empirical Prob. Dist.

Def

P: prob. measure.

$X_1, \cdots, X_n \sim \mathbb{P}$: iid sample.

The $\boxed{\text{empirical probability distribution}}$ $\boxed{\hat{\mathbb{P}}_n}$ is defined by $\hat{\mathbb{P}}_n(A) = \dfrac{\#\{X_i \mid X_i \in A\}}{n}$.

Let $\mathcal{A}$ be a class of subsets of Ω.

We are going to use,

$\mathbb{P}\left(\sup_{A \in \mathcal{A}} |\hat{\mathbb{P}}_n(A) - \mathbb{P}(A)| > \varepsilon\right)$ to quantify the diff. b/w $\mathbb{P}$ and $\hat{\mathbb{P}}_n$ on $\mathcal{A}$.

The Vapnik-Chervonenkis (VC) Theory:

Def

$\mathcal{A}$: a class of subsets of Ω.

$R = \{x_1, \cdots, x_n\} \subseteq \Omega$.

Define

$$\boxed{N_{\mathcal{A}}(R)} = \#\{R \cap A \mid A \in \mathcal{A}\}.$$

R is said to $\boxed{\text{be shattered by } \mathcal{A}}$ if

$N_{\mathcal{A}}(R) = 2^n$. i.e. $\{R \cap A \mid A \in \mathcal{A}\} = 2^R$.

The $\boxed{\text{shatter coefficient}}$ is defined by

$$\boxed{s(\mathcal{A}, n)} = \max_{R \in \mathcal{F}_n} N_{\mathcal{A}}(R),$$

where $\boxed{\mathcal{F}_n} := \{R \subseteq \Omega \mid \#(R) = n\}$.

Thm 2.41 (Vapnik and Chervonenkis, 1971)

$\forall\, \mathbb{P}, n$ and $\varepsilon > 0$,

$$\mathbb{P}\left(\sup_{A \in \mathcal{A}} |\hat{\mathbb{P}}_n(A) - \mathbb{P}(A)| > \varepsilon\right) \le 8\, s(\mathcal{A}, n)\, e^{-n\varepsilon^2/32}.$$

Def

$\mathcal{A}$: a class of subsets of Ω.

Define $\boxed{VC(\mathcal{A})}$, the $\boxed{\text{VC dimension of } \mathcal{A}}$, by

① if $s(\mathcal{A}, n) = 2^n, \forall n$, $VC(\mathcal{A}) := \infty$;

② o.w. $VC(\mathcal{A}) := \max\{k \in \mathbb{N} \mid s(\mathcal{A}, k) = 2^k\}$.

Thm 2.43

$VC(\mathcal{A}) = v < \infty$,

Then $s(\mathcal{A}, n) \le n^v + 1$.

Thus,

$$\mathbb{P}\left(\sup_{A \in \mathcal{A}} |\hat{\mathbb{P}}_n(A) - \mathbb{P}(A)| > \varepsilon\right) \le 8 \cdot (n^v + 1)\, e^{-n\varepsilon^2/32}. \quad (2.44)$$

Ex:

Let $\mathcal{A} = \{(-\infty, x] \mid x \in \mathbb{R}\}$.

It's clear $VC(\mathcal{A}) = 1$ since no set of the form $\{x, y\}$ can be shattered by $\mathcal{A}$.

Thus

$$\mathbb{P}\left(\sup_{A \in \mathcal{A}} |\hat{\mathbb{P}}_n(A) - \mathbb{P}(A)| > \varepsilon\right) \le 8(n+1) \cdot e^{-n\varepsilon^2/32}.$$

Notice LHS $= \mathbb{P}\left(\sup_x |F(x) - \hat{F}_n(x)| > \varepsilon\right)$.

This band is way looser than that given by DKW inequality.

Ref: [L. Wasserman] All of Nonpara. Stat. Sec 2.3.

Q: When do we have $T(\hat{F}_n) \to T(F)$?

Influence Functions.

Def

T : statistical functional.

① The $\boxed{\text{Gateaux derivative}}$ of F in the direction G is defined by

$F \;\varepsilon\; 1-\varepsilon\; G$

$\varepsilon G + (1-\varepsilon) F$

$$\boxed{L_F(G)} := \lim_{\varepsilon \to 0} \frac{T((1-\varepsilon)F + \varepsilon G) - T(F)}{\varepsilon}.$$

② If $G = \delta_x$, the pt mass at x, we write

$\boxed{L_F(x)}$ for $L_F(\delta_x)$. i.e. $L_F(x) := L_F(\delta_x)$.

$L_F(x)$ is called the $\boxed{\text{influence function}}$ of F.

Explicitly, $L_F(x) = \lim_{\varepsilon \to 0} \dfrac{T((1-\varepsilon)F + \varepsilon\delta_x) - T(F)}{\varepsilon}$.

③ The $\boxed{\text{empirical influence function}}$ is defined by

$\boxed{\hat{L}(x)} = L_{\hat{F}_n}(x)$.

i.e. $$\hat{L}(x) := \lim_{\varepsilon \to 0} \frac{T((1-\varepsilon)\hat{F}_n + \varepsilon\delta_x) - T(\hat{F}_n)}{\varepsilon}$$

④ Often, we write $\boxed{L(x)} = L_F(x)$.

Thm 2.22 (Behavior of $T(F)$ for linear T)

$T(F) := \int a(x)\, dF(x)$, a linear functional.

Then:

① $L_F(x) = a(x) - T(F)$ and $\hat{L}(x) = a(x) - T(\hat{F}_n)$.

② For any G, $T(G) = T(F) + \int L_F(x)\, dG(x)$.

③ $\int L_F(x)\, dF(x) = 0$.

④ Denote $\boxed{\tau^2} = \int L_F^2(x)\, dF(x)$. $\left(= \int (a(x) - T(F))^2 dF(x)\right)$.

If $\tau^2 < \infty$, then

$\sqrt{n}(T(F) - T(\hat{F}_n)) \rightsquigarrow N(0, \tau^2)$.

⑤ Denote $\boxed{\hat{\tau}^2} = \frac{1}{n}\sum_{i=1}^{n} \hat{L}^2(X_i)$. $\left(= \frac{1}{n}\sum_{i=1}^{n}(a(X_i) - T(\hat{F}_n))^2\right)$

Denote $\boxed{se} = \sqrt{\mathbb{V}ar(T(\hat{F}_n))}$ and $\boxed{\hat{se}} = \hat{\tau}/\sqrt{n}$.

Then $\hat{\tau}^2 \xrightarrow{P} \tau^2$ and

$\hat{se}/se \xrightarrow{P} 1$.

〈pf〉

① $L_F(x) = \lim_{\varepsilon \to 0} \frac{T((1-\varepsilon)F + \varepsilon\delta_x) - T(F)}{\varepsilon}$

$= \lim_{\varepsilon \to 0} \frac{(1-\varepsilon)\int a(x)dF(x) + \varepsilon a(x) - \int a(x)dF(x)}{\varepsilon}$

$= \lim_{\varepsilon \to 0} \frac{\varepsilon(a(x) - T(F))}{\varepsilon} = a(x) - T(F)$. ⊞

②, ③ comes from ①. ⊞

For ④, since $T(\hat{F}_n) = \frac{1}{n}\sum_{i=1}^{n} a(X_i)$ is the sample mean of the r.v. $a(X)$ (where $X \sim F$), by CLT,

$T(F) = \mathbb{E}_F[a(X)]$.

$\frac{T(\hat{F}_n) - T(F)}{\tau/\sqrt{n}} \rightsquigarrow N(0,1)$.

i.e. $\sqrt{n}(T(\hat{F}_n) - T(F)) \rightsquigarrow N(0, \tau^2)$. ⊞

For ⑤, notice that

$\tau^2 = Var[a(X)] = \mathbb{E}[(a(X) - T(F))^2]$ and

$\hat{\tau}^2 = \frac{1}{n}\sum_{i=1}^{n}(a(X_i) - T(F))^2$.

By LLN, $\hat{\tau}^2 \xrightarrow{P} \tau^2$.

In addition, by computation, $se = \tau/\sqrt{n}$.

$\Rightarrow \hat{se}/se = \frac{\hat{\tau}/\sqrt{n}}{\tau/\sqrt{n}} = \hat{\tau}/\tau \xrightarrow{P} 1$ as long as $\tau^2 < \infty$. ⊞

For nonlinear T, we need:

Def

$\mathcal{F} :=$ {all CDF}. $\mathcal{D} :=$ linear sp. generated by $\mathcal{F}$.

Equip $\mathcal{D}$ w/ a metric d.

T: a statistical functional.

T is called $\boxed{\text{Hadamard differentiable at } F}$

if $\exists$ linear functional L_F on $\mathcal{D}$ s.t.

$\forall\ \varepsilon_n \to 0$ and $\{D, D_1, D_2, \cdots\} \subseteq \mathcal{D}$ w/

$d(D_n, D) \to 0$ and $F + \varepsilon_n D_n \in \mathcal{F}$,

we have

$$\lim_{n\to\infty}\left(\frac{T(F + \varepsilon_n D_n) - T(F)}{\varepsilon_n} - L_F(D_n)\right) = 0.$$

Thm 2.27

T: Hadamard diff. at F w.r.t. $d(F, G) = \sup_x |F(x) - G(x)|$

Then

① $\sqrt{n}(T(\hat{F}_n) - T(F)) \rightsquigarrow N(0, \tau^2)$,

where $\tau^2 = \int L_F^2(x)\, dF(x)$.

② Denote $\hat{\tau}^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{L}^2(X_i)$ and $\hat{se} = \hat{\tau}/\sqrt{n}$.

Then $\frac{T(\hat{F}_n) - T(F)}{\hat{se}} \rightsquigarrow N(0,1)$.

Rmk:

We call the app. $\frac{T(\hat{F}_n) - T(F)}{\hat{se}} \approx N(0,1)$ the

$\boxed{\text{nonparametric delta method}}$.

Chap 8 The Bootstrap.

Idea:

step 1: Estimate $\mathbb{V}_F(T_n)$ w/ $\mathbb{V}_{\hat{F}_n}(T_n)$.

step 2: Approximate $\mathbb{V}_{\hat{F}_n}(T_n)$ using simulation.

8.1 Simulation.

Q: Given a CDF G (known), how do we estimate $\mathbb{E}[h(Y)]$, where h: a fun. and $Y \sim G$?

A:

Generate random sample $Y_1, Y_2, \cdots, Y_B \sim G$.

Then $\frac{1}{B}\sum_{i=1}^{B} h(Y_i) \xrightarrow{P} \mathbb{E}[h(Y)]$ (by LLN).

As long as we make B sufficiently large, we will have sufficiently small error. This process is called $\boxed{\text{simulation}}$.

8.2 Bootstrap Variance Estimation.

Idea: (As before) $\|F-\hat{F}_n\|_\infty \xrightarrow{P} 0$ via simulation

hopefully small, small

$$\mathbb{V}_F(T_n) \approx \mathbb{V}_{\hat{F}_n}(T_n) \approx v_{boot} \leftarrow \text{(as below)}.$$

Algorithm: (Bootstrap var. est.)

1. Draw $X_1^*, \cdots, X_n^* \sim \hat{F}_n$.
   (i.e. draw n observations w/ replacement from $X_1, \cdots, X_n$).
2. Compute $T_n^* = g(X_1^*, \cdots, X_n^*)$.
3. Repeat 1. and 2., B times, to get $T_{n,1}^*, \cdots, T_{n,B}^*$.
4. Let
$$v_{boot} = \frac{1}{B}\sum_{b=1}^{B}\left(T_{n,b}^* - \frac{1}{B}\sum_{r=1}^{B} T_{n,r}^*\right)^2.$$

Rmk:
In terms of languages in 8.1, we are simulating the variance of G, w/ G the CDF representing $g(X_1^*, \cdots, X_n^*)$ w/ $X_1^*, \cdots, X_n^* \sim \hat{F}_n$.

8.3 Bootstrap Confidence Intervals.

Method 1: (The Normal Interval).

$\boxed{\hat{se}_{boot}} := \sqrt{v_{boot}}$.

$C_n := T_n \pm z_{\alpha/2}\, \hat{se}_{boot}$.

Method 2: (Pivotal Interval).

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

$\theta_{n,1}^*, \cdots, \theta_{n,B}^*$:= bootstrap results for B times.

For $\beta \in (0,1)$, $\boxed{\theta_\beta^*}$:= β sample quantile of $(\theta_{n,1}^*, \cdots, \theta_{n,B}^*)$.

Define

$C_n = (2\hat{\theta}_n - \theta^*_{1-\alpha/2},\ 2\hat{\theta}_n - \theta^*_{\alpha/2})$, called the **(1-α) bootstrap pivotal confidence interval**.

Method 3: (Percentile Interval).

$C_n := (\theta^*_{\alpha/2}, \theta^*_{1-\alpha/2})$, called the **(1-α) bootstrap percentile interval**.

Thm 8.3

Under weak conditions on $T(F)$, as $n \to \infty$,

$\mathbb{P}_F(T(F) \in C_n) \to 1-\alpha$, where

C_n is as in Method 2.

Prop

Suppose $\exists$ monotone m s.t. $U = m(T) \sim N(\phi, c^2)$. (for some ϕ and c.)

Then $\mathbb{P}(\theta \in C_n) = 1-\alpha$, where

C_n is as in Method 3.

Chap 9 Parametric Inference.

Note:

We rarely know whether the dist. generating the data is in some parametric model.

Q: Why study parametric model?

A: (Larry Wasserman).

① Background knowledge suggests it.

② Can provide background for understanding certain nonpara. models.

9.1 Parameter of Interest.

Def

The **parameter of interest** is the parameter(s) that we are interested in. Other parameters are called **nuisance parameters**.

9.2 The Method of Moments.

Def

For $X \sim F_\theta(x)$, where $\theta = (\theta_1, \cdots, \theta_k)$ is the parameter,

① $\boxed{\alpha_j} := \mathbb{E}_\theta(X^j) = \int x^j \, dF_\theta(x)$ is called the **jth moment** of X.

$\boxed{\hat{\alpha}_j} := \frac{1}{n}\sum_{i=1}^{n} X_i^j$ is called the **jth sample moment**.

② The **method of moments estimator** $\boxed{\hat{\theta}_n}$ is the value s.t.

$$\alpha_1(\hat{\theta}_n) = \hat{\alpha}_1$$
$$\alpha_2(\hat{\theta}_n) = \hat{\alpha}_2$$
$$\vdots$$
$$\alpha_K(\hat{\theta}_n) = \hat{\alpha}_K.$$

i.e. $\hat{\theta}_n$ is a sol'n of this system of K equations w/ K unknowns.

Ex:

$X_1, \cdots, X_n \sim N(\mu, \sigma^2)$.

$\alpha_1(\mu, \sigma^2) = \mu$.

$\alpha_2(\mu, \sigma^2) = \mathbb{E}[(X-\mu)^2] = \sigma^2 + \mu^2$.

Thus, we need to solve

$$\begin{cases} \hat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i \\ (\hat{\sigma})^2 + (\hat{\mu})^2 = \frac{1}{n}\sum_{i=1}^n X_i^2. \end{cases}$$

i.e.

$$\begin{cases} \hat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i \\ (\hat{\sigma})^2 = \left(\frac{1}{n}\sum_{i=1}^n X_i^2\right) - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2. \end{cases}$$

#

Thm 9.6

$\hat{\theta}_n :=$ the method of moments estimator.

Under appropriate conditions on the model, we have

① $\hat{\theta}_n$ exists w/ prob. $\to 1$.

② $\hat{\theta}_n \xrightarrow{P} \theta$.

③ denote $g_j := \partial \alpha_j^{-1}(\theta)/\partial\theta$

$g := (g_1, \cdots, g_K) \leftarrow K\times 1.$

$Y := (X, X^2, \cdots, X^k)^T \leftarrow 1\times K.$

$\Sigma := g\,\mathbb{E}_\theta(YY^T)g^T, \leftarrow K\times K.$

then $\sqrt{n}(\hat{\theta}_n - \theta) \rightsquigarrow N(0, \Sigma)$.

9.3 Maximum Likelihood.

Def $X_1, \cdots, X_n$ iid w/ PDF $f(x;\theta)$.

$\boxed{\mathcal{L}_n(\theta)} := \prod_{i=1}^n f(X_i;\theta)$ is called the **likelihood function**

$\boxed{\ell_n(\theta)} := \log \mathcal{L}_n(\theta)$ is called the **log-likelihood function**.

Def

The **maximum likelihood estimator** **MLE**, denoted $\hat{\theta}_n$, is the value θ that maximizes $\mathcal{L}_n(\theta)$.

9.4 Properties of MLE.

Under certain conditions on the model, we have

1. MLE is consistent. i.e. $\hat{\theta}_n \xrightarrow{P} \theta$
2. MLE is equivariant. i.e. $\hat{\theta}_n$: MLE of θ $\Rightarrow g(\hat{\theta}_n)$: MLE of $g(\theta)$.
3. MLE is asymptotically normal. i.e. $(\hat{\theta} - \theta)/\hat{se} \rightsquigarrow N(0,1)$.
4. .. .. asymptotically optimal. (also called efficient).
5. .. .. approximately the Bayes estimator.

9.5 Consistency of MLE.

Def

f, g : PDF.

The **Kullback-Leibler distance** b/w f and g is

$$\boxed{D(f,g)} = \int f(x) \log\left(\frac{f(x)}{g(x)}\right) dx.$$

Note: D is NOT a distance since it is NOT symmetric.

Prop

① $D(f,f) = 0$.

② $D(f,g) \geq 0$.

〈Pf〉

① is obvious.

$-\log$ is convex.

② $\int f(x) \log\left(\frac{f(x)}{g(x)}\right) dx = \int -\log\left(\frac{g(x)}{f(x)}\right) dF(x)$

$\geq -\log\left(\int \frac{g(x)}{f(x)} dF(x)\right) = -\log\left(\int \frac{g(x)}{\cancel{f(x)}} \cancel{f(x)}\, dx\right) = 0.$

i.e. $D(f,g) \geq 0$. #

Def

For a parametric model $\mathfrak{F} = \{f(x;\theta) \mid \theta \in \Theta\}$, and $\theta, \psi \in \Theta$, we shall denote

$$\boxed{D(\theta, \psi)} := D(f(x;\theta), f(x;\psi)).$$

The model $\mathfrak{F}$ is called **identifiable** if

$$D(\theta, \psi) > 0, \ \forall\, \theta \neq \psi \text{ in } \Theta.$$

We shall assume from now on that the model $\mathfrak{F}$ is identifiable.

Prop

$\theta_* :=$ true value of θ.

$\boxed{M_n(\theta)} := \frac{1}{n}\sum_i \log \frac{f(X_i;\theta)}{f(X_i;\theta_*)}$, where $X_1,\cdots,X_n$: iid from $f(X;\theta_*)$.

Fix $\theta \in \Theta$.

Then $M_n(\theta) \xrightarrow{P} -D(\theta_*,\theta)$.

⟨Pf⟩

By LLN, $M_n(\theta) \xrightarrow{P} \mathbb{E}_{\theta_*}\left(\log \frac{f(X;\theta)}{f(X;\theta_*)}\right)$.

$= \int \log \frac{f(x;\theta)}{f(x;\theta_*)} f(x;\theta_*)\,dx = -D(\theta_*,\theta)$. #

Thm 9.13

$\theta_* :=$ true value of θ.

$M_n(\theta)$ as above.

$M(\theta) := -D(\theta_*,\theta)$.

Suppose

(9.7) $\underbrace{\sup_{\theta\in\Theta} |M_n(\theta)-M(\theta)| \xrightarrow{P} 0}$ and

unif. con. for the above prop.

(9.8) $\underbrace{\forall \varepsilon>0,\ \sup_{|\theta-\theta_*|\geq\varepsilon} M(\theta) < M(\theta_*)}$.

θ_* is the unique global maximum.

Then the MLE $\hat{\theta}_n \xrightarrow{P} \theta_*$.

⟨Pf⟩

$0 \overset{\text{by (9.8)}}{\leq} M(\theta_*) - M(\hat{\theta}_n)$

$= M_n(\theta_*) - M(\hat{\theta}_n) + M(\theta_*) - M_n(\theta_*)$

$\overset{\because \hat{\theta}_n \text{ is MLE}}{\leq} M_n(\hat{\theta}_n) - M(\hat{\theta}_n) + M(\theta_*) - M_n(\theta_*)$

$\leq |M_n(\hat{\theta}_n) - M(\hat{\theta}_n)| + |M(\theta_*) - M_n(\theta_*)|$.

$\leq 2 \sup_\theta |M_n(\theta)-M(\theta)| \xrightarrow{P} 0$ by 9.7

Thus $M(\hat{\theta}_n) \xrightarrow{P} M(\theta_*)$.

By (9.8), it is clear that $\hat{\theta}_n \xrightarrow{P} \theta_*$. #

9.6 Equivariance of MLE.

5.2 Types of Convergences.

Def

$X_1, X_2, \cdots$: seq. of r.v.

X : r.v.

F_n : CDF of X_n,

F : CDF of X.

① $\boxed{X_n \text{ converges to } X \text{ in probability}}$, written $\boxed{X_n \xrightarrow{P} X}$,

if, $\forall\, \varepsilon > 0$, $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$ as $n \to \infty$.

② $\boxed{X_n \text{ converges to } X \text{ in distribution}}$, written $\boxed{X_n \rightsquigarrow X}$,

if $\lim_{n\to\infty} F_n(t) = F(t)$, $\forall\, t$ w/ F cont. at t.

③ $\boxed{X_n \text{ converges to } X \text{ in quadratic mean}}$ (or $\boxed{\text{in } L^2}$),

written $\boxed{X_n \xrightarrow{qm} X}$, if

$$\mathbb{E}[(X_n - X)^2] \to 0 \quad \text{as} \quad n \to \infty.$$

Thm 5.4

(a) $X_n \xrightarrow{qm} X \Rightarrow X_n \xrightarrow{P} X$.

(b) $X_n \xrightarrow{P} X \Rightarrow X_n \rightsquigarrow X$

(c) $X_n \rightsquigarrow X$ w/ $X = c$ (const.) a.s.

$\Rightarrow X_n \xrightarrow{P} X$.

Rmk:

① (Summary)

$$\begin{tikzcd}
 & L^1 \arrow[dr] & \text{almost surely} \arrow[d] & \text{point-mass dist.} \\
\text{quadratic mean} \arrow[ur] \arrow[rr] & & \text{probability} \arrow[r] & \text{distribution} \arrow[l, bend right, dashed]
\end{tikzcd}$$

② For $\not\Leftarrow$ of (a) and (b) above, see P74-75.

③ $X_n \xrightarrow{P} b$ (const.) $\not\Rightarrow \mathbb{E}[X_n] \to b$.

e.g. X_n s.t.
$$\begin{cases} \mathbb{P}(X_n = n^2) = 1/n \\ \mathbb{P}(X_n = 0) = 1 - 1/n \end{cases}$$

Thm 5.5

X_n, X, Y_n, Y : r.v.

g : cont. Then

(a) $X_n \xrightarrow{P} X$, $Y_n \xrightarrow{P} Y \Rightarrow X_n + Y_n \xrightarrow{P} X + Y$.

(b) $X_n \xrightarrow{qm} X$, $Y_n \xrightarrow{qm} Y \Rightarrow X_n + Y_n \xrightarrow{qm} X + Y$.

(c) $X_n \rightsquigarrow X$, $Y_n \rightsquigarrow c \Rightarrow X_n + Y_n \rightsquigarrow X + c$.

(d) $X_n \xrightarrow{P} X$, $Y_n \xrightarrow{P} Y \Rightarrow X_n Y_n \xrightarrow{P} XY$.

Slutzky's thm.

(e) $X_n \rightsquigarrow X$, $Y_n \rightsquigarrow c \Rightarrow X_n Y_n \rightsquigarrow cX$.

(f) $X_n \xrightarrow{P} X \Rightarrow g(X_n) \xrightarrow{P} g(X)$.

(g) $X_n \rightsquigarrow X \Rightarrow g(X_n) \rightsquigarrow g(X)$.

Def

X_n, X : r.v.

① $\boxed{X_n \text{ converges almost surely to } X}$, written $\boxed{X_n \xrightarrow{as} X}$,

if $\mathbb{P}(\{\omega \mid X_n(\omega) \to X(\omega)\}) = 1$.

② $\boxed{X_n \text{ converges in } L^1 \text{ to } X}$, written $\boxed{X_n \xrightarrow{L^1} X}$, if

$\mathbb{E}[|X_n - X|] \to 0$, as $n \to \infty$.

Thm 5.17

X_n, X : r.v. Then

(a) $X_n \xrightarrow{as} X \Rightarrow X_n \xrightarrow{P} X$.

(b) $X_n \xrightarrow{qm} X \Rightarrow X_n \xrightarrow{L^1} X$

(c) $X_n \xrightarrow{L^1} X \Rightarrow X_n \xrightarrow{P} X$.