

Review on Statistics

Probability Spaces

A probability space is a triple (Ω, \mathcal{F}, P) where Ω is a set of “outcomes,” \mathcal{F} is a set of “events,” and $P : \mathcal{F} \rightarrow [0, 1]$ is a function that assigns probabilities to events.

Definition. Let Ω be a set. A nonempty collection \mathcal{F} of subsets of Ω is called σ -algebra (or field) if

- (i) if $A \in \mathcal{F}$ then $\Omega \setminus A \in \mathcal{F}$, and
- (ii) if $A_1, A_2, \dots \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Example. $\mathcal{F} = \{\emptyset, \Omega\}$ trivial σ -field

$$\mathcal{F} = 2^{\Omega} = \{A \mid A \subset \Omega\} : \text{power set} \implies \sigma\text{-field}$$

Without P , (Ω, \mathcal{F}) is called a measurable space, i.e., it is a space on which we can put a measure.

Definition. A measure is a nonnegative countably additive set function; that is, for an σ -algebra \mathcal{F} , a function $\mu : \mathcal{F} \rightarrow \mathbb{R}$ is a measure if

- (i) $\mu(A) \geq \mu(\emptyset) = 0$ for all $A \in \mathcal{F}$, and
- (iii) For $A_1, A_2, \dots \in \mathcal{F}$ with $A_i \cap A_j = \emptyset$ for any $i \neq j$,

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i).$$

Definition. (1) $\mu(\Omega) < \infty \implies$ finite measure

(2) $\mu(\Omega) = 1 \implies$ probability measure

(3) \exists a partition A_1, A_2, \dots with $\bigcup_{i=1}^{\infty} A_i = \Omega$ and $\mu(A_i) < \infty \implies \sigma$ -finite measure

Theorem ([Durrett(2010), Theorem 1.1.4]). *Let μ be a measure on (Ω, \mathcal{F}) .*

- (i) *Monotonicity.* If $A \subset B$ then $\mu(A) \leq \mu(B)$.
- (ii) *Subadditivity.* If $A \subset \bigcup_{i=1}^{\infty} A_i$ then $\mu(A) \leq \sum_{i=1}^{\infty} \mu(A_i)$.

(iii) *Continuity from below.* $A_n \uparrow A$ (i.e. $A_1 \subset A_2 \subset \dots$ and $A = \bigcup_{i=1}^{\infty} A_i$) then $\mu(A_i) \uparrow \mu(A)$.

(iv) *Continuity from above.* $A_n \downarrow A$ (i.e. $A_1 \supset A_2 \supset \dots$ and $A = \bigcap_{i=1}^{\infty} A_i$) with $\mu(A_1) < \infty$ then $\mu(A_i) \downarrow \mu(A)$.

Definition. Let \mathcal{A} be a class of subsets of Ω . Then $\sigma(\mathcal{A})$ denotes the smallest σ -algebra that contains \mathcal{A} .

For any any \mathcal{A} , such $\sigma(\mathcal{A})$ exists and is unique: [Durrett(2010), Exercise 1.1.1].

Definition. Borel σ -field on \mathbb{R}^d , denoted by \mathcal{R}^d , is the smallest σ -field containing all open sets.

Theorem ([Durrett(2010), Theorem 1.1.2]). *There is a unique measure μ on $(\mathbb{R}, \mathcal{R})$ with*

$$\mu((a, b]) = b - a.$$

Such measure is called Lebesgue measure.

Example ([Durrett(2010), Example 1.1.3]). Product space

$(\Omega_i, \mathcal{F}_i, P_i)$: sequence of probability spaces

Let $\Omega = \Omega_1 \times \dots \times \Omega_n = \{(\omega_1, \dots, \omega_n) \mid \omega_i \in \Omega_i\}$

$\mathcal{F} = \mathcal{F}_1 \times \dots \times \mathcal{F}_n$ =the σ -field generated by $A_1 \times \dots \times A_n$, where $A_i \in \mathcal{F}_i$

$P = P_1 \times \dots \times P_n$ (i.e. $P(A_1 \times \dots \times A_n) = P_1(A_1) \dots P_n(A_n)$)

Distribution and Random Variables

Definition. Let (Ω, \mathcal{F}) and (S, \mathcal{S}) are measurable spaces. A mapping $X : \Omega \rightarrow S$ is a measurable map from (Ω, \mathcal{F}) to (S, \mathcal{S}) if

$$\text{for all } B \in \mathcal{S}, X^{-1}(B) := \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}.$$

If $(S, \mathcal{S}) = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and $d > 1$ then X is called a random vector. If $d = 1$, X is called a random variable.

Example. A trivial but useful example of a random variable is indicator function 1_A of a set $A \in \mathcal{F}$:

$$1_A(\omega) = \begin{cases} 1 & \omega \in A, \\ 0 & \omega \notin A. \end{cases}$$

If X is a random variable, then X induces a probability measure on \mathbb{R} .

Definition. The probability measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ defined as $\mu(A) = P(X \in A)$ for all $A \in \mathcal{B}(\mathbb{R})$ is called the distribution of X .

Remark. The distribution can be defined similarly for random vectors.

The distribution of a random variable X is usually described by giving its distribution function.

Definition. The distribution function $F(x)$ of a random variable X is defined as $F(x) = P(X \leq x)$.

Theorem ([Durrett(2010), Theorem 1.2.1]). *Any distribution function F has the following properties:*

- (i) F is nondecreasing.
- (ii) $\lim_{n \rightarrow \infty} F(x) = 1$, $\lim_{n \rightarrow -\infty} F(x) = 0$.
- (iii) F is right continuous. i.e. $\lim_{y \downarrow x} F(y) = F(x)$.
- (iv) $P(X < x) = F(x-) = \lim_{y \uparrow x} F(y)$.
- (v) $P(X = x) = F(x) - F(x-)$.

Theorem ([Durrett(2010), Theorem 1.2.2]). *If F satisfies (i) (ii) (iii) in [Durrett(2010), Theorem 1.2.1], then it is the distribution function of some random variable. That is, there exists a triple (Ω, \mathcal{F}, P) and a random variable X such that $F(x) = P(X \leq x)$.*

Theorem. *If F satisfies (i) (ii) (iii), then $\exists!$ probability measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that for all $a < b$,*

$$\mu((a, b]) = F(b) - F(a)$$

Definition. If X and Y induce the same distribution μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, we say X and Y are equal in distribution.

We write

$$X \stackrel{d}{=} Y.$$

Definition. When the distribution function $F(x) = P(X \leq x)$ has the form $F(x) = \int_{-\infty}^x f(y)dy$, then we say X has the density function f .

Remark. f is not unique, but unique up to Lebesgue measure 0.

Theorem ([Durrett(2010), Theorem 1.3.2]). *If $X : (\Omega, \mathcal{F}) \rightarrow (S, \mathcal{S})$ and $f : (S, \mathcal{S}) \rightarrow (T, \mathcal{T})$ are measurable maps, then $f(X)$ is measurable.*

Theorem. *$f : (S, \mathcal{S}) \rightarrow (T, \mathcal{T})$ and suppose $\mathcal{S} = \sigma(\text{open sets})$, $\mathcal{T} = \sigma(\text{open sets})$. Then, if f is continuous then f is measurable.*

Theorem ([Durrett(2010), Theorem 1.3.3]). *If X_1, \dots, X_n are random variables and $f : (\mathbb{R}^n, \mathcal{R}^n) \rightarrow (\mathbb{R}, \mathcal{R})$ is measurable, then $f(X_1, \dots, X_n)$ is a random variable.*

Theorem ([Durrett(2010), Theorem 1.3.4]). *If X_1, \dots, X_n are random variables then $X_1 + \dots + X_n$ is a random variable.*

Remark. If X, Y are random variables, then

$$cX \text{ (} c \text{ is scalar), } X \pm Y, XY, \sin(X), X^2, \dots,$$

are all random variables.

Theorem ([Durrett(2010), Theorem 1.3.5]). *$\inf_n X_n, \sup_n X_n, \limsup_n X_n, \liminf_n X_n$ are random variables.*

Integration

Let μ be a σ -finite measure on (Ω, \mathcal{F}) .

Definition. For any predicate $Q(\omega)$ defined on Ω , we say Q is true $(\mu-)$ almost everywhere (or a.e.) if $\mu(\{\omega : Q(\omega) \text{ is false}\}) = 0$

Step 1.

Definition. φ is a simple function if $\varphi(\omega) = \sum_{i=1}^n a_i 1_{A_i}$ with $A_i \in \mathcal{F}$

If φ is a simple function and $\varphi \geq 0$, we let

$$\int \varphi d\mu = \sum_{i=1}^n a_i \mu(A_i)$$

Step 2.

Definition. If f is measurable and $f \geq 0$ then we let

$$\int f d\mu = \sup\{\int h d\mu : 0 \leq h \leq f \text{ and } h \text{ simple}\}$$

Step 3.

Definition. We say measurable f is integrable if $\int |f| d\mu < \infty$

let $f^+(x) := f(x) \vee 0$, $f^-(x) := (-f)(x) \vee 0$ where $a \vee b = \max(a, b)$

We define the integral of f by

$$\int f d\mu = \int f^+ d\mu - \int f^- d\mu$$

we can also define $\int f d\mu$ if $\int f^+ d\mu = \infty$ and $\int f^- d\mu < \infty$, or $\int f^+ d\mu < \infty$ and $\int f^- d\mu = \infty$

Theorem. (1.4.7) Suppose f and g are integrable.

(i) If $f \geq 0$ a.e. then $\int f d\mu \geq 0$

(ii) $\forall a \in \mathbb{R}$, $\int a f d\mu = a \int f d\mu$

(iii) $\int f + g d\mu = \int f d\mu + \int g d\mu$

(iv) If $g \leq f$ a.e. then $\int g d\mu \leq \int f d\mu$

(v) If $g = f$ a.e. then $\int g d\mu = \int f d\mu$

(vi) $|\int f d\mu| \leq \int |f| d\mu$

Independence

Definition. Let (Ω, \mathcal{F}, P) be probability space. Two events $A, B \in \mathcal{F}$ are independent if

$$P(A \cap B) = P(A) \times P(B)$$

Two random variables X and Y are independent if

$$\forall C, D \in \mathcal{R}, P(X \in C, Y \in D) = P(X \in C)P(Y \in D)$$

Two σ -fields \mathcal{F}_1 and $\mathcal{F}_2 (\subset \mathcal{F})$ are independent if

$$\forall A \in \mathcal{F}_1, \forall B \in \mathcal{F}_2, A \text{ and } B \text{ are independent.}$$

Remark. An infinite collection of objects (σ -fields, random variables, or sets) is said to be independent if every finite subcollection is.

Definition. σ -fields $\mathcal{F}_1, \dots, \mathcal{F}_n$ are independent if

$$P\left(\bigcap_{i=1}^n A_i\right) = \prod_{i=1}^n P(A_i), \quad \forall A_i \in \mathcal{F}_i$$

random variables X_1, \dots, X_n are independent if

$$P\left(\bigcap_{i=1}^n \{X_i \in B_i\}\right) = \prod_{i=1}^n P(X_i \in B_i), \quad \forall B_i \in \mathcal{R}$$

Sets A_1, \dots, A_n are independent if

$$P\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} P(A_i) \text{ for all } I \subset \{1, \dots, n\}$$

Remark. the definition of independent events is not enough to assume pairwise independent, which is $P(A_i \cap A_j) = P(A_i)P(A_j)$, $i \neq j$. It is clear that independent events are pairwise independent, but converse is not true.

Example. Let X_1, X_2, X_3 be independent random variables with $P(X_i = 0) = P(X_i = 1) = \frac{1}{2}$

Let $A_1 = \{X_2 = X_3\}$, $A_2 = \{X_3 = X_1\}$ and $A_3 = \{X_1 = X_2\}$. These events are pairwise independent but not independent.

Weak laws of large numbers

Various modes of convergence

$\{X_n\}$ and X are random variables defined on (Ω, \mathcal{F}, P)

Definition. $X_n \rightarrow X$ almost surely (a.s.) (with probability 1(w.p. 1), almost everywhere(a.e.)) if $P\{\omega : X_n(\omega) \rightarrow X(\omega)\} = 1$

Equivalent definition : $\forall \epsilon, \lim_{m \rightarrow \infty} P\{\omega : |X_n(\omega) - X(\omega)| \leq \epsilon \forall n \geq m\} = 1$

or $\forall \epsilon, \lim_{m \rightarrow \infty} P\{\omega : |X_n(\omega) - X(\omega)| > \epsilon \forall n \geq m\} = 0$

Definition. $X_n \rightarrow X$ in probability (in pr, \xrightarrow{p}) if $\lim_{n \rightarrow \infty} P\{|X_n - X| > \epsilon\} = 0$

Theorem. $X_n \rightarrow X$ a.s. $\implies X_n \xrightarrow{p} X$

Remark. $X_n \xrightarrow{p} X \not\Rightarrow X_n \rightarrow X$ a.s.

Definition. $X_n \rightarrow X$ in L_p , $0 < p < \infty$

if $\lim_{n \rightarrow \infty} E(|X_n - X|^p) = 0$ provided $E|X_n|^p < \infty$, $E|X|^p < \infty$.

Theorem. $X_n \rightarrow X$ in $L_p \implies X_n \xrightarrow{p} X$

Theorem. (Chebyshev inequality)

$$P(|X| \geq \epsilon) \leq \frac{E|X|^p}{\epsilon^p}$$

Remark. $X_n \xrightarrow{p} X \not\Rightarrow X_n \rightarrow X$ in L_p

Example. $\Omega = [0, 1]$, $\mathcal{F} = \mathcal{B}[0, 1]$, $P = \text{Unif}[0, 1]$

$$X(\omega) = 0, X_n(\omega) = nI(0 \leq \omega \leq \frac{1}{n})$$

$$\text{Then } P\{|X_n(\omega) - X(\omega)| > \epsilon\} = P\{0 \leq \omega \leq \frac{1}{n}\} = \frac{1}{n} \rightarrow 0$$

$$\text{But } E|X_n - X| = E|X_n| = 1$$

Theorem. $X_n \xrightarrow{p} X$ and there exists a random variables Z s.t.

$$|X_n| \leq Z \text{ and } E|Z|^p < \infty$$

Then $X_n \rightarrow X$ in L_p .

Remark. If $E|X| < \infty$, then

$$\lim_{n \rightarrow \infty} \int_{A_n} |X| dP \rightarrow 0 \text{ whenever } P(A_n) \rightarrow 0$$

2..2.1. L_2 weak law

Theorem. (2.2.3) Let X_1, X_2, \dots be uncorrelated random variables with $EX_i = \mu$ and $\text{Var}(X_i) \leq C < \infty$

Let $S_n = \sum_{i=1}^n X_i$. Then

$\frac{S_n}{n} \rightarrow \mu$ in L_2 and so in pr.

Weak Convergence

Definition. A sequence of distribution function F_n converges weakly to a limit F ($F_n \Rightarrow F$, $F_n \xrightarrow{w} F$)

if $F_n(y) \rightarrow F(y) \forall y$ that are continuity points of F .

Definition. A sequence of random variables $\{X_n\}$ converges weakly or converges in distribution to a limit X

$$(X_n \Rightarrow X, X_n \xrightarrow{w} X, X_n \xrightarrow{d} X)$$

If the distribution function F_n of X_n converges weakly to the distribution of X .

Example. (3.2.1) Let X_1, X_2, \dots be iid with $P(X_1 = 1) = P(X_1 = -1) = \frac{1}{2}$.

Let $S_n = X_1 + \dots + X_n$.

Then $F_n(y) = P(S_n/\sqrt{n} \leq y) \rightarrow \int_{-\infty}^y \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \forall y$

That is, $F_n \Rightarrow N(0, 1)$

Example ([Durrett(2010), Example 3.2.3]). Let $X \sim F$ and $X_n = X + \frac{1}{n}$

Then $F_n(x) = P(X_n \leq x) = F(x - \frac{1}{n}) \rightarrow F(x-)$

Hence $F_n(x) \rightarrow F(x)$ only when $F(x) = F(x-)$

(i.e. x is a continuity point of F)

so $X_n \rightarrow X$

Example ([Durrett(2010), Example 3.2.4]). (3.2.4) $X_p \sim \text{Geo}(p)$ (i.e. $P(X_p \geq m) = (1-p)^{m-1}$)

Then $P(X_p > \frac{x}{p}) = (1-p)^{\frac{x}{p}} \rightarrow e^{-x}$ as $p \rightarrow 0$

Central Limit Theorem

Theorem ([Durrett(2010), Theorem 3.4.1]). Let X_1, X_2, \dots be iid with $EX_i = \mu$ and $\text{Var}(X_i) = \sigma^2 > 0$.

If $S_n = X_1 + \dots + X_n$, then

$$(S_n - n\mu)/(\sqrt{n}\sigma) \xrightarrow{d} N(0,1)$$

Theorem ([Durrett(2010), Theorem 3.4.9]). *Berry-Essen theorem*

Let X_1, X_2, \dots be i.i.d. with $EX_i = 0$, $EX_i^2 = \sigma^2$ and $E|X_1|^3 = \rho < \infty$

Let $F_n(x)$ be the distribution function of $(X_1 + \dots + X_n)/(\sigma\sqrt{n})$ and $\Phi(x)$ be the standard normal distribution.

Then $\sup_x |F_n(x) - \Phi(x)| \leq 3\rho/(\sigma^3\sqrt{n})$

Stochastic Order Notation

The classical order notation should be familiar to you already.

1. We say that a sequence $a_n = o(1)$ if $a_n \rightarrow 0$ as $n \rightarrow \infty$. Similarly, $a_n = o(b_n)$ if $a_n/b_n = o(1)$.
2. We say that a sequence $a_n = O(1)$ if the sequence is eventually bounded, i.e. for all n large, $|a_n| \leq C$ for some constant $C \geq 0$. Similarly, $a_n = O(b_n)$ if $a_n/b_n = O(1)$.
3. If $a_n = O(b_n)$ and $b_n = O(a_n)$ then we use either $a_n = \Theta(b_n)$ or $a_n \asymp b_n$.

When we are dealing with random variables we use stochastic order notation.

1. We say that $X_n = o_P(1)$ if for every $\epsilon > 0$, as $n \rightarrow \infty$

$$\mathbb{P}(|X_n| \geq \epsilon) \rightarrow 0,$$

i.e. X_n converges to zero in probability.

2. We say that $X_n = O_P(1)$ if for every $\epsilon > 0$ there is a finite $C(\epsilon) > 0$ such that, for all n large enough:

$$\mathbb{P}(|X_n| \geq C(\epsilon)) \leq \epsilon.$$

The typical use case: suppose we have X_1, \dots, X_n which are i.i.d. and have finite variance, and we define:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i.$$

1. $\hat{\mu} - \mu = o_P(1)$ (Weak Law of Large Number)
2. $\hat{\mu} - \mu = O_P(1/\sqrt{n})$ (Central Limit Theorem)

As with the classical order notation, we can do some simple “calculus” with stochastic order notation and observe that for instance: $o_P(1) + O_P(1) = O_P(1)$, $o_P(1)O_P(1) = o_P(1)$ and so on.

Asymptotic Theory

We suppose that we obtain a sample $X_1, \dots, X_n \sim P$. Let $\theta(P)$ be a parameter, which is some function of P . Let $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ denote an estimator for θ , which is a function of a sample. We are interested in two questions:

1. Consistency: Does the estimator $\hat{\theta}$ converge in probability to θ , i.e. does $\hat{\theta} \xrightarrow{P} \theta$? More precisely, can we find some function $f(n)$ of the sample size n such that $d(\hat{\theta}, \theta) = O_P(f(n))$? This is analogous to the Law of Large Number.

2. Asymptotic distribution: What can we say about the distribution of $\sqrt{n}(\hat{\theta} - \theta)$? This is analogous to the Central Limit Theorem.

Confidence Set

Suppose we have a statistical model (i.e. a collection of distributions) \mathcal{P} . Let $C_n(X_1, \dots, X_n)$ be a set constructed using the observed data X_1, \dots, X_n . This is a random set. C_n is a $1 - \alpha$ confidence set for a parameter θ if:

$$P(\theta \in C_n(X_1, \dots, X_n)) \geq 1 - \alpha, \text{ for all } P \in \mathcal{P}.$$

This means that no matter which distribution in \mathcal{P} generated the data, the interval guarantees the coverage property described above.

Bootstrap

The bootstrap is a method for estimating standard errors and computing confidence intervals. Let $X_1, \dots, X_n \sim P$, and $T_n = g(X_1, \dots, X_n)$ be a statistic, that is, T_n is any function of the data. Suppose we want to know $\mathbb{V}_P(T_n)$, the variance of T_n , where the notation \mathbb{V}_P emphasizes the dependence on the unknown distribution P . For example, if $T_n = \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ then $\mathbb{V}_P(T_n) = \sigma^2/n$ where $\sigma^2 = \int (x - \mu)^2 dP(x)$ and $\mu = \int x dP(x)$. Let P_n be the empirical measure that puts mass $1/n$ at each data point, thus

$$P_n(A) = \frac{1}{n} \sum_{i=1}^n I(X_i \in A).$$

The bootstrap idea has two steps:

Step 1: Estimate $\mathbb{V}_P(T_n)$ with $\mathbb{V}_{P_n}(T_n)$.

Step 2: Approximate $\mathbb{V}_{P_n}(T_n)$ using Monte Carlo.

Monte Carlo

Suppose we draw an iid sample $Y_1, \dots, Y_B \sim P$, and h is any function with finite mean, i.e., $\mathbb{E}[h(Y)] < \infty$, then by the weak law of large numbers,

$$\frac{1}{B} \sum_{j=1}^B h(Y_j) \xrightarrow{P} \int h(y) dP(y) = \mathbb{E}[h(Y)],$$

as $B \rightarrow \infty$. In particular,

$$\begin{aligned} \frac{1}{B} \sum_{j=1}^B (Y_j - \bar{Y})^2 &= \frac{1}{B} \sum_{j=1}^B Y_j^2 - \left(\frac{1}{B} \sum_{j=1}^B Y_j \right)^2 \\ &\xrightarrow{P} \int y^2 dP(y) - \left(\int y dP(y) \right)^2 = \mathbb{V}_P(Y). \end{aligned}$$

Hence, we can use the sample variance of the simulated values to approximate $\mathbb{V}_P(Y)$.

Bootstrap Variance Estimation

Now, $\mathbb{V}_{P_n}(T_n)$ means “the variance of T_n if the distribution of the data is P_n ”. To compute this, we simulate X_1^*, \dots, X_n^* from P_n and then compute $T_n^* = g(X_1^*, \dots, X_n^*)$. This constitutes one draw from the distribution of T_n . The idea is illustrated in the following diagram:

$$\begin{array}{llll} \text{Real world } P & \Rightarrow & X_1, \dots, X_n & \Rightarrow & T_n = g(X_1, \dots, X_n) \\ \text{Bootstrap world } P_n & \Rightarrow & X_1^*, \dots, X_n^* & \Rightarrow & T_n^* = g(X_1^*, \dots, X_n^*) \end{array}$$

How do we simulate X_1^*, \dots, X_n^* from P_n ? Notice that P_n puts mass $1/n$ at each data point X_1, \dots, X_n . Therefore

drawing an observation from P_n is equivalent to drawing one point at random from the original data set.

Thus, to simulate $X_1^*, \dots, X_n^* \sim P_n$, it suffices to draw n observations with replacement from X_1, \dots, X_n . The algorithm for bootstrap variance estimation is below:

1. Draw $X_1^*, \dots, X_n^* \sim P_n$.
2. Compute $T_n^* = g(X_1^*, \dots, X_n^*)$.
3. Repeat step 1 and 2, B times, to get $T_{n,1}^*, \dots, T_{n,B}^*$.
4. Let

$$v_{boot} = \frac{1}{B} \sum_{b=1}^B \left(T_{n,b}^* - \frac{1}{B} \sum_{r=1}^B T_{n,r}^* \right)^2.$$

Notice that we are using two approximations:

$$\mathbb{V}_P(T_n) \overset{\text{not so small}}{\approx} \mathbb{V}_{P_n}(T_n) \overset{\text{small}}{\approx} v_{boot}.$$

Bootstrap Confidence Intervals

There are several ways to construct bootstrap confidence intervals. We suggest one way here, Pivotal Intervals.

Let $\theta = T(P)$ and $\hat{\theta}_n = T(P_n)$ and define the pivot $R_n = \hat{\theta}_n - \theta$. Let $\hat{\theta}_{n,1}^*, \dots, \hat{\theta}_{n,B}^*$ denote bootstrap replications of $\hat{\theta}_n$. Let $H(r)$ denote the cdf of the pivot:

$$H(r) = \mathbb{P}(R_n \leq r).$$

Define

$$C_n^* = \left(\hat{\theta}_n - H^{-1} \left(1 - \frac{\alpha}{2} \right), \hat{\theta}_n - H^{-1} \left(\frac{\alpha}{2} \right) \right).$$

Then it follows that

$$\begin{aligned} \mathbb{P}(\theta \in C_n^*) &= \mathbb{P} \left(\hat{\theta}_n - H^{-1} \left(1 - \frac{\alpha}{2} \right) \leq \theta \leq \hat{\theta}_n - H^{-1} \left(\frac{\alpha}{2} \right) \right) \\ &= \mathbb{P} \left(H^{-1} \left(\frac{\alpha}{2} \right) \leq \hat{\theta}_n - \theta \leq H^{-1} \left(1 - \frac{\alpha}{2} \right) \right) \\ &= \mathbb{P} \left(H^{-1} \left(\frac{\alpha}{2} \right) \leq R_n \leq H^{-1} \left(1 - \frac{\alpha}{2} \right) \right) \\ &= H \left(H^{-1} \left(1 - \frac{\alpha}{2} \right) \right) - H \left(H^{-1} \left(\frac{\alpha}{2} \right) \right) \\ &= 1 - \alpha. \end{aligned}$$

Hence, C_n^* is an exact $1 - \alpha$ confidence interval for θ . Unfortunately, computing C_n^* depends on the unknown distribution H but we can form a bootstrap estimate of H :

$$\hat{H}(r) = \frac{1}{n} \sum_{b=1}^B I(R_{n,b}^* \leq r),$$

where $R_{n,b}^* = \hat{\theta}_{n,b}^* - \hat{\theta}_n$. Let r_β^* denote the β sample quantile of $(R_{n,1}^*, \dots, R_{n,B}^*)$. It follows that the $1 - \alpha$ bootstrap confidence interval is

$$C_n = \left(\hat{\theta}_n - r_{1-\alpha/2}^*, \hat{\theta}_n - r_{\alpha/2}^* \right).$$

Minimax

When solving a statistical learning problem, there are often many procedures to choose from. This leads to the following question: how can we tell if one statistical learning procedure is better than another? One answer is provided by *minimax theory* which is a set of techniques for finding the minimum, worst case behavior of a procedure.

Definition. Let \mathcal{P} be a set of distributions and let X_1, \dots, X_n be a sample from some distribution $P \in \mathcal{P}$. Let $\theta(P)$ be a parameter, which is some function of P . Let $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ denote an estimator, which is a function of a sample. Given a metric d , the *minimax risk* is

$$R_n \equiv R_n(\mathcal{P}) = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[d(\hat{\theta}, \theta(P))] \quad (1)$$

where the infimum is over all estimators.

For example, $\theta(P)$ could be the mean of P , the variance of P or the density of P . $\hat{\theta}(X_1, \dots, X_n)$ can be the sample mean $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$, the sample variance $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$, the kernel density estimator \hat{p}_h , etc.

Definition. (i) An estimator $\hat{\theta}$ is a minimax estimator if $\sup_{P \in \mathcal{P}} \mathbb{E}_P[d(\hat{\theta}, \theta(P))] = R_n$.

(ii) An estimator $\hat{\theta}$ is a (asymptotic) minimax estimator if $\sup_{P \in \mathcal{P}} \mathbb{E}_P[d(\hat{\theta}, \theta(P))] = O(R_n)$.

Example. Suppose that $\mathcal{P} = \{N(\theta, 1) : \theta \in \mathbb{R}\}$ where $N(\theta, 1)$ denotes a Gaussian with mean θ and variance 1. Consider estimating θ with the metric $d(a, b) = (a - b)^2$. The minimax risk is

$$R_n = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[(\hat{\theta} - \theta)^2]. \quad (2)$$

In this example, θ is a scalar.

The minimax risk is $R_n = 1/n$ and \bar{X}_n is a minimax estimator.

Example. Suppose that \mathcal{P} is the set of densities with uniformly bounded second derivatives. Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be a sample from a distribution P . Let $m(x) = \mathbb{E}_P(Y|X = x) = \int y dP(y|X = x)$ be the regression function. In this case, we might use the metric $d(m_1, m_2) = \int (m_1(x) - m_2(x))^2 dx$ in which case the minimax risk is

$$R_n = \inf_{\hat{m}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\int (\hat{m}(x) - m(x))^2 dx \right]. \quad (3)$$

In this example, θ is a function.

The minimax risk is $R_n = \Theta(n^{-4/(4+d)})$ and the kernel density estimator is a minimax estimator.

Kernel Density Estimation

Definition. A kernel function $K : \mathbb{R}^d \rightarrow \mathbb{R}$ is a function satisfying $\int K(x) dx = 1$.

Remark. It is usually assumed that $K(x) \geq 0$ for all $x \in \mathbb{R}^d$, i.e., nonnegative, which makes the computation much cleaner. However, for faster rate of convergence, it is inevitable to allow negative values to the kernel function.

For 1-dimension, some commonly used kernels are the following:

| | | | |
|----------------------|-----------------------------------|------------------|--|
| Boxcar: | $K(x) = \frac{1}{2}I(x)$ | Gaussian: | $K(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ |
| Epanechnikov: | $K(x) = \frac{3}{4}(1 - x^2)I(x)$ | Tricube: | $K(x) = \frac{70}{81}(1 - x ^3)^3I(x)$ |

where $I(x) = 1$ if $|x| \leq 1$ and $I(x) = 0$ otherwise. These kernels are plotted in Figure 1. Two commonly used multivariate kernels are $\prod_{j=1}^d K(x_j)$ and $K(\|x\|)$.

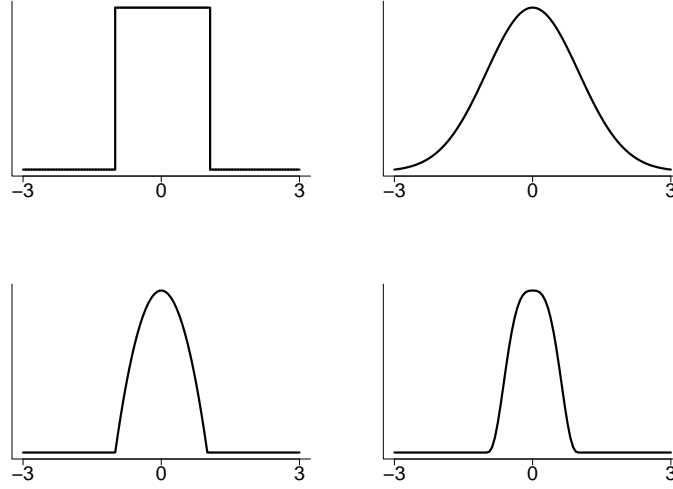


Figure 1: Examples of smoothing kernels: boxcar (top left), Gaussian (top right), Epanechnikov (bottom left), and tricube (bottom right).

Definition. Suppose that $X_1, \dots, X_n \in \mathbb{R}^d$. Given a kernel K and a positive number h , called the bandwidth, the kernel density estimator is defined to be

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h^d} K\left(\frac{\|x - X_i\|}{h}\right). \quad (4)$$

More generally, we define

$$\hat{p}_H(x) = \frac{1}{n} \sum_{i=1}^n K_H(x - X_i),$$

where H is a positive definite bandwidth matrix and $K_H(x) = |H|^{-1/2} K(H^{-1/2}x)$.

For simplicity, we will take $H = h^2 I$ and we get back the previous formula.

Sometimes we write the estimator as \hat{p}_h to emphasize the dependence on h . In the multivariate case the coordinates of X_i should be standardized so that each has the same variance, since the norm $\|x - X_i\|$ treats all coordinates as if they are on the same scale.

The kernel estimator places a smoothed out lump of mass of size $1/n$ over each data point X_i ; see Figure 2. The choice of kernel K is not crucial, but the choice of bandwidth h is important. Small bandwidths give very rough estimates while larger bandwidths give smoother estimates.

0.1 Confidence Bands

To get a confidence band we use the bootstrap. Let P_n be the empirical distribution of X_1, \dots, X_n . The idea is to estimate the distribution

$$F_n(t) = \mathbb{P}\left(\sqrt{nh^d} \|\hat{p}_h(x) - p_h(x)\|_\infty \leq t\right)$$

with the bootstrap estimator

$$\hat{F}_n(t) = \mathbb{P}\left(\sqrt{nh^d} \|\hat{p}_h^*(x) - \hat{p}_h(x)\|_\infty \leq t \mid X_1, \dots, X_n\right)$$

where \hat{p}_h^* is constructed from the bootstrap sample $X_1^*, \dots, X_n^* \sim P_n$. Then

$$\sup_t |F_n(t) - \hat{F}_n(t)| \xrightarrow{P} 0.$$

Here is the algorithm.

1. Let P_n be the empirical distribution that puts mass $1/n$ at each data point X_i .
2. Draw $X_1^*, \dots, X_n^* \sim P_n$. This is called a bootstrap sample.

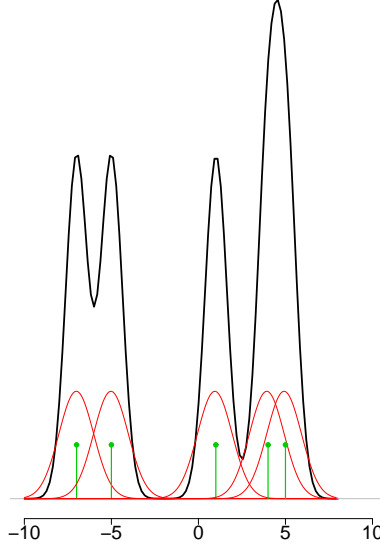


Figure 2: A kernel density estimator \hat{p} . At each point x , $\hat{p}(x)$ is the average of the kernels centered over the data points X_i . The data points are indicated by short vertical bars. The kernels are not drawn to scale.

3. Compute the density estimator \hat{p}_h^* based on the bootstrap sample.
4. Compute $R = \sup_x \sqrt{nh^d} \|\hat{p}_h^* - \hat{p}_h\|_\infty$.
5. Repeat steps 2-4 B times. This gives R_1, \dots, R_B .
6. Let z_α be the upper α quantile of the R_j 's. Thus

$$\frac{1}{B} \sum_{j=1}^B I(R_j > z_\alpha) \approx \alpha.$$

7. Let

$$\ell_n(x) = \hat{p}_h(x) - \frac{z_\alpha}{\sqrt{nh^d}}, \quad u_n(x) = \hat{p}_h(x) + \frac{z_\alpha}{\sqrt{nh^d}}.$$

Theorem. Under appropriate (very weak) conditions, we have

$$\liminf_{n \rightarrow \infty} \mathbb{P}(\ell_n(x) \leq p_h(x) \leq u_n(x) \quad \text{for all } x) \geq 1 - \alpha.$$

See Figure 3.

References

[Durrett(2010)] Rick Durrett. *Probability: theory and examples*, volume 31 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, fourth edition, 2010. ISBN 978-0-521-76539-8. doi: 10.1017/CBO9780511779398. URL <https://doi.org/10.1017/CBO9780511779398>.

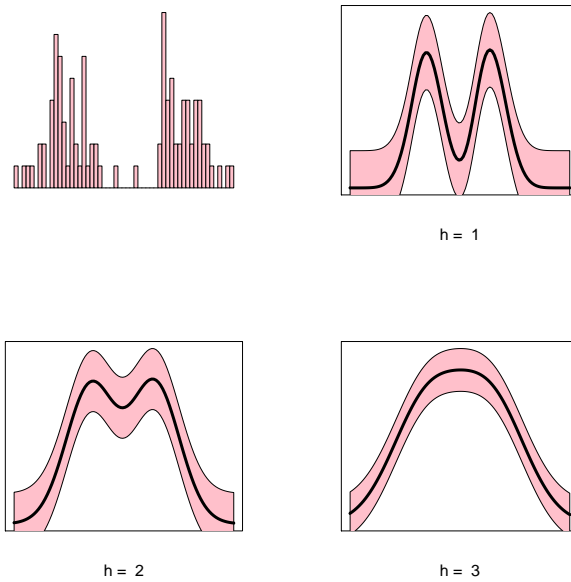


Figure 3: 95 percent bootstrap confidence bands using various bandwidths.