

Random Variables

A **random variable** X is a map $X : \Omega \rightarrow \mathbb{R}$. For $A \subset \mathbb{R}$ we write

$$\mathbb{P}(X \in A) = \mathbb{P}(\{w \in \Omega : X(w) \in A\}) \quad (1)$$

The **cdf** F_X of X is

$$F_X(x) = \mathbb{P}(X \leq x) \quad (2)$$

For continuous X , the **pdf** f_X is a function satisfying

$$\int_A f_X(x) dx = \mathbb{P}(X \in A) \quad (3)$$

Note that $f_X = F'_X$.

Transformations

Let $Y = g(X)$, $\mathcal{X} = \{x : f_X(x) > 0\}$, and $\mathcal{Y} = \{y : y = g(x), x \in \mathcal{X}\}$ (\mathcal{X} and \mathcal{Y} called the *support* of X and Y). Then $\forall A \subset \mathcal{Y}$

$$\mathbb{P}(Y \in A) = \mathbb{P}(X \in \{x : g(x) \in A\}) \quad (4)$$

For the cdf F_Y

$$F_Y(y) = \mathbb{P}(X \in \{x : g(x) \leq y\}) = \int_{\{x : g(x) \leq y\}} f_X(x) dx \quad (5)$$

For g monotonic

$$F_Y(y) = \begin{cases} F_X(g^{-1}(y)) & \text{if } g \text{ increasing} \\ 1 - F_X(g^{-1}(y)) & \text{if } g \text{ decreasing} \end{cases} \quad (6)$$

Additionally, for g monotonic, if $g^{-1}(y)$ has a continuous derivative on \mathcal{Y}

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right| \quad \text{for } y \in \mathcal{Y} \quad (7)$$

Expected Values

The **mean** or **expected value** of $g(X)$ is

$$\mathbb{E}(g(X)) = \int g(x) dF(x) = \int g(x) dP(x) \quad (8)$$

Related properties and definitions:

$$(a) \quad \mu = \mathbb{E}(X) \quad (9)$$

$$(b) \quad \mathbb{E}(\sum_i c_i g_i(X_i)) = \sum_i c_i \mathbb{E}(g_i(X_i)) \quad (10)$$

$$(c) \quad \mathbb{E}\left(\prod_i X_i\right) = \prod_i \mathbb{E}(X_i), \quad X_1, \dots, X_n \text{ indep't} \quad (11)$$

$$(d) \quad \text{Var}(X) = \sigma^2 = \mathbb{E}((X - \mu)^2) \quad \text{is the } \mathbf{variance} \text{ of } X \quad (12)$$

$$(e) \quad \text{Var}(X) = \mathbb{E}(X^2) - \mu^2 \quad (13)$$

$$(f) \quad \text{Var}\left(\sum_i a_i X_i\right) = \sum_i a_i^2 \text{Var}(X_i), \quad X_1, \dots, X_n \text{ indep't} \quad (14)$$

$$(g) \quad \text{Cov}(X, Y) = \mathbb{E}((X - \mu_X)(Y - \mu_Y)) \quad \text{is the } \mathbf{covariance} \quad (15)$$

$$(h) \quad \text{Cov}(X, Y) = \mathbb{E}(XY) - \mu_X \mu_Y \quad (16)$$

$$(i) \quad \rho(X, Y) = \text{Cov}(X, Y) / \sigma_X \sigma_Y, \quad -1 \leq \rho(X, Y) \leq 1 \quad (17)$$

The **conditional expectation** of Y given X is the random variable $g(X) = \mathbb{E}(Y|X)$, where

$$\mathbb{E}(Y|X = x) = \int y f(y|x) dy \quad (18)$$

$$\text{and } f(y|x) = f_{X,Y}(x, y) / f_X(x) \quad (19)$$

The *Law of Total/Iterated Expectation* is

$$\mathbb{E}(Y) = \mathbb{E}[\mathbb{E}(Y|X)] \quad (20)$$

The *Law of Total Variance* is

$$\text{Var}(Y) = \text{Var}[\mathbb{E}(Y|X)] + \mathbb{E}[\text{Var}(Y|X)] \quad (21)$$

The *Law of Total Covariance* is

$$\text{Cov}(X, Y) = \mathbb{E}(\text{Cov}(X, Y|Z)) + \text{Cov}(\mathbb{E}(X|Z), \mathbb{E}(Y|Z)) \quad (22)$$

Moment Generating Function

The **mgf** of X is

$$M_X(t) = \mathbb{E}(e^{tX}) \quad (23)$$

Properties:

$$(a) \quad M_X^{(n)}(t)|_{t=0} = \mathbb{E}(X^n) \quad \text{is the } \mathbf{n^{th} \text{ moment}} \text{ of } X \quad (24)$$

$$(b) \quad M_X(t) = M_Y(t) \quad \forall t \text{ around } 0 \implies X \stackrel{d}{=} Y \quad (25)$$

$$(c) \quad M_{aX+b}(t) = e^{bt} M_X(at) \quad (26)$$

$$(d) \quad M_{\sum_i X_i}(t) = \prod_i M_{X_i}, \quad X_1, \dots, X_n \text{ indep't} \quad (27)$$

Independence

Random variables X and Y are **independent**, written $X \perp\!\!\!\perp Y$, iff

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A) \mathbb{P}(Y \in B) \quad (28)$$

If (X, Y) is a random vector with pdf $f_{X,Y}$, then

$$X \perp\!\!\!\perp Y \iff f_{X,Y}(x, y) = f_X(x) f_Y(y) \quad (29)$$

Distributions

Some discrete distributions:

$$(a) \quad \text{Bernoulli } f(x|p) = p^x (1-p)^{1-x}, \quad x \in \{0, 1\} \\ \text{Mean} = p, \text{Var} = p(1-p) \quad (30)$$

$$(b) \quad \text{Binomial } f(x|n, p) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x \in \{0, 1, \dots, n\} \\ \text{Mean} = np, \text{Var} = np(1-p) \quad (31)$$

$$(c) \quad \text{Poisson } f(x|\lambda) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x \in \{0, 1, 2, \dots\} \\ \text{Mean} = \lambda, \text{Var} = \lambda \quad (32)$$

Some continuous distributions:

$$(a) \quad \text{Uniform } f(x|a, b) = \frac{1}{b-a}, \quad x \in [a, b] \\ \text{Mean} = (b+a)/2, \text{Var} = (b-a)^2/12 \quad (33)$$

$$(b) \quad \text{Normal } f(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}, \quad x \in \mathbb{R} \\ \text{Mean} = \mu, \text{Var} = \sigma^2 \quad (34)$$

$$(c) \quad \text{Gamma } f(x|\alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, \quad x \in \mathbb{R}_+, \alpha, \beta > 0 \\ \text{Mean} = \alpha\beta, \text{Var} = \alpha\beta^2 \quad (35)$$

$$(d) \quad \text{Exponential } f(x|\beta) = \frac{1}{\beta} e^{-x/\beta}, \quad x \in \mathbb{R}_+, \beta > 0 \\ \text{Mean} = \beta, \text{Var} = \beta^2 \quad (36)$$

$$(d) \quad \text{Chi-Squared } f(x|p) = \frac{x^{(p/2)-1} e^{-x/2}}{\Gamma(p/2) 2^{p/2}}, \quad x \in \mathbb{R}_+, p = 1, 2, 3, \dots \\ \text{Mean} = p, \text{Var} = 2p \quad (37)$$

Probability Inequalities

Thm 1 (Gaussian Tail Inequality): Let $X \sim \mathcal{N}(0, 1)$. Then

$$\mathbb{P}(|X| > \epsilon) \leq \frac{2}{\epsilon} e^{-\epsilon^2/2} \quad (38)$$

Additionally:

$$\mathbb{P}(|\bar{X}_n| > \epsilon) \leq \frac{1}{\sqrt{n}\epsilon} e^{-n\epsilon^2/2} \quad (39)$$

Thm 2 (Markov Inequality): Let X be a non-negative random variable s.t. $\mathbb{E}(X)$ exists. Then $\forall t > 0$

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}(X)}{t} \quad (40)$$

Thm 3 (Chebyshev's Inequality): Let $\mu = \mathbb{E}(X)$ and $\sigma^2 = \text{Var}(X)$. Then

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2} \quad (41)$$

$$\mathbb{P}(|(X - \mu)/\sigma| \geq t) \leq \frac{1}{t^2} \quad (42)$$

Lemma 4: Let $\mathbb{E}(X) = 0$ and $a \leq X \leq b$. Then

$$\mathbb{E}(e^{tX}) \leq e^{t^2(b-a)^2/8} \quad (43)$$

Lemma 5: Let X be any random variable. Then

$$\mathbb{P}(X > \epsilon) \leq \inf_{t \geq 0} e^{-t\epsilon} \mathbb{E}(e^{tX}) \quad (44)$$

Thm 6 (Hoeffding's Inequality): X_1, \dots, X_n iid, $\mathbb{E}(X_i) = \mu$, $a \leq X_i \leq b$. Then $\forall \epsilon > 0$

$$\mathbb{P}(|\bar{X} - \mu| \geq \epsilon) \leq 2e^{-2n\epsilon^2/(b-a)^2} \quad (45)$$

Thm 9 (McDiarmid): X_1, \dots, X_n indep't. If $\sup_{x_1, \dots, x_n, x'_i} |g(x_1, \dots, x_n) - g_i^*(x_1, \dots, x_n)| \leq c_i \quad \forall i, \implies$

$$\mathbb{P}(g(X_1, \dots, X_n) - \mathbb{E}(g(X_1, \dots, X_n)) \geq \epsilon) \leq e^{-2\epsilon^2/\sum_i c_i^2} \quad (46)$$

where $g_i^* = g$ with x_i replaced by x'_i .

Shattering

F a finite set, $|F| = n$, and $G \subset F$. \mathcal{A} is a class of sets.

\mathcal{A} **picks out** G if $\exists A \in \mathcal{A}$ s.t. $A \cap F = G$.

Let $S(\mathcal{A}, F) = \#\{G \subset F \text{ picked out by } \mathcal{A}\} \leq 2^n$.

F is **shattered** by \mathcal{A} if $S(\mathcal{A}, F) = 2^n$ (ie if \mathcal{A} picks out all $G \subset F$).

Let \mathcal{F}_n be all finite sets with n elements.

The **shatter coefficient** $s_n(\mathcal{A}) = \sup_{F \in \mathcal{F}_n} s(\mathcal{A}, F) \leq 2^n$.

The **VC dimension** $d(\mathcal{A})$ is the largest n s.t. $s_n(\mathcal{A}) = 2^n$.

Thm 5: $\forall \epsilon > 0, \mathbb{P}(\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| > \epsilon) \leq 8s_n(\mathcal{A})e^{-n\epsilon^2/32}$

Random Samples

For $X_1, \dots, X_n \sim F$ a **statistic** is any $T = g(X_1, \dots, X_n)$.

E.g. $\bar{X}_n, S_n^2 = \sum_i (X_i - \bar{X}_n)^2/(n-1), (X_{(1)}, \dots, X_{(n)})$

Notes: $\mathbb{E}(\bar{X}_n) = \mathbb{E}(X_i), \text{Var}(\bar{X}_n) = \text{Var}(X_i)/n, \mathbb{E}(S_n^2) = \text{Var}(X_i)$

$X_{1, \dots, n} \sim \text{Bern}(p) \implies \sum_i X_i \sim \text{Bin}(n, p)$

$X_{1, \dots, n} \sim \text{Exp}(\beta) \implies \sum_i X_i \sim \Gamma(n, \beta)$

$X_{1, \dots, n} \sim \mathcal{N}(0, 1) \implies \sum_i X_i^2 \sim \chi_n$.

Thm. 1: $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2) \implies \bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$.

Convergence

X, X_1, X_2, \dots random variables.

(1) X_n converges **almost surely** $X_n \xrightarrow{a.s.} X$ if $\forall \epsilon > 0$

$$\mathbb{P}(\lim_{n \rightarrow \infty} |X_n - X| < \epsilon) = 1 \quad (47)$$

(2) X_n converges **in probability** $X_n \xrightarrow{p} X$ if $\forall \epsilon > 0$

$$\lim_{n \rightarrow \infty} \mathbb{P}(|X_n - X| > \epsilon) = 0 \quad (48)$$

(3) X_n converges **in quadratic mean** $X_n \xrightarrow{qm} X$ if

$$\lim_{n \rightarrow \infty} \mathbb{E}[(X_n - X)^2] = 0 \quad (49)$$

(4) X_n converges **in distribution** $X_n \rightsquigarrow X$ if

$$\lim_{n \rightarrow \infty} F_{X_n}(t) = F_X(t) \quad (50)$$

$\forall t$ on which F_X is continuous.

Thm 7: Conv. a.s. and in q.m. imply conv. in prob. All three imply conv. in distribution. Conv. in distribution to a point-mass also implies conv. in prob.

Thm 10a: X, X_n, Y, Y_n random variables. Then

$$(a) \quad X_n \xrightarrow{p} X, Y_n \xrightarrow{p} Y \implies X_n + Y_n \xrightarrow{p} X + Y \quad (51)$$

$$(b) \quad X_n \xrightarrow{p} X, Y_n \xrightarrow{p} Y \implies X_n Y_n \xrightarrow{p} XY \quad (52)$$

$$(c) \quad X_n \xrightarrow{qm} X, Y_n \xrightarrow{qm} Y \implies X_n + Y_n \xrightarrow{qm} X + Y \quad (53)$$

Thm 10b (Slutzky's Thm): X, X_n, Y_n random variables. Then

$$(a) \quad X_n \rightsquigarrow X, Y_n \rightsquigarrow c \implies X_n + Y_n \rightsquigarrow X + c \quad (54)$$

$$(b) \quad X_n \rightsquigarrow X, Y_n \rightsquigarrow c \implies X_n Y_n \rightsquigarrow cX \quad (55)$$

Thm 12 (Law of Large Numbers): X_1, \dots, X_n iid, $\mathbb{E}(X_i) = \mu \implies \bar{X}_n \xrightarrow{qm} \mu$.

Thm 14 (CLT): X_1, \dots, X_n iid, $\mathbb{E}(X_i) = \mu, \text{Var}(X_i) = \sigma^2$

$$\implies \sqrt{n}(\bar{X}_n - \mu)/\sigma \rightsquigarrow \mathcal{N}(0, 1)$$

$$\implies \sqrt{n}(\bar{X}_n - \mu) \rightsquigarrow \mathcal{N}(0, \sigma^2)$$

$$\implies \bar{X}_n \approx \mathcal{N}(\mu, \sigma^2/n)$$

$$\implies \sqrt{n}(\bar{X}_n - \mu)/S_n \rightsquigarrow \mathcal{N}(0, 1)$$

Thm 18 (delta method): If $\sqrt{n}(Y_n - \mu)/\sigma \rightsquigarrow \mathcal{N}(0, 1), g'(\mu) \neq 0$

$$\implies \sqrt{n}(g(Y_n) - g(\mu))/|g'(\mu)|\sigma \rightsquigarrow \mathcal{N}(0, 1)$$

$$\text{ie } Y_n \approx \mathcal{N}(\mu, \sigma^2/n) \implies g(Y_n) \approx \mathcal{N}(g(\mu), g'(\mu)^2 \sigma^2/n)$$

Thm 18b (2nd order delta method):

Sufficiency

If $X_1, \dots, X_n \sim p(x; \theta)$, T **sufficient** for θ if $p(x^n|t; \theta) = p(x^n|t)$.

Thm 9 (factorization): for $X^n \sim p(x; \theta)$, $T(X^n)$ sufficient for θ if the joint probability can be factorized as

$$p(x^n; \theta) = h(x^n) \times g(t; \theta) \quad (56)$$

T is a **minimal sufficient statistic (MSS)** if T is sufficient and $T = g(U)$ for all other sufficient stats U .

Thm 15: T is a MSS if:

$$\frac{p(y^n; \theta)}{p(x^n; \theta)} \text{ is constant in } \theta \iff T(y^n) = T(x^n) \quad (57)$$

Parametric Point Estimation

Method of Moments: Define equations

- (a) $(\sum_i X_i)/n = \mathbb{E}_{\hat{\theta}}(X_i)$
- (b) $(\sum_i X_i^2)/n = \mathbb{E}_{\hat{\theta}}(X_i^2)$
- (c) ...

and solve for $\hat{\theta}$.

Maximum Likelihood (MLE): The MLE is

$$\hat{\theta} = \arg \max_{\theta} L(\theta) = \arg \max_{\theta} l(\theta) \quad (58)$$

Often suffices to solve for θ in $\frac{\partial l(\theta)}{\partial \theta} = 0$.

The MLE is **equivariant** \implies if $\eta = g(\theta)$ then $\hat{\eta} = g(\hat{\theta})$.

Bayes Estimation: For prior $\pi(\theta)$, choose

$$\hat{\theta} = \mathbb{E}(\theta|x^n) = \int \theta \pi(\theta|x^n) d\pi \quad (59)$$

Mean Squared Error (MSE): The MSE is

$$\text{MSE} = \mathbb{E}(\hat{\theta} - \theta)^2 = \int (\hat{\theta} - \theta)^2 p(x^n; \theta) dx^n = \text{bias}(\hat{\theta})^2 + \text{Var}(\hat{\theta}) \quad (60)$$

Defs: **bias**($\hat{\theta}$) = $\mathbb{E}(\hat{\theta}) - \theta$. We say $\hat{\theta}$ is **consistent** if $\hat{\theta} = \hat{\theta}_n \xrightarrow{p} \theta$.

The **standard error** of $\hat{\theta}$, $\text{se}(\hat{\theta})$, is the standard deviation of $\hat{\theta}$.

Risks and Estimators

$L(\theta, \hat{\theta})$ is the **loss** of an estimator $\hat{\theta} = \hat{\theta}(x^n)$ for $x^n \sim p(x^n; \theta)$.

The **risk** of this $\hat{\theta}$ is

$$R(\theta, \hat{\theta}) = \mathbb{E}[L(\theta, \hat{\theta})] = \int L(\theta, \hat{\theta}) p(x^n; \theta) dx^n \quad (61)$$

When $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$, the risk is the MSE.

The **max risk** of $\hat{\theta}$ over a set $\theta \in \Theta$ is

$$\bar{R}(\hat{\theta}) = \sup_{\theta \in \Theta} R(\theta, \hat{\theta}) \quad (62)$$

The **minimax estimator** is

$$\hat{\theta} = \arg \inf_{\hat{\theta}} \bar{R}(\hat{\theta}) \quad (63)$$

The **Bayes risk** of $\hat{\theta}$ given a prior $\pi(\theta)$ is

$$B_{\pi}(\hat{\theta}) = \int R(\theta, \hat{\theta}) \pi(\theta) d\theta \quad (64)$$

The **posterior risk** of $\hat{\theta}$ given a prior $\pi(\theta)$ is

$$r(\hat{\theta}|x^n) = \int L(\theta, \hat{\theta}) \pi(\theta|x^n) d\theta \quad (65)$$

where $\pi(\theta|x^n) = \frac{\mathbb{P}(x^n; \theta) \pi(\theta)}{m(x^n)}$ is the posterior over θ .

The **Bayes estimator** is

$$\hat{\theta} = \arg \inf_{\hat{\theta}} B_{\pi}(\hat{\theta}) = \arg \inf_{\hat{\theta}} r(\hat{\theta}|x^n) \quad (66)$$

which equals the posterior mean $\mathbb{E}(\theta|x^n)$ when $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$, the posterior median when $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$, and the posterior mode when $L(\theta, \hat{\theta}) = \mathbb{I}[\theta \neq \hat{\theta}]$.

Thm 10: If $\hat{\theta}$ is a Bayes estimator for some prior π and $R(\theta, \hat{\theta})$ is constant, then $\hat{\theta}$ is a minimax estimator.

Note: The MLE is approximately minimax (as n increases, if the dimension of the parameter is fixed).

Aymptotic (Large Sample) Theory

A random sequence A_n is:

$$(a) \quad o_p(1) \text{ if } A_n \xrightarrow{p} 0 \quad (67)$$

$$(b) \quad o_p(B_n) \text{ if } A_n/B_n \xrightarrow{p} 0 \quad (68)$$

$$(c) \quad O_p(1) \text{ if } \forall \epsilon > 0, \exists M : \lim_{n \rightarrow \infty} \mathbb{P}(|A_n| > M) < \epsilon \quad (69)$$

$$(d) \quad O_p(B_n) \text{ if } A_n/B_n = O_p(1) \quad (70)$$

If $Y_n \rightsquigarrow Y \implies Y_n = O_p(1)$

If $\sqrt{n}(Y_n - c) \rightsquigarrow Y \implies Y_n = O_p(1/\sqrt{n})$

Distances Between Distributions

For distributions P and Q with pdfs p and q :

$$(a) \quad V(P, Q) = \sup_A |P(A) - Q(A)| \quad \text{total variation distance} \quad (71)$$

$$(b) \quad K(P, Q) = \int p \log(p/q) \quad \text{Kullback-Leibler divergence} \quad (72)$$

$$(c) \quad d_2(P, Q) = \int (p - q)^2 \quad \text{L}_2 \text{ distance} \quad (73)$$

A model is **identifiable** if: $\theta_1 \neq \theta_2 \implies K(\theta_1, \theta_2) > 0$.

Consistency

$\hat{\theta}_n = T(X^n)$ is **consistent** for θ if $\hat{\theta}_n \xrightarrow{p} \theta$ (ie if $\hat{\theta}_n - \theta = o_p(1)$).

To show consistency, can show: $\text{Bias}^2(\hat{\theta}_n) + \text{Var}(\hat{\theta}_n) \rightarrow 0$.

The MLE is consistent under regularity conditions.

MLE not consistent when number of params (or support?) grows.

Score and Fisher Information

The **score function** is $S(\theta) = \frac{\partial}{\partial \theta} l(\theta) = \frac{\partial}{\partial \theta} \sum_{i=1}^n \log p(x_i|\theta)$.

The **Fisher information** is defined as

$$I_n(\theta) = \mathbb{E}_{\theta} [S(\theta)^2] = \text{Var}_{\theta} [S(\theta)] = -\mathbb{E}_{\theta} \left[\frac{\partial^2}{\partial \theta^2} l(\theta) \right] \quad (74)$$

$$\text{and } I_n(\theta) = -n \mathbb{E} \left[\frac{\partial^2}{\partial \theta^2} \log p(X_1; \theta) \right] = n I_1(\theta).$$

The **observed information** $\hat{I}_n(\theta) = -\sum_i \frac{\partial^2}{\partial \theta^2} \log p(X_i; \theta)$.

$$\text{Vector case: } S(\theta) = \left[\frac{\partial l(\theta)}{\partial \theta_i} \right]_{i=1, \dots, K} \quad I_{ij} = -\mathbb{E}_{\theta} \left[\frac{\partial^2 l(\theta)}{\partial \theta_i \partial \theta_j} \right]_{i,j=1, \dots, K}$$

Efficiency and Robustness

For an estimator $\hat{\theta}_n(X^n)$ of θ , where $X^n \stackrel{\text{iid}}{\sim} p(x|\theta)$:

(64) If $\sqrt{n}(\hat{\theta}_n - \theta) \rightsquigarrow \mathcal{N}(0, v^2)$, then v^2 is the **asymptotic-Var**($\hat{\theta}_n$).

E.g. for $\hat{\theta}_n = \bar{X}_n$: $v^2 = \sigma^2 = \text{Var}(X_i) = \lim_{n \rightarrow \infty} n \text{Var}(\bar{X}_n)$.

In general, asymptotic-Var($\hat{\theta}_n$) $v^2 \neq \lim_{n \rightarrow \infty} n \text{Var}(\hat{\theta}_n)$.

Therefore: $\text{Var}(\hat{\theta}_n) = (\text{se})^2 \approx v^2/n$.

(65) For param $\tau(\theta)$, $v(\theta) = \frac{|\tau'(\theta)|^2}{I_1(\theta)}$ is the **Cramer-Rao lower bound** for most estimators $v(\theta) \leq v^2$.

If $\sqrt{n}(\hat{\theta}_n - \tau(\theta)) \rightsquigarrow \mathcal{N}(0, v(\theta))$ (ie if $v^2 = v(\theta)$) $\implies \hat{\theta}_n$ **efficient**.

usually, $\sqrt{n}(\tau(\hat{\theta}_{\text{MLE}}) - \tau(\theta)) \rightsquigarrow \mathcal{N}(0, v(\theta)) \implies$ MLE efficient.

The **standard error** of efficient $\hat{\theta}_n$ is $se = \sqrt{\text{Var}(\hat{\theta}_n)} \approx \sqrt{\frac{1}{I_n(\hat{\theta}_n)}}$.

The **estimated standard error** of efficient $\hat{\theta}_n$ is $\hat{se} \approx \sqrt{\frac{1}{I_n(\hat{\theta}_n)}}$.

For efficient $\hat{\theta}_n$, $\hat{\tau} = \tau(\hat{\theta}_n)$, $se \approx \sqrt{\frac{|\tau'(\hat{\theta}_n)|^2}{I_n(\hat{\theta}_n)}}$, and $\hat{se} \approx \sqrt{\frac{|\tau'(\hat{\theta}_n)|^2}{I_n(\hat{\theta}_n)}}$.

In general, **asymptotic normality** is when:

$$\frac{\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n)}{\sqrt{\text{Var}(\hat{\theta}_n)}} \rightsquigarrow \mathcal{N}(0, 1) \implies \hat{\theta}_n \rightsquigarrow \mathcal{N}(\mathbb{E}(\hat{\theta}_n), \text{Var}(\hat{\theta}_n)).$$

If $\sqrt{n}(W_n - \tau(\theta)) \rightsquigarrow \mathcal{N}(0, \sigma_W^2)$ and $\sqrt{n}(V_n - \tau(\theta)) \rightsquigarrow \mathcal{N}(0, \sigma_V^2)$ \implies **asymptotic relative efficiency** $\text{ARE}(V_n, W_n) = \sigma_W^2 / \sigma_V^2$.

Often there is a tradeoff between efficiency and robustness. (?)

Hypothesis Testing

Null hypothesis $H_0 : \theta \in \Theta_0$, **alternative** $H_1 : \theta \in \Theta_1$.

Type I error: If H_0 true but we reject H_0 .

To construct a test:

1. Choose a test statistic $W = W(X_1, \dots, X_n)$
2. Choose a rejection region R
3. If $W \in R$, reject H_0 otherwise retain H_0

The **power function** $\beta(\theta) = \mathbb{P}_\theta(W \in R)$ for a rejection region R .

Want **level- α** test ($\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha$) that maximizes $\beta(\theta \in \Theta_1)$.

A level- α test with power fn β is **uniformly most powerful** if:
 $\beta(\theta) \geq \beta'(\theta) \quad \forall \theta \in \Theta_1 \quad \forall \beta' \neq \beta$.

Neyman-Pearson Test

For simple $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$, reject H_0 if $\frac{L(\theta_1)}{L(\theta_0)} > k$.
 where k chosen s.t. $\mathbb{P}(\frac{L(\theta_1)}{L(\theta_0)} > k) = \alpha$.

Wald Test

For $H_0 : \theta = \theta_0$ and $H_1 : \theta \neq \theta_0$, reject H_0 if $\left| \frac{\hat{\theta}_n - \theta_0}{se} \right| > z_{\alpha/2}$.
 where $z_{\alpha/2}$ is the inverse standard-normal CDF of $1 - \frac{\alpha}{2}$.
 and $\hat{\theta}_n$ an estimator s.t. $(\hat{\theta} - \theta)/se \rightsquigarrow \mathcal{N}(0, 1)$ eg: $\hat{\theta}_{mle}$
 and $se = \sqrt{\text{Var}(\hat{\theta}_n)}$. Can also use $\hat{se} =_{\text{eg.}} \sqrt{S_n^2/n}$.
 and if $\hat{\theta}_n$ efficient, can approx: $se \approx \sqrt{\frac{1}{I_n(\theta)}}$ or $\hat{se} \approx \sqrt{\frac{1}{I_n(\hat{\theta}_n)}}$.

Likelihood Ratio Test

For $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \notin \Theta_0$, reject H_0 if $\lambda(x^n) = \frac{L(\hat{\theta}_0)}{L(\hat{\theta})} \leq c$.

where $L(\hat{\theta}_0) = \sup_{\theta \in \Theta_0} L(\theta)$ and $L(\hat{\theta}) = \sup_{\theta \in \Theta} L(\theta)$.

and c chosen s.t. $\mathbb{P}(\lambda(x^n) \leq c) = \alpha$.

Thm: under $H_0 : \theta = \theta_0 \implies W_n = -2\log\lambda(X^n) \rightsquigarrow \chi^2_{1,\alpha}$
 \implies reject H_0 if $W_n > \chi^2_{1,\alpha}$.

Also: for $\theta = (\theta_1, \dots, \theta_k)$, if H_0 fixes some of the parameters
 $\implies -2\log\lambda(X^n) \rightsquigarrow \chi^2_\nu$, where $\nu = \dim(\Theta) - \dim(\Theta_0)$.

P-Values

The **p-value** $p(x^n)$ is the smallest α -level s.t. we reject H_0 .

Thm: For a test of the form: reject H_0 when $W(x^n) > c$,
 $\implies p(x^n) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(W(X^n) \geq W(x^n)) = \sup_{\theta \in \Theta_0} [1 - F(W(x^n)|\theta)]$.

Thm: Under $H_0 : \theta = \theta_0$, $p(x^n) \sim \text{Unif}(0, 1)$.

Permutation Test

$X^n \sim F$, $Y^m \sim G$, $H_0 : F = G$, $H_1 : F \neq G$

Let $Z = (X^n, Y^m)$ and $L = (1, \dots, 1, 2, \dots, 2)$.

Let $W = g(L, Z) = |(\text{ave of 1 labeled pts}) - (\text{ave of 2 labeled pts})|$.

Let $p = \frac{1}{N!} \sum \pi \mathbb{I}(g(L_\pi, Z) > g(L, Z)) \implies$ reject H_0 when $p < \alpha$.

Confidence Intervals

We want a $1 - \alpha$ **confidence interval** $C_n = [L(X^n), U(X^n)]$ s.t.

$\mathbb{P}_\theta(L(X^n) \leq \theta \leq U(X^n)) \geq 1 - \alpha, \quad \forall \theta \in \Theta$.

Generally, a $1 - \alpha$ **confidence set** C_n is a random set $C_n \subset \Theta$ s.t.

$\inf_{\theta \in \Theta} \mathbb{P}_\theta(\theta \in C_n(X^n)) \geq 1 - \alpha$.

Using Probability Inequalities

Prob inequalities give (for eg.) $\mathbb{P}(|\hat{\theta}_n - \theta| > \epsilon) \leq g(\exp^{-f(\epsilon)})_{\text{set to}} \alpha$.

solving for ϵ gives $\epsilon = \tilde{f}(\alpha) \implies \mathbb{P}(|\hat{\theta}_n - \theta| > \tilde{f}(\alpha)) \leq \alpha$

$\implies C_n = (\hat{\theta} - \tilde{f}(\alpha), \hat{\theta} + \tilde{f}(\alpha))$.

Inverting a Test

In level- α tests $\mathbb{P}_{\theta_0}(T(x^n) \in R) \leq \alpha \implies$ let $C_n = \{\theta : T(x^n) \in A(\theta_0)\}$.

where $A(\theta_0) = \{T(x^n) \notin R \mid \theta = \theta_0\}$ (ie the accept region if $\theta = \theta_0$).

For Wald: $C_n = \hat{\theta}_n \pm (z_{\alpha/2} \times se) = \hat{\theta}_n \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$.

For LRT: $C_n = \{\theta : \frac{L(\theta)}{L(\hat{\theta})} > c\}$ (for test where reject H_0 if $\frac{L(\theta_0)}{L(\hat{\theta})} \leq c$).

Pivots

$Q(X^n, \theta)$ a **pivot** if the distribution of Q does not depend on θ .

Find a, b s.t. $\mathbb{P}_\theta(a \leq Q(X^n, \theta) \leq b) \geq 1 - \alpha, \quad \forall \theta$.

$\implies C_n = \{\theta : a \leq Q(X^n, \theta) \leq b\} \geq 1 - \alpha\}$.

Nonparametric Inference

The **empirical CDF** is: $\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(X_i \leq x)$

Thm (DKW): $\forall \epsilon > 0, \mathbb{P}(\sup_x |\hat{F}_n(x) - F(x)| > \epsilon) \leq 2e^{-2n\epsilon^2}$

The **kernel density estimator** is: $\hat{p}_n(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\left(\frac{x - X_i}{h}\right)$
 where K a symmetric zero-mean density, and bandwidth $h > 0$.

Thm: The risk $R = \mathbb{E}(\mathcal{L}(p, \hat{p})) = \int (b^2(x) + \text{Var}(x)) dx = \frac{a}{n^{4/5}}$
 for some a , where $\mathcal{L}(p, \hat{p}) = \int (p(x) - \hat{p}(x))^2 dx$, and
 $b^2(x) = \mathbb{E}(\hat{p}(x)) - p(x)$. And this is minimax.

A **statistical functional** $T(F)$ is any function of the CDF.

A **plug-in estimator** of $\theta = T(F)$ is: $\hat{\theta}_n = T(\hat{F}_n)$.

Often, $\hat{\theta}_n \approx \mathcal{N}(T(F), \hat{se}^2)$, where \hat{se} is estimate of $\sqrt{\text{Var}(T(\hat{F}_n))}$.

Bootstrap

The **bootstrap** is a nonparametric way to find standard errors and confidence intervals of estimators of statistical functionals:

(1) Draw $X_1^*, \dots, X_n^* \sim \hat{F}_n$ (by drawing $X_i^* \sim \{X_1, \dots, X_n\}$ unif).

(2) Compute $T_n^* = g(X_1^*, \dots, X_n^*)$

(3) Do (1) and (2) B times to get $T_{n,1}^*, \dots, T_{n,B}^*$

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

Then: $v_{\text{boot}} \xrightarrow{a.s.} \text{Var}_{\hat{F}_n}(T_n)$ as $B \rightarrow \infty$ and $\hat{se}(T_N) = \sqrt{v_{\text{boot}}}$

Bayesian Inference

Frequentists: probability is long-run frequencies. Procedures are random but parameters are fixed, unknown quantities.

Bayesians: probability is a measure of subjective degree of belief.

Everything is random, including parameters.

Using Bayes Thm \nRightarrow Bayesian inference.

For $X_1, \dots, X_n \sim p(x|\theta)$, and prior $\pi(\theta)$, **Bayes Thm** gives:

$$\pi(\theta|X^n) = \frac{p(X^n|\theta)\pi(\theta)}{m(X^n)} = \frac{p(X^n|\theta)\pi(\theta)}{\int p(X^n|\theta)\pi(\theta)d\theta}$$