Random Variables

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

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

The **cdf** F_X of X is

$$F_X(x) = \mathbb{P}(X \le x) \tag{2}$$

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

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

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 $\forall\ A \subset \mathcal{Y}$

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

For the cdf F_Y

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

For g monotonic

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

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

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

Expected Values

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

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

Related properties and definitions:

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

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

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

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

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

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

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

$$(h)\ \ Cov(X,Y) = \mathbb{E}(XY) - \mu_x\mu_Y \tag{16}$$

$$(i)\ \ \rho(X,Y) = Cov(X,Y)/\sigma_x\sigma_y, \quad -1 \le \rho(X,Y) \le 1 \tag{17}$$

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

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

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

The *Law of Total/Iterated Expectation* is

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

The *Law of Total Variance* is

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

The *Law of Total Covariance* is

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

Moment Generating Function

The **mgf** of X is

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

Properties:

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

$$(b)\ \ M_x(t) = M_Y(t)\ \ \forall t \ \text{ around } 0 \implies X \overset{d}{=} Y \tag{25}$$

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

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

Independence

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

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

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

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

Distributions

Some discrete distributions:

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

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

$$\begin{array}{ll}(c) & \text{Poisson } \ f(x|\lambda) = \dfrac{e^{-\lambda}\lambda^x}{x!}, \quad x \in \{0,1,2,\dots\} \\ & \text{Mean} = \lambda, \text{Var} = \lambda\end{array} \tag{32}$$

Some continuous distributions:

$$\begin{array}{ll}(a) & \text{Uniform } \ f(x|a,b) = \dfrac{1}{b-a}, \quad x \in [a,b] \\ & \text{Mean} = (b+a)/2, \text{Var} = (b-a)^2/12\end{array} \tag{33}$$

$$\begin{array}{ll}(b) & \text{Normal } \ f(x|\mu,\sigma^2) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}, \quad x \in \mathbb{R} \\ & \text{Mean} = \mu, \text{Var} = \sigma^2\end{array} \tag{34}$$

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

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

$$\begin{array}{ll}(e) & \text{Chi-Squared } \ f(x|p) = \dfrac{x^{(p/2)-1}e^{-x/2}}{\Gamma(p/2)2^{p/2}}, \quad x \in \mathbb{R}_+, p = 1,2,3,\dots \\ & \text{Mean} = p, \text{Var} = 2p\end{array} \tag{37}$$

Probability Inequalities

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

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

Additionally:

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

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

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

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

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

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

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

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

Lemma 5: Let X be any random variable. Then

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

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

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

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

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

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 $S(\mathcal{A}, F) = \#\{G \subset F$ 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** $d(\mathcal{A})$ = the largest n s.t. $s_n(\mathcal{A}) = 2^n$.
Thm 5: $\forall \epsilon > 0$, $\mathbb{P}(\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| > \epsilon) \leq 8 s_n(\mathcal{A}) e^{-n\epsilon^2/32}$

Random Samples

For $X_1, \ldots, X_n \sim F$ a **statistic** is any $T = g(X_1, \ldots, X_n)$.
E.g. $\overline{X}_n$, $S_n^2 = \sum_i (X_i - \overline{X}_n)^2/(n-1)$, $\left(X_{(1)}, \ldots, X_{(n)}\right)$
Notes: $\mathbb{E}(\overline{X}_n) = \mathbb{E}(X_i)$, $\mathrm{Var}(\overline{X}_n) = \mathrm{Var}(X_i)/n$, $\mathbb{E}(S_n^2) = \mathrm{Var}(X_i)$

$X_{1,\ldots,n} \sim \mathrm{Bern}(p) \implies \sum_i X_i \sim \mathrm{Bin}(n,p)$
$X_{1,\ldots,n} \sim \mathrm{Exp}(\beta) \implies \sum_i X_i \sim \Gamma(n,\beta)$
$X_{1,\ldots,n} \sim \mathcal{N}(0,1) \implies \sum_i X_i^2 \sim \chi_n$.

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

Convergence

$X, X_1, X_2, \ldots$ random variables.
(1) X_n converges **almost surely** $X_n \xrightarrow{a.s.} X$ if $\forall \epsilon > 0$

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

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

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

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

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

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

$$\lim_{n\to\infty} F_{X_n}(t) = F_X(t) \tag{50}$$

$\forall t$ 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

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

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

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

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

$$(a)\ \ X_n \rightsquigarrow X, Y_n \rightsquigarrow c \implies X_n + Y_n \rightsquigarrow X + c \tag{54}$$

$$(b)\ \ X_n \rightsquigarrow X, Y_n \rightsquigarrow c \implies X_n Y_n \rightsquigarrow cX \tag{55}$$

Thm 11 (Cont's Mapping): g a continuous function, then:

$$(a)\ \ X_n \xrightarrow{p} X \implies g(X_n) \xrightarrow{p} g(X) \tag{56}$$

$$(b)\ \ X_n \rightsquigarrow X \implies g(X_n) \rightsquigarrow g(X) \tag{57}$$

Thm 12 (Law of Large Numbers): $X_1, \ldots, X_n$ iid, $\mathbb{E}(X_i) = \mu$ $\implies \overline{X}_n \xrightarrow{\text{a.s.}} \mu$.
Thm 14 (CLT): $X_1, \ldots, X_n$ iid, $\mathbb{E}(X_i) = \mu$ $\mathrm{Var}(X_i) = \sigma^2$

$\implies \sqrt{n}(\overline{X}_n - \mu)/\sigma \rightsquigarrow \mathcal{N}(0,1)$
$\implies \sqrt{n}(\overline{X}_n - \mu) \rightsquigarrow \mathcal{N}(0,\sigma^2)$
$\implies \overline{X}_n \approx \mathcal{N}(\mu, \sigma^2/n)$
$\implies \sqrt{n}(\overline{X}_n - \mu)/S_n \rightsquigarrow \mathcal{N}(0,1)$

Thm 18 (delta method): If $\sqrt{n}(Y_n - \mu)/\sigma \rightsquigarrow \mathcal{N}(0,1)$, $g'(\mu) \neq 0$
$\implies \sqrt{n}(g(Y_n) - g(\mu))/|g'(\mu)|\sigma \rightsquigarrow \mathcal{N}(0,1)$
ie $Y_n \approx \mathcal{N}(\mu, \sigma^2/n) \implies g(Y_n) \approx \mathcal{N}(g(\mu), g'(\mu)^2\sigma^2/n)$

Sufficiency

If $X_1, \ldots, X_n \sim p(x;\theta)$, 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

$$p(x^n;\theta) = h(x^n) \times g(t;\theta) \tag{58}$$

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:

$$\frac{p(y^n;\theta)}{p(x^n;\theta)} \text{ is constant in } \theta \iff T(y^n) = T(x^n) \tag{59}$$

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

$$\hat\theta = \arg\max_\theta L(\theta) = \arg\max_\theta l(\theta) \tag{60}$$

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

$$\hat\theta = \mathbb{E}(\theta|x^n) = \int \theta\pi(\theta|x^n)d\pi \tag{61}$$

Mean Squared Error (MSE): The MSE is

$$\text{MSE} = \mathbb{E}(\hat\theta-\theta)^2 = \int(\hat\theta-\theta)^2 p(x^n;\theta)dx^n = \text{bias}(\hat\theta)^2 + Var(\hat\theta) \tag{62}$$

bias$(\hat\theta) = \mathbb{E}(\hat\theta) - \theta$.

The **standard error** of $\hat\theta$, se$(\hat\theta)$, is the standard deviation of $\hat\theta$:

$$\text{se}(\hat\theta) = \sqrt{\text{Var}(\hat\theta)} \tag{63}$$

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

$$R(\theta,\hat\theta) = \mathbb{E}[L(\theta,\hat\theta)] = \int L(\theta,\hat\theta)p(x^n;\theta)dx^n \tag{64}$$

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

$$\overline{R}(\hat\theta) = \sup_{\theta\in\Theta} R(\theta,\hat\theta) \tag{65}$$

The **minimax estimator** is

$$\hat\theta = \arg\inf_{\hat\theta} \overline{R}(\hat\theta) \tag{66}$$

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

$$B_\pi(\hat\theta) = \int R(\theta,\hat\theta)\pi(\theta)d\theta \tag{67}$$

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

$$r(\hat\theta|x^n) = \int L(\theta,\hat\theta)\pi(\theta|x^n)d\theta \tag{68}$$

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

$$\hat\theta = \arg\inf_{\hat\theta} B_\pi(\hat\theta) = \arg\inf_{\hat\theta} r(\hat\theta|x^n) \tag{69}$$

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:

$$(a)\ \ o_p(1) \text{ if } A_n \xrightarrow{p} 0 \tag{70}$$

$$(b)\ \ o_p(B_n) \text{ if } A_n/B_n \xrightarrow{p} 0 \tag{71}$$

$$(c)\ \ O_p(1) \text{ if } \forall\epsilon>0, \exists M: \lim_{n\to\infty}\mathbb{P}(|A_n|>M)<\epsilon \tag{72}$$

$$(d)\ \ O_p(B_n) \text{ if } A_n/B_n = O_p(1) \tag{73}$$

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:

$$(a)\ \ V(P,Q) = \sup_A |P(A)-Q(A)| \quad \textbf{total variation} \text{ distance} \tag{74}$$

$$(b)\ \ K(P,Q) = \int p\log(p/q) \quad \textbf{Kullback-Leibler} \text{ divergence} \tag{75}$$

$$(c)\ \ d_2(P,Q) = \int (p-q)^2 \quad \mathbf{L_2} \text{ distance} \tag{76}$$

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: $\text{Bias}^2(\hat\theta_n) + \text{Var}(\hat\theta_n) \to 0$.

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

The **Fisher information** is defined as

$$I_n(\theta) = -n\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log p(X_1;\theta)\right] = nI_1(\theta) \tag{77}$$

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

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

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 $X^n \overset{\text{iid}}{\sim} p(x|\theta)$:

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

Eg. for $\hat\theta_n = \overline{X}_n$: $\theta = \mu$, $v^2 = \sigma^2 = \text{Var}(X_i) = \lim_{n\to\infty} n\text{Var}(\overline{X}_n)$.

But in general, asymptotic-Var$(\hat\theta_n)$ $v^2 \neq \lim_{n\to\infty} n\text{Var}(\hat\theta_n)$.

Note that: $\text{Var}(\hat\theta_n) = (\text{se})^2 \approx v^2/n$.

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

For most estimators $v(\theta) \le v^2$.

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 $se = \sqrt{\text{Var}(\hat\theta_n)} \approx \sqrt{\frac{1}{I_n(\theta)}}$.

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

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

$$\begin{aligned} &1.\ \text{Choose a test statistic } W = W(X_1,\dots,X_n)\\ &2.\ \text{Choose a rejection region } R\\ &3.\ \text{If } W \in R, \text{ reject } H_0 \text{ otherwise retain } H_0 \end{aligned} \tag{78}$$

The **power function** $\beta(\theta) = \mathbb{P}_\theta(W \in R)$ for a rejection region R.
Want **level-**α test ($\sup_{\theta\in\Theta_0}\beta(\theta) \le \alpha$) that maximizes $\beta(\theta \in \Theta_1)$.
A level-α test with power fn β is **uniformly most powerful** if: $\beta(\theta) \ge \beta'(\theta)\ \forall\theta \in \Theta_1\ \forall\beta' \ne \beta$.

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 $H_1 : \theta \ne \theta_0$, reject H_0 if $\left|\frac{\hat\theta_n - \theta_0}{se(\hat\theta_n)}\right| > z_{\alpha/2}$.
where $z_{\alpha/2}$ is the inverse standard-normal CDF of $1 - \frac{\alpha}{2}$.
and $\hat\theta_n$ an estimator s.t. $(\hat\theta - \theta)/\text{se} \rightsquigarrow \mathcal{N}(0,1)$ eg: $\theta = \hat\theta_{\text{mle}}$
and $se = \sqrt{\text{Var}(\hat\theta_n)}$. Can also use (for eg.) $\hat{se} = \sqrt{S_n^2/n}$.
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 $\lambda(x^n) = \frac{L(\hat\theta_0)}{L(\hat\theta)} \le c$.
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. $\mathbb{P}(\lambda(x^n) \le c) = \alpha$.
Thm: under $H_0 : \theta = \theta_0 \implies W_n = -2\log\lambda(X^n) \rightsquigarrow \chi^2_1$
$\implies$ 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 $W(x^n) > c$,
$\implies p(x^n) = \sup\limits_{\theta\in\Theta_0} \mathbb{P}_\theta(W(X^n) \ge W(x^n)) = \sup\limits_{\theta\in\Theta_0}[1 - F(W(x^n)|\theta)]$.
Thm: Under $H_0 : \theta = \theta_0$, $p(x^n) \sim \text{Unif}(0,1)$.

Permutation Test

$X^n \sim F$, $Y^m \sim G$, $H_0 : F = G$, $H_1 : F \ne G$
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. $\mathbb{P}_\theta(L(X^n) \le \theta \le U(X^n)) \ge 1 - \alpha, \quad \forall\theta \in \Theta$.
Generally, a $1-\alpha$ **confidence set** C_n is a random set $C_n \subset \Theta$ s.t. $\inf_{\theta\in\Theta}\mathbb{P}_\theta(\theta \in C_n(X^n)) \ge 1 - \alpha$.

Using Probability Inequalities

Prob inequalities give (for eg.) $\mathbb{P}(|\hat\theta_n - \theta| > \epsilon) \le g(\exp^{-f(\epsilon)}) \underset{\text{set to}}{=} \alpha$.
solving for ϵ gives $\epsilon = \tilde f(\alpha) \implies \mathbb{P}\left(|\hat\theta_n - \theta| > \tilde f(\alpha)\right) \le \alpha$
$\implies C_n = \left(\hat\theta - \tilde f(\alpha), \hat\theta + \tilde f(\alpha)\right)$.

Inverting a Test

In level-α tests $\mathbb{P}_{\theta_0}(T(x^n) \in R) \le \alpha \Rightarrow$ 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 $\frac{L(\hat\theta_0)}{L(\hat\theta)} \le c$).

Pivots

$Q(X^n, \theta)$ a **pivot** if the distribution of Q does not depend on θ.
Find a, b s.t. $\mathbb{P}_\theta(a \le Q(X^n,\theta) \le b) \ge 1 - \alpha$, $\forall\theta$.
$\implies C_n = \{\theta : a \le Q(X^n,\theta) \le b) \ge 1 - \alpha\}$.

Large Sample Confidence Intervals

For mle $\hat\theta_n$ with se $\approx 1/\sqrt{I_n(\hat\theta_n)}$, approx $1-\alpha$ confidence sets are:
For Wald: $C_n = \hat\theta_n \pm (z_{\alpha/2} \times \text{se})$
For Wald with delta method: $C_n = \tau(\hat\theta_n) \pm (z_{\alpha/2} \times \text{se}(\hat\theta) \times |\tau'(\hat\theta_n)|)$
For LRT: $C_n = \left\{\theta : -2\log\left(\frac{L(\theta)}{L(\hat\theta)}\right) \le \chi^2_{k,\alpha}\right\}$

Nonparametric Inference

The **empirical CDF** is: $\hat F_n(x) = \frac{1}{n}\sum_{i=1}^n I(X_i \le x)$
Thm (DKW): $\forall\epsilon > 0, \mathbb{P}(\sup_x |\hat F(x) - F(x)| > \epsilon) \le 2e^{-2n\epsilon^2}$
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 $R = \mathbb{E}(\mathcal{L}(p,\hat p)) = \int(\text{b}^2(x) + \text{Var}(x))dx = \frac{a}{n^{4/5}}$
for some a, where $\mathcal{L}(p,\hat p) = \int(p(x) - \hat p(x))^2 dx$, and $\text{b}^2(x) = \mathbb{E}(\hat p(x)) - p(x)$. 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, $\hat\theta_n \approx \mathcal{N}(T(F), \hat{\text{se}}^2)$, where $\hat{\text{se}}$ is estimate of $\sqrt{\text{Var}\left(T(\hat F_n)\right)}$.

Bootstrap

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

$$\begin{aligned} &1.\ \text{Draw } X_1^*,\dots,X_n^* \sim \hat F_n \text{ (via } X_i^* \sim \{X_1,\dots,X_n\} \text{ unif).}\\ &2.\ \text{Compute } T_n^* = g(X_1^*,\dots,X_n^*)\\ &3.\ \text{Do 1. and 2. } B \text{ times to get } T_{n,1}^*,\dots,T_{n,B}^*\\ &4.\ \text{Let } v_{\text{boot}} = \frac{1}{B}\sum_{b=1}^B\left(T_{n,b}^* - \frac{1}{B}\sum_{r=1}^B T_{n,r}^*\right)^2 \end{aligned} \tag{79}$$

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

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

$$\pi(\theta|X^n) = \frac{p(X^n|\theta)\pi(\theta)}{m(X^n)} = \frac{p(X^n|\theta)\pi(\theta)}{\int p(X^n|\theta)\pi(\theta)d\theta} \tag{80}$$

Prediction

For train-data $(X_i, Y_i)|_{i=1,\dots,n}$, want to predict Y given a new X, where $Y \in \{0,1\}$ (**classification**) or $Y \in \mathbb{R}$ (**regression**).
For prediction rule $h(X)$,
classification risk: $R(h) = \mathbb{P}(Y \ne h(X)) = \mathbb{E}(I(Y \ne h(X)))$
regression risk: $R(h) = \mathbb{E}\left((Y - h(X))^2\right)$
Thm 1: $R(h)$ minimized by $m(x) = \mathbb{E}(Y|X = x)$.
The **Bayes classifier** $h_B(x) = I(m(x) \ge 1/2)$

Model Selection

Consider models $\mathcal{M}_{1,\dots,k}$, $\mathcal{M}_j = \{p(y;\theta_j) : \theta_j \in \Theta_j\}$, $\hat\theta_j = \text{mle}(\mathcal{M}_j)$
AIC: choose $j^* = \arg\max|_j\ \text{AIC}(j) = 2\log L_j(\hat\theta_j) - 2\dim(\Theta_j)$
BIC: choose $j^* = \arg\max|_j\ \text{BIC}(j) = \log L_j(\hat\theta_j) - \left(\frac{\dim(\Theta_j)}{2}\log n\right)$
Cross-validation: For train-data $= Y_{1,\dots,n}$ and test-data $= Y^*_{1,\dots,n}$ choose $j^* = \arg\max|_j \hat K_j = \frac{1}{n}\sum_{i=1}^n \log p(Y_i^*;\hat\theta_j)$