Aymptotic (Large Sample) Theory

A random sequence A_n is:

$$(a)\ \ o_p(1) \text{ if } A_n \xrightarrow{p} 0 \tag{1}$$
$$(b)\ \ o_p(B_n) \text{ if } A_n/B_n \xrightarrow{p} 0 \tag{2}$$
$$(c)\ \ O_p(1) \text{ if } \forall \epsilon > 0, \exists M : \lim_{n\to\infty} \mathbb{P}(|A_n| > M) < \epsilon \tag{3}$$
$$(d)\ \ O_p(B_n) \text{ if } A_n/B_n = O_p(1) \tag{4}$$

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:

$$(a)\ \ V(P,Q) = \sup_A |P(A) - Q(A)| \quad \textbf{total variation distance} \tag{5}$$
$$(b)\ \ K(P,Q) = \int p \log(p/q) \quad \textbf{Kullback-Leibler divergence} \tag{6}$$
$$(c)\ \ d_2(P,Q) = \int (p-q)^2 \quad \mathbf{L_2}\ \text{distance} \tag{7}$$

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

$$I_n(\theta) = \mathbb{E}_\theta\left[S(\theta)^2\right] = \text{Var}_\theta\left[S(\theta)\right] = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2} l(\theta)\right] \tag{8}$$

and $I_n(\theta) = -n\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log p(X_1;\theta)\right] = nI_1(\theta)$.
The **observed information** $\hat I_n(\theta) = -\sum_i \frac{\partial^2}{\partial\theta^2}\log p(X_i;\theta)$.
Vector case: $S(\theta) = \left[\frac{\partial l(\theta)}{\partial\theta_i}\right]_{i=1,\ldots,K}$ $\quad I_{ij} = -\mathbb{E}_\theta\left[\frac{\partial^2 l(\theta)}{\partial\theta_i\partial\theta_j}\right]_{i,j=1,\ldots,K}$

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**$(\hat\theta_n)$.
E.g. for $\hat\theta_n = \overline{X}_n$: $\quad v^2 = \sigma^2 = \text{Var}(X_i) = \lim_{n\to\infty} n\text{Var}(\overline{X}_n)$.
In general, asymptotic-Var$(\hat\theta_n)$ $v^2 \neq \lim_{n\to\infty} n\text{Var}(\hat\theta_n)$.
We will use approx: $\text{Var}(\hat\theta_n) \approx v^2/n$.
For param $\tau(\theta)$, $v(\theta) = \frac{|\tau'(\theta)|^2}{I_1(\theta)}$ is the **Cramer-Rao lower bound**.
for most estimators $v^2 \geq v(\theta)$.
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:

1. Choose a test statistic $W = W(X_1, \ldots, X_n)$
2. Choose a rejection region R (9)
3. If $W \in R$, reject H_0 otherwise retain H_0

For rejection region R, the **power function** $\beta(\theta) = \mathbb{P}_\theta(X^n \in R)$.
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:
$\beta(\theta) \geq \beta'(\theta)\ \forall\theta \in \Theta_1\ \forall\beta' \neq \beta$.

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 $\hat\theta_n$ is an unbiased estimator for θ.
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 $H_0 : \theta = \theta_0 \implies W_n = -2\log\lambda(X^n) \rightsquigarrow \chi_1^2$
$\implies$ reject H_0 if $W_n > \chi^2_{1,\alpha}$.
Also: for $\theta = (\theta_1, \ldots, \theta_k)$, if H_0 fixes some of the parameters
$\implies -2\log\lambda(X^n) \rightsquigarrow \chi_\nu^2$, 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 $L = (1, \ldots, 1, 2, \ldots, 2)$.
Let $W = g(L, Z) = |$(ave of 1 labeled pts) $-$ (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.) $\mathbb{P}(|\hat\theta_n - \theta| > \epsilon) \leq g(\exp^{-f(\epsilon)}) \underset{\text{set to}}{=} \alpha$.
solving for ϵ: $\mathbb{P}\left(|\hat\theta_n - \theta| > \tilde f(\alpha)\right) \leq \alpha \Rightarrow C_n = \left(\hat\theta - \tilde f(\alpha), \hat\theta + \tilde f(\alpha)\right)$.

Inverting a Test

In level-α tests $\mathbb{P}_{\theta_0}(T(x^n) \in R) = \alpha \Rightarrow$ let $C_n = \{\theta : T(x^n) \in A(\theta)\}$.
where $A(\theta) = \{T(x^n) \notin R \text{ s.t. } \theta = \theta_0\}$ (accept region if θ is null).
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. $\mathbb{P}_\theta(a \le Q(X^n,\theta) \le b) \ge 1-\alpha$, $\forall\theta$.
$\implies C_n = \{\theta : a \le Q(X^n,\theta) \le b) \ge 1-\alpha\}$.

Random Samples

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

Notes: $\mathbb{E}(\overline{X}_n) = \mathbb{E}(X_i)$, $\mathrm{Var}(\overline{X}_n) = \mathrm{Var}(X_i)/n$, $\mathbb{E}(S_n)^2 = \mathrm{Var}(X_i)$, $X_{1,\dots,n} \sim \mathrm{Bern}(p) \implies \sum_i X_i \sim \mathrm{Bin}(n,p)$, $X_{1,\dots,n} \sim \mathrm{Exp}(\beta) \implies \sum_i X_i \sim \Gamma(n,\beta)$, $X_{1,\dots,n} \sim \mathcal{N}(0,1) \implies \sum_i X_i^2 \sim \chi_n$.

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

Convergence

$X, X_1, X_2, \dots$ random variables.

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

$$\mathbb{P}(\lim_{n\to\infty} |X_n - X| < \epsilon) = 1 \tag{10}$$

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

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| < \epsilon) = 1 \tag{11}$$

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

$$\lim_{n\to\infty} \mathbb{E}[(X_n - X)^2] = 0 \tag{12}$$

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

$$\lim_{n\to\infty} F_{X_n}(t) = F_X(t) \tag{13}$$

$\forall t$ 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

$$(a)\quad X_n \xrightarrow{p} X, Y_n \xrightarrow{p} Y \implies X_n + Y_n \xrightarrow{p} X + Y \tag{14}$$
$$(b)\quad X_n \xrightarrow{p} X, Y_n \xrightarrow{p} Y \implies X_n Y_n \xrightarrow{p} XY \tag{15}$$
$$(c)\quad X_n \xrightarrow{qm} X, Y_n \xrightarrow{qm} Y \implies X_n + Y_n \xrightarrow{qm} X + Y \tag{16}$$

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

$$(a)\quad X_n \rightsquigarrow X, Y_n \rightsquigarrow c \implies X_n + Y_n \rightsquigarrow X + c \tag{17}$$
$$(b)\quad X_n \rightsquigarrow X, Y_n \rightsquigarrow c \implies X_n Y_n \rightsquigarrow cX \tag{18}$$

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

Thm 14 (CLT): $X_1,\dots,X_n$ iid, $\mathbb{E}(X_i) = \mu$ $\mathrm{Var}(X_i) = \sigma^2$
$\implies \sqrt{n}(\overline{X}_n - \mu)/\sigma \rightsquigarrow \mathcal{N}(0,1)$
$\implies \overline{X}_n \rightsquigarrow \mathcal{N}(\mu,\sigma^2/n)$
$\implies \sqrt{n}(\overline{X}_n - \mu)/S_n \rightsquigarrow \mathcal{N}(0,1)$

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

Distributions

Discrete distributions:

$$(a)\quad \text{Bernoulli}\quad f(x|p) = p^x(1-p)^{1-x}, \quad x \in \{0,1\} \tag{19}$$
$$(b)\quad \text{Binomial}\quad f(x|n,p) = \binom{n}{x} p^x(1-p)^{n-x}, \quad x \in \{0,1,\dots,n\} \tag{20}$$
$$(c)\quad \text{Poisson}\quad f(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x \in \{0,1,2,\dots\} \tag{21}$$

Continuous distributions:

$$(a)\quad \text{Uniform}\quad f(x|a,b) = \frac{1}{b-a}, \quad x \in [a,b] \tag{22}$$
$$(b)\quad \text{Normal}\quad f(x|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}, \quad x \in \mathbb{R} \tag{23}$$
$$(c)\quad \text{Gamma}\quad f(x|\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, \quad x \in \mathbb{R}_+, \alpha\ \beta > 0 \tag{24}$$

Expected Values

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

$$\mathbb{E}(g(X)) = \int g(x)dF(x) = \int g(x)dP(x) \tag{25}$$

Related properties and definitions:

$$(a)\quad \mu = \mathbb{E}(X) \tag{26}$$
$$(b)\quad \mathbb{E}(\sum_i c_i g_i(X_i)) = \sum_i c_i \mathbb{E}(g_i(X_i)) \tag{27}$$
$$(c)\quad \mathbb{E}\left(\prod_i X_i\right) = \prod_i \mathbb{E}(X_i), \quad X_1,\dots,X_n \text{ indep't} \tag{28}$$
$$(d)\quad Var(X) = \sigma^2 = \mathbb{E}((X-\mu)^2) \quad \text{is the } \textbf{variance} \text{ of X} \tag{29}$$
$$(e)\quad Var(X) = \mathbb{E}(X^2) - \mu^2 \tag{30}$$
$$(f)\quad Var\left(\sum_i a_i X_i\right) = \sum_i a_i^2 Var(X_i), \quad X_1,\dots,X_n \text{ indep't} \tag{31}$$
$$(g)\quad Cov(X,Y) = \mathbb{E}((X-\mu_X)(Y-\mu_Y)) \text{ is the } \textbf{covariance} \tag{32}$$
$$(h)\quad Cov(X,Y) = \mathbb{E}(XY) - \mu_x\mu_Y \tag{33}$$
$$(i)\quad \rho(X,Y) = Cov(X,Y)/\sigma_x\sigma_y, \quad -1 \le \rho(X,Y) \le 1 \tag{34}$$

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

$$\mathbb{E}(Y|X=x) = \int y f(y|x)dy \tag{35}$$
$$\text{and}\quad f(y|x) = f_{X,Y}(x,y)/f_X(x) \tag{36}$$

The *Law of Total/Iterated Expectation* is

$$\mathbb{E}(Y) = \mathbb{E}[\mathbb{E}(Y|X)] \tag{37}$$

The *Law of Total Variance* is

$$Var(Y) = Var[\mathbb{E}(Y|X)] + \mathbb{E}[Var(Y|X)] \tag{38}$$

The *Law of Total Covariance* is

$$Cov(X,Y) = \mathbb{E}(Cov(X,Y|Z)) + Cov(\mathbb{E}(X|Z),\mathbb{E}(Y|Z)) \tag{39}$$