Aymptotic (Large Sample) Theory

o and O notation.
Distances between distributions.
Consistency of MLE.
Score and Fisher Information.
Efficiency and Asymptotic Normality.
Relative efficiency.
Robustness.

Probability Inequalities

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

$$\mathbb{P}(|X| > \epsilon) \le \frac{2}{\epsilon} e^{-\epsilon^2/2} \tag{1}$$

Additionally:

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

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

$$\mathbb{P}(X > t) \le \frac{\mathbb{E}(X)}{t} \tag{3}$$

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

$$\mathbb{P}(|X-\mu| \ge t) \le \frac{\sigma^2}{t^2} \tag{4}$$

$$\mathbb{P}(|(X-\mu)/\sigma| \ge t) \le \frac{1}{t^2} \tag{5}$$

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

$$\mathbb{E}(e^{tX}) \le e^{t^2(b-a)^2/8} \tag{6}$$

Lemma 5: Let X be any random variable. Then

$$\mathbb{P}(X > \epsilon) \le \inf_{t \ge 0} e^{-t\epsilon}\mathbb{E}(e^{tX}) \tag{7}$$

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

$$\mathbb{P}(|\overline{X} - \mu| \ge \epsilon) \le 2e^{-2n\epsilon^2/(b-a)^2} \tag{8}$$

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

$$\mathbb{P}\left(g(X_1,\dots,X_n) - \mathbb{E}(g(X_1,\dots,X_n)) \ge \epsilon\right) \le e^{-2\epsilon^2/\sum_i c_i^2} \tag{9}$$

where $g_i^* = g$ with x_i replaced by x_i'.
Thm 12 (Cauchy-Schwartz inequality):
Thm 13 (Jensen's inequality):
Ex 15 (Kullback Leibler distance):
Thm 18:
O_p **and** o_p: $X_n = o_p(1)$ if $\forall\ \epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(|X_n| > \epsilon) = 0$.
$X_n = O_p(1)$ if $\forall\ \epsilon > 0$, $\exists\ C > 0$ s.t. $\lim_{n\to\infty} \mathbb{P}(|X_n| > C) \le \epsilon$.
$X_n = o_p(a_n)$ if $X_n/a_n = o_p(1)$ and $X_n = O_p(a_n)$ if $X_n/a_n = O_p(1)$.

Shattering

Note: remember uniform bounds and union bound.
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 $S(\mathcal{A}, F) = |\{G \subset F \text{ picked out by } \mathcal{A}\}| \le 2^n$.
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** $s_n(\mathcal{A}) = \sup_{F\in\mathcal{F}_n} s(\mathcal{A}, F) \le 2^n$.
The **VC dimension** $d(\mathcal{A})$ = the largest n s.t. $s_n(\mathcal{A}) = 2^n$.
Thm 5: $\forall \epsilon > 0$, $\mathbb{P}(\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| > \epsilon) \le 8 s_n(\mathcal{A}) e^{-n\epsilon^2/32}$

Random Samples

For $X_1, \dots, X_n \sim F$ a **statistic** is any $T = g(X_1, \dots, X_n)$.
E.g. $\overline{X}_n$, $S_n = \sum_i (X_i - \overline{X}_n)^2/(n-1)$, $\left(X_{(1)}, \dots, X_{(n)}\right)$
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, $\mathbb{E}(X_i) = \mu$ $\implies \overline{X}_n \xrightarrow{\text{qm}} \mu$.
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 $Y_n \approx \mathcal{N}(\mu, \sigma^2/n) \implies g(Y_n) \approx \mathcal{N}(g(\mu), g'(\mu)^2\sigma^2/n)$
Thm 18b (2nd order delta method):?? Should I include this?

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.

$$p(x^n;\theta) = h(x^n) \times g(t;\theta) \tag{19}$$

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:

$$\frac{p(y^n;\theta)}{p(x^n;\theta)} \text{ constant in } \theta \iff T(y^n) = T(x^n) \tag{20}$$

Parametric Point Estimation

make sure i've defined: $\mathbb{E}_\theta(\hat{\theta})$, bias, sampling distro, standard error, $\hat{\theta}_n$ consistent.

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

$$\hat{\theta} = \arg\max_\theta L(\theta) = \arg\max_\theta l(\theta) \tag{21}$$

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

$$\hat{\theta} = \mathbb{E}(\theta|x^n) = \int \theta\pi(\theta|x^n)d\pi \tag{22}$$

Mean Squared Error (MSE): The MSE is

$$\text{MSE} = \mathbb{E}(\hat{\theta}-\theta)^2 = \int (\hat{\theta}-\theta)^2 p(x^n;\theta)dx^n = \text{bias}(\hat{\theta})^2 + Var(\hat{\theta}) \tag{23}$$

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 $\hat{\theta}$, 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

$$R(\theta,\hat{\theta}) = \mathbb{E}[L(\theta,\hat{\theta})] = \int L(\theta,\hat{\theta})p(x^n;\theta)dx^n \tag{24}$$

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

$$\overline{R}(\hat{\theta}) = \sup_{\theta\in\Theta} R(\theta,\hat{\theta}) \tag{25}$$

The **minimax estimator** is

$$\hat{\theta} = \arg\inf_{\hat{\theta}} \overline{R}(\hat{\theta}) \tag{26}$$

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

$$B_\pi(\hat{\theta}) = \int R(\theta,\hat{\theta})\pi(\theta)d\theta \tag{27}$$

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

$$r(\hat{\theta}|x^n) = \int L(\theta,\hat{\theta})\pi(\theta|x^n)d\theta \tag{28}$$

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

$$\hat{\theta} = \arg\inf_{\hat{\theta}} B_\pi(\hat{\theta}) = \arg\inf_{\hat{\theta}} r(\hat{\theta}|x^n) \tag{29}$$

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 dimension of the parameter is fixed).

Distributions

Discrete distributions:

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

$$(b)\ \text{Binomial}\ \ f(x|n,p) = \binom{n}{x}p^x(1-p)^{n-x},\ \ x \in \{0,1,\ldots,n\} \tag{31}$$

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

Continuous distributions:

$$(a)\ \text{Uniform}\ \ f(x|a,b) = \frac{1}{b-a},\quad x \in [a,b] \tag{33}$$

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

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

Expected Values

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

$$\mathbb{E}(g(X)) = \int g(x)dF(x) = \int g(x)dP(x) \tag{36}$$

Related properties and definitions:

$$(a)\ \ \mu = \mathbb{E}(X) \tag{37}$$

$$(b)\ \ \mathbb{E}(\sum_i c_i g_i(X_i)) = \sum_i c_i \mathbb{E}(g_i(X_i)) \tag{38}$$

$$(c)\ \ \mathbb{E}\left(\prod_i X_i\right) = \prod_i \mathbb{E}(X_i),\quad X_1,\ldots,X_n \text{ indep't} \tag{39}$$

$$(d)\ \ Var(X) = \sigma^2 = \mathbb{E}((X-\mu)^2)\quad \text{is the } \textbf{variance} \text{ of X} \tag{40}$$

$$(e)\ \ Var(X) = \mathbb{E}(X^2) - \mu^2 \tag{41}$$

$$(f)\ \ Var\left(\sum_i a_i X_i\right) = \sum_i a_i^2 Var(X_i),\quad X_1,\ldots,X_n \text{ indep't} \tag{42}$$

$$(g)\ \ Cov(X,Y) = \mathbb{E}((X-\mu_X)(Y-\mu_Y))\ \text{ is the } \textbf{covariance} \tag{43}$$

$$(h)\ \ Cov(X,Y) = \mathbb{E}(XY) - \mu_x\mu_Y \tag{44}$$

$$(i)\ \ \rho(X,Y) = Cov(X,Y)/\sigma_x\sigma_y,\quad -1 \le \rho(X,Y) \le 1 \tag{45}$$

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

$$\mathbb{E}(Y|X=x) = \int yf(y|x)dy \tag{46}$$

$$\text{and}\ \ f(y|x) = f_{X,Y}(x,y)/f_X(x) \tag{47}$$

The *Law of Total/Iterated Expectation* is

$$\mathbb{E}(Y) = \mathbb{E}[\mathbb{E}(Y|X)] \tag{48}$$

The *Law of Total Variance* is

$$Var(Y) = Var[\mathbb{E}(Y|X)] + \mathbb{E}[Var(Y|X)] \tag{49}$$

The *Law of Total Covariance* is

$$Cov(X,Y) = \mathbb{E}(Cov(X,Y|Z)) + Cov(\mathbb{E}(X|Z),\mathbb{E}(Y|Z)) \tag{50}$$