Preliminaries

Risk function : $R(\theta,\delta)=E_\theta(L(\theta,\delta(X)))$.
Admissibility : δ is inadmissible if there exists δ' such that $R(\theta,\delta')\le R(\theta,\delta)$ for all θ and $R(\theta',\delta')<R(\theta',\delta)$ for some θ'.
Moment Generating Function : $m_X(t)=E(e^{tX})=\int e^{tX}dF(x)$, $t\in\mathbb{R}$ and $E\left(X^k\right)=m_X^{(k)}(0)$ when derivative exists in some neighborhood of 0. Properties : 1. $m_{\mu+\sigma X}(t)=e^{\mu t}m_X(\sigma t)$; 2. $m_{X+Y}(t)=m_X(t)m_Y(t)$ if X and Y are independent.
Characteristic functions : $\phi_X(t)=E(e^{itX})=\int e^{itx}dF(x)$.

$$f_X(x)=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-itx}\phi_X(t)dt$$

$$F_X(x)-F_X(y)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{e^{-itx}-e^{-ity}}{-it}\phi_X(t)dt$$

for points of continuity of F, x and y.
[Theorem] Let X and Y be random k-vectors. (i) If $\phi_X(t)=\phi_Y(t)$ for all $t\in\mathbb{R}^k$, then $F_X=F_Y$. (ii) If $m_X(t)=m_Y(t)<\infty$ for all t in a neighbourhood of 0, then $F_X=F_Y$.
Tail behavior : For a scalar random variable $X\sim F$, we say X has an exponential tail of algebraic tail if

$$\lim_{a\to\infty}\frac{-\log(1-F(a))}{Ca^r}=1,\quad \text{for some } C>0, r>0$$

$$\lim_{a\to\infty}\frac{-\log(1-F(a))}{m\log a}=1,\quad \text{for some } m>0$$

Exponential : $F(a)=1-e^{-\lambda a}\to c=\lambda, r=1$; Gaussian : $F(a)=\ldots\to c=2, r=2$.
Some integration properties 1. If $f=0$ a.e., then $\int fd\mu=0$; 2. If $f\ge 0$ and $\int fd\mu=0$, then $f=0$ a.e. 3. If $f=g$ a.e., then $\int fd\mu=\int gd\mu$ whenever either one of the integrals exists ; 4. If $\int 1_{(c,x)}fd\mu=0$ for all $x>c$, then $f(x)=0$ for a.e. $x>c$. The constant c here can be $-\infty$; 5. $f_+(x)=f_-(x)$ if and only if $f_+(x)=f_-(x)=0$ a.e.

Order statistics

$$f_{X_{(j)}}(x)=\frac{n!}{(j-1)!(n-j)!}f_X(x)\left[F_X(x)\right]^{j-1}\left[1-F_X(x)\right]^{n-j},$$

$$F_{X_{(j)}}(x)=P(Y\ge j)=\sum_{k=j}^{n}\binom{n}{k}\left[F_X(x)\right]^k\left[1-F_X(x)\right]^{n-k},$$

$$f_{X_{(i)},X_{(j)}}(u,v)=\frac{n!}{(i-1)!(j-1-i)!(n-j)!}f_X(u)f_X(v)\left[F_X(u)\right]^{i-1}\times\left[F_X(v)-F_X(u)\right]^{j-1-i}\left[1-F_X(v)\right]^{n-j}$$

for $-\infty<u<v<\infty$.

$$f_{X_{(1)},\ldots,X_{(n)}}(x_1,\ldots,x_n)=\begin{cases}n!f_X(x_1)\cdots f_X(x_n) & -\infty<x_1<\cdots<x_n<\infty\\ 0 & \text{otherwise.}\end{cases}$$

Exponential Families

Let μ be a measure on $\mathbb{R}^n$, let $h:\mathbb{R}^n\to\mathbb{R}$ be a nonnegative function, and let $T_1,\ldots,T_s$ be measurable functions from $\mathbb{R}^n$ to $\mathbb{R}$. For $\eta\in\mathbb{R}^s$, define

$$A(\eta)=\log\int\exp\left[\sum_{i=1}^{s}\eta_iT_i(x)\right]h(x)d\mu(x)$$

$$p_\eta(x)=\exp\left[\sum_{i=1}^{s}\eta_iT_i(x)-A(\eta)\right]h(x),\quad x\in\mathbb{R}^n,$$

where $\int p_\eta d\mu=1$. The set $\Xi=\{\eta:A(\eta)<\infty\}$ is called the **natural parameter space**, which is convex. The family of densities $\{p_\eta:\eta\in\Xi\}$ is called an s-parameter exponential family in **canonical form.** Let η be a function from some space Ω into Ξ and define

$$p_\theta(x)=\exp\left[\sum_{i=1}^{s}\eta_i(\theta)T_i(x)-B(\theta)\right]h(x)$$

for $\theta\in\Omega, x\in\mathbb{R}^n$, where $B(\theta)=A(\eta(\theta))$. Families $\{p_\theta:\theta\in\Omega\}$ of this form are called s-parameter exponential families.
⋆ An canonical exponential family is minimal if neither T's nor the η's satisfy a linear constraint. If Ξ contains s-dimensional rectangle then it is full rank, otherwise it is curved such that η's are related non-linearly.
⋆ In an exponential family of full rank, T is complete and minimal. If η's satisfy linear constraint T is sufficient but not minimal. If curved, T will be minimal but not complete.
[Theorem] Let Ξ_f be the set of values for $\eta\in\mathbb{R}^s$ where

$$\int|f(x)|\exp\left[\sum_{i=1}^{s}\eta_iT_i(x)\right]h(x)d\mu(x)<\infty.$$

Then the function

$$g(\eta)=\int f(x)\exp\left[\sum_{i=1}^{s}\eta_iT_i(x)\right]h(x)d\mu(x)$$

is continuous and has continuous partial derivatives of all orders for $\eta\in\Xi_f^\circ$ (the interior of Ξ_f). Furthermore, these derivatives can be computed by differentiation under the integral sign. So $\frac{\partial A(\eta)}{\partial\eta_j}=\int T_j(x)p_\eta(x)d\mu(x)=E_\eta T_j(X)$ $\frac{\partial^2A(\eta)}{\partial\eta_i\partial\eta_j}=\mathrm{Cov}_\eta\left(T_i(x),T_j(x)\right)$.
· If X has density of canonical exponential form, then $M_X(t)=\exp\{A(\eta+t)-A(\eta)\}$.

$\frac{1}{\Gamma(a)b^a}x^{a-1}e^{-x/b}$	$\Gamma(a,b)$	$0<x<\infty$
$\frac{1}{\Gamma(f/2)2^{f/2}}x^{f/2-1}e^{-x/2}$	χ_f^2	$0<x<\infty$
$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$	$B(a,b)$	$0<x<1$
$p^x(1-p)^{n-x}$	$b(p)$	$x=0,1$
$C_n^xp^x(1-p)^{n-x}$	$b(p,n)$	$x=0,1,\ldots,n$
$\frac{1}{x!}\lambda^xe^{-\lambda}$	$P(\lambda)$	$x=0,1,\ldots$
$C_{m+x-1}^{m-1}p^mq^x$	$Nb(p,m)$	$x=0,1,\ldots$

Sufficient Statistics

Suppose X has distribution from a family $\mathcal{P}=\{P_\theta:\theta\in\Omega\}$. Then $T=T(X)$ is a sufficient statistic for $\mathcal{P}$ (or for X, or for θ) if for every t and θ, the conditional distribution of X under P_θ given $T=t$ does not depend on θ.
[Theorem] (equal risk) If $X\sim P_\theta\in\mathcal{P}$ and $T=T(X)$ is a sufficient statistic for $\mathcal{P}$, then for any decision procedure δ, there exists a (possibly randomized) decision procedure of equal risk that depends on X only through $T=T(X)$ only.
⋆ A family of distributions $\mathcal{P}=\{P_\theta:\theta\in\Omega\}$ is dominated if there exists a measure μ with P_θ absolutely continuous with respect to μ, for all $\theta\in\Omega$.
[Theorem] (**Factorization Theorem**). Let $\mathcal{P}=\{P_\theta:\theta\in\Omega\}$ be a family of distributions dominated by μ. A necessary and sufficient condition for a statistic T to be sufficient is that there exist functions $g_\theta\ge 0$ and $h\ge 0$ such that the densities p_θ for the family satisfy $p_\theta(x)=g_\theta(T(x))h(x)$, for a.e. x under μ.
⋆ If T is sufficient and $T=H(U)$, then U is also sufficient. If H is 1 :1 then T and U are equivalent, otherwise T provides greater data reduction. A statistic T is minimal sufficient if T is sufficient, and for every sufficient statistic $\tilde{T}$ there exists a function f such that $T=f(\tilde{T})$ (a.e. $\mathcal{P}$).
[Theorem] Suppose $\mathcal{P}=\{P_\theta:\theta\in\Omega\}$ is a dominated family with densities $p_\theta(x)=g_\theta(T(x))h(x)$. If $p_\theta(x)\propto_\theta p_\theta(y)$ implies $T(x)=T(y)$, then T is minimal sufficient.
Remark : If $p(x;\theta)=C_{x,y}p(y;\theta)$, x and y must be supported by the same θ (support of $X:\{x\in\mathcal{X}:p(x;\theta)>0\}$). Otherwise, the 'constant' $C_{x,y}$ will be θ-dependent.
⋆ A statistic T is complete for a family $\mathcal{P}=\{P_\theta:\theta\in\Omega\}$ if $E_\theta f(T)=c$, for all θ, implies $f(T)=c$ (a.e. $\mathcal{P}$).
[Theorem] (TPE 1.6.12) Let $\mathcal{P}$ be a finite family with densities $p_i, i=0,1,\ldots,k$, all having the same support. Then, the statistic $T(X)=\left(\frac{p_1(X)}{p_0(X)},\frac{p_2(X)}{p_0(X)},\ldots,\frac{p_k(X)}{p_0(X)}\right)$ is minimal sufficient.
[Lemma] (TPE 1.6.14) If $\mathcal{P}$ is a family of distributions with common support and $\mathcal{P}_0\subset\mathcal{P}$, and if T is minimal sufficient for $\mathcal{P}_0$ and sufficient for $\mathcal{P}$, it is minimal sufficient for $\mathcal{P}$.
[Theorem] (**Bahadur**) If T is complete and sufficient, then T is minimal sufficient.
⋆ A statistic V is called ancillary if its distribution does not depend on θ. So, V by itself provides no information about θ.
[Theorem] (**Basu**) If T is complete and sufficient for $\mathcal{P}=\{P_\theta:\theta\in\Omega\}$, and if V is ancillary, then T and V are independent under P_θ for any $\theta\in\Omega$.
[Theorem] (**Jensen's Inequality**). If C is an open interval, f is a convex function on C, $P(X\in C)=1$, and EX is finite, then $f(EX)\le Ef(X)$. If f is strictly convex, the inequality is strict unless $X=E(X)$ (a.e. P_θ).
[Theorem] (**Rao-Blackwell**). Let T be a sufficient statistic for $\mathcal{P}=\{P_\theta:\theta\in\Omega\}$, let δ be an estimator of $g(\theta)$, and define $\eta(T)=E[\delta(X)\mid T]$. If $\theta\in\Omega, R(\theta,\delta)<\infty$, and $L(\theta,\cdot)$ is convex, then $R(\theta,\eta)\le R(\theta,\delta)$. Furthermore, if $L(\theta,\cdot)$ is strictly convex, the inequality will be strict unless $\delta(X)=\eta(T)$ (a.e. P_θ).
[Theorem] (TPE 1.7.10) If L is strictly convex and δ is an admissible estimator of $g(\theta)$, and if δ' is another estimator with the same risk function, that is, satisfying $R(\theta,\delta)=R(\theta,\delta')$ for all θ, then $\delta'=\delta$ with probability 1.

Unbiased Estimation

⋆ An estimator δ is called unbiased for $g(\theta)$ if $E_\theta\delta(X)=g(\theta)$, $\forall\theta\in\Omega$. If an unbiased estimator exists, g is called U-estimable.
⋆ An unbiased estimator δ satisfying $R(\theta,\delta)\le R(\theta,\delta')$ for all $\theta\in\Omega$ and any other unbiased estimator δ' is called a uniformly minimum risk unbiased estimator (UMRUE).
⋆ An unbiased estimator δ is uniformly minimum variance unbiased

(UMVU) if $\mathrm{Var}_\theta(\delta) \le \mathrm{Var}_\theta(\delta^*)$, $\forall \theta \in \Omega$, for any competing unbiased estimator δ^*.

[Theorem] (TPE 2.1.7) (Characterization of UMVUEs) Let $\Delta = \{\delta : E_\theta(\delta^2) < \infty\}$. Then $\delta_0 \in \Delta$ is UMVU for $g(\theta) = E(\delta_0)$ if and only if $E(\delta_0(\theta)u) = 0$ for every $u \in \mathcal{U} = \{E(u) = 0\}$.

⋆ By above theorem, $\forall u \in \mathcal{U}, E((\delta_1 + \delta_2)u) = E(\delta_1 u) + E(\delta_2 u) = 0$. Therefore, $\delta_1 + \delta_2$ is an UMVUE for $g_1(\theta) + g_1(\theta)$.

[Theorem] (Lehmann-Scheffé Theorem) If T is a complete and sufficient statistic, and $\mathbb{E}_\theta\{h(T(X))\} = g(\theta)$, i.e. $h(T(X))$ is unbiased for $g(\theta)$, then $h(T(X))$ is (a) the only function of $T(X)$ that is unbiased for $g(\theta)$; (b) an UMRUE under **any convex loss function**; (c) the unique UMRUE (hence UMVUE), up to a $\mathcal{P}$-null set, under any strictly convex loss function.

[Lemma] (TPE 2.1.5) U is an unbiased estimator of zero if and only if $U(k) = -kU(-1)$ for $k = 0, 1, \ldots$

⋆ Let $X_1, \ldots, X_n$ be i.i.d. from $N(\mu, \sigma^2)$, $\bar{X}$ is the UMVU estimator of μ, $\bar{X}^2 - S^2/n$ is UMVU for μ^2. $V = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.

$$\begin{aligned} ES^r &= E\left[\frac{\sigma^r}{(n-1)^{r/2}} V^{r/2}\right] \\ &= \frac{\sigma^r}{(n-1)^{r/2}} \int_0^\infty \frac{x^{(r+n-3)/2} e^{-x/2}}{2^{(n-1)/2}\Gamma[(n-1)/2]} dx \\ &= \frac{\sigma^r 2^{r/2}\Gamma[(r+n-1)/2]}{(n-1)^{r/2}\Gamma[(n-1)/2]} \end{aligned}$$

From this,

$$\frac{(n-1)^{r/2}\Gamma[(n-1)/2]}{2^{r/2}\Gamma[(r+n-1)/2]} S^r$$

is an unbiased estimate of σ^r.

$$\frac{\bar{X}\sqrt{2}\Gamma[(n-1)/2]}{S\sqrt{n-1}\Gamma[(n-2)/2]}$$

is UMVU for μ/σ.

[Lemma] (TPE 2.2.7) Let the risk be expected squared error. If δ is an unbiased estimator of $g(\theta)$ and if $\delta^* = \delta + b$, where the bias b is independent of θ, then δ^* has uniformly larger risk than δ, in fact,

$$R_{\delta^*}(\theta) = R_\delta(\theta) + b^2.$$

Variance bound and information

Covariance inequality : $\mathrm{Cov}(X, Y) \le \sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}$.

⋆ Let $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ be a dominated family with densities $p_\theta, \theta \in \Omega \subset \mathbb{R}$. As a starting point, $E_{\theta+\Delta}\delta - E_\theta\delta$ gives the same value $g(\theta+\Delta) - g(\theta)$ for any unbiased δ. Here Δ must be chosen so that $\theta + \Delta \in \Omega$. Next, we write $E_{\theta+\Delta}\delta - E_\theta\delta$ as a covariance under P_θ. Assume that $p_{\theta+\Delta}(x) = 0$ whenever $p_\theta(x) = 0$ and define $L(x) = p_{\theta+\Delta}(x)/p_\theta(x)$ when $p_\theta(x) > 0$, and $L(x) = 1$, otherwise. Then we define $\psi(X) = L(X) - 1$, then $E_\theta\psi = 0$ and $E_{\theta+\Delta}\delta - E_\theta\delta = E_\theta L\delta - E_\theta\delta = E_\theta\psi\delta = \mathrm{Cov}_\theta(\delta, \psi)$. Thus $\mathrm{Cov}_\theta(\delta, \psi) = g(\theta+\Delta) - g(\theta)$ for any unbiased estimator δ. With this choice for ψ, the covariance inequality gives

$$\mathrm{Var}_\theta(\delta) \ge \frac{[g(\theta+\Delta) - g(\theta)]^2}{\mathrm{Var}_\theta(\psi)} = \frac{[g(\theta+\Delta) - g(\theta)]^2}{E_\theta\left(\frac{p_{\theta+\Delta}(X)}{p_\theta(X)} - 1\right)^2},$$

called the Hammersley-Chapman-Robbins inequality.
Under suitable conditions to use DCT, we can show that

$$\lim_{\Delta \to 0} \frac{\left\{\frac{g(\theta+\Delta) - g(\theta)}{\Delta}\right\}^2}{E_\theta\left(\frac{\{p_{\theta+\Delta}(x) - p_\theta(x)\}/\Delta}{p_\theta(x)}\right)^2} = \frac{(g'(\theta)^2)}{E_\theta\left(\frac{\partial p_\theta(x)/\partial\theta}{p_\theta(x)}\right)^2}.$$

The denominator here is known as Fisher Information, denoted as $I(\theta)$ and is given by

$$I(\theta) = E_\theta\left(\frac{\partial \log p_\theta(x)}{\partial\theta}\right)^2.$$

Under suitable conditions to interchange integration and differentiation,

$$\begin{aligned} 0 = \frac{\partial}{\partial\theta} 1 &= \frac{\partial}{\partial\theta} \int p_\theta(x) d\mu(x) = \int \frac{\partial}{\partial\theta} p_\theta(x) d\mu(x) \\ &= \int \frac{\partial \log p_\theta(x)}{\partial\theta} p_\theta(x) d\mu(x) = E_\theta \frac{\partial \log p_\theta(X)}{\partial\theta}, \end{aligned}$$

and so

$$I(\theta) = \mathrm{Var}_\theta\left(\frac{\partial \log p_\theta(X)}{\partial\theta}\right).$$

If we can pass two partial derivatives with respect to θ, then

$$\int \frac{\partial^2 p_\theta(x)}{\partial\theta^2} d\mu(x) = E_\theta\left[\frac{\partial^2 p_\theta(X)/\partial\theta^2}{p_\theta(X)}\right] = 0.$$

$$\frac{\partial^2 \log p_\theta(X)}{\partial\theta^2} = \frac{\partial^2 p_\theta(X)/\partial\theta^2}{p_\theta(X)} - \left(\frac{\partial \log p_\theta(X)}{\partial\theta}\right)^2$$

$$I(\theta) = -E_\theta \frac{\partial^2 \log p_\theta(X)}{\partial\theta^2}.$$

[Theorem] (**Cramér-Rao information, bound**) Let $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ be a dominated family with Ω an open set in $\mathbb{R}$ and densities p_θ differentiable with respect to θ. If $E_\theta\psi = 0$, $E_\theta\delta^2 < \infty$, and $g'(\theta) = E_\theta\delta\psi$ for all $\theta \in \Omega$, then

$$\mathrm{Var}_\theta(\delta) \ge \frac{[g'(\theta)]^2}{I(\theta)}, \quad \theta \in \Omega.$$

⋆ Suppose $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ is a dominated family with densities p_θ and Fisher information I. If h is a one-to-one function from Ξ to Ω, then the family $\mathcal{P}$ can be reparameterized as $\tilde{\mathcal{P}} = \{Q_\xi : \xi \in \Xi\}$ with the identification $Q_\xi = P_{h(\xi)}$. Then Q_ξ has density $q_\xi = p_{h(\xi)}$. Letting $\theta = h(\xi)$, by the chain rule, Fisher information $\tilde{I}$ for the reparameterized family $\tilde{\mathcal{P}}$ is given by

$$\begin{aligned} \tilde{I}(\xi) &= \tilde{E}_\xi\left(\frac{\partial \log q_\xi(X)}{\partial\xi}\right)^2 = \tilde{E}_\xi\left(\frac{\partial \log p_{h(\xi)}(X)}{\partial\xi}\right)^2 \\ &= [h'(\xi)]^2 E_\theta\left(\frac{\partial \log p_\theta(X)}{\partial\theta}\right)^2 = [h'(\xi)]^2 I(\theta). \end{aligned}$$

⋆ Suppose we have $X_1, \ldots, X_n \overset{i.i.d}{\sim} p_\theta$, $I_{\mathbf{x}} = I_{X_1}(\theta) + \cdots + I_{X_n}(\theta) = nI_{X_1}(\theta)$. Then $\mathrm{Var}_\theta(\delta) \ge \frac{g'(\theta)}{nI(\theta)}$. ⋆ When the parameter θ takes values in $\mathbb{R}^s$, Fisher information will be a matrix, defined in regular cases by

$$\begin{aligned} I(\theta)_{i,j} &= E_\theta\left[\frac{\partial \log p_\theta(X)}{\partial\theta_i}\frac{\partial \log p_\theta(X)}{\partial\theta_j}\right] \\ &= \mathrm{Cov}_\theta\left(\frac{\partial \log p_\theta(X)}{\partial\theta_i}, \frac{\partial \log p_\theta(X)}{\partial\theta_j}\right) \\ &= -E_\theta\left[\frac{\partial^2 \log p_\theta(X)}{\partial\theta_i \partial\theta_j}\right], \\ I(\theta) &= E_\theta\left(\{\nabla_\theta \log p_\theta(x)\}\{\nabla_\theta \log p_\theta(x)\}^\top\right) \\ &= \mathrm{Cov}(\nabla_\theta \log p_\theta(x)) = -E_\theta \nabla^2_\theta \log p_\theta(x). \end{aligned}$$

The lower bound for the variance of an unbiased estimator δ of $g(\theta)$, where $g : \Omega \to \mathbb{R}$, is $\mathrm{Var}_\theta(\delta) \ge \nabla g(\theta)' I^{-1}(\theta) \nabla g(\theta)$.

Bayes Estimator

[Theorem] (TPE 4.1.1) Let Θ have distribution Λ, and given $\Theta = \theta$, let X have distribution P_θ. Suppose, in addition, the following assumptions hold for the problem of estimating $g(\Theta)$ with non-negative loss function $L(\theta, d)$. (a) There exists an estimator δ_0 with finite risk. (b) For almost all x, there exists a value $\delta_\Lambda(x)$ minimizing $E\{L[\Theta, \delta(x)] \mid X = x\}$. Then, $\delta_\Lambda(X)$ is a Bayes estimator.

[Corollary] (TPE 4.1.2) Suppose the assumptions of Theorem 4.1.1 hold. (a) If $L(\theta, d) = w(\theta)[d - g(\theta)]^2$ then

$$\delta_\Lambda(x) = \frac{\int w(\theta) g(\theta) d\Lambda(\theta \mid x)}{\int w(\theta) d\Lambda(\theta \mid x)} = \frac{E[w(\Theta)g(\Theta) \mid x]}{E[w(\Theta) \mid x]}$$

(b) If $L(\theta, d) = |d - g(\theta)|$, then $\delta_\Lambda(x)$ is anymedian of the conditional distribution of Θ given x. (c) If

$$L(\theta, d) = \begin{cases} 0 \text{ when } |d - \theta| \le c \\ 1 \text{ when } |d - \theta| > c, \end{cases}$$

then $\delta_\Lambda(x)$ is the midpoint of the interval I of length $2c$ which maximizes $P[\Theta \in I \mid x]$.

[Theorem] (TPE 5.2.4) Any unique Bayes estimator is admissible.

[Corollary] (TPE 4.1.4) (**Uniqueness**) If the loss function $L(\theta, d)$ is squared error, or more generally, if it is strictly convex in d, a Bayes solution δ_Λ is unique (a.e. $\mathcal{P}$), where $\mathcal{P}$ is the class of distributions P_θ, provided (a) the average risk of δ_Λ with respect to Λ is finite, and (b) if Q is the marginal distribution of X given by $Q(A) = \int P_\theta(X \in A) d\Lambda(\theta)$, then a.e. Q implies a.e. $\mathcal{P}$.

[Theorem] (TPE 4.2.3) Let Θ have a distribution Λ, and let P_θ denote the conditional distribution of X given θ. Consider the estimation of $g(\theta)$ when the loss function is squared error. Then, no unbiased estimator $\delta(X)$ can be a Bayes solution unless $E[\delta(X) - g(\Theta)]^2 = 0$, where the expectation is taken with respect to variation in both X and Θ.

[Theorem] (TPE 4.3.2) If X has density of canonical exponential family, and η has prior density $\pi(\eta)$, then for $j = 1, \ldots, n$,

$$E\left(\sum_{i=1}^s \eta_i \frac{\partial T_i(\mathrm{x})}{\partial x_j} \mid \mathrm{x}\right) = \frac{\partial}{\partial x_j} \log m(\mathrm{x}) - \frac{\partial}{\partial x_j} \log h(\mathrm{x})$$

where $m(\mathrm{x}) = \int p_\eta(\mathrm{x})\pi(\eta) d\eta$ is the marginal distribution of X. Alternatively, the posterior expectation can be expressed in matrix form as

$$E(\mathcal{T}\eta) = \nabla \log m(\mathrm{x}) - \nabla \log h(\mathrm{x}),$$

where $\mathcal{T} = \{\partial T_i/\partial x_j\}$.
[Corollary] (TPE 4.3.3) If $\mathbf{X} = (X_1, \ldots, X_p)$ has the density

$$p_\eta(\mathbf{x}) = e^{\sum_{i=1}^p \eta_i x_i - A(\eta)} h(\mathbf{x})$$

and η has prior density $\pi(\eta)$, the Bayes estimator of η under the loss $L(\eta, \delta) = \Sigma(\eta_i - \delta_i)^2$ is given by

$$E(\eta_i \mid \mathbf{x}) = \frac{\partial}{\partial x_i} \log m(\mathbf{x}) - \frac{\partial}{\partial x_i} \log h(\mathbf{x}).$$

[Theorem] (TPE 4.3.5) Under the assumptions of Corollary 3.3, the risk of the Bayes estimator in Corollary 4.3.3, under the sum of squared error loss, is

$$\begin{aligned} R[\eta, E(\eta \mid \mathbf{X})] =& R[\eta, -\nabla \log h(\mathbf{X})] \\ &+ \sum_{i=1}^p E\left\{2\frac{\partial^2}{\partial X_i^2} \log m(\mathbf{X}) + \left(\frac{\partial}{\partial X_i} \log m(\mathbf{X})\right)^2\right\} \end{aligned}$$

In addition, $-\nabla \log h(\mathbf{X})$ is an unbiased estimator of η with risk

$$\begin{aligned} R[\eta, -\nabla \log h(\mathbf{X})] &= E_\eta \sum_{i=1}^p \left[\eta_i + \frac{\partial}{\partial X_i} \log h(\mathbf{X})\right]^2 \\ &= E_\eta |\eta + \nabla \log h(\mathbf{X})|^2 \end{aligned}$$

⋆ For the density $p_\eta(x) = e^{\eta x - A(\eta)} h(x), \quad -\infty < x < \infty$, the conjugate prior family is $\pi(\eta \mid k, \mu) = c(k, \mu) e^{k\eta\mu - kA(\eta)}$, where μ can be thought of as a prior mean and k is proportional to a prior variance.
⋆ (Binomial-Beta) Suppose $X \sim \text{Binomial}(n, \theta)$ given $\Theta = \theta$ and that Θ has a prior distribution Beta (α, β), with hyperparameters α and β.

$$\pi(\theta \mid X) \sim \text{Beta}(x + \alpha, n - x + \beta),$$

meaning that the posterior mean of $\Theta \mid X$ is $(x+\alpha)/(n+\alpha+\beta)$.
⋆ (Normal Mean Estimation) Let $X_1, \cdots, X_n \overset{iid}{\sim} N(\Theta, \sigma^2)$, with σ^2 known. Let $\Theta \sim N(\mu, b^2)$ where μ and b^2 are two fixed prior hyperparameters. Then the posterior distribution of $\Theta \mid X$ is

$$\pi(\theta \mid X) \propto \exp\left\{-\frac{1}{2}\left(\frac{n}{\sigma^2} + \frac{1}{b^2}\right)\theta^2 + \left(\frac{n\bar{X}}{\sigma^2} + \frac{\mu}{b^2}\right)\theta\right\}.$$

The posterior distribution of Θ given X is $N(\tilde{\mu}, \tilde{\sigma}^2)$ where

$$\tilde{\mu} = \frac{n\bar{X}/\sigma^2 + \mu/b^2}{n/\sigma^2 + 1/b^2} \quad \text{and} \quad \tilde{\sigma}^2 = \frac{1}{n/\sigma^2 + 1/b^2}.$$

Minimax Estimator

⋆ An estimator δ^M of θ, which minimizes the maximum risk, that is, which satisfies

$$\inf_\delta \sup_\theta R(\theta, \delta) = \sup_\theta R\left(\theta, \delta^M\right),$$

is called a minimax estimator.
⋆ Denote the average risk (Bayes risk) of the Bayes solution δ_Λ by

$$r_\Lambda = r(\Lambda, \delta_\Lambda) = \int R(\theta, \delta_\Lambda)\, d\Lambda(\theta).$$

A prior distribution Λ is least favorable if $r_\Lambda \geq r_{\Lambda'}$ for all prior distributions Λ'.

[Theorem] (TPE 5.1.4) Suppose that Λ is a distribution on Θ such that

$$r(\Lambda, \delta_\Lambda) = \int R(\theta, \delta_\Lambda)\, d\Lambda(\theta) = \sup_\theta R(\theta, \delta_\Lambda).$$

Then : (i) δ_Λ is minimax. (should be Bayes estimator in advance) (ii) If δ_Λ is the unique Bayes solution with respect to Λ, it is the unique minimax procedure. (iii) Λ is least favorable.
[Corollary] (TPE 5.1.5) If a Bayes solution δ_Λ has constant risk, then it is minimax.
[Corollary] (TPE 5.1.6) Let ω_Λ be the set of parameter points at which the risk function of δ_Λ takes on its maximum, that is,

$$\omega_\Lambda = \left\{\theta : R(\theta, \delta_\Lambda) = \sup_{\theta'} R(\theta', \delta_\Lambda)\right\}.$$

Then, δ_Λ is minimax if and only if $\Lambda(\omega_\Lambda) = 1$.
[Lemma] (TPE 5.1.10) Let δ be a Bayes (respectively, UMVU, minimax, admissible) estimator of $g(\theta)$ for squared error loss. Then, $a\delta + b$ is Bayes (respectively, UMVU, minimax, admissible) for $ag(\theta) + b$.
⋆ A sequence of prior distributions $\{\Lambda_n\}$ is least favorable if for every prior distribution Λ we have

$$r_\Lambda \leq r = \lim_{n \to \infty} r_{\Lambda_n},$$

where

$$r_{\Lambda_n} = \int R(\theta, \delta_n)\, d\Lambda_n(\theta)$$

is the Bayes risk under Λ_n.
[Theorem] (TPE 5.1.12) Suppose that $\{\Lambda_n\}$ is a sequence of prior distributions with Bayes risks $\lim_{n\to\infty} r_{\Lambda_n} = r$ and that δ is an estimator for which

$$\sup_\theta R(\theta, \delta) = r.$$

Then (i) δ is minimax and (ii) the sequence $\{\Lambda_n\}$ is least favorable.
[Lemma] (TPE 5.1.13) If δ_Λ is the Bayes estimator of $g(\theta)$ with respect to Λ and if

$$r_\Lambda = E\left[\delta_\Lambda(\mathbf{X}) - g(\Theta)\right]^2$$

is its Bayes risk, then

$$r_\Lambda = \int \text{var}[g(\Theta) \mid \mathbf{x}] dP(\mathbf{x}).$$

In particular, if the posterior variance of $g(\Theta) \mid \mathbf{x}$ is independent of $\mathbf{x}$, then

$$r_\Lambda = \text{var}[g(\Theta) \mid \mathbf{x}].$$

[Lemma] (TPE 5.1.15) Suppose that δ is minimax for a submodel $\theta \in \Omega_0 < \Omega$ and

$$\sup_{\theta \in \Omega_0} R(\theta, \delta) = \sup_{\theta \in \Omega} R(\theta, \delta)$$

Then δ is minimax for the full model $\theta \in \Omega$.
[Theorem] (TPE 5.2.6) Let X be a random variable with mean θ (unbiased) and variance σ^2. Then, $aX + b$ is an inadmissible estimator of θ under squared error loss whenever (i) $a > 1$, or (ii) $a < 0$, or (iii) $a = 1$ and $b \neq 0$.
⋆ If δ is admissible with constant risk, then δ is also minimax.
[Theorem] (TPE 5.5.1) Let $X_i, i = 1, \ldots, r (r > 2)$, be independent, with distributions $N(\theta_i, 1)$ and let the estimator δ_c of θ be given by

$$\delta_c(\mathbf{x}) = \left(1 - c\frac{r-2}{|\mathbf{x}|^2}\right)\mathbf{x}, \quad |\mathbf{x}|^2 = \Sigma x_j^2.$$

Then, the risk function of δ_c, with average squared loss function, is

$$R(\theta, \delta_c) = 1 - \frac{(r-2)^2}{r} E_\theta\left[\frac{c(2-c)}{|\mathbf{X}|^2}\right].$$

Furthermore, the James-Stein estimator δ, which equals δ_c with $c = 1$, dominates all estimators δ_c with $c \neq 1$.

Hypotheses Testing

⋆ Test function/critical function : $\phi(x) \in [0, 1]$

$$\phi(x) = P\left(\delta_\phi(x, u) = \text{Reject } H_0 \mid x\right)$$

where u is a uniform random variable independent of X.
⋆ Power function of a test ϕ is $\beta(\theta) = \mathbb{E}_\theta(\phi(X)) = P_\theta(\text{Reject } H_0)$.
⋆ If $\theta_0 \in \Omega_0$, then $\beta(\theta_0) = R(\theta_0, \delta_\phi) = $ Type I error.
For $\theta_1 \in \Omega_1$, then $\beta(\theta_1) = 1 - R(\theta_1, \delta_\phi) = 1-$ Type II error.
⋆ (Neyman-Pearson Framework) Control the level of significance

$$\sup_{\theta_0 \in \Omega_0} \mathbb{E}_{\theta_0}\phi(X) = \sup_{\theta_0 \in \Omega_0} \beta(\theta_0) \leq \alpha$$

where $\sup_{\theta_0 \in \Omega_0} \beta(\theta_0)$ is called the size of the test. The optimality goal is to find a level α test that maximizes the power $\beta(\theta_1) = \mathbb{E}_{\theta_1}(\phi(X))$ for each $\theta_1 \in \Omega_1$. Such a test is called a uniformly powerful (UMP) test.
[Theorem](Neyman-Pearson Lemma)
(i) Existence. For testing $H_0 : p_0$ vs $H_1 : p_1$, there exists a test $\phi(X)$ and a constant k such that
(a) $E_{p_0}(\phi(X)) = \alpha($ size = level $)$
(b)

$$\phi(x) = \begin{cases} 1, & \text{if } \frac{p_1(x)}{p_0(x)} > k[\text{Rejection}] \\ 0, & \text{otherwise [Acceptance]} \end{cases}$$

such a test is called a likelihood ratio test.
(ii) Sufficiency : If a test satisfies (a) and (b) for some constant k, it is most powerful for testing $H_0 : p_0$ vs $H_1 : p_1$ at level α.
(iii) Necessity : If a test ϕ is MP at level α, then it satisfies (b) for some k, and it also satisfies (a) unless there exists a test if size $< \alpha$ with power 1.
[Corollary] (TSH 3.2.1) Let β denote the power of the most powerful level - α test $(0 < \alpha < 1)$ for testing P_0 against P_1. Then $\alpha < \beta$ unless $P_0 = P_1$.
⋆ The family of densities $\{p_\theta : \theta \in R\}$ has monotone likelihood ratio in $T(x)$ if (1) $\theta \neq \theta'$ implies $p_\theta \neq p_{\theta'}$ (Identifiability); (2) $\theta < \theta'$ implies $p_{\theta'}(x)/p_\theta(x)$ is a non-decreasing function of $T(x)$ (Monotonicity).
[Theorem] (TSH 3.4.1) Let θ be a real parameter, and let the random variable X have probability density $p_\theta(x)$ with monotone likelihood ratio in $T(x)$.
(i) For testing $H : \theta \leq \theta_0$ against $K : \theta > \theta_0$, there exists a UMP test, which is given by

$$\phi(x) = \begin{cases} 1 \text{ when } T(x) > C, \\ \gamma \text{ when } T(x) = C, \\ 0 \text{ when } T(x) < C, \end{cases}$$

where C and γ are determined by $E_{\theta_0}\phi(X) = \alpha$.
(ii) The power function $\beta(\theta) = E_\theta\phi(X)$ of this test is strictly increasing for all points θ for which $0 < \beta(\theta) < 1$.
(iii) For all θ', the test determined by (i) is UMP for testing $H' : \theta \leq \theta'$ against $K' : \theta > \theta'$ at level $\alpha' = \beta(\theta')$.
(iv) For any $\theta < \theta_0$ the test minimizes $\beta(\theta)$ (the probability of an

error of the first kind) among all tests satisfying (i).

⋆ Consider the case with a simple alternative :

$$H_0 : X \sim f_\theta, \theta \in \Omega_0$$
$$H_1 : X \sim g \text{ (unknown)}, \text{ [simple]}$$

We impose a prior distribution Λ on Ω_0. So we consider the new hypothesis :

$$H_\Lambda : X \sim h_\Lambda(x) = \int_{\Omega_0} f_\theta(x) d\Lambda(\theta),$$

where $h_\Lambda(x)$ is the marginal distribution of X induced by Λ. We shall test H_Λ vs H_1. Let β_Λ be the power of the MP level- α test Φ_Λ for testing H_Λ vs. $H_1(g)$. The prior Λ is a <u>least favourable distribution</u> if $\beta_\Lambda \leq \beta_{\Lambda'}$ for any prior Λ'.

[Theorem] (TSH 3.8.1) Suppose Φ_Λ is a MP level- α test for testing H_Λ against g. If ϕ_Λ is level- α for the original hypothesis H_0 (i.e $E_{\theta_0}\Phi_\Lambda(x) \leq \alpha, \forall \theta \in \Omega_0$), then

1. The test Φ_Λ is MP for the original : $H_0 : \theta \in \Omega_0$ vs $H_1 : g$;
2. The distribution Λ is least favourable.

[Corollary] (TSH 3.8.1) Suppose that Λ is a probability distribution over ω and that ω' is a subset of ω with $\Lambda(\omega') = 1$. Let ϕ_Λ be a test such that

$$\phi_\Lambda(x) = \begin{cases} 1 \text{ if } g(x) > k \int f_\theta(x) d\Lambda(\theta), \\ 0 \text{ if } g(x) < k \int f_\theta(x) d\Lambda(\theta). \end{cases}$$

Then ϕ_Λ is a most powerful level- α for testing H against g provided

$$E_{\theta'}\phi_\Lambda(X) = \sup_{\theta \in \omega} E_\theta \phi_\Lambda(X) = \alpha \quad \text{for} \quad \theta' \in \omega'.$$

Proof of N-P Lemma

For $\alpha = 0$ and $\alpha = 1$ the theorem is easily seen to be true provided the value $k = +\infty$ is admitted in (3.8) and $0 \cdot \infty$ is interpreted as 0 . Throughout the proof we shall therefore assume $0 < \alpha < 1$. (i) : Let $\alpha(c) = P_0\{p_1(X) > cp_0(X)\}$. Since the probability is computed under P_0, the inequality needs to be considered only for the set where $p_0(x) > 0$, so that $\alpha(c)$ is the probability that the random variable $p_1(X)/p_0(X)$ exceeds c. Thus $1 - \alpha(c)$ is a cumulative distribution function, and $\alpha(c)$ is nonincreasing and continuous on the right, $\alpha(c^-) - \alpha(c) = P_0\{p_1(X)/p_0(X) = c\}, \alpha(-\infty) = 1$, and $\alpha(\infty) = 0$. Given any $0 < \alpha < 1$, let c_0 be such that $\alpha(c_0) \leq \alpha \leq \alpha(c_0^-)$, and consider the test ϕ defined by

$$\phi(x) = \begin{cases} 1 & \text{when } p_1(x) > c_0 p_0(x) \\ \frac{\alpha - \alpha(c_0)}{\alpha(c_0^-) - \alpha(c_0)} & \text{when } p_1(x) = c_0 p_0(x) \\ 0 & \text{when } p_1(x) < c_0 p_0(x) \end{cases}$$

Here the middle expression is meaningful unless $\alpha(c_0) = \alpha\left(c_0^-\right)$; since then $P_0\{p_1(X) = c_0 p_0(X)\} = 0, \phi$ is defined a.e. The size of ϕ is

$$E_0\phi(X) = P_0\left\{\frac{p_1(X)}{p_0(X)} > c_0\right\} + \frac{\alpha - \alpha(c_0)}{\alpha(c_0^-) - \alpha(c_0)} P_0\left\{\frac{p_1(X)}{p_0(X)} = c_0\right\} = \alpha,$$

so that c_0 can be taken as the k of the theorem. (ii) : Suppose that ϕ is a test satisfying (3.7) and (3.8) and that ϕ^* is any other test with $E_0\phi^*(X) \leq \alpha$. Denote by S^+ and S^- the sets in the sample space where $\phi(x) - \phi^*(x) > 0$ and < 0, respectively. If x is in $S^+, \phi(x)$ must be > 0 and $p_1(x) \geq kp_0(x)$. In the same way $p_1(x) \leq kp_0(x)$ for all x in S^-, and hence

$$\int (\phi - \phi^*)(p_1 - kp_0)\, d\mu = \int_{S^+ \cup S^-} (\phi - \phi^*)(p_1 - kp_0)\, d\mu \geq 0.$$

The difference in power between ϕ and $\phi*$ therefore satisfies

$$\int (\phi - \phi^*) p_1 d\mu \geq k \int (\phi - \phi^*) p_0 d\mu \geq 0,$$

as was to be proved. (iii) : Let ϕ^* be most powerful at level α for testing p_0 against p_1, and let ϕ satisfy (3.7) and (3.8). Let S be the intersection of the set $S^+ \cup S^-$, on which ϕ and ϕ^* differ, with the set $\{x : p_1(x) \neq kp_0(x)\}$, and suppose that $\mu(S) > 0$. Since $(\phi - \phi^*)(p_1 - kp_0)$ is positive on S, it follows from Problem 2.4 that

$$\int_{S^+ \cup S^-} (\phi - \phi^*)(p_1 - kp_0)\, d\mu = \int_S (\phi - \phi^*)(p_1 - kp_0)\, d\mu > 0$$

and hence ϕ is more powerful against p_1 than ϕ^*. This is a contradiction, and therefore $\mu(S) = 0$, as was to be proved.

If ϕ^* were of size $< \alpha$ and power < 1, it would be possible to include in the rejection region additional points or portions of points and thereby to increase the power until either the power is 1 or the size is α. Thus either $E_0\phi^*(X) = \alpha$ or $E_1\phi^*(X) = 1$.

Asymptotic Theory

[Theorem] Let $X_1, X_2, \ldots$, be iid $f(x \mid \theta)$, and let $L(\theta \mid \mathbf{x})$ be the likelihood function. Let $\hat{\theta}$ denote the MLE of θ. Let $\tau(\theta)$ be a continuous function of θ. Under the regularity conditions (A1)-(A4), for every $\epsilon > 0$ and every $\theta \in \Theta$, $\lim_{n\to\infty} P_\theta(|\tau(\hat{\theta}) - \tau(\theta)| \geq \epsilon) = 0$. That is, $\tau(\hat{\theta})$ is a consistent estimator of $\tau(\theta)$.

[Theorem] Let $X_1, X_2, \ldots$, be iid $f(x \mid \theta)$, let $\hat{\theta}$ denote the MLE of θ, and let $\tau(\theta)$ be a continuous function of θ. Under the regularity conditions (A1)-(A6), $\sqrt{n}[\tau(\hat{\theta}) - \tau(\theta)] \to \mathrm{n}[0, v(\theta)]$ where $v(\theta)$ is the Cramér-Rao Lower Bound. That is, $\tau(\hat{\theta})$ is a consistent and asymptotically efficient estimator of $\tau(\theta)$.

[Regularity Conditions]

(A1) We observe $X_1, \ldots, X_n$, where $X_i \sim f(x \mid \theta)$ are iid.

(A2) The parameter is identifiable; that is, if $\theta \neq \theta'$, then $f(x \mid \theta) \neq f(x \mid \theta')$.

(A3) The densities $f(x \mid \theta)$ have common support, and $f(x \mid \theta)$ is differentiable in θ.

(A4) The parameter space Ω contains an open set ω of which the true parameter value θ_0 is an interior point.

(A5) For every $x \in \mathcal{X}$, the density $f(x \mid \theta)$ is three times differentiable with respect to θ, the third derivative is continuous in θ, and $\int f(x \mid \theta) dx$ can be differentiated three times under the integral sign.

(A6) For any $\theta_0 \in \Omega$, there exists a positive number c and a function $M(x)$ (both of which may depend on θ_0) such that $\left|\frac{\partial^3}{\partial\theta^3} \log f(x \mid \theta)\right| \leq M(x)$ for all $x \in \mathcal{X}$, $\quad \theta_0 - c < \theta < \theta_0 + c$, with $\mathrm{E}_{\theta_0}[M(X)] < \infty$.

$E^*, P^* \quad X \sim \mu \quad G^*, G_\theta - T$

$\mathbb{E}_\theta(f(T)) = \int f(T(x)) g_\theta(T(x)) h(x) d\mu(x)$

$= \iint f(t) g_\theta(t) h(x) \, dQ_t(x) \, dG^*(t)$

$= \int f(t) g_\theta(t) w(t) dG^*(t) \quad w(t) = \int h(x) \, dQ_t(x)$

Let $\tilde{Q}(B) = \int_B \frac{h(x)}{w(t)} \, dQ_t(x)$

$\mathbb{E}_\theta(f(X,T)) = \mathbb{E}^*(f(X,T) g_\theta(T) h(x))$

$= \iint f(x,t) g_\theta(t) h(x) \, dQ_t(x) \, dG^*(t)$

$= \iint f(x,t) (h(x)/w(t)) dQ_t(x) g_\theta(t) w(t) \, dG^*(t)$

$= \iint f(x,t) \, d\tilde{Q}_t(x) \, dG_\theta(t) \qquad \tilde{Q}$ independent of θ.

$\bar{X}$

$\theta_i \quad X_i$